Exploratory Data Analysis

In [2]:

```
import pandas as pd
import matplotlib.pyplot as plt
import re
import time
import warnings
warnings.filterwarnings("ignore")
import numpy as np
import seaborn as sns
from collections import Counter, defaultdict
from nltk.corpus import stopwords
from sklearn.decomposition import TruncatedSVD
from sklearn.preprocessing import normalize
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.manifold import TSNE
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix, normalized_mutual_info_score
from sklearn.metrics.classification import accuracy_score, log_loss
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import SGDClassifier
from imblearn.over_sampling import SMOTE
from collections import Counter
from scipy.sparse import hstack
from sklearn.multiclass import OneVsRestClassifier
from sklearn.svm import SVC

from sklearn.calibration import CalibratedClassifierCV
from sklearn.naive_bayes import MultinomialNB, GaussianNB
from sklearn.model_selection import train_test_split, GridSearchCV, StratifiedKFold

import math
from sklearn.ensemble import RandomForestClassifier


from mlxtend.classifier import StackingClassifier

from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
```

Reading Data

Reading gene and Variants Data

In [3]:

```
data_variants = pd.read_csv('C:\\Users\\Khushmeet Singh\\Downloads\\training_variants')
print('Number of data points : ', data_variants.shape[0])
print('Number of features : ', data_variants.shape[1])
print('Features : ', data_variants.columns.values)
data_variants.head()
```

```
Number of data points :  3321
Number of features :  4
Features :  ['ID' 'Gene' 'Variation' 'Class']
```

Out[3]:

	ID	Gene	Variation	Class
0	0	FAM58A	Truncating Mutations	1
1	1	CBL	W802*	2
2	2	CBL	Q249E	2

	ID	Gene	Variation	Class
3	3	CBL	N454D	3
4	4	CBL	L399V	4

Reading Text Data

In [4]:

```
# note the seprator in this file
data_text =pd.read_csv("C:\\Users\\Khushmeet Singh\\Downloads\\training_text",sep="\|\|",engine="p
ython",names=["ID","TEXT"],skiprows=1)
print('Number of data points : ', data_text.shape[0])
print('Number of features : ', data_text.shape[1])
print('Features : ', data_text.columns.values)
data_text.head()
```

```
Number of data points :  3321
Number of features :  2
Features :  ['ID' 'TEXT']
```

Out[4]:

	ID	TEXT
0	0	Cyclin-dependent kinases (CDKs) regulate a var...
1	1	Abstract Background Non-small cell lung canc...
2	2	Abstract Background Non-small cell lung canc...
3	3	Recent evidence has demonstrated that acquired...
4	4	Oncogenic mutations in the monomeric Casitas B...

#Preprocessing

In [5]:

```
# loading stop words from nltk library
stop_words = set(stopwords.words('english'))


def nlp_preprocessing(total_text, index, column):
    if type(total_text) is not int:
        string = ""
        # replace every special char with space
        total_text = re.sub('[^a-zA-Z0-9\n]', ' ', total_text)
        # replace multiple spaces with single space
        total_text = re.sub('\s+',' ', total_text)
        # converting all the chars into lower-case.
        total_text = total_text.lower()

        for word in total_text.split():
        # if the word is a not a stop word then retain that word from the data
            if not word in stop_words:
                string += word + " "

        data_text[column][index] = string
```

In [6]:

```
# Text processing stage.
start_time = time.clock()
for index, row in data_text.iterrows():
    if type(row['TEXT']) is str:
        nlp_preprocessing(row['TEXT'], index, 'TEXT')
    else:
        print("there is no text description for id:",index)
print('Time took for preprocessing the text :',time.clock() - start time, "seconds")
```

```
there is no text description for id: 1109
there is no text description for id: 1277
there is no text description for id: 1407
there is no text description for id: 1639
there is no text description for id: 2755
Time took for preprocessing the text : 47.28520687981337 seconds
```

In [7]:

```
# Merging both gene_variations and text data based on ID
result = pd.merge(data_variants, data_text,on='ID', how='left')
result.head()
```

Out[7]:

	ID	Gene	Variation	Class	TEXT
0	0	FAM58A	Truncating Mutations	1	cyclin dependent kinases cdks regulate variety...
1	1	CBL	W802*	2	abstract background non small cell lung cancer...
2	2	CBL	Q249E	2	abstract background non small cell lung cancer...
3	3	CBL	N454D	3	recent evidence demonstrated acquired uniparen...
4	4	CBL	L399V	4	oncogenic mutations monomeric casitas b lineag...

In [8]:

```
result[result.isnull().any(axis=1)]
```

Out[8]:

	ID	Gene	Variation	Class	TEXT
1109	1109	FANCA	S1088F	1	NaN
1277	1277	ARID5B	Truncating Mutations	1	NaN
1407	1407	FGFR3	K508M	6	NaN
1639	1639	FLT1	Amplification	6	NaN
2755	2755	BRAF	G596C	7	NaN

In [9]:

```
result.loc[result['TEXT'].isnull(),'TEXT'] = result['Gene'] +' '+result['Variation']
```

In [10]:

```
result[result['ID']==1109]
```

Out[10]:

	ID	Gene	Variation	Class	TEXT
1109	1109	FANCA	S1088F	1	FANCA S1088F

Train test split data

In [11]:

```
result.Gene = result.Gene.str.replace('\s+', '_')
result.Variation = result.Variation.str.replace('\s+', '_')
y_true = result[['Class']]
x_true = result.drop(['Class'], axis=1)
```

```
print("Feature columns in dataset: ")
print(x_true.head())
print()
print("Target columns in dataset: ")
print(y_true.head())
```

```
Feature columns in dataset:
   ID    Gene             Variation  \
0   0  FAM58A  Truncating_Mutations
1   1     CBL                W802*
2   2     CBL                Q249E
3   3     CBL                N454D
4   4     CBL                L399V

                                                TEXT
0  cyclin dependent kinases cdks regulate variety...
1  abstract background non small cell lung cancer...
2  abstract background non small cell lung cancer...
3  recent evidence demonstrated acquired uniparen...
4  oncogenic mutations monomeric casitas b lineag...

Target columns in dataset:
   Class
0      1
1      2
2      2
3      3
4      4
```

In [12]:

```
# Split the data into test and train by maintaining same distribution of output varaible 'y_true' [stratify=y_true]
x_train, x_test, y_train, y_test = train_test_split(x_true, y_true, stratify=y_true, test_size=0.2)

# Split the train data into train and cross validation by maintaining same distribution of output varaible 'y_train' [stratify=y_train]
x_train, x_cv, y_train, y_cv = train_test_split(x_train, y_train, stratify=y_train, test_size=0.2)
```

In [13]:

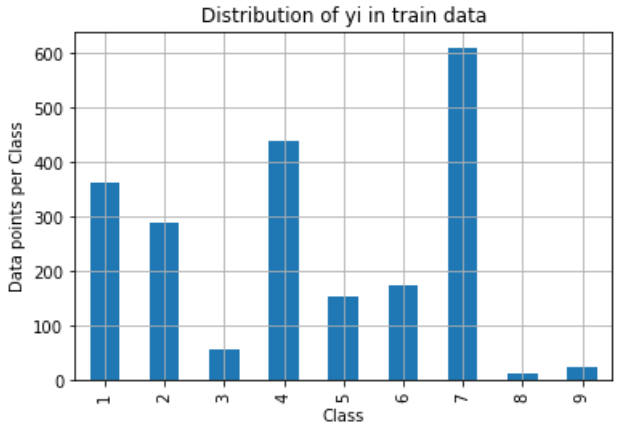
```
print('Number of data points in train data:', x_train.shape[0])
print('Number of data points in test data:', x_test.shape[0])
print('Number of data points in cross validation data:', x_cv.shape[0])
```

```
Number of data points in train data: 2124
Number of data points in test data: 665
Number of data points in cross validation data: 532
```

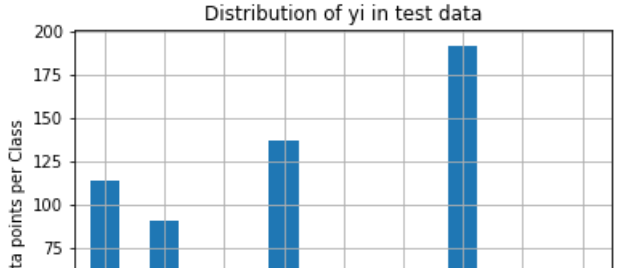
In [14]:

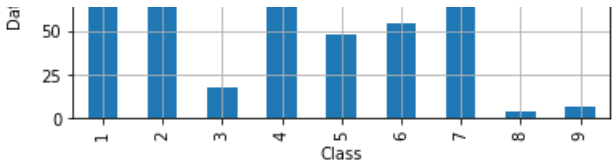
```
def plot_distribution(class_distribution,title,xlabel,ylabel):
    class_distribution.plot(kind='bar')
    plt.xlabel(xlabel)
    plt.ylabel(ylabel)
    plt.title(title)
    plt.grid()
    plt.show()


# it returns a dict, keys as class labels and values as the number of data points in that class
train_class_distribution = y_train['Class'].value_counts().sort_index()
test_class_distribution = y_test['Class'].value_counts().sort_index()
cv_class_distribution = y_cv['Class'].value_counts().sort_index()

plot_distribution(train_class_distribution,
                  'Distribution of yi in train data',
                  'Class',
                  'Data points per Class')

# ref: argsort https://docs.scipy.org/doc/numpy/reference/generated/numpy.argsort.html
# -(train_class_distribution.values): the minus sign will give us in decreasing order
```

```
sorted_yi = np.argsort(-train_class_distribution.values)
for i in sorted_yi:
    print('Number of data points in class', i+1, ':',train_class_distribution.values[i],
          '(', np.round((train_class_distribution.values[i]/x_train.shape[0]*100), 3), '%)')

print('-'*80)



plot_distribution(test_class_distribution,
                  'Distribution of yi in test data',
                  'Class',
                  'Data points per Class')

# ref: argsort https://docs.scipy.org/doc/numpy/reference/generated/numpy.argsort.html
# -(test_class_distribution.values): the minus sign will give us in decreasing order
sorted_yi = np.argsort(-test_class_distribution.values)
for i in sorted_yi:
    print('Number of data points in class', i+1, ':',test_class_distribution.values[i],
          '(', np.round((test_class_distribution.values[i]/x_test.shape[0]*100), 3), '%)')

print('-'*80)



plot_distribution(cv_class_distribution,
                  'Distribution of yi in cross validation data',
                  'Class',
                  'Data points per Class')

# ref: argsort https://docs.scipy.org/doc/numpy/reference/generated/numpy.argsort.html
# -(cv_class_distribution.values): the minus sign will give us in decreasing order
sorted_yi = np.argsort(-cv_class_distribution.values)
for i in sorted_yi:
    print('Number of data points in class', i+1, ':',cv_class_distribution.values[i],
          '(', np.round((cv_class_distribution.values[i]/x_cv.shape[0]*100), 3), '%)')
```

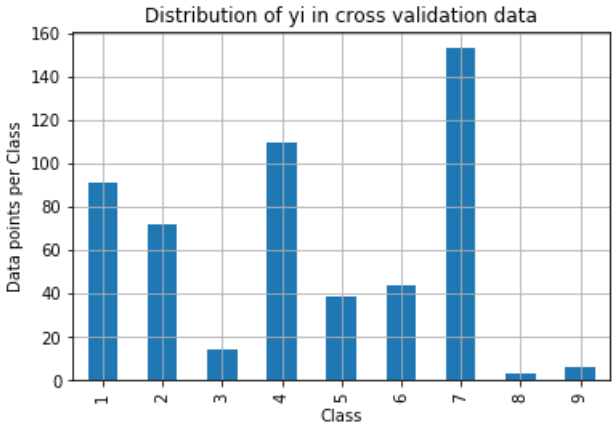



```
Number of data points in class 7 : 609 ( 28.672 %)
Number of data points in class 4 : 439 ( 20.669 %)
Number of data points in class 1 : 363 ( 17.09 %)
Number of data points in class 2 : 289 ( 13.606 %)
Number of data points in class 6 : 176 ( 8.286 %)
Number of data points in class 5 : 155 ( 7.298 %)
Number of data points in class 3 : 57 ( 2.684 %)
Number of data points in class 9 : 24 ( 1.13 %)
Number of data points in class 8 : 12 ( 0.565 %)
--------------------------------------------------------------------------------
```






```
Number of data points in class 7 : 191 ( 28.722 %)
Number of data points in class 4 : 137 ( 20.602 %)
Number of data points in class 1 : 114 ( 17.143 %)
Number of data points in class 2 : 91 ( 13.684 %)
Number of data points in class 6 : 55 ( 8.271 %)
Number of data points in class 5 : 48 ( 7.218 %)
Number of data points in class 3 : 18 ( 2.707 %)
Number of data points in class 9 : 7 ( 1.053 %)
Number of data points in class 8 : 4 ( 0.602 %)
------------------------------------------------------------------------------------
```




```
Number of data points in class 7 : 153 ( 28.759 %)
Number of data points in class 4 : 110 ( 20.677 %)
Number of data points in class 1 : 91 ( 17.105 %)
Number of data points in class 2 : 72 ( 13.534 %)
Number of data points in class 6 : 44 ( 8.271 %)
Number of data points in class 5 : 39 ( 7.331 %)
Number of data points in class 3 : 14 ( 2.632 %)
Number of data points in class 9 : 6 ( 1.128 %)
Number of data points in class 8 : 3 ( 0.564 %)
```

Prediction using Random Model

In [57]:

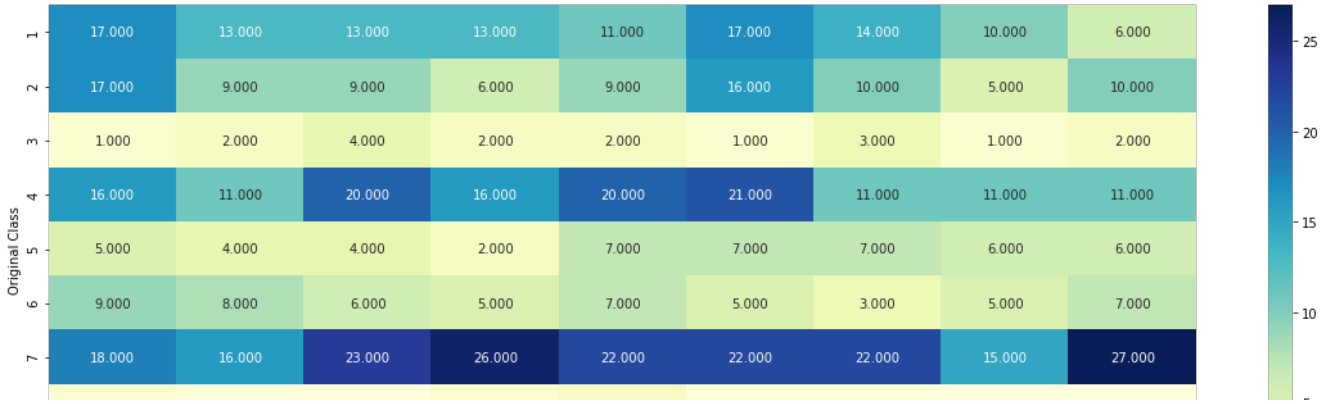
```
def plot_matrix(matrix,labels):
    plt.figure(figsize=(20,7))
    sns.heatmap(matrix, annot=True, cmap="YlGnBu", fmt=".3f", xticklabels=labels, yticklabels=label
s)
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.show()

# This function plots the confusion matrices given y_i, y_i_hat.
def plot_confusion_matrix(test_y, predict_y):
    cm = confusion_matrix(test_y, predict_y)
    # C = 9,9 matrix, each cell (i,j) represents number of points of class i are predicted class j

    recall_table =(((cm.T)/(cm.sum(axis=1))).T)
    # How did we calculateed recall_table :
    # divide each element of the confusion matrix with the sum of elements in that column
    # C = [[1, 2],
    #      [3, 4]]
    # C.T = [[1, 3],
    #        [2, 4]]
    # C.sum(axis = 1)  axis=0 corresonds to columns and axis=1 corresponds to rows in two
diamensional array
    # C.sum(axix =1) = [[3, 7]]
```

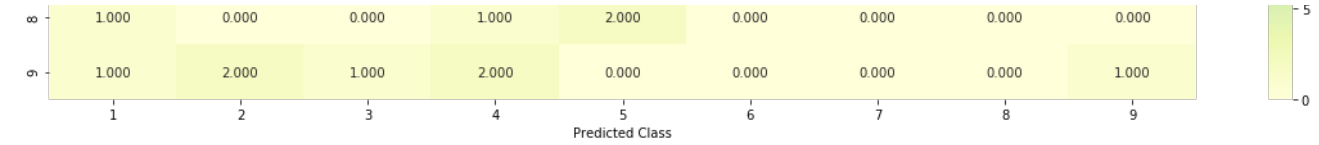
```
    # ((C.T)/(C.sum(axis=1))) = [[1/3, 3/7]
    #                           [2/3, 4/7]]
    # ((C.T)/(C.sum(axis=1))).T = [[1/3, 2/3]
    #                           [3/7, 4/7]]
    # sum of row elements = 1

    precision_table =(cm/cm.sum(axis=0))
    # How did we calculateed precision_table :
    # divide each element of the confusion matrix with the sum of elements in that row
    # C = [[1, 2],
    #     [3, 4]]
    # C.sum(axis = 0)  axis=0 corresonds to columns and axis=1 corresponds to rows in two
diamensional array
    # C.sum(axix =0) = [[4, 6]]
    # (C/C.sum(axis=0)) = [[1/4, 2/6],
    #                      [3/4, 4/6]]

    labels = [1,2,3,4,5,6,7,8,9]
    print()
    print("-"*20, "Confusion matrix", "-"*20)
    plot_matrix(cm,labels)

    print("-"*20, "Precision matrix (Columm Sum=1)", "-"*20)
    plot_matrix(precision_table,labels)

    print("-"*20, "Recall matrix (Row sum=1)", "-"*20)
    plot_matrix(recall_table,labels)
```

In [16]:

```
# one solution is to genarate 9 numbers and divide each of the numbers by their sum
# ref: https://stackoverflow.com/a/18662466/4084039
test_data_len = x_test.shape[0]
cv_data_len = x_cv.shape[0]

# we create a output array that has exactly same size as the CV data
cv_predicted_y = np.zeros((cv_data_len,9))
for i in range(cv_data_len):
    rand_probs = np.random.rand(1,9)
    cv_predicted_y[i] = ((rand_probs/sum(sum(rand_probs)))[0])
print("Log loss on Cross Validation Data using Random Model",log_loss(y_cv,cv_predicted_y, eps=1e-
15))


# Test-Set error.
# We create a output array that has exactly same as the test data
test_predicted_y = np.zeros((test_data_len,9))
for i in range(test_data_len):
    rand_probs = np.random.rand(1,9)
    test_predicted_y[i] = ((rand_probs/sum(sum(rand_probs)))[0])
print("Log loss on Test Data using Random Model",log_loss(y_test,test_predicted_y, eps=1e-15))

predicted_y =np.argmax(test_predicted_y, axis=1)
plot_confusion_matrix(y_test, predicted_y+1)
```

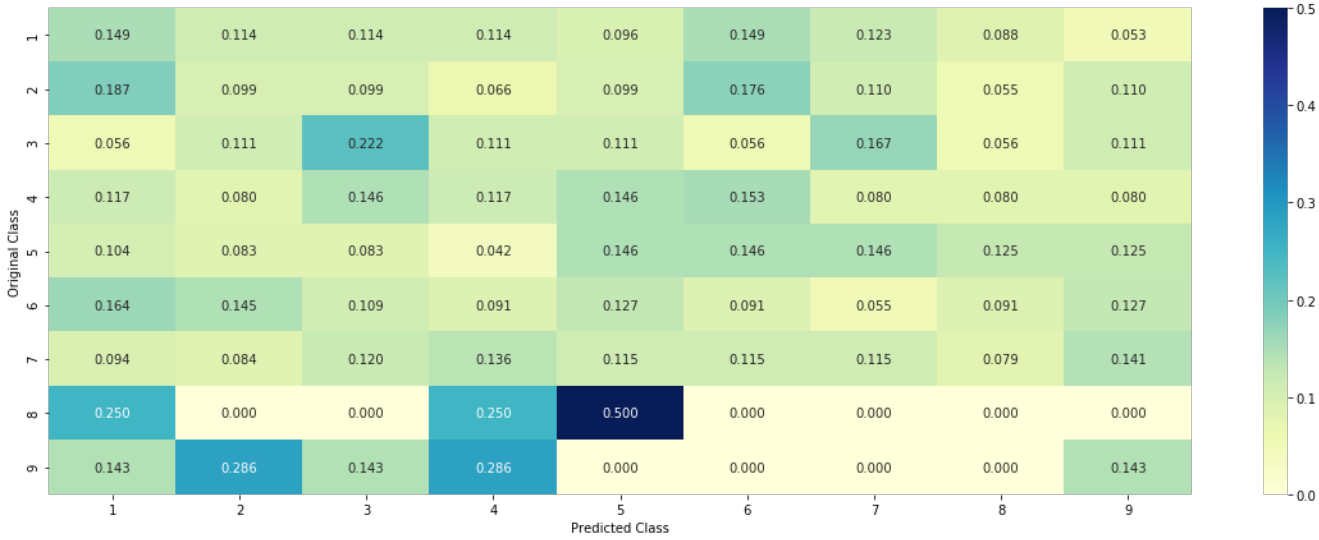
```
Log loss on Cross Validation Data using Random Model 2.497792642083869
Log loss on Test Data using Random Model 2.472101948254495

-------------------- Confusion matrix --------------------
```





---------- Precision matrix (Columm Sum=1) ----------

---------- Recall matrix (Row sum=1) ----------



Univariate Analysis

In [17]:

```
# code for response coding with Laplace smoothing.
# alpha : used for laplace smoothing
# feature: ['gene', 'variation']
# df: ['x_train', 'x_test', 'x_cv']
# algorithm
# ----------
# Consider all unique values and the number of occurances of given feature in train data dataframe
# build a vector (1*9) , the first element = (number of times it occured in class1 + 10*alpha / nu
mber of time it occurred in total data+90*alpha)
# gv_dict is like a look up table, for every gene it store a (1*9) representation of it
# for a value of feature in df:
# if it is in train data:
# we add the vector that was stored in 'gv_dict' look up table to 'gv_fea'
# if it is not there is train:
# we add [1/9, 1/9, 1/9, 1/9,1/9, 1/9, 1/9, 1/9, 1/9] to 'gv_fea'
# return 'gv_fea'
```

```
# ----------------------

# get_gv_fea_dict: Get Gene varaition Feature Dict
def get_gv_fea_dict(alpha, feature, df):
    # value_count: it contains a dict like
    # print(train_df['Gene'].value_counts())
    # output:
    #        {BRCA1      174
    #         TP53       106
    #         EGFR        86
    #         BRCA2       75
    #         PTEN        69
    #         KIT         61
    #         BRAF        60
    #         ERBB2       47
    #         PDGFRA      46
    #         ...}
    # print(train_df['Variation'].value_counts())
    # output:
    # {
    # Truncating_Mutations                     63
    # Deletion                                 43
    # Amplification                            43
    # Fusions                                  22
    # Overexpression                            3
    # E17K                                      3
    # Q61L                                      3
    # S222D                                     2
    # P130S                                     2
    # ...
    # }
    value_count = x_train[feature].value_counts()

    # gv_dict : Gene Variation Dict, which contains the probability array for each gene/variation
    gv_dict = dict()

    # denominator will contain the number of time that particular feature occured in whole data
    for i, denominator in value_count.items():
        # vec will contain (p(yi==1/Gi) probability of gene/variation belongs to perticular class
        # vec is 9 diamensional vector
        vec = []
        for k in range(1,10):
            # print(train_df.loc[(train_df['Class']==1) & (train_df['Gene']=='BRCA1')])
            #         ID   Gene             Variation  Class
            # 2470  2470  BRCA1               S1715C      1
            # 2486  2486  BRCA1               S1841R      1
            # 2614  2614  BRCA1                  M1R      1
            # 2432  2432  BRCA1               L1657P      1
            # 2567  2567  BRCA1               T1685A      1
            # 2583  2583  BRCA1               E1660G      1
            # 2634  2634  BRCA1               W1718L      1
            # cls_cnt.shape[0] will return the number of rows

            cls_cnt = x_train.loc[(y_train['Class']==k) & (x_train[feature]==i)]

            # cls_cnt.shape[0](numerator) will contain the number of time that particular feature o
ccured in whole data
            vec.append((cls_cnt.shape[0] + alpha*10)/ (denominator + 90*alpha))

        # we are adding the gene/variation to the dict as key and vec as value
        gv_dict[i]=vec
    return gv_dict

# Get Gene variation feature
def get_gv_feature(alpha, feature, df):
    # print(gv_dict)
    #     {'BRCA1': [0.20075757575757575, 0.03787878787878788, 0.068181818181818177,
0.13636363636363635, 0.25, 0.19318181818181818, 0.03787878787878788, 0.03787878787878788,
0.03787878787878788],
    #      'TP53': [0.32142857142857145, 0.061224489795918366, 0.061224489795918366,
0.27040816326530615, 0.061224489795918366, 0.066326530612244902, 0.051020408163265307, 0.051020408
163265307, 0.056122448979591837],
    #      'EGFR': [0.056818181818181816, 0.21590909090909091, 0.0625, 0.068181818181818177,
0.068181818181818177, 0.0625, 0.34659090909090912, 0.0625, 0.056818181818181816],
    #      'BRCA2': [0.13333333333333333, 0.060606060606060608, 0.060606060606060608,
0.078787878787878782, 0.1393939393939394, 0.34545454545454546, 0.060606060606060608,
0.060606060606060608, 0.060606060606060608],
```

```
    #       'PTEN': [0.069182389937106917, 0.062893081761006289, 0.069182389937106917,
0.46540880503144655, 0.075471698113207544, 0.062893081761006289, 0.069182389937106917, 0.062893081
761006289, 0.062893081761006289],
    #       'KIT': [0.066225165562913912, 0.25165562913907286, 0.072847682119205295,
0.072847682119205295, 0.066225165562913912, 0.066225165562913912, 0.27152317880794702,
0.066225165562913912, 0.066225165562913912],
    #       'BRAF': [0.066666666666666666, 0.17999999999999999, 0.073333333333333334,
0.073333333333333334, 0.093333333333333338, 0.080000000000000002, 0.29999999999999999,
0.066666666666666666, 0.066666666666666666],
    #       ...
    #     }
    gv_dict = get_gv_fea_dict(alpha, feature, df)
    # value_count is similar in get_gv_fea_dict
    value_count = x_train[feature].value_counts()

    # gv_fea: Gene_variation feature, it will contain the feature for each feature value in the da
ta
    gv_fea = []
    # for every feature values in the given data frame we will check if it is there in the train
data then we will add the feature to gv_fea
    # if not we will add [1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9] to gv_fea
    for index, row in df.iterrows():
        if row[feature] in dict(value_count).keys():
            gv_fea.append(gv_dict[row[feature]])
        else:
            gv_fea.append([1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9])
#             gv_fea.append([-1,-1,-1,-1,-1,-1,-1,-1,-1])
    return gv_fea
```

In [18]:

```
unique_genes = x_train['Gene'].value_counts()
print('Number of Unique Genes :', unique_genes.shape[0])
# the top 10 genes that occured most
print(unique_genes.head(10))
```

```
Number of Unique Genes : 232
BRCA1     169
TP53      104
EGFR       84
BRCA2      83
PTEN       77
BRAF       60
KIT        59
ERBB2      47
PDGFRA     41
ALK        40
Name: Gene, dtype: int64
```

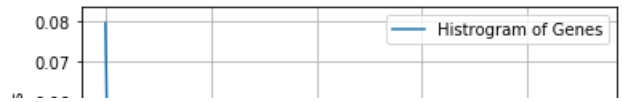
In [19]:

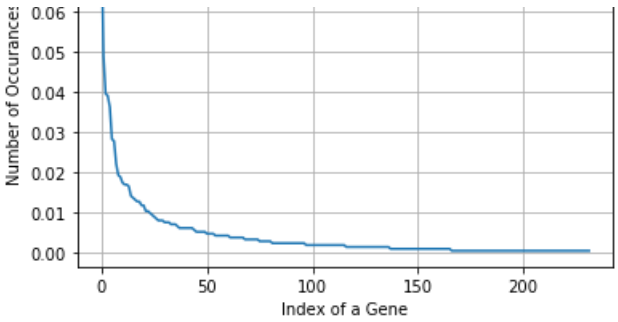
```
print("Ans: There are", unique_genes.shape[0] ,"different categories of genes in the train data, an
d they are distibuted as follows",)
```

```
Ans: There are 232 different categories of genes in the train data, and they are distibuted as fol
lows
```

In [20]:

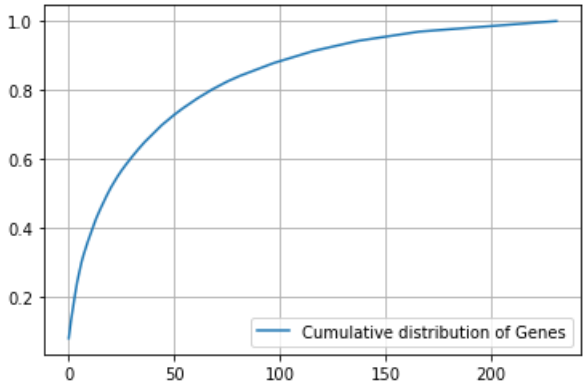
```
s = sum(unique_genes.values);
h = unique_genes.values/s;
plt.plot(h, label="Histrogram of Genes")
plt.xlabel('Index of a Gene')
plt.ylabel('Number of Occurances')
plt.legend()
plt.grid()
plt.show()
```





In [21]:

```
c = np.cumsum(h)
plt.plot(c,label='Cumulative distribution of Genes')
plt.grid()
plt.legend()
plt.show()
```



In [22]:

```
#response-coding of the Gene feature
# alpha is used for laplace smoothing
alpha = 1
# train gene feature
train_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_train))
# test gene feature
test_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_test))
# cross validation gene feature
cv_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_cv))
```

In [26]:

```
print("train_gene_feature_responseCoding is converted feature using respone coding method. The sha
pe of gene feature:", train_gene_feature_responseCoding.shape)
```

```
train_gene_feature_responseCoding is converted feature using respone coding method. The shape of g
ene feature: (2124, 9)
```

In [27]:

```
# one-hot encoding of Gene feature.
gene_vectorizer = TfidfVectorizer()
train_gene_feature_onehotCoding = gene_vectorizer.fit_transform(x_train['Gene'])
test_gene_feature_onehotCoding = gene_vectorizer.transform(x_test['Gene'])
cv_gene_feature_onehotCoding = gene_vectorizer.transform(x_cv['Gene'])
```

In [28]:

```
train_gene_feature_onehotCoding
```

Out[28]:

```
<2124x236 sparse matrix of type '<class 'numpy.float64'>'
 with 2124 stored elements in Compressed Sparse Row format>
```

In [29]:

```
print("train_gene_feature_onehotCoding is converted feature using one-hot encoding method. The sha
pe of gene feature:",
      train_gene_feature_onehotCoding.shape)
```

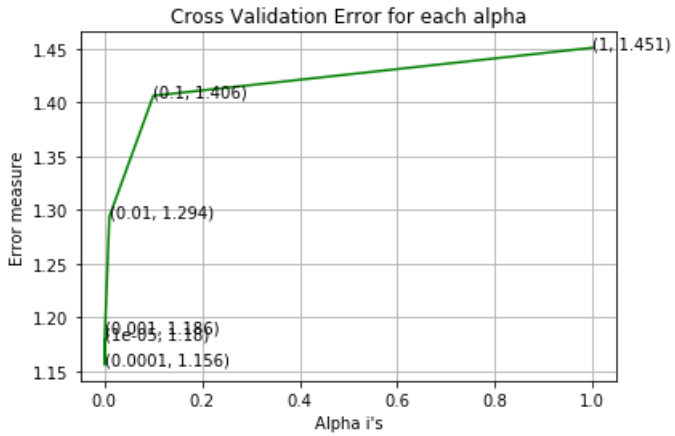
```
train_gene_feature_onehotCoding is converted feature using one-hot encoding method. The shape of g
ene feature: (2124, 236)
```

In [30]:

```
alpha = [10 ** x for x in range(-5, 1)] # hyperparam for SGD classifier.

# read more about SGDClassifier() at http://scikit-
learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_i
ter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0
=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, ...]) Fit linear model with Stochastic Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link:
#-------------------------------


cv_log_error_array=[]
for i in alpha:
    clf = SGDClassifier(alpha=i, penalty='l2', loss='log', random_state=42)
    clf.fit(train_gene_feature_onehotCoding, y_train)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_gene_feature_onehotCoding, y_train)
    predict_y = sig_clf.predict_proba(cv_gene_feature_onehotCoding)
    cv_log_error_array.append(log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
    print('For values of alpha = ', i, "The log loss is:",log_loss(y_cv, predict_y, labels=clf.clas
ses_, eps=1e-15))

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_gene_feature_onehotCoding, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_gene_feature_onehotCoding, y_train)

predict_y = sig_clf.predict_proba(train_gene_feature_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The train log loss is:",
      log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15))

predict_y = sig_clf.predict_proba(cv_gene_feature_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The cross validation log loss is:",
      log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
```

```
predict_y = sig_clf.predict_proba(test_gene_feature_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The test log loss is:",
      log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
For values of alpha =  1e-05 The log loss is: 1.1797694412356643
For values of alpha =  0.0001 The log loss is: 1.155621785595475
For values of alpha =  0.001 The log loss is: 1.185846909120284
For values of alpha =  0.01 The log loss is: 1.2937854756949845
For values of alpha =  0.1 The log loss is: 1.4063938697951321
For values of alpha =  1 The log loss is: 1.4511312472038138
```




```
For values of best alpha =  0.0001 The train log loss is: 1.014604830453934
For values of best alpha =  0.0001 The cross validation log loss is: 1.155621785595475
For values of best alpha =  0.0001 The test log loss is: 1.1818342257320826
```

In [31]:

```
print("Q6. How many data points in Test and CV datasets are covered by the ",
      unique_genes.shape[0], " genes in train dataset?")

test_coverage=x_test[x_test['Gene'].isin(list(set(x_train['Gene'])))].shape[0]
cv_coverage=x_cv[x_cv['Gene'].isin(list(set(x_train['Gene'])))].shape[0]

print('Ans\n1. In test data',test_coverage, 'out of',x_test.shape[0], ":",(test_coverage/x_test.sha
pe[0])*100)
print('2. In cross validation data',cv_coverage, 'out of ',x_cv.shape[0],":" ,(cv_coverage/x_cv.sha
pe[0])*100)
```

```
Q6. How many data points in Test and CV datasets are covered by the  236  genes in train dataset?
Ans
1. In test data 641 out of 665 : 96.39097744360903
2. In cross validation data 521 out of  532 : 97.93233082706767
```

In [32]:

```
unique_variations = x_train['Variation'].value_counts()
print('Number of Unique Variations :', unique_variations.shape[0])
# the top 10 variations that occured most
print(unique_variations.head(10))
```

```
Number of Unique Variations : 1928
Truncating_Mutations    63
Amplification           49
Deletion                47
Fusions                 17
Overexpression           6
G12V                     4
E17K                     3
ETV6-NTRK3_Fusion        2
S308A                    2
Q61H                     2
Name: Variation, dtype: int64
```

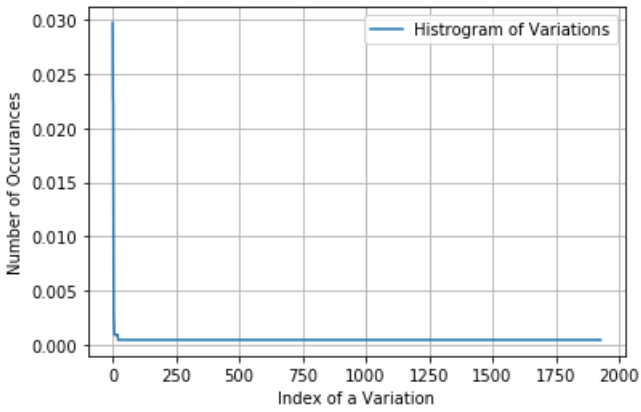
In [33]:

```
print("Ans: There are", unique_variations.shape[0] ,
      "different categories of variations in the train data, and they are distibuted as follows",)
```

```
Ans: There are 1928 different categories of variations in the train data, and they are distibuted
as follows
```

In [34]:

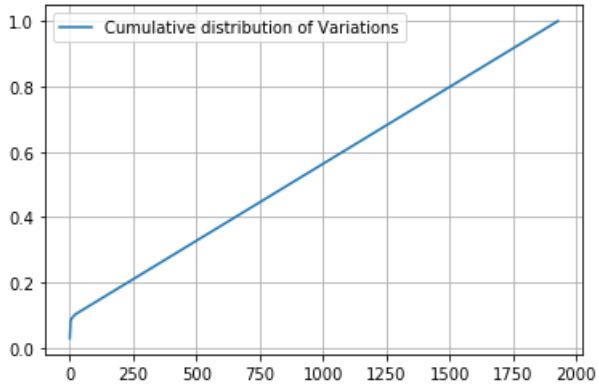
```
s = sum(unique_variations.values);
h = unique_variations.values/s;
plt.plot(h, label="Histrogram of Variations")
plt.xlabel('Index of a Variation')
plt.ylabel('Number of Occurances')
plt.legend()
plt.grid()
plt.show()
```



In [35]:

```
c = np.cumsum(h)
print(c)
plt.plot(c,label='Cumulative distribution of Variations')
plt.grid()
plt.legend()
plt.show()
```

```
[0.02966102 0.0527307  0.07485876 ... 0.99905838 0.99952919 1.        ]
```



In [36]:

```
# alpha is used for laplace smoothing
alpha = 1

# train gene feature
train_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", x_train))
```

```
# test gene feature
test_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", x_test))

# cross validation gene feature
cv_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", x_cv))
```

In [37]:

```
print("train_variation_feature_responseCoding is a converted feature using the response coding met
hod. The shape of Variation feature:",
      train_variation_feature_responseCoding.shape)
```

```
train_variation_feature_responseCoding is a converted feature using the response coding method. Th
e shape of Variation feature: (2124, 9)
```

In [38]:

```
# one-hot encoding of variation feature.
variation_vectorizer = TfidfVectorizer()
train_variation_feature_onehotCoding = variation_vectorizer.fit_transform(x_train['Variation'])
test_variation_feature_onehotCoding = variation_vectorizer.transform(x_test['Variation'])
cv_variation_feature_onehotCoding = variation_vectorizer.transform(x_cv['Variation'])
```

In [39]:

```
print("train_variation_feature_onehotEncoded is converted feature using the onne-hot encoding meth
od. The shape of Variation feature:",
      train_variation_feature_onehotCoding.shape)
```

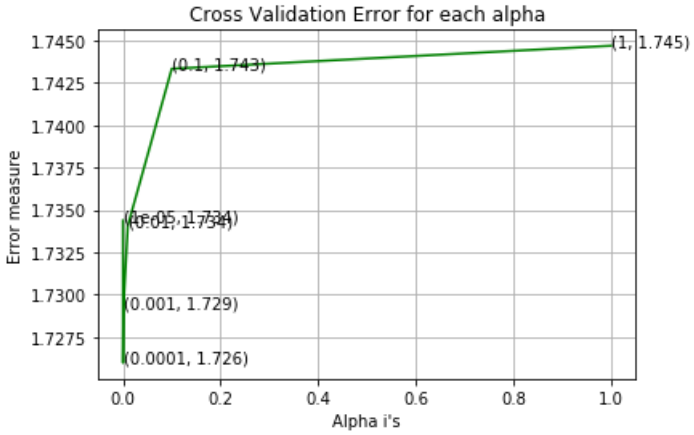
```
train_variation_feature_onehotEncoded is converted feature using the onne-hot encoding method. The
shape of Variation feature: (2124, 1961)
```

In [40]:

```
alpha = [10 ** x for x in range(-5, 1)]

# read more about SGDClassifier() at http://scikit-
learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_i
ter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0
=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, ...]) Fit linear model with Stochastic Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link:
#-------------------------------


cv_log_error_array=[]
for i in alpha:
    clf = SGDClassifier(alpha=i, penalty='l2', loss='log', random_state=42)
    clf.fit(train_variation_feature_onehotCoding, y_train)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_variation_feature_onehotCoding, y_train)
    predict_y = sig_clf.predict_proba(cv_variation_feature_onehotCoding)
    cv_log_error_array.append(log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
    print('For values of alpha = ', i, "The log loss is:",log_loss(y_cv, predict_y, labels=clf.clas
ses_, eps=1e-15))

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],cv_log_error_array[i]))
plt.grid()
```

```
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_variation_feature_onehotCoding, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_variation_feature_onehotCoding, y_train)

predict_y = sig_clf.predict_proba(train_variation_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train,
predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_variation_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_variation_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
For values of alpha =  1e-05 The log loss is: 1.734376903289353
For values of alpha =  0.0001 The log loss is: 1.7259746698767122
For values of alpha =  0.001 The log loss is: 1.7292469555718046
For values of alpha =  0.01 The log loss is: 1.734075581840475
For values of alpha =  0.1 The log loss is: 1.7433309221788011
For values of alpha =  1 The log loss is: 1.7446894243597695
```




```
For values of best alpha =  0.0001 The train log loss is: 0.7511696453332067
For values of best alpha =  0.0001 The cross validation log loss is: 1.7259746698767122
For values of best alpha =  0.0001 The test log loss is: 1.6827263182540182
```

In [41]:

```
print("Q12. How many data points are covered by total ",
      unique_variations.shape[0],
      " genes in test and cross validation data sets?")
test_coverage=x_test[x_test['Variation'].isin(list(set(x_train['Variation'])))].shape[0]
cv_coverage=x_cv[x_cv['Variation'].isin(list(set(x_train['Variation'])))].shape[0]
print('Ans\n1. In test data',test_coverage, 'out of',x_test.shape[0], ":",(test_coverage/x_test.sha
pe[0])*100)
print('2. In cross validation data',cv_coverage, 'out of ',x_cv.shape[0],":" ,(cv_coverage/x_cv.sha
pe[0])*100)
```

```
Q12. How many data points are covered by total  1928  genes in test and cross validation data
sets?
Ans
1. In test data 82 out of 665 : 12.330827067669173
2. In cross validation data 44 out of  532 : 8.270676691729323
```

In [45]:

```
# cls_text is a data frame
# for every row in data fram consider the 'TEXT'
# split the words by space
```

```
# split the words by space
# make a dict with those words
# increment its count whenever we see that word

def extract_dictionary_paddle(cls_text):
    dictionary = defaultdict(int)
    for index, row in cls_text.iterrows():
        for word in row['TEXT'].split():
            dictionary[word] +=1
    return dictionary
```

In [46]:

```
import math
#https://stackoverflow.com/a/1602964
def get_text_responsecoding(df):
    text_feature_responseCoding = np.zeros((df.shape[0],9))
    for i in range(0,9):
        row_index = 0
        for index, row in df.iterrows():
            sum_prob = 0
            for word in row['TEXT'].split():
                sum_prob += math.log(((dict_list[i].get(word,0)+10 )/(total_dict.get(word,0)+90)))
            text_feature_responseCoding[row_index][i] = math.exp(sum_prob/len(row['TEXT'].split()))
            row_index += 1
    return text_feature_responseCoding
```

In [47]:

```
x_train['TEXT'].head()
```

Out[47]:

```
2946    abstract case 84 year old woman diagnosed stag...
1849    introduction ctcf originally identified zf 3 p...
2267    pten phosphatase tensin homolog phosphatase un...
2660    mutations brca1 brca2 account majority heredit...
2176    phosphatase tensin homolog pten inactivating m...
Name: TEXT, dtype: object
```

In [48]:

```
def top_tfidf_feats(row, features, top_n=25):
    ''' Get top n tfidf values in row and return them with their corresponding feature names.'''
    topn_ids = np.argsort(row)[::-1][:top_n]
    top_feats = [(features[i], row[i]) for i in topn_ids]
    df = pd.DataFrame(top_feats)
    df.columns = ['feature', 'tfidf']
    return df

def top_mean_feats(Xtr, features, min_tfidf=0.1, grp_ids=None, top_n=25):
    ''' Return the top n features that on average are most important amongst documents in rows
        indentified by indices in grp_ids. '''
    if grp_ids:
        D = Xtr[grp_ids].toarray()
    else:
        D = Xtr.toarray()

    D[D < min_tfidf] = 0
    tfidf_means = np.mean(D, axis=0)
    return top_tfidf_feats(tfidf_means, features, top_n)
```

In [46]:

```
# building a CountVectorizer with all the words that occured minimum 3 times in train data
text_vectorizer = TfidfVectorizer(min_df=3)
train_text_feature_onehotCoding = text_vectorizer.fit_transform(x_train['TEXT'])

# getting top 1000 feature names (words)
train_text_features = top_mean_feats(train_text_feature_onehotCoding,
                                                text_vectorizer.get_feature_names(),
                                                top_n=1000)['feature'].tolist()
```

In [47]:

```
# train_text_feature_onehotCoding.sum(axis=0).A1 will sum every row and returns (1*number of featu
res) vector
train_text_fea_counts = train_text_feature_onehotCoding.sum(axis=0).A1
train_text_fea_counts
```

Out[47]:

```
array([9.16453893, 8.84428113, 0.03207983, ..., 0.02137892, 0.0252473 ,
       0.04781381])
```

In [48]:

```
# zip(list(text_features),text_fea_counts) will zip a word with its number of times it occured
text_fea_dict = dict(zip(list(train_text_features),train_text_fea_counts))


print("Total number of unique words in train data :", len(train_text_features))
```

```
Total number of unique words in train data : 1000
```

In [49]:

```
dict_list = []
# dict_list =[] contains 9 dictoinaries each corresponds to a class
for i in range(1,10):
    cls_text = x_train[y_train['Class']==i]
    # build a word dict based on the words in that class
    dict_list.append(extract_dictionary_paddle(cls_text))
    # append it to dict_list

# dict_list[i] is build on i'th  class text data
# total_dict is buid on whole training text data
total_dict = extract_dictionary_paddle(x_train)


confuse_array = []
for i in train_text_features:
    ratios = []
    max_val = -1
    for j in range(0,9):
        ratios.append((dict_list[j][i]+10 )/(total_dict[i]+90))
    confuse_array.append(ratios)
confuse_array = np.array(confuse_array)
```

In [50]:

```
#response coding of text features
train_text_feature_responseCoding  = get_text_responsecoding(x_train)
test_text_feature_responseCoding  = get_text_responsecoding(x_test)
cv_text_feature_responseCoding  = get_text_responsecoding(x_cv)
```

In [51]:

```
# https://stackoverflow.com/a/16202486
# we convert each row values such that they sum to 1
train_text_feature_responseCoding =
(train_text_feature_responseCoding.T/train_text_feature_responseCoding.sum(axis=1)).T
test_text_feature_responseCoding =
(test_text_feature_responseCoding.T/test_text_feature_responseCoding.sum(axis=1)).T
cv_text_feature_responseCoding = (cv_text_feature_responseCoding.T/cv_text_feature_responseCoding.
sum(axis=1)).T
```

In [52]:

```
# don't forget to normalize every feature
train_text_feature_onehotCoding = normalize(train_text_feature_onehotCoding, axis=0)

# we use the same vectorizer that was trained on train data
```

```
test_text_feature_onehotCoding = text_vectorizer.transform(x_test['TEXT'])
# don't forget to normalize every feature
test_text_feature_onehotCoding = normalize(test_text_feature_onehotCoding, axis=0)

# we use the same vectorizer that was trained on train data
cv_text_feature_onehotCoding = text_vectorizer.transform(x_cv['TEXT'])
# don't forget to normalize every feature
cv_text_feature_onehotCoding = normalize(cv_text_feature_onehotCoding, axis=0)
```

In [53]:

```
#https://stackoverflow.com/a/2258273/4084039
sorted_text_fea_dict = dict(sorted(text_fea_dict.items(), key=lambda x: x[1] , reverse=True))
sorted_text_occur = np.array(list(sorted_text_fea_dict.values()))
```

In [54]:

```
# Number of words for a given frequency.
print(Counter(sorted_text_occur))
```

```
Counter({0.0218755337978128: 28, 0.08321330026799921: 16, 0.026246636383422933: 12,
0.03145946494793289: 9, 0.02008233341412003: 8, 0.2437365183161721: 7, 0.02011579759256247: 7,
0.32861753250452697: 6, 0.0386363309935239: 6, 0.022546512549612794: 5, 0.06291892989586578: 4, 0.
015370571147195871: 4, 0.0887160069489253: 3, 0.07451658762836698: 3, 0.02781312058705004: 3,
0.019979238214001684: 3, 0.01844305136135315: 3, 0.018410686629147566: 3, 0.015297310298719644: 3,
0.011549291101875801: 3, 0.1537777084761026: 2, 0.09437839484379865: 2, 0.06454052529505916: 2,
0.06283334985894277: 2, 0.060133265939140645: 2, 0.05237655967454377: 2, 0.049356360584032 67: 2, 0
.048303138326722214: 2, 0.043713637180042716: 2, 0.04016466682824006: 2, 0.03692783406253373: 2, 0
.03596656924259874: 2, 0.02636464513067374: 2, 0.023707986130952194: 2, 0.022705613937073377: 2, 0
.02211812559425675: 2, 0.021164474578780082: 2, 0.021128150160064973: 2, 0.0202627207542918: 2, 0.
018545206272600 63: 2, 0.01642455999414605: 2, 0.014571212393347573: 2, 0.013684985888998672: 2, 0.
009974135953923116: 2, 32.84514657439612: 1, 18.647851940744044: 1, 17.865911582840514: 1,
14.823457379467126: 1, 13.06940342597618: 1, 9.164538934836354: 1, 8.844281126162553: 1, 6.76729902
4865416: 1, 5.448277808934394: 1, 4.972552304764364: 1, 4.666932787819262: 1, 4.040415716656921: 1,
3.9434952574611777: 1, 3.938491717685176: 1, 3.830253724961427: 1, 3.6375122591517655: 1,
3.5619449366133824: 1, 3.5171715454521335: 1, 3.475896449824401: 1, 3.4155664617779 11: 1,
3.3548832738738907: 1, 2.8101709025571373: 1, 2.6811017014778953: 1, 2.6297748356762076: 1,
2.544061050883985: 1, 2.4512983052359636: 1, 2.39149946224829: 1, 2.0730191633828294: 1,
2.0636892805009825: 1, 1.972920498808173: 1, 1.967200753895839: 1, 1.887878744287898: 1, 1.78085094
759688: 1, 1.7693665100434608: 1, 1.6408502840504782: 1, 1.6147898468384796: 1,
1.5867328078503498: 1, 1.5622268939806714: 1, 1.4933437569002366: 1, 1.4545499147439498: 1,
1.3387115651608483: 1, 1.3349008691105833: 1, 1.3043750793680835: 1, 1.2894625610039876: 1,
1.2884969173554264: 1, 1.277143361834737: 1, 1.2370516887310659: 1, 1.1645960776841755: 1,
1.1548655827895737: 1, 1.111744617656526: 1, 0.9451063879212047: 1, 0.9439090243657683: 1,
0.9284959457758171: 1, 0.9229780385294615: 1, 0.920515592041798: 1, 0.8901726011115413: 1,
0.8373428642726165: 1, 0.8151065386404605: 1, 0.7672829460423848: 1, 0.7633196512668657: 1,
0.7613834306219481: 1, 0.7512947722205152: 1, 0.7495777476374436: 1, 0.7382376765093839: 1,
0.7378208240336732: 1, 0.7333482568604017: 1, 0.7272311186931664: 1, 0.7265647381984874: 1,
0.7249674171497306: 1, 0.7161162025825365: 1, 0.7046399241176203: 1, 0.6632395495085028: 1,
0.6574522864147084: 1, 0.6515614532100821: 1, 0.6426139133606246: 1, 0.6251564595446479: 1,
0.6183700764951137: 1, 0.6085255353702246: 1, 0.6005748305440448: 1, 0.5898270587161774: 1,
0.5759535580282213: 1, 0.5696087472083877: 1, 0.567498680535008: 1, 0.5662971303695836: 1,
0.5434210343902178: 1, 0.5216231187407915: 1, 0.5043492115497825: 1, 0.5006227299914484: 1,
0.49072881732052653: 1, 0.475904410369326: 1, 0.4694562228121825: 1, 0.44440363838129776: 1,
0.42897047201399563: 1, 0.4283255984227381: 1, 0.4274336899311197: 1, 0.4254504389383318: 1,
0.42515939428060173: 1, 0.42192673277092585: 1, 0.4158498021744742: 1, 0.4041222778459 88: 1,
0.4032690998145317: 1, 0.3993355868543165: 1, 0.39496098176828637: 1, 0.3930864979812366: 1,
0.3903591145103835: 1, 0.38508929366000005: 1, 0.3832137920645352: 1, 0.37759605779259015: 1,
0.3755023644496816: 1, 0.3734594380166687: 1, 0.37163833532963064: 1, 0.3670645037928947: 1,
0.3613775262712479: 1, 0.3547844714679051: 1, 0.35233382893653975: 1, 0.35174114274781054: 1,
0.34980983468006516: 1, 0.3468236060623477: 1, 0.34574686977925806: 1, 0.34449037300950863: 1,
0.3437002738868112: 1, 0.3412610520429364: 1, 0.33835723383158645: 1, 0.33298595516487606: 1,
0.3326093473852135: 1, 0.3323482356490455: 1, 0.3307334275562701: 1, 0.3296089823228605: 1,
0.32912343818922374: 1, 0.3276211520485098: 1, 0.3267638077443173: 1, 0.3244516011076179: 1,
0.3235262393537481: 1, 0.32300674448630995: 1, 0.31835364733969956: 1, 0.3180845947026471: 1,
0.31666306484325474: 1, 0.3161949616616533: 1, 0.3120778909057146: 1, 0.3105660059333294: 1,
0.30810751935272807: 1, 0.30700552849486357: 1, 0.30319416370575664: 1, 0.3029972127606094: 1,
0.30118666867362065: 1, 0.2973466499640328: 1, 0.2946816450679251: 1, 0.29142311583585556: 1,
0.284734532173132: 1, 0.2825120475936878: 1, 0.280806393987563: 1, 0.2758896039632867: 1,
0.27494948140171593: 1, 0.2746659520808447: 1, 0.2743401374914483: 1, 0.27383690502862623: 1,
0.25927961177128184: 1, 0.25893451084207064: 1, 0.25796995509911297: 1, 0.257299156917571: 1,
0.25722407822116583: 1, 0.2533711930253906: 1, 0.2485562388311114: 1, 0.24850353973490566: 1,
0.2483935420568943: 1, 0.24801163804574075: 1, 0.24759518328120533: 1, 0.2472552253601171: 1,
0.2471606161219624: 1, 0.24577894030787834: 1, 0.24456794114159056: 1, 0.24436023939098375: 1,
0.2373369780034949: 1, 0.2366034425639258: 1, 0.2365906625007 1806: 1, 0.2357322686840 4287: 1
```

0.23733697800349491: 1, 0.23660342425639258: 1, 0.23659066250071806: 1, 0.23573226868404287: 1,
0.23280439921722532: 1, 0.2321282869278797: 1, 0.23050820756057194: 1, 0.228716048910103l6: 1,
0.2258393117078849: 1, 0.22286656020472478: 1, 0.22055056530840925: 1, 0.21900090606532044: 1,
0.2157088319999621: 1, 0.21538234040139703: 1, 0.21516836428875188: 1, 0.21487206127130057: 1,
0.2130798357293006: 1, 0.2124576463388124: 1, 0.21202106514125666: 1, 0.20998937684736838: 1,
0.20890904024179907: 1, 0.20758598159837185: 1, 0.2049926565158656: 1, 0.204914433947616O2: 1,
0.20303832696379345: 1, 0.19792725290350824: 1, 0.1972302958940711: 1, 0.1958495004618708: 1,
0.19528642458847326: 1, 0.18953304434671703: 1, 0.1870203840529148: 1, 0.18627528272867344: 1,
0.1858568918895609: 1, 0.18503511605340195: 1, 0.18300439139738123: 1, 0.18273075836735103: 1,
0.1815887470225346: 1, 0.17891188856062776: 1, 0.17789753332064842: 1, 0.17600271711099488: 1,
0.17495866258668077: 1, 0.17372742669813895: 1, 0.17270258020407264: 1, 0.17241600542624805: 1,
0.17120696139292793: 1, 0.1707809175754714: 1, 0.17041277833349666: 1, 0.1702040493308769: 1,
0.1686109434372546: 1, 0.16856577513806156: 1, 0.16840326332877634: 1, 0.16704535221799996: 1,
0.16655123425400492: 1, 0.16649327456616622: 1, 0.16626950273157828: 1, 0.16357101168603708: 1,
0.16287563426988438: 1, 0.16195394885706033: 1, 0.16127189241717554: 1, 0.16047053967158512: 1,
0.15912509186293652: 1, 0.15876700955769457: 1, 0.15827855738266972: 1, 0.1573096577489777: 1,
0.15703629902335875: 1, 0.15633634112985062: 1, 0.15620857766167248: 1, 0.1555028657611047: 1,
0.1530216936015552: 1, 0.1517252990147881: 1, 0.15151267883222555: 1, 0.150804903267620581: 1,
0.15074778922993884: 1, 0.1507219056762946: 1, 0.15066561753153623: 1, 0.15008658822785725: 1,
0.14903317525673396: 1, 0.14762051566165799: 1, 0.14495357693442387: 1, 0.1449227575744902: 1,
0.1445198570531253: 1, 0.14299899304487357: 1, 0.14209271489504452: 1, 0.141107274851191821: 1,
0.14057368970244538: 1, 0.14033807436165108: 1, 0.13955887726113658: 1, 0.1390586135560911: 1,
0.1388103118164267: 1, 0.13851501604707162: 1, 0.13843364715187273: 1, 0.1384084930636046: 1,
0.13715388039076487: 1, 0.13683044073561987: 1, 0.13546067770016829: 1, 0.13286171150553824: 1,
0.13273319522953966: 1, 0.13163108404829113: 1, 0.13156578287297319: 1, 0.13028923708685244: 1,
0.12996659767487947: 1, 0.12942654948080445: 1, 0.12880597409018268: 1, 0.12778578076685093: 1,
0.1260353629190163: 1, 0.12599989687037494: 1, 0.12491948769911984: 1, 0.12446999962708785: 1,
0.12431847646592523: 1, 0.12332972977064247: 1, 0.12208114846289217: 1, 0.12180847599042402: 1,
0.12179502840871212: 1, 0.12082921381333825: 1, 0.12056270118955931: 1, 0.12049400048472017: 1,
0.11979769992894174: 1, 0.11860727220426859: 1, 0.11845026753395824: 1, 0.11765937336595071: 1,
0.1175892819160575: 1, 0.11742419016002383: 1, 0.11616282898800917: 1, 0.11593407639377597: 1,
0.11579137789670696: 1, 0.11552264675078594: 1, 0.11536438222917939: 1, 0.11501773358264118: 1,
0.11451184422517789: 1, 0.11374492928905294: 1, 0.1133383152488079: 1, 0.11327789993641676: 1,
0.11234470127191531: 1, 0.11176516963720405: 1, 0.11161062184521621: 1, 0.11157083885065623: 1,
0.11117815499373597: 1, 0.11110842885809606: 1, 0.11035300761632616: 1, 0.10962895944913521: 1,
0.10747844820416788: 1, 0.10723605283360108: 1, 0.10702681150875984: 1, 0.10672872598916974: 1,
0.10632712764956581: 1, 0.10601571556791575: 1, 0.10536490244791889: 1, 0.10492800728580917: 1,
0.10481784565181534: 1, 0.10375878207625874: 1, 0.10328772133372108: 1, 0.10052335587214126: 1,
0.10047021573823037: 1, 0.1003496256335755: 1, 0.10011177398343966: 1, 0.09975698852529218: 1,
0.09953019886767892: 1, 0.09902705965469037: 1, 0.09752872482525081: 1, 0.09744992174735288: 1,
0.09732200650677685: 1, 0.09725065793298611: 1, 0.09668970185143184: 1, 0.09653937290944675: 1,
0.09647482386837847: 1, 0.09612244503470929: 1, 0.09570462503951097: 1, 0.09511226115040394: 1,
0.09474287761753594: 1, 0.09473968187556298: 1, 0.09416419349621725: 1, 0.09409568437583192: 1,
0.09405158096016104: 1, 0.09334091147931292: 1, 0.09317626430359582: 1, 0.09300494805639128: 1,
0.0924704213799608: 1, 0.09224664811172587: 1, 0.09221102415355191: 1, 0.09220382032565201: 1,
0.09143674264127122: 1, 0.09142693978609143: 1, 0.09140690857508933: 1, 0.09134040944715882: 1,
0.09123091555937443: 1, 0.09122291141963625: 1, 0.09110716684186732: 1, 0.0910003806737879: 1,
0.09057560005531576: 1, 0.08978064067125929: 1, 0.08976236598294678: 1, 0.08962442044124817: 1,
0.08937484438295287: 1, 0.08917665704772632: 1, 0.08868096797617792: 1, 0.08853553381311423: 1,
0.08789999217325564: 1, 0.08778460153644253: 1, 0.08776056735929495: 1, 0.0872590275192522: 1,
0.08723571922267484: 1, 0.08683829165105517: 1, 0.08676936883416361: 1, 0.0860080160786578: 1,
0.08579297088247413: 1, 0.08493796815286914: 1, 0.08490595283857108: 1, 0.08430749864438046: 1,
0.08366229073570049: 1, 0.08341643485839186: 1, 0.08286263099765898: 1, 0.08283802165109048: 1,
0.08276896843688826: 1, 0.08244148608062171: 1, 0.08220856681008992: 1, 0.08217627713014818: 1,
0.08211683945956291: 1, 0.08195807484375214: 1, 0.08151475908007637: 1, 0.08113341606381888: 1,
0.08105415408996851: 1, 0.08100369233509475: 1, 0.08095684318452062: 1, 0.08074890139974068: 1,
0.08044383267891758: 1, 0.07981755265518818: 1, 0.07928391713403149: 1, 0.07913486887992488: 1,
0.07897727103732893: 1, 0.07876966858237615: 1, 0.07850829210608612: 1, 0.07849865138443618: 1,
0.07842614349325963: 1, 0.07810258388892402: 1, 0.07803433025994956: 1, 0.0779811809106243: 1,
0.0778310068329641: 1, 0.07769456014518017: 1, 0.07765771386793253: 1, 0.07724438808359345: 1,
0.07707695100042597: 1, 0.07699846773851132: 1, 0.07661508344444266: 1, 0.07638160929463535: 1,
0.07632584417741252: 1, 0.07606501163424806: 1, 0.07542405228589166: 1, 0.0752807207680466: 1,
0.07506215898053885: 1, 0.07456902503528906: 1, 0.07455950771133044: 1, 0.07451011309146938: 1,
0.07374321625496283: 1, 0.07308205505317869: 1, 0.07297049806220199: 1, 0.07286223408032885: 1,
0.0726465626995079: 1, 0.07256573776879117: 1, 0.07215776327963831: 1, 0.07198619261841786: 1,
0.07181565944355674: 1, 0.07173109505549191: 1, 0.0716149413429744: 1, 0.0715959202488276: 1,
0.07036308139487837: 1, 0.07005522741901424: 1, 0.07005189785743259: 1, 0.0699646388566616: 1,
0.06929841868534549: 1, 0.06865870413171829: 1, 0.06824899753546684: 1, 0.06790179093656264: 1,
0.06781861552275353: 1, 0.06735728009359992: 1, 0.06685228436876499: 1, 0.06666908642794409: 1,
0.06626640227872907: 1, 0.0660184548752471: 1, 0.0659167326170776: 1, 0.06440162089888808: 1,
0.0641228504107781: 1, 0.06309841013098297: 1, 0.062426498076741256: 1, 0.06227919878143356: 1,
0.062150423924939836: 1, 0.062132234799808006: 1, 0.061232529587123524: 1, 0.0607670016753241: 1,
0.060542818086425094: 1, 0.06034739277768741: 1, 0.06030085599272969: 1, 0.06002285649336385: 1,
0.05912363663798351: 1, 0.05906025286511385: 1, 0.059035399407024916: 1, 0.05899887596738334: 1, 0
.058829686682975356: 1, 0.05783877669606921: 1, 0.05770422106392882: 1, 0.0571836793029792: 1,
0.05655994429081651: 1, 0.05647573478276042: 1, 0.0562400368114399: 1, 0.05595255796152597: 1,
0.05591422391110538: 1, 0.05568955375716220 4: 1, 0.05563561881780189: 1, 0.055566717955764566: 1,

0.05514371501895754: 1, 0.055058446157794 6: 1, 0.05497489510207536: 1, 0.05488494820859598: 1,
0.05456686209247832: 1, 0.05427130607112963: 1, 0.05422795775629401: 1, 0.05415318694015095: 1,
0.05409347264209083: 1, 0.05385534483762704: 1, 0.053411417508106726: 1, 0.053311253483365004: 1,
0.05327938227063337: 1, 0.053139802340130604: 1, 0.05292750246279231: 1, 0.05282075188608721: 1, 0
.05219385077667749: 1, 0.052190623962004835: 1, 0.05211215039405402: 1, 0.05186613495784867: 1,
0.05166763604769521: 1, 0.051667131681578625: 1, 0.05164885409930964: 1, 0.05146400961428714: 1, 0
.05140197342398005: 1, 0.050837910366394366: 1, 0.05083271939481851: 1, 0.050052444832315385: 1, 0
.05002557307518568: 1, 0.04958206581943702: 1, 0.04958200877143805: 1, 0.04933072632312338: 1,
0.04929758792035216 6: 1, 0.04924925296572147: 1, 0.04910040829469594: 1, 0.04902996089412871: 1,
0.04892508306915975 5: 1, 0.048913995106530946: 1, 0.04887985276765447 6: 1, 0.04882373292139215: 1,
0.04870794630765675 6: 1, 0.04856389435254343 5: 1, 0.04836688806583705 5: 1, 0.04800014710731826 5: 1
, 0.04794466873064035: 1, 0.047920642032056: 1, 0.04787998160830883: 1, 0.04783149962056617: 1,
0.04777210401496236: 1, 0.04757313139066316: 1, 0.047465743672882366: 1, 0.0473563863058122: 1, 0.
04726774387470321: 1, 0.047117030926372365: 1, 0.04707666043753422: 1, 0.04697579989819141: 1,
0.04692144964442561 4: 1, 0.04688204162274986 5: 1, 0.04641744685304585: 1, 0.04629391448997296 3: 1,
0.04618488652719028 6: 1, 0.04595170563221798: 1, 0.04589193089615893 6: 1, 0.04564112219137302: 1,
0.04559676711416559: 1, 0.0455713113768784: 1, 0.04554077365751159: 1, 0.045513602725401794: 1, 0.
0452794882356915: 1, 0.045106812751048234: 1, 0.045048576230454926: 1, 0.044818980484022206: 1, 0.
04402636877611 97: 1, 0.043937018077952895: 1, 0.0437306098667593: 1, 0.043474908755550365: 1, 0.04
3102202813013797: 1, 0.04309341508511187: 1, 0.043075663294609344: 1, 0.04294620551655883: 1,
0.04285859524631877: 1, 0.04277054043983691: 1, 0.042510934712533206: 1, 0.04250865343636025: 1, 0.
04183578679839856: 1, 0.04158532754238364: 1, 0.04151683377314458: 1, 0.04106890554353136: 1,
0.04105841972978145 6: 1, 0.040953213192398606: 1, 0.040856359217228994: 1, 0.040788633320609606: 1
, 0.040707062380482495: 1, 0.040564454272459013: 1, 0.04023159518512494: 1, 0.04016758596506487: 1
, 0.03995847642800337: 1, 0.0395988175387108: 1, 0.03922448705342252: 1, 0.03910644215620991: 1,
0.03907105539258468: 1, 0.03828540362740261: 1, 0.03808260659547063: 1, 0.03792559177732808: 1,
0.03781407197161881: 1, 0.03768694730748471: 1, 0.03767468443848556: 1, 0.03758217138035813: 1,
0.03720855458553787: 1, 0.036933181613130646: 1, 0.03672007616188774: 1, 0.03665121340082197: 1, 0
.03652044360031685: 1, 0.03642604552555563: 1, 0.03632682978345138: 1, 0.03594798622807317: 1,
0.03589411439739432 4: 1, 0.03588109465535849: 1, 0.0354307061969036: 1, 0.03537963604047291: 1,
0.03499308263193869: 1, 0.034962804134511624: 1, 0.03470950895214318: 1, 0.034508769388343685: 1,
0.03448771603555475 5: 1, 0.034031206398725006: 1, 0.03400019550287834: 1, 0.03392577595260724: 1,
0.03381996025078901: 1, 0.03363413488601709: 1, 0.033486066621028064: 1, 0.03342451035685173: 1, 0
.03294904910258872: 1, 0.03284354154592801: 1, 0.03279182390685831: 1, 0.032716023206563874: 1,
0.03259545303834488: 1, 0.032575642920659134: 1, 0.03231034285042782: 1, 0.03216167375930497: 1, 0
.03207983128065431: 1, 0.03201162364978366: 1, 0.03186660400531431: 1, 0.03183155908336008: 1,
0.03171401965561704: 1, 0.0314743729971579: 1, 0.031229871809625033: 1, 0.03119832749259576: 1, 0.
03116155482925876 7: 1, 0.03102837414924212: 1, 0.030800806560567016: 1, 0.030783425652486714: 1, 0
.03074483557457141: 1, 0.03008676275832728: 1, 0.030017861338024907: 1, 0.030012760344470427: 1, 0
.02990677864649264: 1, 0.029512008186402157: 1, 0.02949848757102523: 1, 0.0293881273486861: 1,
0.02938569576046086 4: 1, 0.029316497073177013: 1, 0.029135174050834864: 1, 0.029096890062135883: 1
, 0.02909505853502635: 1, 0.02908699440468744: 1, 0.02887558265183919: 1, 0.02887410773174913: 1,
0.02875244761505268 3: 1, 0.028707032974651735: 1, 0.028685056395954596: 1, 0.028666268410637578: 1
, 0.028651958272255675: 1, 0.028633619554893595: 1, 0.02838096036800437: 1, 0.02823926270553535: 1
, 0.02813187089508442: 1, 0.02808188338905778: 1, 0.0280608495326868: 1, 0.028029654857262994: 1,
0.02786752280250002 4: 1, 0.027703003209414323: 1, 0.027684770528197004: 1, 0.027570498363693564: 1
, 0.0273513889134136: 1, 0.02728080364607455: 1, 0.027161271968925184: 1, 0.027071836272759518: 1,
0.02700424267483086 6: 1, 0.0269876939964225: 1, 0.02685639248056206: 1, 0.026717043635861366: 1, 0
.02665009878596662: 1, 0.026576632619457162: 1, 0.026489682414224842: 1, 0.026420498706374362: 1,
0.02640866005109584: 1, 0.026395966748484385: 1, 0.02632409140106775 6: 1, 0.0262342220281764: 1, 0
.02622995491291055 6: 1, 0.026057595086292136: 1, 0.026034233115160448: 1, 0.025833818023847606: 1,
0.02579868152761039 7: 1, 0.025796998468367933: 1, 0.02575595948201237: 1, 0.025255719516876537: 1,
0.02505120234954385: 1, 0.025045949542171028: 1, 0.02489644347857038: 1, 0.024828179173348446: 1,
0.02482347277201856: 1, 0.024604243072700663: 1, 0.02444029983913526: 1, 0.02412194928935439: 1, 0
.02409385295966411: 1, 0.023970891407614915: 1, 0.023934406988771688: 1, 0.023894478786823926: 1,
0.02381454501120609 8: 1, 0.023662911762129: 1, 0.023453013337364 01: 1, 0.023400379688179564: 1,
0.02339006147083737: 1, 0.023271880147566 67: 1, 0.023176797947938: 1, 0.023077280791395278: 1, 0.
02299619120515294 2: 1, 0.022782942911339286: 1, 0.022769982457813: 1, 0.022701323923803668: 1,
0.02263143962979933 4: 1, 0.022599405750293795: 1, 0.02241447478317012: 1, 0.021954002272044398: 1,
0.02175260226341811 7: 1, 0.021370433926621156: 1, 0.02135509197424148: 1, 0.02127705721032487: 1,
0.02094217620754762 5: 1, 0.020898916039441058: 1, 0.02079266377119182: 1, 0.020637051044325675: 1,
0.02060873192659823 4: 1, 0.02060622439716793: 1, 0.0205362876588248: 1, 0.02037836119931498: 1,
0.02034948855375850 7: 1, 0.020340924694022655: 1, 0.020312684380292274: 1, 0.020229847280670062: 1
, 0.020223804652288985: 1, 0.019558573625208596: 1, 0.019540833652641386: 1, 0.019447525863613865:
1, 0.019420475469073203: 1, 0.01938331890654901: 1, 0.01932369089279939: 1, 0.019174676959356582:
1, 0.01902494999807025: 1, 0.018605497742259065: 1, 0.018582337053612032: 1, 0.018509481204625974:
1, 0.01845922766385331: 1, 0.0183868731430728: 1, 0.018247224545628916: 1, 0.01823566097580016: 1,
0.01816339635061947: 1, 0.01806359939337258: 1, 0.018003049112048607: 1, 0.017648111584755953: 1,
0.01750827498685431 4: 1, 0.0173959797384234: 1, 0.01724972649106228: 1, 0.017197980555781506: 1, 0
.01717321333199267: 1, 0.01698373511777538: 1, 0.016762949906347562: 1, 0.01665774520287551: 1,
0.01660813582513942 5: 1, 0.016339872157523267: 1, 0.016181815913298404: 1, 0.016122815230490733: 1
, 0.015934249603094128: 1, 0.01567427939593816: 1, 0.015655575976997578: 1, 0.015270503141831739:
1, 0.015199802806232132: 1, 0.015134539542915808: 1, 0.014900269199843866: 1,
0.01483749777847576 8: 1, 0.014811967261893565: 1, 0.014209950595539412: 1, 0.013904686141038474: 1
, 0.013875375402527682: 1, 0.013838944591048192: 1, 0.013706944701421442: 1, 0.01353404770813005:
1, 0.01342368079913377: 1, 0.013209222911430719: 1, 0.013168547361160902: 1, 0.012948337157569394:
1, 0.012916909011923803: 1, 0.012839864721988758: 1, 0.012833054765557048: 1,

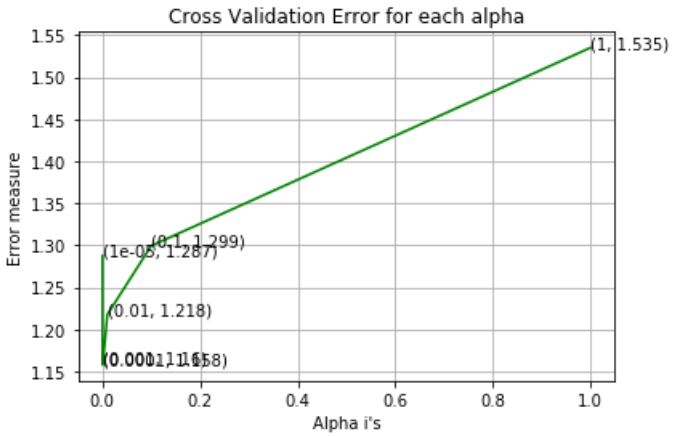
0.012784150717873944: 1, 0.012416020593028286: 1, 0.012395502192859512: 1, 0.012233171428651789: 1
, 0.012046926479832056: 1, 0.011913695252820919: 1, 0.011767598777478983: 1, 0.011663365485919161:
1, 0.011572463211585194: 1, 0.011276921340774581: 1, 0.01124471660549813: 1, 0.011239494184071878:
1, 0.01123861194458317: 1, 0.01117971338897192: 1, 0.01063960624938321: 1, 0.010551718579082853: 1
, 0.010495476337177076: 1, 0.010338960932531324: 1, 0.010216300607101918: 1, 0.009978250046992657:
1, 0.009255346980735824: 1, 0.00895147385324949: 1, 0.007872707983647273: 1, 0.007511195073459762:
1, 0.007409421835076895: 1, 0.006988328739624248: 1, 0.006455539222712627: 1,
0.006439000265004140 4: 1, 0.005615438052827567: 1, 0.005432475040410829: 1})

In [55]:

```
# Train a Logistic regression+Calibration model using text features whicha re on-hot encoded
alpha = [10 ** x for x in range(-5, 1)]

# read more about SGDClassifier() at http://scikit-
learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_i
ter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0
=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, …]) Fit linear model with Stochastic Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link:
#-------------------------------


cv_log_error_array=[]
for i in alpha:
    clf = SGDClassifier(alpha=i, penalty='l2', loss='log', random_state=42)
    clf.fit(train_text_feature_onehotCoding, y_train)

    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_text_feature_onehotCoding, y_train)
    predict_y = sig_clf.predict_proba(cv_text_feature_onehotCoding)
    cv_log_error_array.append(log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
    print('For values of alpha = ', i, "The log loss is:",log_loss(y_cv, predict_y, labels=clf.clas
ses_, eps=1e-15))

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_text_feature_onehotCoding, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_text_feature_onehotCoding, y_train)

predict_y = sig_clf.predict_proba(train_text_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train,
predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_text_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_text_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
For values of alpha =  1e-05 The log loss is: 1.2874760994060528
For values of alpha =  0.0001 The log loss is: 1.1576333991816496
For values of alpha =  0.001 The log loss is: 1.1598307015096385
For values of alpha =  0.01 The log loss is: 1.2175346271743057
```

```
For values of alpha =  0.1 The log loss is: 1.2994051256827142
For values of alpha =  1 The log loss is: 1.5350983698530376
```




```
For values of best alpha =  0.0001 The train log loss is: 0.6644156062524178
For values of best alpha =  0.0001 The cross validation log loss is: 1.1576333991816496
For values of best alpha =  0.0001 The test log loss is: 1.0495859922955166
```

In [56]:

```
def get_intersec_text(df):
    df_text_vec = TfidfVectorizer(min_df=3)
    df_text_fea = df_text_vec.fit_transform(df['TEXT'])

    df_text_features = top_mean_feats(df_text_fea,
                                      df_text_vec.get_feature_names(),
                                      top_n=1000)['feature'].tolist()

    df_text_fea_counts = df_text_fea.sum(axis=0).A1
    df_text_fea_dict = dict(zip(list(df_text_features),df_text_fea_counts))
    len1 = len(set(df_text_features))
    len2 = len(set(train_text_features) & set(df_text_features))
    return len1,len2
```

In [57]:

```
len1,len2 = get_intersec_text(x_test)
print(np.round((len2/len1)*100, 3), "% of word of test data appeared in train data")
len1,len2 = get_intersec_text(x_cv)
print(np.round((len2/len1)*100, 3), "% of word of Cross Validation appeared in train data")
```

```
62.1 % of word of test data appeared in train data
58.7 % of word of Cross Validation appeared in train data
```

Machine Learning Models

In [59]:

```
#Data preparation for ML models.

#Misc. functionns for ML models

def predict_and_plot_confusion_matrix(train_x, train_y,test_x, test_y, clf):
    clf.fit(train_x, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x, train_y)
    pred_y = sig_clf.predict(test_x)

    # for calculating log_loss we willl provide the array of probabilities belongs to each class
    print("Log loss :",log_loss(test_y, sig_clf.predict_proba(test_x)))
    # calculating the number of data points that are misclassified
    print("Number of mis-classified points :", np.count_nonzero((pred_y- test_y))/test_y.shape[0])
    plot_confusion_matrix(test_y, pred_y)
```

In [59]:

```
def report_log_loss(train_x, train_y, test_x, test_y,  clf):
    clf.fit(train_x, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x, train_y)
    sig_clf_probs = sig_clf.predict_proba(test_x)
    return log_loss(test_y, sig_clf_probs, eps=1e-15)
```

In [60]:

```
# this function will be used just for naive bayes
# for the given indices, we will print the name of the features
# and we will check whether the feature present in the test point text or not
def get_impfeature_names(indices, text, gene, var, no_features):
    gene_count_vec = TfidfVectorizer()
    var_count_vec = TfidfVectorizer()
    text_count_vec = TfidfVectorizer(min_df=3)

    gene_vec = gene_count_vec.fit(x_train['Gene'])
    var_vec  = var_count_vec.fit(x_train['Variation'])
    text_vec = text_count_vec.fit(x_train['TEXT'])

    fea1_len = len(gene_vec.get_feature_names())
    fea2_len = len(var_count_vec.get_feature_names())

    word_present = 0
    for i,v in enumerate(indices):
        if (v < fea1_len):
            word = gene_vec.get_feature_names()[v]
            yes_no = True if word == gene else False
            if yes_no:
                word_present += 1
                print(i, "Gene feature [{}] present in test data point [{}]".format(word,yes_no))
        elif (v < fea1_len+fea2_len):
            word = var_vec.get_feature_names()[v-(fea1_len)]
            yes_no = True if word == var else False
            if yes_no:
                word_present += 1
                print(i, "variation feature [{}] present in test data point [{}]".format(word,yes_r
o))
        else:
            word = text_vec.get_feature_names()[v-(fea1_len+fea2_len)]
            yes_no = True if word in text.split() else False
            if yes_no:
                word_present += 1
                print(i, "Text feature [{}] present in test data point [{}]".format(word,yes_no))

    print("Out of the top ",no_features," features ", word_present, "are present in query point")
```

In [61]:

```
# merging gene, variance and text features

# building train, test and cross validation data sets
# a = [[1, 2],
#      [3, 4]]
# b = [[4, 5],
#      [6, 7]]
# hstack(a, b) = [[1, 2, 4, 5],
#                 [ 3, 4, 6, 7]]

train_gene_var_onehotCoding =
hstack((train_gene_feature_onehotCoding,train_variation_feature_onehotCoding))
test_gene_var_onehotCoding =
hstack((test_gene_feature_onehotCoding,test_variation_feature_onehotCoding))
cv_gene_var_onehotCoding = hstack((cv_gene_feature_onehotCoding,cv_variation_feature_onehotCoding)
)

train_x_onehotCoding = hstack((train_gene_var_onehotCoding, train_text_feature_onehotCoding)).tocs
r()
train_y = np.array(list(y_train['Class']))

test_x_onehotCoding = hstack((test_gene_var_onehotCoding, test_text_feature_onehotCoding)).tocsr()
```

```
test_x_onehotCoding = hstack((test_gene_var_onehotCoding, test_text_feature_onehotCoding)).tocsr()
test_y = np.array(list(y_test['Class']))

cv_x_onehotCoding = hstack((cv_gene_var_onehotCoding, cv_text_feature_onehotCoding)).tocsr()
cv_y = np.array(list(y_cv['Class']))


train_gene_var_responseCoding =
np.hstack((train_gene_feature_responseCoding,train_variation_feature_responseCoding))
test_gene_var_responseCoding =
np.hstack((test_gene_feature_responseCoding,test_variation_feature_responseCoding))
cv_gene_var_responseCoding =
np.hstack((cv_gene_feature_responseCoding,cv_variation_feature_responseCoding))

train_x_responseCoding = np.hstack((train_gene_var_responseCoding,
train_text_feature_responseCoding))
test_x_responseCoding = np.hstack((test_gene_var_responseCoding, test_text_feature_responseCoding)
)
cv_x_responseCoding = np.hstack((cv_gene_var_responseCoding, cv_text_feature_responseCoding))
```

In [62]:

```
print("One hot encoding features :")
print("(number of data points * number of features) in train data = ", train_x_onehotCoding.shape)
print("(number of data points * number of features) in test data = ", test_x_onehotCoding.shape)
print("(number of data points * number of features) in cross validation data =", cv_x_onehotCoding
.shape)
```

```
One hot encoding features :
(number of data points * number of features) in train data =  (2124, 55006)
(number of data points * number of features) in test data =  (665, 55006)
(number of data points * number of features) in cross validation data = (532, 55006)
```

In [63]:

```
print(" Response encoding features :")
print("(number of data points * number of features) in train data = ", train_x_responseCoding.shap
e)
print("(number of data points * number of features) in test data = ", test_x_responseCoding.shape)
print("(number of data points * number of features) in cross validation data =",
cv_x_responseCoding.shape)
```

```
 Response encoding features :
(number of data points * number of features) in train data =  (2124, 27)
(number of data points * number of features) in test data =  (665, 27)
(number of data points * number of features) in cross validation data = (532, 27)
```

BaseLine Model(Naive Bayes)

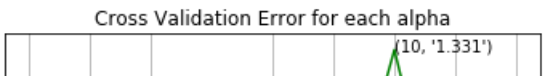
Hyperparameter Tuning

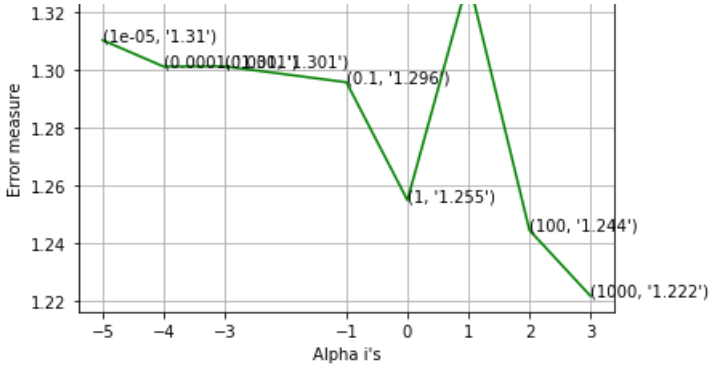
In [64]:

```
# find more about Multinomial Naive base function here http://scikit-
learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html
# -------------------------
# default paramters
# sklearn.naive_bayes.MultinomialNB(alpha=1.0, fit_prior=True, class_prior=None)

# some of methods of MultinomialNB()
# fit(X, y[, sample_weight]) Fit Naive Bayes classifier according to X, y
# predict(X) Perform classification on an array of test vectors X.
# predict_log_proba(X) Return log-probability estimates for the test vector X.
# -----------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/naive-bayes-
algorithm-1/
# -----------------------


# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
```

```
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
# ----------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/naive-bayes-algorithm-1/
# ------------------------


alpha = [0.00001, 0.0001, 0.001, 0.1, 1, 10, 100,1000]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = MultinomialNB(alpha=i)
    clf.fit(train_x_onehotCoding, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_onehotCoding, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
    cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    # to avoid rounding error while multiplying probabilites we use log-probability estimates
    print("Log Loss :",log_loss(cv_y, sig_clf_probs))

fig, ax = plt.subplots()
ax.plot(np.log10(alpha), cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],str(txt)), (np.log10(alpha[i]),cv_log_error_array[i]))
plt.grid()
plt.xticks(np.log10(alpha))
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
clf = MultinomialNB(alpha=alpha[best_alpha])
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)


predict_y = sig_clf.predict_proba(train_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train,
predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-05
Log Loss : 1.3102155807905453
for alpha = 0.0001
Log Loss : 1.3010604173831077
for alpha = 0.001
Log Loss : 1.3011148791442082
for alpha = 0.1
Log Loss : 1.2955981289530853
for alpha = 1
Log Loss : 1.2547428785482246
for alpha = 10
Log Loss : 1.3314835325730479
for alpha = 100
Log Loss : 1.2443518921483896
for alpha = 1000
Log Loss : 1.221568175176729
```



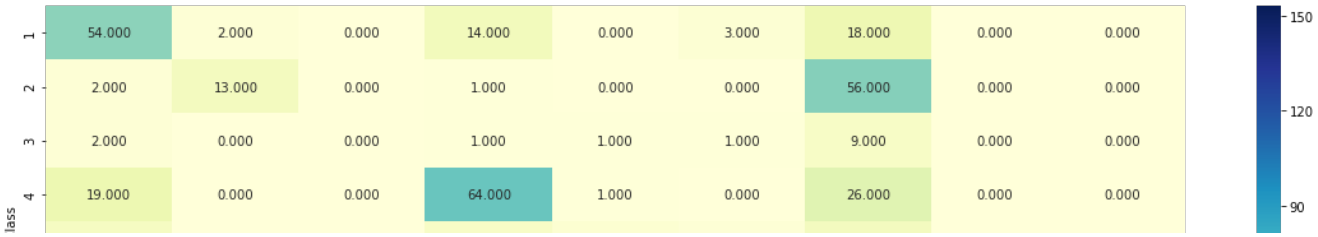


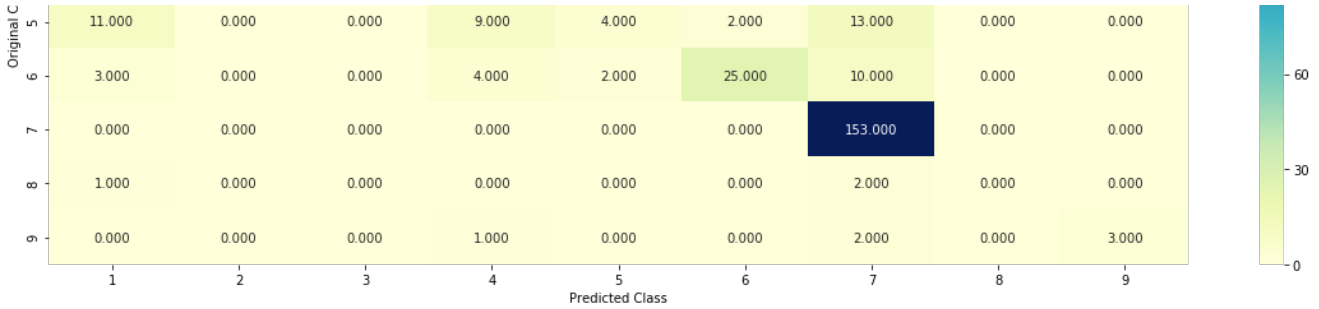

```
For values of best alpha =  1000 The train log loss is: 0.9395398613587067
For values of best alpha =  1000 The cross validation log loss is: 1.221568175176729
For values of best alpha =  1000 The test log loss is: 1.189010329869463
```

In [65]:

```
# find more about Multinomial Naive base function here http://scikit-
learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html
# -------------------------
# default paramters
# sklearn.naive_bayes.MultinomialNB(alpha=1.0, fit_prior=True, class_prior=None)

# some of methods of MultinomialNB()
# fit(X, y[, sample_weight]) Fit Naive Bayes classifier according to X, y
# predict(X) Perform classification on an array of test vectors X.
# predict_log_proba(X) Return log-probability estimates for the test vector X.
# -----------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/naive-bayes-
algorithm-1/
# -----------------------


# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
# ----------------------------

clf = MultinomialNB(alpha=alpha[best_alpha])
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)
sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
# to avoid rounding error while multiplying probabilites we use log-probability estimates
print("Log Loss :",log_loss(cv_y, sig_clf_probs))
print("Number of missclassified point :", np.count_nonzero((sig_clf.predict(cv_x_onehotCoding)- cv
_y))/cv_y.shape[0])
plot_confusion_matrix(cv_y, sig_clf.predict(cv_x_onehotCoding.toarray()))
```

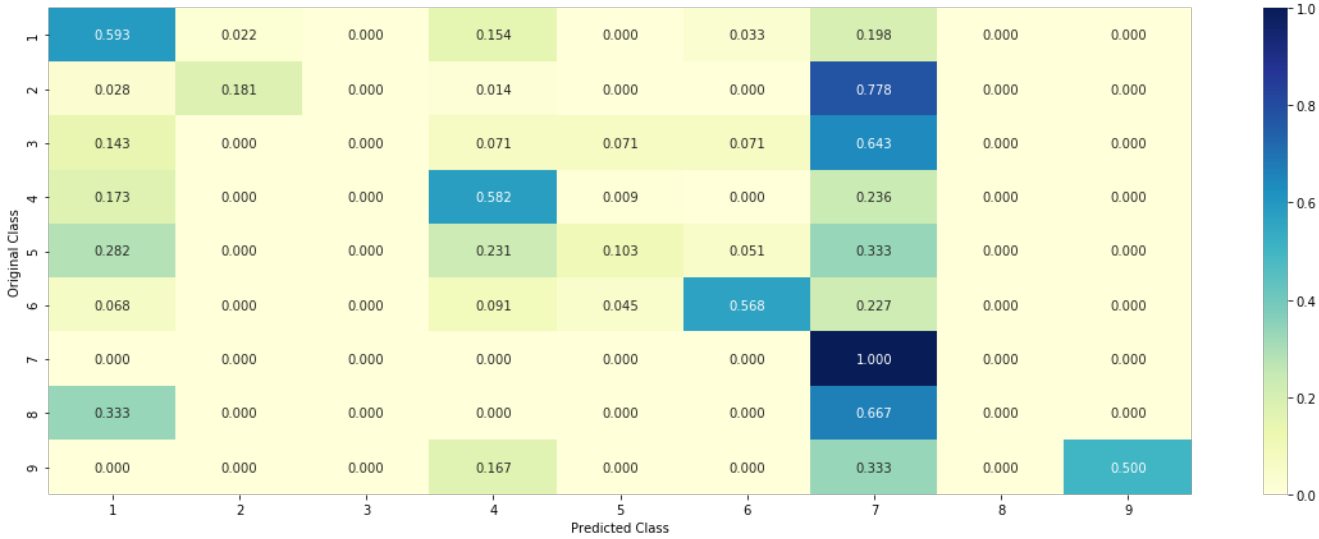
```
Log Loss : 1.221568175176729
Number of missclassified point : 0.40601503759398494

-------------------- Confusion matrix --------------------
```





---------- Precision matrix (Columm Sum=1) ----------

---------- Recall matrix (Row sum=1) ----------



Feature Importance

In [66]:

```
test_point_index = 1
no_feature = 1000
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_index]),4))
print("Actual Class :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
```

```
                    x_test['Variation'].iloc[test_point_index],
                    no_feature)
```

```
Predicted Class : 1
Predicted Class Probabilities: [[0.6539 0.0204 0.0011 0.2357 0.0137 0.0155 0.059  0.0007 0.    ]]
Actual Class : 1
--------------------------------------------------
9 Text feature [protein] present in test data point [True]
10 Text feature [type] present in test data point [True]
13 Text feature [wild] present in test data point [True]
14 Text feature [dna] present in test data point [True]
15 Text feature [one] present in test data point [True]
16 Text feature [therefore] present in test data point [True]
17 Text feature [affect] present in test data point [True]
18 Text feature [results] present in test data point [True]
19 Text feature [two] present in test data point [True]
20 Text feature [binding] present in test data point [True]
21 Text feature [using] present in test data point [True]
22 Text feature [functions] present in test data point [True]
23 Text feature [containing] present in test data point [True]
24 Text feature [either] present in test data point [True]
25 Text feature [also] present in test data point [True]
26 Text feature [control] present in test data point [True]
27 Text feature [table] present in test data point [True]
28 Text feature [possible] present in test data point [True]
29 Text feature [effect] present in test data point [True]
30 Text feature [function] present in test data point [True]
31 Text feature [involved] present in test data point [True]
32 Text feature [following] present in test data point [True]
33 Text feature [region] present in test data point [True]
34 Text feature [determined] present in test data point [True]
35 Text feature [role] present in test data point [True]
36 Text feature [expression] present in test data point [True]
38 Text feature [however] present in test data point [True]
39 Text feature [loss] present in test data point [True]
40 Text feature [specific] present in test data point [True]
41 Text feature [shown] present in test data point [True]
42 Text feature [well] present in test data point [True]
43 Text feature [three] present in test data point [True]
44 Text feature [indicated] present in test data point [True]
45 Text feature [analysis] present in test data point [True]
46 Text feature [present] present in test data point [True]
47 Text feature [ability] present in test data point [True]
48 Text feature [human] present in test data point [True]
49 Text feature [interacts] present in test data point [True]
50 Text feature [gene] present in test data point [True]
51 Text feature [similar] present in test data point [True]
52 Text feature [four] present in test data point [True]
53 Text feature [described] present in test data point [True]
54 Text feature [previously] present in test data point [True]
55 Text feature [amino] present in test data point [True]
59 Text feature [reporter] present in test data point [True]
60 Text feature [result] present in test data point [True]
61 Text feature [25] present in test data point [True]
62 Text feature [may] present in test data point [True]
63 Text feature [complex] present in test data point [True]
64 Text feature [indicate] present in test data point [True]
65 Text feature [10] present in test data point [True]
66 Text feature [transcriptional] present in test data point [True]
67 Text feature [addition] present in test data point [True]
68 Text feature [least] present in test data point [True]
69 Text feature [including] present in test data point [True]
71 Text feature [cancer] present in test data point [True]
72 Text feature [based] present in test data point [True]
73 Text feature [critical] present in test data point [True]
74 Text feature [respectively] present in test data point [True]
75 Text feature [within] present in test data point [True]
76 Text feature [deletion] present in test data point [True]
77 Text feature [proteins] present in test data point [True]
78 Text feature [used] present in test data point [True]
79 Text feature [level] present in test data point [True]
80 Text feature [different] present in test data point [True]
81 Text feature [whereas] present in test data point [True]
82 Text feature [important] present in test data point [True]
83 Text feature [observed] present in test data point [True]
84 Text feature [essential] present in test data point [True]
```

```
85 Text feature [15] present in test data point [True]
86 Text feature [indicating] present in test data point [True]
87 Text feature [previous] present in test data point [True]
88 Text feature [associated] present in test data point [True]
89 Text feature [plays] present in test data point [True]
90 Text feature [significant] present in test data point [True]
91 Text feature [large] present in test data point [True]
93 Text feature [page] present in test data point [True]
94 Text feature [sequences] present in test data point [True]
96 Text feature [contains] present in test data point [True]
98 Text feature [transcription] present in test data point [True]
99 Text feature [according] present in test data point [True]
100 Text feature [many] present in test data point [True]
101 Text feature [terminal] present in test data point [True]
103 Text feature [another] present in test data point [True]
104 Text feature [remaining] present in test data point [True]
105 Text feature [obtained] present in test data point [True]
106 Text feature [levels] present in test data point [True]
107 Text feature [form] present in test data point [True]
108 Text feature [whether] present in test data point [True]
109 Text feature [several] present in test data point [True]
111 Text feature [reveal] present in test data point [True]
112 Text feature [30] present in test data point [True]
113 Text feature [could] present in test data point [True]
114 Text feature [likely] present in test data point [True]
115 Text feature [structure] present in test data point [True]
116 Text feature [conserved] present in test data point [True]
117 Text feature [example] present in test data point [True]
118 Text feature [defined] present in test data point [True]
119 Text feature [data] present in test data point [True]
120 Text feature [together] present in test data point [True]
122 Text feature [length] present in test data point [True]
123 Text feature [compared] present in test data point [True]
125 Text feature [genes] present in test data point [True]
126 Text feature [discussion] present in test data point [True]
127 Text feature [studies] present in test data point [True]
128 Text feature [fraction] present in test data point [True]
129 Text feature [lack] present in test data point [True]
130 Text feature [full] present in test data point [True]
131 Text feature [relative] present in test data point [True]
132 Text feature [domains] present in test data point [True]
133 Text feature [system] present in test data point [True]
134 Text feature [specifically] present in test data point [True]
135 Text feature [although] present in test data point [True]
136 Text feature [thus] present in test data point [True]
137 Text feature [mutation] present in test data point [True]
138 Text feature [fig] present in test data point [True]
140 Text feature [yet] present in test data point [True]
141 Text feature [directly] present in test data point [True]
142 Text feature [eight] present in test data point [True]
143 Text feature [major] present in test data point [True]
144 Text feature [highly] present in test data point [True]
145 Text feature [finally] present in test data point [True]
146 Text feature [sequence] present in test data point [True]
147 Text feature [majority] present in test data point [True]
148 Text feature [domain] present in test data point [True]
149 Text feature [specificity] present in test data point [True]
150 Text feature [high] present in test data point [True]
151 Text feature [methods] present in test data point [True]
153 Text feature [regulation] present in test data point [True]
154 Text feature [analyzed] present in test data point [True]
155 Text feature [16] present in test data point [True]
156 Text feature [remains] present in test data point [True]
157 Text feature [dependent] present in test data point [True]
158 Text feature [37] present in test data point [True]
159 Text feature [furthermore] present in test data point [True]
160 Text feature [indicates] present in test data point [True]
162 Text feature [limited] present in test data point [True]
163 Text feature [32] present in test data point [True]
164 Text feature [contain] present in test data point [True]
165 Text feature [identified] present in test data point [True]
166 Text feature [genetic] present in test data point [True]
167 Text feature [mutant] present in test data point [True]
168 Text feature [additional] present in test data point [True]
169 Text feature [interact] present in test data point [True]
170 Text feature [bind] present in test data point [True]
171 Text feature [sites] present in test data point [True]
```

```
172 Text feature [sufficient] present in test data point [True]
173 Text feature [changes] present in test data point [True]
174 Text feature [single] present in test data point [True]
175 Text feature [central] present in test data point [True]
176 Text feature [among] present in test data point [True]
178 Text feature [figure] present in test data point [True]
179 Text feature [complete] present in test data point [True]
180 Text feature [significantly] present in test data point [True]
181 Text feature [acid] present in test data point [True]
183 Text feature [see] present in test data point [True]
185 Text feature [range] present in test data point [True]
186 Text feature [identify] present in test data point [True]
187 Text feature [six] present in test data point [True]
188 Text feature [increased] present in test data point [True]
189 Text feature [recognizes] present in test data point [True]
192 Text feature [cell] present in test data point [True]
194 Text feature [effects] present in test data point [True]
195 Text feature [lower] present in test data point [True]
196 Text feature [24] present in test data point [True]
197 Text feature [less] present in test data point [True]
198 Text feature [strong] present in test data point [True]
199 Text feature [p53] present in test data point [True]
200 Text feature [presence] present in test data point [True]
202 Text feature [stability] present in test data point [True]
203 Text feature [mean] present in test data point [True]
204 Text feature [displayed] present in test data point [True]
206 Text feature [half] present in test data point [True]
212 Text feature [materials] present in test data point [True]
214 Text feature [cells] present in test data point [True]
215 Text feature [mutations] present in test data point [True]
216 Text feature [wide] present in test data point [True]
217 Text feature [transfection] present in test data point [True]
218 Text feature [amount] present in test data point [True]
220 Text feature [provide] present in test data point [True]
224 Text feature [constructs] present in test data point [True]
225 Text feature [might] present in test data point [True]
226 Text feature [target] present in test data point [True]
227 Text feature [efficiency] present in test data point [True]
228 Text feature [confirmed] present in test data point [True]
229 Text feature [deficient] present in test data point [True]
230 Text feature [serves] present in test data point [True]
231 Text feature [core] present in test data point [True]
232 Text feature [template] present in test data point [True]
233 Text feature [low] present in test data point [True]
234 Text feature [consists] present in test data point [True]
235 Text feature [introduction] present in test data point [True]
236 Text feature [calculated] present in test data point [True]
237 Text feature [consistent] present in test data point [True]
238 Text feature [shows] present in test data point [True]
239 Text feature [found] present in test data point [True]
240 Text feature [study] present in test data point [True]
241 Text feature [fold] present in test data point [True]
242 Text feature [provides] present in test data point [True]
249 Text feature [suggested] present in test data point [True]
251 Text feature [rather] present in test data point [True]
253 Text feature [folding] present in test data point [True]
254 Text feature [structural] present in test data point [True]
255 Text feature [site] present in test data point [True]
256 Text feature [would] present in test data point [True]
257 Text feature [show] present in test data point [True]
258 Text feature [first] present in test data point [True]
261 Text feature [expressed] present in test data point [True]
262 Text feature [change] present in test data point [True]
263 Text feature [surface] present in test data point [True]
266 Text feature [reported] present in test data point [True]
267 Text feature [required] present in test data point [True]
268 Text feature [five] present in test data point [True]
269 Text feature [strongly] present in test data point [True]
270 Text feature [cellular] present in test data point [True]
271 Text feature [functional] present in test data point [True]
272 Text feature [detected] present in test data point [True]
273 Text feature [characterized] present in test data point [True]
274 Text feature [frequency] present in test data point [True]
276 Text feature [29] present in test data point [True]
277 Text feature [80] present in test data point [True]
279 Text feature [100] present in test data point [True]
282 Text feature [negative] present in test data point [True]
```

```
283 Text feature [supplementary] present in test data point [True]
284 Text feature [derived] present in test data point [True]
285 Text feature [repeats] present in test data point [True]
286 Text feature [vitro] present in test data point [True]
287 Text feature [35] present in test data point [True]
288 Text feature [larger] present in test data point [True]
290 Text feature [represent] present in test data point [True]
291 Text feature [co] present in test data point [True]
292 Text feature [translation] present in test data point [True]
293 Text feature [side] present in test data point [True]
295 Text feature [red] present in test data point [True]
296 Text feature [residues] present in test data point [True]
297 Text feature [definition] present in test data point [True]
299 Text feature [fact] present in test data point [True]
302 Text feature [development] present in test data point [True]
303 Text feature [folded] present in test data point [True]
305 Text feature [involves] present in test data point [True]
306 Text feature [promega] present in test data point [True]
308 Text feature [cannot] present in test data point [True]
311 Text feature [expected] present in test data point [True]
312 Text feature [occurs] present in test data point [True]
315 Text feature [small] present in test data point [True]
319 Text feature [considered] present in test data point [True]
320 Text feature [ref] present in test data point [True]
321 Text feature [necessary] present in test data point [True]
323 Text feature [peptide] present in test data point [True]
324 Text feature [50] present in test data point [True]
326 Text feature [absence] present in test data point [True]
327 Text feature [lost] present in test data point [True]
330 Text feature [interestingly] present in test data point [True]
331 Text feature [responsible] present in test data point [True]
333 Text feature [groove] present in test data point [True]
334 Text feature [recently] present in test data point [True]
335 Text feature [40] present in test data point [True]
336 Text feature [destabilize] present in test data point [True]
338 Text feature [unclear] present in test data point [True]
340 Text feature [chain] present in test data point [True]
341 Text feature [interest] present in test data point [True]
343 Text feature [27] present in test data point [True]
344 Text feature [inactivation] present in test data point [True]
345 Text feature [indeed] present in test data point [True]
346 Text feature [life] present in test data point [True]
347 Text feature [contrast] present in test data point [True]
349 Text feature [regions] present in test data point [True]
351 Text feature [little] present in test data point [True]
354 Text feature [correlation] present in test data point [True]
356 Text feature [common] present in test data point [True]
357 Text feature [includes] present in test data point [True]
358 Text feature [cancers] present in test data point [True]
359 Text feature [showed] present in test data point [True]
361 Text feature [without] present in test data point [True]
362 Text feature [mapped] present in test data point [True]
363 Text feature [interactions] present in test data point [True]
364 Text feature [broad] present in test data point [True]
365 Text feature [28] present in test data point [True]
366 Text feature [modulate] present in test data point [True]
367 Text feature [sequencing] present in test data point [True]
368 Text feature [roles] present in test data point [True]
369 Text feature [destabilized] present in test data point [True]
370 Text feature [observation] present in test data point [True]
371 Text feature [represents] present in test data point [True]
372 Text feature [activities] present in test data point [True]
373 Text feature [examined] present in test data point [True]
376 Text feature [fail] present in test data point [True]
378 Text feature [often] present in test data point [True]
379 Text feature [13] present in test data point [True]
380 Text feature [decreased] present in test data point [True]
381 Text feature [plasmid] present in test data point [True]
382 Text feature [underlying] present in test data point [True]
384 Text feature [sequenced] present in test data point [True]
385 Text feature [finding] present in test data point [True]
387 Text feature [relatively] present in test data point [True]
388 Text feature [luciferase] present in test data point [True]
390 Text feature [missense] present in test data point [True]
391 Text feature [controls] present in test data point [True]
392 Text feature [22] present in test data point [True]
393 Text feature [understanding] present in test data point [True]
```

```
396 Text feature [latter] present in test data point [True]
397 Text feature [terminus] present in test data point [True]
398 Text feature [early] present in test data point [True]
404 Text feature [part] present in test data point [True]
405 Text feature [made] present in test data point [True]
408 Text feature [individual] present in test data point [True]
409 Text feature [cause] present in test data point [True]
411 Text feature [cases] present in test data point [True]
412 Text feature [size] present in test data point [True]
414 Text feature [test] present in test data point [True]
416 Text feature [tumor] present in test data point [True]
417 Text feature [overall] present in test data point [True]
419 Text feature [clearly] present in test data point [True]
420 Text feature [order] present in test data point [True]
426 Text feature [repair] present in test data point [True]
427 Text feature [known] present in test data point [True]
430 Text feature [key] present in test data point [True]
432 Text feature [number] present in test data point [True]
433 Text feature [resulting] present in test data point [True]
434 Text feature [reduce] present in test data point [True]
437 Text feature [damage] present in test data point [True]
438 Text feature [close] present in test data point [True]
439 Text feature [signal] present in test data point [True]
440 Text feature [potential] present in test data point [True]
441 Text feature [apparent] present in test data point [True]
443 Text feature [causes] present in test data point [True]
444 Text feature [deletions] present in test data point [True]
445 Text feature [higher] present in test data point [True]
446 Text feature [since] present in test data point [True]
447 Text feature [fragment] present in test data point [True]
449 Text feature [activity] present in test data point [True]
451 Text feature [series] present in test data point [True]
454 Text feature [17] present in test data point [True]
462 Text feature [allows] present in test data point [True]
464 Text feature [act] present in test data point [True]
466 Text feature [include] present in test data point [True]
467 Text feature [largely] present in test data point [True]
468 Text feature [background] present in test data point [True]
471 Text feature [propose] present in test data point [True]
472 Text feature [vector] present in test data point [True]
474 Text feature [upon] present in test data point [True]
476 Text feature [23] present in test data point [True]
477 Text feature [still] present in test data point [True]
478 Text feature [gel] present in test data point [True]
482 Text feature [quantitatively] present in test data point [True]
486 Text feature [assay] present in test data point [True]
489 Text feature [types] present in test data point [True]
492 Text feature [view] present in test data point [True]
494 Text feature [confirm] present in test data point [True]
495 Text feature [thought] present in test data point [True]
496 Text feature [21] present in test data point [True]
497 Text feature [normal] present in test data point [True]
498 Text feature [correct] present in test data point [True]
502 Text feature [disrupt] present in test data point [True]
503 Text feature [2001] present in test data point [True]
506 Text feature [nuclear] present in test data point [True]
509 Text feature [germ] present in test data point [True]
510 Text feature [analyses] present in test data point [True]
512 Text feature [experiments] present in test data point [True]
516 Text feature [measured] present in test data point [True]
517 Text feature [new] present in test data point [True]
518 Text feature [conversely] present in test data point [True]
521 Text feature [transfected] present in test data point [True]
523 Text feature [like] present in test data point [True]
524 Text feature [equal] present in test data point [True]
526 Text feature [nearly] present in test data point [True]
528 Text feature [method] present in test data point [True]
530 Text feature [predict] present in test data point [True]
532 Text feature [phenotype] present in test data point [True]
534 Text feature [cotransfected] present in test data point [True]
536 Text feature [lead] present in test data point [True]
538 Text feature [make] present in test data point [True]
541 Text feature [useful] present in test data point [True]
543 Text feature [modifications] present in test data point [True]
544 Text feature [approximately] present in test data point [True]
545 Text feature [independent] present in test data point [True]
546 Text feature [distinct] present in test data point [True]
```

```
549 Text feature [aggregate] present in test data point [True]
551 Text feature [regulatory] present in test data point [True]
553 Text feature [functionally] present in test data point [True]
554 Text feature [variants] present in test data point [True]
556 Text feature [onto] present in test data point [True]
557 Text feature [relationship] present in test data point [True]
559 Text feature [fall] present in test data point [True]
565 Text feature [44] present in test data point [True]
567 Text feature [value] present in test data point [True]
569 Text feature [efficiently] present in test data point [True]
570 Text feature [incubated] present in test data point [True]
571 Text feature [stress] present in test data point [True]
573 Text feature [26] present in test data point [True]
575 Text feature [34] present in test data point [True]
576 Text feature [work] present in test data point [True]
578 Text feature [importance] present in test data point [True]
580 Text feature [suppressors] present in test data point [True]
581 Text feature [complexes] present in test data point [True]
583 Text feature [program] present in test data point [True]
586 Text feature [construct] present in test data point [True]
587 Text feature [address] present in test data point [True]
588 Text feature [evaluated] present in test data point [True]
590 Text feature [42] present in test data point [True]
591 Text feature [potentially] present in test data point [True]
592 Text feature [tetramerization] present in test data point [True]
594 Text feature [observations] present in test data point [True]
595 Text feature [identical] present in test data point [True]
596 Text feature [nucleus] present in test data point [True]
599 Text feature [precise] present in test data point [True]
600 Text feature [antibody] present in test data point [True]
602 Text feature [induced] present in test data point [True]
603 Text feature [depending] present in test data point [True]
604 Text feature [flanking] present in test data point [True]
607 Text feature [associate] present in test data point [True]
608 Text feature [recognize] present in test data point [True]
609 Text feature [future] present in test data point [True]
610 Text feature [20] present in test data point [True]
611 Text feature [05] present in test data point [True]
612 Text feature [interfere] present in test data point [True]
613 Text feature [far] present in test data point [True]
614 Text feature [reduction] present in test data point [True]
616 Text feature [residue] present in test data point [True]
617 Text feature [groups] present in test data point [True]
618 Text feature [much] present in test data point [True]
619 Text feature [lysate] present in test data point [True]
620 Text feature [demonstrated] present in test data point [True]
623 Text feature [recent] present in test data point [True]
627 Text feature [hypothesis] present in test data point [True]
629 Text feature [powerpoint] present in test data point [True]
632 Text feature [factors] present in test data point [True]
634 Text feature [apoptosis] present in test data point [True]
635 Text feature [hundred] present in test data point [True]
636 Text feature [quantitative] present in test data point [True]
651 Text feature [develop] present in test data point [True]
652 Text feature [biological] present in test data point [True]
658 Text feature [recombinant] present in test data point [True]
659 Text feature [selected] present in test data point [True]
660 Text feature [70] present in test data point [True]
667 Text feature [thereby] present in test data point [True]
668 Text feature [measurements] present in test data point [True]
669 Text feature [11] present in test data point [True]
670 Text feature [manner] present in test data point [True]
671 Text feature [reverse] present in test data point [True]
673 Text feature [18] present in test data point [True]
676 Text feature [similarly] present in test data point [True]
678 Text feature [genotoxic] present in test data point [True]
679 Text feature [noted] present in test data point [True]
681 Text feature [defines] present in test data point [True]
682 Text feature [depend] present in test data point [True]
683 Text feature [nine] present in test data point [True]
684 Text feature [classes] present in test data point [True]
687 Text feature [plotted] present in test data point [True]
688 Text feature [evidence] present in test data point [True]
689 Text feature [framework] present in test data point [True]
692 Text feature [proline] present in test data point [True]
699 Text feature [moderately] present in test data point [True]
700 Text feature [hydrophobic] present in test data point [True]
```

```
702 Text feature [database] present in test data point [True]
703 Text feature [plasmids] present in test data point [True]
708 Text feature [difficult] present in test data point [True]
710 Text feature [perhaps] present in test data point [True]
711 Text feature [values] present in test data point [True]
712 Text feature [binds] present in test data point [True]
715 Text feature [contained] present in test data point [True]
717 Text feature [showing] present in test data point [True]
719 Text feature [53] present in test data point [True]
721 Text feature [occur] present in test data point [True]
722 Text feature [us] present in test data point [True]
724 Text feature [94] present in test data point [True]
727 Text feature [image] present in test data point [True]
728 Text feature [able] present in test data point [True]
730 Text feature [increase] present in test data point [True]
731 Text feature [account] present in test data point [True]
733 Text feature [targets] present in test data point [True]
737 Text feature [family] present in test data point [True]
739 Text feature [14] present in test data point [True]
740 Text feature [native] present in test data point [True]
741 Text feature [defects] present in test data point [True]
743 Text feature [crystallographic] present in test data point [True]
745 Text feature [12] present in test data point [True]
748 Text feature [induce] present in test data point [True]
750 Text feature [sds] present in test data point [True]
751 Text feature [4c] present in test data point [True]
753 Text feature [48] present in test data point [True]
754 Text feature [times] present in test data point [True]
755 Text feature [novel] present in test data point [True]
756 Text feature [nucleotide] present in test data point [True]
761 Text feature [hypothesized] present in test data point [True]
763 Text feature [sharing] present in test data point [True]
765 Text feature [contact] present in test data point [True]
766 Text feature [mouse] present in test data point [True]
767 Text feature [room] present in test data point [True]
768 Text feature [adopts] present in test data point [True]
771 Text feature [screening] present in test data point [True]
774 Text feature [immobilized] present in test data point [True]
778 Text feature [difference] present in test data point [True]
780 Text feature [identification] present in test data point [True]
783 Text feature [model] present in test data point [True]
786 Text feature [stable] present in test data point [True]
787 Text feature [regulating] present in test data point [True]
792 Text feature [inactivates] present in test data point [True]
794 Text feature [blue] present in test data point [True]
799 Text feature [play] present in test data point [True]
801 Text feature [positive] present in test data point [True]
802 Text feature [sought] present in test data point [True]
804 Text feature [explaining] present in test data point [True]
805 Text feature [interacting] present in test data point [True]
806 Text feature [evident] present in test data point [True]
808 Text feature [qiagen] present in test data point [True]
809 Text feature [strands] present in test data point [True]
811 Text feature [differences] present in test data point [True]
820 Text feature [raised] present in test data point [True]
825 Text feature [pbs] present in test data point [True]
826 Text feature [distribution] present in test data point [True]
827 Text feature [peptides] present in test data point [True]
829 Text feature [surrounding] present in test data point [True]
831 Text feature [probably] present in test data point [True]
834 Text feature [diverse] present in test data point [True]
835 Text feature [aggregation] present in test data point [True]
837 Text feature [trypsin] present in test data point [True]
839 Text feature [severely] present in test data point [True]
840 Text feature [ml] present in test data point [True]
843 Text feature [depends] present in test data point [True]
847 Text feature [mechanism] present in test data point [True]
848 Text feature [possess] present in test data point [True]
856 Text feature [substitution] present in test data point [True]
857 Text feature [promoter] present in test data point [True]
860 Text feature [unique] present in test data point [True]
861 Text feature [description] present in test data point [True]
862 Text feature [possibly] present in test data point [True]
864 Text feature [crystal] present in test data point [True]
867 Text feature [accessible] present in test data point [True]
870 Text feature [smaller] present in test data point [True]
871 Text feature [give] present in test data point [True]
```

```
875 Text feature [vectors] present in test data point [True]
876 Text feature [refs] present in test data point [True]
878 Text feature [mutants] present in test data point [True]
880 Text feature [suppress] present in test data point [True]
883 Text feature [nature] present in test data point [True]
884 Text feature [45] present in test data point [True]
887 Text feature [non] present in test data point [True]
888 Text feature [poor] present in test data point [True]
890 Text feature [destabilizing] present in test data point [True]
891 Text feature [41] present in test data point [True]
893 Text feature [remove] present in test data point [True]
895 Text feature [shift] present in test data point [True]
898 Text feature [generic] present in test data point [True]
901 Text feature [disruption] present in test data point [True]
902 Text feature [confirming] present in test data point [True]
903 Text feature [regulates] present in test data point [True]
907 Text feature [5a] present in test data point [True]
911 Text feature [supported] present in test data point [True]
912 Text feature [increases] present in test data point [True]
913 Text feature [staining] present in test data point [True]
915 Text feature [antibodies] present in test data point [True]
917 Text feature [43] present in test data point [True]
918 Text feature [implications] present in test data point [True]
919 Text feature [cycle] present in test data point [True]
921 Text feature [activation] present in test data point [True]
922 Text feature [bears] present in test data point [True]
923 Text feature [per] present in test data point [True]
927 Text feature [induction] present in test data point [True]
930 Text feature [use] present in test data point [True]
933 Text feature [requires] present in test data point [True]
937 Text feature [seen] present in test data point [True]
940 Text feature [average] present in test data point [True]
943 Text feature [proteolysis] present in test data point [True]
944 Text feature [dimensional] present in test data point [True]
950 Text feature [encodes] present in test data point [True]
953 Text feature [introducing] present in test data point [True]
954 Text feature [tumors] present in test data point [True]
958 Text feature [prevent] present in test data point [True]
959 Text feature [mutated] present in test data point [True]
962 Text feature [molecular] present in test data point [True]
966 Text feature [fluorescent] present in test data point [True]
968 Text feature [remain] present in test data point [True]
969 Text feature [localized] present in test data point [True]
971 Text feature [direct] present in test data point [True]
976 Text feature [325] present in test data point [True]
980 Text feature [clear] present in test data point [True]
982 Text feature [tabdownload] present in test data point [True]
983 Text feature [additionally] present in test data point [True]
986 Text feature [frequently] present in test data point [True]
987 Text feature [resolution] present in test data point [True]
988 Text feature [figureopen] present in test data point [True]
991 Text feature [green] present in test data point [True]
992 Text feature [approach] present in test data point [True]
993 Text feature [reviewed] present in test data point [True]
994 Text feature [degree] present in test data point [True]
995 Text feature [occurring] present in test data point [True]
998 Text feature [transactivation] present in test data point [True]
Out of the top  1000  features  588 are present in query point
```

In [67]:

```
test_point_index = 55
no_feature = 1000
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_index]),4))
print("Actual Class :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0142 0.0244 0.0058 0.0115 0.0217 0.0158 0.9031 0.0025 0.0011]]
Actual Class : 7
--------------------------------------------------
15 Text feature [cells] present in test data point [True]
16 Text feature [downstream] present in test data point [True]
17 Text feature [activated] present in test data point [True]
19 Text feature [cell] present in test data point [True]
21 Text feature [contrast] present in test data point [True]
22 Text feature [activation] present in test data point [True]
23 Text feature [presence] present in test data point [True]
24 Text feature [shown] present in test data point [True]
25 Text feature [expressing] present in test data point [True]
26 Text feature [factor] present in test data point [True]
27 Text feature [also] present in test data point [True]
28 Text feature [phosphorylation] present in test data point [True]
29 Text feature [growth] present in test data point [True]
30 Text feature [inhibitor] present in test data point [True]
31 Text feature [however] present in test data point [True]
34 Text feature [suggest] present in test data point [True]
35 Text feature [independent] present in test data point [True]
36 Text feature [found] present in test data point [True]
37 Text feature [mechanism] present in test data point [True]
38 Text feature [similar] present in test data point [True]
39 Text feature [recently] present in test data point [True]
40 Text feature [treatment] present in test data point [True]
41 Text feature [10] present in test data point [True]
42 Text feature [showed] present in test data point [True]
43 Text feature [treated] present in test data point [True]
45 Text feature [compared] present in test data point [True]
46 Text feature [signaling] present in test data point [True]
48 Text feature [1a] present in test data point [True]
49 Text feature [addition] present in test data point [True]
50 Text feature [increased] present in test data point [True]
51 Text feature [potential] present in test data point [True]
52 Text feature [inhibition] present in test data point [True]
53 Text feature [constitutive] present in test data point [True]
54 Text feature [inhibited] present in test data point [True]
55 Text feature [3b] present in test data point [True]
56 Text feature [mutant] present in test data point [True]
57 Text feature [figure] present in test data point [True]
58 Text feature [consistent] present in test data point [True]
59 Text feature [well] present in test data point [True]
63 Text feature [higher] present in test data point [True]
64 Text feature [mutations] present in test data point [True]
65 Text feature [may] present in test data point [True]
66 Text feature [enhanced] present in test data point [True]
67 Text feature [including] present in test data point [True]
68 Text feature [interestingly] present in test data point [True]
69 Text feature [various] present in test data point [True]
72 Text feature [reported] present in test data point [True]
73 Text feature [constitutively] present in test data point [True]
74 Text feature [demonstrated] present in test data point [True]
75 Text feature [fig] present in test data point [True]
76 Text feature [sensitive] present in test data point [True]
77 Text feature [using] present in test data point [True]
78 Text feature [described] present in test data point [True]
79 Text feature [mutation] present in test data point [True]
80 Text feature [inhibitors] present in test data point [True]
81 Text feature [observed] present in test data point [True]
82 Text feature [furthermore] present in test data point [True]
83 Text feature [increase] present in test data point [True]
84 Text feature [detected] present in test data point [True]
85 Text feature [followed] present in test data point [True]
86 Text feature [two] present in test data point [True]
87 Text feature [3a] present in test data point [True]
88 Text feature [expression] present in test data point [True]
89 Text feature [concentrations] present in test data point [True]
90 Text feature [without] present in test data point [True]
91 Text feature [absence] present in test data point [True]
92 Text feature [antibodies] present in test data point [True]
93 Text feature [4a] present in test data point [True]
94 Text feature [small] present in test data point [True]
95 Text feature [approximately] present in test data point [True]
96 Text feature [total] present in test data point [True]
97 Text feature [whether] present in test data point [True]
```

```
98 Text feature [suggesting] present in test data point [True]
99 Text feature [proliferation] present in test data point [True]
100 Text feature [previous] present in test data point [True]
101 Text feature [domain] present in test data point [True]
102 Text feature [expressed] present in test data point [True]
103 Text feature [activating] present in test data point [True]
104 Text feature [antibody] present in test data point [True]
105 Text feature [recent] present in test data point [True]
107 Text feature [induced] present in test data point [True]
108 Text feature [identified] present in test data point [True]
109 Text feature [12] present in test data point [True]
110 Text feature [performed] present in test data point [True]
111 Text feature [mechanisms] present in test data point [True]
112 Text feature [examined] present in test data point [True]
113 Text feature [respectively] present in test data point [True]
114 Text feature [pathways] present in test data point [True]
115 Text feature [role] present in test data point [True]
116 Text feature [molecular] present in test data point [True]
117 Text feature [could] present in test data point [True]
118 Text feature [approved] present in test data point [True]
119 Text feature [together] present in test data point [True]
120 Text feature [different] present in test data point [True]
121 Text feature [study] present in test data point [True]
123 Text feature [leading] present in test data point [True]
125 Text feature [therapeutic] present in test data point [True]
126 Text feature [either] present in test data point [True]
127 Text feature [thus] present in test data point [True]
128 Text feature [discussion] present in test data point [True]
129 Text feature [report] present in test data point [True]
130 Text feature [resulting] present in test data point [True]
131 Text feature [led] present in test data point [True]
132 Text feature [despite] present in test data point [True]
133 Text feature [single] present in test data point [True]
134 Text feature [15] present in test data point [True]
136 Text feature [occur] present in test data point [True]
137 Text feature [high] present in test data point [True]
138 Text feature [tumor] present in test data point [True]
139 Text feature [13] present in test data point [True]
140 Text feature [survival] present in test data point [True]
141 Text feature [suggests] present in test data point [True]
142 Text feature [lines] present in test data point [True]
143 Text feature [findings] present in test data point [True]
145 Text feature [patients] present in test data point [True]
146 Text feature [1b] present in test data point [True]
148 Text feature [activate] present in test data point [True]
149 Text feature [development] present in test data point [True]
150 Text feature [although] present in test data point [True]
151 Text feature [results] present in test data point [True]
152 Text feature [studies] present in test data point [True]
156 Text feature [another] present in test data point [True]
157 Text feature [three] present in test data point [True]
158 Text feature [one] present in test data point [True]
159 Text feature [2b] present in test data point [True]
160 Text feature [common] present in test data point [True]
161 Text feature [anti] present in test data point [True]
162 Text feature [due] present in test data point [True]
163 Text feature [lead] present in test data point [True]
164 Text feature [might] present in test data point [True]
165 Text feature [4b] present in test data point [True]
166 Text feature [next] present in test data point [True]
167 Text feature [revealed] present in test data point [True]
168 Text feature [receptor] present in test data point [True]
169 Text feature [oncogenic] present in test data point [True]
170 Text feature [overexpression] present in test data point [True]
171 Text feature [clinical] present in test data point [True]
172 Text feature [pathway] present in test data point [True]
174 Text feature [similarly] present in test data point [True]
175 Text feature [show] present in test data point [True]
176 Text feature [time] present in test data point [True]
177 Text feature [whereas] present in test data point [True]
178 Text feature [express] present in test data point [True]
179 Text feature [additional] present in test data point [True]
180 Text feature [drug] present in test data point [True]
181 Text feature [vitro] present in test data point [True]
182 Text feature [specific] present in test data point [True]
183 Text feature [trials] present in test data point [True]
184 Text feature [promote] present in test data point [True]
```

```
186 Text feature [positive] present in test data point [True]
187 Text feature [measured] present in test data point [True]
189 Text feature [18] present in test data point [True]
191 Text feature [dependent] present in test data point [True]
193 Text feature [20] present in test data point [True]
195 Text feature [within] present in test data point [True]
196 Text feature [active] present in test data point [True]
197 Text feature [therapies] present in test data point [True]
199 Text feature [several] present in test data point [True]
200 Text feature [suggested] present in test data point [True]
201 Text feature [culture] present in test data point [True]
202 Text feature [analysis] present in test data point [True]
203 Text feature [present] present in test data point [True]
205 Text feature [conditions] present in test data point [True]
206 Text feature [analyzed] present in test data point [True]
207 Text feature [taken] present in test data point [True]
209 Text feature [3c] present in test data point [True]
210 Text feature [target] present in test data point [True]
211 Text feature [human] present in test data point [True]
213 Text feature [result] present in test data point [True]
215 Text feature [new] present in test data point [True]
216 Text feature [contribute] present in test data point [True]
217 Text feature [progression] present in test data point [True]
219 Text feature [lysates] present in test data point [True]
220 Text feature [go] present in test data point [True]
221 Text feature [tumors] present in test data point [True]
222 Text feature [lung] present in test data point [True]
223 Text feature [part] present in test data point [True]
224 Text feature [genomic] present in test data point [True]
225 Text feature [regulated] present in test data point [True]
226 Text feature [targeted] present in test data point [True]
227 Text feature [identification] present in test data point [True]
228 Text feature [gene] present in test data point [True]
229 Text feature [mediated] present in test data point [True]
231 Text feature [major] present in test data point [True]
232 Text feature [less] present in test data point [True]
234 Text feature [collection] present in test data point [True]
235 Text feature [sequencing] present in test data point [True]
238 Text feature [indicate] present in test data point [True]
239 Text feature [derived] present in test data point [True]
240 Text feature [19] present in test data point [True]
241 Text feature [free] present in test data point [True]
243 Text feature [frequently] present in test data point [True]
246 Text feature [lower] present in test data point [True]
247 Text feature [table] present in test data point [True]
249 Text feature [currently] present in test data point [True]
250 Text feature [elevated] present in test data point [True]
251 Text feature [malignant] present in test data point [True]
252 Text feature [cancers] present in test data point [True]
253 Text feature [1c] present in test data point [True]
254 Text feature [indicating] present in test data point [True]
255 Text feature [indicated] present in test data point [True]
256 Text feature [support] present in test data point [True]
257 Text feature [first] present in test data point [True]
258 Text feature [effects] present in test data point [True]
260 Text feature [level] present in test data point [True]
264 Text feature [normal] present in test data point [True]
266 Text feature [distinct] present in test data point [True]
267 Text feature [differences] present in test data point [True]
268 Text feature [primary] present in test data point [True]
269 Text feature [mutants] present in test data point [True]
270 Text feature [somatic] present in test data point [True]
271 Text feature [therapy] present in test data point [True]
272 Text feature [25] present in test data point [True]
273 Text feature [leads] present in test data point [True]
276 Text feature [characterized] present in test data point [True]
278 Text feature [caused] present in test data point [True]
279 Text feature [form] present in test data point [True]
282 Text feature [sequenced] present in test data point [True]
283 Text feature [established] present in test data point [True]
284 Text feature [27] present in test data point [True]
286 Text feature [following] present in test data point [True]
287 Text feature [relative] present in test data point [True]
288 Text feature [experiments] present in test data point [True]
289 Text feature [24] present in test data point [True]
292 Text feature [48] present in test data point [True]
294 Text feature [important] present in test data point [True]
```

```
295 Text feature [investigated] present in test data point [True]
296 Text feature [cause] present in test data point [True]
297 Text feature [determined] present in test data point [True]
299 Text feature [samples] present in test data point [True]
303 Text feature [stimulation] present in test data point [True]
304 Text feature [disease] present in test data point [True]
305 Text feature [resulted] present in test data point [True]
306 Text feature [containing] present in test data point [True]
308 Text feature [seen] present in test data point [True]
309 Text feature [2d] present in test data point [True]
310 Text feature [significantly] present in test data point [True]
311 Text feature [novel] present in test data point [True]
313 Text feature [according] present in test data point [True]
314 Text feature [represents] present in test data point [True]
315 Text feature [days] present in test data point [True]
316 Text feature [keywords] present in test data point [True]
317 Text feature [activity] present in test data point [True]
318 Text feature [harboring] present in test data point [True]
320 Text feature [17] present in test data point [True]
321 Text feature [manner] present in test data point [True]
322 Text feature [cellular] present in test data point [True]
323 Text feature [prepared] present in test data point [True]
325 Text feature [remains] present in test data point [True]
327 Text feature [test] present in test data point [True]
328 Text feature [signal] present in test data point [True]
329 Text feature [levels] present in test data point [True]
331 Text feature [maintained] present in test data point [True]
332 Text feature [indeed] present in test data point [True]
333 Text feature [unlike] present in test data point [True]
334 Text feature [response] present in test data point [True]
335 Text feature [tissues] present in test data point [True]
336 Text feature [representative] present in test data point [True]
337 Text feature [inhibits] present in test data point [True]
338 Text feature [30] present in test data point [True]
340 Text feature [non] present in test data point [True]
341 Text feature [mutated] present in test data point [True]
343 Text feature [evaluated] present in test data point [True]
345 Text feature [pcr] present in test data point [True]
346 Text feature [significant] present in test data point [True]
349 Text feature [highly] present in test data point [True]
352 Text feature [targets] present in test data point [True]
355 Text feature [directly] present in test data point [True]
358 Text feature [advanced] present in test data point [True]
360 Text feature [negative] present in test data point [True]
361 Text feature [possible] present in test data point [True]
362 Text feature [selected] present in test data point [True]
363 Text feature [entire] present in test data point [True]
364 Text feature [download] present in test data point [True]
365 Text feature [therefore] present in test data point [True]
366 Text feature [represent] present in test data point [True]
367 Text feature [increases] present in test data point [True]
368 Text feature [sigma] present in test data point [True]
369 Text feature [cancer] present in test data point [True]
371 Text feature [developed] present in test data point [True]
372 Text feature [4c] present in test data point [True]
374 Text feature [agent] present in test data point [True]
376 Text feature [11] present in test data point [True]
378 Text feature [plays] present in test data point [True]
379 Text feature [14] present in test data point [True]
381 Text feature [oncogene] present in test data point [True]
383 Text feature [identify] present in test data point [True]
384 Text feature [increasing] present in test data point [True]
389 Text feature [used] present in test data point [True]
393 Text feature [subjected] present in test data point [True]
395 Text feature [40] present in test data point [True]
397 Text feature [introduction] present in test data point [True]
401 Text feature [driven] present in test data point [True]
404 Text feature [importance] present in test data point [True]
406 Text feature [low] present in test data point [True]
407 Text feature [22] present in test data point [True]
408 Text feature [50] present in test data point [True]
410 Text feature [rate] present in test data point [True]
412 Text feature [factors] present in test data point [True]
414 Text feature [variety] present in test data point [True]
417 Text feature [along] present in test data point [True]
418 Text feature [generation] present in test data point [True]
419 Text feature [enhance] present in test data point [True]
```

```
420 Text feature [transient] present in test data point [True]
421 Text feature [stable] present in test data point [True]
422 Text feature [confer] present in test data point [True]
423 Text feature [initially] present in test data point [True]
424 Text feature [stage] present in test data point [True]
425 Text feature [fold] present in test data point [True]
428 Text feature [immunoblotting] present in test data point [True]
429 Text feature [understanding] present in test data point [True]
430 Text feature [21] present in test data point [True]
431 Text feature [sample] present in test data point [True]
432 Text feature [fact] present in test data point [True]
434 Text feature [16] present in test data point [True]
437 Text feature [observation] present in test data point [True]
438 Text feature [type] present in test data point [True]
439 Text feature [sensitivity] present in test data point [True]
441 Text feature [strongly] present in test data point [True]
446 Text feature [required] present in test data point [True]
447 Text feature [transfected] present in test data point [True]
450 Text feature [concentration] present in test data point [True]
456 Text feature [cases] present in test data point [True]
457 Text feature [located] present in test data point [True]
458 Text feature [drugs] present in test data point [True]
459 Text feature [associated] present in test data point [True]
460 Text feature [demonstrate] present in test data point [True]
461 Text feature [play] present in test data point [True]
463 Text feature [benefit] present in test data point [True]
465 Text feature [day] present in test data point [True]
466 Text feature [intracellular] present in test data point [True]
469 Text feature [line] present in test data point [True]
470 Text feature [28] present in test data point [True]
471 Text feature [control] present in test data point [True]
472 Text feature [amplification] present in test data point [True]
473 Text feature [regulation] present in test data point [True]
474 Text feature [phase] present in test data point [True]
475 Text feature [molecule] present in test data point [True]
477 Text feature [able] present in test data point [True]
478 Text feature [data] present in test data point [True]
479 Text feature [purchased] present in test data point [True]
481 Text feature [date] present in test data point [True]
482 Text feature [provide] present in test data point [True]
483 Text feature [remained] present in test data point [True]
484 Text feature [number] present in test data point [True]
487 Text feature [carcinoma] present in test data point [True]
496 Text feature [mouse] present in test data point [True]
497 Text feature [alone] present in test data point [True]
498 Text feature [cdna] present in test data point [True]
499 Text feature [resistant] present in test data point [True]
500 Text feature [subsequent] present in test data point [True]
502 Text feature [appears] present in test data point [True]
505 Text feature [respective] present in test data point [True]
508 Text feature [treating] present in test data point [True]
509 Text feature [yet] present in test data point [True]
510 Text feature [finally] present in test data point [True]
511 Text feature [exhibited] present in test data point [True]
515 Text feature [long] present in test data point [True]
516 Text feature [reagent] present in test data point [True]
517 Text feature [conformation] present in test data point [True]
518 Text feature [s3] present in test data point [True]
522 Text feature [decreased] present in test data point [True]
523 Text feature [multiple] present in test data point [True]
524 Text feature [coding] present in test data point [True]
528 Text feature [et] present in test data point [True]
530 Text feature [failed] present in test data point [True]
533 Text feature [forms] present in test data point [True]
534 Text feature [since] present in test data point [True]
536 Text feature [adenocarcinoma] present in test data point [True]
537 Text feature [occurs] present in test data point [True]
538 Text feature [blood] present in test data point [True]
539 Text feature [al] present in test data point [True]
541 Text feature [decrease] present in test data point [True]
542 Text feature [amplified] present in test data point [True]
544 Text feature [nsclc] present in test data point [True]
545 Text feature [observations] present in test data point [True]
547 Text feature [agents] present in test data point [True]
548 Text feature [order] present in test data point [True]
550 Text feature [23] present in test data point [True]
551 Text feature [via] present in test data point [True]
```

```
554 Text feature [regulates] present in test data point [True]
556 Text feature [reports] present in test data point [True]
557 Text feature [trial] present in test data point [True]
558 Text feature [frequent] present in test data point [True]
559 Text feature [properties] present in test data point [True]
560 Text feature [particularly] present in test data point [True]
562 Text feature [course] present in test data point [True]
566 Text feature [inhibiting] present in test data point [True]
567 Text feature [example] present in test data point [True]
574 Text feature [32] present in test data point [True]
575 Text feature [endogenous] present in test data point [True]
576 Text feature [necessary] present in test data point [True]
577 Text feature [wild] present in test data point [True]
578 Text feature [status] present in test data point [True]
579 Text feature [3e] present in test data point [True]
581 Text feature [egfr] present in test data point [True]
583 Text feature [membrane] present in test data point [True]
584 Text feature [short] present in test data point [True]
585 Text feature [potent] present in test data point [True]
589 Text feature [blocked] present in test data point [True]
591 Text feature [establish] present in test data point [True]
594 Text feature [critical] present in test data point [True]
596 Text feature [carried] present in test data point [True]
597 Text feature [involved] present in test data point [True]
598 Text feature [formation] present in test data point [True]
600 Text feature [early] present in test data point [True]
602 Text feature [responsible] present in test data point [True]
604 Text feature [seems] present in test data point [True]
607 Text feature [primers] present in test data point [True]
608 Text feature [s1] present in test data point [True]
609 Text feature [promotes] present in test data point [True]
611 Text feature [epithelial] present in test data point [True]
612 Text feature [nucleotide] present in test data point [True]
613 Text feature [case] present in test data point [True]
615 Text feature [acid] present in test data point [True]
616 Text feature [effect] present in test data point [True]
617 Text feature [range] present in test data point [True]
618 Text feature [often] present in test data point [True]
620 Text feature [generate] present in test data point [True]
624 Text feature [overall] present in test data point [True]
625 Text feature [proliferate] present in test data point [True]
627 Text feature [region] present in test data point [True]
631 Text feature [like] present in test data point [True]
633 Text feature [vectors] present in test data point [True]
636 Text feature [extracted] present in test data point [True]
637 Text feature [direct] present in test data point [True]
639 Text feature [33] present in test data point [True]
641 Text feature [thereby] present in test data point [True]
643 Text feature [inhibitory] present in test data point [True]
646 Text feature [review] present in test data point [True]
647 Text feature [exon] present in test data point [True]
648 Text feature [majority] present in test data point [True]
651 Text feature [viability] present in test data point [True]
652 Text feature [preclinical] present in test data point [True]
656 Text feature [metastatic] present in test data point [True]
660 Text feature [vivo] present in test data point [True]
663 Text feature [additionally] present in test data point [True]
664 Text feature [future] present in test data point [True]
666 Text feature [least] present in test data point [True]
670 Text feature [efficacy] present in test data point [True]
672 Text feature [include] present in test data point [True]
673 Text feature [strategies] present in test data point [True]
677 Text feature [protein] present in test data point [True]
678 Text feature [hybridization] present in test data point [True]
679 Text feature [bearing] present in test data point [True]
681 Text feature [many] present in test data point [True]
682 Text feature [peroxidase] present in test data point [True]
683 Text feature [corresponding] present in test data point [True]
688 Text feature [capable] present in test data point [True]
690 Text feature [collectively] present in test data point [True]
691 Text feature [proteins] present in test data point [True]
698 Text feature [prolonged] present in test data point [True]
699 Text feature [reverse] present in test data point [True]
701 Text feature [whole] present in test data point [True]
703 Text feature [supports] present in test data point [True]
704 Text feature [right] present in test data point [True]
705 Text feature [wide] present in test data point [True]
```

706 Text feature [groups] present in test data point [True]
711 Text feature [notion] present in test data point [True]
712 Text feature [immunoblot] present in test data point [True]
713 Text feature [harbored] present in test data point [True]
714 Text feature [hand] present in test data point [True]
717 Text feature [regulate] present in test data point [True]
718 Text feature [500] present in test data point [True]
719 Text feature [29] present in test data point [True]
720 Text feature [experimental] present in test data point [True]
721 Text feature [full] present in test data point [True]
722 Text feature [model] present in test data point [True]
723 Text feature [exons] present in test data point [True]
724 Text feature [reagents] present in test data point [True]
725 Text feature [acute] present in test data point [True]
726 Text feature [providing] present in test data point [True]
728 Text feature [inhibit] present in test data point [True]
730 Text feature [sites] present in test data point [True]
731 Text feature [contain] present in test data point [True]
732 Text feature [design] present in test data point [True]
735 Text feature [left] present in test data point [True]
737 Text feature [targeting] present in test data point [True]
738 Text feature [blocks] present in test data point [True]
739 Text feature [resistance] present in test data point [True]
747 Text feature [hypothesized] present in test data point [True]
750 Text feature [exposure] present in test data point [True]
751 Text feature [known] present in test data point [True]
753 Text feature [potentially] present in test data point [True]
754 Text feature [alterations] present in test data point [True]
755 Text feature [maintain] present in test data point [True]
756 Text feature [spectrum] present in test data point [True]
758 Text feature [polyclonal] present in test data point [True]
763 Text feature [substitution] present in test data point [True]
764 Text feature [schematic] present in test data point [True]
765 Text feature [numerous] present in test data point [True]
768 Text feature [laboratory] present in test data point [True]
770 Text feature [mean] present in test data point [True]
771 Text feature [certain] present in test data point [True]
775 Text feature [six] present in test data point [True]
776 Text feature [statistical] present in test data point [True]
777 Text feature [top] present in test data point [True]
778 Text feature [particular] present in test data point [True]
779 Text feature [roche] present in test data point [True]
781 Text feature [promoting] present in test data point [True]
785 Text feature [emerged] present in test data point [True]
786 Text feature [prognosis] present in test data point [True]
790 Text feature [2000] present in test data point [True]
791 Text feature [36] present in test data point [True]
794 Text feature [become] present in test data point [True]
797 Text feature [origin] present in test data point [True]
799 Text feature [mice] present in test data point [True]
803 Text feature [indicates] present in test data point [True]
805 Text feature [density] present in test data point [True]
806 Text feature [primarily] present in test data point [True]
807 Text feature [end] present in test data point [True]
812 Text feature [chemotherapy] present in test data point [True]
822 Text feature [amino] present in test data point [True]
830 Text feature [evaluation] present in test data point [True]
831 Text feature [among] present in test data point [True]
832 Text feature [reduced] present in test data point [True]
834 Text feature [pattern] present in test data point [True]
835 Text feature [needed] present in test data point [True]
837 Text feature [poor] present in test data point [True]
839 Text feature [adjacent] present in test data point [True]
844 Text feature [develop] present in test data point [True]
847 Text feature [involve] present in test data point [True]
848 Text feature [see] present in test data point [True]
852 Text feature [components] present in test data point [True]
854 Text feature [generally] present in test data point [True]
856 Text feature [sequences] present in test data point [True]
857 Text feature [promising] present in test data point [True]
860 Text feature [accounts] present in test data point [True]
862 Text feature [acquired] present in test data point [True]
863 Text feature [tested] present in test data point [True]
864 Text feature [toward] present in test data point [True]
866 Text feature [substantial] present in test data point [True]
867 Text feature [better] present in test data point [True]
868 Text feature [panel] present in test data point [True]

```
870 Text feature [essential] present in test data point [True]
871 Text feature [glutamine] present in test data point [True]
876 Text feature [genes] present in test data point [True]
878 Text feature [1d] present in test data point [True]
879 Text feature [37] present in test data point [True]
880 Text feature [chain] present in test data point [True]
883 Text feature [immunoprecipitation] present in test data point [True]
884 Text feature [especially] present in test data point [True]
885 Text feature [types] present in test data point [True]
886 Text feature [growing] present in test data point [True]
887 Text feature [modified] present in test data point [True]
893 Text feature [gain] present in test data point [True]
897 Text feature [38] present in test data point [True]
898 Text feature [system] present in test data point [True]
899 Text feature [length] present in test data point [True]
901 Text feature [twice] present in test data point [True]
903 Text feature [actin] present in test data point [True]
913 Text feature [dual] present in test data point [True]
918 Text feature [mg] present in test data point [True]
920 Text feature [gastrointestinal] present in test data point [True]
921 Text feature [processes] present in test data point [True]
923 Text feature [technologies] present in test data point [True]
925 Text feature [outcome] present in test data point [True]
927 Text feature [deletion] present in test data point [True]
928 Text feature [repeated] present in test data point [True]
931 Text feature [confers] present in test data point [True]
934 Text feature [biology] present in test data point [True]
936 Text feature [methods] present in test data point [True]
939 Text feature [markedly] present in test data point [True]
942 Text feature [underlying] present in test data point [True]
949 Text feature [synthesized] present in test data point [True]
950 Text feature [therapeutics] present in test data point [True]
951 Text feature [roles] present in test data point [True]
955 Text feature [05] present in test data point [True]
956 Text feature [molecules] present in test data point [True]
957 Text feature [causes] present in test data point [True]
964 Text feature [provides] present in test data point [True]
966 Text feature [site] present in test data point [True]
968 Text feature [preferentially] present in test data point [True]
970 Text feature [rank] present in test data point [True]
972 Text feature [selectively] present in test data point [True]
984 Text feature [60] present in test data point [True]
989 Text feature [basal] present in test data point [True]
991 Text feature [analyses] present in test data point [True]
996 Text feature [complex] present in test data point [True]
998 Text feature [specifically] present in test data point [True]
Out of the top  1000  features  580 are present in query point
```

KNN Model

With Hyperparameter Tuning

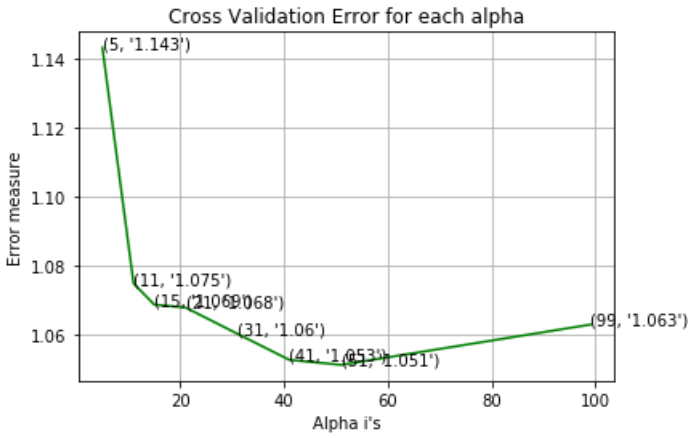
In [68]:

```
# find more about KNeighborsClassifier()
# here http://scikit-
learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
# -------------------------
# default parameter
# KNeighborsClassifier(n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2,
# metric='minkowski', metric_params=None, n_jobs=1, **kwargs)

# methods of
# fit(X, y) : Fit the model using X as training data and y as target values
# predict(X):Predict the class labels for the provided data
# predict_proba(X):Return probability estimates for the test data X.
#--------------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/k-nearest-ne
ighbors-geometric-intuition-with-a-toy-example-1/
#--------------------------------------


# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
```

```
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
#-------------------------------------
# video link:
#-------------------------------------


alpha = [5, 11, 15, 21, 31, 41, 51, 99]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = KNeighborsClassifier(n_neighbors=i)
    clf.fit(train_x_responseCoding, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_responseCoding, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_responseCoding)
    cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    # to avoid rounding error while multiplying probabilites we use log-probability estimates
    print("Log Loss :",log_loss(cv_y, sig_clf_probs))

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],str(txt)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
clf.fit(train_x_responseCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_responseCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_responseCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The train log loss is:",
      log_loss(y_train, predict_y, labels=clf.classes_,eps=1e-15))

predict_y = sig_clf.predict_proba(cv_x_responseCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The cross validation log loss is:",
      log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))

predict_y = sig_clf.predict_proba(test_x_responseCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The test log loss is:",
      log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 5
Log Loss : 1.143315403858183
for alpha = 11
Log Loss : 1.0748031716504756
for alpha = 15
Log Loss : 1.068635924092476
for alpha = 21
Log Loss : 1.067819110078448
for alpha = 31
Log Loss : 1.0601802267389477
for alpha = 41
Log Loss : 1.0525983320073617
for alpha = 51
Log Loss : 1.05114447442798
for alpha = 99
```

```
Log Loss : 1.0628356617640367
```



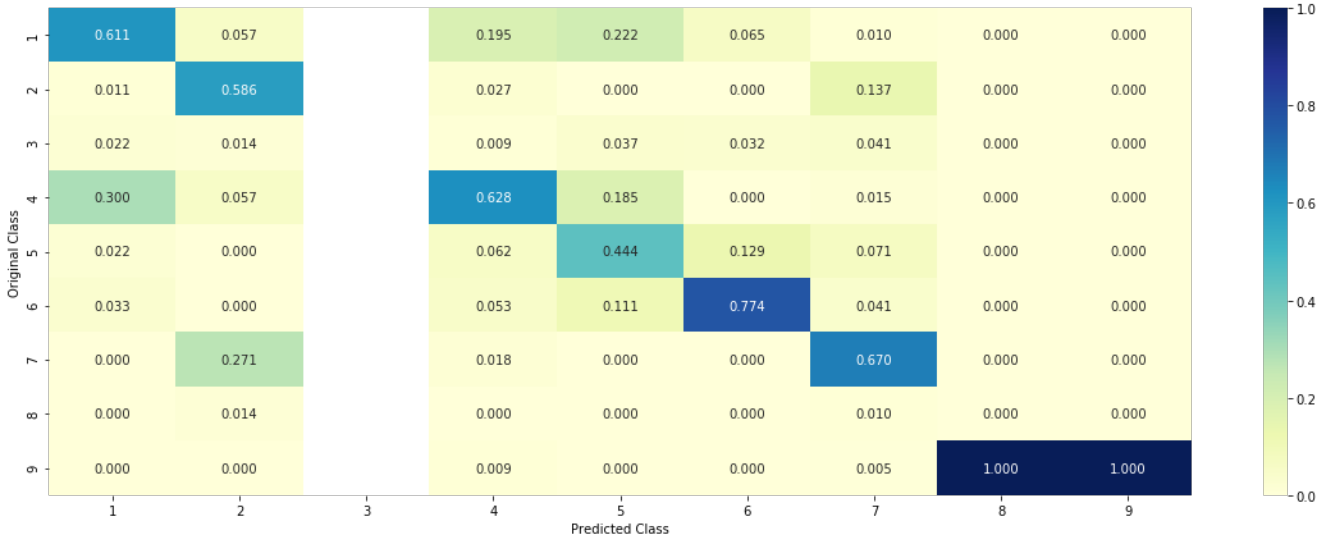

```
For values of best alpha =  51 The train log loss is: 0.8972936640409283
For values of best alpha =  51 The cross validation log loss is: 1.05114447442798
For values of best alpha =  51 The test log loss is: 1.0395637894469603
```

In [69]:

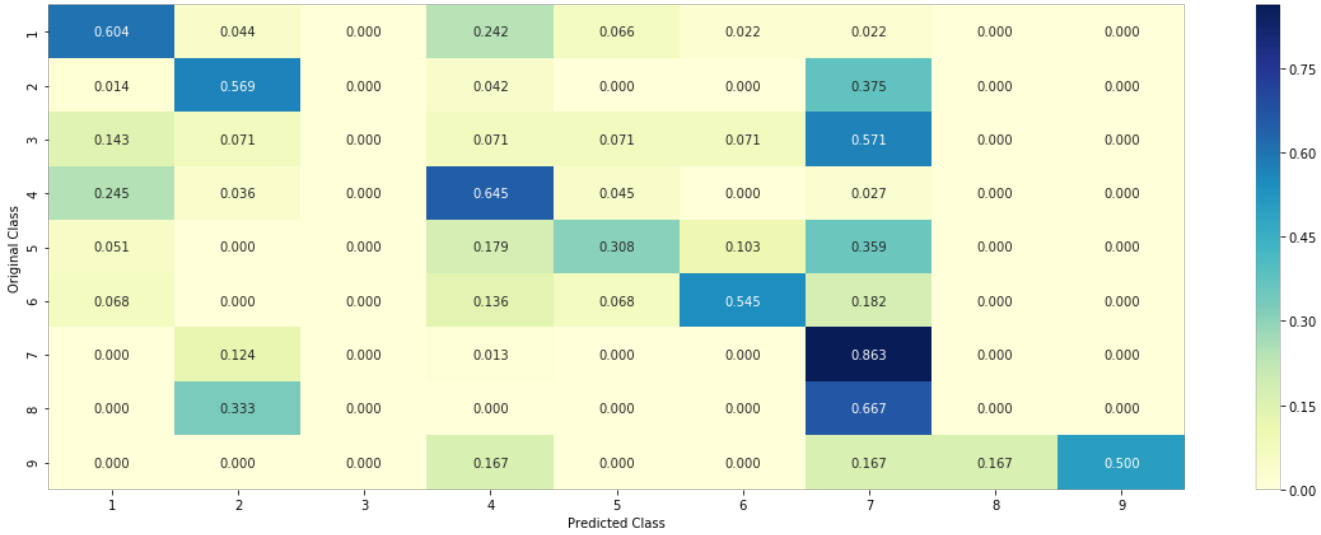
```
# find more about KNeighborsClassifier()
# here http://scikit-
learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
# -------------------------
# default parameter
# KNeighborsClassifier(n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2,
# metric='minkowski', metric_params=None, n_jobs=1, **kwargs)

# methods of
# fit(X, y) : Fit the model using X as training data and y as target values
# predict(X):Predict the class labels for the provided data
# predict_proba(X):Return probability estimates for the test data X.
#-------------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/k-nearest-ne
ighbors-geometric-intuition-with-a-toy-example-1/
#-------------------------------------
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
predict_and_plot_confusion_matrix(train_x_responseCoding, train_y, cv_x_responseCoding, cv_y, clf)
```

```
Log loss : 1.05114447442798
Number of mis-classified points : 0.36466165413533835

-------------------- Confusion matrix --------------------
```


```
-------------------- Precision matrix (Column Sum=1) --------------------
```



---------- Recall matrix (Row sum=1) ----------



In [70]:

```
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
clf.fit(train_x_responseCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_responseCoding, train_y)

test_point_index = 1
predicted_cls = sig_clf.predict(test_x_responseCoding[0].reshape(1,-1))
print("Predicted Class :", predicted_cls[0])
print("Actual Class :", test_y[test_point_index])
neighbors = clf.kneighbors(test_x_responseCoding[test_point_index].reshape(1, -1), alpha[best_alpha
])
print("The ",alpha[best_alpha]," nearest neighbours of the test points belongs to classes",train_y
[neighbors[1][0]])
print("Fequency of nearest points :",Counter(train_y[neighbors[1][0]]))
```

```
Predicted Class : 1
Actual Class : 1
The  51  nearest neighbours of the test points belongs to classes [1 1 4 1 1 1 1 1 1 1 4 6 1 1 2 6
1 1 1 4 1 1 1 1 1 4 4 1 4 4 1 4 1 1 1 1 4
 5 1 4 1 5 1 1 4 1 1 4 4 4 4]
Fequency of nearest points : Counter({1: 31, 4: 15, 6: 2, 5: 2, 2: 1})
```

In [71]:

```
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
clf.fit(train_x_responseCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
```

```
sig_clf.fit(train_x_responseCoding, train_y)

test_point_index = 100

predicted_cls = sig_clf.predict(test_x_responseCoding[test_point_index].reshape(1,-1))
print("Predicted Class :", predicted_cls[0])
print("Actual Class :", test_y[test_point_index])
neighbors = clf.kneighbors(test_x_responseCoding[test_point_index].reshape(1, -1), alpha[best_alpha
])
print("the k value for knn is",alpha[best_alpha],"and the nearest neighbours of the test points be
longs to classes",train_y[neighbors[1][0]])
print("Fequency of nearest points :",Counter(train_y[neighbors[1][0]]))
```

```
Predicted Class : 4
Actual Class : 4
the k value for knn is 51 and the nearest neighbours of the test points belongs to classes [4 4 4
4 4 4 4 4 4 4 1 1 4 4 4 4 4 1 4 4 4 4 4 4 4 1 4 4 4 4 4 4 4 4 4 4 4 4 4
 4 4 4 4 4 4 4 4 4 4 4 4 4 4]
Fequency of nearest points : Counter({4: 47, 1: 4})
```

Logistic Regression

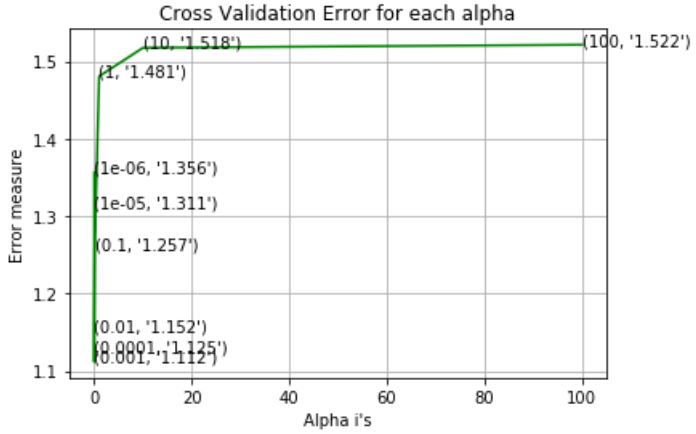
With Class Balancing

In [72]:

```
# read more about SGDClassifier() at http://scikit-
learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_i
ter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0
=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, ...]) Fit linear model with Stochastic Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/geometric-in
tuition-1/
#-------------------------------


# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
#-----------------------------------
# video link:
#-----------------------------------

alpha = [10 ** x for x in range(-6, 3)]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = SGDClassifier(class_weight='balanced', alpha=i, penalty='l2', loss='log', random_state=42
)
    clf.fit(train_x_onehotCoding, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_onehotCoding, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
    cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    # to avoid rounding error while multiplying probabilites we use log-probability estimates
    print("Log Loss :",log_loss(cv_y, sig_clf_probs))
```

```
fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],str(txt)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', ran
dom_state=42)
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The train log loss is:",
      log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15))

predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The cross validation log loss is:",
      log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))

predict_y = sig_clf.predict_proba(test_x_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The test log loss is:",
      log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-06
Log Loss : 1.3564523765925833
for alpha = 1e-05
Log Loss : 1.3108055239663265
for alpha = 0.0001
Log Loss : 1.1254493920067241
for alpha = 0.001
Log Loss : 1.1116768856075288
for alpha = 0.01
Log Loss : 1.1516152184972408
for alpha = 0.1
Log Loss : 1.2572363740930683
for alpha = 1
Log Loss : 1.4807968868899772
for alpha = 10
Log Loss : 1.5176515480817374
for alpha = 100
Log Loss : 1.5215916360402555
```




```
For values of best alpha =  0.001 The train log loss is: 0.5353964942439243
For values of best alpha =  0.001 The cross validation log loss is: 1.1116768856075288
For values of best alpha =  0.001 The test log loss is: 0.9771005085130103
```

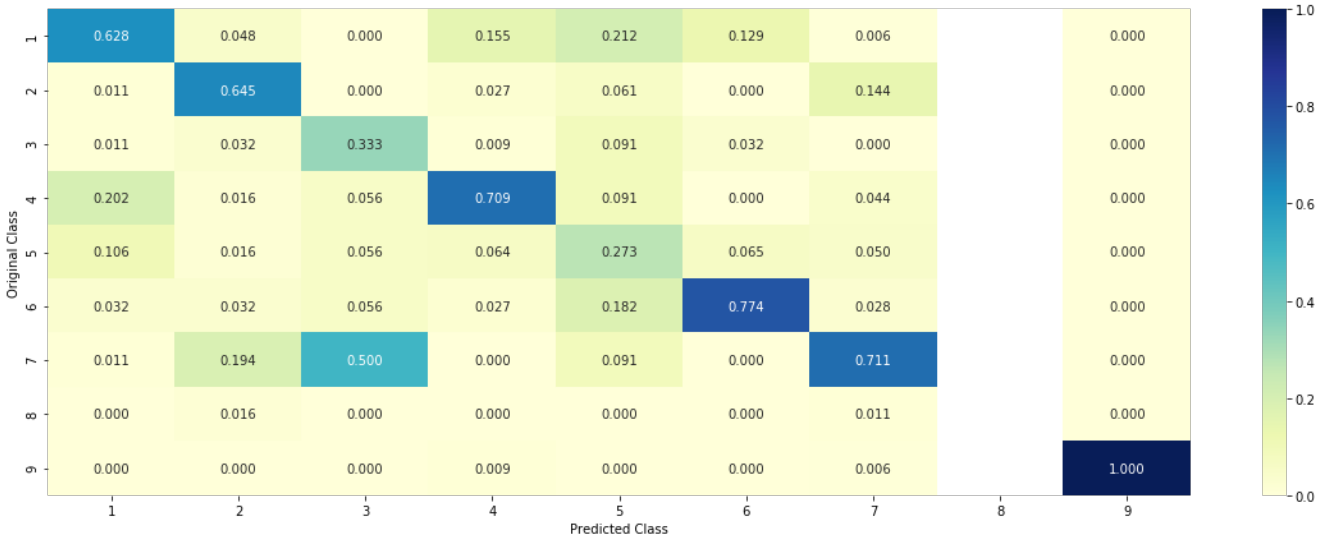
for values of best alpha = 0.001 the test log loss is: 0.9771005065150105

In [73]:

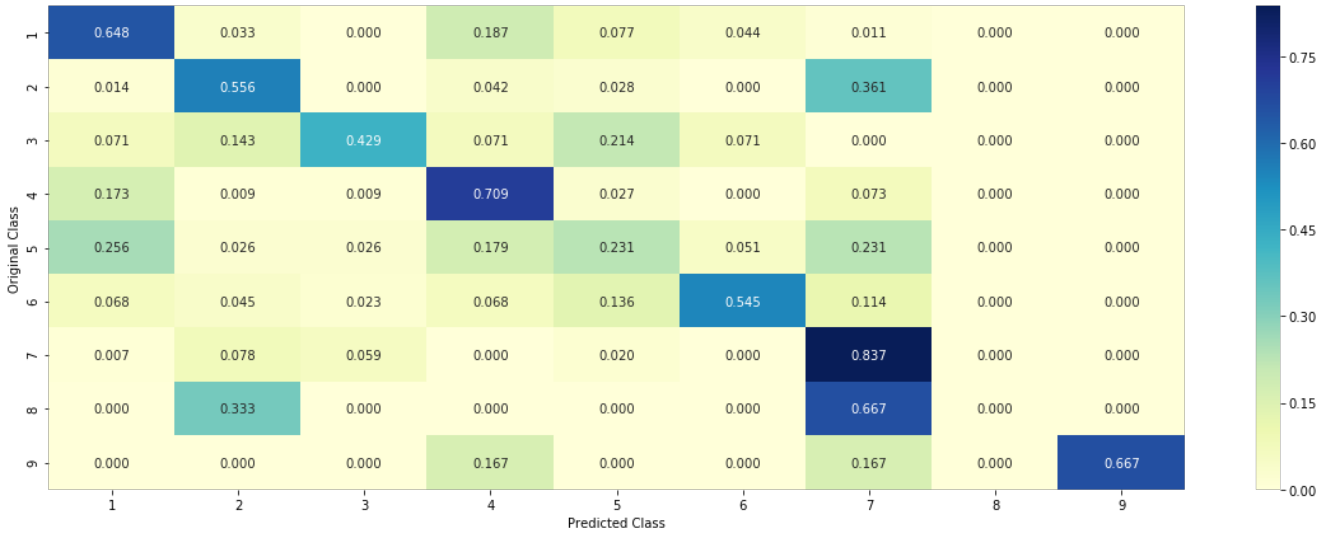
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, ...]) Fit linear model with Stochastic Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/geometric-intuition-1/
#-------------------------------
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y, cv_x_onehotCoding, cv_y, clf)
```

```
Log loss : 1.1116768856075288
Number of mis-classified points : 0.3458646616541353

-------------------- Confusion matrix --------------------
```


```
-------------------- Precision matrix (Columm Sum=1) --------------------
```



```
-------------------- Recall matrix (Row sum=1) --------------------
```



Feature Importance

In [74]:

```
def get_imp_feature_names(text, indices, removed_ind = []):
    word_present = 0
    tabulte_list = []
    incresingorder_ind = 0
    for i in indices:
        if i < train_gene_feature_onehotCoding.shape[1]:
            tabulte_list.append([incresingorder_ind, "Gene", "Yes"])
        elif i< 18:
            tabulte_list.append([incresingorder_ind,"Variation", "Yes"])
        if ((i > 17) & (i not in removed_ind)) :
            word = train_text_features[i]
            yes_no = True if word in text.split() else False
            if yes_no:
                word_present += 1
            tabulte_list.append([incresingorder_ind,train_text_features[i], yes_no])
        incresingorder_ind += 1
    print(word_present, "most importent features are present in our query point")
    print("-"*50)
    print("The features that are most importent of the ",predicted_cls[0]," class:")
    print (tabulate(tabulte_list, headers=["Index",'Feature name', 'Present or Not']))
```

In [75]:

```
# from tabulate import tabulate
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', ran
dom_state=42)
clf.fit(train_x_onehotCoding,train_y)
test_point_index = 1
no_feature = 1000
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_index]),4))
print("Actual Class :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 1
Predicted Class Probabilities: [[0.6416 0.0674 0.0016 0.1509 0.0155 0.1025 0.0145 0.0037 0.0024]]
Actual Class : 1
--------------------------------------------------
```

```
249 Text feature [immobilized] present in test data point [True]
289 Text feature [aggregation] present in test data point [True]
423 Text feature [synergy] present in test data point [True]
483 Text feature [aggregate] present in test data point [True]
513 Text feature [gal] present in test data point [True]
560 Text feature [destabilized] present in test data point [True]
613 Text feature [fas] present in test data point [True]
691 Text feature [saos] present in test data point [True]
712 Text feature [aggregates] present in test data point [True]
780 Text feature [deficient] present in test data point [True]
836 Text feature [unfolded] present in test data point [True]
958 Text feature [tetramerization] present in test data point [True]
973 Text feature [r282] present in test data point [True]
975 Text feature [r273] present in test data point [True]
997 Text feature [misfolded] present in test data point [True]
Out of the top  1000  features  15 are present in query point
```

In [76]:

```
test_point_index = 55
no_feature = 1000
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_index]),4))
print("Actual Class :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0043 0.0213 0.0012 0.0041 0.0058 0.0027 0.9569 0.0021 0.0017]]
Actual Class : 7
--------------------------------------------------
21 Text feature [constitutive] present in test data point [True]
33 Text feature [doublet] present in test data point [True]
37 Text feature [downstream] present in test data point [True]
41 Text feature [activated] present in test data point [True]
66 Text feature [oncogene] present in test data point [True]
71 Text feature [constitutively] present in test data point [True]
104 Text feature [overexpression] present in test data point [True]
129 Text feature [activation] present in test data point [True]
154 Text feature [papillary] present in test data point [True]
182 Text feature [proliferate] present in test data point [True]
200 Text feature [cap] present in test data point [True]
235 Text feature [adenocarcinoma] present in test data point [True]
236 Text feature [expressing] present in test data point [True]
271 Text feature [topical] present in test data point [True]
272 Text feature [temporary] present in test data point [True]
310 Text feature [wako] present in test data point [True]
360 Text feature [joseph] present in test data point [True]
446 Text feature [kras] present in test data point [True]
478 Text feature [fibrosarcoma] present in test data point [True]
481 Text feature [nf] present in test data point [True]
490 Text feature [enhancing] present in test data point [True]
494 Text feature [prolonged] present in test data point [True]
611 Text feature [cysteine] present in test data point [True]
621 Text feature [ectopically] present in test data point [True]
636 Text feature [lung] present in test data point [True]
738 Text feature [nanomolar] present in test data point [True]
805 Text feature [inhibited] present in test data point [True]
857 Text feature [activating] present in test data point [True]
987 Text feature [activate] present in test data point [True]
991 Text feature [competing] present in test data point [True]
998 Text feature [sf] present in test data point [True]
Out of the top  1000  features  31 are present in query point
```

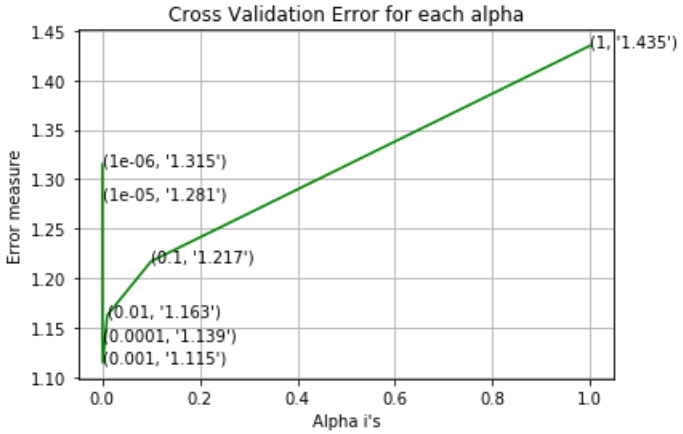
Without Class Balancing

In [77]:

```
# read more about SGDClassifier() at http://scikit-
learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_i
ter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0
=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, …]) Fit linear model with Stochastic Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/geometric-in
tuition-1/
#-------------------------------



# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
#-------------------------------------
# video link:
#-------------------------------------

alpha = [10 ** x for x in range(-6, 1)]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = SGDClassifier(alpha=i, penalty='l2', loss='log', random_state=42)
    clf.fit(train_x_onehotCoding, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_onehotCoding, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
    cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    print("Log Loss :",log_loss(cv_y, sig_clf_probs))

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],str(txt)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The train log loss is:",
      log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15))

predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The cross validation log loss is:",
      log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
```

```
predict_y = sig_clf.predict_proba(test_x_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The test log loss is:",
      log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-06
Log Loss : 1.3150491995633427
for alpha = 1e-05
Log Loss : 1.280635740646622
for alpha = 0.0001
Log Loss : 1.1385262297972858
for alpha = 0.001
Log Loss : 1.1148520648245654
for alpha = 0.01
Log Loss : 1.1628632235687217
for alpha = 0.1
Log Loss : 1.2172747068611212
for alpha = 1
Log Loss : 1.4346594882701027
```




```
For values of best alpha =  0.001 The train log loss is: 0.5346832885161039
For values of best alpha =  0.001 The cross validation log loss is: 1.1148520648245654
For values of best alpha =  0.001 The test log loss is: 0.985931777777684
```

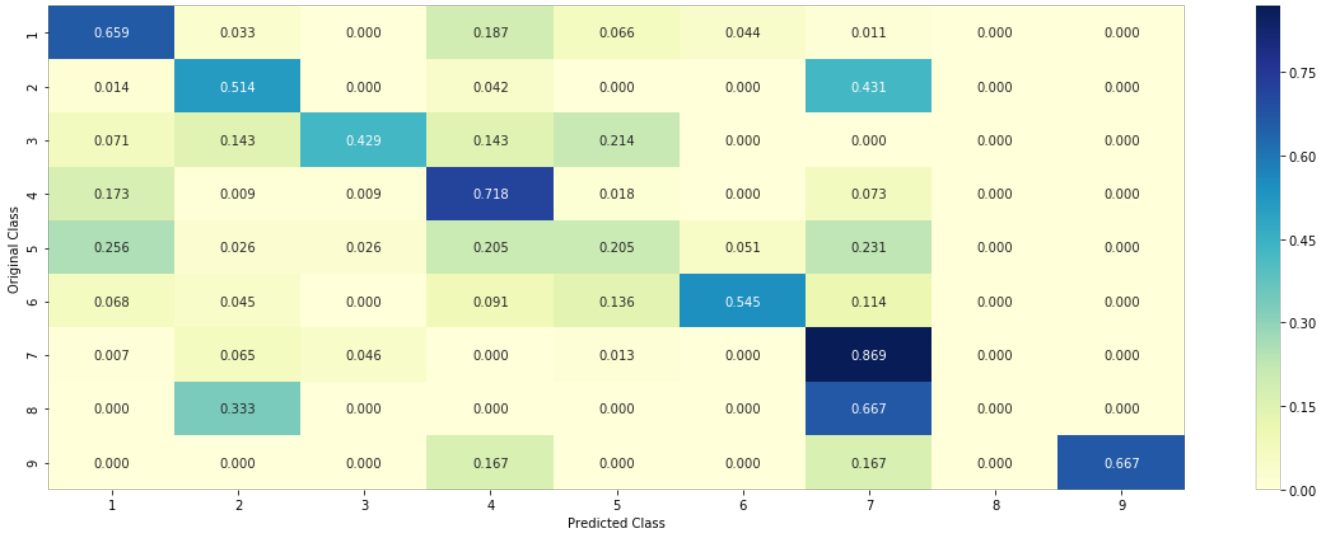
In [78]:

```
# read more about SGDClassifier() at http://scikit-
learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_i
ter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0
=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, …]) Fit linear model with Stochastic Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link:
#-------------------------------

clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y, cv_x_onehotCoding, cv_y, clf)
```

```
Log loss : 1.1148520648245654
Number of mis-classified points : 0.34022556390977443

-------------------- Confusion matrix --------------------
```

---------- Precision matrix (Columm Sum=1) ----------

---------- Recall matrix (Row sum=1) ----------



FeatureImportance

In [79]:

```
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_x_onehotCoding,train_y)
test_point_index = 100
no_feature = 1000
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
```

```
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_index]),4))
print("Actual Class :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 4
Predicted Class Probabilities: [[0.0533 0.037  0.0023 0.7504 0.0129 0.0065 0.1332 0.0031 0.0013]]
Actual Class : 4
--------------------------------------------------
119 Text feature [suppressor] present in test data point [True]
130 Text feature [devoid] present in test data point [True]
167 Text feature [shen] present in test data point [True]
191 Text feature [ovary] present in test data point [True]
312 Text feature [mgcl] present in test data point [True]
405 Text feature [germline] present in test data point [True]
427 Text feature [nonidet] present in test data point [True]
431 Text feature [eb1] present in test data point [True]
433 Text feature [bsteii] present in test data point [True]
472 Text feature [tgf] present in test data point [True]
475 Text feature [asymmetric] present in test data point [True]
501 Text feature [phosphatases] present in test data point [True]
502 Text feature [tagged] present in test data point [True]
507 Text feature [localization] present in test data point [True]
556 Text feature [bardeesy] present in test data point [True]
562 Text feature [6a] present in test data point [True]
574 Text feature [material] present in test data point [True]
616 Text feature [bind] present in test data point [True]
639 Text feature [ionizing] present in test data point [True]
669 Text feature [pten] present in test data point [True]
678 Text feature [inactivating] present in test data point [True]
689 Text feature [strengthen] present in test data point [True]
743 Text feature [billaud] present in test data point [True]
756 Text feature [tumorigenesis] present in test data point [True]
763 Text feature [cdk] present in test data point [True]
773 Text feature [heterozygosity] present in test data point [True]
783 Text feature [dominant] present in test data point [True]
790 Text feature [hereditary] present in test data point [True]
807 Text feature [hcl] present in test data point [True]
816 Text feature [mammalian] present in test data point [True]
819 Text feature [stk11] present in test data point [True]
836 Text feature [arrest] present in test data point [True]
845 Text feature [11b] present in test data point [True]
874 Text feature [analysed] present in test data point [True]
908 Text feature [missense] present in test data point [True]
928 Text feature [flip] present in test data point [True]
988 Text feature [diluent] present in test data point [True]
Out of the top  1000  features  37 are present in query point
```

In [80]:

```
test_point_index = 1
no_feature = 1000
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_index]),4))
print("Actual Class :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 1
Predicted Class Probabilities: [[6.622e-01 5.980e-02 7.000e-04 1.696e-01 1.230e-02 7.600e-02 1.690
e-02
  2.100e-03 4.000e-04]]
```

```
Actual Class : 1
--------------------------------------------------
237 Text feature [immobilized] present in test data point [True]
317 Text feature [aggregation] present in test data point [True]
386 Text feature [synergy] present in test data point [True]
494 Text feature [gal] present in test data point [True]
524 Text feature [aggregate] present in test data point [True]
560 Text feature [fas] present in test data point [True]
614 Text feature [destabilized] present in test data point [True]
698 Text feature [saos] present in test data point [True]
753 Text feature [deficient] present in test data point [True]
758 Text feature [aggregates] present in test data point [True]
808 Text feature [lysate] present in test data point [True]
876 Text feature [fail] present in test data point [True]
877 Text feature [unfolded] present in test data point [True]
914 Text feature [deletion] present in test data point [True]
939 Text feature [folded] present in test data point [True]
948 Text feature [feasibility] present in test data point [True]
Out of the top  1000  features  16 are present in query point
```

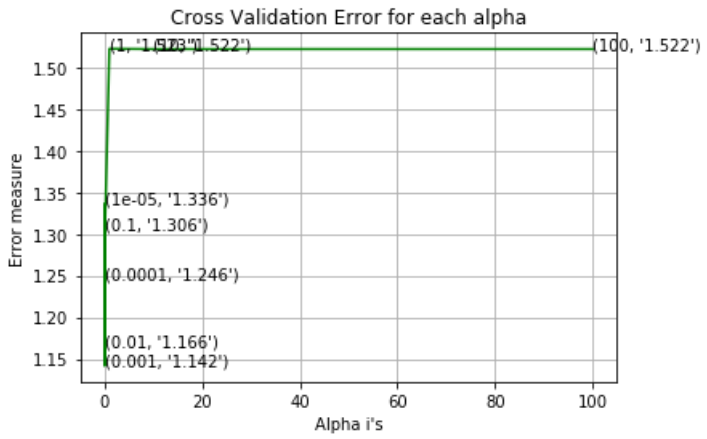
Linear Support Vector Machine

In [81]:

```
# read more about support vector machines with linear kernals here http://scikit-
learn.org/stable/modules/generated/sklearn.svm.SVC.html

# ----------------------------------
# default parameters
# SVC(C=1.0, kernel='rbf', degree=3, gamma='auto', coef0=0.0, shrinking=True, probability=False, t
ol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', ra
ndom_state=None)

# Some of methods of SVM()
# fit(X, y, [sample_weight]) Fit the SVM model according to the given training data.
# predict(X) Perform classification on samples in X.
# ----------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-
online/lessons/mathematical-derivation-copy-8/
# ----------------------------------



# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
#-------------------------------------
# video link:
#-------------------------------------

alpha = [10 ** x for x in range(-5, 3)]
cv_log_error_array = []
for i in alpha:
    print("for C =", i)
#     clf = SVC(C=i,kernel='linear',probability=True, class_weight='balanced')
    clf = SGDClassifier( class_weight='balanced', alpha=i, penalty='l2', loss='hinge', random_state
=42)
    clf.fit(train_x_onehotCoding, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_onehotCoding, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
    cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    print("Log Loss :",log_loss(cv_y, sig_clf_probs))

fig, ax = plt.subplots()
```

```
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],str(txt)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
# clf = SVC(C=i,kernel='linear',probability=True, class_weight='balanced')
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='hinge', r
andom_state=42)
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The train log loss is:",
      log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15))

predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The cross validation log loss is:",
      log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))

predict_y = sig_clf.predict_proba(test_x_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The test log loss is:",
      log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for C = 1e-05
Log Loss : 1.3358237402874038
for C = 0.0001
Log Loss : 1.2459074882326469
for C = 0.001
Log Loss : 1.1415185478977379
for C = 0.01
Log Loss : 1.1658462928531612
for C = 0.1
Log Loss : 1.3057530723351634
for C = 1
Log Loss : 1.5226488626928387
for C = 10
Log Loss : 1.5222364199812715
for C = 100
Log Loss : 1.5222364511280635
```



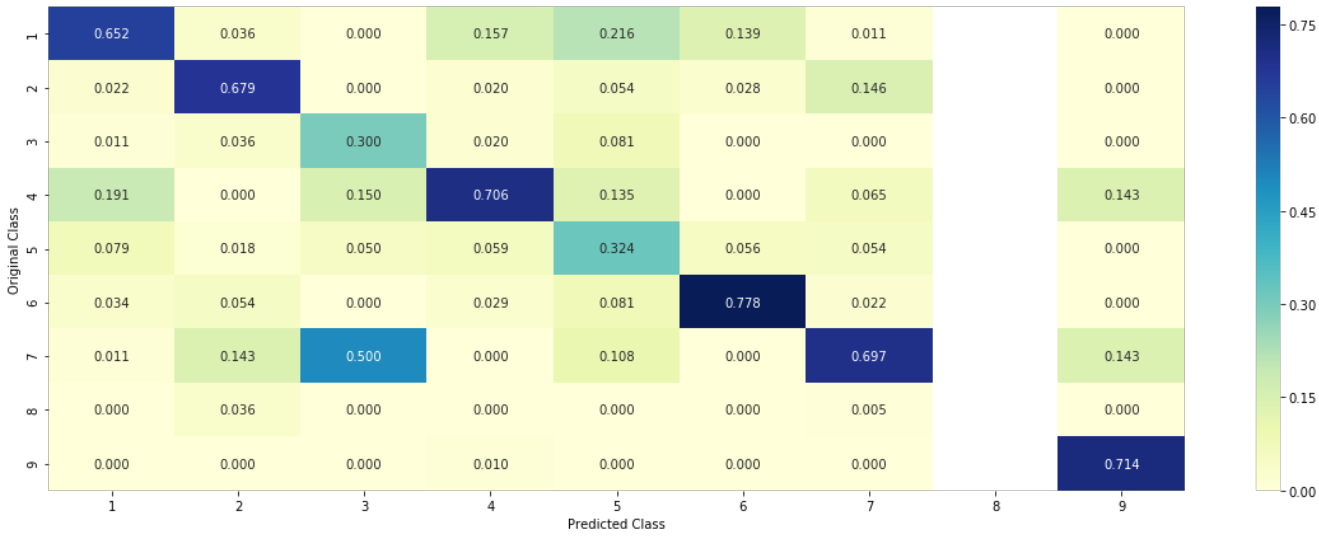

```
For values of best alpha =  0.001 The train log loss is: 0.5825419176978665
For values of best alpha =  0.001 The cross validation log loss is: 1.1415185478977379
For values of best alpha =  0.001 The test log loss is: 1.0158806372000406
```

In [82]:

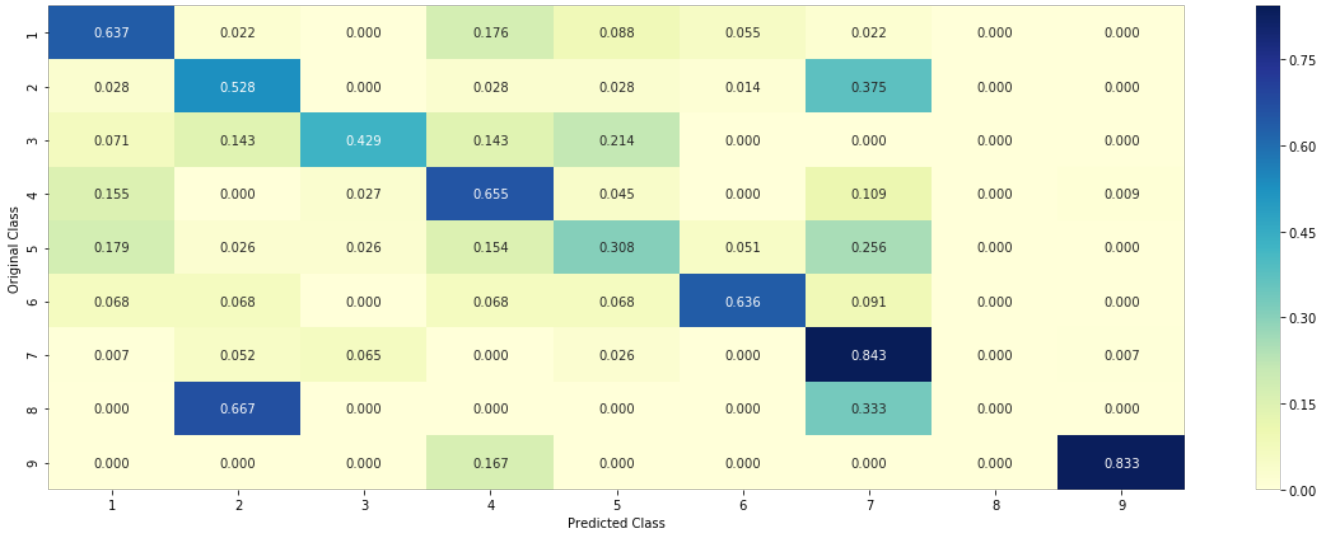
```
# read more about support vector machines with linear kernals here http://scikit-
learn.org/stable/modules/generated/sklearn.svm.SVC.html

# ----------------------------------
# default parameters
# SVC(C=1.0, kernel='rbf', degree=3, gamma='auto', coef0=0.0, shrinking=True, probability=False, t
ol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', ra
ndom_state=None)

# Some of methods of SVM()
# fit(X, y, [sample_weight]) Fit the SVM model according to the given training data.
# predict(X) Perform classification on samples in X.
# ----------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-
online/lessons/mathematical-derivation-copy-8/
# ----------------------------------


# clf = SVC(C=alpha[best_alpha],kernel='linear',probability=True, class_weight='balanced')
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='hinge',
random_state=42,class_weight='balanced')
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y,cv_x_onehotCoding,cv_y, clf)
```

```
Log loss : 1.1415185478977379
Number of mis-classified points : 0.3458646616541353

-------------------- Confusion matrix --------------------
```


```
-------------------- Precision matrix (Columm Sum=1) --------------------
```



```
-------------------- Recall matrix (Row sum=1) --------------------
```



Feature Importance

In [83]:

```
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='hinge', random_state=42)
clf.fit(train_x_onehotCoding,train_y)
test_point_index = 1
# test_point_index = 100
no_feature = 1000
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_index]),4))
print("Actual Class :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 1
Predicted Class Probabilities: [[0.3312 0.1303 0.0108 0.1193 0.0408 0.2191 0.1314 0.0052 0.0119]]
Actual Class : 1
--------------------------------------------------
392 Text feature [cancerous] present in test data point [True]
394 Text feature [aggregation] present in test data point [True]
404 Text feature [fas] present in test data point [True]
408 Text feature [immobilized] present in test data point [True]
462 Text feature [synergy] present in test data point [True]
536 Text feature [gal] present in test data point [True]
571 Text feature [lysate] present in test data point [True]
573 Text feature [germ] present in test data point [True]
716 Text feature [aggregates] present in test data point [True]
781 Text feature [fail] present in test data point [True]
842 Text feature [aggregate] present in test data point [True]
911 Text feature [feasibility] present in test data point [True]
941 Text feature [destabilized] present in test data point [True]
986 Text feature [unfolded] present in test data point [True]
Out of the top  1000  features  14 are present in query point
```

In [85]:

```
test_point_index = 55
no_feature = 1000
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_index]),4))
print("Actual Class :", test_y[test_point_index])
```

```
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0142 0.0597 0.0065 0.0326 0.0149 0.0036 0.8618 0.0024 0.0044]]
Actual Class : 7
--------------------------------------------------
195 Text feature [mir] present in test data point [True]
213 Text feature [constitutive] present in test data point [True]
246 Text feature [topical] present in test data point [True]
260 Text feature [oncogene] present in test data point [True]
273 Text feature [downstream] present in test data point [True]
278 Text feature [doublet] present in test data point [True]
288 Text feature [constitutively] present in test data point [True]
311 Text feature [activated] present in test data point [True]
348 Text feature [overexpression] present in test data point [True]
421 Text feature [expressing] present in test data point [True]
445 Text feature [kreb] present in test data point [True]
481 Text feature [adenocarcinoma] present in test data point [True]
498 Text feature [activation] present in test data point [True]
522 Text feature [colon] present in test data point [True]
553 Text feature [wako] present in test data point [True]
584 Text feature [2d] present in test data point [True]
591 Text feature [nanomolar] present in test data point [True]
624 Text feature [forestomach] present in test data point [True]
700 Text feature [membrane] present in test data point [True]
707 Text feature [26s] present in test data point [True]
722 Text feature [ectopically] present in test data point [True]
745 Text feature [confirmation] present in test data point [True]
786 Text feature [3a] present in test data point [True]
790 Text feature [enhancing] present in test data point [True]
792 Text feature [cells] present in test data point [True]
807 Text feature [elevated] present in test data point [True]
825 Text feature [proliferate] present in test data point [True]
875 Text feature [exemplified] present in test data point [True]
894 Text feature [2h] present in test data point [True]
895 Text feature [3b] present in test data point [True]
946 Text feature [nf] present in test data point [True]
951 Text feature [cap] present in test data point [True]
972 Text feature [gliomas] present in test data point [True]
982 Text feature [inhibited] present in test data point [True]
990 Text feature [enhance] present in test data point [True]
995 Text feature [absence] present in test data point [True]
998 Text feature [polyubiquitinated] present in test data point [True]
Out of the top  1000  features  37 are present in query point
```

Random Forest Classifier

With one-Hot Encoding

In [86]:

```
# ------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_s
amples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_
impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None,
verbose=0, warm_start=False,
# class_weight=None)

# Some of methods of RandomForestClassifier()
# fit(X, y, [sample_weight]) Fit the SVM model according to the given training data.
# predict(X) Perform classification on samples in X.
# predict_proba (X) Perform classification on samples in X.

# some of attributes of  RandomForestClassifier()
# feature_importances_ : array of shape = [n_features]
```

```
# The feature importances (the higher, the more important the feature).

# ---------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/random-fores
t-and-their-construction-2/
# ---------------------------------


# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
#-------------------------------------
# video link:
#-------------------------------------

alpha = [100,200,500,1000,2000]
max_depth = [5, 10]
cv_log_error_array = []
for i in alpha:
    for j in max_depth:
        print("for n_estimators =", i,"and max depth = ", j)
        clf = RandomForestClassifier(n_estimators=i, criterion='gini', max_depth=j, random_state=42
, n_jobs=-1)
        clf.fit(train_x_onehotCoding, train_y)
        sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
        sig_clf.fit(train_x_onehotCoding, train_y)
        sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
        cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
        print("Log Loss :",log_loss(cv_y, sig_clf_probs))

'''fig, ax = plt.subplots()
features = np.dot(np.array(alpha)[:,None],np.array(max_depth)[None]).ravel()
ax.plot(features, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[int(i/2)],max_depth[int(i%2)],str(txt)),
(features[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()
'''

best_alpha = np.argmin(cv_log_error_array)
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='gini', max_depth=max
_depth[int(best_alpha%2)], random_state=42, n_jobs=-1)
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_onehotCoding)
print('For values of best estimator = ',
      alpha[int(best_alpha/2)],
      "The train log loss is:",
      log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15))

predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
print('For values of best estimator = ',
      alpha[int(best_alpha/2)],
      "The cross validation log loss is:",
      log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))

predict_y = sig_clf.predict_proba(test_x_onehotCoding)
print('For values of best estimator = ',
      alpha[int(best_alpha/2)],
      "The test log loss is:",
      log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

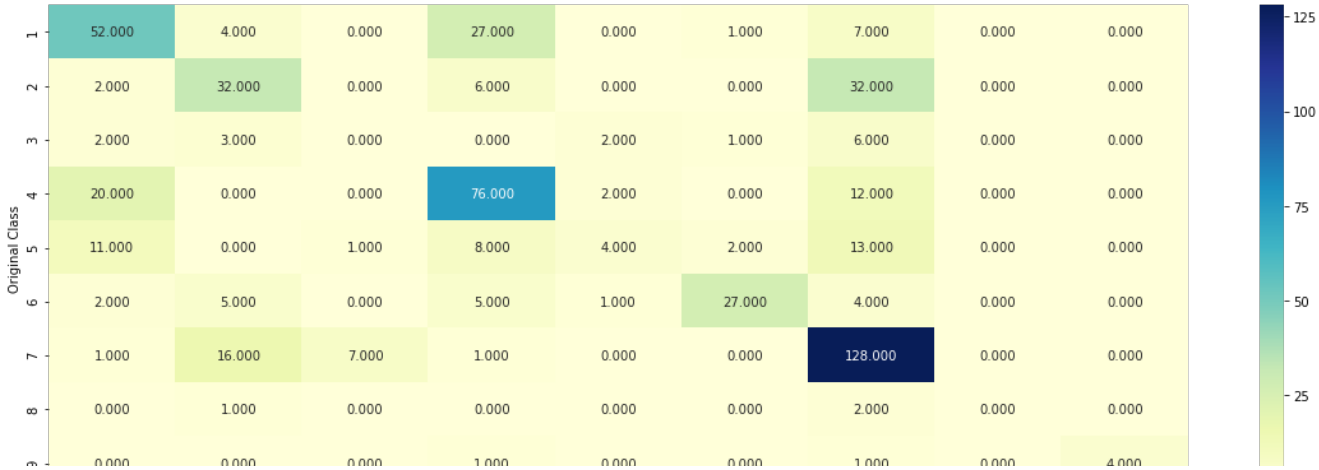
```
for n_estimators = 100 and max depth =  5
```

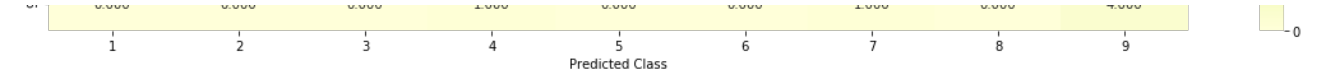
```
for n_estimators = 100 and max depth =  5
Log Loss : 1.239319569407836
for n_estimators = 100 and max depth =  10
Log Loss : 1.17375183690259
for n_estimators = 200 and max depth =  5
Log Loss : 1.2257761185916058
for n_estimators = 200 and max depth =  10
Log Loss : 1.1683034288206005
for n_estimators = 500 and max depth =  5
Log Loss : 1.2217249673939707
for n_estimators = 500 and max depth =  10
Log Loss : 1.166330073803563
for n_estimators = 1000 and max depth =  5
Log Loss : 1.2112895823718948
for n_estimators = 1000 and max depth =  10
Log Loss : 1.164698704776365
for n_estimators = 2000 and max depth =  5
Log Loss : 1.2093849266147776
for n_estimators = 2000 and max depth =  10
Log Loss : 1.1642484433800526
For values of best estimator =  2000 The train log loss is: 0.6651723996745306
For values of best estimator =  2000 The cross validation log loss is: 1.1642484433800528
For values of best estimator =  2000 The test log loss is: 1.1285626688221757
```

In [87]:

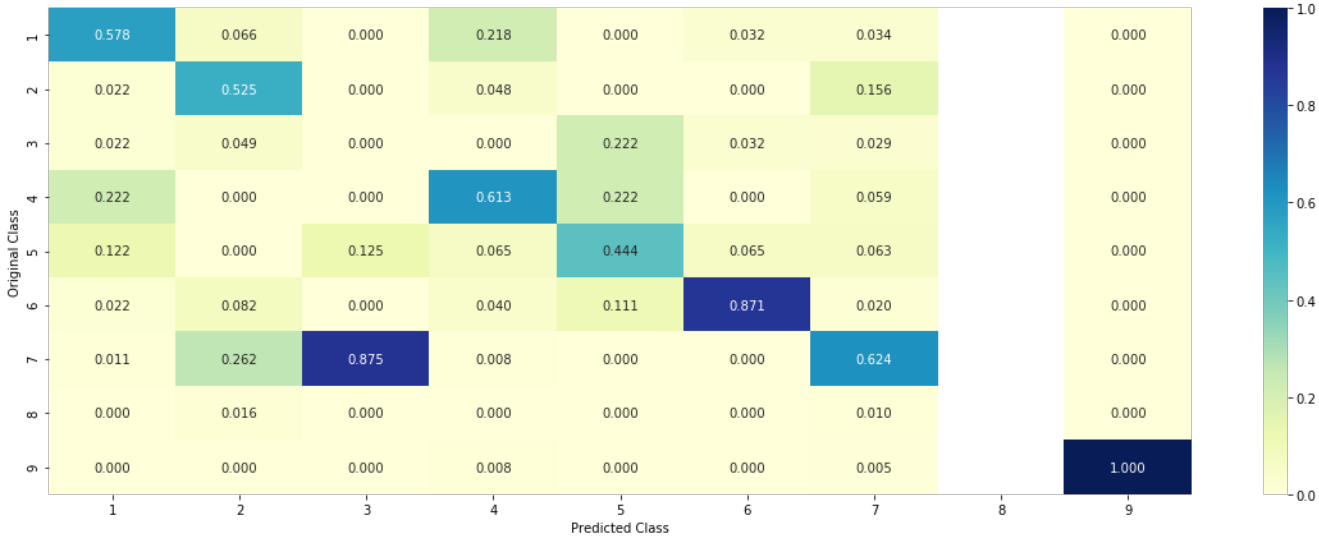
```
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_samples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None, verbose=0, warm_start=False,
# class_weight=None)

# Some of methods of RandomForestClassifier()
# fit(X, y, [sample_weight]) Fit the SVM model according to the given training data.
# predict(X) Perform classification on samples in X.
# predict_proba (X) Perform classification on samples in X.

# some of attributes of  RandomForestClassifier()b
# feature_importances_ : array of shape = [n_features]
# The feature importances (the higher, the more important the feature).

# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/random-forest-and-their-construction-2/
# --------------------------------

clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='gini', max_depth=max_depth[int(best_alpha%2)], random_state=42, n_jobs=-1)
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y,cv_x_onehotCoding,cv_y, clf)
```

```
Log loss : 1.1642484433800528
Number of mis-classified points : 0.39285714285714285

-------------------- Confusion matrix --------------------
```





---------- Precision matrix (Columm Sum=1) ----------



---------- Recall matrix (Row sum=1) ----------

Feature Importance

In [89]:

```
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='gini', max_depth=max
_depth[int(best_alpha%2)], random_state=42, n_jobs=-1)
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)

test_point_index = 1
no_feature = 1000
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_index]),4))
print("Actual Class :", test_y[test_point_index])
indices = np.argsort(-clf.feature_importances_)
print("-"*50)
get_impfeature_names(indices[:no_feature],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 1
Predicted Class Probabilities: [[0.5287 0.076  0.015  0.1944 0.0512 0.0572 0.0694 0.003  0.005 ]]
Actual Class : 1
--------------------------------------------------
0 Text feature [kinase] present in test data point [True]
1 Text feature [activating] present in test data point [True]
3 Text feature [activation] present in test data point [True]
4 Text feature [inhibitor] present in test data point [True]
5 Text feature [oncogenic] present in test data point [True]
6 Text feature [tyrosine] present in test data point [True]
7 Text feature [treatment] present in test data point [True]
8 Text feature [activated] present in test data point [True]
9 Text feature [downstream] present in test data point [True]
10 Text feature [phosphorylation] present in test data point [True]
12 Text feature [suppressor] present in test data point [True]
15 Text feature [drug] present in test data point [True]
17 Text feature [inhibited] present in test data point [True]
18 Text feature [loss] present in test data point [True]
19 Text feature [therapy] present in test data point [True]
20 Text feature [missense] present in test data point [True]
22 Text feature [function] present in test data point [True]
24 Text feature [growth] present in test data point [True]
25 Text feature [treated] present in test data point [True]
26 Text feature [trials] present in test data point [True]
27 Text feature [inhibition] present in test data point [True]
28 Text feature [cells] present in test data point [True]
30 Text feature [kinases] present in test data point [True]
33 Text feature [therapeutic] present in test data point [True]
36 Text feature [activate] present in test data point [True]
37 Text feature [yeast] present in test data point [True]
38 Text feature [protein] present in test data point [True]
39 Text feature [resistance] present in test data point [True]
40 Text feature [amplification] present in test data point [True]
41 Text feature [stability] present in test data point [True]
42 Text feature [proliferation] present in test data point [True]
43 Text feature [expressing] present in test data point [True]
44 Text feature [ligand] present in test data point [True]
47 Text feature [clinical] present in test data point [True]
48 Text feature [variants] present in test data point [True]
51 Text feature [defective] present in test data point [True]
55 Text feature [patients] present in test data point [True]
57 Text feature [active] present in test data point [True]
58 Text feature [functional] present in test data point [True]
63 Text feature [potential] present in test data point [True]
64 Text feature [factor] present in test data point [True]
68 Text feature [phosphorylated] present in test data point [True]
70 Text feature [oncogene] present in test data point [True]
71 Text feature [cell] present in test data point [True]
74 Text feature [retained] present in test data point [True]
82 Text feature [daily] present in test data point [True]
84 Text feature [proteins] present in test data point [True]
85 Text feature [response] present in test data point [True]
87 Text feature [pathway] present in test data point [True]
91 Text feature [repair] present in test data point [True]
92 Text feature [ability] present in test data point [True]
93 Text feature [null] present in test data point [True]
97 Text feature [effective] present in test data point [True]
102 Text feature [sensitivity] present in test data point [True]
104 Text feature [dna] present in test data point [True]
106 Text feature [presence] present in test data point [True]
110 Text feature [progression] present in test data point [True]
112 Text feature [use] present in test data point [True]
114 Text feature [independent] present in test data point [True]
117 Text feature [mechanism] present in test data point [True]
120 Text feature [nuclear] present in test data point [True]
121 Text feature [affect] present in test data point [True]
124 Text feature [chronic] present in test data point [True]
125 Text feature [inactivation] present in test data point [True]
126 Text feature [p53] present in test data point [True]
132 Text feature [classified] present in test data point [True]
137 Text feature [resistant] present in test data point [True]
138 Text feature [classify] present in test data point [True]
139 Text feature [sensitive] present in test data point [True]
140 Text feature [useful] present in test data point [True]
141 Text feature [expression] present in test data point [True]
142 Text feature [therapies] present in test data point [True]
```

```
143 Text feature [core] present in test data point [True]
145 Text feature [tumors] present in test data point [True]
146 Text feature [sequencing] present in test data point [True]
148 Text feature [achieved] present in test data point [True]
153 Text feature [functions] present in test data point [True]
155 Text feature [assay] present in test data point [True]
156 Text feature [primary] present in test data point [True]
157 Text feature [results] present in test data point [True]
161 Text feature [absence] present in test data point [True]
164 Text feature [likelihood] present in test data point [True]
167 Text feature [receptors] present in test data point [True]
171 Text feature [bind] present in test data point [True]
173 Text feature [mutant] present in test data point [True]
175 Text feature [mammalian] present in test data point [True]
178 Text feature [activity] present in test data point [True]
181 Text feature [approved] present in test data point [True]
182 Text feature [potency] present in test data point [True]
184 Text feature [abl] present in test data point [True]
185 Text feature [recently] present in test data point [True]
186 Text feature [interaction] present in test data point [True]
187 Text feature [sequence] present in test data point [True]
188 Text feature [enhanced] present in test data point [True]
193 Text feature [conservation] present in test data point [True]
195 Text feature [molecular] present in test data point [True]
196 Text feature [abrogate] present in test data point [True]
200 Text feature [concentrations] present in test data point [True]
203 Text feature [likely] present in test data point [True]
204 Text feature [11] present in test data point [True]
205 Text feature [information] present in test data point [True]
209 Text feature [contrast] present in test data point [True]
210 Text feature [folding] present in test data point [True]
213 Text feature [arrest] present in test data point [True]
214 Text feature [kit] present in test data point [True]
215 Text feature [based] present in test data point [True]
221 Text feature [novel] present in test data point [True]
223 Text feature [terminal] present in test data point [True]
224 Text feature [partial] present in test data point [True]
227 Text feature [bcr] present in test data point [True]
229 Text feature [family] present in test data point [True]
230 Text feature [laboratories] present in test data point [True]
231 Text feature [study] present in test data point [True]
232 Text feature [defined] present in test data point [True]
238 Text feature [small] present in test data point [True]
242 Text feature [conserved] present in test data point [True]
245 Text feature [pathways] present in test data point [True]
246 Text feature [majority] present in test data point [True]
247 Text feature [12] present in test data point [True]
249 Text feature [currently] present in test data point [True]
253 Text feature [database] present in test data point [True]
254 Text feature [24] present in test data point [True]
255 Text feature [domain] present in test data point [True]
257 Text feature [expected] present in test data point [True]
259 Text feature [damage] present in test data point [True]
260 Text feature [mice] present in test data point [True]
261 Text feature [large] present in test data point [True]
262 Text feature [wild] present in test data point [True]
265 Text feature [binding] present in test data point [True]
270 Text feature [21] present in test data point [True]
272 Text feature [loop] present in test data point [True]
273 Text feature [mutants] present in test data point [True]
276 Text feature [14] present in test data point [True]
279 Text feature [days] present in test data point [True]
285 Text feature [common] present in test data point [True]
286 Text feature [structural] present in test data point [True]
287 Text feature [hours] present in test data point [True]
288 Text feature [known] present in test data point [True]
289 Text feature [antibodies] present in test data point [True]
290 Text feature [anti] present in test data point [True]
291 Text feature [used] present in test data point [True]
292 Text feature [line] present in test data point [True]
294 Text feature [target] present in test data point [True]
301 Text feature [levels] present in test data point [True]
302 Text feature [higher] present in test data point [True]
305 Text feature [experiments] present in test data point [True]
307 Text feature [13] present in test data point [True]
308 Text feature [site] present in test data point [True]
309 Text feature [also] present in test data point [True]
```

```
312 Text feature [well] present in test data point [True]
313 Text feature [conformation] present in test data point [True]
314 Text feature [transcriptional] present in test data point [True]
316 Text feature [showed] present in test data point [True]
321 Text feature [potent] present in test data point [True]
324 Text feature [hcl] present in test data point [True]
325 Text feature [34] present in test data point [True]
326 Text feature [effect] present in test data point [True]
327 Text feature [several] present in test data point [True]
328 Text feature [transfected] present in test data point [True]
331 Text feature [type] present in test data point [True]
334 Text feature [disrupt] present in test data point [True]
342 Text feature [initial] present in test data point [True]
346 Text feature [16] present in test data point [True]
347 Text feature [suppressors] present in test data point [True]
348 Text feature [28] present in test data point [True]
349 Text feature [vitro] present in test data point [True]
353 Text feature [gene] present in test data point [True]
354 Text feature [human] present in test data point [True]
356 Text feature [inhibitory] present in test data point [True]
358 Text feature [residues] present in test data point [True]
359 Text feature [co] present in test data point [True]
361 Text feature [substitutions] present in test data point [True]
362 Text feature [including] present in test data point [True]
364 Text feature [many] present in test data point [True]
365 Text feature [specific] present in test data point [True]
368 Text feature [membrane] present in test data point [True]
370 Text feature [fold] present in test data point [True]
371 Text feature [53] present in test data point [True]
373 Text feature [amino] present in test data point [True]
374 Text feature [relative] present in test data point [True]
375 Text feature [01] present in test data point [True]
376 Text feature [localization] present in test data point [True]
378 Text feature [demonstrated] present in test data point [True]
379 Text feature [identify] present in test data point [True]
380 Text feature [region] present in test data point [True]
386 Text feature [antibody] present in test data point [True]
387 Text feature [within] present in test data point [True]
388 Text feature [first] present in test data point [True]
390 Text feature [full] present in test data point [True]
391 Text feature [vector] present in test data point [True]
394 Text feature [conversely] present in test data point [True]
395 Text feature [using] present in test data point [True]
396 Text feature [containing] present in test data point [True]
397 Text feature [significant] present in test data point [True]
399 Text feature [positive] present in test data point [True]
403 Text feature [68] present in test data point [True]
406 Text feature [shown] present in test data point [True]
407 Text feature [genes] present in test data point [True]
408 Text feature [acid] present in test data point [True]
411 Text feature [deficient] present in test data point [True]
414 Text feature [64] present in test data point [True]
419 Text feature [however] present in test data point [True]
424 Text feature [reporter] present in test data point [True]
427 Text feature [obtained] present in test data point [True]
430 Text feature [46] present in test data point [True]
431 Text feature [mutations] present in test data point [True]
433 Text feature [43] present in test data point [True]
434 Text feature [indicated] present in test data point [True]
435 Text feature [overall] present in test data point [True]
437 Text feature [17] present in test data point [True]
439 Text feature [promote] present in test data point [True]
440 Text feature [values] present in test data point [True]
441 Text feature [signal] present in test data point [True]
442 Text feature [observed] present in test data point [True]
447 Text feature [expressed] present in test data point [True]
449 Text feature [identified] present in test data point [True]
451 Text feature [ng] present in test data point [True]
455 Text feature [data] present in test data point [True]
458 Text feature [changes] present in test data point [True]
462 Text feature [methods] present in test data point [True]
463 Text feature [number] present in test data point [True]
464 Text feature [22] present in test data point [True]
465 Text feature [serves] present in test data point [True]
466 Text feature [thus] present in test data point [True]
470 Text feature [compounds] present in test data point [True]
473 Text feature [significantly] present in test data point [True]
```

```
474 Text feature [promising] present in test data point [True]
479 Text feature [model] present in test data point [True]
480 Text feature [group] present in test data point [True]
481 Text feature [cases] present in test data point [True]
484 Text feature [limited] present in test data point [True]
485 Text feature [transcription] present in test data point [True]
486 Text feature [found] present in test data point [True]
491 Text feature [surface] present in test data point [True]
492 Text feature [trans] present in test data point [True]
493 Text feature [figure] present in test data point [True]
494 Text feature [destabilized] present in test data point [True]
504 Text feature [flanking] present in test data point [True]
505 Text feature [whether] present in test data point [True]
509 Text feature [15] present in test data point [True]
514 Text feature [27] present in test data point [True]
515 Text feature [25] present in test data point [True]
517 Text feature [effects] present in test data point [True]
521 Text feature [differences] present in test data point [True]
523 Text feature [studies] present in test data point [True]
524 Text feature [gain] present in test data point [True]
525 Text feature [culture] present in test data point [True]
528 Text feature [50] present in test data point [True]
529 Text feature [length] present in test data point [True]
530 Text feature [three] present in test data point [True]
531 Text feature [somatic] present in test data point [True]
532 Text feature [required] present in test data point [True]
533 Text feature [estimated] present in test data point [True]
534 Text feature [indicate] present in test data point [True]
536 Text feature [60] present in test data point [True]
538 Text feature [published] present in test data point [True]
540 Text feature [42] present in test data point [True]
541 Text feature [approximately] present in test data point [True]
542 Text feature [free] present in test data point [True]
543 Text feature [20] present in test data point [True]
544 Text feature [drugs] present in test data point [True]
545 Text feature [increased] present in test data point [True]
547 Text feature [discussion] present in test data point [True]
550 Text feature [substitution] present in test data point [True]
556 Text feature [interact] present in test data point [True]
557 Text feature [structure] present in test data point [True]
558 Text feature [tumor] present in test data point [True]
559 Text feature [defects] present in test data point [True]
560 Text feature [may] present in test data point [True]
562 Text feature [cancer] present in test data point [True]
565 Text feature [70] present in test data point [True]
566 Text feature [analysis] present in test data point [True]
567 Text feature [suppression] present in test data point [True]
568 Text feature [although] present in test data point [True]
570 Text feature [occur] present in test data point [True]
571 Text feature [previously] present in test data point [True]
577 Text feature [induced] present in test data point [True]
578 Text feature [fell] present in test data point [True]
581 Text feature [agents] present in test data point [True]
582 Text feature [chemotherapy] present in test data point [True]
583 Text feature [show] present in test data point [True]
585 Text feature [mechanisms] present in test data point [True]
586 Text feature [sequences] present in test data point [True]
587 Text feature [frequency] present in test data point [True]
589 Text feature [table] present in test data point [True]
592 Text feature [35] present in test data point [True]
593 Text feature [18] present in test data point [True]
594 Text feature [identification] present in test data point [True]
595 Text feature [inactivate] present in test data point [True]
596 Text feature [cycle] present in test data point [True]
597 Text feature [40] present in test data point [True]
602 Text feature [two] present in test data point [True]
605 Text feature [evidence] present in test data point [True]
607 Text feature [deletion] present in test data point [True]
608 Text feature [approach] present in test data point [True]
609 Text feature [range] present in test data point [True]
610 Text feature [according] present in test data point [True]
611 Text feature [controls] present in test data point [True]
614 Text feature [determined] present in test data point [True]
615 Text feature [indicates] present in test data point [True]
616 Text feature [important] present in test data point [True]
619 Text feature [interacts] present in test data point [True]
621 Text feature [highly] present in test data point [True]
```

```
622 Text feature [repeats] present in test data point [True]
623 Text feature [remaining] present in test data point [True]
625 Text feature [similarly] present in test data point [True]
629 Text feature [mutation] present in test data point [True]
631 Text feature [fig] present in test data point [True]
636 Text feature [case] present in test data point [True]
637 Text feature [involved] present in test data point [True]
639 Text feature [mg] present in test data point [True]
640 Text feature [increase] present in test data point [True]
643 Text feature [stage] present in test data point [True]
644 Text feature [resource] present in test data point [True]
646 Text feature [allows] present in test data point [True]
648 Text feature [introduction] present in test data point [True]
649 Text feature [one] present in test data point [True]
650 Text feature [examined] present in test data point [True]
653 Text feature [different] present in test data point [True]
655 Text feature [new] present in test data point [True]
664 Text feature [high] present in test data point [True]
667 Text feature [addition] present in test data point [True]
668 Text feature [compared] present in test data point [True]
670 Text feature [made] present in test data point [True]
671 Text feature [control] present in test data point [True]
673 Text feature [require] present in test data point [True]
675 Text feature [among] present in test data point [True]
676 Text feature [introduced] present in test data point [True]
678 Text feature [would] present in test data point [True]
682 Text feature [role] present in test data point [True]
686 Text feature [frequently] present in test data point [True]
688 Text feature [point] present in test data point [True]
689 Text feature [dominant] present in test data point [True]
690 Text feature [furthermore] present in test data point [True]
691 Text feature [constructs] present in test data point [True]
694 Text feature [reported] present in test data point [True]
695 Text feature [critical] present in test data point [True]
698 Text feature [sites] present in test data point [True]
700 Text feature [lysis] present in test data point [True]
701 Text feature [lack] present in test data point [True]
703 Text feature [10] present in test data point [True]
704 Text feature [100] present in test data point [True]
708 Text feature [include] present in test data point [True]
711 Text feature [29] present in test data point [True]
713 Text feature [central] present in test data point [True]
717 Text feature [washed] present in test data point [True]
718 Text feature [strategies] present in test data point [True]
719 Text feature [complex] present in test data point [True]
720 Text feature [properties] present in test data point [True]
722 Text feature [evaluate] present in test data point [True]
723 Text feature [interestingly] present in test data point [True]
726 Text feature [times] present in test data point [True]
728 Text feature [respectively] present in test data point [True]
729 Text feature [single] present in test data point [True]
730 Text feature [types] present in test data point [True]
734 Text feature [deletions] present in test data point [True]
737 Text feature [detected] present in test data point [True]
739 Text feature [lost] present in test data point [True]
740 Text feature [selected] present in test data point [True]
743 Text feature [less] present in test data point [True]
747 Text feature [still] present in test data point [True]
750 Text feature [106] present in test data point [True]
751 Text feature [either] present in test data point [True]
753 Text feature [suppressive] present in test data point [True]
756 Text feature [class] present in test data point [True]
760 Text feature [form] present in test data point [True]
761 Text feature [possible] present in test data point [True]
762 Text feature [fact] present in test data point [True]
763 Text feature [vivo] present in test data point [True]
764 Text feature [four] present in test data point [True]
767 Text feature [result] present in test data point [True]
775 Text feature [steric] present in test data point [True]
776 Text feature [72] present in test data point [True]
777 Text feature [cause] present in test data point [True]
778 Text feature [specificity] present in test data point [True]
780 Text feature [resulting] present in test data point [True]
781 Text feature [could] present in test data point [True]
782 Text feature [least] present in test data point [True]
783 Text feature [essential] present in test data point [True]
784 Text feature [described] present in test data point [True]
```

785 Text feature [factors] present in test data point [True]
786 Text feature [represent] present in test data point [True]
793 Text feature [regulatory] present in test data point [True]
794 Text feature [xenograft] present in test data point [True]
795 Text feature [90] present in test data point [True]
796 Text feature [recognizes] present in test data point [True]
797 Text feature [fraction] present in test data point [True]
799 Text feature [analyzed] present in test data point [True]
800 Text feature [administration] present in test data point [True]
802 Text feature [degradation] present in test data point [True]
803 Text feature [domains] present in test data point [True]
804 Text feature [alterations] present in test data point [True]
805 Text feature [low] present in test data point [True]
808 Text feature [possibly] present in test data point [True]
809 Text feature [finally] present in test data point [True]
811 Text feature [potentially] present in test data point [True]
813 Text feature [without] present in test data point [True]
814 Text feature [activates] present in test data point [True]
815 Text feature [typically] present in test data point [True]
817 Text feature [similar] present in test data point [True]
819 Text feature [reports] present in test data point [True]
820 Text feature [screening] present in test data point [True]
822 Text feature [non] present in test data point [True]
824 Text feature [buffer] present in test data point [True]
825 Text feature [targets] present in test data point [True]
826 Text feature [randomly] present in test data point [True]
827 Text feature [molecule] present in test data point [True]
828 Text feature [example] present in test data point [True]
832 Text feature [hydrophobic] present in test data point [True]
833 Text feature [partially] present in test data point [True]
836 Text feature [67] present in test data point [True]
840 Text feature [system] present in test data point [True]
841 Text feature [genetic] present in test data point [True]
842 Text feature [apoptosis] present in test data point [True]
843 Text feature [half] present in test data point [True]
845 Text feature [considered] present in test data point [True]
846 Text feature [transfection] present in test data point [True]
848 Text feature [screen] present in test data point [True]
850 Text feature [later] present in test data point [True]
852 Text feature [inhibit] present in test data point [True]
853 Text feature [eight] present in test data point [True]
854 Text feature [distribution] present in test data point [True]
856 Text feature [26] present in test data point [True]
857 Text feature [directly] present in test data point [True]
859 Text feature [51] present in test data point [True]
862 Text feature [side] present in test data point [True]
863 Text feature [76] present in test data point [True]
864 Text feature [32] present in test data point [True]
865 Text feature [consistent] present in test data point [True]
867 Text feature [44] present in test data point [True]
868 Text feature [probably] present in test data point [True]
871 Text feature [allele] present in test data point [True]
873 Text feature [regulation] present in test data point [True]
874 Text feature [evaluation] present in test data point [True]
875 Text feature [promoter] present in test data point [True]
876 Text feature [hypothesis] present in test data point [True]
877 Text feature [roles] present in test data point [True]
878 Text feature [interactions] present in test data point [True]
879 Text feature [mouse] present in test data point [True]
880 Text feature [importance] present in test data point [True]
885 Text feature [plasmid] present in test data point [True]
887 Text feature [previous] present in test data point [True]
888 Text feature [decreased] present in test data point [True]
890 Text feature [mutated] present in test data point [True]
891 Text feature [suggested] present in test data point [True]
893 Text feature [might] present in test data point [True]
894 Text feature [individual] present in test data point [True]
898 Text feature [ii] present in test data point [True]
902 Text feature [sequenced] present in test data point [True]
903 Text feature [characterized] present in test data point [True]
904 Text feature [unable] present in test data point [True]
908 Text feature [subset] present in test data point [True]
909 Text feature [predict] present in test data point [True]
911 Text feature [virus] present in test data point [True]
914 Text feature [six] present in test data point [True]
915 Text feature [associated] present in test data point [True]
918 Text feature [following] present in test data point [True]

```
920 Text feature [present] present in test data point [True]
928 Text feature [blue] present in test data point [True]
930 Text feature [correlation] present in test data point [True]
932 Text feature [measured] present in test data point [True]
933 Text feature [seeded] present in test data point [True]
935 Text feature [average] present in test data point [True]
938 Text feature [provide] present in test data point [True]
939 Text feature [evaluated] present in test data point [True]
940 Text feature [effectors] present in test data point [True]
941 Text feature [motif] present in test data point [True]
942 Text feature [phenotype] present in test data point [True]
946 Text feature [dependent] present in test data point [True]
947 Text feature [derived] present in test data point [True]
950 Text feature [indeed] present in test data point [True]
951 Text feature [events] present in test data point [True]
954 Text feature [sds] present in test data point [True]
955 Text feature [lower] present in test data point [True]
956 Text feature [tp53] present in test data point [True]
958 Text feature [whereas] present in test data point [True]
961 Text feature [complete] present in test data point [True]
963 Text feature [wide] present in test data point [True]
964 Text feature [account] present in test data point [True]
965 Text feature [rather] present in test data point [True]
966 Text feature [52] present in test data point [True]
967 Text feature [level] present in test data point [True]
970 Text feature [confirmed] present in test data point [True]
971 Text feature [per] present in test data point [True]
973 Text feature [cultured] present in test data point [True]
977 Text feature [cancers] present in test data point [True]
978 Text feature [classes] present in test data point [True]
980 Text feature [nucleus] present in test data point [True]
984 Text feature [ml] present in test data point [True]
988 Text feature [close] present in test data point [True]
990 Text feature [negative] present in test data point [True]
993 Text feature [change] present in test data point [True]
996 Text feature [review] present in test data point [True]
997 Text feature [normal] present in test data point [True]
998 Text feature [examine] present in test data point [True]
Out of the top  1000  features  495 are present in query point
```

In [91]:

```
test_point_index = 15
no_feature = 1000
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_index]),4))
print("Actuall Class :", test_y[test_point_index])
indices = np.argsort(-clf.feature_importances_)
print("-"*50)
get_impfeature_names(indices[:no_feature],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.052  0.2635 0.0166 0.0537 0.0492 0.0481 0.5095 0.003  0.0043]]
Actuall Class : 2
--------------------------------------------------
0 Text feature [kinase] present in test data point [True]
1 Text feature [activating] present in test data point [True]
2 Text feature [constitutive] present in test data point [True]
3 Text feature [activation] present in test data point [True]
4 Text feature [inhibitor] present in test data point [True]
5 Text feature [oncogenic] present in test data point [True]
6 Text feature [tyrosine] present in test data point [True]
7 Text feature [treatment] present in test data point [True]
8 Text feature [activated] present in test data point [True]
9 Text feature [downstream] present in test data point [True]
10 Text feature [phosphorylation] present in test data point [True]
11 Text feature [receptor] present in test data point [True]
13 Text feature [inhibitors] present in test data point [True]
14 Text feature [nonsense] present in test data point [True]
```

```
15 Text feature [drug] present in test data point [True]
17 Text feature [inhibited] present in test data point [True]
18 Text feature [loss] present in test data point [True]
19 Text feature [therapy] present in test data point [True]
20 Text feature [missense] present in test data point [True]
21 Text feature [constitutively] present in test data point [True]
22 Text feature [function] present in test data point [True]
23 Text feature [erk] present in test data point [True]
24 Text feature [growth] present in test data point [True]
25 Text feature [treated] present in test data point [True]
27 Text feature [inhibition] present in test data point [True]
28 Text feature [cells] present in test data point [True]
30 Text feature [kinases] present in test data point [True]
31 Text feature [akt] present in test data point [True]
33 Text feature [therapeutic] present in test data point [True]
35 Text feature [advanced] present in test data point [True]
36 Text feature [activate] present in test data point [True]
38 Text feature [protein] present in test data point [True]
41 Text feature [stability] present in test data point [True]
42 Text feature [proliferation] present in test data point [True]
43 Text feature [expressing] present in test data point [True]
45 Text feature [phospho] present in test data point [True]
47 Text feature [clinical] present in test data point [True]
52 Text feature [ic50] present in test data point [True]
53 Text feature [serum] present in test data point [True]
55 Text feature [patients] present in test data point [True]
57 Text feature [active] present in test data point [True]
58 Text feature [functional] present in test data point [True]
60 Text feature [ras] present in test data point [True]
63 Text feature [potential] present in test data point [True]
64 Text feature [factor] present in test data point [True]
66 Text feature [predicted] present in test data point [True]
68 Text feature [phosphorylated] present in test data point [True]
70 Text feature [oncogene] present in test data point [True]
71 Text feature [cell] present in test data point [True]
78 Text feature [mek] present in test data point [True]
79 Text feature [mitogen] present in test data point [True]
81 Text feature [egfr] present in test data point [True]
84 Text feature [proteins] present in test data point [True]
85 Text feature [response] present in test data point [True]
87 Text feature [pathway] present in test data point [True]
93 Text feature [null] present in test data point [True]
96 Text feature [erk1] present in test data point [True]
97 Text feature [effective] present in test data point [True]
99 Text feature [survival] present in test data point [True]
100 Text feature [starved] present in test data point [True]
101 Text feature [nsclc] present in test data point [True]
102 Text feature [sensitivity] present in test data point [True]
103 Text feature [patient] present in test data point [True]
104 Text feature [dna] present in test data point [True]
105 Text feature [lines] present in test data point [True]
106 Text feature [presence] present in test data point [True]
107 Text feature [benefit] present in test data point [True]
109 Text feature [metastatic] present in test data point [True]
110 Text feature [progression] present in test data point [True]
111 Text feature [pten] present in test data point [True]
113 Text feature [trial] present in test data point [True]
114 Text feature [independent] present in test data point [True]
115 Text feature [variant] present in test data point [True]
117 Text feature [mechanism] present in test data point [True]
119 Text feature [membranes] present in test data point [True]
121 Text feature [affect] present in test data point [True]
123 Text feature [uncertain] present in test data point [True]
124 Text feature [chronic] present in test data point [True]
128 Text feature [acquired] present in test data point [True]
130 Text feature [assays] present in test data point [True]
131 Text feature [doses] present in test data point [True]
132 Text feature [classified] present in test data point [True]
139 Text feature [sensitive] present in test data point [True]
141 Text feature [expression] present in test data point [True]
142 Text feature [therapies] present in test data point [True]
143 Text feature [core] present in test data point [True]
146 Text feature [sequencing] present in test data point [True]
147 Text feature [responses] present in test data point [True]
153 Text feature [functions] present in test data point [True]
155 Text feature [assay] present in test data point [True]
156 Text feature [primary] present in test data point [True]
```

```
157 Text feature [results] present in test data point [True]
161 Text feature [absence] present in test data point [True]
163 Text feature [median] present in test data point [True]
166 Text feature [atp] present in test data point [True]
172 Text feature [biopsy] present in test data point [True]
173 Text feature [mutant] present in test data point [True]
175 Text feature [mammalian] present in test data point [True]
178 Text feature [activity] present in test data point [True]
181 Text feature [approved] present in test data point [True]
183 Text feature [braf] present in test data point [True]
185 Text feature [recently] present in test data point [True]
187 Text feature [sequence] present in test data point [True]
195 Text feature [molecular] present in test data point [True]
198 Text feature [stimulation] present in test data point [True]
200 Text feature [concentrations] present in test data point [True]
201 Text feature [lung] present in test data point [True]
203 Text feature [likely] present in test data point [True]
204 Text feature [11] present in test data point [True]
209 Text feature [contrast] present in test data point [True]
211 Text feature [raf] present in test data point [True]
214 Text feature [kit] present in test data point [True]
215 Text feature [based] present in test data point [True]
216 Text feature [il] present in test data point [True]
217 Text feature [leukemia] present in test data point [True]
220 Text feature [epidermal] present in test data point [True]
221 Text feature [novel] present in test data point [True]
222 Text feature [classification] present in test data point [True]
230 Text feature [laboratories] present in test data point [True]
231 Text feature [study] present in test data point [True]
233 Text feature [leading] present in test data point [True]
238 Text feature [small] present in test data point [True]
239 Text feature [pi3k] present in test data point [True]
245 Text feature [pathways] present in test data point [True]
246 Text feature [majority] present in test data point [True]
247 Text feature [12] present in test data point [True]
249 Text feature [currently] present in test data point [True]
251 Text feature [individuals] present in test data point [True]
254 Text feature [24] present in test data point [True]
255 Text feature [domain] present in test data point [True]
257 Text feature [expected] present in test data point [True]
261 Text feature [large] present in test data point [True]
262 Text feature [wild] present in test data point [True]
265 Text feature [binding] present in test data point [True]
270 Text feature [21] present in test data point [True]
272 Text feature [loop] present in test data point [True]
274 Text feature [reduced] present in test data point [True]
276 Text feature [14] present in test data point [True]
283 Text feature [western] present in test data point [True]
285 Text feature [common] present in test data point [True]
288 Text feature [known] present in test data point [True]
289 Text feature [antibodies] present in test data point [True]
290 Text feature [anti] present in test data point [True]
291 Text feature [used] present in test data point [True]
292 Text feature [line] present in test data point [True]
294 Text feature [target] present in test data point [True]
297 Text feature [strand] present in test data point [True]
301 Text feature [levels] present in test data point [True]
302 Text feature [higher] present in test data point [True]
305 Text feature [experiments] present in test data point [True]
307 Text feature [13] present in test data point [True]
308 Text feature [site] present in test data point [True]
309 Text feature [also] present in test data point [True]
312 Text feature [well] present in test data point [True]
316 Text feature [showed] present in test data point [True]
321 Text feature [potent] present in test data point [True]
322 Text feature [combined] present in test data point [True]
324 Text feature [hcl] present in test data point [True]
325 Text feature [34] present in test data point [True]
326 Text feature [effect] present in test data point [True]
327 Text feature [several] present in test data point [True]
328 Text feature [transfected] present in test data point [True]
330 Text feature [tissue] present in test data point [True]
331 Text feature [type] present in test data point [True]
342 Text feature [initial] present in test data point [True]
345 Text feature [technology] present in test data point [True]
346 Text feature [16] present in test data point [True]
348 Text feature [28] present in test data point [True]
```

353 Text feature [gene] present in test data point [True]
354 Text feature [human] present in test data point [True]
355 Text feature [determine] present in test data point [True]
356 Text feature [inhibitory] present in test data point [True]
359 Text feature [co] present in test data point [True]
360 Text feature [assessment] present in test data point [True]
362 Text feature [including] present in test data point [True]
364 Text feature [many] present in test data point [True]
365 Text feature [specific] present in test data point [True]
367 Text feature [time] present in test data point [True]
370 Text feature [fold] present in test data point [True]
373 Text feature [amino] present in test data point [True]
374 Text feature [relative] present in test data point [True]
378 Text feature [demonstrated] present in test data point [True]
379 Text feature [identify] present in test data point [True]
380 Text feature [region] present in test data point [True]
386 Text feature [antibody] present in test data point [True]
387 Text feature [within] present in test data point [True]
388 Text feature [first] present in test data point [True]
389 Text feature [tested] present in test data point [True]
390 Text feature [full] present in test data point [True]
391 Text feature [vector] present in test data point [True]
393 Text feature [49] present in test data point [True]
395 Text feature [using] present in test data point [True]
396 Text feature [containing] present in test data point [True]
397 Text feature [significant] present in test data point [True]
399 Text feature [positive] present in test data point [True]
402 Text feature [alk] present in test data point [True]
404 Text feature [2b] present in test data point [True]
406 Text feature [shown] present in test data point [True]
407 Text feature [genes] present in test data point [True]
408 Text feature [acid] present in test data point [True]
411 Text feature [deficient] present in test data point [True]
414 Text feature [64] present in test data point [True]
418 Text feature [event] present in test data point [True]
419 Text feature [however] present in test data point [True]
420 Text feature [homozygous] present in test data point [True]
422 Text feature [2a] present in test data point [True]
427 Text feature [obtained] present in test data point [True]
431 Text feature [mutations] present in test data point [True]
434 Text feature [indicated] present in test data point [True]
435 Text feature [overall] present in test data point [True]
437 Text feature [17] present in test data point [True]
438 Text feature [due] present in test data point [True]
439 Text feature [promote] present in test data point [True]
440 Text feature [values] present in test data point [True]
441 Text feature [signal] present in test data point [True]
442 Text feature [observed] present in test data point [True]
443 Text feature [developed] present in test data point [True]
447 Text feature [expressed] present in test data point [True]
449 Text feature [identified] present in test data point [True]
450 Text feature [79] present in test data point [True]
451 Text feature [ng] present in test data point [True]
455 Text feature [data] present in test data point [True]
458 Text feature [changes] present in test data point [True]
459 Text feature [history] present in test data point [True]
460 Text feature [1a] present in test data point [True]
461 Text feature [various] present in test data point [True]
462 Text feature [methods] present in test data point [True]
463 Text feature [number] present in test data point [True]
464 Text feature [22] present in test data point [True]
466 Text feature [thus] present in test data point [True]
470 Text feature [compounds] present in test data point [True]
473 Text feature [significantly] present in test data point [True]
474 Text feature [promising] present in test data point [True]
475 Text feature [included] present in test data point [True]
476 Text feature [general] present in test data point [True]
477 Text feature [33] present in test data point [True]
480 Text feature [group] present in test data point [True]
481 Text feature [cases] present in test data point [True]
482 Text feature [pcr] present in test data point [True]
484 Text feature [limited] present in test data point [True]
485 Text feature [transcription] present in test data point [True]
486 Text feature [found] present in test data point [True]
487 Text feature [36] present in test data point [True]
490 Text feature [oncogenes] present in test data point [True]
493 Text feature [figure] present in test data point [True]

500 Text feature [none] present in test data point [True]
505 Text feature [whether] present in test data point [True]
506 Text feature [31] present in test data point [True]
509 Text feature [15] present in test data point [True]
510 Text feature [multiple] present in test data point [True]
514 Text feature [27] present in test data point [True]
515 Text feature [25] present in test data point [True]
516 Text feature [available] present in test data point [True]
517 Text feature [effects] present in test data point [True]
523 Text feature [studies] present in test data point [True]
525 Text feature [culture] present in test data point [True]
527 Text feature [performed] present in test data point [True]
528 Text feature [50] present in test data point [True]
530 Text feature [three] present in test data point [True]
532 Text feature [required] present in test data point [True]
534 Text feature [indicate] present in test data point [True]
538 Text feature [published] present in test data point [True]
540 Text feature [42] present in test data point [True]
541 Text feature [approximately] present in test data point [True]
542 Text feature [free] present in test data point [True]
543 Text feature [20] present in test data point [True]
544 Text feature [drugs] present in test data point [True]
545 Text feature [increased] present in test data point [True]
546 Text feature [samples] present in test data point [True]
547 Text feature [discussion] present in test data point [True]
548 Text feature [suggesting] present in test data point [True]
560 Text feature [may] present in test data point [True]
562 Text feature [cancer] present in test data point [True]
565 Text feature [70] present in test data point [True]
566 Text feature [analysis] present in test data point [True]
567 Text feature [suppression] present in test data point [True]
568 Text feature [although] present in test data point [True]
570 Text feature [occur] present in test data point [True]
571 Text feature [previously] present in test data point [True]
574 Text feature [acids] present in test data point [True]
577 Text feature [induced] present in test data point [True]
579 Text feature [exon] present in test data point [True]
581 Text feature [agents] present in test data point [True]
583 Text feature [show] present in test data point [True]
584 Text feature [improved] present in test data point [True]
585 Text feature [mechanisms] present in test data point [True]
587 Text feature [frequency] present in test data point [True]
589 Text feature [table] present in test data point [True]
591 Text feature [specimens] present in test data point [True]
592 Text feature [35] present in test data point [True]
593 Text feature [18] present in test data point [True]
594 Text feature [identification] present in test data point [True]
596 Text feature [cycle] present in test data point [True]
597 Text feature [40] present in test data point [True]
598 Text feature [mutational] present in test data point [True]
601 Text feature [findings] present in test data point [True]
602 Text feature [two] present in test data point [True]
603 Text feature [coding] present in test data point [True]
605 Text feature [evidence] present in test data point [True]
607 Text feature [deletion] present in test data point [True]
608 Text feature [approach] present in test data point [True]
609 Text feature [range] present in test data point [True]
610 Text feature [according] present in test data point [True]
611 Text feature [controls] present in test data point [True]
614 Text feature [determined] present in test data point [True]
615 Text feature [indicates] present in test data point [True]
616 Text feature [important] present in test data point [True]
618 Text feature [comparison] present in test data point [True]
620 Text feature [rate] present in test data point [True]
621 Text feature [highly] present in test data point [True]
625 Text feature [similarly] present in test data point [True]
626 Text feature [support] present in test data point [True]
628 Text feature [significance] present in test data point [True]
629 Text feature [mutation] present in test data point [True]
636 Text feature [case] present in test data point [True]
640 Text feature [increase] present in test data point [True]
641 Text feature [vehicle] present in test data point [True]
642 Text feature [54] present in test data point [True]
643 Text feature [stage] present in test data point [True]
645 Text feature [targeted] present in test data point [True]
648 Text feature [introduction] present in test data point [True]
649 Text feature [one] present in test data point [True]

650 Text feature [examined] present in test data point [True]
653 Text feature [different] present in test data point [True]
654 Text feature [need] present in test data point [True]
655 Text feature [new] present in test data point [True]
660 Text feature [final] present in test data point [True]
663 Text feature [targeting] present in test data point [True]
664 Text feature [high] present in test data point [True]
665 Text feature [lysates] present in test data point [True]
666 Text feature [report] present in test data point [True]
667 Text feature [addition] present in test data point [True]
668 Text feature [compared] present in test data point [True]
669 Text feature [material] present in test data point [True]
671 Text feature [control] present in test data point [True]
675 Text feature [among] present in test data point [True]
678 Text feature [would] present in test data point [True]
682 Text feature [role] present in test data point [True]
685 Text feature [standard] present in test data point [True]
687 Text feature [suggests] present in test data point [True]
688 Text feature [point] present in test data point [True]
689 Text feature [dominant] present in test data point [True]
690 Text feature [furthermore] present in test data point [True]
694 Text feature [reported] present in test data point [True]
695 Text feature [critical] present in test data point [True]
696 Text feature [characterization] present in test data point [True]
699 Text feature [population] present in test data point [True]
700 Text feature [lysis] present in test data point [True]
701 Text feature [lack] present in test data point [True]
703 Text feature [10] present in test data point [True]
704 Text feature [100] present in test data point [True]
709 Text feature [statistical] present in test data point [True]
711 Text feature [29] present in test data point [True]
715 Text feature [kras] present in test data point [True]
717 Text feature [washed] present in test data point [True]
718 Text feature [strategies] present in test data point [True]
719 Text feature [complex] present in test data point [True]
720 Text feature [properties] present in test data point [True]
722 Text feature [evaluate] present in test data point [True]
723 Text feature [interestingly] present in test data point [True]
724 Text feature [panel] present in test data point [True]
726 Text feature [times] present in test data point [True]
727 Text feature [position] present in test data point [True]
728 Text feature [respectively] present in test data point [True]
729 Text feature [single] present in test data point [True]
732 Text feature [4a] present in test data point [True]
735 Text feature [applied] present in test data point [True]
736 Text feature [driven] present in test data point [True]
737 Text feature [detected] present in test data point [True]
740 Text feature [selected] present in test data point [True]
741 Text feature [disease] present in test data point [True]
743 Text feature [less] present in test data point [True]
744 Text feature [epithelial] present in test data point [True]
745 Text feature [studied] present in test data point [True]
749 Text feature [characteristics] present in test data point [True]
751 Text feature [either] present in test data point [True]
752 Text feature [assessed] present in test data point [True]
755 Text feature [38] present in test data point [True]
756 Text feature [class] present in test data point [True]
758 Text feature [elevated] present in test data point [True]
759 Text feature [medium] present in test data point [True]
760 Text feature [form] present in test data point [True]
761 Text feature [possible] present in test data point [True]
764 Text feature [four] present in test data point [True]
765 Text feature [19] present in test data point [True]
766 Text feature [stop] present in test data point [True]
769 Text feature [provided] present in test data point [True]
771 Text feature [total] present in test data point [True]
772 Text feature [related] present in test data point [True]
777 Text feature [cause] present in test data point [True]
778 Text feature [specificity] present in test data point [True]
781 Text feature [could] present in test data point [True]
782 Text feature [least] present in test data point [True]
783 Text feature [essential] present in test data point [True]
784 Text feature [described] present in test data point [True]
785 Text feature [factors] present in test data point [True]
786 Text feature [represent] present in test data point [True]
787 Text feature [set] present in test data point [True]
789 Text feature [heterozygosity] present in test data point [True]

```
795 Text feature [90] present in test data point [True]
797 Text feature [fraction] present in test data point [True]
798 Text feature [generated] present in test data point [True]
799 Text feature [analyzed] present in test data point [True]
805 Text feature [low] present in test data point [True]
807 Text feature [47] present in test data point [True]
808 Text feature [possibly] present in test data point [True]
811 Text feature [potentially] present in test data point [True]
813 Text feature [without] present in test data point [True]
814 Text feature [activates] present in test data point [True]
816 Text feature [phase] present in test data point [True]
817 Text feature [similar] present in test data point [True]
819 Text feature [reports] present in test data point [True]
820 Text feature [screening] present in test data point [True]
822 Text feature [non] present in test data point [True]
824 Text feature [buffer] present in test data point [True]
825 Text feature [targets] present in test data point [True]
827 Text feature [molecule] present in test data point [True]
837 Text feature [papillary] present in test data point [True]
838 Text feature [thyroid] present in test data point [True]
841 Text feature [genetic] present in test data point [True]
845 Text feature [considered] present in test data point [True]
848 Text feature [screen] present in test data point [True]
853 Text feature [eight] present in test data point [True]
854 Text feature [distribution] present in test data point [True]
855 Text feature [discovery] present in test data point [True]
856 Text feature [26] present in test data point [True]
859 Text feature [51] present in test data point [True]
863 Text feature [76] present in test data point [True]
864 Text feature [32] present in test data point [True]
865 Text feature [consistent] present in test data point [True]
871 Text feature [allele] present in test data point [True]
873 Text feature [regulation] present in test data point [True]
876 Text feature [hypothesis] present in test data point [True]
879 Text feature [mouse] present in test data point [True]
880 Text feature [importance] present in test data point [True]
884 Text feature [added] present in test data point [True]
887 Text feature [previous] present in test data point [True]
888 Text feature [decreased] present in test data point [True]
889 Text feature [chromosome] present in test data point [True]
890 Text feature [mutated] present in test data point [True]
891 Text feature [suggested] present in test data point [True]
893 Text feature [might] present in test data point [True]
902 Text feature [sequenced] present in test data point [True]
908 Text feature [subset] present in test data point [True]
911 Text feature [virus] present in test data point [True]
912 Text feature [led] present in test data point [True]
914 Text feature [six] present in test data point [True]
915 Text feature [associated] present in test data point [True]
916 Text feature [established] present in test data point [True]
918 Text feature [following] present in test data point [True]
919 Text feature [driver] present in test data point [True]
920 Text feature [present] present in test data point [True]
922 Text feature [proportion] present in test data point [True]
926 Text feature [genome] present in test data point [True]
927 Text feature [ratios] present in test data point [True]
928 Text feature [blue] present in test data point [True]
932 Text feature [measured] present in test data point [True]
938 Text feature [provide] present in test data point [True]
940 Text feature [effectors] present in test data point [True]
941 Text feature [motif] present in test data point [True]
944 Text feature [components] present in test data point [True]
946 Text feature [dependent] present in test data point [True]
947 Text feature [derived] present in test data point [True]
949 Text feature [exons] present in test data point [True]
951 Text feature [events] present in test data point [True]
952 Text feature [focused] present in test data point [True]
954 Text feature [sds] present in test data point [True]
955 Text feature [lower] present in test data point [True]
956 Text feature [tp53] present in test data point [True]
960 Text feature [consideration] present in test data point [True]
961 Text feature [complete] present in test data point [True]
962 Text feature [s1] present in test data point [True]
964 Text feature [account] present in test data point [True]
966 Text feature [52] present in test data point [True]
967 Text feature [level] present in test data point [True]
969 Text feature [greater] present in test data point [True]
```

```
970 Text feature [confirmed] present in test data point [True]
973 Text feature [cultured] present in test data point [True]
974 Text feature [rearrangements] present in test data point [True]
977 Text feature [cancers] present in test data point [True]
982 Text feature [sample] present in test data point [True]
985 Text feature [testing] present in test data point [True]
986 Text feature [investigate] present in test data point [True]
990 Text feature [negative] present in test data point [True]
992 Text feature [features] present in test data point [True]
994 Text feature [note] present in test data point [True]
997 Text feature [normal] present in test data point [True]
998 Text feature [examine] present in test data point [True]
Out of the top  1000  features  488 are present in query point
```

With RESPONSE CODING

In [93]:

```
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_s
amples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_
impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None,
verbose=0, warm_start=False,
# class_weight=None)

# Some of methods of RandomForestClassifier()
# fit(X, y, [sample_weight]) Fit the SVM model according to the given training data.
# predict(X) Perform classification on samples in X.
# predict_proba (X) Perform classification on samples in X.

# some of attributes of  RandomForestClassifier()
# feature_importances_ : array of shape = [n_features]
# The feature importances (the higher, the more important the feature).

# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/random-fores
t-and-their-construction-2/
# --------------------------------


# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
#-----------------------------------
# video link:
#-----------------------------------

alpha = [10,50,100,200,500,1000]
max_depth = [2,3,5,10]
cv_log_error_array = []
for i in alpha:
    for j in max_depth:
        print("for n_estimators =", i,"and max depth = ", j)
        clf = RandomForestClassifier(n_estimators=i, criterion='gini', max_depth=j, random_state=42
, n_jobs=-1)
        clf.fit(train_x_responseCoding, train_y)
        sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
        sig_clf.fit(train_x_responseCoding, train_y)
        sig_clf_probs = sig_clf.predict_proba(cv_x_responseCoding)
        cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
        print("Log Loss :",log_loss(cv_y, sig_clf_probs))
'''
fig, ax = plt.subplots()
features = np.dot(np.array(alpha)[:,None],np.array(max_depth)[None]).ravel()
ax.plot(features, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[int(i/4)],max_depth[int(i%4)],str(txt))
```

```
    ax.annotate((alpha[int(i/4)],max_depth[int(i%4)]),str(txt))
(features[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()
'''

best_alpha = np.argmin(cv_log_error_array)
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/4)], criterion='gini', max_depth=max
_depth[int(best_alpha%4)], random_state=42, n_jobs=-1)
clf.fit(train_x_responseCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_responseCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_responseCoding)
print('For values of best alpha = ',
      alpha[int(best_alpha/4)],
      "The train log loss is:",
      log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15))

predict_y = sig_clf.predict_proba(cv_x_responseCoding)
print('For values of best alpha = ',
      alpha[int(best_alpha/4)],
      "The cross validation log loss is:",
      log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))

predict_y = sig_clf.predict_proba(test_x_responseCoding)
print('For values of best alpha = ',
      alpha[int(best_alpha/4)],
      "The test log loss is:",
      log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for n_estimators = 10 and max depth =  2
Log Loss : 2.1855266967685165
for n_estimators = 10 and max depth =  3
Log Loss : 1.9010938678427525
for n_estimators = 10 and max depth =  5
Log Loss : 1.5302721692598613
for n_estimators = 10 and max depth =  10
Log Loss : 2.1547146678012976
for n_estimators = 50 and max depth =  2
Log Loss : 1.6860676000646313
for n_estimators = 50 and max depth =  3
Log Loss : 1.4886297793635241
for n_estimators = 50 and max depth =  5
Log Loss : 1.4171539065905359
for n_estimators = 50 and max depth =  10
Log Loss : 1.7664497977690754
for n_estimators = 100 and max depth =  2
Log Loss : 1.5512688350319863
for n_estimators = 100 and max depth =  3
Log Loss : 1.526330133228856
for n_estimators = 100 and max depth =  5
Log Loss : 1.293317563092307
for n_estimators = 100 and max depth =  10
Log Loss : 1.808932008021683
for n_estimators = 200 and max depth =  2
Log Loss : 1.5669229288951905
for n_estimators = 200 and max depth =  3
Log Loss : 1.4912703411536
for n_estimators = 200 and max depth =  5
Log Loss : 1.3589150203553668
for n_estimators = 200 and max depth =  10
Log Loss : 1.7873934945503012
for n_estimators = 500 and max depth =  2
Log Loss : 1.6458704732571374
for n_estimators = 500 and max depth =  3
Log Loss : 1.5538691687767354
for n_estimators = 500 and max depth =  5
Log Loss : 1.4021273311053173
for n_estimators = 500 and max depth =  10
Log Loss : 1.7236286784642227
for n_estimators = 1000 and max depth =  2
Log Loss : 1.6293163927223382
for n_estimators = 1000 and max depth =  3
```

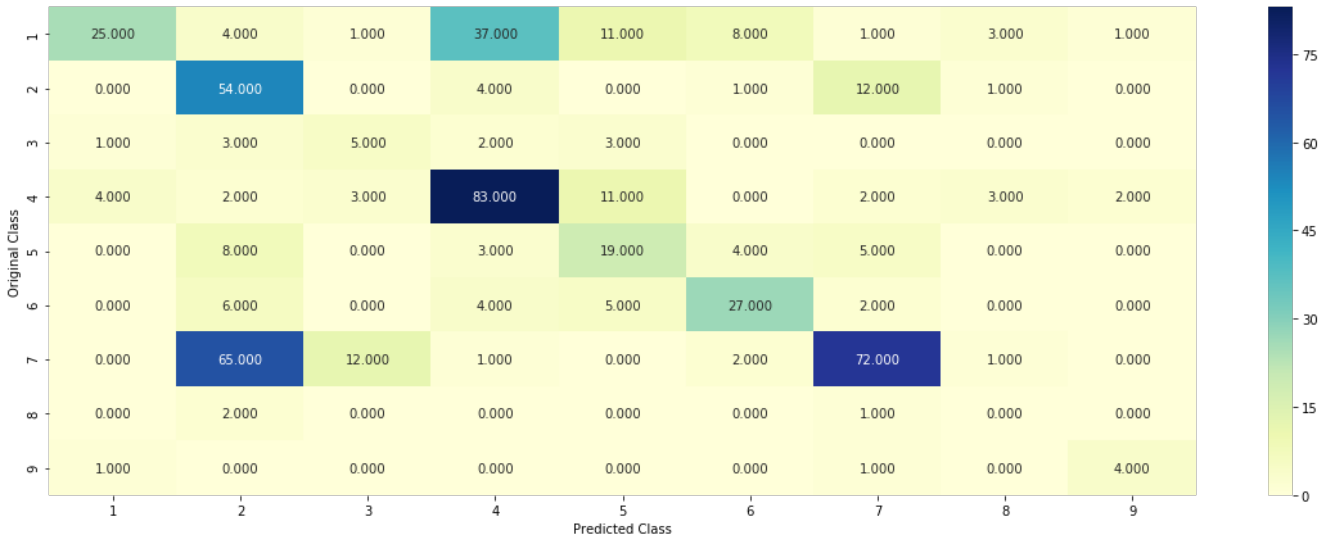
```
Log Loss : 1.5461433652380694
for n_estimators = 1000 and max depth =  5
Log Loss : 1.380841879065371
for n_estimators = 1000 and max depth =  10
Log Loss : 1.7144491281602734
For values of best alpha =  100 The train log loss is: 0.0540709413077688
For values of best alpha =  100 The cross validation log loss is: 1.293317563092307
For values of best alpha =  100 The test log loss is: 1.2518685898035833
```

In [94]:

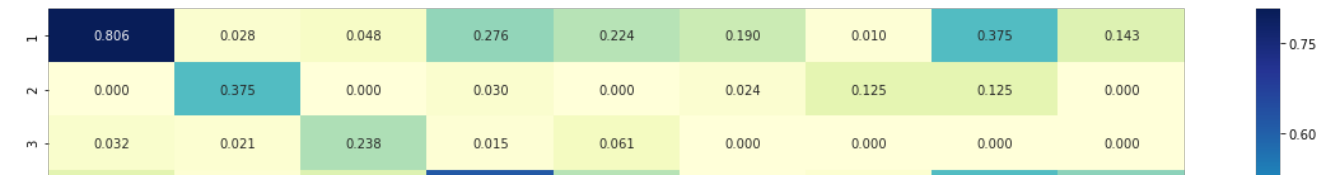
```
# ------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_s
amples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_
impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None,
verbose=0, warm_start=False,
# class_weight=None)

# Some of methods of RandomForestClassifier()
# fit(X, y, [sample_weight]) Fit the SVM model according to the given training data.
# predict(X) Perform classification on samples in X.
# predict_proba (X) Perform classification on samples in X.

# some of attributes of  RandomForestClassifier()
# feature_importances_ : array of shape = [n_features]
# The feature importances (the higher, the more important the feature).

# ------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/random-fores
t-and-their-construction-2/
# ------------------------------

clf = RandomForestClassifier(max_depth=max_depth[int(best_alpha%4)],
n_estimators=alpha[int(best_alpha/4)], criterion='gini', max_features='auto',random_state=42)
predict_and_plot_confusion_matrix(train_x_responseCoding, train_y,cv_x_responseCoding,cv_y, clf)
```

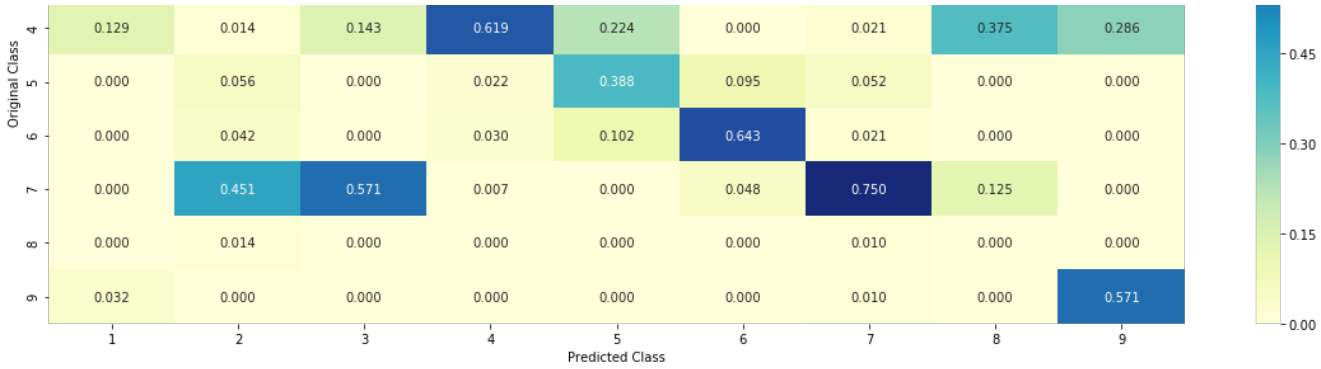
```
Log loss : 1.293317563092307
Number of mis-classified points : 0.4567669172932331
```

```
-------------------- Confusion matrix --------------------
```

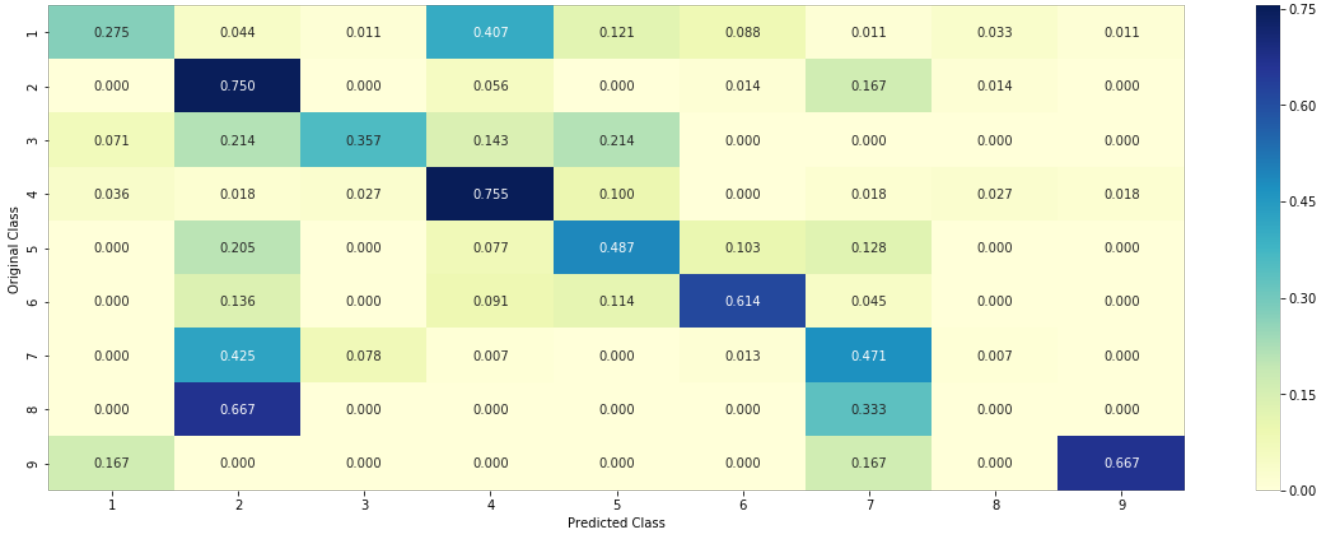



```
-------------------- Precision matrix (Column Sum=1) --------------------
```





---------- Recall matrix (Row sum=1) ----------



Feature Importance

In [95]:

```
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/4)], criterion='gini', max_depth=max
_depth[int(best_alpha%4)], random_state=42, n_jobs=-1)
clf.fit(train_x_responseCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_responseCoding, train_y)


test_point_index = 100
no_feature = 1000
predicted_cls = sig_clf.predict(test_x_responseCoding[test_point_index].reshape(1,-1))
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_responseCoding[test_point_index].reshape(1,-1)),4))
print("Actual Class :", test_y[test_point_index])
# indices = np.argsort(-clf.feature_importances_)
# print("-"*50)
# for i in indices:
#     if i<9:
#         print("Gene is important feature")
#     elif i<18:
#         print("Variation is important feature")
#     else:
#         print("Text is important feature")
```

```
Predicted Class : 4
Predicted Class Probabilities: [[0.1201 0.0289 0.1662 0.494  0.04   0.0589 0.009  0.0435 0.0394]]
Actual Class : 4
```

In [96]:

```
test_point_index = 31
predicted_cls = sig_clf.predict(test_x_responseCoding[test_point_index].reshape(1,-1))
print("Predicted Class :", predicted_cls[0])
```

```
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_responseCoding[test_point_index].reshape(1,-1)),4))
print("Actual Class :", test_y[test_point_index])
# indices = np.argsort(-clf.feature_importances_)
# print("-"*50)
# for i in indices:
#     if i<9:
#         print("Gene is important feature")
#     elif i<18:
#         print("Variation is important feature")
#     else:
#         print("Text is important feature")
```

```
Predicted Class : 4
Predicted Class Probabilities: [[0.0929 0.0338 0.1887 0.5164 0.0356 0.0437 0.0076 0.0393 0.042 ]]
Actual Class : 4
```

Stack The Models

In [97]:

```
# read more about SGDClassifier() at http://scikit-
learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_i
ter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0
=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, …]) Fit linear model with Stochastic Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/geometric-in
tuition-1/
#-------------------------------


# read more about support vector machines with linear kernals here http://scikit-
learn.org/stable/modules/generated/sklearn.svm.SVC.html
# ---------------------------------
# default parameters
# SVC(C=1.0, kernel='rbf', degree=3, gamma='auto', coef0=0.0, shrinking=True, probability=False, t
ol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', ra
ndom_state=None)

# Some of methods of SVM()
# fit(X, y, [sample_weight]) Fit the SVM model according to the given training data.
# predict(X) Perform classification on samples in X.
# ---------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-
online/lessons/mathematical-derivation-copy-8/
# ---------------------------------


# read more about support vector machines with linear kernals here http://scikit-
learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
# ---------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_s
amples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_
impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None,
verbose=0, warm_start=False,
# class_weight=None)

# Some of methods of RandomForestClassifier()
# fit(X, y, [sample_weight]) Fit the SVM model according to the given training data.
```

```
# predict(X) Perform classification on samples in X.
# predict_proba (X) Perform classification on samples in X.

# some of attributes of  RandomForestClassifier()
# feature_importances_ : array of shape = [n_features]
# The feature importances (the higher, the more important the feature).

# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/random-forest-and-their-construction-2/
# --------------------------------


clf1 = SGDClassifier(alpha=0.001, penalty='l2', loss='log', class_weight='balanced', random_state=0)
clf1.fit(train_x_onehotCoding, train_y)
sig_clf1 = CalibratedClassifierCV(clf1, method="sigmoid")

clf2 = SGDClassifier(alpha=0.01, penalty='l2', loss='hinge', class_weight='balanced', random_state=0)
clf2.fit(train_x_onehotCoding, train_y)
sig_clf2 = CalibratedClassifierCV(clf2, method="sigmoid")


clf3 = MultinomialNB(alpha=1000)
clf3.fit(train_x_onehotCoding, train_y)
sig_clf3 = CalibratedClassifierCV(clf3, method="sigmoid")

sig_clf1.fit(train_x_onehotCoding, train_y)
print("Logistic Regression :  Log Loss: %0.2f" % (log_loss(cv_y, sig_clf1.predict_proba(cv_x_onehotCoding))))
sig_clf2.fit(train_x_onehotCoding, train_y)
print("Support vector machines : Log Loss: %0.2f" % (log_loss(cv_y, sig_clf2.predict_proba(cv_x_onehotCoding))))
sig_clf3.fit(train_x_onehotCoding, train_y)
print("Naive Bayes : Log Loss: %0.2f" % (log_loss(cv_y, sig_clf3.predict_proba(cv_x_onehotCoding))))
print("-"*50)
alpha = [0.0001,0.001,0.01,0.1,1,10]
best_alpha = 999
for i in alpha:
    lr = LogisticRegression(C=i)
    sclf = StackingClassifier(classifiers=[sig_clf1, sig_clf2, sig_clf3], meta_classifier=lr, use_probas=True)
    sclf.fit(train_x_onehotCoding, train_y)
    print("Stacking Classifer : for the value of alpha: %f Log Loss: %0.3f" % (i, log_loss(cv_y, sclf.predict_proba(cv_x_onehotCoding))))
    log_error =log_loss(cv_y, sclf.predict_proba(cv_x_onehotCoding))
    if best_alpha > log_error:
        best_alpha = log_error
```

```
Logistic Regression :  Log Loss: 1.11
Support vector machines : Log Loss: 1.17
Naive Bayes : Log Loss: 1.22
--------------------------------------------------
Stacking Classifer : for the value of alpha: 0.000100 Log Loss: 2.173
Stacking Classifer : for the value of alpha: 0.001000 Log Loss: 1.994
Stacking Classifer : for the value of alpha: 0.010000 Log Loss: 1.406
Stacking Classifer : for the value of alpha: 0.100000 Log Loss: 1.091
Stacking Classifer : for the value of alpha: 1.000000 Log Loss: 1.221
Stacking Classifer : for the value of alpha: 10.000000 Log Loss: 1.565
```

In [98]:

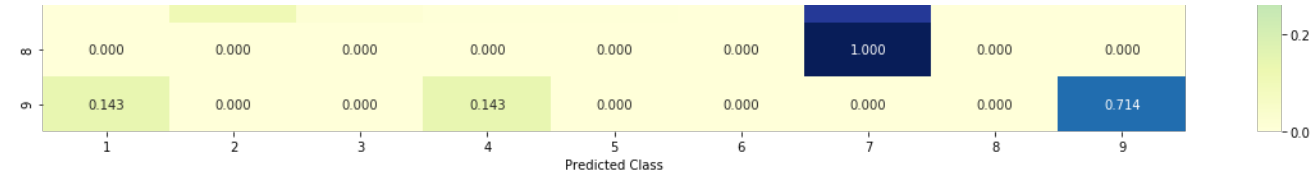
```
lr = LogisticRegression(C=0.1)
sclf = StackingClassifier(classifiers=[sig_clf1, sig_clf2, sig_clf3], meta_classifier=lr, use_probas=True)
sclf.fit(train_x_onehotCoding, train_y)

log_error = log_loss(train_y, sclf.predict_proba(train_x_onehotCoding))
print("Log loss (train) on the stacking classifier :",log_error)

log_error = log_loss(cv_y, sclf.predict_proba(cv_x_onehotCoding))
print("Log loss (CV) on the stacking classifier :",log_error)

log_error = log_loss(test_y, sclf.predict_proba(test_x_onehotCoding))
```

```
log_error = log_loss(test_y, sclf.predict_proba(test_x_onehotCoding))
print("Log loss (test) on the stacking classifier :",log_error)

print("Number of missclassified point :", np.count_nonzero((sclf.predict(test_x_onehotCoding)-
test_y))/test_y.shape[0])
plot_confusion_matrix(test_y=test_y, predict_y=sclf.predict(test_x_onehotCoding))
```

```
Log loss (train) on the stacking classifier : 0.6042061255058566
Log loss (CV) on the stacking classifier : 1.09119049360419
Log loss (test) on the stacking classifier : 1.0249739372234026
Number of missclassified point : 0.34135338345864663

-------------------- Confusion matrix --------------------
```


```
-------------------- Precision matrix (Columm Sum=1) --------------------
```


```
-------------------- Recall matrix (Row sum=1) --------------------
```

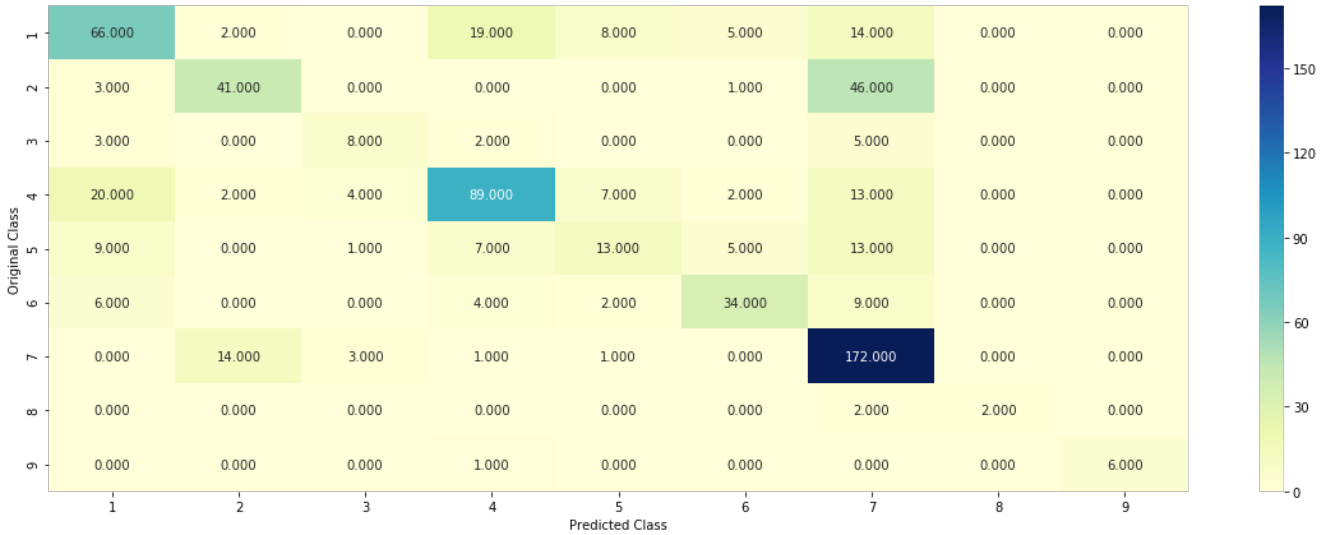


Maximum Voting Classifier

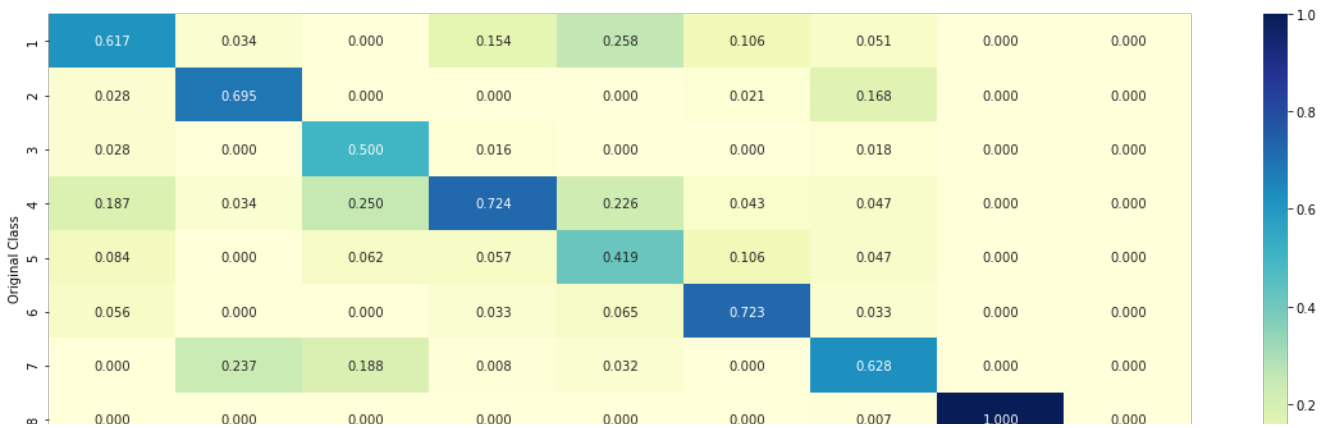
In [99]:

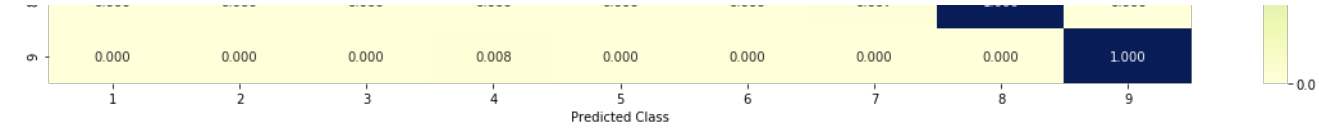
```
#Refer:http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html
from sklearn.ensemble import VotingClassifier
vclf = VotingClassifier(estimators=[('lr', sig_clf1), ('svc', sig_clf2), ('rf', sig_clf3)], voting=
'soft')
vclf.fit(train_x_onehotCoding, train_y)
print("Log loss (train) on the VotingClassifier :", log_loss(train_y,
vclf.predict_proba(train_x_onehotCoding)))
print("Log loss (CV) on the VotingClassifier :", log_loss(cv_y,
vclf.predict_proba(cv_x_onehotCoding)))
print("Log loss (test) on the VotingClassifier :", log_loss(test_y,
vclf.predict_proba(test_x_onehotCoding)))
print("Number of missclassified point :", np.count_nonzero((vclf.predict(test_x_onehotCoding)-
test_y))/test_y.shape[0])
plot_confusion_matrix(test_y=test_y, predict_y=vclf.predict(test_x_onehotCoding))
```

```
Log loss (train) on the VotingClassifier : 0.6893888872250667
Log loss (CV) on the VotingClassifier : 1.0677015127965879
Log loss (test) on the VotingClassifier : 0.9811707230674391
Number of missclassified point : 0.3518796992481203

-------------------- Confusion matrix --------------------
```

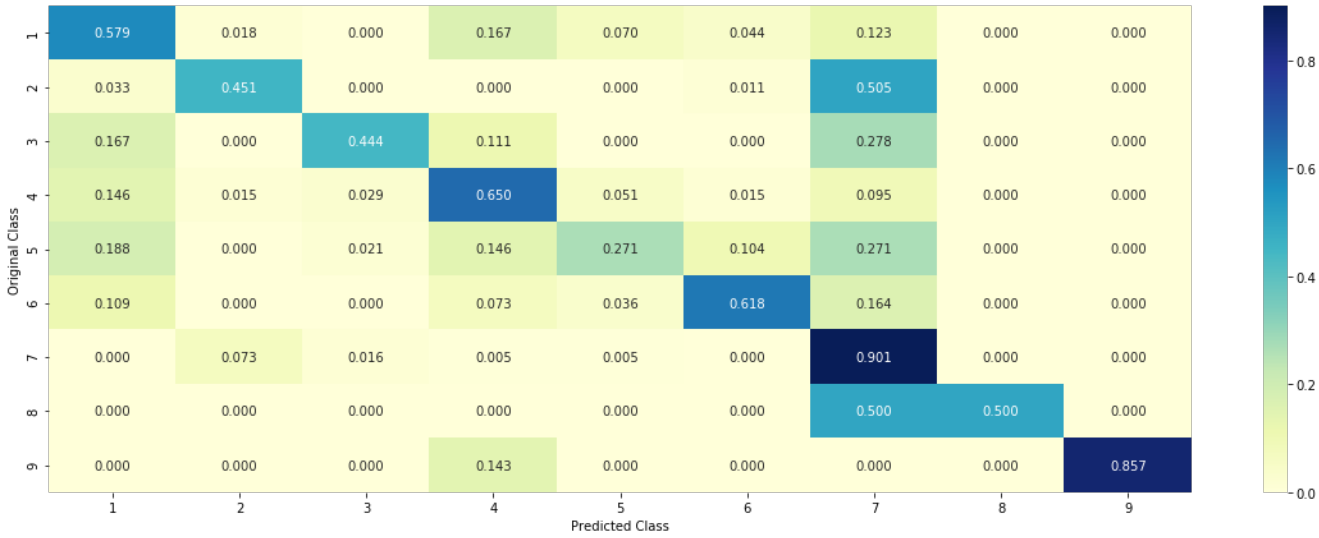



```
-------------------- Precision matrix (Column Sum=1) --------------------
```






```
-------------------- Recall matrix (Row sum=1) --------------------
```



LR with class Balancing On BOW

In [23]:

```
#response-coding of the Gene feature
# alpha is used for laplace smoothing
alpha = 1

# train gene feature
train_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_train))

# test gene feature
test_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_test))

# cross validation gene feature
cv_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_cv))
```

In [33]:

```
# one-hot encoding of Gene feature.
gene_vectorizer = CountVectorizer(ngram_range=(1, 4),min_df=5)
train_gene_feature_onehotCoding = gene_vectorizer.fit_transform(x_train['Gene'])
test_gene_feature_onehotCoding = gene_vectorizer.transform(x_test['Gene'])
cv_gene_feature_onehotCoding = gene_vectorizer.transform(x_cv['Gene'])

# don't forget to normalize every feature
train_gene_feature_onehotCoding = normalize(train_gene_feature_onehotCoding, axis=0)
test_gene_feature_onehotCoding = normalize(test_gene_feature_onehotCoding, axis=0)
cv_gene_feature_onehotCoding = normalize(cv_gene_feature_onehotCoding, axis=0)
```

In [26]:

```
# alpha is used for laplace smoothing
alpha = 1

# train gene feature
train_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", x_train))

# test gene feature
test_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", x_test))

# cross validation gene feature
cv_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", x_cv))
```

```
cv_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", x_cv))
```

In [34]:

```
# one-hot encoding of variation feature.
variation_vectorizer = CountVectorizer(ngram_range=(1, 4),min_df=5)
train_variation_feature_onehotCoding = variation_vectorizer.fit_transform(x_train['Variation'])
test_variation_feature_onehotCoding = variation_vectorizer.transform(x_test['Variation'])
cv_variation_feature_onehotCoding = variation_vectorizer.transform(x_cv['Variation'])


# don't forget to normalize every feature
train_variation_feature_onehotCoding = normalize(train_variation_feature_onehotCoding, axis=0)
test_variation_feature_onehotCoding = normalize(test_variation_feature_onehotCoding, axis=0)
cv_variation_feature_onehotCoding = normalize(cv_variation_feature_onehotCoding, axis=0)
```

In [ ]:

```

```

In [105]:

```
# building a CountVectorizer with all the words that occured minimum 3 times in train data
text_vectorizer = CountVectorizer(min_df=5,ngram_range=(1, 4))
train_text_feature_onehotCoding = text_vectorizer.fit_transform(x_train['TEXT'])

# getting all the feature names (words)
train_text_features= text_vectorizer.get_feature_names()

# train_text_feature_onehotCoding.sum(axis=0).A1 will sum every row and returns (1*number of featu
res) vector
train_text_fea_counts = train_text_feature_onehotCoding.sum(axis=0).A1

# zip(list(text_features),text_fea_counts) will zip a word with its number of times it occured
text_fea_dict = dict(zip(list(train_text_features),train_text_fea_counts))


print("Total number of unique words in train data :", len(train_text_features))
```

```
Total number of unique words in train data : 760847
```

In [106]:

```
#response coding of text features
train_text_feature_responseCoding  = get_text_responsecoding(x_train)
test_text_feature_responseCoding  = get_text_responsecoding(x_test)
cv_text_feature_responseCoding  = get_text_responsecoding(x_cv)

# https://stackoverflow.com/a/16202486
# we convert each row values such that they sum to 1
train_text_feature_responseCoding =
(train_text_feature_responseCoding.T/train_text_feature_responseCoding.sum(axis=1)).T
test_text_feature_responseCoding =
(test_text_feature_responseCoding.T/test_text_feature_responseCoding.sum(axis=1)).T
cv_text_feature_responseCoding = (cv_text_feature_responseCoding.T/cv_text_feature_responseCoding.
sum(axis=1)).T
```

In [107]:

```
# don't forget to normalize every feature
train_text_feature_onehotCoding = normalize(train_text_feature_onehotCoding, axis=0)

# we use the same vectorizer that was trained on train data
test_text_feature_onehotCoding = text_vectorizer.transform(x_test['TEXT'])
# don't forget to normalize every feature
test_text_feature_onehotCoding = normalize(test_text_feature_onehotCoding, axis=0)

# we use the same vectorizer that was trained on train data
cv_text_feature_onehotCoding = text_vectorizer.transform(x_cv['TEXT'])
# don't forget to normalize every feature
cv_text_feature_onehotCoding = normalize(cv_text_feature_onehotCoding, axis=0)
```

In [108]:

```
# merging gene, variance and text features

# building train, test and cross validation data sets
# a = [[1, 2],
#      [3, 4]]
# b = [[4, 5],
#      [6, 7]]
# hstack(a, b) = [[1, 2, 4, 5],
#                 [ 3, 4, 6, 7]]

train_gene_var_onehotCoding =
hstack((train_gene_feature_onehotCoding,train_variation_feature_onehotCoding))
test_gene_var_onehotCoding =
hstack((test_gene_feature_onehotCoding,test_variation_feature_onehotCoding))
cv_gene_var_onehotCoding = hstack((cv_gene_feature_onehotCoding,cv_variation_feature_onehotCoding)
)

train_x_onehotCoding = hstack((train_gene_var_onehotCoding, train_text_feature_onehotCoding)).tocs
r()
train_y = np.array(list(y_train['Class']))

test_x_onehotCoding = hstack((test_gene_var_onehotCoding, test_text_feature_onehotCoding)).tocsr()
test_y = np.array(list(y_test['Class']))

cv_x_onehotCoding = hstack((cv_gene_var_onehotCoding, cv_text_feature_onehotCoding)).tocsr()
cv_y = np.array(list(y_cv['Class']))


train_gene_var_responseCoding =
np.hstack((train_gene_feature_responseCoding,train_variation_feature_responseCoding))
test_gene_var_responseCoding =
np.hstack((test_gene_feature_responseCoding,test_variation_feature_responseCoding))
cv_gene_var_responseCoding =
np.hstack((cv_gene_feature_responseCoding,cv_variation_feature_responseCoding))

train_x_responseCoding = np.hstack((train_gene_var_responseCoding,
train_text_feature_responseCoding))
test_x_responseCoding = np.hstack((test_gene_var_responseCoding, test_text_feature_responseCoding)
)
cv_x_responseCoding = np.hstack((cv_gene_var_responseCoding, cv_text_feature_responseCoding))
```

In [109]:

```
print("One hot encoding features :")
print("(number of data points * number of features) in train data = ", train_x_onehotCoding.shape)
print("(number of data points * number of features) in test data = ", test_x_onehotCoding.shape)
print("(number of data points * number of features) in cross validation data =", cv_x_onehotCoding
.shape)
```

```
One hot encoding features :
(number of data points * number of features) in train data =  (2124, 763150)
(number of data points * number of features) in test data =  (665, 763150)
(number of data points * number of features) in cross validation data = (532, 763150)
```

In [110]:

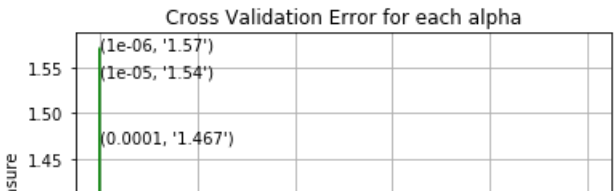
```
print(" Response encoding features :")
print("(number of data points * number of features) in train data = ", train_x_responseCoding.shap
e)
print("(number of data points * number of features) in test data = ", test_x_responseCoding.shape)
print("(number of data points * number of features) in cross validation data =",
cv_x_responseCoding.shape)
```

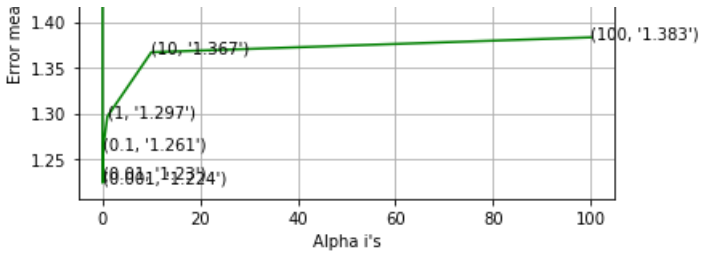
```
 Response encoding features :
(number of data points * number of features) in train data =  (2124, 27)
(number of data points * number of features) in test data =  (665, 27)
(number of data points * number of features) in cross validation data = (532, 27)
```

In [113]:

```
alpha = [10 ** x for x in range(-6, 3)]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = SGDClassifier(class_weight='balanced', alpha=i, penalty='l2', loss='log', random_state=42
)
    clf.fit(train_x_onehotCoding, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_onehotCoding, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
    cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    # to avoid rounding error while multiplying probabilites we use log-probability estimates
    print("Log Loss :",log_loss(cv_y, sig_clf_probs))

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],str(txt)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', ran
dom_state=42)
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The train log loss is:",
      log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15))

predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The cross validation log loss is:",
      log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))

predict_y = sig_clf.predict_proba(test_x_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha], "The test log loss is:",
      log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-06
Log Loss : 1.570454443220152
for alpha = 1e-05
Log Loss : 1.5400714645577374
for alpha = 0.0001
Log Loss : 1.4672563139746189
for alpha = 0.001
Log Loss : 1.2240840129069792
for alpha = 0.01
Log Loss : 1.23034888384657
for alpha = 0.1
Log Loss : 1.261355413871351
for alpha = 1
Log Loss : 1.296571179562698
for alpha = 10
Log Loss : 1.3667555951358168
for alpha = 100
Log Loss : 1.3830018860983684
```






```
For values of best alpha =  0.001 The train log loss is: 0.7893966095831639
For values of best alpha =  0.001 The cross validation log loss is: 1.2240840129069792
For values of best alpha =  0.001 The test log loss is: 1.1302352381408354
```

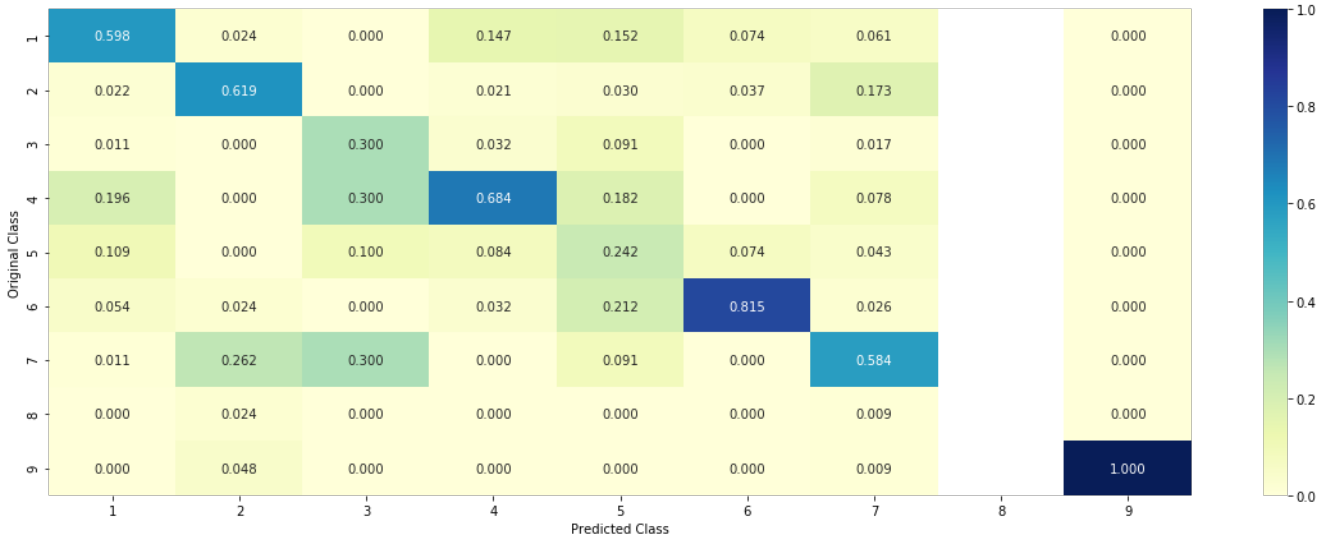
In [112]:

```
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', ran
dom_state=42)
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y, cv_x_onehotCoding, cv_y, clf)
```

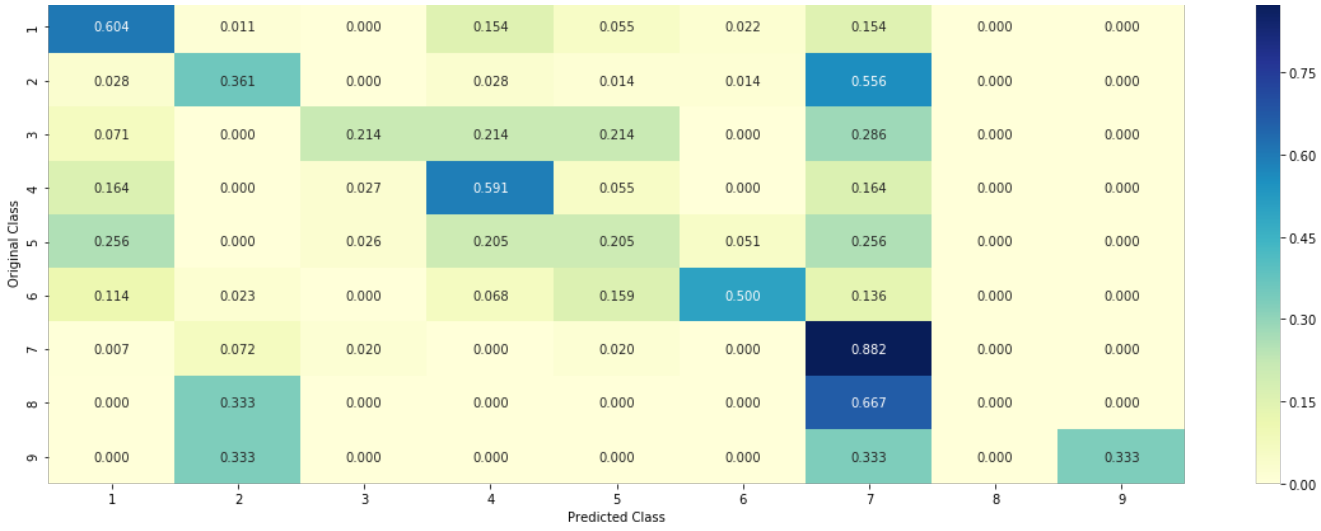
```
Log loss : 1.2240840129069792
Number of mis-classified points : 0.40601503759398494
```

```
--------------------- Confusion matrix ----------------------
```


```
-------------------- Precision matrix (Columm Sum=1) --------------------
```




```
-------------------- Recall matrix (Row sum=1) --------------------
```



ApplyingTfidf (1,4)ngrams using LR Model with Class Balancing

In [36]:

```
#response-coding of the Gene feature
# alpha is used for laplace smoothing
alpha = 1

# train gene feature
train_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_train))

# test gene feature
test_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_test))

# cross validation gene feature
cv_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_cv))
```

In [38]:

```
# one-hot encoding of Gene feature.
gene_vectorizer = TfidfVectorizer(ngram_range=(1, 4),min_df=5)
train_gene_feature_onehotCoding = gene_vectorizer.fit_transform(x_train['Gene'])
test_gene_feature_onehotCoding = gene_vectorizer.transform(x_test['Gene'])
cv_gene_feature_onehotCoding = gene_vectorizer.transform(x_cv['Gene'])

# don't forget to normalize every feature
train_gene_feature_onehotCoding = normalize(train_gene_feature_onehotCoding, axis=0)
test_gene_feature_onehotCoding = normalize(test_gene_feature_onehotCoding, axis=0)
cv_gene_feature_onehotCoding = normalize(cv_gene_feature_onehotCoding, axis=0)
```

In [39]:

```
# alpha is used for laplace smoothing
alpha = 1

# train gene feature
train_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", x_train))

# test gene feature
test_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", x_test))

# cross validation gene feature
cv_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", x_cv))
```

In [40]:

```
# one-hot encoding of variation feature.
variation_vectorizer = TfidfVectorizer(ngram_range=(1, 4),min_df=5)
train_variation_feature_onehotCoding = variation_vectorizer.fit_transform(x_train['Variation'])
```

```
test_variation_feature_onehotCoding = variation_vectorizer.transform(x_test['Variation'])
cv_variation_feature_onehotCoding = variation_vectorizer.transform(x_cv['Variation'])


# don't forget to normalize every feature
train_variation_feature_onehotCoding = normalize(train_variation_feature_onehotCoding, axis=0)
test_variation_feature_onehotCoding = normalize(test_variation_feature_onehotCoding, axis=0)
cv_variation_feature_onehotCoding = normalize(cv_variation_feature_onehotCoding, axis=0)
```

In [41]:

```
# building a CountVectorizer with all the words that occured minimum 3 times in train data
text_vectorizer = TfidfVectorizer(min_df=5,ngram_range=(1, 4),max_features=3000)
train_text_feature_onehotCoding = text_vectorizer.fit_transform(x_train['TEXT'])

# getting all the feature names (words)
train_text_features= text_vectorizer.get_feature_names()

# train_text_feature_onehotCoding.sum(axis=0).A1 will sum every row and returns (1*number of featu
res) vector
train_text_fea_counts = train_text_feature_onehotCoding.sum(axis=0).A1

# zip(list(text_features),text_fea_counts) will zip a word with its number of times it occured
text_fea_dict = dict(zip(list(train_text_features),train_text_fea_counts))


print("Total number of unique words in train data :", len(train_text_features))
```

```
Total number of unique words in train data : 3000
```

In [50]:

```
#response coding of text features
train_text_feature_responseCoding  = get_text_responsecoding(x_train)
test_text_feature_responseCoding  = get_text_responsecoding(x_test)
cv_text_feature_responseCoding  = get_text_responsecoding(x_cv)

# https://stackoverflow.com/a/16202486
# we convert each row values such that they sum to 1
train_text_feature_responseCoding =
(train_text_feature_responseCoding.T/train_text_feature_responseCoding.sum(axis=1)).T
test_text_feature_responseCoding =
(test_text_feature_responseCoding.T/test_text_feature_responseCoding.sum(axis=1)).T
cv_text_feature_responseCoding = (cv_text_feature_responseCoding.T/cv_text_feature_responseCoding.
sum(axis=1)).T
```

In [51]:

```
# don't forget to normalize every feature
train_text_feature_onehotCoding = normalize(train_text_feature_onehotCoding, axis=0)

# we use the same vectorizer that was trained on train data
test_text_feature_onehotCoding = text_vectorizer.transform(x_test['TEXT'])
# don't forget to normalize every feature
test_text_feature_onehotCoding = normalize(test_text_feature_onehotCoding, axis=0)

# we use the same vectorizer that was trained on train data
cv_text_feature_onehotCoding = text_vectorizer.transform(x_cv['TEXT'])
# don't forget to normalize every feature
cv_text_feature_onehotCoding = normalize(cv_text_feature_onehotCoding, axis=0)
```

In [52]:

```
# merging gene, variance and text features

# building train, test and cross validation data sets
# a = [[1, 2],
#      [3, 4]]
# b = [[4, 5],
#      [6, 7]]
# hstack(a, b) = [[1, 2, 4, 5],
#                 [ 3, 4, 6, 7]]
```

```
train_gene_var_onehotCoding =
hstack((train_gene_feature_onehotCoding,train_variation_feature_onehotCoding))
test_gene_var_onehotCoding =
hstack((test_gene_feature_onehotCoding,test_variation_feature_onehotCoding))
cv_gene_var_onehotCoding = hstack((cv_gene_feature_onehotCoding,cv_variation_feature_onehotCoding)
)

train_x_onehotCoding = hstack((train_gene_var_onehotCoding, train_text_feature_onehotCoding)).tocs
r()
train_y = np.array(list(y_train['Class']))

test_x_onehotCoding = hstack((test_gene_var_onehotCoding, test_text_feature_onehotCoding)).tocsr()
test_y = np.array(list(y_test['Class']))

cv_x_onehotCoding = hstack((cv_gene_var_onehotCoding, cv_text_feature_onehotCoding)).tocsr()
cv_y = np.array(list(y_cv['Class']))


train_gene_var_responseCoding =
np.hstack((train_gene_feature_responseCoding,train_variation_feature_responseCoding))
test_gene_var_responseCoding =
np.hstack((test_gene_feature_responseCoding,test_variation_feature_responseCoding))
cv_gene_var_responseCoding =
np.hstack((cv_gene_feature_responseCoding,cv_variation_feature_responseCoding))

train_x_responseCoding = np.hstack((train_gene_var_responseCoding,
train_text_feature_responseCoding))
test_x_responseCoding = np.hstack((test_gene_var_responseCoding, test_text_feature_responseCoding)
)
cv_x_responseCoding = np.hstack((cv_gene_var_responseCoding, cv_text_feature_responseCoding))
```

In [53]:

```
print("One hot encoding features :")
print("(number of data points * number of features) in train data = ", train_x_onehotCoding.shape)
print("(number of data points * number of features) in test data = ", test_x_onehotCoding.shape)
print("(number of data points * number of features) in cross validation data =", cv_x_onehotCoding
.shape)
```

```
One hot encoding features :
(number of data points * number of features) in train data =  (2124, 3110)
(number of data points * number of features) in test data =  (665, 3110)
(number of data points * number of features) in cross validation data = (532, 3110)
```

In [54]:

```
print(" Response encoding features :")
print("(number of data points * number of features) in train data = ", train_x_responseCoding.shap
e)
print("(number of data points * number of features) in test data = ", test_x_responseCoding.shape)
print("(number of data points * number of features) in cross validation data =",
cv_x_responseCoding.shape)
```

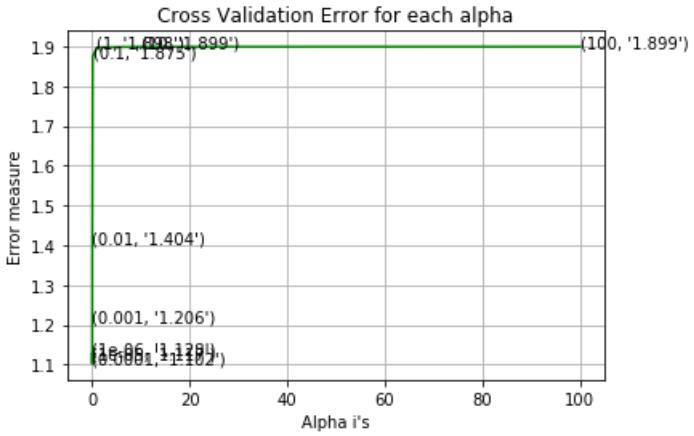
```
 Response encoding features :
(number of data points * number of features) in train data =  (2124, 27)
(number of data points * number of features) in test data =  (665, 27)
(number of data points * number of features) in cross validation data = (532, 27)
```

In [55]:

```
alpha = [10 ** x for x in range(-6, 3)]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = SGDClassifier(class_weight='balanced', alpha=i, penalty='l2', loss='log', random_state=42
)
    clf.fit(train_x_onehotCoding, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_onehotCoding, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
    cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    # to avoid rounding error while multiplying probabilites we use log-probability estimates
```

```
    print("Log Loss :",log_loss(cv_y, sig_clf_probs))

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],str(txt)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', ran
dom_state=42)
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The train log loss is:",
      log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15))

predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha],
      "The cross validation log loss is:",
      log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))

predict_y = sig_clf.predict_proba(test_x_onehotCoding)
print('For values of best alpha = ',
      alpha[best_alpha], "The test log loss is:",
      log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-06
Log Loss : 1.1280550427061118
for alpha = 1e-05
Log Loss : 1.1168260289264693
for alpha = 0.0001
Log Loss : 1.1021510973378432
for alpha = 0.001
Log Loss : 1.2062195613530553
for alpha = 0.01
Log Loss : 1.4036264719331117
for alpha = 0.1
Log Loss : 1.8750828231591583
for alpha = 1
Log Loss : 1.8975111023902147
for alpha = 10
Log Loss : 1.898903007968171
for alpha = 100
Log Loss : 1.8990305702120678
```




```
For values of best alpha =  0.0001 The train log loss is: 0.6483471736807301
For values of best alpha =  0.0001 The cross validation log loss is: 1.1021510973378432
```

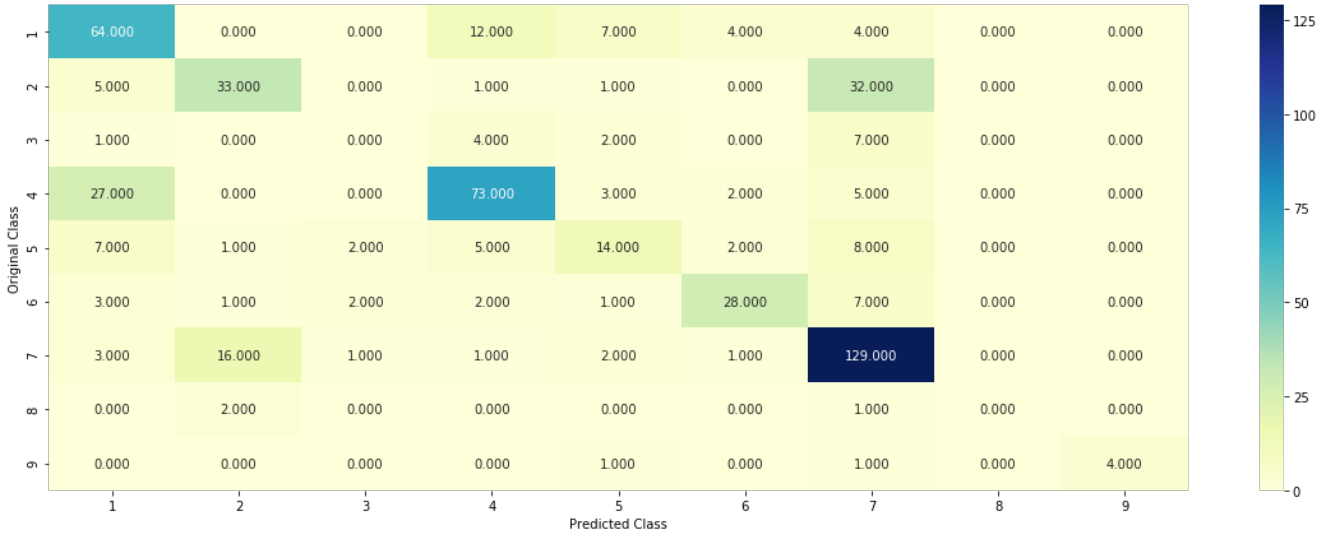
```
For values of best alpha =  0.0001 The test log loss is: 1.0788439583554366
```

In [60]:

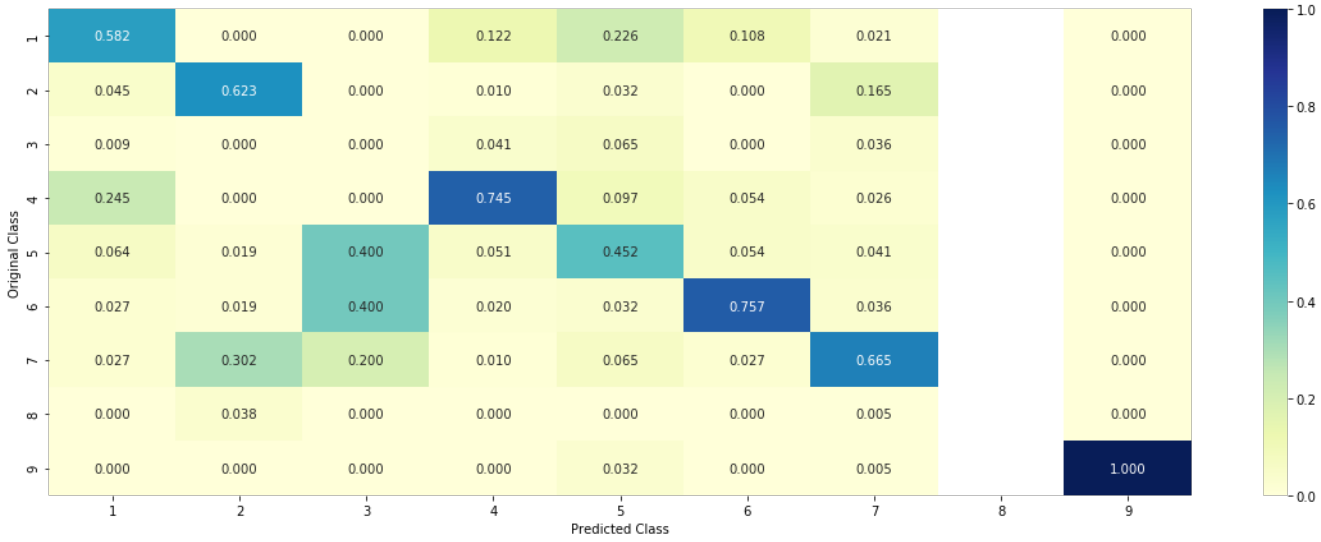
```
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', ran
dom_state=42)
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y, cv_x_onehotCoding, cv_y, clf)
```

```
Log loss : 1.1021510973378432
Number of mis-classified points : 0.35150375939849626
```

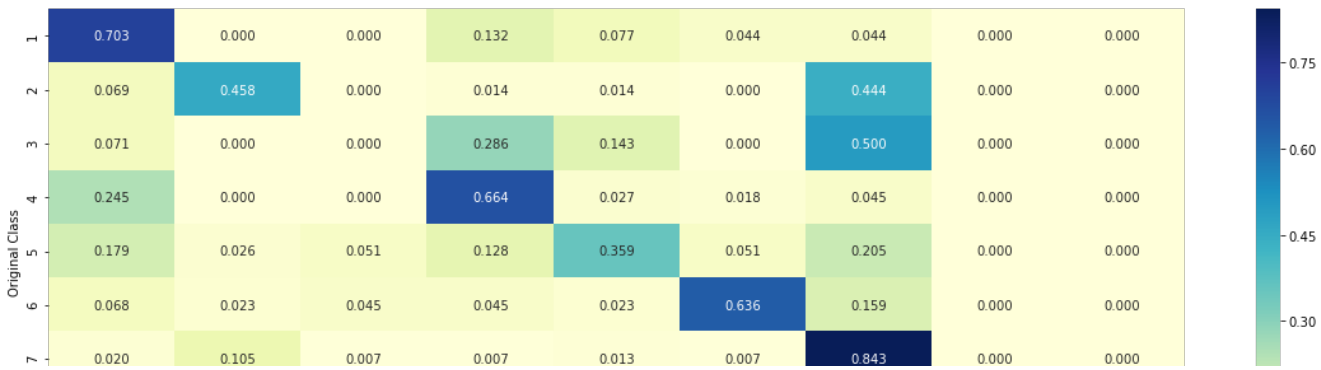
---------- Confusion matrix ----------

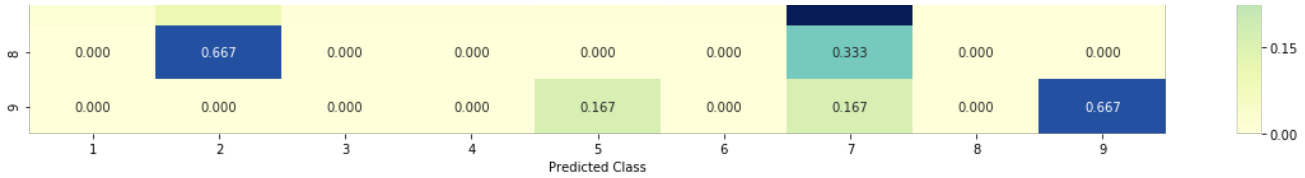


---------- Precision matrix (Columm Sum=1) ----------



---------- Recall matrix (Row sum=1) ----------





Pretty Table

In [67]:

```
from prettytable import PrettyTable
ptable = PrettyTable()
ptable.title = "*** Model Summary *** [Performance Metric: Log-Loss]"
ptable.field_names=["Model Name","Train","CV","Test","% Misclassified Points"]
ptable.add_row(["Naive Bayes","0.939","1.221","1.189","40"])
ptable.add_row(["KNN","0.897","1.051","1.039","36"])
ptable.add_row(["Logistic Regression With Class balancing","0.535","1.11","0.977","34"])
ptable.add_row(["Logistic Regression Without Class balancing","0.534","1.11","0.985","34"])
ptable.add_row(["Linear SVM","0.582","1.141","1.015","34"])
ptable.add_row(["Random Forest Classifier With One hot Encoding","0.665","1.164","1.128","39"])
ptable.add_row(["Random Forest Classifier With Response Coding","0.059","1.2933","1.251","45"])
ptable.add_row(["Stack Models:LR+NB+SVM","0.604","1.091","1.024","34"])
ptable.add_row(["Maximum Voting classifier","0.689","1.067","0.9811","35"])
ptable.add_row(["Logistic Regression Bow ","0.789","1.22","1.13","40"])
ptable.add_row(["Logistic Regression Tfidf with 3000 features","0.648","1.102","1.07","35"])
print(ptable)
print()
```

```
+-----------------------------------------------------------------------------------------------------+
|                         *** Model Summary *** [Performance Metric: Log-Loss]
|
+-----------------------------------------------+-------+--------+--------+-----------------------
|                   Model Name                  | Train |   CV   |  Test  | % Misclassified Points
|
+-----------------------------------------------+-------+--------+--------+-----------------------
|                  Naive Bayes                  | 0.939 | 1.221  | 1.189  |           40
|                      KNN                      | 0.897 | 1.051  | 1.039  |           36
|    Logistic Regression With Class balancing   | 0.535 |  1.11  | 0.977  |           34
|
|  Logistic Regression Without Class balancing  | 0.534 |  1.11  | 0.985  |           34
|
|                  Linear SVM                   | 0.582 | 1.141  | 1.015  |           34
| Random Forest Classifier With One hot Encoding | 0.665 | 1.164  | 1.128  |           39
|
| Random Forest Classifier With Response Coding  | 0.059 | 1.2933 | 1.251  |           45
|
|             Stack Models:LR+NB+SVM            | 0.604 | 1.091  | 1.024  |           34
|
|           Maximum Voting classifier           | 0.689 | 1.067  | 0.9811 |           35
|
|            Logistic Regression Bow            | 0.789 |  1.22  |  1.13  |           40
|
|  Logistic Regression Tfidf with 3000 features | 0.648 | 1.102  |  1.07  |           35
|
+-----------------------------------------------+-------+--------+--------+-----------------------
```

I have managed to get test logloss less than one in 2 cases when I applied Logistic regression Model for 1000 Tfidf features ,and another Time I got test logloss less than 1 in Maximum Voting Classifier. As instructed I have applied Logistic Regression on tfidf features with 3000 feature with min_df=5 and ngrams=(1,4),results were decent ,test logloss is equal to 1 ,missclassified points are 35 percent which is better than many models here. By looking on above table I can conclude that Logistic regression with class balancing on Tfidf with 1000 featues performed best among all.