2017 IEEE International Conference on Cloud Computing in Emerging Markets

POSTER PAPER

Experimental Exploration of Support Vector Machine for Cancer Cell Classification

Kevin Joy Dsouza
Dept. of CSE
Yenepoya Institute of Technology, Moodbidri
kevin@yenepoya.edu.in

Zahid Ahmed Ansari
Dept. of CSE
P.A. College of Engineering, Mangalore
zahid_cs@pace.edu.in

***Abstract*— text classification is the task of automatically categorizing collections of electronic textual documents into their predefined classes, based on their contents. Due to the increase in the amount of text data in these recent years, document classification has emerged in the form of text classification systems. They have been widely implemented in a large number of applications such as spam filtering, e-mails, knowledge repositories and ontology mapping. The main essence is to propose a text classification technique based on the feature selection and reduction of the feature vector dimensionality and increase the classification accuracy using pre-processing. This paper gives the detailed study on how support vector machine (SVM) can be used to classify uncertain data. SVM is a powerful and supervised learning sample based on the lowest structural risk principle. During training, this algorithm creates a hyperplane for separating positive and negative samples. The type of kernel used for SVM classifier will be having a major impact on classification results. In this paper Breast Cancer Wisconsin (Diagnostic) Data Sets are used in order to classify using four types of SVM kernel methods such as linear, polynomial, sigmoid and radial. Classification results obtained reveal that radial kernel method is best-suited data sets. In order to measure the suitability of kernel method, various factors are compared from classification results such as accuracy, kappa value, sensitivity, specificity precision etc.**

Keywords*— *Text Classification, Soft-Computing, R-Tool, SVM;

I. Introduction

Data to discover previously unknown, relationships and valid patterns in data set. These tools can include statistical models, mathematical algorithm and machine learning methods. Data mining consists of more than collection and managing data; it also includes analysis and prediction. Classification technique is capable of processing a wide variety of data than regression and is growing in popularity mining [1] involves the use of sophisticated data analysis tools.

Text classification [2][3][14] is the automated assigning of natural language texts to predefined categories based on their content. As the amount of online text is increasing, the need for text classification to aid the analysis and management of text is in high demand. The text is cheap but managing information in the form of knowledge, what classes a text related to is difficult to predict. Automatic classification of text can be used to provide this required knowledge but the classifiers must be built with expensive human effort or trained from texts which have otherwise been manually classified. Text Classification is the primary requirement of Text retrieval systems, which retrieve texts based on a query provided by the user and the Text Understanding systems which transform text in some way such as producing summaries, answering questions or extracting data.

Soft computing [3][4] is an emerging approach to computing which parallels remarkable capability of human mind to reason and learn in an environment of incomplete, uncertainty and imprecision. Soft Computing consists of several computing techniques like Neural Networks, Fuzzy Logic, and Genetic algorithms. Soft Computing uses enhancement of these techniques. A hybrid technique would inherit all the advantages of constituent techniques. Thus the components of Soft Computing are complementary [3], not competitive, offering their own advantages and techniques to provide solutions, otherwise unsolvable problems in text classification.

SVM is the recent classification technique in data mining. The objective of data mining is to automatic discovery of useful information from large databases and pre-processes them for further use. Classification is a data mining technique used to make a prediction of group membership for data instances. It is the process of finding a model that can describe and distinguish data classes or concepts. The model is then used to predict the class of objects whose class label is unknown. The model is built based on the analysis of training data. So, the classification involves two steps- training the classifier model based on training data and testing the accuracy of the model for the testing data whose class label is unknown

The paper is structured as follows: Section 1 gives a brief description about, how data mining can be used explore useful information from textual documents. Section 2 gives a brief literature survey, of how text mining techniques can be used in knowledge discovery. Especially in this paper soft computing based SVM method for text classification is used. Section 3 gives a primer about preprocessing, feature extraction and classification of text using SVM method, Section 4 gives experimental results obtained by the classifier. Section 5 provides conclusion from this study.

II. LITERATURE SURVEY

Soft computing refers to algorithms that are able to cope with uncertainty and incomplete information and that are still capable of discovering approximate, good solutions to

978-1-5386-2450-0/17 $31.00 © 2017 IEEE
DOI 10.1109/CCEM.2017.15

complex computational problems, and doing so faster from a computational standpoint. These algorithms include neural nets, fuzzy logic, rough set, Bayesian algorithms, and evolutionary computing and support vector machines. While extensive coverage of these algorithms can indeed be found in the following section, a brief description of different techniques which can be used for text classification is given in the following section.

Support vector machines[7][6] (SVM) are a collection of supervised machine learning algorithms that are based on statistical learning theory (kernel-based techniques) and the Vapnik-Chervonenkis(VC) dimension. Primary applications of SVM include classification and regression. Most of the aforementioned soft computing techniques have already found their way into computational toolboxes offered by vendors such as 'The Mathworks' or as open source implementations by researchers. Support Vector Machine is learning method that makes use of a hypothesis space of separating functions in a high dimensional feature space. One of the successful uses of SVM algorithm is the task of text categorization into a fixed number of predefined categorization based on their content.

The classification problem addressed in [6] restricted by considering the two-class problem without loss of generality. The multiclass problem can be addressed using multiple binary classifications. The ultimate goal is to produce a classifier that will work well on unseen examples, i.e., it generalizes well.

SVM based classification can be implemented with the help of tools such as LIBSVM, SVM Light. A tool SVM Light [7][8] is the implementation of SVM in C language.

Aurangabadkar et al[5] discussed SVM based classification of sports articles in the form of text files using SVM Light tool. To perform classification using SVM Light, all necessary pre-processing of sports articles is implemented as part of the system. The performance of the classifier is also generated as a part of the system using post-processing of the result of classification. The system classifies the sports articles as Cricket relevant and others and shows the same result. So, the proposed system performs SVM based classification of sports articles in the form of text documents as well as generates the predictions for the classification.

Bhaskar et al [6] showed distinct versions of SVM use different kernel functions to handle different types of data sets. Linear and non-linear kernels are supported. This paper is more helpful for researchers in the field of Data Mining to understand how SVM classification attempts to separate the target classes with the widest possible margin. This work recommends that researchers use SVM algorithm for preparation of data to optimize the real world problems such as text and image classification, hand-writing recognition, and bioinformatics and biosequence analysis.Also, this paper explains the previous work, which has been reviewed for the understanding of research in the area of Data Mining technologies.

Ya GAO Et Al [11] compared the performance of linear And nonlinear kernels of Support Vector Machines (SVM) used for text classification. The study is motivated by the viewpoint that linear SVM performs better than a nonlinear one, and that, although there are many investigations have proved that SVM performs well in text classification, there is no serious investigation on the comparison between linear SVM and nonlinear SVM. In this study, two experiments are carried out with different datasets and use grid-search on the selection of kernel parameters. Empirical results show that, in fact, nonlinear SVM performs better than linear SVM as long as with appropriate kernel parameters. The conclusion from this study provides useful guidance for people applying SVM to text classification and other corresponding fields.

III. SVM Based Text Classification Method

Support Vector Machines (SVMs)[9] is a supervised method of learning applied for classification and regression tasks usually used in data mining approach. SVM is a universal classification model that produces non-overlapping partitions and usually makes use of all attributes. The object space is partitioned in a single pass so that flat and linear partitions are produced. SVMs are founded on maximum margin linear discriminants, and are similar to probabilistic approaches, but do not consider the dependencies among attribute.

A classification task involves training and test sets which consist of data instances. Each instance in the training set contains one target value called as a class label and several attributes called as features. The aim of a classifier is to prepare a model which will be able to predict target class labels of data instances in the testing set, for which only the attributes are given. The classification problem can be viewed as a two-class problem in which one's objective is to separate the two classes by a function introduced from available examples. The objective is to prepare classifier that generalizes well, i.e. that works well on unidentified classes.

Algorithm:

Let D be a classification dataset with n different points in a d-dimensional space, $D = \{(x_i, y_i)\}$, with $i = 1, 2, ..., n$ and let there be only two class labels such that y_i is either +1 or -1. A hyperplane $h(x)$ gives a linear discriminant function in d dimensions and splits the original space into two half-spaces.

$$h(x) = \omega^T x + b = \omega_1 x_1 + \omega_2 x_2 + ... + \omega_d x_d + b \qquad (1)$$

Where ω are a d-dimensional weight vector and b is a scalar bias, Points on the hyperplane have $h(x) = 0$, i.e. the hyperplane is defined by all points for which $\omega^T x = -b$.

if the dataset is separable linearly, a separating hyperplane can be derived such that for all points with the label -1, $h(x) < 0$ and for all points labeled +1, $h(x) > 0$. In this context, h(x) results to be a linear classifier or linear discriminant that predicts the class for any point. The weight vector w is orthogonal to the hyperplane, thus gives the direction that is normal to it, but the bias b fixes the offset of the hyperplane in the d-dimensional space.

Given a hyperplane h(x) = 0, it is possible to measure the distance between each point x_i and the hyperplane by the equation (1)

$$\delta_i = \frac{y_i h(x_i)}{\| \omega \|} \qquad (2)$$

The margin of the linear classifier is the minimum distance of all n points to the separating hyperplane.

$$\delta^* = \frac{\min}{x_i}\{\frac{y_i h(x_i)}{\| \omega \|}\} \qquad (3)$$

All points (vectors x_i^*) that obtain this minimum distance are termed the support vectors for the linear classifier. There for support vector is a point that lies precisely on the margin of the classifying hyperplane.

In a canonical representation of the hyperplane, for each support vector x_i with label, y_i we have that. A similar way to any point that is not a support vector, we have that $y_i h(x_i) > 1$ .since by definition it must be farther from the hyperplane than a support vector. Therefore we have that $y_i h(x_i) \geq 1, \forall x_i \in D$.

IV. PERFORMANCE ANALYSIS

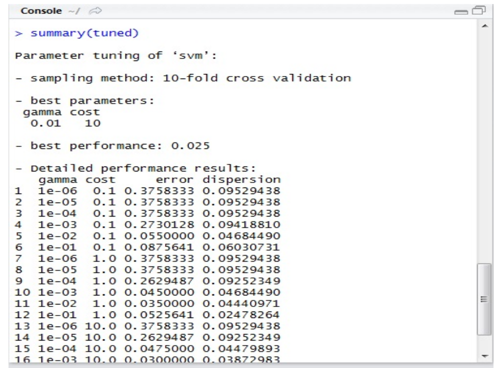
In this work, R programming is used in order to build the classifier. Data sets are collected from UCI machine learning repository. Diagnostic Wisconsin Breast Cancer Datasets are used for the experiment. In this dataset, there are 569 instances and 32 attributes for each instance. The first attribute is the identification of instance, the second is the label for the instance class, which can be M (malignant tumor) or B (benign tumor). The remaining 30 attributes are real-valued input features that are obtained from a digitized image of a fine needle aspirate (FNA) of a breast mass. Finally, there are 357 benign instances and 212 malignant instances in the dataset.

A. Preparing the DataSet

To prepare datasets for classification, datasets are divided into two subsets, one with about 70% of the instances for training, and another with around the remaining 30% of instances to testing.

B. Choosing Parameters

In this experiment tune() function is used to do a grid search over the supplied parameter ranges (C - cost, γ gamma), using the train set. The range of gamma parameter is between 0.000001 and 0.1. For cost parameter, the range is from 0.1 until 10. The snapshot of R console is shown Fig1.



```
Console ~/
> summary(tuned)

Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 gamma cost
  0.01   10

- best performance: 0.025

- Detailed performance results:
   gamma cost     error dispersion
1  1e-06  0.1 0.3758333 0.09529438
2  1e-05  0.1 0.3758333 0.09529438
3  1e-04  0.1 0.3758333 0.09529438
4  1e-03  0.1 0.2730128 0.09418810
5  1e-02  0.1 0.0550000 0.04684490
6  1e-01  0.1 0.0875641 0.06030731
7  1e-06  1.0 0.3758333 0.09529438
8  1e-05  1.0 0.3758333 0.09529438
9  1e-04  1.0 0.2629487 0.09252349
10 1e-03  1.0 0.0450000 0.04684490
11 1e-02  1.0 0.0350000 0.04440971
12 1e-01  1.0 0.0525641 0.02478264
13 1e-06 10.0 0.3758333 0.09529438
14 1e-05 10.0 0.2629487 0.09252349
15 1e-04 10.0 0.0475000 0.04479893
16 1e-03 10.0 0.0300000 0.03872983
```

Fig.1 Parameter tuning of SVM

C. Training and Testing the model

To prepare the SVM model to predict breast cancer cell category using the parameters C=10 and gamma=0.001, which are proved to be the best values according to the tune () function showed in Fig1.

The training phase is the one of a crucial phase of making the model to build with enough information about the features. This always influences the accuracy in prediction. In text classification, a text learner takes the advantage of examples (data) to record characteristics of interest of their unknown underlying probability distribution. Our goal in text classification is high accuracy on testing data or new data. Data can be seen as examples that illustrate relations between observed variables. A major focus of text classification is to automatically learn to devise a model to recognize complex patterns and make intelligent decisions based on data information. The difficulty lies in the fact that the set of all possible behaviors given possible inputs is too huge to be covered by the set of observed examples (training data). Hence the learner must provide a generalized model from the given examples, so as to be able to produce a useful output in new data (testing data).

Once the model is built and trained accordingly to the feature vectors, next step is to test model predict the untrained features. Prediction accuracy depends on how well the model is trained. The Table1 shows the confusion matrix for output prediction.

Table 1
Confusion Matrix

	True B	M
Pred **B**	102	3
M	1	64

This means that there are 102 benign instances in the test set and one of them were predicted as benign instances. On the other hand, there are 64 malign instances in the test set, 61 were predicted rightly and 3 as benign instances.

In order to evaluate the performance of the classifier, the following is conventions and criteria are used.

Let:

TP: true positive, i.e. malign instances predicted rightly.

FP: false positive, i.e. benign instances predicted as malign.

TN: true negative, i.e. benign instances predicted rightly.

|N|: a total of benign instances.

|P|: a total of malign instances.

a) Sensitivity = $\frac{TP}{|p|}$

b) Specificity = $\frac{TN}{|N|}$

c) Precision = $\frac{TP}{TP+FP}$

Following results are obtained by applying values obtained from the confusion matrix.

a) Sensitivity = $\frac{64}{64+3}$ = 0.95.

b) Specificity = $\frac{102}{103}$ = 99.

c) Precision = $\frac{64}{64+1}$ = 98.

From following results it is observed Classification results are suitable for the built model.

V. COMPARISION of KERNEL METHODS

In order to find the better-suited SVM model for classification, different kernel methods are used such as linear, polynomial, sigmoid, radial.

A. Linear kernel method

The Linear kernel is the simplest kernel function. It is given by the Inner Product <x, y> plus optional constant 'c'. Kernel algorithms using a linear kernel are often equivalent to their non-kernel counterparts [12]

$$K(x,y)= x^{T}+c \tag{4}$$

The results obtained for this parameter setting is listed in Table 1. Where kappa is The Kappa statistic (or value) is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance).

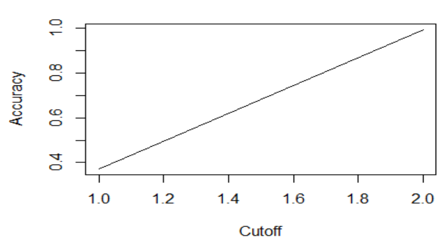


Fig.2 Accuracy graph for linear kernel

The Fig.2 shows accuracy graph for linear kernel classifier performance. It is observed from the graph the accuracy obtained is close to 96%. For the same result parameters such as sensitivity which is also called as true positive(TP) and sensitivity which is also called a false negative(FP) is shown in the Fig.3.

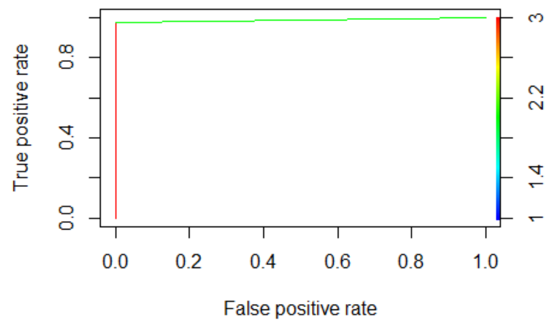


Fig.3 TPFP graph for linear kernel

B. Polynomial kernel function.

The Polynomial kernel is a non-stationary kernel. Polynomial kernels are better suited for data where all the training data is normalized. For degree polynomials, the polynomial kernel is defined as

$$K(x,y) = \alpha\, x^{T}y+c \tag{5}$$

Where x and y are vectors computed from training or test samples, C$\geq$ 0 is a constant trading of the influence of higher-order versus lower-order terms in the polynomial and when C $\geq$ 0, the kernel is called Homogeneous.

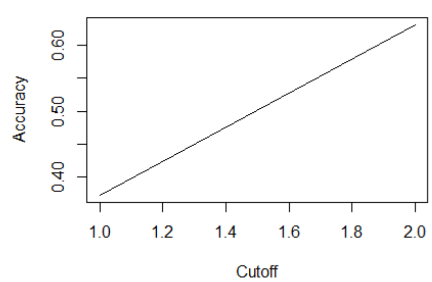


Fig.4 Accuracy graph for polynomial kernel.

Fig.4 shows the accuracy of polynomial kernel classifier. Results of sensitivity and specificity and corresponding graph plot can be seen in Fig.5.

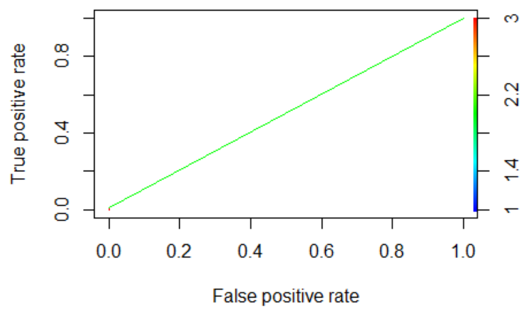


Fig.5 TPFP graph for polynomial kernel

C. *Sigmoid Kernel function.*

SVM model using a sigmoid kernel function is similar to a two-layer, perceptron neural network. This kernel was quite popular for support vector machines due to its origin from neural network theory. It has been found to perform well in practice [12].

$$K(x,y) = \tanh(\alpha\, x^T y + c) \qquad (5)$$

The classification accuracy for the sigmoid kernel is shown Fig.6 and also the sensitivity and specificity plot is shown in the Fig.7.

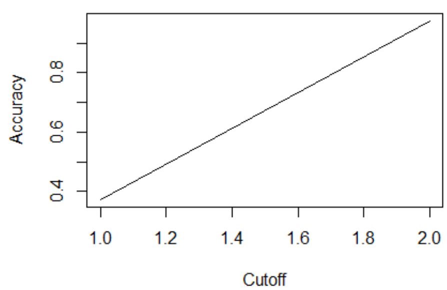


Fig.6 Accuracy graph for sigmoid kernel

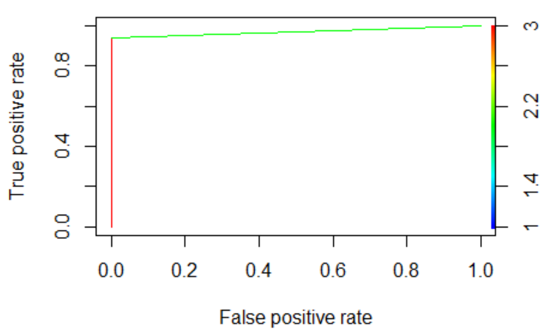


Fig.7 TPFP graph for polynomial kernel

D. *Radial Kernel function*

The RBF is by far the most popular choice of kernel types used in Support Vector Machines. This is mainly because of their localized and finite responses across the entire range of the real x-axis. There are different types of Radial Basis Kernel Function such as Linear Radial Basis Function, Gaussian Radial Basis Function, and Multiquadrics Radial Basis Function.

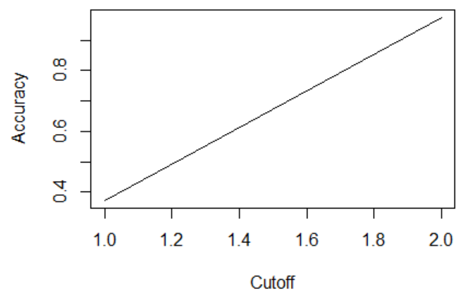


Fig.8 Accuracy graph for sigmoid kernel

The accuracy of the radial kernel is about 96% as shown in the Fig.8. The corresponding TPFP plot is shown in the Fig.9.

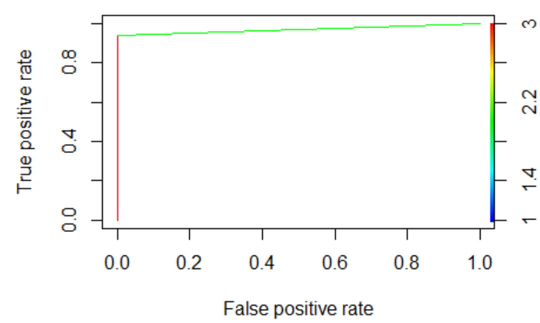


Fig.9 TPFP graph for radial kernel

It is evident from the experimental results that suitability of particular type of kernel depends on input data. The number of observations and correlation between features influences in great extent to classify the category of data. In the meantime, the metrics that can be used evaluate the performance of classifier are sensitivity and specificity. The overall experimental outcome for various type of kernel is tabulated in Table 2. This parameter is used to measure the performance and suitability of classifier.

Table 2
Experimental Results

parameters	Type of kernel			
	linear	**polynomial**	**radial**	**sigmoid**
accuracy	0.9647	0.6235	0.9647	0.9588
kappa	0.9282	0.0189	0.8257	0.9588
sensitivity	0.9789	1.0000	0.9257	0.9909
specificity	0.9467	0.01538	0.9265	0.9000
Positive prediction	0.9588	0.62130	0.9528	0.9478
Negative prediction	0.9726	1.0000	0.9844	0.9818
prevalence	0.5588	0.61765	0.6000	0.6471
Detection rate	0.5471	0.61765	0.5941	0.6412
Detection prevalence	0.5706	0.99412	0.6235	0.6765
Balance accuracy	0.9628	0.50769	0.9583	0.9455
Positive class	B	B	B	B

VI. CONCLUSION

This study provides various aspects of text classification and use of soft computing based SVM classification technique for text classification. A detailed framework of how SVM can be used for classification task is explained. Further different parameters to be set for SVM classifier and influence of this on classifiers performance are recorded. A major challenge in text classification is building the model for the classifier, which is discussed in detail with SVM method. At its simplest, classifier model provides a quick and easy way to explore and analyze data for classification. The results obtained by the model produce justifiable values for sensitivity, specificity, and precision, hence the results are suitable for the text classifier. This study is helpful for the researchers to understand the suitability of SVM kernel selection methods for the classification task.

REFERENCES

[1] Kecman, Vojislav. *Learning and soft computing: support vector machines, neural networks, and fuzzy logic models*. MIT press, 2001.

[2] Thanh, Vo Duy, et al. "Text classification based on semi-supervised learning."Soft Computing and Pattern Recognition (SoCPaR), 2013 International Conference of. IEEE, 2013.

[3] Saad, Ashraf. "An overview of hybrid soft computing techniques for classifier design and feature selection." *Hybrid Intelligent Systems, 2008. HIS'08. Eighth International Conference on*. IEEE, 2008.

[4] Wermter, Stefan. "Neural network agents for learning semantic text classification." *Information Retrieval* 3.2 (2000): 87-103.

[5] Aurangabadkar, Sumedha, and M. A. Potey. "Support Vector Machine based classification system for classification of sport articles." *Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on*. IEEE, 2014.

[6] Bhaskar, Sachin, Vijay Bahadur Singh, and A. K. Nayak. "Managing data in SVM supervised algorithm for data mining technology." *IT in Business, Industry and Government (CSIBIG), 2014 Conference on*. IEEE, 2014.

[7] Omer, Galal, et al. "Performance of support vector machines and artificial neural network for mapping endangered tree species using WorldView-2 data in Dukuduku forest, South Africa." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8.10 (2015): 4825-4840.

[8] Aurangabadkar, Sumedha, and M. A. Potey. "Support Vector Machine based classification system for classification of sports articles." *Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on*. IEEE, 2014.

[9] Steinwart, Ingo, and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[10] Zaki, Mohammed J., and Wagner Meira Jr. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

[11] Gao, Ya, and Shiliang Sun. "An empirical evaluation of linear and nonlinear kernels for text classification using support vector machines." *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on*. Vol. 4. IEEE, 2010.

[12] Hossain, Md Mosharaf, and Mohammad Sujan Miah. "Evaluation of different SVM kernels for predicting customer churn." Computer and Information Technology (ICCIT), 2015 18th International Conference on. IEEE, 2015.

[13] Prajapati, Gend Lal, and Arti Patle. "On performing classification using SVM with radial basis and polynomial kernel functions." Emerging Trends in Engineering and Technology (ICETET), 2010 3rd International Conference on. IEEE, 2010.

[14] Dsouza, Kevin Joy, and Zaheed Ahmed Ansari. "A Novel Data Mining Approach for Multi Variant Text Classification." *Cloud Computing in Emerging Markets (CCEM), 2015 IEEE International Conference on*. IEEE, 2015.