2016 16th International Conference on Control, Automation and Systems (ICCAS 2016)
Oct. 16-19, 2016 in HICO, Gyeongju, Korea

Medical Decision Making Diagnosis System Integrating k-means and Naïve Bayes algorithms

°Aigerim Altayeva[1], °Suleimenov Zharas[1], Young Im Cho[2*]
Department of Computer Engineering, Gachon University, Seongnam, Seoul
1) (tel: +82-31-750-5800; E-mail: aikosha1703@gmail.com, nonpensavo@gmail.com)
2)* (tel: +82-31-750-5800; E-mail: yicho@gachon.ac.kr) corresponding author

Abstract
In this paper, by using data mining we can evaluate many patterns which will be use in future to make intelligent systems and decisions By data mining refers to various methods of identifying information or the adoption of solutions based on knowledge and data extraction of these data so that they can be used in various areas such as decision-making, the prediction value for the prediction and calculation. In our days the health industry has collected vast amounts of patient data, which, unfortunately, is not "produced" in order to give some hidden information, and thus to make effective decisions, which are connected with the base of the patient's data and are subject to data mining. This research work has developed a Decision Support in Heart Disease Prediction System (HDPS) using data mining modelling technique, namely, Naïve Bayes and K-means clustering algorithms that are one of the most popular clustering techniques; however, where the initial choice of the centroid strongly influences the final result. Using of medical data, such as age, sex, blood pressure and blood sugar levels, chest pain, electrocardiogram, analyzes of different study patient, etc. graphics can predict the likelihood of the patient. This paper shows the effectiveness of unsupervised learning techniques, which is a k-means clustering to improve teaching methods controlled, which is naive Bayes. It explores the integration of K-means clustering with naive Bayes in the diagnosis of disease patients. It also investigates different methods of initial centroid selection of the K-means clustering such as range, inlier, outlier, random attribute values, and random row methods in the diagnosis of heart disease patients. The results indicate that the integration of the K-means clustering with naïve Bayes with different initial centroid selecting naive Bayesian improve accuracy in diagnosis of the patient.

Keywords: Data Mining, Naïve Bayes, K-Means Clustering, Artificial Intelligence.

1. Introduction

Data mining this is discovery process in the raw data previously unknown, non-trivial, practically useful, the interpretation of the available knowledge necessary for decision-making in the various spheres of human activity. This search for relationship with existing large associated data that are hidden among large amounts of data and refers to the "mining" knowledge from large amounts of data. Existing systems are used to assist in decision-making, referred to as data mining. These systems represent an iterative sequence of pre-processing as cleaning, data integration, and data selection is correct the pattern identification of data mining and knowledge representation. Data mining is the search for relationships and global patterns that exist in large databases, but hidden among the large amounts of data. Computer diagnosis of diseases is the doctor for the same instrument, the calculations for an engineer: design diagnostics does not replace the doctor, but it helps. Therefore, important to develop mathematical methods of diagnostics and compare their effectiveness. For example, prior art works used for diagnosing diagnostic features defined in one day, mainly during hospitalization. Signs in subsequent days are not involved in the calculations, i.e. not taken into account the dynamics of the disease the most important factor when making a diagnosis. This example illustrates the method development relevance, taking into account the dynamics of diseases, as is done in this paper. Most hospitals today use decision-support systems, but to get the results of the disease are largely limited. They can answer simple questions such as "What is the average age of the patients with cardiovascular disease?" "After surgery, many patients must remain for more than one week?" "Determine based on gender, marital status and who have been treated for heart failure." Solutions are always made in a hospital based on intuition and experience of doctors, and not on the rich knowledge data that are hidden in the database. This process leads to undesirable biases, errors and unnecessary health care costs, which affects the quality of services provided to patients. Machine learning can be used to determine the automatic conclusion of diagnostic rules from the past descriptions, successfully treat a patient, as well as experts and specialists will help make the diagnostic process more objective and more reliable. Intelligent decision support systems are defined as interactive computer systems to help make decisions in the use of data sets and models to find problems, solve problems and make decisions [1]. The proposed system uses the analysis to integrate and make the right decision at the clinic with a computer system. This patient record can reduce the number of patients to improve the safety of medical decisions errors, reduces unwanted changes in practice and improve patient outcomes. This proposal is promising, as modeling and analysis instruments, such as data mining, have the ability to generate knowledge-rich environment that can help to significantly improve the quality of clinical decisions [2]

In section 2, we talk K-means clustering algorithm. In section 3, we explain the overall structure of K-means and Naïve Bayes algorithms, and in section 4, we explain of our algorithm and shows the experimental results. Finally, we will summary our results.

2. K-Means Clustering

2.1 Data Source

978-89-93215-12-0/16/$31.00 ⓒICROS

Clinical database has accumulated a wealth of information about patients and their medical conditions. Archives of Medical Information at the disposal of various medical institutions contain a huge stock of information about different cases of each particular disease. Removing the "hidden" from the data set of laws - one of the tasks of many medical subjects' research. To solve these problems apply automated analysis methods by accounting for almost get knowledge from the "debris" of information. Cardiovascular diseases are the main cause of loss in the world. The term "cardiovascular disease" includes a very wide range of services that affect cardiovascular disease and blood vessels. This is a serious illness, disability and death. Medical record attribute set has been obtained from the database Cleveland heart disease. With the help of a set of data patterns is vital for predicting heart attack removed. Records were divided equally into two data sets: training dataset and testing dataset [4]. To avoid bias, for each set were randomly selected. Attribute "Diagnosis" is known as the predictable attribute with a value of "1" for patients with heart disease and the value of "0" for patients without cardiovascular disease. "Subject" is used in the recording, the last attribute as output data, and the other input attributes. It is assumed that this problem was solved like missing conflicting and redundant data

Predictable attribute

1. Diagnosis (value 0: <50% diameter narrowing (no disease); value 1: >50% diameter narrowing (has disease)) [3]

Table 1. Input attributes and values data

		Value 0	Value 1	Value 2	Value 3	Value 4	Value 5
1	Age in Year		-	-	-	-	-
2	Sex	Male	Female		-	-	-
3	Chest Pain Type	-	Typical type 1 angina	Typical type 2 angina	Non-angina pain	Asymptomatic	-
4	Fasting Blood Sugar	<120 mg/dl	>120 mg/dl	-	-	-	-
5	Resting	normal	Having ST-T wave abnormality	Showing probable or definite left ventricular hypertrophy	-	-	-
6	Exercise induced angina	No	Yes				
7	Slope – the slope of the peak exercises segment		Unsloping	Flat	Down sloping		
8	CA – number of major vessels colored by fluoroscopy		0	1	2	3	-
9	Thal – the heart status		Normal	Fixed defect	Reversible defect		
10	Trust Blood Pressure	Resting blood pressure					
11	Cholesterol	Serum Cholesterol					
12	Halacha	Maximum heart rate achieved					
13	Old peak	ST Depression Induced by exercise					
14	Heart Disease Present	No	Yes				

K-means Clustering is often referred to as unsupervised learning. Because there is no need for a marked data, learning algorithms without a teacher are suitable for many applications where the labeled data are difficult to obtain. Uncontrolled tasks such as clustering, as often used, which would investigate and characterize the data set before starting a controlled learning objectives. Since clustering is performed without using a class label, some idea of the similarity must be determined on the basis of object attributes. Description of similarities and method in which dots are grouped differ based clustering algorithm is used. K-means algorithm is a simple iterative clustering algorithm, which divides a particular set of data on the number of clusters, k specified by the user. The algorithm is simple to implement and run relatively fast, easily adaptable and common in practice. It is historically one of the most important data mining algorithms.
The flow chart in Figure 1 below shows the different steps of k-means clustering algorithm [3].

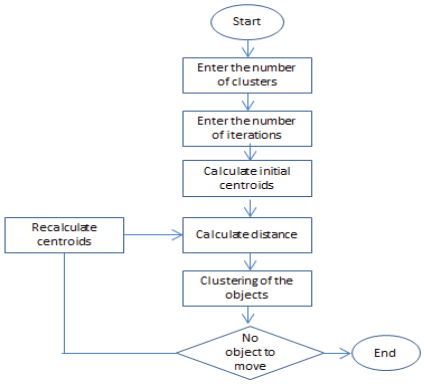

Figure 1. K-means clustering algorithm

The steps of the above flow chart are as follows:
Step1: Enter the number of clusters; this is "k" value.
Step2: After calculate the initial centroids from the actual datasets. Divide datapoints into "k" clusters.
Step3: By using Euclidean's distance formula (1) move the datapoints into clusters and recalculate new centroids. These centroids are calculated on the basis of average of means.
Step4: Repeat step 3until no datapoints is to be moved.

$$r(x_i, x_j) = \sqrt{\sum_{k=1}^{N} (x_{ik} - x_{jk})^2} \quad (1)$$

Where $r(x_i, x_j)$ is the distance between x_i and x_j. x_i and x_j are the attributes of a given data, where i, j and k vary from 1 to N. N is total number of attributes of that given data. After we compute new cluster centers as centroids, and iterate this process until the cluster members are stabilized. This algorithm uses a square-error criterion for re-assignment of any sample from one cluster to another. The sum of square-error value "E" is shown in equation 2.

$$E = \sum_{i=1}^{X} \sum_{j=1}^{Y} \sum_{k=1}^{Z} (A - B)^2 \quad (2)$$

Where $(A - B)^2$ is the distance between the data points. The main problems in k-means clustering algorithm shows below: the user needs and must be specify the number of clusters k; the algorithm is valid only to datasets; the algorithm is sensitive to initial seeds and to outliers; the k - means algorithm is not valid for discovering clusters that are not hyper-spheres [4]. This is a main problem of k-means clustering algorithm.

2.2 Naive Bayesian Classification Algorithm

Bayesian classification presents supervised learning method, and the method of statistical classification. It assumes a basic model of probability, and it allows us to capture some uncertainty about the model by determining the probability of the outcome source. It can solve problems of diagnosis and prognostic. Bayesian approach to classification is based on the theorem, which states that if the density distribution of each of the classes is known, the desired algorithm can be written in explicit analytic form. Furthermore, this algorithm is optimal, that is, has the minimum error probability. It calculates the precise probability of the hypothesis, and it is resistant to the noise at the input [5] data.
Naïve Bayesian method is particularly relevant for the problems of high dimensionality of the input space, in case of problems with a large number of input variables. Despite its simplicity, Naive Bayes often superior to other more sophisticated classification methods and this model takes on different characteristics of the patients with different diseases, define and determines the probability of each input source for a predictable state. Bayes's Rules one of the main and basic algorithms of machine learning and data mining methods. The algorithm is used to create models with predictive capabilities. It provides all possible new ways of learning and understanding the data.
Why to implement the preferred algorithm Naive Bayes:
1) When working with very large datasets.
2) When the attributes are independent of each other.
3) When we expect that in comparison to other production methods, a more effective output.
Bayes' formula allows you to "rearrange the cause and effect": event on the known fact to calculate the probability that it was due to this reason.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (3)$$

Where $P(A)$ - is a priori probability of the hypothesis, $P(A|B)$- the probability of a hypothesis A when an event occurs in B; $P(B|A)$ - the probability of an event B at the truth of the hypothesis A; $P(B)$ – the total probability of occurrence B.

2.2.1 Naive Bayesian Classifier

A Naive Bayesian Classifier is a a simple probabilistic classifier based on applying Bayes' theorem to strict (naive) assumptions about independence. Depending on the exact nature of the probabilistic model, Naive Bayes classifier

can be trained very effectively. The classifier, which uses the Naive Bayesian formula to calculate the probability of each class A given the values B_i of all attributes for an instance to be classified, the conditional independence of the attributes given the class:

$$P(A|B_1..B_n) = P(A)\prod_i \frac{P(A|B_i)}{P(A)} \tag{4}$$

The new instance is classified into the class with a maximal calculated probability. This makes the Naïve Bayesian classifier more accurate.

2.2.2 Semi-naive Bayesian Classifier

Semi-naive Bayesian Classifier: Kononenko (1991) in his research deeper learning and expanding Naïve Bayes's algorithm and developed the semi-naive Bayesian classifier that explicitly searches for dependencies between the values of different attributes. If such dependency is discovered between two values B_i and B_j of two different attributes, in this case the data are not considered as conditionally independent [5]. Accordingly, we have the term:

$$\frac{P(A|\,B_i)}{P(A)} \times \frac{P(A|B_j)}{P(A)} \tag{5}$$

In Equation (5) is replaced with

$$\frac{P(A|B_i, B_j)}{P(A)} \tag{6}$$

In this case a positive identity of the conditional probability $P(A|B_i, B)$ is necessary. Therefore, the accommodation of algorithms between the non-naivety and the veracity of identity of probabilities. Technique Naive Bayes classifier is especially designed for large input data. Despite its simplicity, Naive Bayes algorithm often outperform more sophisticated classification methods, and is often used to compute the solution of problems with high probability.

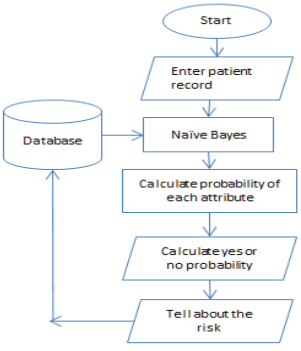

Figure 2. Realization of Naïve Bayes algorithm

Figure 2 illustrates the working schema of the system using Naïve Bayes algorithm of patient data. The program receives patient record from the interface and uses this data to Naïve Bayes algorithm. By using the Naïve Bayes method, possible attributes will be determined and probability of each attribute will be calculated. Then yes or no probability of each attribute will be computed, and depending on these results the information about risk will be returned [6]. The received results will be written into the database, and it will supplement the knowledge base, also it will be used to solve next problems.

3. Overall structure of K-means and Naïve Bayes algorithms

Figure 3 illustrates the working scheme of the system, after starting and giving the input parameters the system will ask to choose train or test. If train was chosen, then the system starts to learn and supplement the knowledge base.

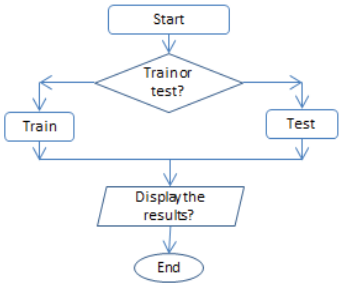

Figure 3. Working schema of the system.

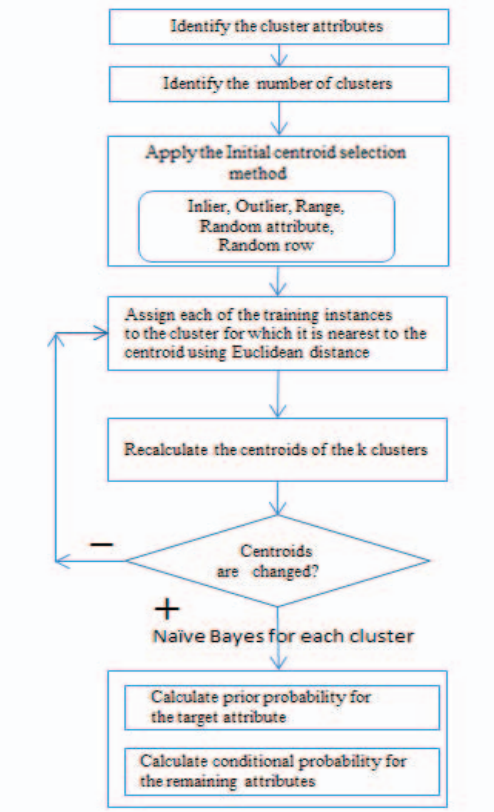

Figure 4. K-means clustering and Naïve Bayes integrated training algorithm

The steps of the training that used in k-means clustering and Naïve Bayes method are shown in Figure 4. K-means clustering algorithm is one of the main basic clustering techniques, because it is simple to use and it has good behaviors in many applications.

In our days many researchers have identified that age, blood pressure and cholesterol are critical risk factors

associated with heart disease. In recognizing some attributes that will be used the clustering, these attributes are evident by clustering those attributes for heart disease of patients. The number of clusters used in the k-means in this investigation ranged between two and five clusters.

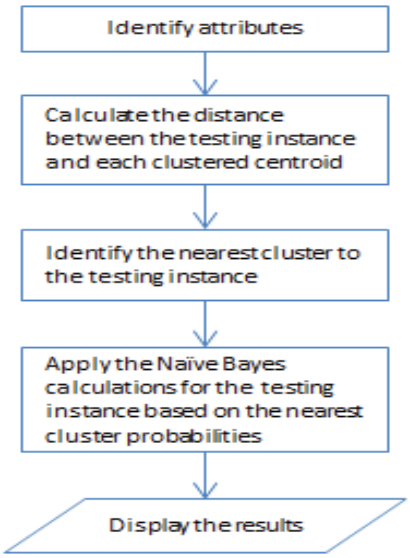

Figure 5. K-means clustering and Naïve Bayes integrated testing algorithms

Figure 5 presents testing part of the system integrating K-Means clustering and Naïve Bayes algorithm. It consists of several expressions such as, compute distance between the testing instance and each clustered centroid, identify the nearest cluster, and apply the Naïve Bayes calculations, in the end it returns the received results.

4. Learning model of agent based intelligent decision making system

Different types of decision making systems were simulated for different parameter settings.

Table 2. Experiment results of simple and intelligent decision making systems

	Sensitivity	Specificity	Accuracy	FPR	FNR	Training time (seconds)
Simple decision-making system	28%	34%	32%	42%	65%	956
Intelligent decision-making system	54%	39%	56%	54%	20%	3872

Intelligent decision making system uses Big Data analysis and Simple decision making system use own static database. The database contains diagnosis results of patients. Experimental results of decision making systems are shown in the Table 2 and Figure 6.

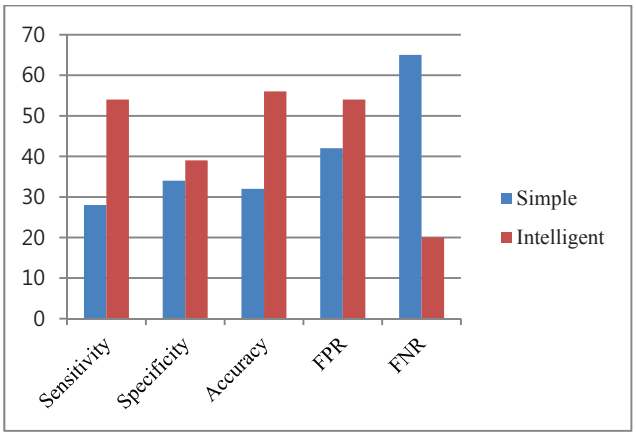

Figure 6. Experiment results of simple and intelligent decision making systems

As it shown in the Table 3, intelligent decision making system training requires more time consuming (about four times) because of more data, however it gives more accuracy of diagnosis.

In our system we use human's knowledge and artificial agents. The human's knowledge based medical dates that have been set of doctor during examination of a patient and some dates that were previously known. Artificial agents are part of the system that have own knowledge base and can communicate with main database and also agents will help to solve part of the main problem which showed up in the result of dividing several sub-problems of the diagnosis problem [7].

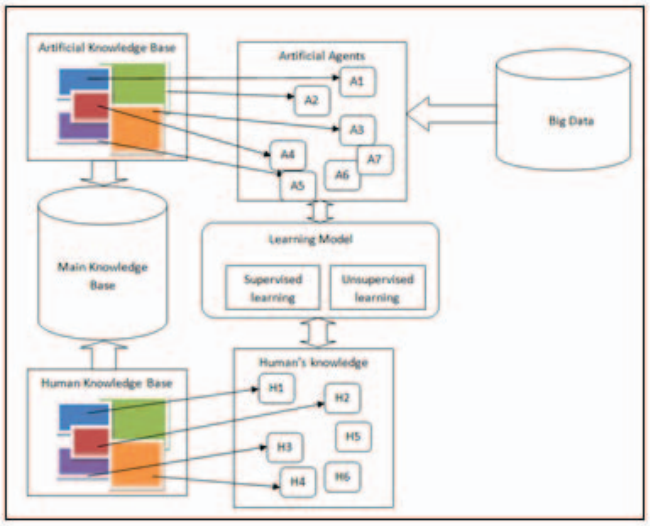

Figure 7. Overall structure of artificial agents and human's knowledge based learning model.

As an illustration of sub-problem of a main problem we can say that collect medical information about patients, making an interim decision, analysis necessity for accuracy of disease identification, and identification of the disturbance of an internal organ of patient, based on different information taken from medical survey. Figure 7 constructs on human's knowledge and artificial agent based decision making system model using big data processing. A1, A2, …, An are the artificial agent participants of the scheme, and H1, H2, …, Hn are human's knowledge based on medical data. In own case artificial agents can be process

unstructured data from several source of information. Each artificial agent has own knowledge base, and it served to complete the main knowledge base. Learning results of each artificial agent will be contributed to supplement of knowledge base. Initially, the system needs supervised learning to elaborate the knowledge base. It helps to get better knowledge and give better decision.

5. Experimental Results

The number of clusters could increase their accuracy. In our experiment, we compared the accuracy rates of each method, for the caste number of clusters from 2 till 5. As the results show, most of the methods return high accuracy for the case that clusters number equal to 2 or 3.

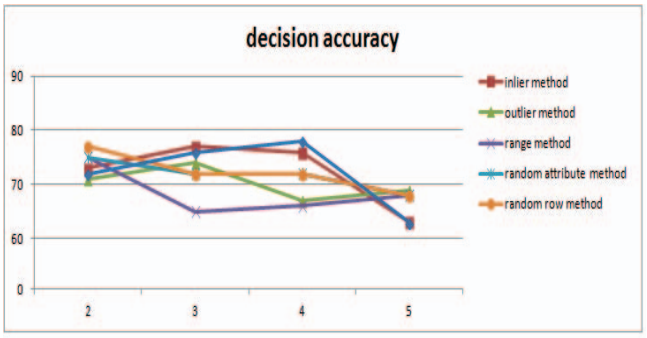

Figure 8. Accuracy rate changes of each method depending on number of clusters

Using the result of each method we build the graph of integration of all the methods. So, it looks like Figure 8.

Table 3. Accuracy rates of each method

Inlier method	73%	77 %	76%	63 %
Outlier method	71%	74%	67%	69 %
Range method	75%	65%	66%	67 %
Random attribute method	75%	72%	72%	68 %
Random row method	76%	72%	72%	68 %

Using the received results, we created Table 3 that presents accuracy rate of each method and by using their maximum values Figure 8 was created. Figure 9 shows that the changing of accuracy rate depends on number of clusters.

6. Conclusion

Today exists many expert systems that operate on the basis of different methods and are used in many areas of medicine. However, a common feature of medical Expert Systems can be identified lack of a unified technology of their creation. For the most part, in development and in the full operating systems, based on different operation algorithms. In this article, we show the effectiveness of unsupervised learning techniques of data, which used K-means clustering algorithm to improving supervised learning technique which is Naive Bayes. It is explores the integration of K-means clustering with naive Bayes in the structure of the decision making system in medicine. It also investigates different methods of initial centroid selection of the K-means clustering such as range, inlier, outlier, random attribute values, and random row methods in the diagnosis of the patients. The results indicate that the integration of the K-means clustering with naïve Bayes algorithm with different initial selection can improve the accuracy in diagnosis of patient. Also, Big Data processing through artificial agents and human's knowledge based shows, and their integration were presented. Agents based on k-means clustering by using Bayesian, its architecture, and data mining methods are considered. Data mining approach for processing Big Data can solve many problems of analyzing large and growing data sets. During the research agent based learning model had been created. Then, for future work, we will improve this intelligent decision making system by using other new models and apply them to other environments.

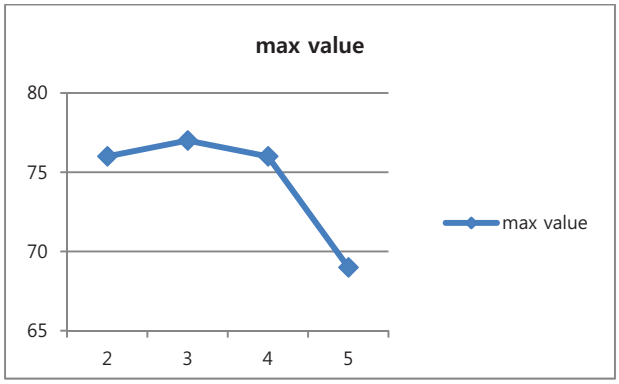

Figure 9. The changing of accuracy rate depends on number of clusters

7. Acknowledgements

This paper is supported by NRF project "Intelligent Smart City Convergence Platform. Project number is 20151D1A1A01061271

References

[1] Cao, L., Zhang, Z., Gorodetsky, V., & Zhang, C. Interaction between agents and data mining, 2008
[2] Davidson, Ian. "Understanding K-Means Non-hierarchical Clustering", SUNY Albany-Technical Report 2002.
[3] http://www.ijarcce.com/
[4] Xindong Wu · Vipin Kumar · J. Ross Quinlan "Top 10 algorithms in data mining", 4 December 2007, Springer-Verlag London Limited 2007.
[5] Igor Kononenko, "Semi-Naïve Bayesian Classifier", Springer Berlin Heidelberg, March 6–8, 1991,p 206-219
[6] K. Ming Leung, "Naive Bayesian Classifier", November 28, 2007
[7] Peng, Y., Kou, G. Vol. 7, Issue: 4, Page 639-682, 2008. International Journal of Information Technology and Decision Making System.