IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 1, January 2015.
www.ijiset.com
ISSN 2348 – 7968

K-Means Clustering and Naive Bayes Classifier For Categorization Of Diabetes Patients

L.Pandeeswari*, K.Rajeswari, M.Sc., M.Phil.**

*M.Phil(Computer Science), Research Scholar,
Vivekanandha College for Women, Unjanai, Tiruchengode. India

**Assistant Professor in Computer Science
Vivekanandha College for Women, Unjanai, Tiruchengode. India

Abstract

Data mining is essentially the discovery of valuable information and patterns from huge chunks of available data. This paper presents the development of a hybrid model for classifying Pima Indian diabetic database (PIDD). Clustering and classification are two important techniques of data mining. Classification is a supervised learning problem of assigning an object to one of several pre-defined categories based upon the attributes of the object. While, clustering is an unsupervised learning problem that group objects based upon distance or similarity. We make use of a large database 'Pima Indian diabetic database (PIDD)' containing 768 instances and 7 attributes to perform an integration of clustering and classification techniques of data mining. The model consists of two stages. In the first stage, the K-means clustering is used to identify and eliminate incorrectly classified instances. In the second stage a fine tuned classification is done using Naive Bayes by taking the correctly clustered instance of first stage. Experimental results signify the cascaded K-means clustering and Naive Bayes has enhanced classification accuracy. We compared the results of simple classification technique (Naive Bayes algorithm) with the results of integration of clustering (K-Means) and classification techniques based upon various parameters using WEKA (Waikato Environment for Knowledge Analysis) a data mining tool. The results of the experiment show that integration of clustering and classification gives promising results with utmost accuracy rate even when the dataset contains missing values.

Keyword: Data Mining, KDD, K-Means, Naïve Bayes,WEKA

1. Introduction:

Data mining refers to extracting or "mining" knowledge from large amounts of data. Data Mining is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data with the wide use of databases and the explosive growth in their sizes. Data mining is the search for the relationships and global patterns that exist in large databases but are hidden among large amounts of data.

Data mining is the task of discovering interesting patterns from large amounts of data, where the data can be stored in databases, data warehouses, or other information repositories. It is a young interdisciplinary field, drawing from areas such as database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high-performance computing.

ISSN 2348 – 7968

2. Diabetes Dataset:

Diabetes mellitus is a disease in which the body is unable to produce or unable to properly use and store glucose (a form of sugar). Glucose backs up in the bloodstream causing one's blood glucose or "sugar" to rise too high. Most of the food we eat is converted to glucose, or sugar which is used for energy. The pancreas secretes insulin which carries glucose into the cells of our bodies, which in turn produces energy for the perfect functioning of the body.

General Symptoms of Diabetes:

- ✓ Increased thirst.
- ✓ Increased urination - Weight loss
- ✓ Increased appetite – Fatigue
- ✓ Nausea and/or vomiting - Blurred vision
- ✓ Slow-healing infections - Impotence in men

3. Related Work:

Pardha Repalli, in their research work predicted how likely the people with diverse age groups are affected by diabetes based on their activities. They also found out factors responsible for the individual to be diabetic. Statistics given by the Centers for Disease Control states that 26.9% of the population affected by diabetes are people whose age is greater than 65, 11.8% of all men aged 20 years or older are affected by diabetes and 10.8% of all women aged 20 years or older are affected by diabetes. The dataset used for analysis and modeling has 50784 records with 37 variables.

Joseph L. Breault in his research work used the publicly available Pima Indian diabetic database (PIDD) at the UCIrvine Machine Learning Lab. They tested data mining algorithms to predict their accuracy in predicting diabetic status from the 8 variables given. Out of 392 complete cases, guessing all are non-diabetic gives an accuracy of 65.1%.

Padmaja et al., in their research aimed at finding out the characteristics that determine the presence of diabetes and to track the maximum number of women suffering from diabetes. They used Data mining functionalities like clustering and attribute oriented induction techniques to track the characteristics of the women suffering from diabetes. Information related to the study was obtained from National Institute of Diabetes, Digestive and Kidney Diseases.

Samir Kumar Sarangi and Vivek Jaglan proposed the simple k-means clustering algorithm and this integration technique were applied on "Diabetes Diagnosis" data set. From our observation and analysis it was concluded that the integration of K-means (clustering) + J48 (classification) have zero MAE and RMSE error and it also takes less time to build the model. So the performance of K-means+J48 is better than other algorithms.

4. Methodology

4.1 K-Means Clustering Algorithm:

K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far

away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center.

When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroid as bar center of the clusters resulting from the previous step. After we have these k new centroid, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this K-means algorithm aims at minimizing an objective function namely sum of squared error (SSE). SSE is defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2$$

Where E is sum of the square error of objects with cluster means for k cluster. p is the object belong to a cluster Ci and mi is the mean of cluster Ci.

Algorithm: *k*-means

The *k*-means algorithm for partitioning, where each cluster center is represented by the mean value of the objects in the cluster.

Input:

k: the number of clusters,

D: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;

(2) repeat

(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

(4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;

(5) until no change;

4.2 Naive Bayes Classifiers:

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes models are also known under a variety of names in the literature, including simple Bayes and independence Bayes. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naive Bayes is not (necessarily) a Bayesian method. Russell and Norvig note that "[naive Bayes] is sometimes called a Bayesian classifier, a somewhat careless usage that has prompted true Bayesians to call it the idiot Bayes model."

How Naive Bayesian Works:

1. Let *D* be a training set of tuples and their associated class labels. As usual, each tuple is represented by an *n*-dimensional attribute vector, ***X*** = (*x*1, *x*2, : : : , *xn*), depicting *n* measurements made on the tuple from *n* attributes, respectively, *A*1, *A*2, : : : , *An*.

ISSN 2348 – 7968

2. Suppose that there are *m* classes, *C*1, *C*2, : : : , *Cm*. Given a tuple, ***X***, the classifier will predict that ***X*** belongs to the class having the highest posterior probability, conditioned on ***X***. That is, the naive Bayesian classifier predicts that tuple ***X*** belongs to the class *Ci* if and only if

$$P(Ci|\boldsymbol{X}) > P(Cj|\boldsymbol{X}) \text{ for } 1 <= j <= m; j \neq i$$

Thus we maximize $P(Ci|\boldsymbol{X})$. The class*Ci* for which $P(Ci|\boldsymbol{X})$ is maximized is called the

Maximum posteriori hypothesis. By Bayes' theorem

$$P(Ci|\boldsymbol{X}) = P(\boldsymbol{X}|Ci)P(Ci)/P(\boldsymbol{X})$$

3. As $P(\boldsymbol{X})$ is constant for all classes, only $P(\boldsymbol{X}|Ci)\ P(Ci)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C1) = P(C2) = \ldots\ldots = P(Cm)$, and we would therefore maximize $P(\boldsymbol{X}|Ci)$. Otherwise, we maximize $P(\boldsymbol{X}|Ci)P(Ci)$. Note that the class prior probabilities may be estimated by $P(Ci)=|Ci,D|/|D|$,where $|Ci,D|$ is the number of training tuples of class *Ci* in *D*.

4. Given data sets with many attributes, it would be extremely computationally expensive to compute $P(\boldsymbol{X}|Ci)$. In order to reduce computation in evaluating $P(\boldsymbol{X}|Ci)$, thenaive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes).

5. Experiments and Results

Diabetes Dataset where object corresponds to Diabetic result and object class label corresponds to results of diabetes. Every diabetes result consists of various parameters which are used to predict the result of diabetic. Apply clustering technique on the original data set using WEKA tool and we come up with a number of clusters. It also adds an attribute "cluster" to the data set. Apply classification technique on the resulting data set which is obtained after clustering. Then compare the results of simple classification and an integration of clustering and classification. In this paper, we identified the finest classification rules through experimental study for the task of classifying Diabetic result as tested_positive or tested_negative using WEKA data mining tool.

Data Preprocessing:

It use of "Diabetes dataset" which consist of 768 instances and 7 attributes to perform integration of clustering and classification algorithm. The data is often presented in a spreadsheet or database. However Weka's native data storage method is ARFF format. Data can be easily converted from spreadsheet to ARFF format. The bulk of an ARFF file consists of a list of the instances, and the attribute values for each instance are separated by commas.

Attribute Description	
Name	**Description**
Plas	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Pres	Diastolic blood pressure(mm Hg)
Age	Age (years)
Pedi	Diabetes pedigree function
Class	Class variable (0 or 1)

ISSN 2348 – 7968

Cluster:

After loading the dataset and choosing the classifier, the results are displayed in the results panel of the tool. By using simple-k-means the dataset is grouped into number of clusters based on the value of k specified by the user. The value can provided by the user by clicking on the classifier text field, and providing the value of k in the number of clusters field. The results of the cluster can viewed in the figure below. The results of the classifier can be viewed in the cluster output panel. The number of clusters formed (the value of k is taken as) 4. In cluster centroid the number of attributes and the mean for each attribute in each cluster are provided. The first column displays the means of the complete dataset. And under each cluster number, the means of the clusters i.e., means of the instances in the cluster are provided. The numbers of instances that are formed are also mentioned and also the percentage of each cluster (based on the number of instances) is provided.

Steps for K-Means Method:

Step1: Select the number of clusters. Let this number be k.

Step2: Pick k seeds as centroid of the k clusters. The seeds may be picked randomly unless the user has some insight into the data.

Step3: Compute the Euclidean distance of each object in the dataset from each of the centroid.

Step4: Allocate each object to the cluster it is nearest to based on the distances computed in the previous step.

Step5: Compute the centroid of the clusters by computing the means of the attribute Values of the objects in each cluster.

Step6: Check if the stopping criterion has been met. If yes, go to step7. If not, go to step 3.

Step7: One may decide to stop at this stage or to split a cluster or combine two clusters heuristically until a stopping criterion is met.

Classification:

Naive bayes is an implementation of that builds decision trees from a set of training data in the same way using the concept of Information Entropy. The training data is a set S = s1, s2... of already classified samples. Each sample si = x1, x2... is a vector where x1, x2…represent attributes or features of the sample. Decision tree are efficient to use and display good accuracy for large amount of data.

The rules generated by the naive bayes classifier are given below.

1. If Plasma=low then class=> Tested Negative.

2. If Plasma =medium & Age=low & Pedigree =low then Class => Tested Negative.

3. If Plasma =medium & Age=low & Pedigree =medium & Diastolic BP=medium then Class=> Tested Negative.

4. If Plasma=medium & Age=low & Pedigree =medium & Diastolic BP =low then Class=> Tested Negative.

5. If Plasma=medium & Age=low & Pedigree =medium &Diastolic BP =high then Class=> Tested Positive.

6. If Plasma =medium & age=high then Class => Tested Positive.

7. If Plasma =medium & Age=low & Pedigree =high then Class=> Tested Positive.

Performance Evaluation:

ISSN 2348 – 7968

To measure the performance, the concepts such as TP Rate or Recall, FP Rate, F-Measure, Precision are measured.

The confusion matrix is used to identify the correctly and incorrectly classified instances. The 2×2 matrix representation.

Confusion matrix	
TP	TN
FP	FN

Confusion Matrix

TP Rate: (True Positive Rate)

It is simply the ratio of true positives to true positives plus false negatives. In an ideal world we want the TPR to be one. It can be defined as:

TPR = TP / TP + FN

FP Rate: (False Positive Rate)

It is simply the ratio of false positives to false positives plus true negatives. In an ideal world we want the FPR to be zero. It can be defined as:

FPR = FP / FP+TN

Precision:

In information retrieval positive predictive value is called precision. It is calculated as number of correctly classified instances belongs to X divided by number of instances classified as belonging to class X; that is, it is the proportion of true positives out of all positive results. In an ideal world we want the precision to be one. It can be defined as:

Precision = TP / TP+FP

F-Measure:

F-Measure is a way of combining recall and precision scores into a single measure of performance. The formula for it is:

Recall * 2 * Precision / Recall + Precision

6. Conclusion and Future Enhancement:

A comparative study of data mining classification technique and an integration of clustering and classification technique helps in identifying large data sets. It technique gives more accurate results than simple Classification technique to classify data sets whose attributes and classes are given to us.

This integrated technique of clustering and classification gives a promising classification results with utmost accuracy rate and robustness. In future work binary classifiers and try to find the results from the integration of classification, clustering and association technique of data mining.

7. Reference:

[1] J. Han and M. Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufmann, 2000.

[2] Ritu Chauhan, Harleen Kaur, M.Afshar Alam, “Data Clustering Method for Discovering Clusters in Spatial Cancer Databases”, International Journal of Computer Applications (0975 – 8887) Volume 10– No.6, November 2010.

[3] Thair Nu Phyu “Survey of Classification Techniques in Data Mining”, International MultiConference of Engineers and Computer Scientists 2009 Vol I , March 18 - 20, 2009.

[4] Jain, A.K., Murty M.N., and Flynn P.J.” Data Clustering: A Review” (1999).

[5] Rakesh Agrawal,Tomasz Imielinski and Arun Swami,” Data mining : A Performance perspective“. IEEE Transactions on Knowledge and Data Engineering , 5(6):914-925, December 1993.

ISSN 2348 – 7968

[6] J.R Quinlan and R..L. Rivest. Inferring decision tress using minimum description length principle. Information and computation, 1989.