Performance Enhancement of Classifiers using Integration of Clustering and Classification Techniques

G.Keerthana
PG Student
Department of Computer Science
Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore, Tamilnadu, India
Email ID : keerthana24zeal@gmail.com

Dr. V.Srividhya
Assistant Professor
Department of Computer Science
Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore, Tamilnadu, India
Email ID : vidhyavasyu@gmail.com

***Abstract* - Medical professionals need a reliable prediction methodology to diagnose Diabetes. Data mining is the process of analysing data from different perspectives and summarizing it into useful information. The main goal of data mining is to discover new patterns for the users and to interpret the data patterns to provide meaningful and useful information for the users. Data mining is applied to find useful patterns to help in the important tasks of medical diagnosis and treatment. In this paper, performance comparison of simple classification algorithms and integrated clustering and classification algorithms are carried out. It was found that the integrated clustering-classification technique was better than the simple classification technique. Data mining tool used is WEKA. PIMA INDIANS DIABETES dataset is used.**

Keywords : *Data mining, classification, integrated clustering-classification, WEKA, Pima Indians Diabetes dataset.*

I. INTRODUCTON

Data mining is the process of automatic classification of cases based on data patterns obtained from a dataset. A number of algorithms have been developed and implemented to extract information and discover knowledge patterns. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD). The KDD method is applied on large amount of data from stored database/data warehouse/any data repository for extracting patterns, relationships, changes, anomalies and hidden or core information using algorithms and techniques.

Classification is a basic task in the data analysis that requires the construction of a classifier, that is, a function that assigns a class label to instances described by a set of attributes. It is one of the important applications of data mining. This technique predicts categorical class labels. It is a supervised learning technique. Classification of data is very typical task in data mining. The goal of classification is to correctly predict the value. Classification Algorithms used are NavieBayes, BayesNet, OneR and J48.

Clustering is a data mining technique to group the similar data into a cluster and dissimilar data into different clusters**.** Clustering is an unsupervised learning technique; it deals with finding a structure in a collection of unlabeled data. Clustering is the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. Clustering Algorithms used are Simple K-Means and Density Based Clustering.

Integration of clustering and classification technique is useful even when the dataset contains missing values. Performance of classifiers has been improved due to integrating.

II. PROPOSED METHOD

Classification is the process which finds the common properties among a set of objects in a database and classifies them into different classes, according to a classification model. Clustering is the task of segmenting a diverse group into a number of similar subgroups or clusters. In this proposed technique first the clustering algorithm is applied on the dataset with the help of any clustering algorithms such as Simple K-Means and Density Based Clustering. Clustering algorithm adds the attribute “cluster on the dataset. After that, classification algorithm is applied on this clustered dataset. This approach gives results with a better accuracy than the simple classification technique.

1.DATASET:

Weka uses data set (Attribute-relationship) file of “.arff” format. This data set consists of attribute names, types, values and the data. In this paper, “Diabetes Diagnosis” is used. Data set contains eight attributes, one class attribute and 768 instances.

Table 1: Diabetes Dataset Attributes

S. No	Attributes	Type
1	Number of times pregnant	Continuous
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Continuous
3	Diastolic blood pressure (mm Hg)	Continuous
4	Triceps skin fold thickness (mm)	Continuous
5	2-Hour serum insulin (mu U/ml)	Continuous
6	Body mass index (kg/m)^2)	Continuous
7	Diabetes pedigree function	Continuous
8	Age (years)	Continuous
9	Class variable (0 or 1)	Discrete

2. CLUSTERING:

Clustering is an unsupervised method of data mining. In clustering user needs to define their own classes according to class variables, here no predefined classes are present. In weka number of clustering algorithms are present like cobweb, DBSCAN, FarthestFirst, SimpleK-Means etc.

***Simple K-Means algorithm*:**

The basic step of k-means clustering is simple. In the beginning, we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects can also serve as the initial centroids. Then the K means algorithm will do the four steps below until convergence.

1) Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2) Assign each object to the group that has the closest centroid.
3) When all objects have been assigned, recalculate the positions of the K centroids.
4) Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Density Based Clustering:

Density-based clustering algorithms try to find clusters based on density of data points in a region. The key idea of density-based clustering is that for each instance of a cluster the neighbourhood of a given radius (Eps) has to contain atleast a minimum number of instances (MinPts). To find a cluster, it starts with an arbitrary instance (p) in data set (D) and retrieves all instances of D with respect to Eps and Min Pts. The algorithm makes use of a spatial data structure to locate points within Eps distance from the core points of the clusters. Algorithm is given below.

1) Start with an arbitrary starting point that has not been visited.

2) Extract the neighborhood of this point using ε (All points which are within the ε distance are neighborhood).

3) If there is sufficient neighborhood around this point then clustering process starts and point is marked as visited else this point is labeled as noise (Later this point can become the part of the cluster).

4) If a point is found to be a part of the cluster then its ε neighborhood is also the part of the cluster and the above procedure from step 2 is repeated for all ε neighborhood points. This is repeated until all points in the cluster are determined.

5) A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

6) This process continues until all points are marked as visited.

4. CLASSIFICATION:

Data classification is a very important task in data mining. For classifying a data mining problem, numbers of classification algorithms are used like Bayes, Functions, misc, Rules, Trees etc. The aim of classification is to calculate the values of each variable and assign those variables to matched predefined classes. In this paper, four different classification algorithms have been used, which have been listed below:

NaïveBayes (NB): An independent feature probability model, it is based on the Bayes theorem and is thus a probabilistic classifier.

BayesNet(BN) : By using the bayes theorem BayesNet can be developed. To structure a Baysian network first conditional probability of every node must be calculated. Acyclic graphs are used to represent the network.

OneR : OneR, short for "One Rule", is a simple, accurate, classification algorithm that generates one rule for each predictor in the data, then selects the rule with the smallest total error as its "one rule".

J48: It is an optimized version of C4.5 algorithm. When any specific data item is classified, it will be divided in different levels starting from root node to the leaf or terminal node in a hierarchical manner.

III. EXPERIMENTAL RESULTS

The experiments performed on the dataset gave the results as shown below. This table shows the accuracy measure of the simple classification algorithms, integration of clustering and classification using Simple K-Means and integration of clustering and classification using Density Based Clustering Algorithm. From the table, it is clear that the performance of classifiers has been improved after clustering.

Table 2 : Comparative Table for Accuracy

Algorithms	**Simple Classification (in %)**	**K-Means + Classification (in %)**	**Density Based + Classification (in %)**
BayesNet	74.349 %	94.5313 %	95.7031%
NavieBayes	**76.3021 %**	**95.8333 %**	**96.3542%**
OneR	71.4844 %	89.976 %	91.4063%
J48	73.8281%	95.7031 %	95.0521%

The following graph illustrates the tabulated results shown above:

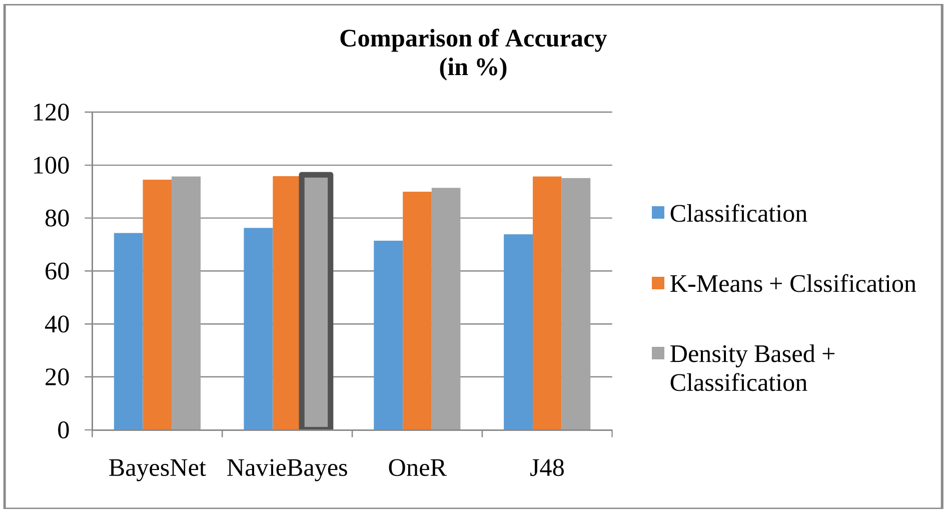

Figure I : Classification and Integration of Clustering and Classification Algorithms

IV. CONCLUSION AND FUTURE WORK

In this paper four different classifiers are integrated with the simple k-means clustering algorithm and density based clustering algorithm. This integration technique was applied on “Diabetes” data set. From the observation and analysis it was concluded that the performance of Density Based + NavieBayes is better than other algorithms because of the following features:

1. Number of correctly classified instances is more
2. Absolute errors are less.

There are large numbers of classifiers present and many other data mining tools are present. So the future work will be based on other classifiers that can be applied on the data set and also to apply other data mining tools on the data set such that the best techniques can be identified. Above algorithms can be applied to other datasets in order to observe whether the same algorithm gives the highest accuracy.

V. REFERENCES

[1] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, 2006.

[2] Varun Kumar and Nisha Rathee , ITM University, “Knowledge discovery from database Using an integration of clustering and classification”, International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, March 2011.

[3] Weka 3-Data Mining with open source machine learning software available, http://www.cs.waikato. ac.nz/ml/ weka/

[4] UCI Repository - https://archive.ics.uci.edu/ml/datasets.html

[5] Narendra Sharma, Aman Bajpai , Mr. Ratnesh Litoriya, Jaypee University of Engg. & Technology, “Comparison the various clustering algorithms of WEKA Tool”, International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, Volume 2, Issue 5, May 2012)