Volume 5, Issue 5, May 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Classifiers Performance Improvement through Integration of Clustering Technique

Sambasiva Rao Voleti, Kiran Kumar Reddi
Department of CS & Krishna University,
Andhra Pradesh India

Abstract— Doctors need reliable automatic classification tools that reduce burden on them in predicting diseases. But the issue is how to find these tools. We can use data mining for this purpose. Data mining is the process of analyzing data from diverse perspectives and presenting it into useful information. The main objective of data mining is to investigating the new patterns for the users and to present these new patterns to provide meaningful and useful information. Data mining is applied to discover new patterns to help in the important tasks of medical diagnosis and treatment. There exist different data mining techniques, in that two important techniques are clustering and classification. In this paper we integrated both clustering and classification techniques. We compared the Performance of simple classification algorithms and integrated clustering-classification algorithms.. It was observed that the integrated clustering-classification technique was superior to the simple classification technique. We used the WEKA Data mining tool for results. And the data sets used were Wisconsin Diagnostic Breast Cancer dataset and Pima Indians Diabetes dataset.

Keywords— Data Mining, Classification, Integrated Clustering-Classification, WEKA, Wisconsin Diagnostic Breast Cancer Dataset, Pima Indians Diabetes Dataset.

I. INTRODUCTION

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 is 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 [1].

Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. In classification the first step is to divide the data in two portions known as training set and testing set. In these datasets, one attribute must be necessarily defined as class. According to Han et al. [1], the two steps of the classification task are model construction and model usage. In this task, the model is built with the help of trained dataset and then this trained model is used to allocate the unseen records as precisely as possible. While training dataset is used to build and train the model, the testing dataset is used to validate and test the model accuracy [1].

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 Algorithm used here is Simple K-Means.

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

One of the dataset used in this study was “The Pima Indians Diabetes Data Set” which was taken from the UCI Machine Learning Repository [2]. The original owner of this data set is the National Institute of Diabetes and Digestive and Kidney Diseases. Several constraints were placed on the selection of this dataset from larger database. In particular, all patients selected are females at least 21 years old of Pima Indian heritage.

Another dataset used in this study was “Wisconsin Diagnostic Breast Cancer (WDBC) data Set” which was taken from the UCI Machine Learning Repository [2]. The original owner of this data set is the University of Wisconsin. Several constraints were placed on the selection of this dataset from larger database.

WEKA software package was used throughout this study. WEKA software is a collection of machine learning algorithms for data mining tasks. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. WEKA system is open source software issued under GNU General Public License, where it can be modified by anybody for use [3].

II. 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, Farthest First, Simple K-Means etc. K-Means is the simplest technique and gives more accurate result than others [9].

***K-Means algorithm*:**

1) Select number of clusters to be divided.
2) Select initial value of centroid for each cluster randomly from the set of instances.
3) Object clustering
    I. Measure the Euclidean distance (Manhattan distance median value) of each object form the centroid.
    II. Place the object in the nearest cluster, whose centroid distance is minimal.
4) After placing all objects calculate the MEAN value.
5) If changes found in centroid value and newly calculated mean value
    I. Then make the MEAN value new centroid.
    II. Count the repetitions.
    III. Go to step 3.
6) Else stop this process.

III. CLASSIFICATION ALGORITHMS

Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known.

A. BayesNet:

BayesNet [6] learns Bayesian networks under the presumptions: nominal attributes (numeric one are pre-discredited) and no missing values (any such values are replaced globally). There are two different parts for estimating the conditional probability tables of the network. In this study we run BayesNet with the Simple Estimator and K2 search algorithm without using ADTree. K2 algorithm is a greedy search algorithm that works as follows. Suppose we know the total ordering of the nodes. Initially each node has no parents. The algorithm then incrementally adds the parent whose addition increases most of the score of the resulting structure. When no addition of a single parent can increase the score, it stops adding parents to the node. Since an ordering of the nodes is known beforehand, the search space under this constraint is much smaller than the entire space. And we do not need to check for cycles, since the total ordering guarantees that there is no cycle in the deduced structures. Furthermore, based on some appropriate assumptions, we can choose the parents for each node independently.

B. Naive Bayes Algorithm:

Bayesian Classifiers are statistical classifiers based on Bayes theorem. Bayesian classification is very simple and it shows high accuracy and speed when applied to large data bases. It works on one assumption that is the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence [1]. Bayesian classification can predict class membership probabilities, such as probability that a given tuple belongs to a particular class [1]. The Naïve Bayesian classification predicts that the tuple *X* belongs to the class *Ci*.
Using the formula
$P(Ci / X) = P(X / Ci) P(Ci) / P(X)$
Where P (*Ci* /*X*) is *maximum posteriori hypothesis* for the class *Ci*.
As *P*(*X*) is constant for all classes, only *P*(*X*/ *Ci*) *P* (*Ci*) needed to be maximized.
If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is,

$P(C1) = P(C2) = \ldots = P(Cm)$.
$P(Ci/X) = P(Xj/Ci)$.

Otherwise

$P(Ci/X) = P(X/Ci)P(Ci)$.

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*.
Given datasets with many attributes, it would be extremely computationally expensive to compute *P*(*X*/*Ci*). In order to reduce computation in evaluating *P*(*X*/*Ci*), the naive 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.
Thus, $P(X / Ci) = {}_{k=1}\Pi^{n} P(Xk / Ci)$
$= P(X1|Ci) \times P(X2|Ci) \times \ldots P(Xn|Ci)$
Probabilities *P*(*X*1/*Ci*), *P*(*X*2/*Ci*),…. are easily estimated from the training tuples. Recall that that here *Xk* refers to the value of attribute *Ak* for tuple *X* which may be categorical or continuous-valued.

C. C4.5 Algorithm:

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set *S* = S1, S2,... of already classified samples. Each sample S*i* = X1, X2,... is a vector where X1,X2,... represent attributes or features of the sample.

The training data is augmented with a vector *C* = C1, C2,.. Where C1, C2, represent the class to which each sample belongs. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision.

• All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.

• None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.

• Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

D. Back propagation Algorithm:

The Back Propagation Algorithm is a multi-layered Neural Networks for learning rules [5], credited to Rumelhart and McClelland. It produces a prescription for adjusting the initially randomized set of synaptic weights such that to maximize the difference between the neural network's output of each input fact and the output with which the given input is known (or desired) to be associated. Back propagation is a supervised learning algorithm and is mainly used by Multi-Layer- perception to change the weights connected to the net's hidden neuron layer(s).

The back propagation algorithm uses a computed output error to change the weight values in backward direction [8]. To get this net error, a forward propagation phase must have been done before. The neurons are being activated using the sigmoid activation function while propagating in forward direction.

E. Support Vector Machines (SVM) Algorithm:

A Support Vector Machine (SVM) separates the data into two categories of performing classification and constructing an N-dimensional hyper plane. These models are closely related to neural networks. In fact, this model uses a sigmoid kernel function which is equivalent to a two-layer, perception neural network.

These models are closely related to classical multilayer perception neural networks. By using a kernel function, these are an alternative training method for polynomial, radial basis function and multi-layer perception classifiers in which the weights of the network are found by solving a quadratic programming problem with linear constraints, rather than by solving a non-convex, unconstrained minimization problem as in standard neural network training.

In the SVM literature, a predictor variable which is called an attribute and a transformed attribute that is used to define the hyper plane is called a feature [7]. Here, choosing the most suitable representation can be taken as feature selection. A set of features that describes one case (i.e., a row of predictor values) is called a vector. The goal of this modelling is to find the optimal hyper plane which separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other side of the plane. The vectors near the hyper plane are the support vectors.

IV. RESULTS AND DISCUSSION

Performance of Selected classification algorithms were evaluated with The Pima Indians Diabetes Dataset and Wisconsin Diagnostic Breast Cancer (WDBC) data Set. The Pima Indians Diabetes dataset contains 768 patient records with 8 attributes as shown in Table 1. And the Wisconsin Diagnostic Breast Cancer (WDBC) dataset contains 699 patient records with 10 attributes as shown in Table 2. All attributes are numerical values.

Table 1: The Pima Indians Diabetes Dataset Attributes

S.No	Name	Type
1	Number of times pregnant	Numeric
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Numeric
3	Diastolic blood pressure (mm Hg)	Numeric
4	Triceps skin fold thickness (mm)	Numeric
5	2-Hour Serum insulin (mu U/ml)	Numeric
6	Body mass index (weight in kg / (height in m)2)	Numeric
7	Diabetes pedigree function Numeric	Numeric
8	Age (years)	Numeric

Table 2: Wisconsin Diagnostic Breast Cancer (Wdbc) Data Set

S.No	Name	Type
1	Sample code number	Numeric
2	Clump Thickness	Numeric
3	Uniformity of Cell Size	Numeric
4	Uniformity of Cell Shape	Numeric
5	Marginal Adhesion	Numeric
6	Single Epithelial Cell Size	Numeric
7	Bare Nuclei	Numeric
8	Bland Chromatin	Numeric
9	Normal Nucleoli	Numeric
10	Mitoses	Numeric

For the purpose of experimentation, we used Weka Data Mining open source machine learning software [3]. With each algorithm, we have observed Accuracy that can be defined as follows:

Accuracy: The accuracy of a classifier is the percentage of the test set tuples that are correctly classified by the classifier.

Accuracy = (Number of true positives + Number of true negatives) / (Number of true positives + false positives + false negatives + true negatives)

Here we used 10 fold cross validation in our analysis with each of the selected algorithms. That is, each dataset was divided into ten parts out of which nine parts were used as training set and the remaining part is used as testing set. Repeating these ten folds ensures that each part is used for training and testing thus minimizing the sample bias.

First the data set needs to be normalized. After normalization, the Integration technique was applied. First we have to apply K-means clustering algorithm, this divides the data set into number of clusters. The number of clusters generated is same as the number of class labels in the dataset in order to obtain a useful model. So, here we have taken the number of clusters as two. After clustering, the result needs to be saved in .arff format for applying the classification algorithm to make integration.

The performance of Bayes Net, Naïve Bayes, C 4.5, Back Propagation, and SVM Classification algorithms are analyzed with Pima Indians Diabetes dataset and Wisconsin Diagnostic Breast Cancer (WDBC) data Set. The features are ranked based on priority using the ranking algorithm available in WEKA tool. The experiments performed on the datasets gave the results as shown below in the Tables 3 & 4. These tables show the accuracy measure of the simple classification algorithms, integration of clustering and classification using Simple K-Means algorithm. From the tables, it is clear that the performance of classifiers has been improved after clustering.

Table III: Comparison of Results For Pima Indians Diabetes Dataset

Algorithm	Simple Classification Result	K-Means + Classification Result
Bayes Net	74.34	100
Naïve Bayes	76.30	99.3
Back Propagation	75.39	100
SVM	77.34	100
C 4.5	73.82	100

Table IV: Comparison of Results For Wisconsin Diagnostic Breast Cancer Data Set

Algorithm	Simple Classification Result	K-Means + Classification Result
Bayes Net	97.1	98.9
Naïve Bayes	96.0	97.6
Back Propagation	95.4	100
SVM	96.9	99.1
C 4.5	94.6	99.3

The following graphs illustrates the tabulated results shown above

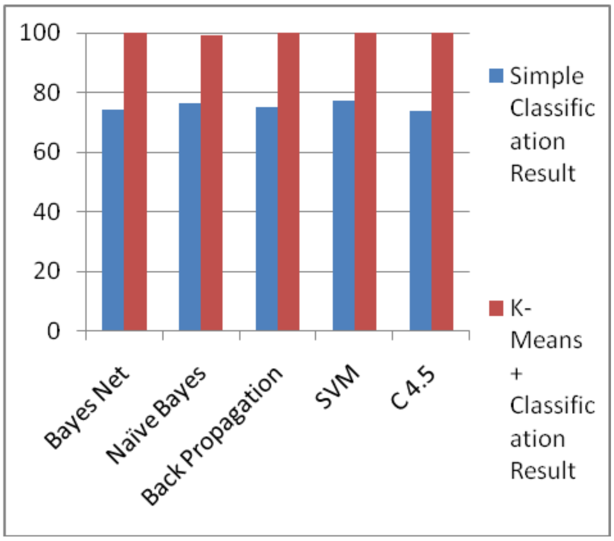

Figure 1: Classification and Integration of Clustering-Classification for Pima Indians Diabetes dataset

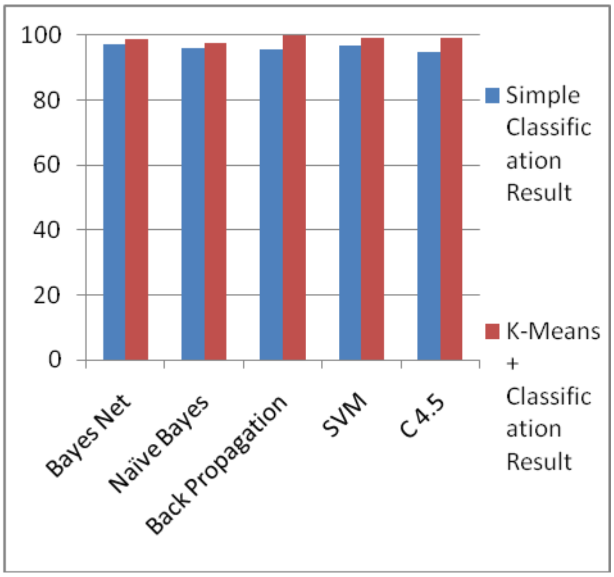

Figure 2: Classification and Integration of Clustering- Classification for Wisconsin Diagnostic Breast Cancer data set

V. CONCLUSIONS

In this paper five different classifiers are integrated with the simple k-means clustering algorithm. This integration technique was applied on "Diabetes" data set and "Breast Cancer" data set. From the observation and analysis it was concluded that the performance of simple k-means + classification is better than simple classification algorithms.

REFERENCES

[1] Jiawei Han and Micheline Kamber, *Data Mining Concepts and Techniques*, second edition, Morgan Kaufmann Publishers an imprint of Elsevier.

[2] UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets.html]. Irvine, CA: University of California, School of Information and Computer Science.

[3] WEKA, by university of Waikato, http://www.cs.waikato.ac.nz/ml/weka/

[4] http://www.scribd.com/doc/28249613/Data-Mining-Tutorial.

[5] Paul R. Harper, *A review and comparison of classification algorithms for medical decision making*.

[6] G. H. John and P. Langley, *Estimating Continuous Distributions in Bayesian Classifiers*, Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, San Francisco, 1995, pp. 338-345.

[7] Michael J. Sorich, John O. Miners, Ross A. McKinnon, David A. Winkler, Frank R.Burden, and Paul A. Smith *Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP- Glucuronosyltransferase Isoforms*.

[8] Paul R. Harper, A review and comparison of classification algorithms for decision making.

[9] Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya, *Comparison the various clustering algorithms of weka tools*, International Journal of Emerging Technology and Advanced Engineering, (ISSN 2250-2459, Volume 2, Issue 5, May 2012).

[10] Bendi Venkata Ramana, Prof. M.Surendra Prasad Babu and Prof.N. B. Venkateswarlu: *A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis*. In Proceedings of the International Journal of Database Management Systems (IJDMS), Vol.3, No.2, pages 101- 114, May 2011.

[11] Yugal kumar, Hindu College of Engineering and G. Sahoo, Birla Institute of Technology, *Analysis of Bayes Neural Network and Tree Classifier of Classification Technique in Data Mining using WEKA*.

ABOUT AUTHOR

Sambasiva Rao Voleti obtained his M.Tech degree in Computer Science and Engineering from JNTU, Hyderabad.He is pursuing Ph.D in KrishnaUniversity, Andhra Pradesh. He has more than 15 years of teaching experience. At present he is working as Vice-Principal, Kamakshi College of Engineering & Technology, SuryaPet, NalgondaDt.

Dr. Kiran Kumar Reddi obtained his Ph.D from Acharya Nagarjuna University, Guntur and M.Tech from JNTU, Kakinada. At Present he is working as HOD, Dept. of CSE, Krishna University, Machilipatnam. He has more than 17 years of teaching experience. He has more than 40 publications in various national and international journals. He is member of ISTE, IETE and CSI.