International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org, editorijaiem@gmail.com
Volume 2, Issue 1, January 2013 ISSN 2319 - 4847

A Hybrid Approach to Improve Classification with Cascading of Data Mining Tasks

D.Lavanya[1], Dr.K.Usha Rani[2]

[1] Associate Professor, Department of Computer Science and Engineering, Rayalaseema school of Engineering, Research and Management, Tirupati, Andhra Pradesh, India
[2] Associate Professor, Department of Computer Science, Sri Padmavathi Mahila Visvavidyalayam, Tirupati, Andhra Pradesh, India

ABSTRACT

Data mining plays an important role to find the interesting patterns from databases. Medical data mining is very much useful to medical practitioners. To diagnose the patient's disease Classification, one of the data mining tasks, plays a significant role. Cascading classification with some other data mining tasks improves classification accuracy. In this study a hybrid approach which is a combination of CART decision tree classifier with clustering and feature selection has been proposed on breast cancer data sets. The effectiveness of hybrid approach has been compared against CART with Feature Selection, Classification with Clustering and without Feature Selection in terms of accuracy.

Keywords: Data Mining, Classification, Clustering, Feature Selection, K-Means, Breast Cancer.

1. INTRODUCTION

Data Mining is the discovery of previously unknown, potentially useful and hidden knowledge in databases. The various Data mining tasks are:

- Classification
- Regression
- Clustering
- Rule generation
- Discovering Association rules
- Summarization
- Dependency modeling
- Sequence analysis.

Classification is a supervised Machine Learning technique which assigns labels or classes to different objects or groups. Classification is a two step process: In the first step, through the analysis of the training records of a database a model is constructed. In the second step, the constructed model is used for classification. The classification accuracy is estimated by the percentage of test samples or records that are correctly classified.

The classification task can be carried out by various techniques such as: Decision Tress, Bayesian classify and Bayesian networks (Belief Networks), Neural Networks, Rule induction, K-nearest neighbor, Genetic algorithms, Rough sets, Fuzzy logic and so on. By merging some classification techniques new techniques also have been developed (ex: Fuzzy rule induction, Fuzzy decision trees, Neuro-fuzzy networks, etc.). The choice of the best technique to a specific problem can be decided by experimenting many possibilities based on the measures such as accuracy, speed, robustness, scalability and interpretability. Classification is extensively used in various application domains such as retail target marketing, fraud detection, design of telecommunication service plan, medical diagnosis, etc.[6][9].

Clustering is a Data mining technique which segments a heterogeneous data into a number of homogeneous subgroups or clusters. In clustering, the records are grouped together based on self-similarity without any predefined classes or examples. Similarity is nothing but how close the objects are in space based on a distance function. The quality of a cluster is measured by cluster diameter which is the distance between any two objects in a cluster. An alternative measure of cluster quality is centroid distance, which is the average distance of each cluster object from the cluster centroid (i.e., average object or average point in space of cluster). The effectiveness of this technique depends on the nature of the data. Once the proper clusters have been defined it is easy to find simple patterns with each cluster.

Clustering is often used as preliminary step of some other form of data mining models to enhance their accuracy or performance. It is also known as unsupervised learning or learning by observation.

Clustering is very much required in many areas such as Statistics, Biology and Machine Learning etc. Out of many clustering algorithms the major clustering methods can be classified as: Partitioning methods, Hierarchical methods, Density based methods, Grid based methods, and Model based methods. The choice of the clustering algorithm depends both on available data and on the particular purpose of the application. In partitioning methods, the most popular and well known algorithms are k-Means and k-Medoids. In this study k-means clustering is chosen because of its popularity and it's proved effectiveness in the literature [5].

The organization of the paper is: A brief overview of related work, the theory of decision tree, feature selection and K-Means Algorithm is presented in section 2. The Section 3 presents Experimental Results and section 4 concludes the study.

2. BACKGROUND

2.1 Overview of Related Work

Shekhar R.Gaddam et al.[19] performed a work on cascading of clustering and classification to classify the activities in a computer network, an active electronic circuit and a mechanical mass beam system. K-Means algorithm of Clustering and Decision Tree algorithm ID3 was used.

Chin-Yuan Fan et al. [7] proposed a hybrid model by integrating a Case based Data Clustering method and a Fuzzy Decision Tree to classify the liver disorder and Breast cancer datasets.

Themis P.Exarchos et al. [20] proposed a method for detection and classification of transient events in Electroencephalographic recordings. The method implemented here is Association Rule Mining and Classification.

Asha Gowda Karegowda et al. [5] performed a work on classification Tuberculosis data with cascading clustering and classification. K-Means algorithm for clustering and for classification was used.

P. Rajendran et al. [16] proposed a method to classify the brain tumor in the CT scan brain images. CT scan brain images were preprocessed using median filtering process and features are extracted using canny edge detection technique. Frequent patterns from the CT scan images were generated using frequent pattern tree algorithm. The decision tree algorithm was used to classify the medical images for diagnosis as normal, benign and malignant. This proposed method proved more accurate than a conventional method.

Asha.T et al. [4] proposed a hybrid model to classify the diabetic patients data. Hybrid model encompasses k-means Clustering, k-nearest neighbor classification and correlation feature selection.

2.2 Decision Trees

Decision tree, is one of the classification method popularly used for classifying the data. These are popular because to construct decision tree parameter setting or domain knowledge is not required. The structure of decision tree is similar to a tree. A decision tree [10] is induced in two steps:

- Build phase
- Prune phase

Build phase: To construct a tree attributes (nodes) has to be known. This Attribute selection is done based on measures such as Information Gain, Gain Ratio and Gini Index etc. These measures are useful to choose a best attribute that differentiates the tuples belonging to a class. The best attributes number of distinct values is drawn as arcs from the node and each distinct value is written on arcs. Further to decide the best splitting attribute the above is continued until all the records are belonging to a same class. The class nodes are called as leaf nodes.

Prune phase: it is used to avoid anomalies (overfitting, noise) in the tree. Prepruning, Post pruning, cost complexity pruning are the methods to prune the tree. Thus an optimized tree is induced. CART decision tree classifier uses Gini Index attribute selection measure and cost complexity pruning. Decision tree classifiers are used extensively for medical diagnosis such as Breast tumour in ultrasonic images, Ovarian cancer, Heart sound diagnosis and so on [2], [11].

2.3 Feature Selection

Feature selection (FS) is a process of identifying the subsets of attributes which are most significant or relevant for the data mining task. Advantages of Feature Selection:

- Reduces the size of dataset
- Reduces the computational time
- Reduces the feature measuring cost

Several Feature Selection methods are available in the literature [15]. A feature selection method is a combination of searching algorithm and evaluation measure. Searching algorithm generates subsets of attributes. Evaluation measures are used to evaluate subsets of attributes. Evaluation measures are distance, information, dependency, consistency and classifier error rate. The value produced by the evaluation measure is used to test whether the generated subsets are optimal or not. Feature Selection improves the accuracy of classification. To classify the medical data accurately feature selection can be embedded in to the classification task. A medical Practitioner normally considers the dominant features to diagnose a patient's disease rather than entire features. Feature Selection with decision tree classification greatly enhances the quality of the data in medical diagnosis [3],[1], [12], [8], [15].

2.4 K-means clustering

The k-means algorithm is one of the most commonly used clustering algorithms. It is a centroid based technique. In k-means algorithm, k is an input parameter based on which the set of n objects are clustered into k groups such that similarity in intracluster is high. The working of the k-means is given in three steps:

The algorithm randomly selects k data points to be the initial seeds of clusters.

Assign each record to the closest seed.

The third step is to calculate the centroids of the clusters which efficiently cluster the objects than initial seeds.

2.5 Breast Cancer

Breast Cancer is the leading cause of death in women in developing countries and a second cause in developed countries as per the statistics of National Cancer institute. Breast Cancer is a malignant tumor which grows from the cells of the breast. This can occur in both male and female. But the occurrence is high in female. The exact causes of breast cancer are not known. Some of the risk factors in female are: Ageing, Family history, Genetic risk factors, Menstrual periods, Obesity, Not having children, etc., [22], [17], [18].

Early diagnosis and treatment will reduce breast cancer deaths. The patients can be predicted as Benign group (non cancerous) or Malignant group (cancerous). And another group is prognosis in which patients are those whose cancer has been surgically removed. Such group of patients is observed for recurrence of the disease. The accurate and reliable diagnosis is required to distinguish between benign and malignant tumors. Most frequently adopted methods for early detection breast cancer are self-examination, mammography, sonography, and biopsy.

Data Mining has become a popular technology with different technical approaches. An accurate predictive model developed with an appropriate learning method is highly recommendable. In this context, a popular and commonly used Decision Tree model is proved best [2], [11] in the literature for classification to predict the belongingness of a new case to a particular group.

3. EXPERIMENTAL RESULTS

Three Breast Cancer datasets, which are publicly available from UCI Machine Learning Repository [21], are experimented using the decision tree classifier CART. The description of Data sets is presented in the Table 1.

Table 1: Description of Datasets

Dataset	No. of Attributes	No. of Instances	No. of Classes	Missing values
Breast Cancer	10	286	2	yes
Breast Cancer Wisconsin (Original)	11	699	2	yes
Breast Cancer Wisconsin (Diagnostic)	32	569	2	no

CART was proved as the best classifier among the three decision tree classifiers ID3, C4.5 and CART in the medical domain [13]. Datasets contain attributes that are irrelevant to the data mining task. These attributes can be removed by applying feature selection methods. Feature selection methods improve the accuracy of classification. Several Feature Selection methods were experimented on the Breast Cancer datasets and inferred that a particular feature selection method is not the best one for all data sets. The best feature selection methods for the datasets were obtained in our study [14]. CART with Feature Selection enhanced the classification accuracy than CART alone.

To improve classification performance various Data Mining techniques are combined. Only few studies are available related to various applications and very few on medical applications in the literature. It is observed that clustering and classification applied on some applications such as Computer Network activities, Liver disorder diseases & Breast Cancer data, Tuberculosis, EEG and so on. In those studies only cascading clustering and classification was applied.

In this section, a new hybrid method is proposed to enhance the accuracy of classifier CART with cascading Feature Selection and Clustering. For clustering purpose, k-Means algorithm is used because of its popularity.
Proposed Algorithm: Cascading Feature Selection, Clustering and Classification.

1. Apply preprocessing techniques to eliminate missing values in the data.
2. Select significant attributes by using best Feature Selection method to a specific Breast Cancer Data set.
3. Cluster the data set using K-Means algorithm.
4. Train the classifier CART by taking the clusters formed in step3.

The process diagram of the proposed hybrid approach is shown in the figure 1.

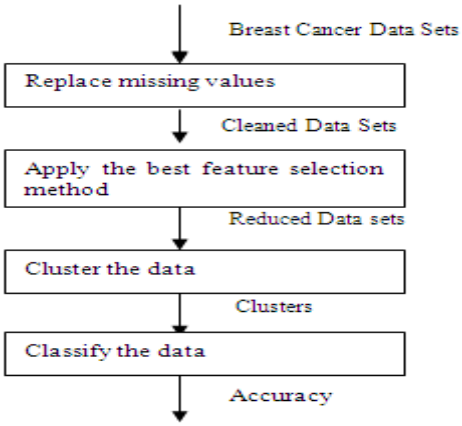

Figure 1. Process diagram of hybrid approach

The algorithm is implemented on the breast cancer datasets, the results are tabulated in Table 2 and the accuracy is compared with CART with FS.

Table2: Accuracy (%) of CART with FS and Hybrid Approach

Data Set	CART with FS	**Hybrid Approach**
Breast Cancer	**73.07**	**100**
Breast Cancer Wisconsin (Original)	**96.99**	**98.71**
Breast Cancer Wisconsin (Diagnostic)	**94.72**	**98.06**

Table 2 depicts that hybrid approach has better accuracy rates than CART with FS**.** Accuracy of CART with FS and hybrid approach is shown graphically in figure 2.

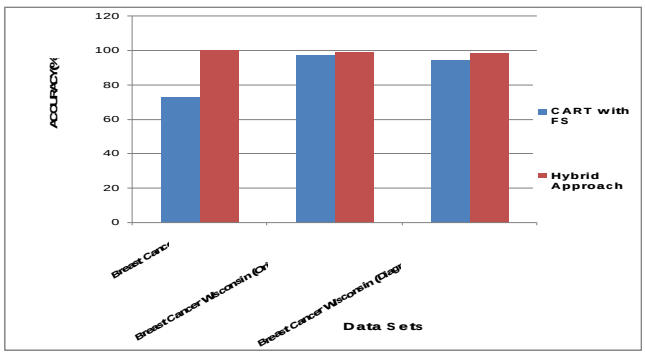

Figure 2. Accuracy of Hybrid Approach and CART with FS.

Further Accuracy of hybrid approach is compared against Cascading Clustering and Classification approach. The results are tabulated in Table 3.

Table 3: Accuracy (%) of Hybrid Approach and Cascading Clustering and Classification

Data Set	Hybrid Approach	Cascading Clustering and Classification
Breast-Cancer	**100**	89.51
Breast-Cancer Wisconsin(Original)	**98.71**	96.13
Breast-Cancer Wisconsin(Diagnostic)	**98.06**	94.90

From these observations it is very clear that the proposed hybrid approach (new hybrid algorithm) has outperformed. Accuracy of hybrid approach and cascading of clustering and classification is presented graphically in the figure 3.

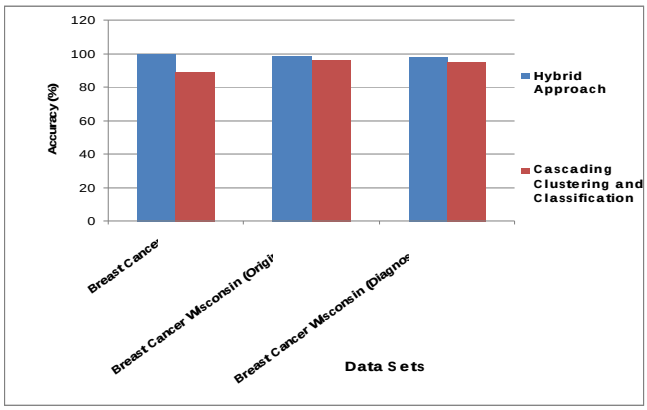

Figure 3. Accuracy of Hybrid Approach and Cascading Clustering and Classification.

From the above comparison, it is recommended that the new Hybrid algorithm is the best classifier for the breast cancer data sets.

4. CONCLUSION

In this study, it is observed that combination of data mining tasks achieved better accuracy rather than a single data mining task. To classify the breast cancer data an effective Hybrid Approach has been proposed. The Hybrid Approach–cascading feature selection, clustering and classification improve the classification accuracy. Experimental results demonstrate that the hybrid approach is better than CART with FS and cascading of classification and clustering without FS.

REFERENCES

[1] Aboul Ella Hassaneian, "Classification and Feature Selection of Breast Cancer Data based on Decision Tree Algorithm", Studies and Informatics Control, vol12, no1, March 2003.

[2] Antonia Vlahou, John O. Schorge, Betsy W.Gregory and Robert L.Coleman,"Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data", Journal of Biomedicine and Biotechnology • 2003:5(2003) 308–314.

[3] Asha Gowda Karegowda, M.A.Jayaram and A.S. Manjunath, "Feature Subset Selection Problem using Wrapper Approach in Supervised Learning", International Journal of Computer Applications 1(7):13–17, February 2010.

[4] Asha.T, S.Natarajan and K.N.B.Murthy, "A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification", 2011.

[5] Asha Gowda Karegowda , M.A. Jayaram, A.S. Manjunath," Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients", International Journal of Engineering and Advanced Technology (IJEAT)ISSN: 2249 – 8958, Volume-1, Issue-3, February 2012.
[6] R. Brachman, T. Khabaza, W.Kloesgan, G.Piatetsky Shapiro and E. Simoudis, "Mining Business Databases",Comm. ACM, Vol. 39, no. 11, pp. 42-48, 1996.
[7] Chin-Yuan Fan, Pei-Chann Chang , Jyun-Jie Lin and J.C. Hsieh, "A Hybrid model combining Case-based reasoning and Fuzzy Decision Tree for Medical Data Classification", Applied Soft Computing, vol.11, issue 1, pp. 632–644, 2011.
[8] Deisy.C, Subbulakshmi.B, Baskar.S and Ramaraj.N, "Efficient Dimen-sionality Reduction Approaches for Feature Selection, Conference on Computational Intelligence and Multimedia Applications", 2007.
[9] U.M. Fayyad, G. Piatetsky Shapiro and P. Smyth, "From Data Mining to knowledge Discovery in Databases", AIMagazine, vol 17, pp. 37-54, 1996.
[10] J. Han and M. Kamber, "Data Mining; Concepts and Technques", Morgan Kaufmann Publishers, 2001.
[11] Kuowj, Chang RF, Chen DR and Lee CC," Data Mining with Decision Trees for Diagnosis of Breast Tumor in Medical Ultrasonic Images", March 2001.
[12] Kemal Polat, Seral Sahan, Halife Kodaz and Salih Günes, "A New Classification Method for Breast Cancer Diagnosis: Feature Selection Artificial Immune Recognition System (FS-AIRS)", In Proceedings of ICNC (2)'2005. pp.830~838.
[13] D.Lavanya, Dr.K.Usha Rani, "Performance Evaluation of Decision Tree Classifiers on Medical Datasets". International Journal of Computer Applications 26(4):1-4, July 2011.
[14] D.Lavanya, Dr.K.Usha Rani,.," Analysis of feature selection with classification: Breast cancer datasets",Indian Journal of Computer Science and Engineering (IJCSE),October 2011.
[15] Mark A. Hall and Lloyd A. Smith, "Feature Subset Selection: A Correlation Based Filter Approach", In 1997 International Conference on Neural Information Processing and Intelligent Information Systems (1997), pp. 855-858.
[16] P. Rajendran and M.Madheswaran, "Hybrid Medical Image Classification Using Association Rule Mining with Decision Tree Algorithm", Journal of Computing, volume 2, issue 1, January 2010, ISSN 2151-9617.
[17] Shelly Gupta, Dharminder Kumar, Anand Sharma, " Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis", Indian Jouranl of Computer science and Engineering, Vol. 2 No. 2 Apr-May, 2011.
[18] EI-Sebakhy A. Emad, Faisal Abed Kanaan, Helmy T, Azzedin F and AI-Sushaim F, " Evaluation of Breast Cancer tumor classification with unconstrained functional networks classifier", Computer Systems and Applications, IEEE, International Conference, 2006, pp. 281-287.
[19] Shekhar R.Gaddam, Vir V. Phoha and Kiran S. Balagani, "KMeans+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-Means Clustering and ID3 Decision Tree Learning Methods", IEEE Transactions on Knowledge and Data Engineering.
[20] Themis P. Exarchos, Alexandros T. Tzallas, Dimitrios I. Fotiadis, Spiros Konitsiotis, and Sotirios Giannopoulos, "EEG Transient Event Detection and Classification Using Association Rules", IEEE Transactions on Information Technology in Biomedicine, vol. 10, no. 3, july 2006.
[21] [UCIMLP] UCIrvine Machine Learning Repository www.ics.uci.edu/~mlearn/ MLRepository.html
[22] Wen-Jai Kuo, Ruey-Feng Chag, Dar-Ren Chen and Cheng Chun Lee," Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images", Breast Cancer Research and Treatment 66: 51-57, 2001, Kulwer Academic Publishers, Printed in the Netherlands.