Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 1, Number 2; November, 2014 pp. 82-85
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/acsit.html

Hybrid Algorithm of Clustering and Classification in data Mining: A Survey

KanchanJha[1] and Divakar Singh[2]

[1,2]BUIT BU Bhopal, (MP)
E-mail: [1]kkanchan2007jha@gmail.com, [2]divakar_singh@rediffmail.com

Abstract: *Now day's data mining is used in every field. Data mining is the knowledge extraction from large data, there are various techniques in data mining in which Clustering and classification is very important, Clustering is a data mining technique used to place data elements into related groups without advance knowledge of the group definitions. Popular clustering techniques include k-means clustering c-means, x-means clustering etc. Clustering method will use for make the clusters of similar groups to extract the features or properties. Classification is a data mining technique used to predict group membership for data instances. Popular classification techniques include decision trees and neural networks etc. Decision tree method will use for choose to decide the optimal decision to extract the valuable information. In this survey paper, we study about clustering and classification technique to mine the data by combining both techniques.*

1. INTRODUCTION

Data mining is the important step for discover the knowledge in knowledge discovery process in data set. Data mining provide us useful pattern or model to discovering important and useful data from whole database[2].

To mine the data Classification use to classify the data items into the predefined classes and find the model to analysis. Regression identifies real valued variables. Clustering use to describe the data and categories into similar objects in groups. Find the dependencies Between variables. Mine the data using tools. Clustering and classification are two of the mostly used methods of data mining which provide us much more convenience in researching information data [3].

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A good clustering algorithm is able to identity clusters irrespective of their shapes. Other requirements of clustering algorithms are scalability, ability to deal with noisy data, insensitivity to the order of input records, etc. Data mining is a multi-step process. It requires accessing and preparing data for a data mining algorithm, mining the data, analyzing results and taking appropriate action. The accessed data can be stored in one or more operational databases, a data warehouse or a flat file. In data mining the data is mined using two learning approaches i.e. supervised learning or unsupervised clustering.

Supervised Learning

In this training data includes both the input and the desired results. These methods are fast and accurate. The correct results are known and are given in inputs to the model during the learning process. Supervised models are neural network, Multilayer Perceptron, Decision trees.

Unsupervised Learning

The model is not provided with the correct results during the training. It can be used to cluster the input data in classes on the basis of their statistical properties only. Unsupervised models are different types of clustering, distances and normalization, k-means, self organizingmaps[4].

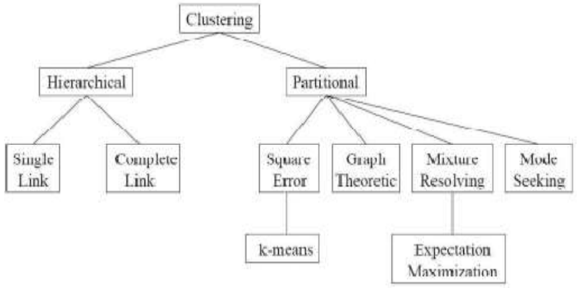

Fig. 1: Clustering

Classification is an important task in data Mining. It belongs to directed learning and the main methods include decision tree, neural network and genetic algorithm. Decision tree build its optimal tree model by selecting important association features. While selection of test attribute and partition of sample sets are two parts in building trees. Different decision tree methods will adopt different technologies to settle these problems. Algorithms include ID3, C4.5, CART and SPRINT etc.

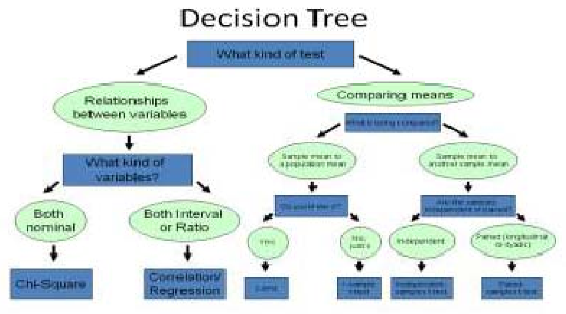

Fig. 2: Decision Tree

Hybrid techniques of clustering and classification method for large dimensional dataset. Clustering analysis is an important and popular data analysis technique that is large variety of fields**.** Clustering and classification are the mostly used methods of data mining. Clustering can be used for describing and decision tree can be applied to analyzing [5].

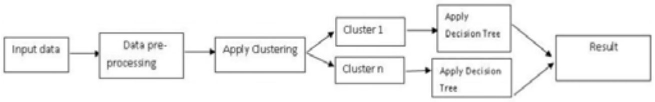

Fig. 3: Hybrid model

2. LITERATURE REVIEW

Dr. Ela Kumar, Arun Solanki proposed a knowledge management platform which is used for data mining process Whenever a new user run into trouble in data mining process, he/she can ask questions to the platform, and then get automatically answers from the knowledge base [2].

Heena Sharma, Navdeep Kaur Kaler compare the effectiveness of two stage clustering and decision tree data mining algorithms by applying them to data sets. Experiment results will show that like two stage algorithms produce the best classification accuracy, and also show the higher robustness and generalization ability compared to the other traditional algorithms. Their approach can remove the shortcoming of hybridization of algorithms and improve the results on applying them to data sets. Their approach gives effective results, better performance and reduces the error rate than the traditional algorithms of clustering and classification in data mining [3].

LeventBolelli, Seyda Ertekin1, Ding Zhou, C. Lee Giles report results on dataset for two clustering criterion functions of K-Means, averaged over ten runs. The first clustering algorithm is the Euclidean K-Means that makes the cluster assignment decisions based on the euclidean distances between the document vectors. The second algorithm they used is the Spherical K-Means that uses the cosine distances between documents as the similarity metric. For both clusterings, they experimented with two separate initialization schemes. In the first scheme, each document is assigned a random cluster ID upon the initialization of K-Means. The second scheme chooses one of the cluster centroids as the farthest point from the center of the whole data set, and all cluster centroids are well separated. [4].

N. Sivaram,K. Ramar, use data sets and the input attributes to determine through knowledge engineering in an IT industry. The process involves defining the problem, identifying relevant stake holders, and learns about current solutions to the problem [5].

NorulHidayah Ibrahim, Aida Mustapha, RozilahRosli, NurdhiyaHazwaniHelmee, apply hybrid classification model refers to a combination of two data mining tasks, which are clustering and classification in effort to obtain higher accuracy result. Previously, hybrid classification models have been applied to predict patients who have higher risk in having diabetes by looking at the patients' profiles. This information is useful to help overcome a predictable diabetic patient[7].

D.Lavanya, Dr.K.Usha Rani apply a hybrid approachwhich 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, Classificationwith Clustering and without Feature Selection in terms of accuracy[8].

Md. Hedayetul Islam ShovonMahfuzaHaque apply K-means clustering algorithm on the training data so that they can group the students in three classes "High" "Medium" and "Low" according to their new grade. New grade is calculated from the previous semester grade that means external assessment and internal assessment. Then apply decision tree to make correct decision about the student which is need to take by the instructor[9].

Varun Kumar, NishaRathee experimented and equated the outcome of a simple classification method (J48 classification) with the outcome of integrated clustering and classification method . As a result they found, the integration of clustering and classification techniques gives more accurate and robust result than applying either of them alone.

After clustering they have applied the classification method to assign the attributes to these clustered classes, because at the time of clustering decision rules are obtained, which are very useful in classification [10].

Table 1: Comparative study

	Title	Association	Clustering	Classification	Hybrid	Remark
1	"Data Mining with Improved and Efficient Mechanism in Clustering Analysis and Decision Tree as a Hybrid Approach" [2]	No	yes	yes	Stage 1:Clustering(K-Means, SOM,HAC) Stage 2: Classification(CHAID, C4.5)	Experiment results will show that the best accuracy, and also show the higher robustness and generalization ability compared to the other algorithms.
2	"K-SVMeans: A Hybrid Clustering Algorithm for Multi-Type Interrelated Datasets"[3]	No	Yes	yes	Stage 1:K-Means Stage2: SVM	Traditional clustering algorithms do not handle rich structured data, it works only on homogeneous data type. K-SVMeanshandle heterogeneous
3	"Applicability of Clustering and Classification Algorithms for Recruitment Data Mining"[4]	No	Yes	Yes	Stage 1:Decision tree Stage 2: clustering	
4	"Integrating human knowledge within a hybrid clustering-classification scheme for detecting patterns within large movement data sets"[5]	No	Yes	Yes	Stage 1:Clustering Stage 2:Classification	
5	"A Hybrid Model of Hierarchical Clustering and Decision Tree for Rule-based Classification of Diabetic Patients"[6]	No	Yes	Yes	Stage 1:clustering Stage 2:Decision tree	
6	"A Hybrid Approach to Improve Classification with Cascading of Data Mining Tasks"[7]	No	Yes	Yes	Stage 1:Clustering Stage 2:Classification	hybrid approach is better than CART with FS and cascading of classification and clustering Without FS.
7	"An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree"[8]	No	Yes	Yes	Stage 1:clustering Stage 2:Decision tree	
8	"Knowledge discovery from database Using an integration of clustering and Classification"[9]	No	Yes	Yes	Stage1:Classification Stage 2: clustering	

3. CONCLUSION

In this study paper we read various hybrid algorithm of data mining and I conclude that hybrid approach of clustering and classification may be the best algorithm for the property tax system, which may be helpful to reduce the problem of property tax gap and raise the property tax revenue.

RFERENCES

[1] RichaDhiman, ShevetaVashisht, KapilSharma"A CLUSTER ANALYSIS AND DECISION TREE HYBRID APPROACH IN DATA MINING TO DESCRIBING TAX AUDIT" International Journal of Computers & Technology Volume 4 No. 1, Jan-Feb, 2013 ISSN 2277-3061

[2] Dr. Ela Kumar, Arun Solanki "A Combined Mining Approach and Application in Tax Administration."International Journal of Engineering and Technology Vol.2(2), 2010.

[3] Heena Sharma, Navdeep Kaur Kaler , "Data Mining with Improved and Efficient Mechanism in Clustering Analysis and Decision Tree as a Hybrid Approach" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-5, April 2013.

[4] LeventBolelli, Seyda Ertekin1, Ding Zhou, C. Lee Giles, "K-SVMeans: A Hybrid Clustering Algorithm for Multi-Type Interrelated Datasets."EEE/WIC/ACM International Conference on Web Intelligence 2007.

[5] N. Sivaram ,K. Ramar, "Applicability of Clustering and Classification Algorithms for Recruitment Data Mining"International Journal of Computer Applications (0975–8887) Volume 4–No.5, July 2010.

[6] René Enguehard,, Benjamin Fowler , Orland Hoeber, RodolpheDevillers, Wolfgang Banzhaf "Integrating human knowledge within a hybrid clustering-classification scheme for detecting patterns within large movement data sets". AGILE 2012–Avignon, April 24-27, 2012.

[7] NorulHidayah Ibrahim, Aida Mustapha, RozilahRosli, NurdhiyaHazwaniHelmee , "A Hybrid Model of Hierarchical Clustering and Decision Tree for Rule-based Classification of Diabetic Patients"NorulHidayah Ibrahim et.al / International Journal of Engineering and Technology (IJET) Vol 5 No 5 Oct-Nov 2013.

[8] D.Lavanya, Dr.K.Usha Rani, "A Hybrid Approach to Improve Classification with Cascading of Data Mining Tasks" International Journal of Application or Innovation in Engineering & Management (IJAIEM) Volume 2, Issue 1, January 2013.

[9] Md. Hedayetul Islam ShovonMahfuzaHaque, "An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol.3, No. 8, 2012.

[10] Varun Kumar, NishaRathee, "Knowledge discovery from database Using an integration of clustering and classification" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, March 2011.