



International Journal of Engineering, Business and Enterprise Applications (IJEBA)

www.iasir.net

A Review of Clustering and Classification Techniques in Data Mining

Samir Kumar Sarangi

Department of Computer Science & Engineering,
Amity School of Engineering and Technology,
Amity University Haryana, Gurgaon (Manesar) – 122 413
Haryana, India

Dr. Vivek Jaglan

Department of Computer Science & Engineering,
Amity School of Engineering & Technology,
Amity University, Haryana, Gurgaon (Manesar) – 122 413
Haryana, India

Yajnaseni Dash

Department of Computer Science & Engineering, Amity
School of Engineering & Technology,
Amity University, Sec-125, Noida, U.P., 201301, India

Abstract: Data mining is the process of extracting hidden analytical information from large databases using multiple algorithms and techniques. This technology allows companies to focus on the most important information in their data warehouses. Generally organizations collect and process huge amount of data. Data mining techniques can be applied rapidly on existing software and hardware platforms to increase the value of existing information resources, and can be integrated with new products and systems. Previously data analysis process often involved manual work and during which interpretation of data was slow, costly, and highly immanent. Data Mining, generally named as 'knowledge discovery in large data', which facilitate firms and organizations to take decisions by analyzing and accessing corporate data. Varieties of tools are available for data mining process, each with their own strengths and weaknesses. People try to solve many important data mining problems using machine learning techniques for real world applications such as business, social, healthcare, predictive toxicology, and bio, chemo-informatics etc. Machine learning techniques like Instance-based Learning algorithms, Decision Trees, Genetic algorithm, Rule Induction, Artificial Neural Networks, and Fuzzy logic are used over some open datasets. The data mining tools gather the data, while the machine learning techniques are used for taking decisions based on the data collected. There are two main methods of Data mining: Clustering and Classification. In many cases the concept of classification is confused by means of clustering, but there is difference between these two methods. According to the perspective of Machine learning clustering method is unsupervised learning and tries to group sets of objects having relationship between them, whereas classification method is supervised and assigning objects to sets of predefined classes. This paper focuses on an ongoing development and research activities of classification and clustering techniques for data mining and provides a review of machine learning algorithms used in data mining.

Keywords: data mining; machine learning; knowledge discovery in databases, clustering, classification, J48, WEKA

I. INTRODUCTION

Although data mining name is introduced in very late but the concept and technology is known since long time. Initially this concept was used manually in an organization but now we use these concepts using super computers over massive database tables, stored in flat files, spreadsheets, or some other storage format.

Data mining is a mechanism to find appropriate patterns in a database, using distinct approach and algorithms to look into current and past data that can be analyzed to predict future behaviours. Because data mining tools predict future veers and behaviours by reading through databases for hidden patterns, they allow organizations to make positive, knowledge-driven decisions and respond question that were previously too time-consuming to resolve [1]. Mainly the Data Mining conceptions are used for the observation of any collected information, evaluate their behaviour, and also find out unknown relationship between critical data in a variety of ways [2]. There are different types of Data Mining technology available and applied over the process for retrieving information from raw data and these technologies are used for different purposes like for security, marketing, and information gathering. This Data mining technology is used for verifying sample data, analyze their paths, to verify and validate their modules for specific general and business utilization. For example an insurance industry uses these techniques for finding out their rate of risks for their customers. In general application of

data mining tools are in the area of marketing, fraud fortifications, and observations [3]. Different types of algorithms and tools are used for data evaluation and they are also providing different results depending on specific algorithms [4]. The purpose of data mining is to discover valid novel, possibly helpful and understandable relationships and patterns in the existing data [5]. Data mining is to extract useful information; this is also known with different names like knowledge extraction, information discovery, information harvesting, and data pattern processing [6]. The name “data mining” is mainly used by database researchers, statisticians, and business communities. Knowledge Discovery in Databases (KDD) is usually used to denote the process of finding out useful information from database, where data mining gives the detailed information in this process [6,7,8]. The processes in the KDD, such as data grounding, data collection, data clean-up, and proper understanding of the outcome of the data mining method, make sure that useful information is resulting from the data [9]. The KDD have different steps [10].

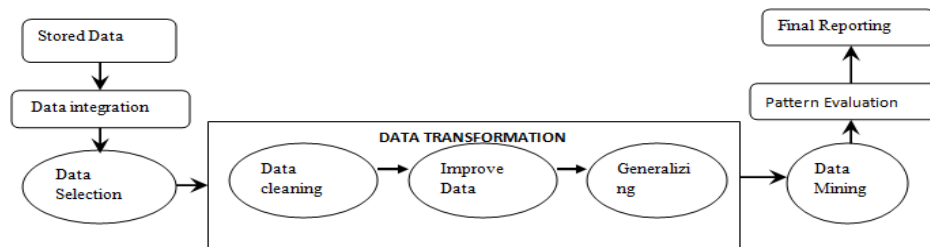


Figure 1: The KDD Process

- Stores data: Where data are stored in any data repository.
- Data integration: Multiple or heterogeneous data sources are integrated into a single unit.
- Data selection: Data retrieve from the database as needed for the KDD method.
- Data transformation: It is a process of data normalization where data are transformed and joined together into a form that is appropriate for mining process. Sub stages of this data transformation are,
 - Data cleaning: It handles noisy, erroneous and irrelevant data.
 - Improve data: Improve quality of data by adding new information, missing values to available data.
 - Generalizing data: Applying operations on data in order to prepare for a machine learning approach.
- Data mining: This is a very important step in mining process. Here intelligent methods are applied in order to extract data patterns.
- Pattern evaluation: This process is to identify the truly interesting patterns that are presented in knowledgebase.
- Knowledge presentation: It is final reporting of KDD process. Where visualization and knowledge representation techniques are used to represent the mined knowledge to the user.

Classification: Classifying data by using classification techniques in data mining is a very distinctive task. There are different types of classification methods are used in data mining such as bayes, functions, rules, trees etc. The main aim of classification is to properly calculate the value of each class variables [16]. This classification process is divided into two steps. The first step is to build the model from the training set, i.e. randomly samples are selected from the data set. In the second step the data values are assigned to the model and verify the model's accuracy [17].

Clustering: Clustering is the most significant task of data mining. It is an unsupervised method of machine learning application. Like classification where classes are previously defined, in clustering the classes are divided according to class variable. Numbers of popular clustering algorithms are available such as K-Means, DBSCAN, EM, Hierarchical clusterer, farthest first, Cobweb, OPTICS etc. K-Means algorithm is simple and performance is better than other [18].

II. LITERATURE SURVEY

According to analyst the process of data mining goes through several cycles. Data mining process is getting the paramount result and new important information from a huge searching on databases. There are two main data mining goals, predictive and descriptive. Predictive model is to retrieve information from the known system model and descriptive model is to retrieve the new, remarkable information. Data mining can be done through different methods like statistical method, machine learning, etc. The statistical methods are mathematical

models, where as machine learning is an algorithm. Statistical model gives emphasis to structure of data and machine learning is to analyze and manage collection of data [15]. Data mining provides some major properties:

- ☐ Extract, transform, and load transaction data onto the data warehouse system.
- ☐ Store and manage the data in a multidimensional database system.
- ☐ Provides facility for data access by business analysts and information technology professionals.
- ☐ Analyze the data & Present the data in a useful format, such as a graph or table by application software.

Business organization has very large data bases, so it is important to apply the data mining processes to have accurate maintenance of those data. Different data mining techniques are applied for extracting core information from different databases. There are varieties of classification algorithms, which represent problems in different way. It looks very difficult to plan and build up a data mining model which can work for every field [14]. A single classification model cannot be capable of making fine distinctions over the data sets. Multiple classification models are combined which specify that a hybrid intelligent system or multiple classifier system provides accurate result than a single classification algorithm applied [13]. The multiple classifier system is having better performance for any specific application in data mining. It has been seen in another study that a specific characteristic of a data set would be preferred for improving the models correctness [11]. According to Varun and Nisha the integration of clustering and classification techniques gives more accurate and robust result than applying either of them alone. In their study, they have shown that clustering is an unsupervised learning method and it creates classes by partitioning number of clusters according to their instances. 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 [12]. Analyst discussed about different machine learning algorithms that are applied for data mining. They classified these machine learning methods as “symbolic and sub-symbolic” for data mining process. Decision trees, case-based reasoning, instance-Based learning’s are under symbolic classification; and neural network, fuzzy rules, and genetic algorithms are under sub-symbolic classification. In their study they suggested a hybrid approach i.e. combination of both case-based reasoning (symbolic) and neural network (sub-symbolic) and applied them for data mining [10].

S.No.	Authors and Title Name	Explanation and Conclusion
1	Wu Yuntian Based on Machine Learning of Data Mining to Further Explore	This paper emphasizes on data gathered, use data information, development data information resources also discussed the role of machine learning in use of data mining. SQL query is trying to retrieve the hidden and not apparent information. This query language is helpful when it is known which information need to retrieve, but data mining is useful when knowledge is not clear about the core part. Therefore both these applications are complementary for any data repository.
2	Neelamadhab Padhy, Dr. Pragnyan Mishra, Rasmita Panigrahi The Survey of Data Mining Applications And Feature Scope	According to authors there are varieties of methods, approaches which are applied for data mining process in different fields. Each application have their own properties which are very important and useful for do research. In their research they focussed on the scope and benefits of number of applications. The domain specific methods are distinguished as special technique for data mining, because these techniques are intended for particular purposes and to produce the precise information.
3	Areej Shhab, Gongde Guo and Daniel Neagu A Study on Applications of Machine Learning Techniques in Data Mining	They discussed about the machine learning techniques for data mining problems. In this paper they applied different artificial intelligence techniques on toxicity prediction data sets. This paper specifies that multiple classification methods are valuable for any certain application. Particularly they mentioned that artificial neural network method has higher capability for solving data mining problems.
4	Jyothi Bellary, Bhargavi Peyakunta,	Authors discussed about different machine learning methods, which are applied for data mining process. They classified these

	Sekhar Konetigari Hybrid Machine Learning Approach In Data Mining	machine learning methods as “symbolic and sub-symbolic”. Decision trees, case-based reasoning, instance-Based learning’s are under symbolic classification; and neural network, fuzzy rules, and genetic algorithms are under sub-symbolic classification. In their study they proposed a hybrid approach, which is combination of both case-based reasoning (symbolic) and neural network (sub-symbolic) are applied them for data mining problems.
5	Neagu C.D., Guo G., Trundle P.R., Cronin M.T.D. A comparative study of machine learning algorithms applied to predictive toxicology data mining	In their research they carried a relative study of commonly applied machine learning algorithms to predictive toxicology data set. The outcome of their research was a single classification model cannot be capable of making fine distinctions over the data sets. Multiple classification models are combined which specify that a hybrid intelligent system or multiple classifier system provides accurate result than a single classification algorithm applied. Multiple classifier system is having better performance for any specific application in data mining. It has been seen in another study that a specific characteristic of a data set would be preferred for improving the models correctness.
6	Yogendra Kumar, Upendra An efficient intrusion detection based on decision tree classifier using feature reduction	They examined four machine learning tools (J48, BayesNet, OneR, NB) in data mining for finding instructions and compare their performances. They concluded that J48 is the worthiest algorithm having maximum classification accurateness with high true positive rate (TPR) and low false positive rate (FTR) also less computation time with accuracy.
7	Yugal kumar, G. Sahoo Analysis of bayes, neural network and tree classifier of classification technique in data mining using WEKA	They focused on data mining classification technique. They used six classification machine learning algorithms (bayesnet, naviebayes, navie bayes uptable, multilayer perceptron, voted perceptron and j48 classifiers) on dataset and compared their performance with respect to their mean absolute error, root mean-squared error and time required to build the model using the WEKA data mining tool. In their result they found that the presentation of the J48 technique is reasonably better than other techniques.
8	Aaditya Desai, Dr. Sunil Rai Analysis of Machine Learning Algorithms using WEKA	In his research they compare of performances of different machine learning algorithms with regard to their time needed to construct model and detection of appropriate instances. According to them Multilayer Perceptron provides better outcome than others.
9	Jung-Ying Wang Data Mining Analysis (breast-cancer data)	In his research compared different machine learning tools like WEKA, MATLAB 6.1, LIBSVM on “breast-cancer dataset”. He equated above learning algorithm’s running time, 10 fold cross value of data sets and also applied the training rules in their experiment. They have also mentioned training errors of different algorithms and compared these with other papers.
10	Varun Kumar, Nisha Rathee Knowledge discovery from database Using an integration of clustering and classification	Authors experimented and equated the outcome of a simple classification method (J48 classification) with the outcome of integrated clustering and classification method with the WEKA (Waikato Environment for Knowledge Analysis) data mining tool. As a result they found, the integration of clustering and

		classification techniques gives more accurate and robust result than applying either of them alone. In their study, they have shown that clustering is an unsupervised learning method and it creates classes by partitioning number of clusters according to their instances. 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.
11	Patrick Ozer Data Mining Algorithms for Classification	In this research author described the exercise of different types of classification trees (J48 tree, REP tree, Random tree, LMT tree) methods with different pruning techniques to analyze the performance of algorithms. As a result he found LMT tree algorithm's performance was best among these four classification tree methods.
12	Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya Comparison the various clustering algorithms of weka tools	Authors studied various clustering algorithms of weka, also tried to know their pros and cons. In their research they compared the results of various clustering algorithm to determine the appropriate one for the user. Each algorithm has its individual consequence and that are use according to the behaviour of the data, but K-means algorithm is the simplest one as compared to other techniques.
13	Sam Drazin, Matt Montag	This paper presents the application of J48 decision tree algorithm in the Weka interface for testing data sets. In their research they conclude that, online pruning is useful when it need to reduce the size of the decision tree, but all the time this pruning method gives an inaccurate result. Also found that, for obtaining a suitable value for any training set the confidence factor must be tested at the time of generating a decision tree.

III. CONCLUSION

In this study some of main provisions of data mining and machine learning were used with the intention to identify the common characteristics, applications, and probable mystification generated by a variety of points of observations. There are several research papers worked upon either on data mining classification or clustering and even on combination of both. Some of them have used the application of many machine learning algorithms in their experiment. However, till now to the best of our knowledge none of the studies have used multiple algorithms along with combined approach of data mining classification or clustering. Therefore, in the present research the above approach will be implemented to compare the performance of different machine learning algorithms by doing a case study upon the integrated approach of data mining i.e. combined classification and clustering.

REFERENCES

- [1] <http://www.theiaa.org/intAuditor/itaudit/archives/2006/august/data-mining-101-tools-and-techniques/>
- [2] <http://www.wisegEEK.com/what-are-the-most-important-data-mining-concepts.htm>
- [3] <http://www.wisegEEK.com/what-are-the-different-types-of-data-mining-technology.htm>
- [4] <http://www.wisegEEK.com/what-are-the-different-types-of-data-mining-analysis.htm>
- [5] Chung, H. M., Gray, P. (1999), "Special Section: Data Mining". Journal of Management Information Systems, (16:1), 11-17.
- [6] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, R (1996). "The KDD Process for Extracting Useful Knowledge from Volumes of Data," Communications of the ACM, (39:11), pp.27-34.

- [7] Peacock, P. R. (1998a) "Data Mining in Marketing: Part 1", Marketing Management, Winter, 9-18.
- [8] Han, J., Kamber, M. (2001), "Data Mining: Concepts and Techniques", Morgan-Kaufmann Academic Press, San Francisco.
- [9] J.Jackson, "Data Mining : A Conceptual Overview" , Communications of the Association for Information Systems (Volume 8, 2002) 267-296 267.
- [10] Jyothi Bellary, Bhargavi Peyakunta, Sekhar Konetigari "Hybrid Machine Learning Approach In Data Mining", 2010 Second International Conference on Machine Learning and Computing
- [11] D.Neagu, G.Guo, P.R.Trundle, M.T.Cronin, "A Comparative study of machine learning algorithms applied to predictive toxicology data mining", Department of Computing, University of Bradford, Bradford,UK.
- [12] 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.
- [13] Guo, G., Neagu, D. (2005). Similarity-based Classifier Combination for Decision Making. Proc. of IEEE International Conference on Systems, Man and Cybernetics, pp. 176-181.
- [14] Neelamadhab Padhy, Dr. Pragnyaban Mishra and Rasmita Panigrahi, "The Survey of Data Mining Applications And Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCEIT), June 2012
- [15] Wu Yuntian, Shaanxi University of Science and Technology, "Based on Machine Learning of Data Mining to Further Explore", 2012 International Conference on Computer Science and Information Processing (CSIP)
- [16] Daniel Grossman and Pedro Domingos (2004), "Learning Bayesian Network Classifiers by Maximizing Conditional Likelihood", 21st International Conference on Machine Learning, Banff, Canada.
- [17] J.-J. Aucouturier and F. Pachet, "Improving Timbre Similarity: How high's the sky?," in Journal of Negative Research Results in Speech and Audio Sciences, 2004.
- [18] 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).