Parametric Comparisons of Classification Techniques in Data Mining Applications

Geeta Kashyap[1], Ekta Chauhan[2]
[1]Student of Masters of Technology, [2]Assistant Professor,
Department of Computer Science and Engineering, AP Goyal Shimla University, India

***Abstract* - Data mining (DM) means to extract the hidden knowledge from large repositories of data with the use of techniques and tools. Day to day growing data in every sector which requires automatic data analysis techniques i.e. the techniques which automatically analyze the data for different perspectives and provide accurate results in terms of several parameters i.e. speed, efficiency and cost. Due to increasing interest in Data mining it become emerging research topic for research community. The various techniques of Data mining like classification, clustering can be applied to bring out hidden knowledge from the large databases. In this paper we focus on comparative analysis of different Data Mining (DM) Classification techniques for Data Mining applications i.e. in Retail Industry for Marketing Data analysis, Financial Data Analysis, Educational Data analysis as well as for the analysis of Biomedical Data. Classification is a Data Mining techniques used to predict group membership for data instance. Our research study compares the accuracy of these Classification techniques on Weka tool for Data mining applications. The compared Classification techniques and their algorithms are presented together with some experimental data that give rise to the final conclusion.**

***Index Term* - C5.O, C4.2, Decision Tree, Data Mining Applications, ID3, J48, KDD (Knowledge Discovery in Databases), Naïve Bayes, Neural Network, Weka.**

I. INTRODUCTION

Data Mining (DM)(knowledge Discovery in databases) is the process of extraction of interesting(non-trivial, implicit, previously unknown and potentially useful) pattern or information from large databases using various data mining techniques and tools such as classification, clustering, association rule etc which helps in various decision making or Data Mining is the process of analyzing the data for different perspective that can help in finding the patterns and relationships within data that are collected from various databases and data warehouses.

The essential difference between the data mining and the traditional data analysis (such as query, reporting and on-line application of analysis) is that the data mining is to mine information and discover knowledge on the premise of no clear assumption [1].The various data mining techniques are developed day to day which automatically analyze the databases for different perspectives and also these techniques require less time as compared to traditional data analysis techniques and can be repeated frequently. In this way data mining techniques overcome the problems which are faced traditionally.

Now a days, Data Mining tools and techniques can be successfully applied in various fields in various forms. Many Organizations now start using Data Mining as a tool, to deal with the competitive environment for data analysis. By using Mining tools and techniques, various fields of business get benefit by easily evaluate various trends and pattern of market and to produce quick and effective market trend analysis [2].

1.1 What are patterns?

Patterns are a set of items, subsequences, or substructures that occur frequently together (or strongly correlated) in a data set. Patterns represent intrinsic and important properties of data sets. **Pattern Discovery** is the process of uncovering patterns from massive data set. There are following patterns which are hidden in databases these are:

- Sequential patterns, Time series Patterns, Periodic Patterns etc.

II. DIFFERENT PHASES OF DATA MINING PROCESS:

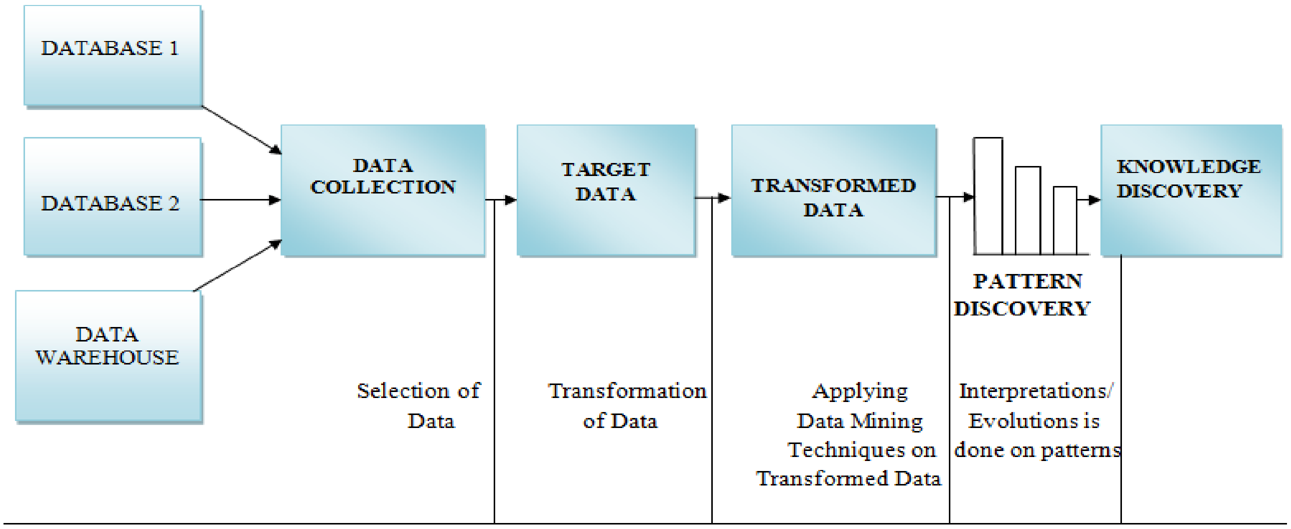


Diagramm1 (Data Mining Phases)

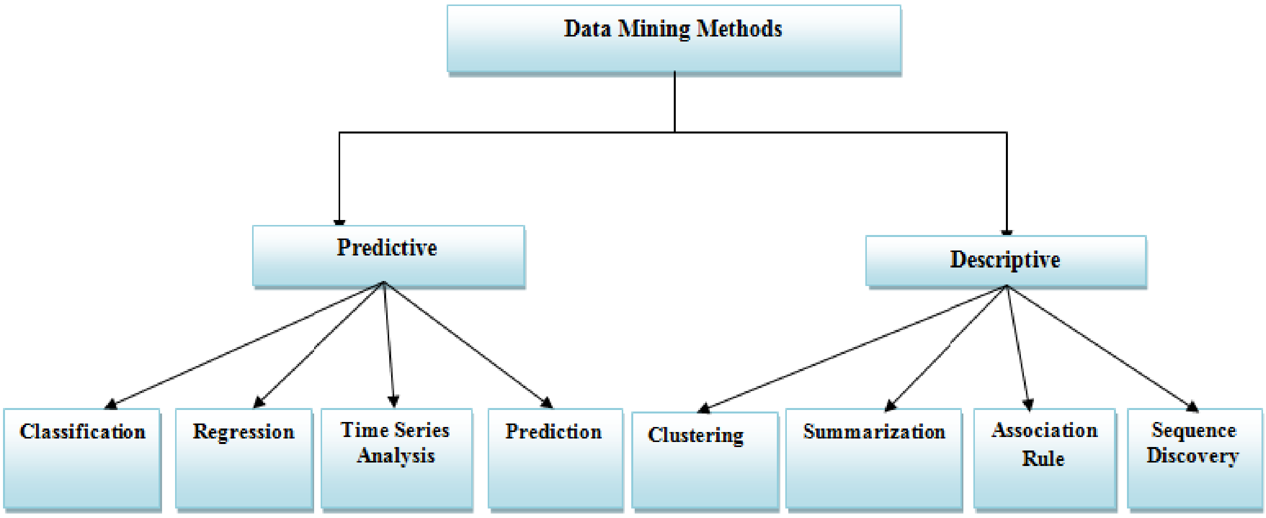
The Data Mining (Knowledge Discovery in Databases (KDD)) process is commonly defined with the following Phases some of the phases of Data Mining Process are iterative in nature:

(1) Data Cleaning
(2) Data Integration
(3) Data Selection
(4) Data Transformation
(5) Data Mining
(6) Pattern Evolution
(7) Knowledge Presentation (Interpretation/Evaluation)

1) ***Data Cleaning:*** In this phase handling of noisy, erroneous, missing, or irrelevant data is done.
2) ***Data Integration:*** In this phase multiple heterogeneous data source may be integrated into one.
3) ***Data Selection:*** In this phase where data relevant to the analysis task are retrieved from the databases.
4) ***Data Transformation****:* In this phase where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
5) ***Data Mining:*** Data Mining is essential process where different data mining methods and tools are applied to extract data patterns.
6) ***Pattern Evolution:*** This phase identify the truly interesting patterns representing knowledge based on some interestingness measures are evolved after data mining phase.
7) ***Knowledge Representation:*** In this phase visualization and knowledge representation techniques are used to present mined knowledge to the user.

Where visualization and knowledge representation techniques are used to present mined knowledge to the user with the widely available relational databases systems and data warehouses, the first four processes: Data Cleaning, Data Integration, Data Selection and Data Transformation, can be performed by constructing Data warehouses and performing some OLAP operations on the constructed data warehouses. The Data Mining, Pattern Evaluation, and Knowledge Presentation processes are sometimes integrated into one (Possibly Iterative) process referred as Data Mining [2, 5].

2.1 Data Mining Methods (Tasks):



Data Mining (DM) not applies only Predictive and Descriptive methods i.e. Classification, clustering, and association analysis, but also proposes, develop and apply methods and tasks drawn from the variety of areas related to Data Mining (statistics, machine learning etc):

2.1.1 PREDICTIVE: Predictive Data Mining methods is used for the analysis of pattern or information in the data set which are unknown and also done the prediction of future values to know what would be likely to happen in future by analyzing current dataset. Predictive Data Mining involves several techniques describes as follows:

- ➢ ***Classification:*** The classification methods classify the data according to predefined categorical classes. It is a supervised learning technique where the target value is known. There are several algorithms used for classifications are:

1. Decision tree
2. Naive biased classification
3. Generalized Linear Models (GLM)
4. Super vector machine etc.

- ➢ ***Regression:*** The regression methods are used to find out the relationship between the dependent variables and outcome variable where the target value is known. It basically used for classify linearly separable data set. There are two types of regression methods:

1. Linear regressions.
2. Non linear regressions.

- ➢ ***Time Series Analysis:*** The time series analysis is done to categories the time series data in order to extract the meaningful knowledge from that data set. This type of analysis is used in banking transactions or in some other type applications that may have a time series dataset.

- ➢ ***Prediction:*** It is a Supervised learning task where the data are used directly (no explicit model is created) to predict the class value of a new instance.

2.1.2 DISCRIPTIVE: Descriptive data mining methods is a Data Mining method which is used to analyze the sequence and behaviors in the dataset .This type of methods are unsupervised learning methods. Descriptive Data Mining involves several techniques describes as follows:

- ➢ ***Clustering:*** In clustering methods the dataset is divided into various groups (Clusters). As per the clustering phenomena the data point of one cluster is more similar to other data point of same cluster and should be dissimilar to data point of another cluster. There are several clustering algorithms describes as follows:
  - K-Means Clustering
  - K-Medoids clustering
  - Hierarchical Clustering.
  - Grid Based Clustering.
  - Density Based Clustering.
  - Optics Clustering.

- ***Association Rule***: This technique is used to find out the association between the variables in a data set, to know how one variable is related to other variable in a dataset. In Association Rule the Support and Confidence are calculated to find out the relationship between the variables. There are several association rule algorithms as follows:
  - Apriori Algorithm.
  - Predictive Apriori Algorithm.
  - Tertius and Filter Associator Algorithm etc
- ***Summarization:*** Summarization is a Data Mining task in which a larger dataset is summarized into a smaller data set in a meaningful manner that describes the general overview of data.

- ***Sequence Discovery:*** Sequence Discovery is Data Mining task where there is analysis of sequential patterns in dataset. For e.g.: In marketing applications it is used to know the particular item is sold out in which sequence i.e. regularly, weekly, monthly, yearly.

III. ISSUES IN DATA MINING:

Data Mining has evolved into an important and active area of research because of the theoretical challenges and practical applications associated with the problem of discovering interesting and previously unknown knowledge from real world databases. The main challenges to the Data Mining and the corresponding consideration in designing the algorithms are as follows:

1. Massive datasets and high dimensionality.
2. Over fitting and assessing the statistical significance.
3. Understandability of patterns.
4. Non-Standard incomplete data and data integration.
5. Handling of redundant data, Uncertain Data.
6. Performance issues: These include efficiency, scalability, and parallelization of Data mining algorithms [5].

IV. DATA MINING APPLICATIONS:

Data Mining applications are widely used in Health industry, Auditing, Telecommunication Industry, Retail industry etc due to day to day growing of data in various fields. Many Organizations now start using Data Mining as a tool, to deal with the competitive environment for data analysis. By using Data Mining tools and techniques, various fields of business get benefit by easily evaluate various trends and pattern of market and to produce quick and effective market trend analysis.

4.1 Data Mining for Biomedical and DNA data analysis:

In recent years, Data Mining has been widely used in area of Medical science such as Biomedical, DNA, Genetics and Medicine etc. In the area of Genetics, the important goal is to understand the mapping relationship between the variation in human DNA sequences and the disease susceptibility. Data Mining is very important tool to help improve the diagnosis, prevention and treatment of the diseases.

4.2 Data Mining for Financial Data Analysis:

Financial data is mainly collected from banks and from other financial sectors. This financial data is usually reliable, complete and has high quality. Financial data need a systematic method for data analysis. Data Mining plays an important in analysis of financial data. Data Mining follows steps such as data collection and understanding, data refinement, model building and model evaluation and deployment. These steps help to deal with analysis of financial data. The proper analysis of financial data enables us to better decisions making capabilities according to the market analysis.

4.3 Data Mining for the Retail Industry:

Data Mining plays an important role in the retail industry also. Retail industry involves large amount of data that includes transportation, sales and consumptions of goods and services. This data grows rapidly due to increase in purchase and sales in business. These days, E-commerce is growing fast with the growth of companies and also improving the online experience. The Data mining in retail industry helps in identifying customer behavior, shopping patterns and distribution policies etc. As retail data in a very large in quantity, so we design data warehouse to store this large data and effective analysis of data. The main decision has to take while designing the data warehouse is dimension, level and preprocessing to perform the quality and efficient data mining [6].

V. COMPARISION OF DATA MINING CLASSIFICATION ALGORITHMS IN DATA MINING APPLICATIONS:

5.1 A Comparison of Different Classification Techniques for Bank Direct Marketing

In this paper K. Wisaeng [2013] present the comparison of different classification techniques in open source data mining software which consists of a decision tree methods and machine learning methods for a set of bank direct marketing dataset. They implemented Decision Tree algorithms i.e. J48-graft and LAD tree and also Machine Learning algorithms i.e. Radial Basis Function Network (RBFN) and Support Vector Machine (SVM) on bank direct marketing datasets. The bank is marketing department can use data mining to analyze customer datasets and develop statistically profiles of individual customer preference for product and service. After that comparison result of Decision Tree and Machine Learning Algorithms is shown along with experimental results [3].

5.2 Predictive Analytics in Higher Education

In this paper Jindal Rajni and Dutta Borah Malaya [2015] implemented a prediction analysis method that can help to improve the education quality in higher education for ensuring organization success at all level. They used the C5.0, C4.5-A2, C4.5-A1 algorithms for prediction analysis, after that they compare their results. The result of C5.0 is best in performance. Then they applied NN (Neural Network) and CRT algorithms on same data set for prediction analysis. After that they compared the result of C5.0 with Neural Network and CRT algorithms result. This paper analyzes the accuracy of algorithm in two ways; the first is by comparing the result of C5.0 with C4.5-A2, and C4.5-A1. After that the C5.0 algorithm is comes out to be best algorithm in accuracy. Then its result is compared with NN (Neural Network) and CRT [7].

5.3 A Survey on the Classification Techniques In Educational Data Mining

In this paper, N. Upadhyay and V.Katiyar [2014] focus on the educational Data Mining Classification techniques to analyze attributes for the prediction of student's behavior and academic performance .They implemented various Classification methods like Decision Trees, C4.5 algorithm, ID3 algorithm, Decision Tree Model and Neural Network Model are used to predict student’s behavior on WEKA open source data mining tool. After that the comparative result of these algorithms is shown along with some experimental data [8].

5.4 Empirical Comparison by Data mining Classification algorithms (C 4.5 & C 5.0) for Thyroid Cancer data set

In this paper A.Upadhayay, S. Shukla, S. Kumar [2013] implemented the Data Mining Classification techniques to classify the thyroid cancer patient data sets. They collect database of 2800 patients’ records each with 29 attributes from which they considered 400 patients records for their research. Than classification algorithms C4.5, C5.0 is used to classify the datasets. After that comparative analysis of these algorithms is shown with experimental results [9].

5.5 Analysis on Classification Techniques in Mammographic Mass Data Set

In this paper Mrs. K. K. Kavitha, Dr. A Kangaiammal; Mr. K. Satheesh [2015] implemented a various Data Mining Classification techniques such as Decision, Naïve Bayes, and K-Nearest Neighbour (KNN) Classifiers on weka tool to classify the Mammographic mass dataset. After that the comparative results of these algorithms is shown with some experimental results [10].

5.6 *COMPARISION TABLE*

PAPER	AUTHOR	DATA SET	TOOL USED	CLASSFICATION ALGORITHM USED	ACCURACY	ADVANTAGE
omparison of ferent ssification hniques for k Direct keting	K. Wisaeng	Bank Direct Marketing Data Set	Weka	J48-Graft	76.52%	SVM Produces very accurate Results .It provides accurate result for both Online and Offline Data Sets. Less over fitting, robust to noise. Especially popular in text classification problems where very high-dimensional Spaces are the norm. Memory-intensive.
				LADT	76.08%	
				RBFN	74.34 %	
				SVM	86.95%	
lictive lytics in her Education	Jindal Rajni and Dutta Borah Malaya	Education Data set	Weka	C5.0	99.91%	C5.0 is better in speed, memory and efficiency than C4.5-A1, C4.5-A2 and the C5.0 algorithm provides maximum Information Gain
				Neural Network(NN)	98.09%	
				CRT	98.72%	

urvey on the ssification hniques In cational Data ing	N. Upadhyay and V.Katiyar	Education Data Set	Weka	ID3	93%	ID3 and Decision Tree Model produces the more accuracy result than the other Algorithm in terms of efficiency and specificity. ID3 algorithm generally uses nominal attributes for classification with no missing Values. It produces false alarm rate and omission rate decreased, increasing the detection rate and reducing the space Consumption
				J48	78.6%	
				Naïve Bayes classifiers	75%	
				Decision tree model	85.188%	
				Neural network model	83.875%	
pirical nparison by mining ssification orithms .5 & C 5.0) Thyroid cer data set.	A.Upadhayay S. Shukla S. Kumar	Thyroid Cancer Patient Data Set	Weka	C4.5	90%	C5.0 Tree generated more accurate results or rule set. Running Time of C5.0 was Small as compare to C4.5. Train error in case of C5.0 was small in compare to the C4.5
				C5.0	95%	
lysis on ssification hniques in mmographic s Data Set	Mrs. K. K. Kavitha, Dr. A.Kangaiammal and Mr. K. Satheesh	Mammo - -graphic Mass Data Set	Weka	Decision Tree	75.8%	KNN is providing accurate results where data is not linearly separable. KNN also provides accurate results where there is a ambiguity and uncertainty in data set
				Naïve Bayes	75.1%	
				KNN	80%	

VI. CONCLUSION:

This paper describes goals of Data Mining (DM), phases of Data Mining, Data Mining Methods (Tasks) as well as the Issues of Data Mining and Data Mining Applications. In this paper, we did the comparative study of different Data Mining Classification techniques with their algorithms on various data sets using Weka tool. We also did the comparative analysis on the basis of accuracy percentage. We also analyzed the advantages of algorithms that applied to various data sets. It is difficult to say that which Classification technique of Data Mining is best because each technique has its own advantages and limitations and it also depend upon the purpose for which data is to be mined.

But according to our comparative study on Data Mining Classification Techniques on Weka tool we can say that Decision Tree algorithms C5.0, ID3 is provide accurate result for the Classification of Structured Educational Data Set but to Classify the Unstructured Educational Data Set the Super Vector Machine(SVM), Naïve Bayes Classification as well as the Neural Network (NN) Classification provide the accurate results in terms of several parameters Speed, Efficiency and in Bio-Medical Data Analysis the Decision Tree algorithm C5.0 provide the better result as compared to C4.5 algorithm . The Neural Network (NN) Classifier provides the more accurate result as compared to Decision Tree and Naïve Bayes Classifier in terms of efficiency for the analysis of Mammographic Mass Data Set. In Bank Direct Marketing the Support Vector Machine (SVM) Classifier provides more accurate results in terms of speed and efficiency as compared to other Classifiers. Our research study will be beneficial to the researchers in Data Mining for the selection of Data Mining Classification techniques with their algorithm according to their Data Mining Applications. So we can say that this paper will provide a beneficial glance of existing solution for Data mining Classification techniques with their accuracy and advantages in Data Mining Applications.

VII. REFERENCES

[1] G. Kashyap, E. Chauhan, Review on Educational Data Mining Techniques, *International Journal of Advanced Technology in Engineering and Science, 3(11),* 2015, 308–316.

[2] M. Soundarya, R. Balakrishnan, Survey on Classification Techniques in Data mining, *International Journal of Advanced Research in Computer and Communication Engineering, 3(7),* 2014, 7550–7552.

[3] K. Wisaeng, A Comparison of Different Classification Techniques for Bank Direct Marketing. *International Journal Of Soft Computing and Engineering (IJSCE), 3(4),* 2013, 116–119.

[4] B. Pannu, P. Sharma, a Comparative Data Mining Technique on Education, *International Journal for Technological Research in Engineering, 2(9),* 2015, 2124–2127.

[5] H.Sahu, S.Shrma, S.Gondhalakar, A Brief Overview on Data Mining Survey, *International Journal of Computer Technology and Electronics Engineering (Ijctee), 1(3),* 2008, 114–121.

[6] S. Bagga, G. Singh, Applications of Data Mining, *International of Science and Emerging Technologies with the Latest Trends, 1(1),* 2012, 19-23.

[7] R.Jindal, M.D Borah (2015), Predictive Analytics in Higher Education, 1520-9202/15/$31.00 © 2015 IEEE

[8] N. Upadhyay, V.Katiyar, A Survey on the Classification Techniques in Educational Data Mining, *International Journal of Computer Applications Technology and Research, 3(11),* 2014, 725–728.

[9] A.Upadhayay, S. Shukla, S. Kumar, Empirical Comparison by data mining Classification algorithms (C 4.5 & C 5.0) For thyroid cancer data set, *International Journal of Computer Science & Communication Networks, 3(1),* 2013, 64–68.

[10] K. Kavitha, A. Kangaiammal, & K. Satheesh, Analysis on Classification Techniques in Mammographic Mass Data Set, *International Journal of Engineering Research and Applications 5(7), 2015, 32–35.*

[11] R.Jindal, M.D Borah, A Survey on Educational Data Mining and Research trends, *International Journal of Database Management System (IJDMS), 5(3),* 2013, 53–73.

[12] Baradwaj, B. Kumar, Mining Educational Data to Analyze Students Performance. *International Journal of Advanced Computer Science and Applications (IJACSA), 2(6),* 2011, 63-69.