Survey on Classification Techniques Used in Data Mining and their Recent Advancements

Saranya Vani.M[1], Dr.S. Uma[2],Sherin.A[3], Saranya.K[4]

[1] PG Scholar, PG CSE Department, Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, India
[2]Head of the Department, PG CSE Department, Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, India
[3]PG Scholar, PG CSE Department, Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, India
[4]PG Scholar, PG CSE Department, Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, India

***Abstract*—Data Mining is an emerging field which has attracted a large number of information industries due to huge volume of data managed in recent days. Efficient data mining requires a good understanding of the data mining techniques to improve business opportunity and to improve the quality of service provided. In response to such needs, this paper provides a review of traditional classification techniques used for data mining. The main attention is on classification techniques like decision tree induction, Bayesian networks, rule based classification, k-nearest neighbor classification techniques which are used to mine databases. A comprehensive review is done on the issues, recent advancements and research works on these techniques.**

***Index Terms*—classification technique, decision tree induction, k nearest neighbor classifier, Bayesian network, and rule based classification.**

I.INTRODUCTION

Data mining refers to extracting or mining knowledge from large amounts of data. It is an essential process where intelligent methods are applied in order to extract the data patterns. Data mining, popularly known as Knowledge Discovery in Databases (KDD), is an extraction of previously unknown and potentially useful information from data in databases [10]. It can also be said as a process of finding the hidden information/pattern of the repositories. Efficiently mining data considering system scalability and performance poses numerous challenges to researchers and developers which had led to huge development in data mining field. The data mining techniques can be utilized to mine interesting sets of data from huge databases and applied to information management, query processing, decision making and many other applications.

Data mining is used extensively as huge amount of data is available with very little information and that there is a need to extract useful information from the data and to interpret the data. Data mining automates the process of finding relationships and patterns in raw data and the results can be either utilized in an automated decision support system or assessed by a human analyst. Data mining is used in science and business areas which need to analyse large amounts of data to discover trends. Data mining is applied in fields such as Financial Data Analysis, Biological Data Analysis, Retail Industry, Telecommunication Industry, Other Scientific Applications, Intrusion Detection and many more.

II. CLASSIFICATION OF DATA MINING SYSTEM

Data mining is an interdisciplinary field which includes database systems, visualization, statistics and many others. Therefore based on the kinds of data to be mined or on the given data mining applications we can incorporate techniques from other fields like pattern recognition, image analysis and so on. Hence it is very important to understand the classification of data mining system to take the advantage of diversity of research going on in other disciplinary fields contributing to data mining.

The data mining systems can be categorized into many ways as the following.

- Kinds of databases mined
- kinds of knowledge mined
- kinds of techniques used
- kinds of applications adapted

Mining databases for knowledge can be of any type like the relational databases, object oriented databases and other kind of databases. Classification

ISSN: 2278 – 7798

based on the knowledge mined utilizes the data mining functionalities like classification, prediction, outlier clustering and many others. The system categorization based on techniques makes use of methods of data analysis employed like machine learning, statistics, and visualization. Classification of data mining systems can also be done based on kinds of applications adapted like finance, stock markets and many other applications.

Data mining can be considered as a synonym for knowledge Discovery of data or as a step in the process of knowledge discovery. Knowledge discovery is an iterative process and it consists of seven basic steps. The first four steps are used to preprocess the data i.e. data is prepared for mining. The steps can be elaborated as follows:

- Data cleaning
- Data integration
- Data selection
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge presentation

The preprocessing of data involves removing of noise, inconsistent data, combining multiple data sources, selection of appropriate data for analysis, consolidating data and so on. Data mining is the step where intelligent methods are applied to mine patterns and pattern evaluation identifies the interesting patterns according to the user requirements. Knowledge presentation is the final step where data is presented to the user using representation techniques that are available.

3. DATA MINING FUNCTIONALITIES

The search of data patterns in a database by users is not predictable. Sometimes the user would have no idea on what kind of patterns in their data is interesting which could lead them to search for various kinds of patterns in parallel. Hence the mining systems should be capable of mining multiple kinds of patterns based on the user expectations. Data mining functionalities are used to specify the kind of patterns to be mined in data mining systems. Data mining functionalities and the kinds of patterns they discover can be grouped as follows:

- Characterization and discrimination
- Classification and prediction
- Mining frequent patterns, Associations and correlations
- Cluster analysis
- Outlier analysis

Data mining tasks can be classified in two categories as descriptive and predictive. The descriptive mining tasks characterize the general properties of the data in database and use it for classification whereas predictive mining tasks perform inference on the current data in order to make predictions.

Classification and prediction are two forms of data analysis that can be used to extract models describing the important data classes or to predict the future data trends whose class labels are known in advance. Such analysis can help to provide us with a better understanding of the data at large. The classification predicts categorical (discrete, unordered) labels, prediction models continuous valued function. Prediction is used for missing or unavailable numeric data values. Prediction may refer to both numeric as well as class label prediction. Regression analysis is a statistical method used for numeric prediction.

Clustering is also a data analysis process which classifies the data without consulting a known class label. The objects are classified by comparing the similarity between the tuples and the tuples in the training set. In certain cases the data objects may not comply with the group behavior of the model. Such data objects are called outliers and analysis of such patterns is called outlier mining. Outlier mining in most methods is considered as noise in data, but in some cases analysis of such data could disclose interesting patterns in data.

4. CLASSIFICATION TECHNIQUES

Classification is a data mining technique used to predict group membership for data instances. There are many traditional classification methods like decision tree induction, k-nearest neighbor classifier, Bayesian networks, support vector machines, rule based classification, case-based reasoning, fuzzy logic techniques, genetic algorithm, rough set approach and so on. The basic difference between the algorithms depends on whether they are lazy learners or eager learners. The decision tree classifiers, Bayesian classifier, support vector classifier are eager learners as they use training tuples to construct the data model whereas nearest neighbor classifiers are lazy learners as they wait until a test tuple arrives for classification to perform generalization. A brief discussion about the decision tree algorithm, Bayesian networks, Rule based classification, K Nearest neighbor and other classification techniques, their issues and recent works to overcome these issues is done in this section.

ISSN: 2278 – 7798

A. Decision tree induction

The decision tree induction is learning about decision tree which is a flow chart like structure. Each node in the structure denotes a test on an attribute value and each branch represents an outcome of the test. The tree leaves represent the class distribution. A decision tree is a predictive model most often used for classification. Each interior node in the decision tree tests the value of some input variable against a classification question, and the branches from the node are labeled with the possible results of the test. The leaf nodes intimate the class to return if that leaf node is reached. The classification of an input instance is performed by starting at the root node and based on the results of the investigation appropriate branches are traversed until a leaf node is reached. Thus a Decision Tree can be viewed as a tree-shaped diagram used to determine a statistical probability. The construction of decision classifiers is quite simple as it does not require any domain knowledge for exploring statistics of data.

Decision tree classifiers allocate objects to predefined classes as and when they come across test data. The decision trees are popular tools for mining streams of data i.e. data which are of infinite set. The main aspect in classification indecision trees for mining data streams is to select the best attribute based on which the grouping of the objects is to be done. Many of the methods used for attribute selection for mining data streams are either wrongly mathematically justified or time-consuming, like the Hoeffding tree algorithm and McDiarmidtree algorithm [2]. The Gaussian decision tree algorithm proposed in [2] is a modification of Hoeffding tree algorithm which tries to choose the best attribute for classification for the current set of data elements as the best for the whole data stream.

Decision tree techniques are very famous and efficient and they can be applied to large scale applications. They are also recognized as highly unstable classifiers with respect to minor changes in the training data presenting high variance in classification of data sets. Fuzzy logic can be used to improve these aspects due to the elasticity of fuzzy sets formalism [3]. Selecting representative samples of data such that a learning algorithm can have a reduced computational cost and an improved learning accuracy is of high importance. [4] proposes a new sample selection method by adding principle of maximal classification ambiguity for selecting the right representative samples and proves that this method is superior when compared to random sample selection methods.

Restructuring decision tree structure for large volumes of data which need to be classified in an online fashion i.e. data from applications like sensor networks and ecommerce is an issue encountered in distributed databases. Most of the classification approaches proposed requires that the new instance to be fully labeled and also the agent who is in charge of the classification suffers to reconstruct the decision tree each time it observes a new data. Since the classification agent does not see the dataset as a whole while constructing the decision tree the accuracy of the model built is compromised sometimes. [5] proposes a new approach in which the agent responsible for restructuring decides when to get a new model. It verifies by either borrowing it from another agent, or can be done by inducing a new classifier.

B. Bayesian Network

A Bayesian classifier is a classification technique used to determine if the given tuple belongs to a particular class or not. The classification is based on Bayes' theorem. Bayes' theorem measures the probability that a given data tuple belongs to a particular class. It is used as a statistical inference to the given set of data. The Bayesian classifiers exhibit high accuracy on when applied to large datasets. Bayesian classification can be a Naïve Bayesian classification or Bayesian network classification. The difference between the two is that the former assumes that the effect of an attribute value on a given class is independent of the values of the other attributes while the later allows representation of dependencies among the subset of attributes.

Constructing Bayesian network classifier structure depends on whether the attribute values on a given class are independent of each other or not. The current methods exchange mutual information to estimate the dependency among variables, which lacks theoretical basis leading to low reliability. Research works are carried out to classify data based on dependency analysis and hypothesis testing [6].
Another issue for constructing a Bayesian network classifier is to learn an accurate Bayesian network structure. K2 algorithm is considered to be most efficient for learning Bayesian network classifier. This algorithm requires variable ordering in advance to construct the Bayesian network structure. Existing methods neglect the information of variables selected for classification. To overcome this difficulty [7] proposes an L1 regularized Bayesian network classifier (L1-BNC) which defines a variable ordering by the LARS (Least Angle Regression)

ISSN: 2278 – 7798

method. It then makes use of this classifier with K2 to construct a Bayesian network classifier.

C. Rule Based Classification

In a rule based classification model the information is represented as IF-THEN rules format. The if-part is called the rule antecedent and the else part is the rule consequent. If the condition in the rule antecedent part is true it means that the rule covers the tuple and if it is false then the rule does not cover the tuple. A rule is assessed by it coverage and accuracy. Rules coverage is the percentage of tuples covered by the rule and its accuracy is the percentage of correctness in classification. The if-then rules can be extracted from decision tree as the if-then classification rules are easy to interpret by humans. These rules can also be extracted from sequential covering algorithms.

Most of the rule based classification methods focus on performance rather than interpretability of data. ROUSER is a rule based classification method which focuses on human understandable decision rules from data. It uses a rough set approach to select attribute value pair for the IF condition and it is said to be more efficient compared to other rule based classification methods [11]

The fuzzy rule-based classification system is difficult to deal with due to exponential growth of the fuzzy rule search space when the number of patterns becomes high. This makes the learning process more difficult and it leads to problems of scalability (in terms of the time and memory consumed) and/or complexity (with respect to the number of rules obtained and the number of variables included in each rule). Fuzzy association rule-based classification method for high-dimensional problems has been proposed to obtain an accurate and compact fuzzy rule-based classifier with a low computational cost [12].

D. K-Nearest neighbor classifier

Nearest neighbor classifiers is a lazy learner's method and is based on learning by analogy. It is a supervised classification technique which is used widely. Unlike the previously described methods the nearest neighbor method waits until the last minute before doing any model construction on a given tuple. In this method the training tuples are represented in N-dimensional space. When given an unknown tuple, k-nearest neighbor classifier searches the k training tuples that are closest to the unknown sample and places the sample in the nearest class.

The K nearest neighbor method is simple to implement when applied to small sets of data, but when applied to large volumes of data and high dimensional data it results in slower performance. The algorithm is sensitive to the value of k and it affects the performance of the classifier. New Field Programmable Gate Arrays (FGPA) architectures of KNN classifiers have been proposed in [8] to overcome this difficulty of classifier to easily adapt to different values of k.

Accuracy in data classification is a major issue in data mining and in order to improve the accuracy of classification, improvements have been made to the K nearest neighbor method. Weighted nearest neighbor classifier (wk-NNC) is one such method which adds a weight to each of the neighbors used for classification. Hamamoto's bootstrapped training set can also be used instead of the training patterns where training pattern is replaced by a weighted mean of a few of its neighbors from its own class of training patterns. This method proves to improve the accuracy of classification. However the time to create the bootstrapped set is O (n^2) where n is the number of training patterns. K-Nearest Neighbor Mean Classifier (k-NNMC) proposed in [9] finds k nearest neighbors for each class of training patterns separately. The classification is done based to the nearest mean pattern. This improvisation proves to show better accuracy of classification in when compared to other techniques using Hamamoto's bootstrapped training set.

E. Alternative methods of classification

Besides the above mentioned classification methods other classical methods also exits. To name a few of them, classification by back propagation, rough set approach, support vector machines, genetic algorithms, classification by association rule analysis, fuzzy set and many more. The methods of classification are chosen based on the user and data classification needs. Some of the issues faced by these techniques are reviewed below.

Back propagation is a neural network learning algorithm and it learns by iteratively processing a data set of training tuples. It compares the network prediction of each tuple with class labels. Back Propagation networks are ideal for simple Pattern Recognition and Mapping Tasks and they are used for classification of data. One of the well-known issues in back propagation method is the problem encountered with local maxima [16].Many variations to back propagation method have been proposed to

ISSN: 2278 – 7798

overcome the issues faced due to growing network size.

Support vector machines were designed for binary classification of data i.e. classifying data into two classes. Several research works are carried out to extend these binary classification methods to multiclass classification. As solving multiclass classification is computationally expensive many comparative study to improve this technique has been done [17]. The support machine vector is also used for classification of linear and nonlinear data. It transforms the data in a higher dimension from where it can find a hyperplane for separation of data using training tuples called support vector.

The genetic algorithms, rough set approach and fuzzy sets are algorithms which are not used often. But their logics are applied to other classification techniques. The Fuzzy set theory can be used in systems which perform rule based classification. The rough sets can be used for attribute selection or feature reduction where attributes that do not contribute for a classification can be identified and removed.

Associative classification uses association mining techniques for mining data. Classification rule mining and association rule mining are two important data mining techniques. Association rule mining is one which finds the rules in the database that satisfy some minimum constraints whereas classification rule mining aims to discover a small set of rules in the database to form an accurate classifier. Both classification rule mining and association rule mining are vital to practical applications. It would be of value to the user if the two mining techniques can be integrated. Research works have been done to integrate the two frameworks, called associative classification [18].

V. CONCLUSION

There are numerous comparative studies of the various classification techniques, but it has not been found that one single method is superior compared to others. Issues like accuracy, scalability, training time and many others contribute to choosing the best technique to classify data for mining. The search for best technique for classification still remains a research topic. In this paper we have made a comprehensive study of the classification techniques used in data mining and their recent advances and issues faced due to the growing volumes of data these days.

VI.REFERENCES

[1] Jaiwei Han and MichelineKamber.Data Mining Concepts and Techniques – Second edition,.

[2] Rutkowski, L.,Jaworski, M. ,Pietruczuk, L. ,Duda, P., "*Decision Trees for Mining Data Streams Based on the Gaussian Approximation*"- Knowledge and Data Engineering, IEEE Transactions on (Volume:26 , Issue: 1) Jan 2014

[3] A complete fuzzy decision tree technique - Cristina Olaru∗, Louis Wehenkel, Elseiver – Fuzzy sets and systems,February 2003

[4] Xi-Zhao Wang; Ling-Cai Dong ; Jian-Hui Yan , "*Maximum Ambiguity-Based Sample Selection in Fuzzy Decision Tree Induction*", Knowledge and Data Engineering, IEEE Transactions on (Volume:24 , Issue: 8) - Aug2012

[5]Bazzan, A.L.C.,"*Cooperative induction of decision trees*" Intelligent Agent (IA), 2013 IEEE Symposium on April 2013

[6] Sun Wenjing ; Yang Youlong ; Li Yangying, "*Learning Bayesian Network Classifier Based on Dependency Analysis and Hypothesis Testing*", Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2013 5th International Conference on Aug 2013

[7] Ying Wang; Hao Wang ; Kui Yu ; Hongliang Yao , "*L1 regularized ordering for learning Bayesian network classifiers*", Natural Computation (ICNC), 2011 Seventh International Conference on July 2011

[8] Hussain, H.M. ; Benkrid, K. ; Seker, H. ,"*An adaptive implementation of a dynamically reconfigurable K-nearest neighbor classifier on FPGA*" , Adaptive Hardware and Systems (AHS), 2012 NASA/ESA Conference on June 2012

[9]Viswanath, P. ; Sarma, T.H.,"*An improvement to k-nearest neighbor classifier*", Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE , Sept. 2011

[10] Ming-Syan Chen; Jiawei Han ; Yu, P.S. ,"*Data Mining : An overview from database perspective*", Knowledge and Data Engineering, IEEE Transactions on (Volume:8 , Issue: 6),dec1996

[11] Chia-Chi Liao; Kuo-Wei Hsu ,"*A rule-based classification algorithm: A rough set approach*", Computational Intelligence and Cybernetics (CyberneticsCom), 2012 IEEE International Conference on July 2012

[12]Alcala-Fdez, J. ; Alcala, R. ; Herrera, F. ,"*A Fuzzy Association Rule-Based Classification Model for High-Dimensional Problems With Genetic Rule Selection and Lateral Tuning*", Fuzzy Systems, IEEE Transactions on (Volume:19 , Issue: 5), Oct 2011

[13]-J. HUYSMANS, B. BAESENS, D.MARTENS, K. DENYS and J. VANTHIENEN,"*New Trends in Data Mining*" Vol. L, 4, 2005

[14] Nikita Jain, Vishal Srivastava, "*DATA MINING TECHNIQUES: A SURVEY PAPER*"–IJRET – Nov 2013

[15] Qiang yang and Xindongwu, "*10 CHALLENGING PROBLEMS IN DATA MINING RESEARCH*", International Journal of Information Technology & Decision MakingVol. 5, No. 4 (2006) 597–604

[16] R. Rojas: "*The Backpropagation Algorithm*", Neural Networks, Springer-Verlag, Berlin, 1996

[17]Chih-Wei Hsu; Chih-Jen Lin, "*A comparison of methods for multiclass support vector machines*", Neural Networks, IEEE Transactions on (Volume:13 , Issue: 2) on Mar 2002

[18]Bing Liu Wynne Hsu Yiming Ma; "Integrating Classification and Association Rule Mining", American Association for Artificial Intelligence -1998

[19]Pernkopf,F.;Wohlmayr,M.; Tschiatschek, S. **; "***Maximum Margin Bayesian Network Classifiers";*Pattern Analysis and Machine Intelligence, IEEE Transactions on (Volume:34 , Issue: 3)March 2012

[20]Barros, R.C. ; de Carvalho, A.C.P.L.F. ; Basgalupp, M.R. ; Quiles, M.G.;"*A clustering-based decision tree induction algorithm*", Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on Nov 2011

ISSN: 2278 – 7798

[21] P. Domingos and G. Hulten, "Mining high-speed data streams",Proc. 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 71-80, 2000.
[22] V. Garcia, E. Debreuve, and M. Barlaud, "Fast k nearest neighbor search using GPU," in Proc. of 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Anchorage, Alaska, USA, June 23-28, 2008, pp. 1-6.
[23]Masud,M.M. ,Jing Gao ; Khan, L. ; Jiawei Han ; Thuraisingham, Bhavani, "A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data ",Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on 15-19 Dec. 2008
[24] A. Bifet and R. Kirkby, Data Stream Mining a Practical Approach,University of WAIKATO, Technical Report, 2009.
[25] James Dougherty, Ron Kohavi, and MehranSahami.Supervised and unsupervised discretization of continuous features. In International Conference on Machine Learning, pages 194–202, 1995.
[26] David Hand, HeikkiMannila and Padhraic Smyth,Principles of Data Mining, MIT press, 2001
[27] Thair Nu Phyu, Survey of Classification Techniques in DataMining, Proceedings of the International MultiConference of Engineers and Computer Scientists 2009
[28] E.W.T. Ngai a,*, Li Xiu b, D.C.K. Chau a, Application of data mining techniques in customer relationship management: A literature review and classification, Expert Systems with Applications (2009)
[29] R. Agrawal, T. Imielinski, and A. Swami. Database mining: A performance perspective. IEEE Trans. on Knowledge and Data Engineering, 5(6), Dec. 1993.
[30] E. Manolakos and I. Stamoulias, "IP-cores design for the kNN classifier," in Proc. of IEEE International Symposium on Circuits and Systems, Paris, France, May 30- June 2, 2010, pp. 4133 – 4136.

Brief Author biography

SaranyaVani M received Bachelor of Technology degree in Information Technology from Amrita Vishwa Vidyapeetham, Coimbatore, India. Now she is pursuing Masters of Engineering in Computer Science & Engineering from Hindusthan Institute of Technology, Coimbatore, Tamilnadu, India affiliated to Anna University.

Dr S.Umais Professor and Head of PG Department of Computer Science and Engineering at Hindusthan Institute of Technology, Coimbatore, Tamilnadu, India. She received her B.E., degree in Computer Science and Engineering in First Class with Distinction from PSG College of technology in 1991 and the M.S., degree from Anna University, Chennai, Tamilnadu, India. She received her Ph.D., in Computer Science and Engineering Anna University, Chennai, Tamilnadu, India with High Commendation. She has nearly 24 years of academic experience. She has organized many National Level events like seminars, workshops and conferences. She has published many research papers in National and International Conferences and Journals. She is a potential reviewer of International Journals and life member of ISTE professional body. Her research interests are pattern recognition and analysis of non-linear time series data.

Sherin A received Bachelor of Technology degree in Computer Science from Government College of Engineering, Sreekrishnapuram, Palakkad, Kerala, India affiliated to Calicut University. Now she is pursuing Masters of Engineering in Computer Science & Engineering from Hindusthan Institute of Technology, Coimbatore, Tamilnadu, India affiliated to Anna University.

Saranya K received Bachelor of Technology degree in Information Technology from Sri Ramakrishna Institute of Technology, Coimbatore, India. Now sheis pursuingMasters of Engineering in Computer Science &Engineering from Hindusthan Institute of Technology, Coimbatore, Tamilnadu, India affiliated to Anna University.

ISSN: 2278 – 7798