International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 05 | May-2016 www.irjet.net p-ISSN: 2395-0072

A Survey on Different Techniques of Classification in Data mining

Sanket S. Ranaware[1], Dr. G.P. Potdar[2]

[1]Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India.
[2]Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India

---

Abstract - *Now a days managing the huge volume of data has become challenging task. Data mining is the emerging field which attracted many industries to manage such amount of data. To improve the business opportunity and the quality of service provided efficient data mining must be implemented, and good understanding of the data mining techniques is required for that. Data classification is one of the effective and challenging technique of data mining. Goal of this survey is to provide review of various classification techniques in data mining. The main focus is on classification techniques like decision tree induction, Bayesian network, K-nearest neighbor classification, Rule based classification techniques, which are widely used for data mining.*

Key Words: classification techniques, .decision tree induction, bayesian network, k nearest neighbour classification, and rule based classification.

1. INTRODUCTION

Data mining is basically analysis step of the knowledge discovery in the databases process. The goal of the data mining is to extract the knowledge and the patterns from the huge amount of data. It does not mean extracting the data itself. This is basically the process of analyzing the data from various perspectives and then summarizing it into meaningful information. This information can be useful in various ways, like increasing profits, reducing costs or both. Data mining techniques are used to extract meaningful interesting sets of data for huge databases and then supplied to information management, decision making, query processing and in many more applications.

Data mining is mostly used when large amount of data is present with very small amount of information. And ii is necessary to extract such information to perform various operations. Science and business areas mostly used data mining, because there is plenty of data is present and that need to be analyzed. Data mining is applied in fields like Telecommunication industry, Biological Data Analysis, Retail industry, Financial Data Analysis, Various Scientific applications and many more.

2. DIFFERENT DATA MINING SYSTEMS

Data mining is related to many fields such as database system, statistics, visualization and many others. So based on the applications requirements or kind of data required to mined we can use fields from other fields like image analysis, pattern recognition etc. Therefore, it is very important to recognize the classification of data mining system to take advantage of diversity of research going on in other related fields contributing to data mining.

We can consider following points while categorizing data mining techniques.

- Which kind of applications gone adapt it.
- Which kind of databases need to be mined.
- Which kind of techniques/approaches used.
- Which kind of knowledge to be mined.

Categorization of data mining is done based on kinds of applications adapted like scientific applications, finance applications etc. Mining techniques can be classified based on type of databases to be mined, like object oriented database, relational databases, temporal database and many more. In another way mining techniques can be differentiated based on the methods used to analyze the data. That methods can be visualization, statistics, machine learning,

Data mining can be considered as a step in the knowledge discovery process. Knowledge discovery is an iterative process which is composed of seven basic steps. The first four steps are of data preprocessing. This seven steps are as follows:

- Data Cleaning
- Data Integration
- Data selection
- Data Transformation
- Data extraction/mining
- Pattern evaluation
- Knowledge presentation

The preprocessing steps includes activities of removing of unnecessary contents, noise, inconsistent data, combining multiple data sources, selection of appropriate data for analysis and so on. Data extraction/mining is the step where different intelligent methods are applied to extract the patterns and evaluation of such patterns will identify the interesting pattern. Using different knowledge representation techniques data is displayed to the user in the Knowledge

3. DATA MINING FUNCTIONALITIES

User searches different data patterns in the database. Sometimes user would have no idea which kind of pattern he need to extract from data. So the mining system should be capable of mining the multiple kind of patterns which would satisfy user expectations. Data mining functionalities are used to specify the kind of patterns to be mined. Data mining functionalities are as follows:

- Cluster Analysis
- Mining frequent items
- Associations and Correlations.
- Characterization and discrimination.
- Classification and prediction.

We can classify data mining tasks into two categories as predictive and descriptive. In predictive tasks inference is performed on the current data and the predictions are made. Whereas in descriptive data mining tasks general properties of the data are characterized and used it for classification.

4. CLASSIFICATION TECHNIQUES

Classification is one of the data mining technique and used to predict group membership for data instances. Currently, there are many classification techniques present like Decision tree induction, Bayesian network, k-nearest neighbor classifier, Support vector machines, fuzzy logic techniques, rule based classification etc. We can differentiate this algorithms based on whether they are lazy learners or eager learners. The decision tree classifier, support vector machine, Bayesian networks are eager learners. Basically they use training tuples to define and construct the data model. Whereas nearest neighbor classifiers are lazy learners, as they wait for test tuple to arrive for classification to perform generalization.

In this section a brief discussion on decision tree algorithm, Bayesian networks, and rule based classification, K nearest neighbor classification and other classification techniques, their issues, recent work to overcome those issues is done.

4.1 Decision tree induction

This method is learning about the decision tree which is structure similar to the flow chart. Each node in this tree denotes the condition applied on the attribute value and each branch of the tree denotes outcome of that condition. The leaf nodes of the tree denotes the final classes of the data. A decision tree is the predictive model of classification. Every interior node of the decision tree performs the conditions checking of input variable according to the classification problem. And the branches from those nodes are labeled with the possible outcome that condition check. Whereas the leaf nodes represents the class which returns if expected leaf node is reached. Here the classification of input started at root node of decision tree and based on the results of the investigation respective branches are traversed until a leaf node is reached. So the decision tree is shaped diagram used determine a statistical probability. Constructing a decision tree is very simple task it does not require any specific domain knowledge for finding statistics of data.

Decision tree classifiers assigns objects to predefined classes when then came across condition checking or test data. Decision trees are most widely used for mining streams of data which are infinite sets. The main aspect in decision tree classification is to select most suitable attribute for grouping of objects. Most of the approaches used for attribute selection for mining data are either wrongly mathematically calculated or they are time consuming like the Hoeffding tree algorithm and McDiarmidtree algorithm [1]. In[1] The Gaussian decision tree algorithm is proposed, which is a modification of Hoeffding tree algorithm. Which attempts to choose the best possible attribute for classification for current set of data element as the best for entire data stream.

Decision tree classifiers are very popular and widely used. They are also known as highly unstable classifiers with respect to smaller changes in training dataset. According to [4] fuzzy logic due to its elasticity of fuzzy sets can improve these aspects. Authors in [5] proposes a new sample selection method which adds the principle of maximal classification ambiguity for selecting the right samples and proves that method is superior when compare to random sample selection method. Most of the classification approaches proposed needs that the new instances to be fully labeled and there is need to construct the decision tree each time. As the classification agent does not see the dataset as whole creating decision tree the accuracy of the approach is degraded most of the times. Authors in [6] proposed approach in which classification agent decides when to restructure new model.

4.2 Rule Based Classification

Basically, in rule based classification, classification model is represented in the form of IF-THEN rules. Such rules have two part, the IF part is known as the rule antecedent and the else part is known as consequent. This structure tells that the antecedent part covers the tuples which satisfies the rule and part of consequent covers the tuples which does not satisfy the rule. Coverage and accuracy are the measures used to assess the rules. The percentage of tuple covered under rule is the coverage of that rule and percentage of correctness is the accuracy of that rule. This IF-THEN classification rules can be extracted from decision tree as this rules are easily interpreted by humans.

Many of the rule based classification techniques focus on performance of the model rather than interpretability of the data. Rule based classification method named as ROUSER mainly focuses on human understandable decision rules from data. To select an attribute value pair for If condition this method uses rough set approach and this approach is said to be more efficient than other rule based classifiers.

The fuzzy rule based classifier is difficult to deal with. This is because when number of patterns high, there is exponential growth in fuzzy rule search space. Due to this learning process becomes more difficult and scalability and complexity problems arises. In [7], Fuzzy association rule-based classification technique has been proposed for high dimensional problems. This method was intended to obtain a compact and accurate fuzzy

rule-based classification method with minimum computational cost.

4.3 K-Nearest Neighbor (KNN) Classification

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

The KNN technique is easy to implement when it is applied to small sets of data, but it gives slower performance when it is applied to large volume and high dimensional data. The value of k affects the performance of the classifier. So this algorithm is sensitive to the value of k. To overcome this difficulty in [9] authors proposed new Field Programmable Gate Array (FPGA) architecture of KNN classifier, which easily adapts the various values of k.

Accuracy in data classification is the major issue in the data mining. So that to improve it improvement are made in the knn method. Weighted nearest neighbor is one method in which weight is added to each neighbor used for classification. This method proves to increase the performance by increasing the accuracy of the classification. In [10] authors proposed K-Nearest Neighbor Mean Classifier (k-NNMC), which find k nearest neighbors for each class of training patterns. In this nearest mean patterns are used for classification. This improvement provides better accuracy of classification when compared to other techniques.

4.4 Bayesian Network

Bayesian classifier is basically used to determine whether the given tuple belongs to particular class or not, This technique is based on Bayes theorem which is measures the probability that given data tuple belongs to a particular class. When applied to large datasets Bayesian classifiers exhibits high accuracy. Bayesian classification can be Bayesian network classification or Naïve Bayesian classification. The difference between this two is later one assumes that the effect of attribute value on given class is independent of the other attributes while the former allows representation of dependencies among the subsets of attributes.

Constructing the Bayesian network classifier structure depends on whether the attribute values on a given class are independent or not. The existing methods interchange the common information to estimate the dependency among variables, which misses theoretical basis landing to low reliability. One more issue regarding building Bayesian network classifier is it is necessary to learn an accurate Bayesian network structure. For learning Bayesian network classifier K2 is considered to be most efficient algorithm. This algorithms demands variable ordering in advance to build the Bayesian structure. Existing methods don't consider the information of selected variables for classification. To overcome this difficulty author an L1 regularized Bayesian network classifier (L1-BNC) in [8]. This method defines a variable ordering by the Least Angle Regression (LARS) method. It then makes the use of this classifier with K2 to construct the Bayesian network.

4.5 Other methods of Classification

Apart from above mentioned classification methods other classical methods also present. Few of them are support vector machines (SVM), genetic algorithms, rough set, back propagation, fuzzy set and many more. This methods for classification are chosen based on the requirement of user and data classification. Some of the issues faced by this methods are reviewed below.

Foe classifying the data into two classes Support Vector machines were designed. This type of classification is known as Binary classification. Multiple research works were carried out for extending this binary classification to multiclass classification. However solving multiclass classification is computationally expensive. The SVM also used for classification of nonlinear and linear data. It transforms the data in higher dimension form, where it can hyperplane for separation of data using training tuples called support vector.

The genetic algorithms, here rough set approach and fuzzy set approaches are not used often. But sometimes their logics are used in other classification algorithms. Systems which performs rule based classification can use Fuzzy set theory. Rough sets are basically used to select attributes and reduce the features where attribute that don't contribute for a classification can be identified and removed.

Back propagation is a neural network learning algorithm and it learns by iteratively processing a data set of training tuples. For mapping tasks and simple pattern recognition Back propagation networks are ideal and they are used for data classification. One of the famous issue in back propagation is the problem encountered with the local maxima [16]. Due to growing size of the network this issue arises to overcome this many variants of back propagation were introduced.

5. CONCLUSION

There are many comparative studies over different classification methods present. But it has not been found that any single method is superior compared to others. To choose the best technique for data classification issue like accuracy, scalability, training time, complexity etc. may help. Different classification techniques have their own merits and demerits. So finding a classification technique which is superior to all is still a research topic. In this paper, we have done a comprehensive survey of the classification techniques used in data mining with their recent advances and issues which are caused by growing data sizes day by day.

REFERENCES

[1] Micheline Kamber and Jaiwei Han, 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," IEEE transaction on Knowledge and Engineering, vol. 26, Jan. 2014,

[3] M. Young, The Technical Writer's Handbook. Mill 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.

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

[5] 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

[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] 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

[8] 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

[9] 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

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

[11] R. Agrawal, T. Imielinski, and A. Swami. Database mining: A performance perspective. IEEE Trans. on Knowledge and Data Engineering, 5(6), Dec. 1993.

[12] K. Alsabti, S. Ranka, and V. Singh, "CLOUDS: A decision tree classifier for large datasets," in Proc. of 4th Intl. Conf. on Knowledge Discovery and Data Mining, Aug 1998.

[13] Carlos J. Mantas, Joaquín Abellán*, "Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data", ELSEVIER- Journal on Expert Systems with Applications, 2014, p.p. 4625-4637

[14] Rodrigo C. Barros, M´arcio P. Basgalupp, Andr´e C. P. L. F. de Carvalho and Alex A. Freitas, "A Survey of Evolutionary Algorithms for Decision Tree Induction", IEEE Transactions on System, MAN and Cybernetics,2013

BIOGRAPHIES

Sanket S. Ranaware has received Diploma in Computer Engineering from Dr. B.A.T. University, Lonere. He has completed B.E. in Computer Engineering from Mumbai University. Currently He is pursuing M.E. in Computer engineering from Pune Institute of Computer Technology, Savitribai Phule Pune University. His research interest includes Mobile Computing and Data Mining.

Dr. Girish Potdar presently working as a professor and Head of Department in the Department of Computer Engineering at Pune Institute of Computer Technology, Pune, India. His research area includes Design Analysis and Algorithm, Data Structured, System Programming.