[Sagunthaladevi, 3(7): July 2016]
DOI-10.5281/zenodo.58513

ISSN 2348 – 8034
Impact Factor- 4.022

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

CLASSIFICATION TECHNIQUES IN DATA MINING: AN OVERVIEW

Mrs.Sagunthaladevi.S[*1] and Dr.Bhupathi Raju Venkata Rama Raju[2]

[*1]Research Scholar, Dept of Computer Science, Mahatma Gandhi University, Meghalaya-793101, Tamilnadu, India.
[2]Professor, Dept of Computer Science, IEFT College of Engineering, Villupuram-605108, Tamilnadu, India.

ABSTRACT

Classification is a widely used technique in the data mining domain, where scalability and efficiency are the immediate problems in classification algorithms for large databases. Classification is a supervised learning technique in data mining where training data is given to classifier that builds with classification rules. Later if test data is given to classifier, it will predict the values for unknown attributes. Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome usually called prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. In order to predict the things correctly, attribute construction is needful for solving the problems. Hence, the motivation of this study is to propose a new attribute generation approach to extend a small data set into a purpose oriented and a higher dimension feature space to enhance the classification accuracy. This paper reviews all the data mining techniques especially classification domain for attribute construction.

Keywords: *Classification, Decision trees, Genetic Algorithms, Fuzzy sets, Support Vector Machine (SVM) Rough Sets and Neural Networks.*

I. INTRODUCTION

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the classification algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute not yet known. Algorithm analyses the input and produces a prediction. The prediction accuracy defines how 'good' the algorithm is. To build a correct classification model, sufficient amount of training data is required. But in real world, there are many situations when organizations must work with small data sets. Learning from a given data set to build a classification model becomes difficult when available sample size is small. This paper aims to develop a new attribute construction to extend the data set into higher dimensional feature space to extract more attribute information. Data quantity is the main issue in the small data set problem, because usually insufficient data will not lead to a robust classification performance. A small data set is very much a relative and subjective concept that needs to be defined.

For instance, with a classifier, it is hard to make accurate forecasts because small data sets not only make the modeling procedure to over fitting, but also cause problems in predicting specific correlations between the inputs and outputs. This paper proposes a new attribute construction approach which converts the original data attributes into a higher dimensional feature space to extract more attribute information. Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data. Many studies present ways to improve the model accuracy of small data sets analysis.

II. DATA MINING PROCESS

Basically, the data mining methods are used for extracting patterns from data. The patterns that can be discovered depend upon the data mining tasks applied. Generally, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on available data. Data mining can be done on data which are in quantitative, textual, or multimedia forms. Data mining applications can use different kind of parameters to examine the data. Data mining involves some of the following key steps-

ISSN 2348 – 8034
Impact Factor- 4.022

(1) **Problem definition:** The first step is to identify goals. Based on the defined goal, the correct series of tools can be applied to the data to build the corresponding behavioral model.

(2) **Data exploration:** If the quality of data is not suitable for an accurate model then recommendations on future data collection and storage strategies can be made at this. For analysis, all data needs to be consolidated so that it can be treated consistently.

(3) **Data preparation:** The purpose of this step is to clean and transform the data so that missing and invalid values are treated and all known valid values are made consistent for more robust analysis.

(4) **Modeling:** Based on the data and the desired outcomes, a data mining algorithm or combination of algorithms is selected for analysis. These algorithms include classical techniques such as statistics, neighborhoods and clustering but also next generation techniques such as decision trees, networks and rule based algorithms. The specific algorithm is selected based on the particular objective to be achieved and the quality of the data to be analyzed.

(5) **Evaluation and Deployment:** Based on the results of the data mining algorithms, an analysis is conducted to determine key conclusions from the analysis and create a series of recommendations for consideration.

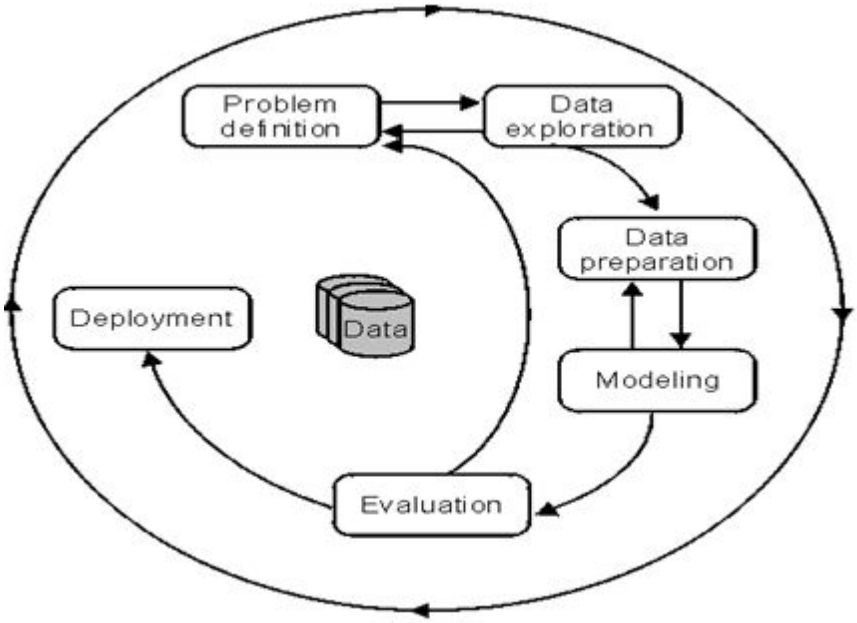

Figure: 1 Data Mining Process

III. CLASSIFICATION

Classification is a learning function that maps a given data item into one of several predefined classes. It is a data analysis technique to extract models describing important data classes and predict future values. Data mining uses classification techniques with machine learning, image processing, natural language processing, statistical and visualization techniques to discover and present knowledge in an understandable format. Most of the classification algorithms in literature are memory resident, typically assuming a small data size. Recent data mining research has built on such techniques, developing scalable and robust classification techniques capable of handling large disk - resident data. Classification has numerous applications including trajectory classification, fraud detection, target marketing, performance prediction, manufacturing and medical diagnosis.

The performance of the classification techniques is measured by the metrics like accuracy, speed, robustness, scalability, comprehensibility, time and interpretability. In the early time of a new system development,

[Sagunthaladevi, 3(7): July 2016] ISSN 2348 – 8034
DOI-10.5281/zenodo.58513 Impact Factor- 4.022
data on hand are not enough, hence, data characteristics such as data distribution, mean, and variance are unknown. As well as a decision is hard to make under the limit data condition.

IV. STEPS INVOLVED IN CLASSIFICATION PROCESS

The steps involved in classification process are as follows

- Deriving a classifier model
- Testing the derived model

Deriving a Classifier Model:

A classifier is derived using the predetermined set of data classes or labels. This is also known as training step or learning step, where a classification technique builds a classifier model using the training dataset. It can also be viewed as a mapping function to predict the associated class labels. This mapping function is usually represented in terms of decision trees, classification rules or mathematical equations.

Testing the derived model

In this stage, the trained model is used for classification of test data, and the predictive accuracy of the classifier model is estimated.

V. DATA MINING CLASSIFICATION METHODS

The data mining consists of various methods. Different methods serve different purposes, each method offering its own advantages and disadvantages. In data mining, classification is one of the most important tasks. It maps the data in to predefined targets. Classification is a form of data analysis which is used to extract models describing important data classes. It is a supervised learning process consisting of learning step and classification step. In learning step, Training data's are analyzed and in classification step test data's are used to estimate accuracy. The main aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. The commonly used methods for data mining classification tasks can be classified into the following groups.

Decision Trees

A decision tree is a tree where each non-terminal node represents a test or decision on the considered data item. It is a flowchart like tree structure which has a choice of a certain branch depends upon the outcome of the test. To classify a particular data item, we start at the root node and follow the assertions down until we reach a terminal node. A decision is made when a terminal node is approached. Decision trees can also be interpreted as a special form of a rule set, characterized by their hierarchical organization of rules.

Support Vector Machine

Support vector machine (SVM) is an algorithm that attempts to find a linear separator between the data points of two classes in multidimensional space. It is a method for the classification of both linear and nonlinear data. SVMs are well suited to dealing with interactions among features and redundant features. It can be used for numeric prediction as well as classification.

Genetic Algorithms

Genetic algorithms are easily parallelizable and it can be used for classification as well as other optimization problems. Genetic algorithms and evolutionary programming are algorithmic optimization strategies that are inspired by the principles observed in natural evolution. Of a collection of potential problem solutions that compete with each other, the best solutions are selected and combined with each other. In doing so, one expects that

ISSN 2348 – 8034
Impact Factor- 4.022

the overall goodness of the solution set will become better and better, similar to the process of evolution of a population of organisms. Genetic algorithms and evolutionary programming are used in data mining to formulate hypotheses about dependencies between variables, in the form of association rules or some other internal formalism.

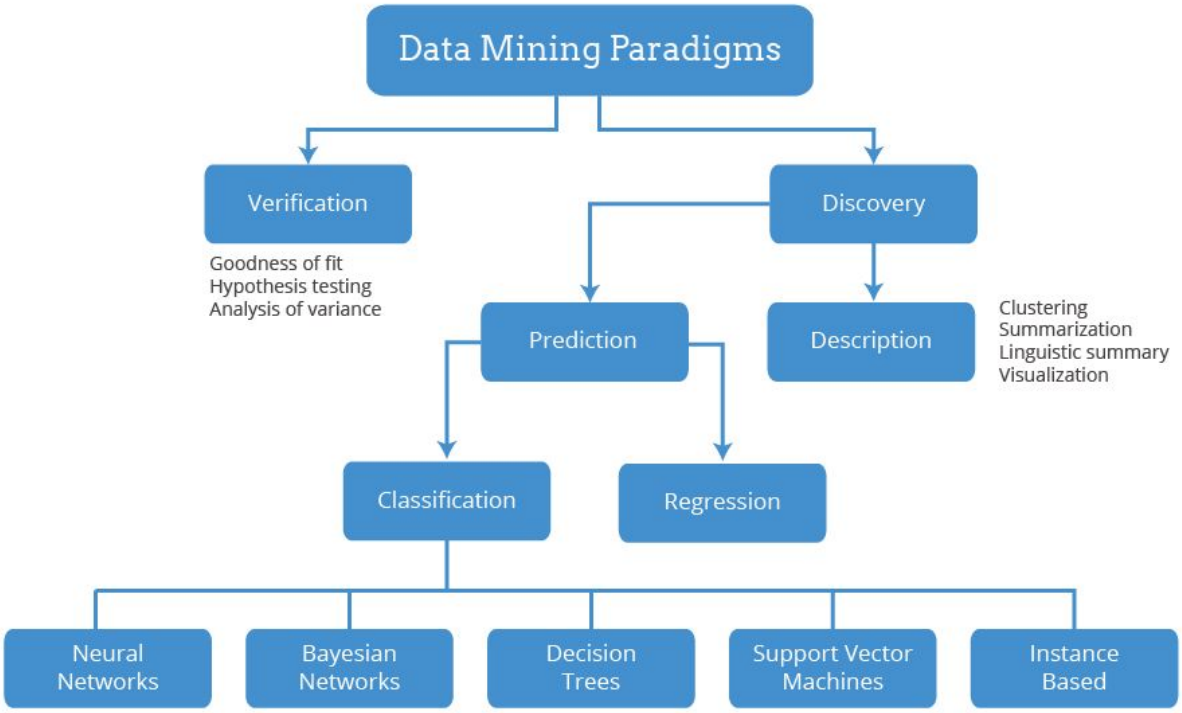

Figure 2: Steps involved in Data Mining Process

Fuzzy Sets

Fuzzy sets form a key methodology for representing and processing uncertainty. Uncertainty arises in many forms in today's databases: imprecision, non-specificity, inconsistency, vagueness, etc. Fuzzy sets exploit uncertainty in an attempt to make system complexity manageable. As such, fuzzy sets constitute a powerful approach to deal not only with incomplete, noisy or imprecise data, but may also be helpful in developing uncertain models of the data that provide smarter and smoother performance than traditional systems.

Rough Sets

Rough set can be used to discover structural relationships within imprecise or noisy data. A rough set is determined by a lower and upper bound of a set. Every member of the lower bound is a certain member of the set. Every non-member of the upper bound is a certain non-member of the set. The upper bound of a rough set is the union between the lower bound and the so-called boundary region. A member of the boundary region is possibly a member of the set. Therefore, rough sets may be viewed as with a three-valued membership function. Rough sets are a mathematical concept dealing with uncertainty in data. It applied to discrete valued attributes. If it can be used for continuous valued attributes, it must be discredited before use. They are usually combined with other methods such as rule induction, classification, or clustering methods.

ISSN 2348 – 8034
Impact Factor- 4.022

Bayes Classification

Bayes classification predicts class membership probabilities such as the probability that a given attribute belongs to a particular class. Generally, Bayes classifiers are said to be statistical classifiers. It exhibits high accuracy and speed when applied to large set of databases and it have minimum error rate in comparison to all other classifiers.

Neural Networks

Neural networks are those systems modeled based on the human brain working. As the human brain consists of millions of neurons that are interconnected by synapses, a neural network is a set of connected input/output units in which each connection has a weight associated with it. The network learns in the learning phase by adjusting the weights so as to be able to predict the correct class label of the input.

VI. CONCLUSION AND FUTURE WORK

This paper concentrated on various classification techniques used in data mining, especially in attribute construction. Every technique described in this paper has its own advantage and disadvantage. On working on performance, many attributes have been tested and some of them are found effective on the performance prediction. This paper deals with possibility of building a classification model for predicting the attribute performance. In order to increase the performance, this paper proposes a Genetic Programming algorithm which is used for attribute construction. This algorithm constructs new attributes out of the original attributes of the data set, performing an important preprocessing step for the subsequent application of a data mining algorithm. In future work, Genetic algorithm will be considered to process the attribute construction for effective and exact manner.

REFERENCES

[1] Qasem A. Al-Radaideh, Eman Al Nagi, "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 2012.

[2] Mohd. Mahmood Ali, Mohd. S. Qaseem, Lakshmi Rajamani, A. Govardhan, "Extracting Useful Rules through Improved Decision Tree Induction Using Information Entropy", International Journal of Information Sciences and Techniques (IJIST) Vol.3, No.1, January 2013.

[3] Andreas G.K. Janecek, Wilfried N. Gansterer, "On the Relationship between Feature Selection and Classification Accuracy", JMLR: Workshop and Conference Proceedings 4: 90-105, 2008.

[4]P.Niyogi,F.Girosi,andP.Tomaso,"IncorporatingPrior Information in Machine Learning by Creating Virtual Examples," Proc. IEEE, vol. 86, no. 11, pp. 2196-2209, Nov. 1998.

[5] Limère, A., Laveren, E., and Van Hoof, K. "A classification model for firm growth on the basis of ambitions, external potential and resources by means of decision tree induction", Working Papers 2004 027, University of Antwerp, Faculty of Applied Economics.

[6] Hoi, S. C., Lyu, M. R., and Chang, E. Y. (2006). "Learning the unified kernel machines for classification, In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining", pp. 187-196.

[7] Xu, J. W., Paiva, A. R., Park, I., and Principe, J. C. (2008). A reproducing kernel Hilbert space framework for information-theoretic learning, IEEE Transactions on Signal Processing, Volume 56, Issue 12, pp.5891-5902.

[8] Shilton, A., and Palaniswami, M. (2008). "A Unified Approach to Support Vector Machines", In B. Verma, & M. Blumenstein (Eds.), Pattern Recognition Technologies and Applications: Recent Advances, pp. 299-324.

[Sagunthaladevi, 3(7): July 2016] ISSN 2348 – 8034
DOI-10.5281/zenodo.58513 Impact Factor- 4.022
[9] Kumar, A., Niculescu-Mizil, A., Kavukcuoglu, K., and Daume III, H. (2012). "A binary classification framework for two-stage multiple kernel learning". arXiv preprint arXiv:1206.6428, Appears in Proceedings of the 29th International Conference on Machine Learning.

[10] Takeda, A., Mitsugi, H., and Kanamori, T. (2012). "A unified robust classification model",arXiv preprint arXiv:1206.4599. J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publish, 2001

[11] C. Kim and C.H. Choi, "A Discriminant Analysis Using Composite Features for Classification Problems," Pattern Recognition, vol. 40, no. 11, pp. 2958-2966, 2007

[12] Raj Kumar, Dr. Rajesh Verma, "Classification Algorithms for Data Mining: A Survey", International Journal of Innovations in Engineering and Technology (IJIET), Vol. 1 Issue 2 August 2012, ISSN: 2319 – 1058, pg: 7-14

[13] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[14] Data Mining Concepts and Techniques, Third Edition ISBN: 978-0-12-381479-1, Morgan Kaufmann Publishers, 225Wyman Street, MA 02451 (USA), 2012.