SURVEY ON CLASSIFICATION METHODS IN DATA MINING

Mrs.Sagunthaladevi.S
Research Scholar, Dept of Computer Science
Mahatma Gandhi University
Meghalaya-793101, India
E-Mail:sagunthaladevis@gmail.com

Dr.Bhupathi Raju Venkata Rama Raju
Professor, Dept of Computer Science
IEFT College of Engineering
Villupuram-605108
Tamilnadu, India

Abstract — Data mining is the analysis step of the "Knowledge Discovery in database" process and it is an interdisciplinary computational process which is used to discover patterns in large datasets involving methods at the intersection of artificial brainpower, machine learning, figures and relevant data and database systems. Classification is a data mining technique based on machine learning which is used to classify each item in a set of data into a set of predefined classes or groups. It is a model finding process that is used for portioning the data into different classes according to some constrains. The accuracy of the classification result will more for the new dataset. This paper provides an inclusive survey of different classification algorithms such as k-nearest neighbor classifier, Naive Bayes, SVM, Apriori, C4.5 and also mention their advantages and disadvantages.

Keywords -- Data Mining; Data Mining Methods; Classification; Support Vector Machines (SVM); Knowledge Discovery in Database (KDD); C4.5 and Feature Extraction.

I. INTRODUCTION

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 and it maps the data in to predefined targets. It is a supervised learning as targets are predefined. The 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. A small data set sizes are very important in machine learning problems, because too few samples will contain incomplete information. For instance, with a classifier, it is hard to make accurate forecasts because small data sets not only make the modeling procedure prone to over fitting, but also cause problems in predicting specific correlations between the inputs and outputs.

A. The Knowledge Discovery Process:

The terms Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably. KDD is the process of turning the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are often treated as equivalent words but in real data mining is an important step in the KDD process. The following figure shows data mining as a step in an iterative knowledge discovery process.

B. Feature Extraction:

Feature extraction is a technique which has the capability to project the original features into a lower feature space to reduce the number of data dimensions and improve analytical efficiency. Generally, there are two steps included in the feature extraction method.

- First, the relevant information for classification is extracted from raw data with the original feature vector, m dimensions.
- Second, a new feature vector with n dimensions (n<m) is created from the parameter vector.

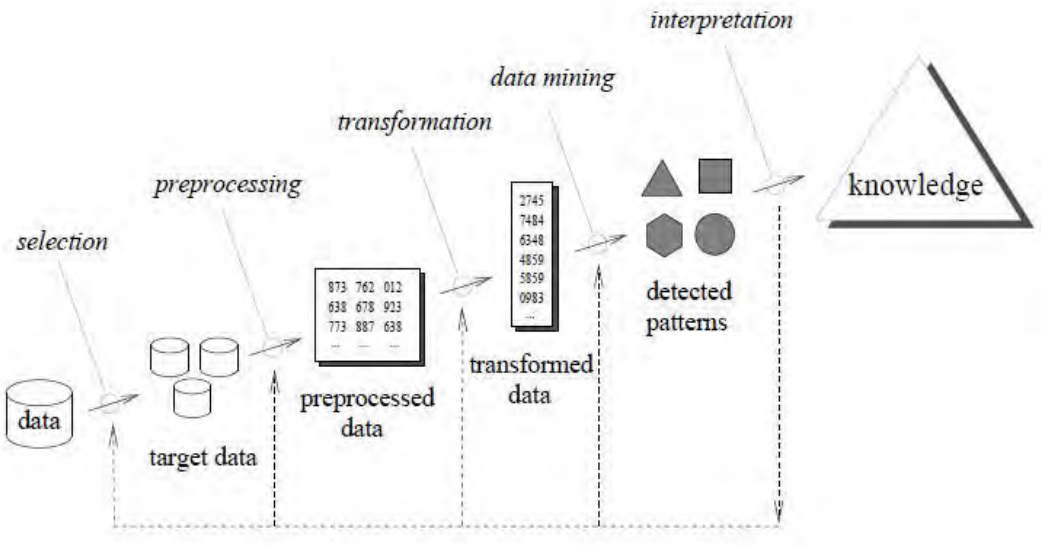

Figure 1: Knowledge Discovery in Database Process

The commonly used methods of feature extraction include principal component analysis, kernel independent component analysis (KICA), canonical correlation analysis (CCA), and kernel principal component analysis (KPCA). Based on the type of transformation function, feature extraction techniques can be classified into two types: linear and nonlinear.

- Linear methods, such as principal component analysis, reduce dimensionality by performing linear transformations on the input data and find the globally defined flat subspace. These methods are most effective if the input patterns are distributed more or less throughout the subspace.
- Nonlinear methods, such as KPCA, try to find the locally defined flat subspace by nonlinear transformation when the structure of the input data is highly nonlinear.

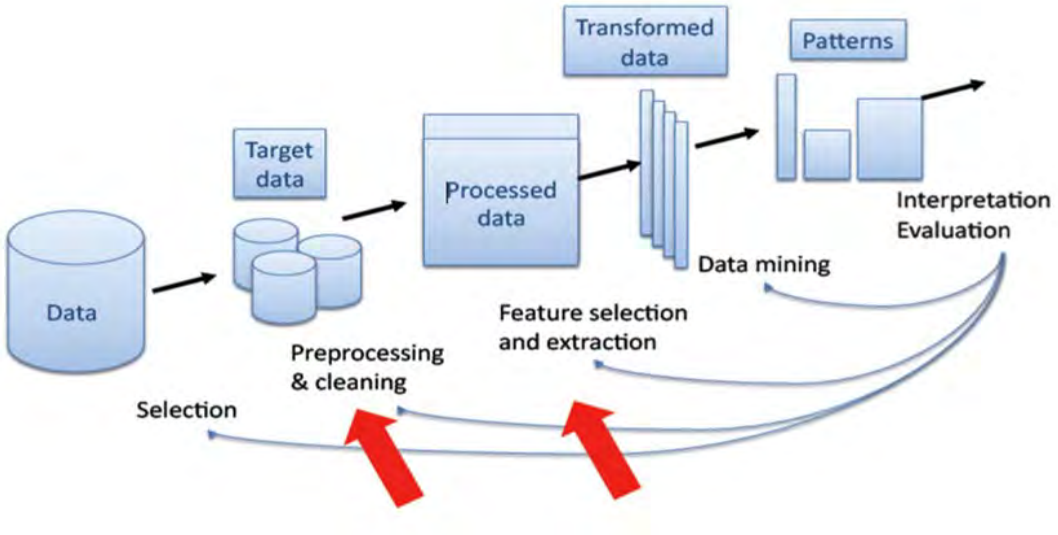

Figure 2: Feature Extraction Process

II. LITERATURE REVIEW

Classification technique is based on the inductive learning principle that analyzes and finds the patterns from the database. If the nature of an environment is dynamic, then the model must be adaptive i.e. it should be able to learn and map efficiently.

A model has been presented by Limère et al. [2004] for firm growth with decision tree induction principle and it gives interesting results and fits the model to economic data like growth competence and resources, growth potential and growth ambitions. Growths ambitions are also positively related to firm growth.

They conclude, if the average profitability is above a certain percentage, firms are classified as strongly growing firms.

A novel framework of learning the unified kernel machines for both labeled and unlabeled data developed by Hoi et al. [2006]. This framework includes semi supervised learning, supervised learning and active learning. Also, a spectral kernel is proposed, where it classifies the given labeled data and unlabeled data efficiently. Though the kernel methods have many interesting features, reducing the training time and classification time are the two major issues concerned with practitioners and researchers. To speed up the training time and classification performance, many techniques have been proposed in the literature. Support Vector Machine is originally used to symbolize popular and modern classifiers that have a well-defined theoretical foundation to provide some enviable performances.

Reproducing kernel Hilbert space framework for information theoretic learning was proposed by Xu et al. [7]. The framework uses the symmetric nonnegative definite kernel function i.e. cross information potential. Though this framework gives better result than the previous RKHS frameworks, still there is an issue to choose an appropriate kernel function for a particular domain.

Shilton and Palaniswami [8] defined a unified approach to support vector machines. This unified approach is formulated for binary classification and later on extended to one - class classification and regression. Some of the techniques that have been proposed to speed up the training time are sequential minimal optimization, modified sequential minimal optimization, decomposition method and low rank kernel matrix construction method.

The classification time of SVM primarily depends on the number of Support Vectors (SVs) involved in the system. So, it is necessary to minimize the number of support vectors that can improve the efficiency and minimize the computation time of the classification process.

Kumar et al. [9] explored a binary classification framework for two stage multiple kernel learning. The distinct advantage of this binary classification framework is that it is easier to leverage research in binary classification and to develop scalable and robust kernel based algorithms. However, kernel methods are processed by operations to the kernel function (such as Gaussian and polynomial kernels) for the data, ignoring both the structure of the input data and the dimensionality problem, and thus cannot always guarantee that the transformed space is useful for classification. The commonly used kernels are the so-called all-function or general purpose ones, such as the Gaussian and polynomial.

Takeda et al. [10] proposed a unified robust classification model that optimizes the existing classification models like SVM, Min-Max probability machine and fisher discriminant analysis. It provides several benefits like well - defined theoretical results extends the existing techniques and clarifies relationships among existing models. Basically, Support vector machines (SVM) are considered as a must try it offers one of the most robust and accurate methods among all well-known algorithms. It has a sound theoretical foundation, requires only a dozen examples for training, and is insensitive to the number of dimensions. In addition, efficient methods for training SVM are also being developed at a fast pace. In a two-class learning task, the aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data. The metric for the concept of the best classification function can be realized geometrically.

Raj Kumar, Dr. Rajesh Verma [11] viewed the pattern classification as an ill - posed problem, it is a prerequisite to develop a unified theoretical framework that classifies and solves the ill posed problems. Recent literature on classification framework has reported better results for binary class datasets alone. For multiclass datasets, there is a lack in accuracy and robustness. So, developing an efficient classification framework for multiclass datasets is still an open research problem.

Table: 1 Advantage and Disadvantages of Classification Algorithms

Algorithm	Advantages	Disadvantages
Naive Bayes Algorithm	• Improves performance by filtering unwanted data • Performance wise good and also takes less time for processing	• Stores all sample data's whatever it takes • Classifier requires more no. of datasets to produce good results
k-Nearest Neighbour	• Easy to implement • Fits for multi model classes	• Memory limitation • Slow running time

Algorithm	• Algorithm is simplest when compare to other algorithms	
Support Vector Machine Algorithm	• Accurate classifier with less over fitting • Less Memory usage	• Slow running time • Expensive
C4.5 Algorithm	• Respond time is short • Shows accurate results	• Insufficient and empty branches

III CONCLUSION

This Survey deals with various classification techniques used in data mining and a study on each of them. Data mining is a wide area that integrates techniques from various fields such as machine learning, artificial intelligence, statistics and pattern recognition. Classification methods are typically strong in modeling interactions. Hence these classification methods show that how a data can be determined and grouped when a new set of data is available. Each technique has got its own Advantages and Disadvantages as given in the paper. Compare to k-nearest neighbors, Decision trees and Bayesian Network (BN) generally have different operational profiles, when one is very accurate the other is not and vice versa. The role of classification is to generate more precise and accurate system results. Constructing new attributes provides a better and faster data classification. The accuracy of the classification result will more for the new dataset.

IV REFERENCE

[1] S.Archana, Dr. K.Elangovan, "Survey of Classification Techniques in Data Mining", International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February-2014, ISSN: 2321-8363,pg: 65-71

[2] M.Soundarya, R.Balakrishnan, "Survey on Classification Techniques in Data mining", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 7, July 2014, ISSN: 2278-1021, pg: 7550-7552

[3] DelveenLuqman AbdAl.Nabi, ShereenShukri Ahmed, "Survey on Classification Algorithms for Data Mining: (Comparison and Evaluation)" (ISSN 2222-2863) Vol.4, No.8, 2013

[4] NeelamadhabPadhy, Dr. Pragnyaban Mishra, and RasmitaPanigrahi, "The Survey of Data Mining Applications and Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012, pg: 43-58

[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.

[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] 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

[12] 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.