World Wide Journal of Multidisciplinary Research and Development

WWJMRD 2016; 2(7): 57-59
www.wwjmrd.com
Impact Factor MJIF: 4.25
e-ISSN: 2454-6615

Sapna Mittal
Assistant Professor in Computer Science Post Graduate Deptt of Computer Science and Applications R.S.D College, Ferozepur City, India

Correspondence:
K. Sivaiah
Assistant Professor in Computer Science Post Graduate Deptt of Computer Science and Applications R.S.D College, Ferozepur City, India

A Review on Various Approaches of Classification Based On Attributes Selection in Data Mining

Sapna Mittal

Abstract
Data mining is the process of extraction of information from various datasets on the basis of different attributes. Mining has to be done to extract hidden relationship between various database entities. On the basis of these entities, different types of decisions are taken for the extraction of different relationships. In the customer relationship management, different relational attributes are available in the dataset. This dataset contains the information about the relations of the customer with an enterprise. The dataset has to be classified using rules for extraction of information. Mainly Churn, appetency, up selling and score are the major entities which will be considered in the proposed work. To overcome the problems of CRM database a new hybrid algorithm is introduced which will be the combination of GA and Fuzzy KNN classification.

Keywords: Data Mining, Fuzzy KNN, SVM, CRM, Classification.

Introduction

1.1 Data Mining

Data mining is crucial for extracting and identifying useful information from a large amount of data that is why retailing companies operate purchase databases in a long way, such that all transactions are stored in arranged order. A record-of-transaction database typically contains the transaction date and the products bought in the course of a given transaction. Usually, each record also contains shopper ID, particularly when the purchase was made using a credit card or a frequent buyer card. Therefore, the purchasing sequence of an e-shopper in the database that has made repeated purchase can easily be determined. This purchase sequence provides a description of the changes in an e-shopper's preferences over time, because purchase sequence can reveal the changes of e-shopper' preferences over time.

1.2 Customer Relationship Management (CRM)

CRM is the core business strategy that integrates internal processes and functions of the organization, to create and deliver value to targeted customers at a profit. It is mainly grounded on high quality customer related data and enabled by information technology. CRM is an information industry term that helps an enterprise to manage customer relationships in an organized way and helping the company to provide better services to its customers. CRM is the process of managing all aspects of interaction of the company with its existing customers and the new customers, including prospecting, sales and service. CRM applications try to provide insight into and help in improving the company/customer relationship by combining all these views of customer interaction into one picture.

1.3 Key Phases In Data Mining Process

1.3.1 Information association: The one of the most familiar and straightforward feature of this system is that here we made association between two or more items or often of the same type to formulate particular pattern. Like it is very well known etiological association between smoking and lung cancer. We have to collect data concerned with smoking habit details including numbers of smoke per day, duration of smoking, type of smoking either bid is, cigarettes, specific brands, lifestyle and age of patient Etc.

1.3.2 Information classification: This is the second phase in this we can classify the collected information according to our objectives like etiological factors, investigation purpose, drug treatment plans and results.

For example the etiological information collected from lung cancer patients can be classified on the basis of duration of smoking habit, type of exposure, number of exposure, age of patient etc.
Information clustering it is easy to locate data trend if we put collected data in the form of cluster by using computer graphics. These identified trends may helpful in preparation of prediction system [3].

1.3.3 Pattern Sequencing This is the next step in module preparation. The pattern sequencing can be prepared with the help of readymade software packages available in market.

1.3.4 Preparation of decision tree: this is final step of prediction system.

1.3.5 Implementation: This is directly concerned with last step. You may have option either long term or short term data processing. Each data mining system has their different objectives. Data mining process are broadly formulated either as supervised or unsupervised learning. Supervised learning is that type of learning in which a training set is used to learn model parameters but in Unsupervised learning no training set is used.

1.4 Types of CRM
1.4.1 Operational CRM The primary goal of CRM systems is to integrate and automate sales, marketing, and customer support. Therefore, these systems typically have a dashboard that gives an overall view of the three functions on a single page for each customer that a company may have. The dashboard may provide client information, past sales, previous marketing efforts, and more, summarizing all of the relationships between the customer and the firm. Operational CRM is made up of 3 main components: sales force automation, marketing automation, and service automation.

1.4.2 Analytical CRM: The role of analytical CRM systems is to analyze customer data collected through multiple sources, and present it so that business managers can make more informed decisions. Analytical CRM systems use techniques such as data mining, correlation, and pattern recognition to analyze the customer data. These analytics help improve customer service by finding small problems which can be solved, perhaps, by marketing to different parts of a consumer audience differently. For example, through the analysis of a customer base's buying behavior, a company might see that this customer base has not been buying a lot of products recently. After scanning through this data, the company might think to market to this subset of consumers differently, in order to best communicate how this company's products might benefit this group specifically.

2. Review of literature

S.Ummugulthum Natchiar et al [1] "Customer Relationship Management Classification Using Data Mining Techniques" conclude Customer Relationship Management possess Business Intelligence by incorporating information acquisition, information storage, and decision support functions to provide customized customer service. It enables customer representatives to analyze and classify data to address customer needs in order to promote greater customer satisfaction and retention, but in reality we have learned CRM classification models are outdated , substandard because of noisy and imbalanced data set. In this paper, a new feature selection method is proposed to resolve such CRM data set with relevant features by incorporating an efficient data mining techniques to improve data quality and feature relevancy after preprocessing. Finally it enhances the performance of classification.
Nedaabdelhamid et al [2] "Emerging trends in associative classification data mining" studied emerging trends in associative classification in data mining. Utilising association rule discovery to learn classifiers in data mining is known as associative classification. In the last decade AC algorithms proved to be effective in devising high accurate classification system from various types of supervised datasets. Yet, there are new emerging trends and that can further enhance the performance of current ac method or necessitate the development of new methods. This paper sheds the light on four possible new research trends within AC that could enhance the predictive performance of the classifier or their quality in terms of rules. These possible research directions are considered starting research points for other scholar in rule based classification in data mining.
Sankaranarayanan, S. et al [3] "Diabetic Prognosis through Data Mining Methods and Techniques" concluded a-priori and FP-growth are used for application to diabetes dataset. Data mining now-a-days plays an important role in prediction of diseases in health care industry. Data mining is the process of selecting, exploring, and modelling large amounts of data to discover unknown patterns or relationships useful to the data analyst. Medical data mining has emerged impeccable with potential for exploring hidden patterns from the data sets of medical domain. These patterns can be utilized for fast and better clinical decision making for preventive and suggestive medicine. However raw medical data are available widely distributed, heterogeneous in nature and voluminous for ordinary processing. Data mining and Statistics can collectively work better towards discovering hidden patterns and structures in data. In this paper, two major Data Mining techniques v.i.z., FP-Growth and A-priori have been used for application to diabetes dataset and association rules are being generated by both of these algorithms.
Wang, Guoyin et al [4] "Granular computing based data mining in the views of rough set and fuzzy set" described data mining is performed at granual level using rough set as fuzzy sets. Data mining is considered as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. In our data-driven data mining model, knowledge is originally existed in data, but just not understandable for human. Data mining is taken as a process of transforming knowledge from data format into some other human understandable format like rule, formula, theorem, etc. In order to keep the knowledge unchanged in a data mining process, the knowledge properties should be kept unchanged during a knowledge transformation process. Many real world data mining tasks are highly constraint-based and domain-oriented. Thus, domain prior knowledge should also be a knowledge source for data mining. The control of a user to a data mining process could also be taken as a kind of dynamic input of

the data mining process. Thus, a data mining process is not only mining knowledge from data, but also from human. This is the key idea of Domain- oriented Data-driven Data Mining (3DM).

Tzung-Pei Hong et al [5] "Using divide-and-conquer GA strategy in fuzzy data mining" investigated that Data mining is most commonly used in attempts to induce association rules from transaction data. Transactions in real-world applications, however, usually consist of quantitative values. This work thus proposes a fuzzy data-mining algorithm for extracting both association rules and membership functions from quantitative transactions. A GA-based framework for finding membership functions suitable for mining problems is proposed. The fitness of each set of membership functions is evaluated using the fuzzy-supports of the linguistic terms in the large 1-itemsets and the suitability of the derived membership functions. The proposed framework thus maintains multiple populations of membership functions, with one population for one item's membership functions. The final best set of membership functions gathered from all the populations is used to effectively mine fuzzy association rules.

3. Approaches used

K-nearest neighbor classifiers: K-nearest neighbor classifiers (see e.g. Hand et al. (2002) [347-352]) represent a totally different approach to classification. They do not build any explicit global model, but approximate it only locally and implicitly. The main idea is to classify a new object by examining the class values of the K most similar data points. The selected class can be either the most common 21 class among the neighbors or a class distribution in the neighborhood. The only learning task in K-nearest neighbor classifiers is to select two important parameters: the number of neighbors K and distance metric d. An appropriate K value can be selected by trying different values and validating the results in a separate test set. When data sets are small, a good strategy is to use leave–one–out cross-validation. If K is fixed, then the size of the neighborhood varies. In sparse areas the nearest neighbors are more remote than in dense areas. However, defining different Ks for different areas is even more difficult.

Support vector machines Support: vector machines (SVMs) (Vapid, 1998) are an ideal method, when the class boundaries are non-linear but here is too little data to learn complex non-linear models. The underlying idea is that when the data is mapped to a higher dimension, the classes become linearly separable. In practice, the mapping is done only implicitly, using kernel functions. 23 SVMs concentrate on only the class boundaries; points which are any way easily classified, are skipped. The goal is to find the "thickest hyper plane" (with the largest margin), which separates the classes. Often, better results are achieved with "soft" margins, which allow some misclassified data points. When the optimal margin is determined, it is enough to save the support vectors, i.e. data points which define the class boundaries. The main advantage of SVMs is that they find always the global optimum, because there are no local optima in maximizing the margin. Another benefit is that the accuracy does not depend on the dimensionality of data and the system is very robust to over fitting.

Conclusion

Data mining is the process for extraction of different data values on the basis of features. In this process different approaches have been used for classification on the basis of features. The different attributes have been used for classification on the basis distance between dataset attributes of training data and testing data. In this process these approaches measure distance from the test samples to the each training samples and on the basis of the distance these test samples have been labels with a single class label. In the purposed work the fuzzy based membership function has been used for classification this approach assigns the weight age fir dataset attributes on the basis of fuzzy membership rules. After the assignment of weight age the data set distance has been computed using Euclidian classifier and the function has been used that use both weight age and distance factor for prediction of class label to a single dataset.

Refrences

1. S.Ummugulthum Natchiar "Customer Relationship Management Classification Using Data Mining Techniques", International Conference on Science, Engineering and Management Research, 2014, pp 223-234.
2. Nedaabdelhamid, Aladdin Ayesh and FadiThabtah "Emerging trends in associative classification data mining" International journal of electronics and electrical engineering Volume 3, Issue 1, Feb 2015.
3. Sankaranarayanan, S. "Diabetic Prognosis through Data Mining Methods and Techniques", International Conf. on Intelligent Computing Applications (ICICA), 2014, pp. 162 – 166.
4. Wang, Guoyin "Granular computing based data mining in the views of rough set and fuzzy set" IEEE Conf. on Granular Computing, 2008, pp. 67.
5. Tzung-Pei Hong "Using divide-and-conquer GA strategy in fuzzy data mining" IEEE Conf. on Computers and Communications, 2004, pp. 116 - 121 Vol.1.
6. Tzung-Pei Hong "GA-based item partition for data mining" IEEE Conf. on Systems, Man, and Cybernetics (SMC), 2011, pp. 2238 – 2242.
7. Jo-Ting Wei "Customer relationship management in the hairdressing industry: An application of data mining techniques", IEEE Conf. on Expert Systems with Applications, 2013, pp Pages 7513–7518.
8. Wen-Yu Chiang "Applying data mining with a new model on customer relationship management systems: a case of airline industry in Taiwan", Conf. on Data Mining, 2014, pp 89-97.
9. Alexander Tuzhilin "Customer relationship management and Web mining: the next frontier", Springer conf. on CRM & WM, 2012, pp 584-612.
10. Siavash Emtiyaz "Customers Behavior Modeling by Semi-Supervised Learning in Customer Relationship Management", Advances in information Sciences and Service Sciences (AISS), 2011, PP 56-67.
11. Shu-Hsien Liao "Data mining techniques and applications – A decade review from 2000 to 2011", Expert Systems with Applications, 2012, PP 11303–11311.
12. Farnoosh Khodakarami "Exploring the role of customer relationship management (CRM) systems in customer knowledge creation", Conf. on CRM, 2014, PP 56-70.