International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Data Mining, Its Issues, and Approaches: A Survey

Simarpreet Kaur[1], Jyoti Arora[2]

[1]Student, Desh Bhagat University

[2]Assistant Professor, Desh Bhagat University

Abstract: ***Data mining is the process of extraction of information from various datasets on the basis of different attributes. 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. On the basis of these values features have been selected from database. SVM, Naïve bayes, J48 has been used for the classification of database.***

Keywords: Data Mining, CRM, SVM, Naïve bayes

1. Introduction

Customer Relationship Management (CRM) has become one of centre point for many industries such as Banking, Retail, Telecommunication, and Insurance. True to the saying "Customer is the King", has now been made possible .CRM takes customer as the focal point and optimizes the business process. But in the real-world application there are major challenges for building high performance CRM classification models. Since data quality is a significant issue for CRM classifications in that various types of data anomaly complicate the data preparation and classification methods. It is difficult to find one methodology that rectifies all data mining problems in the CRM data set such as High dimensional, Heterogeneous, Severe data anomaly and Imbalanced.

Normally the data set is not having all the data because of missing data by reluctant clients who do not furnish all information, misconception and human errors.

Figure 1.1: CRM

High dimensional data may contain large number of redundant and irrelevant information which might affect the performance of learning algorithms. Therefore, feature selection becomes very important for machine learning tasks. Heterogeneous data is collected from any number of sources, mainly unknown and unlimited, and in many different formats either numeric or nominal.

1.1 Issues in Data Mining

Data mining has evolved into an important and active area of research because of the theoretical challenges and practical applications associated with the problem of discovering interesting and previously unknown knowledge from real-world databases. The main challenges to the data mining and the corresponding considerations in designing the algorithms are as follows:

1) Massive datasets and high dimensionality.
2) Over fitting and assessing the statistical significance.
3) Understanding of patterns.
4) Non-standard incomplete data and data integration
5) Mixed changing and redundant data [5].

1.2 What Kind of Information are We Collecting

We have been collecting a myriad of data, from simple numerical measurements and text documents, to more complex information such as spatial data, multimedia channels, and hypertext documents. Here is a non-exclusive list of a variety of information collected in digital form in databases and in flat files.

- **Business transactions**: Every transaction in the business industry is (often) "memorized" for perpetuity. Such transactions are usually time related and can be inter-business deals such as purchases, exchanges, banking, stock, etc., or intra business operations such as management of in-house wares and assets. Large department stores, for example, thanks to the widespread use of bar codes, store millions of transactions daily representing often terabytes of data. Storage space is not the major problem, as the price of hard disks is continuously dropping, but the effective use of the data in a reasonable time frame for competitive decision- making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world.
- **Scientific data**: Whether in a Swiss nuclear accelerator laboratory counting particles, in the Canadian forest studying readings from a grizzly bear radio collar, on a South Pole iceberg gathering data about oceanic activity, or in an American university investigating human psychology, our society is amassing colossal amounts of scientific data that need to be analyzed. Unfortunately, we can capture and store more new data faster than we can analyze the old data already accumulated [6].
- **Medical and personal data:** From government census to personnel and customer files, very large collections of information are continuously gathered about individuals and groups. Governments, companies and organizations such as hospitals, are stockpiling very important quantities of personal data to help them manage human resources,

better understand a market, or simply assist clientele. Regardless of the privacy issues this type of data often reveals, this information is collected, used and even shared.

- **Surveillance video and pictures**: With the amazing collapse of video camera prices, video cameras are becoming ubiquitous. Video tapes from surveillance cameras are usually recycled and thus the content is lost. However, there is a tendency today to store the tapes and even digitize them for future use and analysis [7].
- **Satellite sensing:** There is a countless number of satellites around the globe: some are geo-stationary above a region, and some are orbiting around the Earth, but all are sending a non-stop stream of data to the surface. NASA, which controls a large number of satellites, receives more
- data every second than what all NASA researchers and engineers can cope with. Many satellite pictures and data are made public as soon as they are received in the hopes that other researchers can analyze them.
- **Games:** Our society is collecting a tremendous amount of data and statistics about games, players and athletes. From hockey scores, basketball passes and car-racing lapses, to swimming times, boxer's pushes and chess positions, all the data are stored. Commentators and journalists are using this information for reporting, but trainers and athletes would want to exploit this data to improve performance and better understand opponents.

2. Literature Review

S.Ummugulthum Natchiar et al [1] In this paper "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 (2015) [2] in this paper "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 (2014) [3] in this paper "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 (2008) [4] in the paper " 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 (2004) [5] in the paper "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.

Tzung-Pei Hong et al (2011) [6] in his paper "GA-based item partition for data mining" explained when a mining procedure is directly executed on very large databases; the computer memory may not allow the processing in memory. In the past, we adopted a branch-and-bound search strategy to divide the domain items as a set of groups. Although it works well in partitions the items, the time is quite time consuming. In this paper, we thus propose a GA-based approach to speed up the partition process. A new encoding representation and a transformation scheme are designed to help the search process. Experimental results also show that the algorithm can get a proper partition with good efficiency.

3. Approaches Used

3.1 Genetic Algorithm

Genetic algorithms are inspired by Darwin's theory about evolution. Solution to a problem solved by genetic algorithms is evolved. Algorithm is started with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population. This is motivated by a hope, that the new population will be better than the old one. Solutions which are selected to form new solutions (offspring) are selected according to their fitness - the more suitable they are the more chances they have to reproduce [8].

This is repeated until some condition (for example number of populations or improvement of the best solution) is satisfied.
Outline of the Basic Genetic Algorithm

- [Start] Generate random population of n chromosomes (suitable solutions for the problem)
- [Fitness] Evaluate the fitness f(x) of each chromosome x in the population
- [New population] Create a new population by repeating following steps until the new population is complete
- [Selection] Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)
- [Crossover] With a crossover probability cross over the parents to form a new offspring (children). If no crossover was performed, offspring is an exact copy of parents.
- [Mutation] With a mutation probability mutate new offspring at each locus (position in chromosome) [9].
- [Accepting] Place new offspring in a new population
- [Replace] Use new generated population for a further run of algorithm
- [Test] If the end condition is satisfied, stop, and return the best solution in current population
- [Loop] Go to step 2

3.2 Fuzzy KNN (K nearest Neighbour)

A "Fuzzy KNN" algorithm utilizes strength of test sample into any class called fuzzy class membership and thus produces fuzzy classification rule. An example of Fuzzy set is the set of real numbers much larger than zero, which can be defined with a membership function as follows [10]:

Numbers less than zero are not in the set because value of membership function for those is zero. While numbers larger than zero are in the set based on strength of numbers with respect to zero. This makes Fuzzy Set a useful tool for classification of samples having imprecise boundary. Fuzzy Set gives degree of presence of any sample into specific class.

4. Conclusion

Data mining is the process of extraction of information from various datasets on the basis of different attributes.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. In our work we will design a classifier for optimal data mining of structured dataset. Then Fuzzy KNN algorithm for classification of dataset & Genetic Algorithm to optimize the classification of Fuzzy KNN (Fuzzy K –Nearest Neighbor) is implemented.

References

[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.
[13] E.W.T. Ngai "Application of data mining techniques in customer relationship management: A literature review and classification" Expert Systems with Applications , ELSEVIER, Volume 36, Issue 2, Part 2, March 2009, Pages 2592–2602.
[14] Mikhailov, L "Method for fuzzy rules extraction from numerical data" IEEE Conf. on Intelligent Control, 1997, pp 61 – 66.
[15] Robert E. Marmelstein "Application of Genetic Algorithms to Data Mining" MAICS-97 Proceedings, 1997 AAAI , pp. 53-57