www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 4 April 2016, Page No. 16257-16262

Comparison between Various Approaches for Customer Relationship Management in Data Mining

Surabhi Aggarwal[1], Er. Neena Madan[2]

[1] Student
Guru Nanak Dev University, RC Jalandhar
Email: Surabhiaggarwal.1111@gmail.com

[2] Faculty of computer science
Guru Nanak Dev University, RC Jalandhar
Email: nmadan70@rediffmail.com

Abstract- ***In this paper various techniques used for CRM in data mining are defined and compared with each other. Data mining is a useful and powerful tool for any organization especially for marketing people. Data mining is used in managing relationships with customers. Data mining process can be extremely useful for Medical practitioners for extracting hidden medical knowledge. It would otherwise be impossible for traditional pattern matching and mapping strategies to be so effective and precise in prognosis or diagnosis without data mining techniques.***

Keywords: CRM dataset, Fuzzy KNN, Genetic algorithm, K-means clustering.

1. INTRODUCTION

1.2 DATA MINING

Data mining process can be extremely useful for Medical practitioners for extracting hidden medical knowledge. It would otherwise be impossible for traditional pattern matching and mapping strategies to be so effective and precise in prognosis or diagnosis without data mining techniques. This work aims at correlating various diabetes input parameters for efficient classification of Diabetes dataset and onward to mining useful patterns. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare systems too. Data preprocessing and transformation is required before one can apply data mining to clinical data. Knowledge discovery and data mining is the core step, which results in discovery of hidden but useful knowledge from massive databases.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system, neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

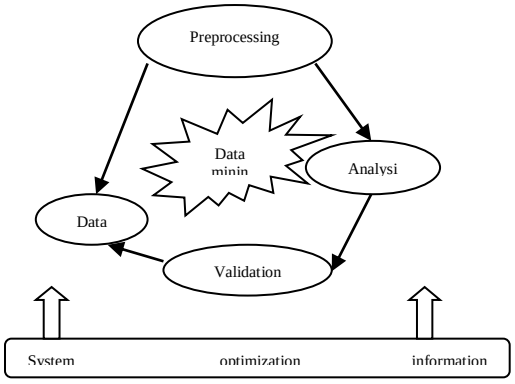

Figure 1: Data Mining

1.2 APPLICATIONS OF DATA MINING

1.2.1 Data Mining Applications in Sales/Marketing

Data mining enables businesses to understand the hidden patterns inside historical purchasing transaction data, thus helping in planning and launching new marketing campaigns in prompt and cost effective way. The following illustrates several data mining applications in sale and marketing.

- Data mining is used for market basket analysis to provide information on what product combinations were purchased together, when they were bought and in what sequence. This information helps businesses promote their most profitable products and maximize the profit. In

addition, it encourages customers to purchase related products that they may have been missed or overlooked.

- Retail companies' uses data mining to identify customer's behavior buying patterns.

1.2.2 Data Mining Applications in Banking / Finance

- Several data mining techniques e.g., distributed data mining have been researched, modeled and developed to help credit card fraud detection.
- Data mining is used to identify customer's loyalty by analyzing the data of customer's purchasing activities such as the data of frequency of purchase in a period of time, total monetary value of all purchases and when was the last purchase. After analyzing those dimensions, the relative measure is generated for each customer. The higher of the score, the more relative loyal the customer is.
- To help bank to retain credit card customers, data mining is applied. By analyzing the past data, data mining can help banks predict customers that likely to change their credit card affiliation so they can plan and launch different special offers to retain those customers.
- Credit card spending by customer groups can be identified by using data mining.
- The hidden correlations between different financial indicators can be discovered by using data mining.
- From historical market data, data mining enables to identify stock trading rules.

1.2.3 Data Mining Applications in Health Care and Insurance

The growth of the insurance industry entirely depends on the ability of converting data into the knowledge, information or intelligence about customers, competitors and its markets. Data mining is applied in insurance industry lately but brought tremendous competitive advantages to the companies who have implemented

- Data mining is applied in claims analysis such as identifying which medical procedures are claimed together.
- Data mining enables to forecasts which customers will potentially purchase new policies.
- Data mining allows insurance companies to detect risky customers' behavior patterns.
- Data mining helps detect fraudulent behavior.

1.2.4 Data Mining Applications in Transportation

- Data mining helps determine the distribution schedules among warehouses and outlets and analyze loading patterns.

1.2.5 Data Mining Applications in Medicine

- Data mining enables to characterize patient activities to see incoming office visits.
- Data mining helps identify the patterns of successful medical therapies for different illnesses.

1.2 HEART DISEASE PREDICTION

Medical data mining has high potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis for widely distributed in raw medical data which is the above discussion; it is regarded as the primary reason behind deaths in adults. Heart disease kills one person every 34 seconds in the United States. The following paper reviewed about predicting of heart disease using data mining technique [8].

Signs and symptoms

Angina (chest pain) that occurs regularly with activity, after heavy meals, or at other predictable times is termed stable angina and is associated with high grade narrowing's of the heart arteries. The symptoms of angina are often treated with beta-blocker therapy such as metoprolol or atenolol. Nitrate preparations such as nitroglycerin, which come in short-acting and long-acting forms are also effective in relieving symptoms but are not known to reduce the chances of future heart attacks. Many other more effective treatments, especially of the underlying athermanous disease, have been developed. Angina that changes in intensity, character or frequency is termed unstable. Unstable angina may precede myocardial infarction. About 80% of chest pains have nothing to do with the heart.

2. REVIEW OF LITERATURE

Thuraisingham, B.et al [1] "Data Mining for Malicious Code Detection and Security Applications " In this paper author want to say that data mining is the process of posing queries and fetching patterns from large quantities of data using pattern matching or some other reasoning techniques. Data mining has many applications in security including for national security as well as for cyber security. Threats include in national security attacking buildings, destroying critical infrastructures such as power grids and telecommunication systems. Data mining techniques are being investigated to find out who the suspicious people are and who is capable of carrying out terrorist activities. Cyber security is involved with protecting the computer and network systems against corruption due to Trojan horses, worms and viruses. Data mining is also being applied to provide solutions such as intrusion detection and auditing.

Thuraisingham, B. et al [2] "Data mining for security applications" Author want to propose that the presentation will provide an overview of data mining and security threats and then discuss the applications of data mining for cyber security and national security including in intrusion detection and biometrics. Privacy considerations including a discussion of privacy preserving data mining will also be given.

Asghar, S. et al [3] "Automated Data Mining Techniques: A Critical Literature Review " In this paper author want to proposed that data mining has emerged as one of the major research domain in the recent decades in order to extract implicit and useful knowledge. This knowledge can be comprehended by humans easily. This knowledge extraction was computed and evaluated manually using statistical techniques. Subsequently, semi-automated data mining techniques emerged because of the advancement in the

technology. Such advancement was also in the form of storage which increases the demands of analysis. In such case, semi-automated techniques have become inefficient .So automated data mining techniques were introduced to synthesis knowledge efficiently.

M.Akhil jabbar et al [4] "Heart Disease Prediction System using Associative Classification and Genetic Algorithm" Author want to propose that Heart disease is a single largest cause of death in developed countries and one of the main contributors to disease burden in developing countries. Data from the registrar general of India shows that heart disease are a major cause of death in India, and in Andhra Pradesh coronary heart disease cause about 30%of deaths in rural areas. So there is a very much requirement to develop a decision support system for predicting heart disease of a patient. Hare we used an efficient associative classification algorithm using genetic approach for heart disease prediction. The main motivation for using genetic algorithm in the discovery of high level prediction rules is that the discovered rules are highly comprehensible, having high predictive accuracy and high interesting values.

Data set:

- **Restaurant & consumer data Set**: The dataset was obtained from a recommender system prototype. The task was to generate a top-n list of restaurants according to the consumer preferences.
- **Consumer Panel Data**: The Consumer Panel Data include information about product purchases made by a panel of consumer households across all retail outlets in all US markets. The data include purchases from all Nielsen-tracked categories, including food, non-food grocery items, health and beauty aids, and selects general merchandise. The data represent approximately 40,000 - 60,000 US households that continually provide information about the makeup of their households, the products they buy, as well as when and where they make purchases.
- **ISMS Durable Goods Dataset 1- Purchases history:** There are 19,936 households who made 173,262 transactions involving durable goods purchases and related services from 1176 different stores of a major U.S. electronics chain. The transactions took place between December 1998 and November 2004.
- **ISMS Durable Goods Dataset 2 - Response to promotion:** This dataset records customer response to a Christmas promotion campaign offered by a major U.S. consumer electronics retailer. There are 176,961 customers in the database, 88,336 of whom were mailed the promotion; 88,625 of whom were not. Retail sales during the promotion period are available for both sets of customers. There are 152 variables for each observation, most of which represent each customer's purchase history before the promotion.

Customer Relationship Prediction: The Consumer Panel Data include information about product purchases made by a panel of consumer households across all retail outlets in all US markets. The data include purchases from all Nielsen-tracked categories, including food, non-food grocery items, health and beauty aids, and selects general merchandise. The data represent approximately 40,000 -60,000 US households that continually provide information about the makeup of their households, the products they buy, as well as when and where they make purchases.

3. COMPARISON B/W APPROACHES USED

s.no	Approaches	Advantage	Disadvantage
1.	**Fuzzy KNN** "Fuzzy KNN" algorithm utilizes strength of test sample into any class called fuzzy class membership and thus produces fuzzy classification rule.	The fuzzy class membership can be used instead of crisp class frequency and this fuzzy membership-label neuron provides another perspective of a feature map. This fuzzy class membership can be also used to select training samples in a support vector machine (SVM) classifier. This method allows us to reduce the training set as well as support vectors without significant loss of classification performance	In the presence of full knowledge of the underlying probabilities, Bayes decision theory gives optimal error rates. In those cases where this information is not present, many algorithms make use of distance or similarity among samples as a means of classification. The *K*-nearest neighbor decision rule has often been used in these pattern recognition problems.

2.	**Genetic Algorithm** A genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic particle algorithms approximate the target probability distributions by a large cloud of random samples termed particles or individuals	1) Concepts are easy to understand. 2) Easily distributed. 3) Less time required for special applications. 4) GA is intrinsically parallel.	1) Genetic Algorithm is very slow. 2) They cannot always find exact solution but they always find best solution.
3.	**K-mean clustering** A method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *K-means* clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Verona cells.	The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets 1) If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls. 2) K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.	1) Difficult to predict K-Value. 2) With global cluster, it didn't work well. 3) Different initial partitions can result in different final clusters. 4) It does not work well with clusters (in the original data) of Different size and Different density. 5)Handling Empty Clusters 6)Outliers 7) Reducing the SSE with Post processing.
4.	**K-Nearest Neighbors (KNN)** algorithm is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:	Simplicity effectiveness intuitiveness competitive classification performance in many domain it is robust to noisy training data and is effective if training the data is large	KNN can have poor run time performance when the training set large it is very sensitive to irrelevant or redundant features coz all features contribute to the similarity and thus to the classification.

3.1 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. K-Nearest Neighbours algorithm is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output

depends on whether k-NN is used for classification or regression:

- In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
- In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

3.2 Fuzzy k-Nearest Neighbor Algorithm (FKNN):

The k-nearest neighbor algorithm (KNN) is one of the oldest and simplest non parametric pattern classification methods. In the KNN algorithm a class is assigned according to the most common class amongst its k nearest neighbors. According to his approach, rather than individual classes as in KNN, the fuzzy memberships of samples are assigned to different categories according to the following formulation.

3.3 Genetic Algorithm:

A genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic particle algorithms approximate the target probability distributions by a large cloud of random samples termed particles or individuals. During the mutation transition, the particles evolve randomly around the space independently and to each particle is associated a fitness weight function. During the selection transitions, such an algorithm duplicates particles with high fitness at the expense of particles with low fitness which die. These genetic type particle samplers belong to the class of mean field particle methods.

3.4 K-mean clustering:

K-means **clustering** is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *K-means* clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however, *k*-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

4. CONCLUSION

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. On the basis of these values features have been selected from database. SVM, Naïve bayes, J48 has been used for the classification of database. These algorithms do not provide better classification of data due to incompatibility with dataset.

To overcome the problems of CRM database a new hybrid algorithm is introduced which will be the combination of GA and fuzzy KNN classification.

REFERENCES

[1] Thuraisingham "Data Mining for Malicious Code Detection and Security Applications"**,** 978-0-7695-4406-9, pp. 4–5, IEEE, 2011.

[2] Thuraisingham "Data Mining for Malicious Code Detection and Security Applications", 978-1-4244-5331-3, pp. 6–7, IEEE, 2013.

[3] Asghar, S. "Automated Data Mining Techniques: A Critical Literature Review", 978-0-7695-3595-1, pp. 75–79, IEEE, 2009.

[4] M.Akhil jabbar a "Heart Disease Prediction System Using Associative Classification and Genetic Algorithm", IEEE, 2012.

[5] Ashish Kumar Sen1 "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level", ISSN 2319-7242 Volume 2, pp. 2663-2671, IEEE, 2013.

[6] Pramod Kumar Yadav1, K.L.Jaiswal2, "Intelligent Heart Disease Prediction Model Using Classification Algorithms", IJCSMC, Vol. 2, Issue, pg.102 – 107, IEEE, 2013.

[7] Shamsher Bahadur Patel 1, Pramod Kumar Yadav2, Dr. D. P.Shukla3 "Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques" ISSN: 2319-2380, 2319-2372. Volume 4, Issue 2, pp. 61-64, IEEE, 2013.

[8] Dr.Motilal C. "Role of Data Mining Techniques in Healthcare sector in India", 2320-6691, 158-160, IEEE, 2013.

[9] Nassar, O.A. "The integrating between web usage mining and data mining techniques" pp. 243 –247, IEEE, 2013.

[10] Y. Li, C. Zhang, and J.R. Swan, "An Information Filtering Model on the Web and Its Application in Jobagent," Knowledge-Based Systems, pp. 285-296, 2000.

[11] Xiang Wang "Topic Mining over Asynchronous Text Sequences, Knowledge and Data Engineering," IEEE Transactions on Date of Publication, 2012.

[12] S. Loh, L. K. "Concept-based knowledge discovery in texts extracted from the Web," SIGKDD Explorations, pp. 29–39, July 2000.

Author Profile

Surabhi Aggarwal received B.TECH degree in Computer Science from Punjab Technical University, Jalandhar, India in 2014. Currently, she is in final year of M.TECH (Computer Science) in Guru Nanak Dev University, Regional Campus, Jalandhar (Punjab), India. Her research interests include Data Mining.
Email: surabhiaggarwal.1111@gmail.com