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

A Review on Data Mining, Its Applications and Approaches

Anu Verma[1], Jyoti Arora[2]

[1]M. Tech Student, Desh Bhagat University

[2]Assistant Professor, Desh Bhagat University

Abstract: ***Data mining is the process of extracting useful information from the large amount of database at any time and at any place. The normalization is a process of organizing the data in database to avoid data redundancy, insertion anomaly, update anomaly & deletion anomaly. If a database design is not perfect, it may contain anomalies, which are like a bad dream for any database administrator. Managing a database with anomalies is next to impossible. In the proposed work, data set of normalization is analyzed. The analyzation process is done to remove the missing values from the database. Then an encryption function is developed. Different classifiers are implemented for accuracy of data.***

Keywords: Data Mining, Encryption, Normalization

1. Introduction

1.1 Data Mining

Data mining is becoming more popular due to the concept of "big data". Data mining is the process of extracting useful information from the large amount of database at any time and at any place. But along with the data mining, it is necessary to save and protect the data from unwanted hands and from attacks i.e. to save the data from being stolen and from the intrusions. Thus original data needs Privacy Preserving Data Mining (PPDM) so that the data can be protected from the unauthenticated as well as unauthorized person. The attacks or intrusions take place in the data due to its linking to other databases, loss of one's personal as well as basic information due to negligence of the data holder or any other kind of issue. Therefore, it becomes essential to apply various privacy preserving data mining techniques which can be efficient enough to hide the data from the third party person so that original data cannot be leaked to anyone or to be misused or misinterpreted.

1.2 Working of Data Mining

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries.

1.2.1 Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

1.2.2 Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

1.2.3 Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

1.2.4 Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

1.3 Issues in data mining

There is lot of issues in Data Mining. These are:

1.3.1 Security and social issues: Security is an important issue with any data collection that is shared and/or is intended to be used for strategic decision-making. In addition, when data is collected for customer profiling, user behavior understanding, correlating personal data with other information, etc., large amounts of sensitive and private information about individuals or companies is gathered and stored.

1.3.2 User interface issues: The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. Good data visualization eases the interpretation of data mining results, as well as helps users better understand their needs. Many data exploratory analysis tasks are significantly facilitated by the ability to see data in an appropriate visual presentation. There are many visualization ideas and proposals for effective data graphical presentation.

1.3.3 Mining methodology issues: These issues pertain to the data mining approaches applied and their limitations. Topics such as versatility of the mining approaches, the diversity of data available, the dimensionality of the domain, the broad analysis needs (when known), the assessment of the knowledge discovered, the exploitation of background knowledge and metadata, the control and handling of noise in data, etc. are all examples that can dictate mining methodology choices. For instance, it is often desirable to

have different data mining methods available since different approaches may perform differently depending upon the data at hand.

1.3.4 Performance issues: Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the very large data sets data mining is dealing with today. Terabyte sizes are common. This raises the issues of scalability and efficiency of the data mining methods when processing considerably large data. Algorithms with exponential and even medium-order polynomial complexity cannot be of practical use for data mining. Linear algorithms are usually the norm. In same theme, sampling can be used for mining instead of the whole dataset. However, concerns such as completeness and choice of samples may arise. Other topics in the issue of performance are *incremental updating*, and parallel programming.

1.3.5 Data source issues: There are many issues related to the data sources, some are practical such as the diversity of data types, while others are philosophical like the data glut problem. We certainly have an excess of data since we already have more data than we can handle and we are still collecting data at an even higher rate.

1.4 Applications of Data Mining

1.4.1 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.

1.4.2 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.

1.4.3 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.

1.4.4 Data Mining Applications in Transportation

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

1.4.5 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.

1.5 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, nonfood 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, nonfood 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.

2. Review of Literature

Sheryl Parmar et al. [1] in "**A Coherent technique for Privacy Preservation in Data Mining using Classification**" elaborated the protection of the sensitive and confidential data using the tangent hyperbolic transfiguration technique. The transfigured values of the original data convert the data in a format that is unrecognizable by the users. After the conversion, the data mining techniques are applied to check the accuracy of the data.

Nedaabdelhamid et al [2] in "Emerging trends in associative classification data minning" 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.

S.UmmugulthumNatchiar et al [3] in "Customer Relationship Management Classification Using Data Mining Techniques" Customer Relationship Management possesses Business Intelligence by incorporating information acquisition, information storage, and decision support functions to provide customized customer service. It enables customer representatives to analyse 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 out dated, substandard because of noisy and unbalanced 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 pre-processing. Finally it enhances the performance of classification.

Sankaranarayanan, S. et al [4] in "Diabetic Prognosis through Data Mining Methods and Techniques" concluded apriori 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 Apriori have been used for application to diabetes dataset and association rules are being generated by both of these algorithms.

C. M. Velu et al [5] in "Visual Data Mining Techniques for Classification of Diabetic Patients" that clustering technique is quite often used by many researchers for classifications due to its' being simple and easy to implement. It uses Expectation-Maximization (EM) algorithm for sampling. The study of classification of diabetic patients was main focus of this research work. Diabetic patients were classified by data mining techniques for medical data obtained from Pima Indian Diabetes (PID) data set. This research was based on three techniques of EM Algorithm, h-means+ clustering and Genetic Algorithm (GA). These techniques were employed to form clusters with similar symptoms. Result analyses proved that h-means+ and double crossover genetics process based techniques were better on performance comparison scale. The simulation tests were performed on WEKA software tool for three models used to test classification. The hypothesis of similar patterns of diabetes case among PID and local hospital data was tested and found positive with correlation coefficient of 0.96 for two types of the data sets. About 35% of a total of 768 test samples were found with diabetes presence.

3. Approaches Used

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 extracting useful information from the large amount of database at any time and at any place. The normalization is a process of organizing the data in database to avoid data redundancy, insertion anomaly, update anomaly & deletion anomaly. First of all we will analyze the data set for normalization. Normalization is creation of shifted and scaled versions of statistics where the intention is that these normalized values allow the comparison of corresponding normalized values for different datasets. Then develop an encryption function that converts originality of the data for security aspects. In Encryption we encode messages or information in such a way that only authorized parties can read it. Then implement different classifiers for the prediction of the accuracy of the original as well as the encrypted data.

References

[1] Sheryl parmar "A Coherent technique for Privacy Preservation in Data Mining using Classification", International conf. on Coherent technique for Privacy**,** 2015, pp 32-43.
[2] E.W.T. Ngai "Application of data mining techniques in customer relationship management: A literature review and classification" IEEE Conf. on Expert Systems with Applications 2009, Pages 2592–2602.
[3] Nedaab delhamid, "Emerging trends in associative classification data mining" International journal of electronics and electrical engineering, Feb 2015, pp 56-62..
[4] Robert E. Marmelstein "Application of Genetic Algorithms to Data Mining" IEEE Conf. on MAICS-97 Proceedings, 1997 , pp. 53-57.
[5] Jagannathan, G. **"**Visual Data Mining Techniques for Classification of Diabetic Patients**",** IEEE Conf. on Data Mining Workshops, 2007, pp. 1-3.
[6] S. Ummugulthum Natchiar "Customer Relationship Management Classification Using Data Mining Techniques" International Conf. on Science Engineering and Management Research (ICSEMR), 2014, pp. 1 – 5.
[7] 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.
[8] Tzung-Pei Hong "GA-based item partition for data mining" IEEE Conf. on Systems, Man, and Cybernetics (SMC), 2011, pp. 2238 – 2242.
[9] Wang, Guoyin "Granular computing based data mining in the views of rough set and fuzzy set" IEEE Conf. on Granular Computing, 2008, pp. 67.
[10] E.W.T.Nagy "Application of data mining techniques in customer relationship management: A literature review

and classification" IEEE Conf. on Expert Systems with Applications 2009, Pages 2592–2602.
[11] Nedaab delhamid, "Emerging trends in associative classification data mining" International journal of electronics and electrical engineering, Feb 2015, pp 56-62.
[12] Hossin, M "A Review on Evaluation Metrics for Data Classification Evaluations", International Journal of Data Mining & Knowledge Management Process, 2015, pp 1-6.
[13] Asghar, S. "Automated Data Mining Techniques: A Critical Literature Review" 978-0-7695-3595-1, 75 – 79, IEEE, 2009.