A Comparative Study of Various Data Mining Techniques Including Data Reduction in Business Data Processing

Mohd Ahtesham Farooqui[1], Dr. Jitendra Sheetlani[2]

[1] *Research scholar, Department of Computer Science & Application*
Sri Satya Sai University of Technology & Medical Sciences, Sehore (M.P.)
INDIA
[2] *Associate Professor, Department of Computer Science & Application.*
Sri Satya Sai University of Technology & Medical Sciences, Sehore (M.P.)
INDIA

ABSTRACT

The data-mining is achieving the attention in varieties of research domains because of their limitless implementations and applications in order to extract the data within a proper way. The data-mining is one of the methods used for extracting the essential information on web. In this paper data-mining approach is used, within this approach few special features are there like analyze usage data, surfing information, and Marketing information. This research paper discusses several techniques of data-mining within the educational sector that has the greater impact to improve the students' academic performance and experience. The major objective of this research work is to analyze the various applications of the data-mining within e-commerce, retail, and other types of business domain through targeting on the structured and the un-structured collection of data from different types of resources.

Keyword: - *Data Mining, Data Reduction, Business Data Processing, e-commerce*

1. INTRODUCTION

In this era of technology, the Data-mining application within the e-commerce domain is related with the combination of databases, statistics and the AI along with few other domains on order to produce a latest concept or the latest combination of technology for the optimal decision-making process. [1] Due to the rapid and drastic growth of E-commerce domain, the manner in which companies are doing businesses has been changed drastically. E-commerce that is mainly characterized by doing business electronically with the help of the Internet has provided the effective and cheapest method for running a business. The information achieved can be used in the field of Medical science, Education, Business, and Agriculture etc. Since the large volume of data is get collected and stored within the huge databases. And the Data-mining or knowledge discovery has become the area of growing significance because it helps in investigating the data from various domains and then concluding them into the essential for of information.

The data-mining concept is related with the introducing the approaches which generate the knowledge among the data from various fields like educational environments, retail, marketing, etc. The data can be collected from various educational institutes that reside in their databases. [2] The data can be personal or academic which can be used to understand students' behavior, to assist instructors, to improve teaching, to evaluate and improve e-learning systems, to improve curriculums and many other benefits. Within this paper, the detailed analysis is performed over the applications of data-mining within the health-care domains also, and various kinds of data are utilized along with the details or information that is extracted. The data-mining approaches are implemented within the healthcare field also which performs a vital role within the prediction and also in diagnosis of several diseases.

The Data-mining concept utilizes separate types of families of the statistical, computational, and the machine-learning approaches which may consists of statistical-analysis, neural-networks, decision trees, rule-induction and the refinement along with the graphic-visualization, in order to carefully represented the data in order to disclose the

complicated relationships which are presented. Moreover the machine-learning approaches that are exist since long time, so the establishment of the advanced and the user-friendly platform for the business purposed [5] that have made the data-mining approach to be more interesting and to be applicable for the organizations. If these types of pattern-extraction approaches are utilized in a correct way so they may be proved to be the efficient approaches for extracting the essential information among the data.

The honorable process of the data-mining has started with recognizing the correct business circumstances. Unluckily, if there are various types of appropriate statisticians along with the adequate type of analysts having the work which is wasted as they are resolving the issues which never support the business purposes and the best data-miners needs to prevent these types of conditions?

2. APPLICATION

2.1 Using Data-mining in Marketing and E-Commerce -
The major issue in the marketing domain is to properly define the good prospect also the method which is applied by various companies to find this is just anyone that may have even showing the curiosity to be like a customer. And the data-mining approach is implemented in this issue through defining first its meaning which needs to be good-prospect then the defining the finding-rules which may enable the people along with those types of features that are focused [11].

The data-mining approach is also applied within the e-commerce domain which is the significant pattern of switching the e-commerce industry in order to help the enterprise along with the needed details that are related with the business. Currently, various companies have adopted the e-commerce platform and are being have huge amount of data as a database.

2.2 Application Data-mining in Business –
The data-mining approach have also applied within the retail, insurances or banking sectors, for the functioning such as the customer-segmentation and the customer-retention, for performing the market-basket analysis or even the fraud-detection. This is very significant approach for business that is successfully implemented the data-mining approach, which is provided within the Retail sector. The retail data-mining may provide support for recognizing the buying behaviors of the customer, and may introduce the shopping patterns of customer along with their trends, also it enhance the quality of the customer service, to obtain the optimal customer-retention and customer-satisfaction [5].

2.3 Application of Data-mining in Education –
The approach of data-mining is also be implemented in the educational sector for grouping the students, anticipating their performance, for planning and also scheduling the courses or may be for understanding behavior of student in order to achieve the better level of education.

2.4 Application of Data-mining in Science and engineering –

The data-mining techniques have also been implemented for the fields such as astronomy, bio-informatics, medicine, genetics, telecommunications, electrical power, or climate data gathering etc.

2.5 Application of Data-mining in Banking-
Several domains are there where the data-mining approaches are implemented like financial domains such as customer-segmentation and the profitability, anticipating the payment-default, credit-analysis, fraudulent-transactions, marketing, ranking-investments, cash-management-system, high-risk-loan candidates, etc [6].

2.6 Application of Data-mining in Healthcare –

Within the Healthcare domain the data-mining is an emerging area of research within the information-technology. The Data-mining may have the great opportunities for the health-care-management system in order to enable the health-system to properly utilize the data and then analyze them in order to enhance the care and also decrease the cost simultaneously that may be implemented as 30 percent of the entire health-care expenses [1].

3. DATA-MINING TECHNIQUES

The Data-mining approach is one of the techniques of finding out the meaningful signatures and the correlation through sifting from huge volume of data that is recorded within the databases. Various types of tools are there for

this type of data generation process that is consisting of abstractions process, aggregations process, summarization process and then characteristics of the data [6]. In the Data-mining various techniques that are implemented in order to operate over the huge data for discovering the hidden patterns along with the relationships for decision-making process are described below:

3.1 Neural Networks-
Within this technique that may be utilized for the classification of large complex data. It can be used to study course selection by students, student course satisfaction, and specialization selection. In paper [8] state the neural network denoted every type of cluster through a neuron or by the proto-type. The input data is also represents by neurons which are connected to the prototype neurons. Each such connection has a weight, which is learned adaptability during learning.

3.2 Decisions Tree-
Within the Decision-tree technique which is one of the data-mining techniques that can be used for classification and prediction of large data. Decision tree is used for profiling customers. Decision tree is also called rule induction technique. In the paper [8] the data, in decision tree, is represented by "a hierarchical tree where each leaf refers to a concept and contains a probability description of the concept."

3.3 Clustering-
In the clustering approach it is represented as the un-supervised type of learning approach which is finding through analyzing only the independent variables whereas the supervised type of learning approach may observe the independent variables and the dependent variables both. This approach is separate from the classification technique that is the supervised type of learning approach. Clustering has no any pre-defined type of classes. Due to this cause, the clustering technique can be applied for the analysis of the introductory behavior, exclusively in case those types of analysis may comprises of the huge volume of data.

3.4 Regression-
Among various data-mining approaches the regression approach is a significant approach. By the help of this approach, it is easy to recognize those types of functions which are essential for representing the correlation within several types of variables. This is basically a mathematical type of approach which is created by the support of the training data-set.

3.5 Classification-
And the Classification technique is the very famous and applied approach of the Data-mining techniques which is applied within the Healthcare domain. This technique split the data-samples into the target-classes. And the classification approach anticipated the target-class for every type of data-points. Along with the classification technique the risk-factor may also be related with the patients through observing the symptoms of the diseases.

4. DATA REDUCTION TECHNIQUES

There are number of methods for reducing the amount of data. By using more "intelligent" capture and storage algorithms, the amount of data initially recorded is reduced. Post processing of the data can reduce the data that a human would have to interpret, by applying various mathematical functions or using statistical presentation techniques. Artificial intelligence tools like fuzzy logic, neural nets, and expert systems can take over the tasks of interpreting the data and present the results as a list of recommendations.

Data reduction schemes used when samples are unneeded or null values is sensed by sensors and unnecessary for transmission. Data acquisition is requiring when reducing the energy spent by sensing subsystems. In-network processing is necessary when data aggregation occurs in between sensor and sink. In this way the amount of data is reduced while traversing network towards the base station. Moreover data compression can be applied to reduce amount of data sent by source nodes by involving encoding at nodes and decoding at sink.

4.1 Adaptive Filter:
Filter is a device that reduces unnecessary features from a signal. Filter has two types stationary and non-stationary. Stationary filter doesn't allow variation of its component and non-stationary filters are required to track time variation and allow changing its coefficient.

4.2 Tree Based Methods:

It sometimes also based on in-network data reduction. There are many tree based methods of data reduction in WSN. For mobility, we can use distance proportional energy consumption model.

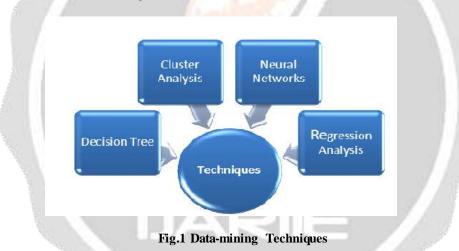**4.3 Cluster-Head (CH) Based Reduction:**

Cluster-Head (CH) Based Reduction plays an important role to reduce energy in WSN thus life of network as well as scalability can be improved. Clustering is necessary when high-density networks because it can easily manage a set of cluster head from each cluster than whole nodes.

4.4 Data Stream Based Reduction:

When the data arrives in online as unordered is called data streams. Data stream reduction mostly requires when real time application used in WSN and not meet its deadline due to delay. Therefore reduction is necessary while transmission and to obtain less data delay in wireless sensor network. Sampling data stream techniques should be used to reduce data traffic.

5. DATA-MINING TECHNIQUE IN BUSINESS PROCESSING

Within the IT sector the Data-mining is one of the huge sector for the scientist and doctors for managing the large volume of data-sets of patients or the researches like for making sense of the complicated diagnostic tests or reports, for interpreting the existing results also for merging the various data with each other. Basically the clinics or hospitals' decision is generated through the medical experts based on the observations or the prior knowledge other than information achieved from the huge data-bases [3].

Fig.1 Data-mining Techniques

This paper also recognizes four areas of application of data-mining in education: improve students' model, discovering models of the knowledge structure, studying the pedagogical support provided by learning software and scientific discovery about learning and learners.

The data-mining approaches are also very essential within the health-care sector. They also enable the optimal medical facilities for the patients in order to support the health-care firms in several medical-management related decisions. Few services are also offered by data-mining approaches within the health-care domain like days number to be recorded of stay within a hospital, ranking of the hospitals, efficient treatments, any type of fraud insurance-claims by any providers, etc [2].

This paper study examines the educational mining using a case study from dataset from students' behavior. It explains how and what data could be collected how it should be processed. And the result of the analysis will make the educational domain people to be benefited. Clustering and classification based on the students' behavior patterns is one of the effective ways to predict the accuracy. It gives us meaningful information based on various evaluation factors.

Various Data-mining Techniques for Business Data Processing:

The challenges faced in processing Big Data technologies are overcome by using various techniques [4]. The most popular techniques used in educational data-mining are listed below.

5.1 Regression – Regression is used in predicting values of a dependant variable by estimating the relationship among variables using statistical analysis

5.2 Nearest Neighbor – In this technique the values are predicted based on the predicted values of the records that are nearest to the record that needs to be predicted.

5.3 Clustering – Clustering involves grouping of records that are similar by identifying the distance between them in an n-dimensional space where n is the number of variables.

5.4 Classification – Classification is the identification of the category/class to which a value belongs to, on the basis of previously categorized values.

6. LITERATURE REVIEW

In this paper [9] presented the way to observe several data-mining implementation within the domain of health-care in order to find out the latest range of the patterns and information. In the medical sector that may have large volume of data-set as the group regarding the diagnosis and the details of patient along with the medications. For making these types of data as an essential pattern also for predicting the upcoming trends of the data-mining techniques that are applied within the health-care sector have been introduced in this paper. Various types of data-mining applications and tools are present for the health-care diagnosis-systems which may be defined clearly within this paper which is also a cost-effective method for this domain.

In this paper [10] determined the application of the data-mining technique within the education section. Data-mining is used in industries for a variety of applications such as predicting and classifying customers and clustering customer's characteristics and prepares marketing strategies to segmented customers for the achievement of profitability. Similarly, universities can also apply data-mining for predicting enrolment of students into various courses. Data-mining can be applied for classifying and clustering students characteristics based on demographic, psychographic and behavioral variables. Data-mining can also be applied by using if-then rule. In addition, it can describe the profile of successful and unsuccessful students based of GPA achieved during the semesters. It can also be used for dropout student, students' academic performance, teachers' performance, and students' complaints.

In this paper [11] represented the results of the suggested approach that may leads to actionable knowledge that can be readily employed by decision makers. A center piece of our work is the proposal of S-PLSA, Using S-PLSA as a means of "summarizing" sentiment information from reviews. In this paper also have developed ARSA, a model for predicting sales performance based on the sentiment information and the products past sales performance. Here also further considered the role of review quality in sales performance prediction. with the analysis and from the dataset we are getting the sales performance and customers behavior with his full information, with company dataset we are getting his current sales details how it is varies from 247-300 units or 105-350 units with respect to year, week and as per the algorithm the sales details are varies with the movie type with respect to year and month.

Within this paper [12] represented the overview of the data-mining technique within the electronic-commerce domain also this explains few type of general issues which may be resolved through the techniques of data-mining which is used this domain. Here also have described an overview of few commonly applied machine-learning approaches within the data-mining concept. Also here selected to be determining the best association-rules mining approach in detail that have discussed related to the issues along with their application. Within this paper the future work is explained as giving the proper descriptions along with the discussion of any other type of techniques within data-mining approach in order to implement it into e-commerce.

This paper [13] described the result of analysis that a complicated requirement is there for any automatic technique in order to effectively and efficiently used the huge volume of the financial information in order to help the companies and the organizations or individual within the strategic-planning approach also within the investment in decision-making process. The data-mining approaches have also been applied here in order to expose the hidden-patterns also to anticipate the future needs and behaviors that are required within the financial-markets. This is the most effective benefit which is obtained by the data-mining approach which consists of raised revenue, less cost

along with enhanced market-place awareness. Therefore, this paper suggested several types of organizations to implement the data-mining approaches in order to resolve all the complicated issues.

In this paper [14] describes the latest processing concept on the E-Commerce sites evaluation which is the new DBSCAN algorithm combining factor analysis with different densities. Compared with the traditional DBSCAN algorithm, the results of evaluating websites are more reasonable and interpretable with the improved DBSCAN algorithm. This paper improves the clustering accuracy and reasonableness of the evaluation by combining factor analysis with DBSCAN. However, the data processed by factor analysis have characteristics of uneven density. The traditional DBSCAN is improved to partition the data with different densities and cluster these sites.

Within this paper [15] have targeted on representing the implementation of the data-mining approaches within the business or industry. This have provided an overview of the data-mining approach, which is giving the proper definition of this concept, also representing the six type of data-mining approaches along with describing the major domains that may implement the data-mining approaches. In this paper also given the major business domains that may are benefited by the applications of the various data-mining tools, and their uses within the banking, retail, or insurance sector. Also the various commercially available data-mining features or platforms are there that are represented in this paper. In this suggested architecture for various solutions of data-mining for implementing within the business may enhance the effectiveness of the company, through enabling the important decision-making information in order to reduce the operating costs also achieving the competitive benefits.

Within this paper [16], described the applications of the data-mining approaches along with text-mining approaches within the business analysis also how these techniques may help the business experts within e-businesses domains have also been discussed here. This paper presented the difference in between the data-mining business analysis and the business experts. Increasing data resources are expected to drive a growth in business analytics and thus data-mining. Businesses are beginning to realize that implementation of the data-mining techniques along with the text-mining have enables them a competitive benefit. With data increasing in an exponential manner, the ability to use data-mining to sieve through massive amount of data and identify relevant patterns will become a strategic tool in improving key areas of the business such as customers, operations and the supply chain.

This paper [17] suggested that the retail industries make profits out of efficiency from their supply chain managements. BI tools help in supporting their supply chain management which makes enterprises have better management over the supplies. When integrated in IMS or WMS, BI tools help in finding patterns, possible over production, shortages, and under productions and so on; but in most cases quick response towards demand spike which are very important to catch. Originally, models that are complexly mathematical were used for logistical problems and supply chain management. Logistics are capable of affecting inventories specially to lack of precise forecasting and spikes in demands from slow response.

In this paper [18] the main objective of the data-mining approach is to predict, so the predictive-data-mining is one of the usual patterns of the data-mining approach described in this paper also in this paper one technique which has direct applications in business have also been described. Customer segmentation is a core process for assisting a marketing strategy. Huge amount of customer data are continuously generated .The paper aimed to study different DM techniques used so far for customer segmentation in online retail Industry. To identify the gaps from previous studies which can be helpful for further research in the field.

Here in this paper [19] determines the huge range of implementations of data-mining within the various domains of industry in which the data is produced. This is also have been found out that the data-mining technology is the very significant aspect within the database also in the information-management-systems. In this it is suggested that data-mining is also a most promising concept for the development of IT industries. The Data-mining approach is a promising discipline and has wide applicability. It can be applied within several sectors. And the data-mining as the confluence of multiple intertwined disciplines, including statistic, machine-learning approach for pattern-recognition, database system, information retrieval world wide web, visualization, and many other application domains, has made great progress in the past decade. In this paper we have discussed the various industry wide applications of data.

7. CONCLUSIONS

In this paper it is concluded that any approach of data-mining have its separate accuracy and the performance, which is having an importance in specific domain. Because of the few factors affecting any type of algorithm may loss the

accuracy and also the performance of the approach. The result of the review of the paper work represented that techniques of data-mining are efficient method for identifying the performance of the students' in education domain. And data-mining approach have also been implemented in to the e-commerce sector to learn the searching behavior of the users in order to examine the success of the marketing strategies, also to enhance the designing of the better e-commerce platforms. In the present study this paper has described the various data-mining techniques which can support education system via generating strategic information. And this paper is also objected to compare various applications of data-mining within healthcare domain for deriving the essential details. This analysis also help in predicting the diagnosis of various diseases by the use of Data-mining techniques which is also a complicated work also this may decrease the effort by rising the accuracy of the system.

8. REFERENCES

[1] Dhanashree S.Medhekar,Mayur P.Bote,Shruti D.Deshmukh,"Heart Disease Prediction using Naïve Bayes", International Journal Of Enhanced Research In Science Technology & Engineering Vol.2 Issue 3, March 2013.

[2] Giudici, Paolo. Applied Data Mining: Statistical Methods for Business and Industry, John Wiley and Sons Ltd., 2013

[3] He, Wu. "Examining students' online interactions in alive video streaming environment using data mining and text mining" Computers in Human Behaviour, 29 (2013): 90-102

[4] C. McGregor, C. Christina and J. Andrew, "A process mining driven framework for clinical guideline improvement in critical care", Learning from Medical Data Streams 13th Conference on Artificial Intelligence in Medicine (LEMEDS). http://ceur-ws. org, vol. 765, (2012).

[5] Abdul Hamid, et al, "A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data mining", IDEE, Proceedings of the 2014 Workshop on Interaction Design in Educational Environments ISBN: 978-1- 4503-3034-3 Pages 106, ACM, 2014.

[6] Chen, H., Chiang, R. and Storey, V. (2012). Business Intelligence and Analytics: From Big Data to Big Impact, MIS Quarterly, 36 (4), 1165-1188.

[7] Wu, M., Zhang, H. and Li, Y. (2013) Data Mining Pattern Valuation in Apparel Industry E-Commerce Cloud. IEEE 4th International Conference on Software Engineering and Service Science (ICSESS), 689-690.

[8] Rokach, Lior. "A survey of clustering algorithm," in O.Maimon, L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, 2nd Edition, Springer, 2010

[9] K. Rajalakshmi Dr. S. S. Dhenakaran, "Analysis of Datamining Prediction Techniques in Healthcare Management System", Volume 5, Issue 4, April 2015 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.

[10] Dr. Mohd Maqsood Ali, "ROLE OF DATA MINING IN EDUCATION SECTOR", International Journal of Computer Science and Mobile Computing A Monthly Journal of Computer Science and Information Technology ISSN 2320–088X IJCSMC, Vol. 2, Issue. 4, April 2013, pg.374 – 383.

[11] Navneet Singh#1, Muhammad Rukunuddin Ghalib*2, "An Effective E-Commerce Management using Mining Techniques", International Journal of Scientific and Research Publications, Volume 3, Issue 8, August 2013 1 ISSN 2250-3153.

[12] César Astudillo1 , Matthew Bardeen2 and Narciso Cerpa, "Editorial: Data Mining in Electronic Commerce – Support vs. Confidence", Journal of Theoretical and Applied Electronic Commerce Research ISSN 0718–1876 Electronic Version VOL 9 / ISSUE 1 / JANUARY 2014 / I-VII.

[13] 1 Radhakrishnan B, 2 Shineraj G, 3 Anver Muhammed K.M, "Application of Data Mining In Marketing", IJCSN International Journal of Computer Science and Network, Volume 2, Issue 5, October 2013 ISSN (Online) : 2277-5420.

[14] Yongyi Cheng1,2 , Yumian Yang3 , Jianhua Jiang3 and GaoChao Xu, "Cluster Analysis of E-Commerce Sites with Data Mining □Approach", International Journal of Database Theory and Application Vol.8, No.3 (2015), pp.343-354.

[15] Ruxandra-Stefania Petre, "Data Mining Solutions for the Business Environment", Database Systems Journal, 2013, vol. 4, issue 4, pages 21-29.

[16] Pui Mun Lee, "Use Of Data Mining In Business Analytics To Support Business Competitiveness", Review of Business Information Systems – Second Quarter 2013 Volume 17, Number 2 2013

[17] Prachi Agarwal, "Benefits and Issues Surrounding Data Mining and its Application in the Retail Industry", International Journal of Scientific and Research Publications, Volume 4, Issue 7, July 2014 1 ISSN 2250-3153.

[18] Mrs. Pradnya Muley* Dr. Anniruddha Joshi, "Application of Data Mining Techniques for Customer Segmentation in Real Time Business Intelligence", International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 4, Volume 2 (April 2015).

[19] Eze, Udoka Felista (PhD)1 , Adeoye, Olufemi Sunday2 , Ikemelu, Chinelo Rose-keziah3, "Industry Wide Applications of Data Mining", International Journal of advanced studies in Computer Science and Engineering IJASCSE, Volume 3, Issue 2, 2014.