International Journal of Recent Trends in Engineering & Research
ISSN(ONLINE) : 2455-1457
IMPACT FACTOR : 3.344

A Study on Data Mining Horizons

Sunny Sharma
Department of computer Science, Guru Nanak Dev University

Abstract— Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Among the rapid pace of data with the need of analysis as well as summarizing, the data mining techniques are applicable in insurance and banking, auditing, opinion mining & sentiment analysis, pharmaceutical sales & research, email spam filters, packaging industries, detection of unusual patterns, student performance analysis & counseling system, terrorist activities & fraudulent behavior, retail industries, telecom & public sector and in medical field. The advancement in different medical fields leads to the discovery of various critical diseases and provides the guidelines for their cure. This whole task is possible only with the help of data mining. This paper describes the data mining horizons in different areas, & highlights the areas where data mining has not mined the data.

Keywords— Decision tree Induction, Rule Based Classification or mining, Support vector machine, stochastic classification, Logistic regression, Naïve bayes, Artificial Neural Network & Fuzzy Logic, Genetic Algorithms

I. INTRODUCTION

Data mining has acknowledged an enormous deal of attention in field of Agriculture, Mathematics, Computer Science, Finance, Chemistry, Economics and especially in areas of Medical Sciences and Bio-informatics. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both [14],[15]. DATA mining is the process of finding correlations, relation between data or patterns. The main aim of this process is find the patterns that were previously unknown [16], [17]. Once these relations are found these can be used in decision making. Multistep Approach involved in data mining which can be expressed as Figure 1.

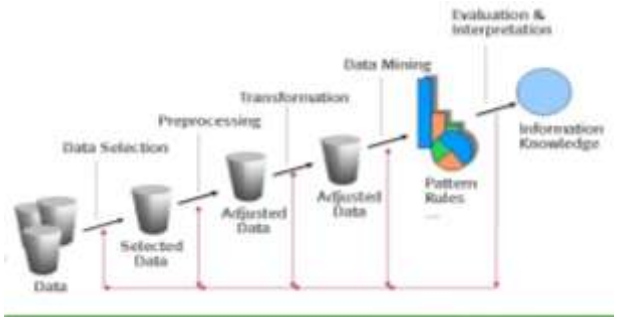

Figure 1: Data mining Process

This Multistep Approach can be expressed as: Firstly data integration: At this initial stage data is being collected from all heterogeneous sources. Secondly data selection: At this stages select the data which is relevant one or in homogenous format. Then data cleaning: which is to remove the bugs or errors from data as data gathered is not in clean form. Then data transformation: which describe data after cleaning is not ready for mining as we need to transform data into forms appropriate for mining. These techniques are for smoothing, aggregation, normalization. Then data mining: now data is

ready for applying techniques on data to locate or search the interesting patterns. Then pattern evaluation & knowledge presentation: This step involves visualization, transformation etc. finally deployment: which express decision /use of discovered knowledge [16].

II. DATA MINING TECHNIQUES

The data mining concept is originated from statistics, machine learning as well as from artificial intelligence. Lots of efforts have been done on data mining techniques. Broadly defined techniques in Data mining can be classified into two classes: Predictive & Descriptive techniques. In predictive data mining the focus is done on discovering a relationship between independent as well as between dependent and independent variables. Predictive data mining can be used to forecast explicit values based on patterns in the data. Descriptive data mining describes a data set in a brief but comprehensive way and gives interesting characteristics of the data without having any predefined target. The various data mining techniques are shown in Figure 2.

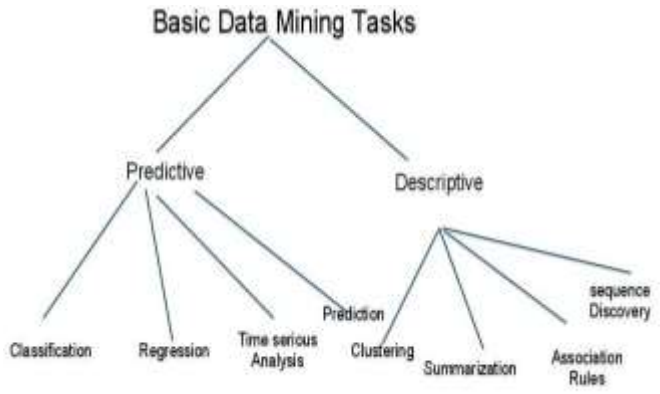

Figure 2: Data Mining Techniques

III. CLASSIFICATION TECHNIQUES

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. Classifications are discrete and do not imply order. Continuous, floating-point values would indicate a numerical, rather than a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm. Classification is a two step process: Learning or training Phase or Supervised Learning & Classification [18] [19]. Various Classification Methods are used these are Decision Tree, Rule-based Methods, Neural Networks, Naïve Bayes, Instance Based Learning, Bayesian Belief Networks, Support Vector Machines [20],[21].

VI. LITERATURE SURVEY

In 2010 A. Deen et al. [11] use rule based classification and discuss the association rule based classification techniques using various algorithms on Primary Tumor, Breast Cancer etc & attain the accuracy of 82%. In 2010 A. K Banerjee et al. use stochastic classification and describe the relation among physiochemical properties of proteins keeping in view hydrophobicity of AGC kinase super family. In 2010 V. Christina et. al. [5] uses multilayer perceptron classifier and describes that it out performs than other classifiers and the false positive rate also very low compared to other algorithms. Authors tells email spam filters using this approach can be adopted either at mail server or at mail client side to reduce the amount of spam messages and to reduce the risk of productivity loss, bandwidth and storage usage. In 2010 V. Christina et. al. [5] finds out multilayer perceptron classifier importance over other classifiers and detects the false positive rate very low compared to other algorithms. The email spam filters using this approach can be adopted either at mail server or at

mail client side to reduce the amount of spam messages and to reduce the risk of productivity loss, bandwidth and storage usage. In 2011 M. Singh et al. [12] also used rule based classification and describe the immense use of rule based mining to predict the protein function. In 2011 J. Sony et al. [8] use Genetic Algorithms, ANN, decision trees for the prediction of heart diseases & found decision trees outperform & achieve the accuracy 99.2%. In 2012 S. ChandraKala et.al [3] performed survey and they found many of the appliances of opinion mining are based on bag of words, which do not capture context which is essential for sentiment analysis. In 2012 R. Changala1 et. al. uses Decision tree Induction on different datasets and express their importance and pitfalls. In 2012 M.Singh et. al.[10] proposed the way to predict the protein function prediction from sequence derived features using See5 tool and design the decision tree through see5 decision rules & attain the accuracy of 64%. In 2013 S. Singh et al. [9] predict staging of cervical cancer with genetic algorithms.In 2015 G. Kaur et. al. [4] use ANN & Fuzzy Logic for data mining dependent filtering which has power to remove the combined noise in far more proficient manner. Author describe the future for enhance data mining dependent fuzzy filter for removing the high density of disturbance. In 2015 N. Handa [7] describes the role of data mining for digital library which is collection by acquiring, describing, storing and delivering resources. Many digital objects can be delivered directly over the Web, while some may require special software for viewing various applications. In 2015 B. Singh et. al. [13] describe that Phylogenetic tree construction can be useful in analyzing the distances between more than two sequences and distance matrix methods such as neighbour joining or UPGMA. In 2016 N. Koul et al [2] use Naïve bayes method and authors did survey on opinion mining & sentiment analysis in text & they compare naïve Bayes, max entropy and support vector vachines and proposed their significant way. In 2016 R. Banswal et. al. [1] ANN & Fuzzy Logic and design students performance analysis and counseling system, authors did the analysis using Apriori algorithm which satisfies the minimum support and minimum confidence threshold. In 2016 N. Koul et al. [2] use support vector machine and authors did survey on opinion mining & sentiment analysis in text & they compare naïve Bayes, max entropy and support vector machines and proposed their significant way. The comparative literature survey for various classification techniques are shown in Table 1.

Table 1: Comparative Literature Survey

Methodology	Year of Publication	Author	Results Discussion
Rule Based Classification	2010	A. A Deen et al. [11]	Authors discuss the association rule based classification techniques using various algorithms on Primary Tumor, Breast Cancer etc & attain the accuracy of 82%.
	2011	M. Singh et al. [12]	Authors describe the immense use of rule based mining to predict the protein function.
Support vector machine	2016	N. Koul et al.[2]	Authors did survey on Opinion Mining & Sentiment Analysis in text & they compare naïve Bayes, Max Entropy and Support Vector Machines and proposed their significant way.
	2012	S. ChandraKala et.al[3]	Authors performed survey and they found many of the appliances of Opinion Mining are based on bag of words, which do not capture context which is essential for Sentiment Analysis.
Stochastic classification	2010	A. K Banerjee et al.	Authors describe the relation among physiochemical properties of proteins keeping in view hydrophobicity of AGC kinase super family.

Naïve bayes	2016	N. Koul et al.[2]	Authors did survey on Opinion Mining & Sentiment Analysis in Text & they compare naïve Bayes, Max Entropy and Support Vector Machines and proposed their significant way.
	2010	V. Christina et. al.[5]	Multilayer Perceptron classifier out performs other classifiers and the false positive rate also very low compared to other algorithms. Email spam filters using this approach can be adopted either at mail server or at mail client side to reduce the amount of spam messages and to reduce the risk of productivity loss, bandwidth and storage usage.
ANN & Fuzzy Logic	2016	R. Banswal et. al.[1]	Students Performance Analysis and Counseling System using ANN and Fuzzy Logic was implemented and authors did the analysis using Apriori algorithm which satisfies the minimum support and minimum confidence threshold.
	2015	G. Kaur et. al[4]	The use regarding data mining dependent filtering has power to remove the combined noise in far more proficient manner. Author describe the future for enhance data mining dependent fuzzy filter for removing the high density of disturbance.
	2010	V. Christina et. al.[5]	Authors find out multilayer perceptron classifier importance than other classifiers and detect the false positive rate very low compared to other algorithms. The email spam filters using this approach can be adopted either at mail server or at mail client side to reduce the amount of spam messages and to reduce the risk of productivity loss, bandwidth and storage usage.
Genetic Algorithms	2011	J. Sony et al. [8]	Authors use the Genetic Algorithm, ANN, decision trees for the prediction of heart diseases & found decision trees outperform & achieve the accuracy 99.2%
	2013	S. RP. Singh et al.[9]	Authors predict staging of cervical cancer with genetic algorithms.
Decision tree Induction	2012	R. Changala1et. al.	Author use the decision tree approach on different datasets and express their importance and pitfalls.
	2012	M.Singh et. al.[10]	Authors proposed the way to predict the protein function prediction from sequence derived features using See5 tool and design the decision tree through see5 decision rules & attain the accuracy of 64%.

VII. CONCLUSION

This paper provides a review over the efficacy of data mining, and provide the review towards the fields where data mining is broadly applicable these days like in banking, auditing, opinion mining & sentiment analysis, pharmaceutical sales and research, email spam filters, packaging industries, to detect unusual patterns, student's performance analysis and counseling system terrorist activities and fraudulent behavior retail industries, telecom and public sector, and specially in protein prediction, detection of breast cancer, stroke, blood pressure, diabetic, & heart diseases. These are the core areas where lots of data has already been mined. Still lots of areas are unrevealed like hypertension, physical disorder, dentistry, digestive disorder etc. So we can say still lots of efforts are needed to mine the human health related problems.

REFERENCES

1. Ritu Banswal, Vishu, "SPACS: Students' Performance Analysis and Counseling System using Fuzzy Logic and Association Rule Mining", International Journal of Computer Applications, vol 134, no. 3, 2016.
2. N. Koul, S. Hassan "A Survey of Machine Learning Approaches to Opinion Mining & Sentiment Analysis in Text", International Journal of Computer Applications, vol 134, no. 3, 2016.
3. S. ChandraKala, C. Sindhu," OPINION MINING AND SENTIMENT CLASSIFICATION: A SURVEY", ICTACT JOURNAL ON SOFT COMPUTING, VOLUME: 03, ISSUE: 01, OCTOBER 2012.
4. G. kaur, G. kaur," THE STUDY OF DATA MINING TECHNIQUES AND FILTERS FOR IMAGE PROCESSING", International Journal of Advance Foundation And Research In Science & Engineering, Volume 1, Special Issue ,2015.
5. V.Christina, S.Karpagavalli, G.Suganya," Email Spam Filtering using Supervised Machine Learning Techniques", International Journal on Computer Science and Engineering, Vol. 02, No. 09, 2010, 3126-3129.
6. R. Changala1, A. Gummadi, G.Yedukondalu, U. Raju, "Classification by Decision Tree Induction Algorithm to Learn Decision Trees from the class-Labeled Training Tuples", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012
7. N, Handa, "Impact of Data Mining and Data Warehousing in Digital Era", International Conference on Communication, Information and Computing Technology, ICCICT-15, 2015.
8. J. soni, U Ansari, "Predictive data mining for medical diagnosis: An overview of heart diseases prediction", International journal of computer applications, vol. 17, no. 8, 2011
9. S. RP. Singh, G. S. Randhawa, R. S. Virk, "Efficacy of genetic algorithms in staging of cervical cancer ", International journal of cancer research, vol. 47, issue 2, 2013.
10. M. Singh, G. Singh, S. Sharma, "Human protein function prediction from sequence derived features using See5", International journal of scientific & engineering research, volume 3, issue 7, July 2012.
11. A. A Deen, M. Nofal, S. B. Ahmad, "Classification based on association rule based techniques: A general survey and empirical comparative evaluation," Ubiquitous computing and communication journal, Volume 5, No. 3, 2010.
12. M. Singh, G. Singh, Development of predictor for sequence derived features from amino acid sequence using associative rule mining, International journal of computer science and security (2011), vol. 5 issue 1.
13. B. Singh, S. Kaur, T. Kaur," A Review: Data Mining Models for Phylogenetic Tree", International Conference on Communication, Information and Computing Technology (ICCICT-15), 12-13 May, 2015.
14. http://www.zentut.com/data-mining/data-mining-techniques
15. Jiawei Han and Micheline Kamber "Data Mining:Concepts and Techniques", 2nd Ed.
16. http://dataminingwarehousing.blogspot.in/2008/10/data-mining-steps-of-data-mining.html
17. http://en.wikipedia.org/wiki/Data_mining
18. http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm
19. http://ptucse.loremate.com/dw/node/13
20. S. Sharma, " A Review on Efficacy of Artificial Neural Networks in Medical & Business Areas ", International Journal of Recent Trends in Engineering & Research, Volume 02, Issue 04, April 2016
21. S. Sharma, " Cervical Cancer stage prediction using Decision Tree approach of Machine Learning ", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 4, April 2016.

