A STUDY ON DATA MINING TECHNIQUES, METHODS, TOOLS AND APPLICATIONS IN VARIOUS INDUSTRIES

Dr. Neelam Bhardwaj
Research Scholar
OPJS University
Rajasthan.

Dr. **Yashpal Singh**
Associate Professor
OPJS University
Rajasthan.

ABSTRACT

Data Mining is special and important technique utilized to have better business solutions. General survey on Data Mining Techniques, Methods, Tools and Challenges of Data Mining in all the domains, is at most important and so much in demand, applying machine learning concepts and techniques in medical field is also most essential in the present scenario. In this paper the following chapters are presented- various data mining techniques, merits and demerits of data mining, open source data mining tools available, and also domains or industries applied DM.

Key Words: Data mining, DM tools, HealthCare software, DM techniques, Classification, Clustering, Decision tree, DM challenges

INTRODUCTION

The survey conduction and publishing the survey result is most expected one, in this research report. DM Techniques and how much the DM technique contributed in the healthcare field [14] is presented.

II. TYPES OF DM TECHNIQUES:

1. CLASSIFICATION:

It is important DM Technique; Classification predicts a certain output based on a set of pre classified examples and it is the mostly used data mining technique. Classification can be broadly divided into supervised and unsupervised algorithms .Major classification method are decision tree induction, Bayesian networks, linear programming, neural network and fuzzy logic technique. [13][26][27][28][29][30].

2. CLUSTERING:

It is important DM Technique; Clustering groups similar and dissimilar objects. There are number of clustering models which can be used for different applications. It can also be used as a pre-processing approach for attribute subset selection and classification [4].Clustering mainly used for pattern recognition, machine learning and information retrieval.

2.1 K-MEANS CLUSTERING:

It is very Simple clustering approach, less complex method and also efficient. In advance it requires number of cluster to proceed further. It is having problem in handling categorical attributes. It will not predict the cluster with non-convex shape. Outcome varies in the presence of outlier.

2.2 HIERARCHICAL CLUSTERING:

Easy to implement and having good visualization capability. Not necessary to mention the number of clusters in advance. It has cubic time complexity in many cases so it is slower. Once a decision is made it cannot be withdrawn. It will not work proper in the presence of noise. It is not scalable one.

2.3 DENSITY BASED CLUSTERING:

No need to specify number of cluster in advance. It is very simple to handle cluster with arbitrary shape. It will work well in the presence of noise. It will not handle the data points with varying densities. Results will be based on the distance measure.

3. PREDICTION:

It is important DM Technique; The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. Unfortunately many real world problems are not simply prediction. For instance, stock price, sales volumes are difficult to predict, therefore we need more complex techniques like logistic regression, decision tress and neural networks [4].

4. ASSOCIATION:

It is important DM Technique; Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in heart disease prediction as it tell us the relationship of different attributes used for analysis and sort out the patient with all the risk factor which are required for prediction of disease[4].

5. NEURAL NETWORK:

It is important DM Technique; Neural networks models have been studied for many years in the hope of achieving human like performance in several fields. As the human brain consists of millions of neurons that are interconnected by synapses, a neural network is a set of connected input/output units in which each connection has a weight associated with it. The network learns in the learning phase by adjusting the weights so as to be able to predict the correct class label of the input. They are the best at identifying patterns or trends in data and well suited for prediction or forecasting needs .due to their performance neural networks have been widely used in experiments and adopted for critical biomedical classification and clustering problems. Its merits and demerits are it will easily identify complex relationships between dependent and independent variables. Able to handle noisy data, Local minima and over-fitting. The processing of ANN network is difficult to interpret and require high processing time if there are large neural networks[12].

6. DECISION TREE(DT):

It is important DM Technique; Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making. The goal is to create a model that predicts the value of a target variable based on several input variables. Its merits and demerits are there are no necessities of domain knowledge in the construction of decision tree. It minimizes the ambiguity of complicated decisions and assigns exact values to outcomes of various actions. It can easily process the data with high dimension. It is easy to interpret. Decision tree also handles both numerical and categorical data. It is restricted to one output attribute. It generates categorical output. It is an unstable classifier i.e. performance of classifier is depend upon the type of dataset. If the type of dataset is numeric than it generates a complex decision tree[9].

7. SUPPORT VECTOR MACHINE (SVM):

It is important DM Technique; Support vector machine is an algorithm that attempts to find a linear separator (hyperplane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing with interactions among features and redundant features. Its merits and demerits are Better accuracy as compare to other classifier. Over fitting problem is not as much as other methods. Easily handle complex nonlinear data points. It is computationally expensive. The main problem is the selection of right kernel function. For every dataset different kernel function shows different results. As compare to other methods training process take more time. SVM was designed to solve the problem of binary class. It solves the problem of multi class by breaking it into pair of two classes such as one- against-one and one-against- all [9].

8. GENETIC ALGORITHMS (GAS) AND EVOLUTIONARY PROGRAMMING (EP):

It is important DM Technique; Genetic algorithms and evolutionary programming are algorithmic optimization strategies that are inspired by the principles observed in natural evolution. Of a collection of potential problem solutions that compete with each other, the best solutions are selected and combined with each other. In doing so, one expects that the overall goodness of the solution set will become better and better, similar to the process of evolution of a population of organisms. Genetic algorithms and evolutionary programming are used in data mining to formulate hypotheses about dependencies between variables, in the form of association rules or some other internal formalism [12].

9. FUZZY SETS (FS):

It is important DM Technique; Fuzzy sets form a key methodology for representing and processing uncertainty. Uncertainty arises in many forms in today's databases: imprecision, non-specificity, inconsistency, vagueness, etc. Fuzzy sets exploit uncertainty in an attempt to make system complexity manageable. As such, fuzzy sets constitute a powerful approach to deal not only with incomplete, noisy or imprecise data, but may also be helpful in developing uncertain models of the data that provide smarter and smoother performance than traditional systems[12].

10. ROUGH SETS(RS):

It is important DM Technique; in this a rough set is determined by a lower and upper bound of a set. Every member of the lower bound is a certain member of the set. Every non-member of the upper bound is a certain non-member of the set. The upper bound of a rough set is the union between the lower bound and the so-called boundary region

11. K-NEAREST NEIGHBOR (KNN):

It is important DM Technique. It is very easy to implement. Training is done in faster manner. It requires large storage space. It is very sensitive to noise. Testing is slow [14].

12. BAYESIAN BELIEF NETWORK (BBN)

It is important DM Technique, It makes computations process easier. Have better speed and accuracy for huge datasets. It does not give accurate results in some cases where there exists dependency among variables [3].

III. DATA MINING MERITS AND DEMERITS

MERITS:

- It predicts future trends
- It helps in decision making
- It helps to improve company revenue
- It is mainly used in Market Analysis
- It is effectively utilized in fraud detection
- DM techniques applied in health care insurers to detect fraud and abuse.
- DM helps the Physicians to identify effective treatments and best practices through healthcare software's.
- Using data mining it is possible to speed up the work in large data sets.
- DM facilitates generation of quicker reports and faster analysis, which will increase operational efficiency and also diminishes operating cost.
- Data mining can extract predictive information from large database, which is a very important.

DEMERITS:

- Heterogeneity of data volume and complexity will create unnecessary mathematical categorization.
- Must consider Ethical legal and Social issues.
- Dealing Data ownership, Lawsuits
- Privacy and Security of Human Data Administrative Issues – Medical data.
- Other general Security Issues.
- Misuse of information or in accurate information.

IV. OPEN SOURCE DATA MINING TOOLS AVAILABLE:

Table 1: Open Source Tools and its strengths [2][7].

The comparison of DM tools: KNIME, RapidMiner- presented in the table no. 1

Sl No	Attributes	KNIME	RapidMiner
1	Partitioning of dataset to training and testing sets	Only limited partitioning abilities	Less/ Limited partitioning abilities
2	Type	Enterprise Reporting, Business Intelligence	Statistical Analysis, Data mining, Predictive Analytics, Clustering.
3	Scaling	Facility Available	Available with this facility
4	Language and OS	Linux ,OS X, Windows	Cross Platform Language Independent
5	Selection	Wrapper methods	Available with this facility
6	Parameter optimization of machine learning/ statistical methods	Does not have automatic facility	Available with this facility
7	Model validation using crossvalidation and/ or independent validation set	Only Less error measurement methods	Available with this facility
8	Advantages	Molecular analysis, Mass spectrometry.	Visualization, Statistical, Attribute Selection, Outlier detection, Parameter Optimization are all possible.
9	Limitations	Limited error measurements	Requires sound knowledge of dealing with database

Table 2: DM – Open Source Tools and its strengths [2], [7]

Sl No	Attributes	Weka	Orange
1	Partitioning of dataset to training and testing sets	Less partitioning abilities	Limited partitioning abilities
2	Type	Machine Learning.	Machine Learning, Data mining
3	Scaling	Not possible	Not possible
4	Language and Operating System	Cross Platform mainly using Java	Cross Platform mainly using Python C++,C
5	Selection	Possible Partially	Not possible
6	Parameter optimization of machine learning/ statistical methods	Not automated	Automatic facility not possible
7	Model validation using cross validation and/or independent validation set	Partially possible	Partially possible
8	Advantages	Simple to use	Better debugger, Shortest Scripts possible.
9	Limitations	Poor documentation	Limited reporting capabilities

The comparison of DM tools: Weka, ORANGE- presented in the table no. 2

V. DATA MINING TECHNIQUES APPLIED DOMAINS:

Data mining is an interdisciplinary field and with wide diverse applications. There be nontrivial gaps between data mining principles and domain-specific applications, few application domains of Data Mining are listed below,

- Healthcare
- Finance
- Retail industry
- Telecommunication
- Text Mining
- Web Mining
- Higher Education , etc

Tremendous results and reports received in all the above fields by the effective utilization of DM[30].

VI. CONCLUSION:

In this paper we presented various data mining techniques that have been employed for medical data mining. Data mining techniques have higher utility in medical data mining as there is voluminous data in this industry. Due to the rapid growth of medical data, it has become in dispensable to use data mining technique to help decision support and prediction system in the field of health care. This paper has provided the summary of data mining techniques used for all the domains.

ACKNOWLEDGEMENT

We acknowledge the immense help from scholars whose articles are cited and included in references of this manuscript.

REFERENCES

1. Sheetal L. Patil "Survey of Data Mining Techniques in Healthcare" International Research Journal of Innovative Engineering Volume1, Issue 9 of September 2015.
2. Shubpreet Kaur1 and Dr. R.K.Bawa2 "Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System", International Journal of Energy, Information and Communications Vol.6, Issue 4 (2015), pp.17-34.
3. Kittipol Wisaeng, "An Empirical Comparison of Data Mining Techniques in Medical Databases", International Journal of Computer Applications (0975 – 8887) Volume 77– No.7, September 2013.
4. Divya Tomar and Sonali Agarwal, "A survey on Data Mining approaches for Healthcare", International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266.
5. Dhanalakshmi D., Dr. J. Komala Lakshmi, "A Survey on Data Mining Research Trends", International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 3, Issue 10 October, 2014 Page No. 8911-8919.
6. Feixiang Huang, Shengyong Wang, and Chien-Chung Chan, "Predicting Disease By Using Data Mining Based on Healthcare Information System", 2012 IEEE International Conference on Granular Computing.
7. Beant Kaur, Williamjeet Singh, " Review on Heart Disease Prediction System using Data Mining Techniques", International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 IJRITCC | October 2014.

8. Shelly Gupta et al., "Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis", Indian Journal of Computer Science and Engineering (IJCSE) ISSN : 0976-5166 Vol. 2 No. 2 Apr-May 2011.
9. Monali Dey et al, "Study and Analysis of Data mining Algorithms for Healthcare Decision Support System", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 470-477.
10. V. Gayathri et al, "A Survey Of Data Mining Techniques On Medical Diagnosis And Research", Singaporean Journal of Scientific Research(SJSR) Vol.6.No.6 2014 Pp. 301-310.
11. Nidhi Bhatla Kiran Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, October – 2012.
12. Reetu1, Narender Kumar, "Licensed Under Creative Commons Attribution CC BY Medical Diagnosis for Liver Cancer using Classification Techniques", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Volume 4 Issue 5, May 2015.
13. Dhanya P Varghese, Tintu P B, "A Survey On Health Data Using Data Mining Techniques" , International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 02 Issue: 07 | Oct-2015
14. Mohammed Abdul Khaleel Sateesh Kumar Pradham G.N. Dash, "A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases", Volume 3, Issue 8, August 2013 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.
15. K. Rajalakshmi Dr. S. S. Dhenakaran, "Analysis of Datamining Prediction Techniques in Healthcare Management System", International Journal of Advanced Research in Computer Science and Software Engineering Volume 5, Issue 4, April 2015
16. Ms. Shinde Swati B. Prof. Amrit Priyadarshi, "Decision Support System on Prediction of Heart Disease Using Data Mining Techniques", International Journal of Engineering Research and General Science Volume 3, Issue 2, March-April, 2015 ISSN 2091-2730.
17. Umair Shafique, Fiaz Majeed, Haseeb Qaiser, and Irfan Ul Mustafa, "Data Mining in Healthcare for Heart Diseases", International Journal of Innovation and Applied Studies ISSN 2028-9324 Vol. 10 No. 4 Mar. 2015, pp. 1312-1322.
18. Durairaj M, Revathi V, "Prediction Of Heart Disease Using Back Propagation MLP Algorithm", International Journal Of Scientific & Technology Research Volume 4, Issue 08, August 2015 Issn 2277-8616.
19. T.Georgeena.S. Thomas, Siddhesh.S. Budhkar et al, " Heart Disease Diagnosis System Using Apriori Algorithm", Volume 5, Issue 2, February 2015 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.
20. Mohan Kumar S , et al. "Statistical Features Based Classification of Micro calcification in Digital Mammogram using Stocastic Neighbour Embedding", International Journal of Advanced Information Science and Technology, Volume 07, Issue 07 , November 2012- Page Numbers: 20-26, ISSN:2319-2682.
21. Mohan Kumar S , et al. "Breast Cancer Diagnostic system based on Discrete Wavelet Transformation and stochastic neighbour Embedding", European Journal of Scientific Research, Volume 87, Issue 03 , October 2012, Page Numbers:301-310,ISSN:1450-216X.
22. Mohan Kumar S , et al. "Classification of Microclacification in digital mammogram using SNE and KNN classifier", International Journal of Computer Applications, IJCA 03,2013, ISSN: 0975 - 8887.
23. Mohan Kumar S, et al. "The Performance Evaluation of the Breast Mass classification CAD System Based on DWT, SNE AND SVM", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 10, October 2013, Page Numbers: 581-587, ISSN 2250–2459.
24. Mohan Kumar S, et al. "Categorization of Benign And Malignant Digital Mammograms Using Mass Classification – SNE and DWT", Karpagam Journal of Computer Science, Volume-07, Issue-04, June-July-2013 - Numbers: 237-243. ISSN No: 0973-2926.
25. Mohan Kumar S, et al. "Classification of Micro Calcification And Categorization Of Breast Abnormalities - Benign and Malignant In Digital Mammograms Using SNE And DWT", Karpagam Journal of Computer Science, Volume-07, Issue-05, July-Aug, 2013- Page Numbers: 253 to 259, ISSN No: 0973-2926.
26. Mohan Kumar S, et al. "The Performance Evaluation of the Breast Microcalcification CAD System Based on DWT, SNE AND SVM", CiiT International Journal of Digital Image Processing, November 2013, DOI: DIP112013005, ISSN 0974 – 9691
27. Dr. Mohan Kumar S, et al. "Classification of Breast Mass Classification – CAD System and Performance Evaluation Using SSNE", IJISET – International Journal of Innovative Science, Engineering & Technology, 417-425, ISSN 2348 – 7968 Vol. 2 Issue 9, September 2015.
28. Dr. Mohan Kumar S, et al. "Classification of Breast Mass classification – CAD System with Performance Evaluation , International Journal of Engineering And Computer Science", International Journal of Engineering And Computer Science, 14187-14193, ISSN 2319-7242Volume 4 Issue 09 September 2015.
29. Dr. Mohan Kumar S, et al. "Classification of Breast Microcalcification- CAD System and Performance Evaluation Using SSNE", International Journal of Advanced Research in Computer Science and Software Engineering, 824-830, Volume 5, Issue 9, September 2015 ISSN: 2277 128X.
30. R. Jaya et al. "Ayurveda Medicine Roles in Healthcare Medicine, and Ayurveda Towards Ayurinformatics", International Journal of Computer Science and Mobile Computing, Volume 4 Issue 12 publishes on 5th Dec to 30th Dec 2015 ISSN 2320-088X, Paper ID: V4I12201512.