International Journal of Civil Engineering and Technology (IJCIET)
Volume 8, Issue 12, December 2017, pp. 1013–1020, Article ID: IJCIET_08_12_110
Available online at http://http://www.iaeme.com/ijciet/issues.asp?JType=IJCIET&VType=8&IType=12
ISSN Print: 0976-6308 and ISSN Online: 0976-6316

© **IAEME** Publication **Scopus** Indexed

AN OVERVIEW OF DATA MINING TECHNIQUES AND ITS APPLICATIONS

Kavitha G and Dr. Elango N.M
School of Information Technology and Engineering,
VIT, Vellore, Tamilnadu, India

ABSTRACT

Storing the enormous amount of raw data into database will not be able to provide the meaningful information, rather those data should be analyzed and the hidden knowledge must be extracted by the use of datamining which is a main phase of the knowledge discovery process. This paper aims to explore information related to various datamining techniques and their relevant applications. Specifically it will elaborate the information regarding preprocessing and post processing steps in datamining techniques such as association rule mining, clustering, classification, neural networks, visualization and their applications such as e-services, education, business, security and agriculture will be discussed along with issues and challenging tasks.

Key words: computer science, data mining, knowledge discovery.

Cite this Article: Kavitha G and Dr. Elango N.M, An Overview of Data Mining Techniques and its Applications. *International Journal of Civil Engineering and Technology*, 8(12), 2017, pp. 1013-1020.
http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=8&IType=12

1. INTRODUCTION

Ask Handling huge amount of data where it tends to tedious task but with the help of statistical tools we can make a assumption driven approach such that the hypothesis are formed and then it has been validated over the data. But in later process we need to discover knowledge from the given data hence the trend has been moved towards the datamining where it can be as a discovery driven approach in which pattern and hypothesis are been generated from data. Rather we extracting useful pattern for the data it should be translated into logical rules where this specify that dataminng is human centered approach when it was defined by Goebel and Gruenwald[1] they state that knowledge discovery in databases(KDD) as the non-trivial process of identifying valid, novel ,potentially useful and ultimately understandable patterns in data and data mining as the extraction of patterns or models from observed data.

States the major steps in datamining are such as data gathering, data sorting, data cleaning and finally loading the filtered data into the datawares house and those filtered data can be used to discover the useful pattern or information which can be done in preprocessing steps of data mining which is discussed in sections . so processing the data in dataminnig can be done with the help of certain techniques which is discussed in the section 3 and similarly the real time applications of datamining is specified in section 4 along with this future challenges in each application has been stated.

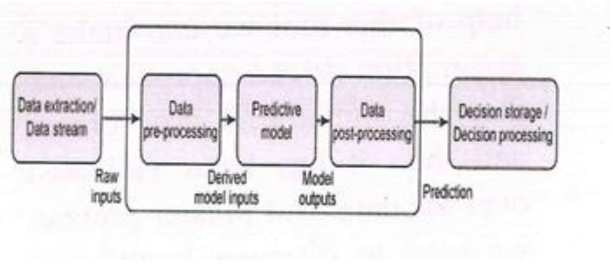

Figure 1 Phases of data mining process

2. DATA MINING PROCESS STEPS

The preparation of data will act as a phase in the datamining process but when we consider on applying the datamining algorithm over on the real world application then the preparation of data will be a major process. Preparing data for mining we need certain set of preprocessing steps[2] through this we can make prediction over on the prepared datasets.

2.1. Preprocessing Steps

It is the steps where it should be taken for preparing the data before applying data mining algorithm for mining process the following steps in Figure 2.

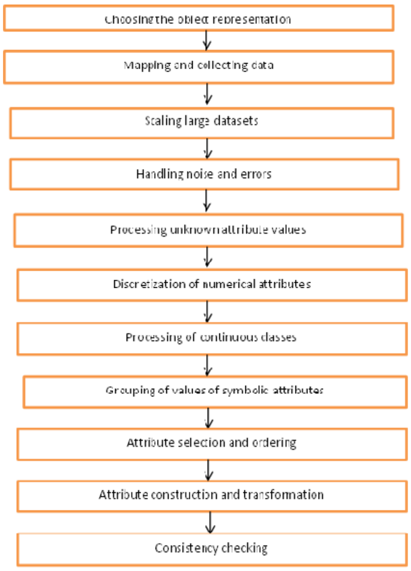
Choosing the object representation
Mapping and collecting data
Scaling large datasets
Handling noise and errors
Processing unknown attribute values
Discretization of numerical attributes
Processing of continuous classes
Grouping of values of symbolic attributes
Attribute selection and ordering
Attribute construction and transformation
Consistency checking

Figure 2 Preprocessing steps for mining data.

2.2. Choosing the Object Representation

An object will be depicted as a unit of the given problem where it should be described by as collection of elementary descriptions such as an elementary descriptions such as an attribute representations and their values should be appropriate for choosing object representations.

2.3. Scaling Large Datasets

Generally the algorithms are assumed that the data stored in main memory and those algorithms do not pay attention on how to deal with large databases thus we use some kind of methods such windowing, batch-incremental mode.

2.4. Handling Noise and Errors

The errors could be handle with both internal (poor prosperities of learning data mining system) and external (random errors, noise errors) we need a to set criteria on hangling those type of error.

2.5. Processing known Attribute Values

There are certain attribute values it mismatch with other type of data or some values may be missing so suggestion on those kind of data is considered don’t care values

2.6. Discretization of Numerical Attribute

Some of the algorithm which may apply only for symbolic or categorical data but the real world problems which will contain both type of data such that he need to discretize the numerical attributes.

2.7. Processing of Continues Classes

In general we consider classes for mining but there will be more challenges on continuous classes such that we can process it by either off-line or online splitting.

2.8. Grouping

The attributes which are been obtained from the given datasets may contain multiple values for single attribute hence we need to group based upon their subjects.

2.9. Attribute Selection and Ordering

The attribute which are been selected from the data are not be so informative hence we need a process fo selecting only relevant attribute data then we should select relatively small subset of selected attributes.

2.10. Attribute Construction and Transformation

The attribute which are derived from datasets muse be adapt with the target problem which we have defined if it is not possible we need the capable of attribute construction and transformation into difficult models based upon the problem.

2.11. Consistency Checking

In this we need to identify the inconsistency in the data process such that we neither use preprocessor nor loop facility of the knowledge discovery process. The above steps are been used of preprocessing the given data in order to mine the process

2.12. Post Processing Steps

Once preprocessing steps are been over then we determine that the filtered data are ready for mining so that we could identify the hidden information but after use of mining techniques the least step in data mining is post processing setps[2] where here we are verifying the discovered knowledge is apt for the target problem the following steps are in Figure 3

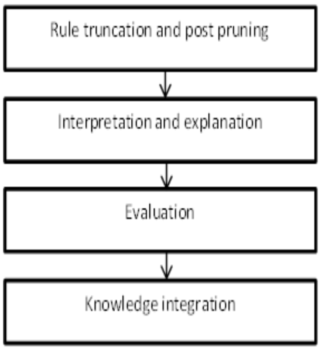
Rule truncation and post pruning
Interpretation and explanation
Evaluation
Knowledge integration

Figure 3 Post processing steps

2.13. Knowledge Filtering: Rule truncation and post processing

When we make a verification process over on the training data, if the datas are supposed to be too noisy then by the use of data mining algorithm generating the decision rules which will be in inconsistent manner so that we either post process or pruncation the discovered rules

2.14. Interpretation and Explanation

Once the acquired knowledge has been derived then those has to be use by end user and they may expect either documentation or visualize the knowledge and moreover we should incorporate domain specifies knowledge for given task.

2.15. Evaluation

Once the hypotheses(models) that are generated by the process then we need to evaluate(test) it by setting up the criteria such as classification accuracy.

2.16. Knowledge Integration

Previously the decision making system will dependent on single model but how ti has the capability to single model but how it has the capability to combine or refine results obtained from several models of the process.

3. DATA MINING TECHNIQUES

Till know we have seen how they are making preprocessing for the given set of datasets, then we mine the dataset for identifying hidden pattern and then finally how we are post processing the discovered knowledge. But the intermediate step is mining he datasets is can be done by the use of data mining techniques in this section we are going to elaborate on those techniques.

3.1. Association Rule Mining

Through this technique we could find the relationship between the attributes and tuples. The pattern which we discover will contain a confidential value and support value. Is is also used to identify the degree e of dependence of the given data and they can be found using two values are support and confidence value. Support is the ration of the number of item-sets satisfying both antecedent(LHS) and consequent(RHS) to the total number of transactions[2].Confidence is derived from a subset of the transactions in which two entitites are realate[2].algorithms such as Apriori, TIDFP Growth, Scalable Algorithm, PDM, CCPD, CHX are used to derive the rules for mining process.

3.2. Clustering

It is a technique in which we can made as a group of similar subset into different classes from the large amount of data, it is useful [4] dimensionality reduction, cluster sampling and to look for patterns from the derived subsets which is from the given datasets. It can be classified into two types[3], the first one is hierarchal clustering where we build a tree of classes called a dendogram, each node in the tree represent a class and the other one is un hierarchal clustering is supposed to be build the classes that are not subclasses of each other. The various clustering algorithm that we are using are Hierarchal clustering algorithm[4],agglomerative, divisive algorithm, portioning algorithm, density based methods some of the application are marketing, insurance, medical sciences and miscelious applications. In this we attend to organize or categorizes by the way of assigning a new element to one of the predefined classes. The methods[5] that are derived can be categorized into decision treem nearest neigh hour, probabilistic models.

3.2.1. Decision tree

It is a nonparametric classification and prediction models. Organized in the form of a rooted tree with two types of nodes called decision nodes and class nodes. It can be represented under the it-ten-else rule condition.

3.2.2. Nearest Neighbor

It typically defines the proximity between instance and then assigns it the label for the majority class of its neighbors.

3.2.3. Probabilistic Models

Which calculate probabilities for hypotheses base on bayes theorem. The various algorithm such as ID3,C4.5

3.3. Neural Network

An Artificial neural network is a system made up of a several simple calculating units functioning in parallel. The various applications are speaker identification, biometrics miscellious applications such as weather prediction, traffic prediction re-routing in transportation and communication network, customer ranking, medical imaging, automatic defects detecting in manufactured items.

3.4. Visualization

Is a process of transforming data, information and knowledge into visual representation and provides on interface between two information processing. With the help of this technique we make a communication medium between the users and computer. And also using computer graphics and other techniques that can be used as samples, variables for find the relation between them. This can be used by the field such as scientific computing and other one is information visualization. It used multidimensional scalar data but the datamining which is depicted on information visualization. Some of the emerging techniques are drill-down, virtualization technique are been used.

3.5. Support Vector Machine

Is a relatively based on the [4] latest advance in statistical learning theory, it will provide simple ideas and a clear intuition about the datasets. It gives high performance on written character recognintion, image classification and bioscience analysis.

4. APPLICATIONS

4.1. E-services

Now day’s people like to makes their financial transactions, services processed through online are act as a major role in the society. But we also need to solve kind of problem such as web mining[7] in order to extract information from the web and also mining on web intelligence. The various e-services are e-banking, search engines, online auctions, social networking blog analysis are can be applicable by the use of data mining where we are mining the useful pattern based upon the services.

4.2. Health Care

In Medical Application there have been various techniques are successfully applied and it caused to identify the diagnosis or prognosis of certain diseases and their relevant medicines, By the use of datamining we can predict the minimal causes of diseases by giving the symptoms of each disease as data into the system.

4.3. Education

Our educational system has been widely improved either on student perspective or faculty perspective[8]. Based upon the student attitude, interests are been utilized for predicting on making better evaluation of the students and based on faculty we can use datamining for enhancing various teaching methods and quality offered study materials.

4.4. Business

The major aspects of business is design the product and their manufacturing steps and benefits. That was attained by the end user which depict on the feedback or product satisfaction. The datamining techniques are been used[9] in production process, fault detection, maintenance, decision support, product quality improvement, job-shop scheduling, load time estimation.

4.5. Security Applications

When we speak about the security it specify[10] national security(surveillance) along with cyber security(virtualization). If we consider the national security they include attacking buildings and destroying infrastructure of vaious sectors are power, telecommunication. Data mining are applied for identifying either individual or groups capable of carrying out terrorist activities. Another kind of security is cyber security where data and auditing. The real-time threads can be overcome by within time limit by the use of techniques of datamining. Threats may refer to cause damage the information about data of an organization either from outside or inside it can be leveraged to detect and also prevent such attacks.

4.6. Agriculture

As we know already that india is well known for the agriculture products and its development is mainly focused how to improve our agriculture based products. In general the geospatial data [11] which comprises of agricultural operations are regarding crops, family pesticides, weather parameters, climatic condition, flexibility of soil and these data can be applied to frame a model through data mining techniques. This model which depict on what kind of climatic conditions are been attained. Another important aspect is about identifying what kind of pesticides can be applied for particular type of crop which we are cultivating[12,13,14]. It is done through occur datamining technique such as data clustering, classification correlation and association rules are used for pattern identification based upon a spatial data related with crop cultivation and as well as analyzing on climatic changes[15,16,17].

5. CURRENT TRENDS IN DATA MINING

In the above section we have been some of datamining technique which is used for mining the information pattern such that now we consider the current trends that are multimedia mining, text mining, image mining, video mining and web mining.

5.1. Multimedia Mining

The tasks which consists of combination of different data types such as text, audio, video hence mining of such kind information will be the current trend. These kind of data but there are some kind of application which will consider only text type of data are electronic book, articles, unstructured free text it is meaning for extract useful information.

5.2. Image Mining

Here they consider image databases such that is entirely differ from structured data types and this especially preferred by medical field by on searching, retrieving and comparing of query image with the stored image.

5.3. Web Mining

Due to the complexity of the web pages and dynamic nature of data stored in the interest we need a techniques for mining such information it is said to be webminnig where it automatically retrieve, extract and evaluate.

6. FUTURE DIRECTIONS

In this section we would like to explore some kind of future research directions based upon of applications of dataminng, initially when we need a database are constantly updated and adding new information into an existing database, A method need for incrementally updating and producing a incremental pattern where updating should be deliberated. When we consider education field we need to develop methods that which incorporate to identify student perspective on their academics.

Business perspective how to improve quality to product act as a research area along with all kind of business transactions, services lead to future research work. In Agriculture how we can improve the quality of product based upon desired pattern on pesticide and how to manage the climatic changes, natural calamities

7. CONCLUSIONS

At the end of this we are concluding that through this paper we could understand purpose of emergence of datamining, their techniques along with relevant applications. We also examine about the information that how datamining can be applicable to that application. Then we moved forward to current trend of mning techniques based upon the type of datasets that are been process and finally we specify about the future research directions in datamining.

REFERENCES

[1] M.Geobel and L.Gruenwald, A survey of data mining in education environment and knowledge discovery software tools, SKIDD Explorations,1999 june,978-953-184111.

[2] “Insight into Data Mining Theory and practice”, K.P Somam, shyam Diwakar, V.Ajay., PHI Learning Pvt. Ltd.,2006,ISBN-81203289 73,9788120328976

[3] Faouzi Mhamdi and Mourad Elloumi,” A New Surevey on knowledge Discovery and Data Mining”, 2008 Second International Conference on Research Challenges in Information Science, 427 432, DOI: 10.1109/RCIS.2008.4632134

[4] “Data Mining Methods”,Rajan Chattamvelli, Data Mining Methods”,Rajan Chattamvelli, Alpha Science International, 2009, ISBN 184265523X,9781842655238

[5] www.SciRP.org/journal Date accessed: 20/5/2016

[6] Kamalakannan, S. "G., Balajee, J., Srinivasa Raghavan.,“Superior content-based video retrieval system according to query image”." International Journal of Applied Engineering Research 10, no. 3 (2015): 7951-7957

[7] Hailiang Jin, Huije Lu, Research on Visualization Techniques in Data Minning”,IEEE, Computational Intelligence and Software Engineering, 2009. CiSE 2009,DOI: 10.1109/CISE.2009.5365927 .

[8] www.emeraldinsight.com/1468-4257.html Date Accessed 25/6/2016.

[9] M.Vranic, D.Pintar, A.Skocir, “The use of datamining in education”,2007

[10] Kamalakannan, S. "G., Balajee, J., Srinivasa Raghavan.,“Superior content-based video retrieval system according to query image”." International Journal of Applied Engineering Research 10, no. 3 (2015): 7951-7957.

[11] Keqin Wang, Shurong Tong, Fourth International Conference on Fuzzy Systesms and Knowledge Discovery, 2007, Volume:4,Pages: 613 -618, DOI: 10.1109/FSKD.2007.482

[12] Bhavani Thraisingham, Latifur Khan, Mohammad M.Msuad,KevinW.hamlen,978-0-7695-3492,2008.

[13] Amiya Kumar Tripathy, Adinne and D.Sudharsan,”Geospatial Data Mining 2009 ,17th International Conference on Geoinformatics Pages: 1DOI: 10.1109/GOINFORMATICS.2009.5293296

[14] Zhongfeizang, florentmassegila, Ramesh jan,Albert Del Bimbo, Introducion to the Special Issue on Data Mining”, International

[15] Agriculture Pest Management aFramework”, Journal of Advanced Research in Data Mining and Cloud Computing Vol.1, Issue 1, July 2013.

[16] Ranjith, D., J. Balajee, and C. Kumar. "In premises of cloud computing and models." International Journal of Pharmacy and Technology.

[17] Jeyakumar and BalaKrishnan P, Action Recongnition in Video Survillance Using Hipi and Map Reducing Model, International Journal of Mechanical Engineering and Technology 8(11), 2017,pp. 368–375.

[18] M.A. Shanti and K. Saravanan, Knowledge Data Map - A Framework for the Field of Data Mining and Knowledge Discovery. International Journal of Computer Engineering & Technology , 8(5), 2017, pp. 67 – 77

[19] Dr. E V Ramana, S Sapthagiri and P Srinivas, Data Mining Approach for Quality Prediction of Injection Molding Process Through Statistica SVM, KNN and GC & RT Techniques. International Journal of Mechanical Engineering and Technology, 7(6), 2016, pp. 22–30.

[20] Senthil Kumar J, Venkataraman V, Meganathan S and Meena V, A Data Mining approach to classify Higher Education Sector data using Bayesian Classifier, International Journal of Mechanical Engineering and Technology 8(9), 2017, pp. 95–103.