International Research Journal of Multidisciplinary Science & Technology
www.irjmrs.com ISSN : 2455-930X

Tools, Methods and Algorithm for Data Mining Operations

MAI Navid[1], NH Niloy[2]
[1]Ruhea College, Bangladesh Corresponding author's
E-mail: niloynh1997@gmail.com

Abstract - *Getting useful patterns, data and trends is the purpose of Data mining by using classification, regression, clustering and association. Many tools that support these techniques and algorithms are available in various applications. This paper summarizes different tools and algorithms of Data Mining. To find the tool to use for the user's need and requirement, comparisons has been done for different tools and validation of the outcomes of the data is also summarized.*

Key Words: Classification, Clustering, Regression, Data Mining, Algorithms.

1.INTRODUCTION

The process of cleansing the data is also called as noise elimination or noise reduction or feature elimination (PhridviRaj and GuruRao, 2013). Data mining is the process of extracting useful information. Data mining is a part of a bigger framework, referred to as knowledge discovery in databases (KDD) that covers a complex process from data preparation to knowledge modeling (Srivastava, 2014). Basically it is the process of discovering hidden patterns and information from the existing data. In data mining, one needs to primarily concentrate on cleansing the data so as to make it feasible for further processing. This can be done by using various tools available supporting various techniques. The important consideration in data mining is whether the data to be handled static or dynamic. In general, static data is easy to handle as it is known earlier and stored. Dynamic data refers to high voluminous and continuously changing information which is not stored earlier for analyzing and processing like static data. It is difficult to maintain dynamic data as it changes with time. Many algorithms are used to analyze the data of interest. Data can be sequential, audio signal, video signal, spatio -temporal, temporal, time series etc. Main data mining task is classification which has main work to assign each record of a database to one of the predefined classes. The next is clustering which works in the way that it finds groups of records instead of only one record that are close to each other according to metrics defined by user. The next task is association which defines implication rules on the basis of that subset of record attributes can be defined. Data mining is the main important step to reach the knowledge discovery. Normally for data preprocessing it goes through various process such as data cleaning, data integration, data selection and data transformation and after these it is prepared for mining task. Its main contribution is in the fields of traditional sciences as astronomy, biology, high engineering physics, medicine and investigations. Various algorithms and tools can be used according to the application as given by Soni and Ganatra (Soni and Ganatra, 2012). According to JSTOR the term data clustering first appeared in the title of a 1954 article dealing with anthropological data (Demšar and Zupan, 2013). The cluster analysis is as old as a human life and has its roots in many fields such as statistics, machine learning, biology and artificial intelligence. Cluster analysis is therefore known as differently in the different field such as a Q-analysis, typology, clumping, numerical taxonomy, data segmentation, unsupervised learning, data visualization, learning by observation etc. (Jain *et al*., 1999; Han and Kamber, 2001 and Rao, 2003). To analyze large amount of data, data mining came into picture and is also called as KDD process. To complete this process various techniques developed so far are explained in this section. KDD is the overall process which is shown in figure 1:

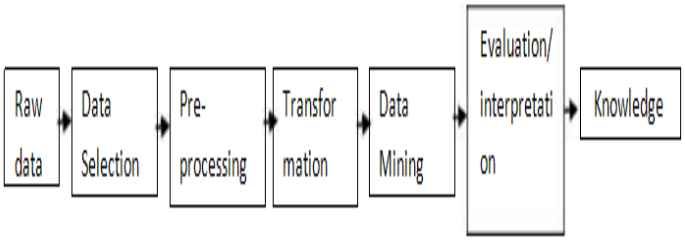

Fig.1 Knowledge Discovery Process (Mitra *et al*., 2002)

In KDD the main and important step is data mining. KDD will turn the low level data into high level data. Data mining is the filed in which useful outcome that is being predicted from large database. It uses already built tools to get out the useful hidden patterns, trends and prediction of future can be obtained using the techniques. Data mining involves model to discover patterns which consists of various components.

2.Classification

Classification is a supervised kind of machine learning in which there is provision of labeled data in advance. Classification is one of the data mining technique which is useful for predicting group membership for data instances. By providing training the data can be trained and we can predict the future of data. Prediction is in the form of predicting the class to which data can belong. Training is based on the training sample provided. Basically there are two types of attributes available that are output or dependent attribute and input or the independent attribute [9]. In the supervised classification, there is mapping of input

data set to finite set of discrete class labels. Input data set X € Ri , where i is the input space dimensionally and discrete class label Y € 1.....T, where T is the total number of class types. And this is modeled in the term of equation Y=Y(x, w), w is the vector of adjustable parameters.

2.1.Classification methods in data mining are as follows:

- Decision tree induction: From the class labeled tuples the decision tree is build. Decision tree is tree like structure in which there are internal node, branch and leaf node. Internal node specifies the test on attribute, branch represents the outcome of the test and leaf node represents the class label. Two steps that are learning and testing are simple and fast. The main goal is to predict the output for continuous attribute but decision tree is less appropriate for estimating tasks. There may be errors in predicting the classes by using decision tree approach. Pruning algorithms are expensive and building decision tree is also an expensive task as at each level there is splitting of node.
- Rule- based classification: It is represented by set of IF-THEN rules. First of all how many of these rules are examined and next care is about how these rules are build and can be generated from decision tree or it may be generated from training data using sequential covering algorithm. Expression for rule is:
- IF condition THEN conclusion Now we define accuracy and coverage of S by following expression (Baker and Yacef, 2009).

  Coverage (R) = /I I

  Coverage (R) = /

- Classification by backpropagation: Backpropagation is a neural network learning algorithm. Neural network learning is often called connectionist learning as it builds connections. It is feasible for that application where long times training is required. The most popular neural network algorithm is backpropagation. This algorithm proceeds in the way that it iteratively performs processing of data and it learns by comparing the results with the target value given earlier.
- Lazy learners: Eager learner is the form in which generalization model is being developed earlier before new tuple is being received for classifying. In lazy learner approach when given a training tuple it simply stores it and waits until a test tuple is given.It supports incremental learning. Some of the examples of lazy learner are K-nearest neighbor classifier and case-based reasoning classifiers (Kumar *et al.*, 2012).

3.Clustering

Unsupervised classification that is called as clustering or it is also known as exploratory data analysis in which there is no provision of labeled data. The main aim of clustering technique is to separate the unlabeled data set into finite and discrete set of natural and hidden data structures. There is no provision of providing accurate characterization of unobserved samples that are generated from by same probability distribution (Zhao and Fränti, 2014; Rui and Donald, 2005). Broadly clustering has two areas based on which it can be categorized as follows:

- Hard clustering: In hard clustering same object can belong to single cluster.
- Soft clustering: In this clustering same object can belong to different clusters.

Given there is set of input patterns Y= {y1, ... , yi, ... yN},where yi= (yi1, ... yid)T € Rd and each is yjd known as variable, feature, dimension or attribute.

- Hard partitioning gives result: C= {C1,.....,CK} where (K≤N) and

  i. Ci ≠ ϕ, i=1, 2,...., N
  ii. ∪i=1 K =
  iii. ∩= ∅, , = 1,2, … . .≠

Hierarchical clustering has different perspective of representing the output that is tree like structure, partition of Y, = 1, … … *h* ≤ ∈ ∈ and l > m imply ∈ , ≠ , = 1,2, … .

3.1.Clustering Process

The clustering process includes various steps and it is a step by step process in which the results can be verified. The main four steps followed are as below:

- Feature selection or extraction: As pointed out by (Jain *et al.*, 1999), feature selection is selecting distinguishing feature form set of candidates and extracting means which it utilizes in the transformation to generate the useful and novel features from original ones (Kleinberg, 2002).
- Clustering algorithm design: Every clustering algorithm is affected by measures. Next is to optimize the clustering solutions. As said by J. Klienberg that " It has been very difficult to develop a unified framework for reasoning about it (clustering) at a technical level, and profoundly diverse approaches to clustering" (Jain and Dubes, 1988).

- Validation: Validations of clusters are in the sense whether the groups formed are valid or not, the data is correctly identified according to groups. These all can be checked by main three indices which are known as testing criteria and these are as follows:

    i. External indices
    ii. Internal indices
    iii. Relative indices

These indices are defined on different clustering structures that are known as partitioning clustering, hierarchal clustering and individual clusters (Abbas, 2008).

- Result interpretation: Next step is to provide accuracy to user and provide a meaningful insight form original data so that efficient results can be provided.

3.2.Methods of clustering

Broadly clustering methods can be divided into two main categories which have number of instances. There are various methods for clustering which act as a general strategy to solve the problem and to complete this, an instance of method is used called as algorithm. On the basis of that we have hierarchical and partitioning based methods. In hierarchical based clustering, the data sets of n elements are divided into hierarchy of groups which has tree like structure. In partitioning based methods the output is like k partitions of N dataset elements.

Hierarchical methods: There is a tree like structure in this method. There are two approaches which are agglomerative and divisive (Kotsiantis and Pintelas, 2004).

- Agglomerative is also known as bottom up approach
- Divisive is the top down approach

3.3.Partitioning methods:

This method simply partitions the dataset into n objects. K-partitions with n objects such that ≤. Different types of approaches are there:

- Grid based method uses grid data structure and at each step grid like structure is being followed (Jain, 2010).
- Subspace based uses subspace of actual document and its main aim is to work with high dimensional data.
- Density based, its general concept is to increase the given cluster to cover the neighborhood exceeds some threshold value.
- Relocation based methods have strategy on the conceptual point of view in which it identify the unknown parameters of the clusters (Kleinberg, 2002; Rao and Nagaraj, 2014).

Various clustering algorithms are discussed so far. These all are compared based on parameters differentiating them like the algorithms supported, type and size of dataset supported.

3.4.Regression

Regression is another data mining technique which is based on supervised learning and is used to predict a continuous and numerical target. It predicts number, sales, profit, square footage, temperature or mortgage rates. All these can be predicted by using regression techniques. Regression starts with data set value already known. It is based on training process. It estimates the value by comparing already known and predicted values. These values can be summarized in some model (Sansgiry *et al.*, 2006). Error is also called as residual which is difference between expected and predicted value. Main aim is to reduce the error so that we get with accurate result.

3.5.Regression techniques or methods:

There are two types of regression techniques namely linear and non –linear.

- Linear regression: Linear regression is used where the relationship between target and predictor can be represented in straight line. = 1 x + P2 + e Multivariate linear regression: The regression line cannot be visualized in two dimensional space. = 1 + 2 1 + 3 2 + ⋯ + − 1 +
- Non- Linear Regression: In this case non linear relationship can be there and this cannot be represented as straight line. This can be represented as linear reaction by preprocessed the data.

4.WORK DONE ON APPLICATIONS OF DATA MINING

There is a study of various factors that affect academic performance and for that the data of pharmacy students have taken focusing on which will help students to improve their performance (Kriegel *et al.*, 2007). Data mining techniques are used in many applications. The effect and future trends have been stated. Many users have designed prediction systems using these techniques. A paper by Radaideh and Nagi (2012) focuses on building the classification model to predict the performance of employees. Many factors have been included and on the basis of that the experiment has done by Vijiyarani S, Sudha S (2013). Another paper by Velmurugan (2014) is on the prediction of diseases as heart diseases, diabetics etc. by using data mining techniques. By using classification techniques like decision tree, naïve bayes a prediction model is designed (Huang, 1998).

Use of K-means algorithm is very useful in designing many applications. Extension of K-means algorithm can be done to improve the performance (Ngai *et al*., 2011). In a paper by Ngai et al. a review of the classification scheme for the application of financial fraud detection using data mining technique is done (André *et al*., 2011. A survey by André et al. shows different perspectives that in the data obtained by partitioning done by clustering ensembles, data can be improved by applying more steps and this all could be done through genetic programming approach (Aviad and Roy, 2012). As in unsupervised learning, there is no target attribute known in advance and there may be some time no comparison and correction in building groups. So to improve this new concept come into picture that is bounded rationality to reveal feature saliency in clustering problem designed (Combes and Azema, 2013). The new approach is being introduced for elder people living in old age homes to improve their way of living and to improve their health standards (Sandeep *et al*., 2014). Comparison of various partition based clustering algorithms is done to distinguish among type of algorithms best suited for user's application (Oyelade *et al*., 2010). Analysis of student performance can also be done by Kmeans algorithm where the predicting power of clustering algorithms and Euclidean distance for sum of squared errors, again academic data is taken and algorithms are applied (Rao and Ramachandra, 2014; Adhikari and Rao, 2008). For validation of clusters different types of parameters are identified on the basis of which clustering is done and relation between WB, Xu and Calinski-Harabasz index (Rui and Donald, 2005). On large dataset the factors that affect performance can be taken care. This study is related to improve the shortcomings of csiFCM i.e. cluster size intensive fuzzy c mean algorithm. New method introduced is slibFCM i.e. cluster size insensitive integrity based FCM method (Jacques and Preda, 2014). For multivariate functional data the new model based clustering algorithms being proposed (Angelis and Dias, 2014). Using hybrid clustering approach mining of categorical sequences from data can be done (Xiao and Fan, 2014). A paper by Xiao and Fan focus on analyzing the large data in BAS building automation system and also improve the building operational performance (Irpino *et al*., 2014). One of the paper works for histogram data by using Dynamic Clustering Algorithm with an automatic weighting step of the variables by using adaptive distances (Liu *et al*., 2014). Different type of prediction model for internet user are also proposed. Novel link prediction that is super edge prediction is being applied to create a super network model introduced by Liu *et al*. (2014).

5.DATA MINING TOOLS

There are various open source tools available for data mining. Some of tools work for clustering, some for classification, regression, association and some for all. There are various algorithms for each technique as discussed in section 2. This section describes features of different tools and which tools can be used to implement which algorithm.

5.1.Tools features

(a) Tool 1-Orange Orange is the Open source data visualization and analysis tool. Data mining is done through visual programming or Python scripting. Regression method is also being used in Orange where ensembles are basically wrappers around learners (Demšar and Zupan, 2013).

(b) Tool 2- WEKA WEKA stands for Waikato Environment for Knowledge Analysis. It is developed in Java programming language. It contains tools for data preprocessing, classification, clustering, association rules and visualization. It is not capable for multi relational data mining. Data file can be used in any format like ARFF (attribute relation file format), CSV(comma separated values), C4.5 and binary and can be read form a URL or from SQL database as well by using JDBC. One additional feature is that data sources, classifiers etc are called as beans and these can be connected graphically (Srivastava, 2014).

(c) Tool 3-SCaVis Scientific Computation and Visualization Environ-ment. It provides environment for scientific computation, data analysis and data visualization designed for scientists, engineers and students. The program incorporates many open source software packages into a coherent interface using the concept of dynamic scripting. It provides freedom to choose a programming language, freedom to choose an operating system and freedom to share code. There is provision of multiple clipboards, multidocument support and multiple Eclipse-like bookmarks Extensive LaTeX support: a structure viewer, a build-in Bibtex manager, LaTeX equation editor and LatexTools (Basili *et al*., 1999).

(d) Tool 4- Apache Mahout Its goal is to build machine learning library scalable to large data set. For Classification following algorithms are included: Logistic Regression, Naive Bayes/ Complementary Naive Bayes, Random Forest, Hidden Markov Models, Multilayer Perceptron. For Clustering following algorithms are included: Canopy Clustering, k-Means Clustering, Fuzzy k-Means, Streaming k-Means, Spectral Clustering by Sean Owen and Sebastian Schelter (Ferrucci and Lally, 2004).

(e) Tool 5- R Software Environment R provides free software environment for statistical computing and graphics mostly for UNIX platforms, Windows and MacOS. It is an integrated suite of software facilities like data manipulation, calculation and graphical display. It provides a wide variety of graphical techniques as well as statistical like linear and nonlinear modeling, classical statistical tests, classification, clustering (Baker and Yacef, 2009).

(f) Tool 6- ML Flex ML uses machine learning algorithms to derive models from independent variables with the purpose of predicting the values of a dependent (class) variable.

(g) Tool 7- Databionic ESOM (Emergent Self Organizing Maps) tool On can do Preprocessing, Training, Visualization, Data analysis, Clustering, Projection, Classification using this tool. Training data is set of points from a high dimensional space called data space.The two most common training algorithms are online and batch training. Both of these training algorithms will search the closest prototype for each data point that is best match. Online training, there is immediately update of best match but in batch training all the best matches are being collected and then update if performed collectively (Baker and Yacef, 2009).

(h) Tool 8-NLTK (Natural Language Tool Kit) NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora. It also provides lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning. NLTK is available for Windows, Mac OS X, and Linux. NLTK is a free, open source, community-driven project. It defines various classifier classes: Conditional Exponential Classifier, DecisionTree Classifier, Maxent Classifier, NaiveBayes Classifier, Weka Classifier (Oyelade *et al*., 2010).

(i) Tool 9-ELKI (Environment for Developing KDDApplications Supported by Index- Structures) ELKI is open source data mining software written in Java. The focus of ELKI is research in algorithms, with an emphasis on unsupervised methods in cluster analysis and outlier detection. ELKI offers many data index structures such as the R*-tree that can provide major performance gain and in order to achieve high performance and scalability. The approach used is the independence of file parsers or database connections, data types, distances, distance functions, and data mining algorithms (Padmaja and Fatima, 2013; Low *et al*., 2010).

(j) Tool 10-UIMA (Unstructured Information Management Architecture) diagram Large amount of unstructured information can be analyzed to get relevant information. It enables application to be decomposed into components. Working of framework is to manage these components and flow between them. Basic availability is frameworks, components and infrastructure.

(k) Tool 11-GraphLab GraphLab has several algorithms already implemented in its toolkit. One can also implement one's own algorithm on top of our graph programming API.

(l) Tool 12-mlpy machine learning Python It has algorithms of regression and classification. Cluster analysis can also be done for dimensionally reduction and wavelet transform. Various different algorithms like feature ranking, resampling algorithm, peak finding algorithm, error evaluation are also available.

(m) Tool 13-KEEL (Knowledge Extraction Evolutionary Learning) KEEL is open source. It uses java software which have license of GPLv3(General Public License version 3). It allows users to have the access of behavior of evolutionary learning and basic soft computing based techniques for various kinds of data mining problems to be handled.

(n) Tool 14-Scikit-learn Scikit-learn is also a free package. It is in Python which extends the functionality of NumPy and SciPy packages. It also uses the matplotlib package for plotting charts. The package supports most of the core DM algorithms except including classification rules and association rules.

5.2.Comparison of various tools on different perspectives

Different factors on which categorization of tools have been stated below:

Table 1. General introduction of tools

Tool	Aim
Orange	Visual data analysis
WEKA	General ML package
Kernlab	Kernel based classification/ Dimensionality reduction
Dlib	Portability, correctness
Nieme	Linear regression, Classification
Java-ML	Feature selection
pyML	Kernel methods
Shogun	General Purpose ML Package with particular focus on large scale learning; Kernel Methods;
Mlpy	Basic algorithms
Torch7	Neural networks
Pybrain	Reinforcement learning
Scikitlearn	General Purpose with simple API /scipy idioms

Table 2. Comparison on the basis of general features

General	Features Tools
GUI	Weka,dlib,nimene,orange,torch7,p ybrain
One class classification	Shogun. weka.kernlab. dlib. pyML scikit-learn
Multi class classification	Shogun,weka,kernlab,nieme,javaml,pyML,mlpy,Pybrain,tor ch3,scik it-learn
Pre-processing	Shogun,weka,kernlab,nieme,javaml,pyML,mlpy,Pybrain,tor ch3,scik it-learn
Regression	Pybrain,torch3,scikitlearn,shogun,weka,kernlab,dlib,nie me,orange,pyML,java-ML
Structured output learning	Shogun, nieme
Visualization	Weka,nieme,orange,pyML,mlpy,p ybrain,torch3,scikit-learn
Test framework	Shogun, weka, dlib, nieme, javaML, scikit-learn
Large scale learning	Shogun, dlib, nieme, mlpy
Semi- supervised learning	Scikit-learn
Multitask learning	Shogun
Serialization	Shogun,weka,kernlab,dlib,nieme,or ange,javaml,pyML,mlpy,pybrain,s cikit-learn
Image processing	Dlib

6.CONCLUSIONS

The survey provided in this paper summarizes the comparison of these tools on the basis of operating system and file formats supported, general features and language bindings. This is useful for various users to select the tool best suitable for their application. All the tools do not support all the data mining operations. Data mining techniques can be widely classified into classification, regression and clustering. There are various applications of each of these. Also there are many tools available which provide methods to do different operations like WEKA, Shogun, Orange, Scikit-learn etc. WEKA and Shogun supports all the three operations wiz. classification, regression and clustering while Scikit-learn supports regression and clustering operations. Orange tool supports classification and clustering. A number of applications developed by different users have been summarized which clearly shows the importance of data mining in real life. Defining the problem statement and executing it this is the overall process. For solving the problem or executing the research, platform is important so to choose it we have different comparisons stated above.

REFERENCES

[1] Abbas OA (2008) Comparisons between Data Clustering Algorithms. International Journal of Information Technology 5: 320-325

[2] Adhikari A, Rao PR (2008) Efficient clustering of databases induced by local patterns. Decision Support Systems 44:925–943 [34] Lin PL, Po-Huang PW ,Kuo PH , Lai YH (2014) A size-insensitive integrity-based fuzzy c-means method for data clustering. Pattern Recognition 47:2042–2056

[3] Angelis LD, Dias JG (2014) Mining categorical sequences from data using a hybrid clustering method. European Journal of Operational Research 234:720–730

[4] André L.V. Coelho, , Everlândio Fernandes, Katti Faceli (2011) Multi-objective design of hierarchical consensus functions for clustering ensembles via genetic programming Decision Support Systems 51:794-809

[5] Aviad B, Roy G (2012) A decision support method, based on bounded rationality concepts, to reveal feature saliency in clustering problems. Decision Support Systems 54: 292–303

[6] Baker RID, Yacef K (2009) The State of Educational Data Mining:A Review and Future Visions. JEDM - Journal of Educational Data Mining, 1: 3-16

[7] Basili, R., Di Nanni, M. and Pazienza, M. T. (1999) Engineering of IE systems: an object oriented approach. In: Pazienza, editor, Information Exctraction, LNAI 1714, pp. 134–164

[8] Breese JS, Heckerman D, Kadie C (1998) Empirica Analysis of Predictive Algorithms for Collaborative Filtering Microsoft Research, Morgan Kaufmann Publishers, pp. 1-18.

[9] Combes C, Azema J (2013) Clustering using principal component analysis applied to Autonomy – disability of elderly people. Decision Support Systems 55:578–586

[10] Demšar J, Zupan B (2013) Orange: Data Mining Fruitful and Fun - A Historical Perspective. Informatica 37:55– 60

[11] Ferrucci D, Lally A (2004) UIMA: an architectural approach to unstructured information processing in the corporate research Environment. Natural Language Engineering 10:327 – 348

[12] Gupta GK (2012) Introduction to data mining with case studies PHI, New Delhi

[13] Huang Z (1998) Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Acsys CRC, CSIRO

[14] Han J, Kamber M (2001) Data Mining. Kaufmann Publishers, Morgan

[15] Irpino A, Verde R, Francisco de A.T, Carvalho (2014) Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. Expert Systems with Applications 41:3351–3366

[16] Jacques J, Preda C (2014) Model-based clustering for multivariate functional data. Computational Statistics and Data Analysis 71:92–106

[17] Jain AK (2010) Data Clustering: 50 Years Beyond KMeans. Pattern Recognition Letters, 31(8): 651-666

[18] Jain AK, Murty MN, Flynn PJ (1999) Data Clustering: A Review. ACM Computing Surveys, 31:264-323

[19] Jain A, Dubes R (1988) Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall

[20] Kleinberg J (2002) An impossibility theorem for clustering. Conf. Advances in Neural Information Processing Systems, 15: 463–470

[21] Kotsiantis SB, Pintelas PB (2004) Recent Advances in Clustering: A Brief Survey. WSEAS Transactions on Information Science and Applications, 1(1): 73–81

[22] Kriegel HK, Borgwardt KM, Kröger P, Pryakhin A, Schubert M, Zimek A (2007) Future trends in data mining. Data Mining and Knowledge Discovery. 15:87–97

[23] Kumar R, Kapil AK, Bhatia (2012) A Modified tree classification in data mining. Global Journals Inc. 12, 12: 58-63

[24] Liu Y, Qianqian Li, Tang X, Ning Ma, Tian R (2014) Superedge prediction:What opinions will be mined based on an opinion supernetwork model. Decision Support Systems 64:118–129

[25] Low Y, Gonzalez J, Kyrola A, Bickson A, Guestrin C, Berkeley UC (2010) GraphLab: A NewFramework For Parallel Machine Learning
[26] Ngai EWT, Yong Hu, Wong YH,Chen Y, Sun X (2011) The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems 50:559-569
[27] Oyelade OJ, Oladipupo OO, Obagbuwa, IC (2010) Application of k-Means Clustering algorithm for prediction of Students' Academic Performance. IJCSIS 7: 292-295
[28] Padmaja S and Fatima SS (2013) Opinion Mining and Sentiment Analysis –An Assessment of People's Belief: A Survey. International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) 4(1)
[29] PhridviRaj MSB., GuruRao CV (2013) Data mining – past, present and future – a typical survey on data streams. INTER-ENG Procedia Technology 12:255 – 263
[30] Rao IKR (2003) Data Mining and Clustering Techniques DRTC Workshop on Semantic Web, pp. 23-30 [8] Mitra S, Pal KS, Mitra P (2002) Data Mining in Soft Computing Framework: A Survey. IEEE, 13: 3-14
[31] Srivastava S (2014) Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining. International Journal of Computer Applications (0975 – 8887) 88:.10
[32] Soni N, Ganatra A (2012) Categorization of Several Clustering Algorithms from Different Perspective: A Review. IJARCSSE
[33] Rui Xu, Donald CW II (2005) Survey of Clustering Algorithms. IEEE Transactions on neural Networks, 16: 645-678
[34] Radaideh Q, Nagi E (2012) Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance. IJACSA 3:144-151
[35] Rao GN, Nagaraj S (2014) A Study on the Prediction of Student's Performance by applying straight-line regression analysis using the method of least squares. IJCSE 3: 43-45
[36] Rao GN, Ramachandra M (2014) A Study on the Academic Performance of the Students by Applying K-Means Algorithm. IJETCAS 14-180
[37] Romero C ,Ventura S (2007) Educational data mining: A survey. Expert Systems with Applications 33: 135–146
[38] Sandeep, Priyanka, Bansal R (2014) Performance Comparison of Various Partition based Clustering Algorithms. IJEMR pp. 216-223
[39] Sansgiry SS, Bhosle M, Sail K (2006) Factors That Affect Academic Performance Among Pharmacy Students. American Journal of Pharmaceutical Education 70 (5) Article 104
[40] Vijiyarani S, Sudha S (2013) Disease prediction in data mining- A survey. IJCAIT (2).
[41] Velmurugan T (2014) Performance based analysis between k-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data. Applied Soft Computing 19 pp.134–146
[42] Xiao FU, Fan C (2014) Data mining in building automation system for improving building operational performance. Energy and Buildings 75: 109–118
[43] Zhao Q, Fränti P (2014) WB-index: A sum-of-squares based index for cluster validity. Data & Knowledge Engineering 92:77–89