Design and Development of a Robust Algorithm for Information Extraction using K-Means and AGNES

Nancy
Research Scholar, Chandigarh University
Gharuan, Mohali
jindalnancy88@gmail.com

Arvind Kaur
Professor, Chandigarh University
Gharuan, Mohali
arvindcse.cgc@gmail.com

Abstract

Today, data mining has become a burning issue of research in computer and information science with the perspective to find knowledge from large datasets. a modern document contains not only text but also audios, videos, images as well. several tools, techniques and algorithms are available for the extraction of knowledge from the dataset. in this paper, a comparative analysis of various clustering techniques along with their features, pros and cons has been done which helps us to give an insight about these techniques in detail. A hybrid algorithm has been proposed to merge the advantages of these two approaches k-means and AGNES so that various disadvantages of both can be removed. In this paper clustering algorithms are implemented on an open source versatile tool, MATLAB (MATrix LABoratory) and comparison has been done on the basis of certain parameters like accuracy, precision, recall, fscore, true positive, true negative, false positive, false negative for prediction on the Iris dataset.

1. Introduction

Data is limitless and is present everywhere in the universe. It can exist in several forms such as text on a piece of paper or bits or bytes stored in an electronic memory or as the facts stored in human mind. In simple words, data is presented everywhere just like blood in veins. Information processes data to be useful and answers "who", "what", "where" and "when" questions. However, Knowledge is the application of both data and information and answers "how" questions. The process of converting data into knowledge is known as "mining of data" or Data Mining. It is the process of extracting the valid and understandable patterns in data from large quantities of data. Fig 1 presents transformation of data into decision making.

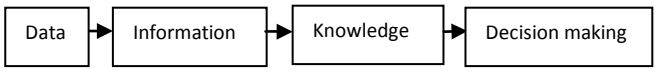

Figure 1. Flow of Data to decision making

Data Mining can also be defined as the process that start from apparently unstructured data tries to extract knowledge and/or unknown interesting patterns. In today's competitive world of information, data is considered as the most precious resource pertaining to potential advantages for the business and corporate world. There are several tools used in Data Mining analysis which have their own significance. It has various techniques or disciplines for advanced database applications such as association generation, clustering and classification. In section 2 the related work has been presented which concludes the research by various authors. Section 3 and section 4 discusses various techniques in detail with its features, advantages and disadvantages along with the comparison of clustering techniques. The research findings along with emperical results have been reported in section 5 whereas conclusion has been provided in section 6. The various references have been discussed in the section 7.

2. Related Work

This section discusses the related work of various authors. Shah et al., [1] emphasised on clustering technique which is a partitioning of data into similar groups; each similar group is called a cluster. Clustering algorithms can be implemented via number of different approaches. A comparison between different clustering algorithms - hierarchical method, density based method and Partitioning method using parameter time taken, cluster instance, sum of squared errors, iterations, etc. for prediction of forest fire is carried out on WEKA. K-Means proved to be a superior algorithm as compared to others because it takes lowest time other than two and also the distribution of cluster instance is fair enough.

Verma et al. [2] gives a detail description about the data, evolution of data and the various flow charts related to it. They also proposed a technique that provides an efficient way as compared to existing methods. The proposed technique is proved to be beneficial in the various cases where proficient information retrieval is required.

Zhang et al., [3] proposed a novel clustering algorithm based on symmetric neighborhood of micro-clusters in large database by compressing the data at first of k-means algorithm to produce micro-clusters and then calculation both neighbors and reverse neighbors of micro-clusters has been carried out to estimate their densities distribution and to gain the clustering results. The algorithm is able to discover arbitrary shape, different densities and also needs fewer input parameters as compared to the existing k-means clustering algorithm. The efficiencies and effectiveness of the algorithm has been validated through the test of IRIS testing dataset and synthetic data set.

Gao et al., [4] proposed an improved simulated algorithm that is able to mine duplicate tasks, invisible tasks and some of non-free-choice structures. Comparing with existing techniques, this algorithm is better than α algorithm which needs to be extended to mine one kind of complicated constructs at a time and it leads to better efficiency than genetic algorithm from computational time point of view. In addition, it has some limitations that it's not effective enough to tackle non-free-choice structures, especially when applied to logs containing both duplicate tasks and non free-choice constructs. In spite of these limitations or disadvantages, it is feasible for process mining.

Zai-an et al., [5] proposed Weka4WS, a framework that is open source Weka toolkit to support distributed data mining on WSRF enabled Grids and had a try at solving the problem of distributed clustering. It also introduced the concepts of Admixture Probability and achieves the distributed clustering algorithm with Weka Library, designs a distributed data mining architecture oriented-services in grid environment combining grid with web services. It proved the validity of the algorithm and feasibility of the system.

Xiaodan et al., [6] discusses the present Data mining technologies that provide relevant algorithms of classification analysis and cluster analysis. The common characteristics of these algorithms through the research of current frame of Data mining technologies to understand the further software development for the programmers are also summed up.

Purwaret al., [7] described the seven most important issues in data mining like feature selection, outlier detection, cluster analysis of high dimensional data, missing value imputation, imbalanced classes in classification, privacy of data and mining from complex/distributed data along with their existing solutions. The analysis of a total of 50 papers from the perspective of performance is measured to verify the outcomes related to specific data mining issue.

Velu et al., [8] explained the classification techniques-EM Algorithm, hmeans+ clustering and Genetic Algorithm (GA) of diabetic patients obtained from Pima Indian Diabetes (PID) data set. These techniques were employed to form clusters with similar symptoms. The simulation tests have been performed on WEKA tool for three models used to test classification. A hypothesis for two different data sets has been also tested. The simulation results performed that h-means+ algorithms performed little better as compared to Expectation-maximization algorithm. This may be because of the fact that EM is not very accurate for high dimensional data sets due to numerical imprecision.

Lekhal et al., [9] mentioned various case studies pertaining to mushroom, breast cancer, larynx cancer. Other datasets are studied to find the utility of association rule mining using Weka tool. Three association algorithms - Apriori, Predictive Apriori and Tertius Algorithms are employed to discuss different case studies. A comparative study of the three algorithms has also made.

Nassar et al., [10] proposed the relationships between data mining techniques and Web usage mining. The integration between the both has been presented for processes at different stages along with the pattern discovery phases and introduces bank cases that have analytical mining technique. Data Mining techniques can be very helpful to the banks for fraud detection in real time, better performance, acquiring new customers, providing segment based products and analysis of the customers purchase patterns over time whereas web documents are structured and attempts to discover the model underlying the link structures of the Web.

Jiang et al., [11] discussed the various privacy issues related to data mining from a wider perspective and investigate various methods that can help to protect sensitive information. Basically, there are four kinds of users involved in data mining applications i.e., Data collector, Data provider, Data miner and decision maker. For each kind of user, various methods and privacy concerns that can be adopted to protect the sensitive information has been discussed.

Owets et al., [12] presents a short history of the big data, definitions, characteristics and tools. Data mining types, techniques that support searching, extracting and analysing are also discussed. The study investigates the most effective Big Data Mining techniques and their applications in various medical, scientific and social fields.

Al-Odan et al., [13] explained the five of the most popular free and open source software tools such as KNIME, Rapid miner, Weka, RStudio and Orange. These tools are compared on a side-by-side manner of both user's acceptance and technical specifications level. After performing several experiments comparative results on the basis of the performance, functionality, flexibility, configuration and framework of the tool are given.

Abdulmohsen et al., [14] proposed an alternative approach for relevance feature discovery in text documents. A method to find and classify low-level features based on both the appearance in the higher-level patterns and their specificity has also been given. A method to select irrelevant documents for weighting features has been introduced.

3. Techniques

There are several techniques used for information retrieval - Association, clustering and classification are the major categories. These major techniques sub categorises a number of algorithms.

3.1. Association

Association occurs in the process of finding relationships between different attributes in large customer databases. The idea in the association rule is to find the nature of the causalities between the values of the different attributes. It can also be generalized to do classification of high dimensional data.

3.2. Classification

Classification is very closely related to the clustering, and is referred to as supervised learning, as opposed to the clustering problem which is referred to as unsupervised learning. This model is used in order to predict the class label of a test example in which only the feature attributes are known.

3.3. Clustering

Clustering is a group similar records together in a large database of multidimensional records. This creates segments of the data which have considerable similarity within a group of points. Depending upon the application, each of these segments may be treated differently. For example, in image and video databases, clustering can be used to detect interesting spatial patterns and features and support content based retrievals of images and videos using low-level features such as texture, color, histogram, shape descriptions, etc.

4. Comparisons of Clustering Techniques

In this section, four major categories of clustering have been discussed.

in Data Mining. The EM iteration alternates between performing an expectation (E) step, which computes the expectation of the log likelihood evaluated using the current estimate for the parameters, and

Table 1. Comparison of Algorithms

Category	Name	Features	Advantages	Disadvantages
Rule based Classification	PART	DIV_{CON}	HE_{DT} E_G EH_{MN} E_I P_{DT} C_{NI}	NA
	RIPPER	C_{TM}	HE_{DT} E_G EH_{MN} E_I C_{NI} P_{DT}	NA
Function based learning	Regression Analysis	ER_{VAR}	USD_{PF} UR_{DINP}	NC_{SE} NL_{VP}
	ANN	SYS_{NEU}	$U_{NSM}P_{LR}$	$C_{PT}C$ PO_F EMP_{DEV}
	SVM	$A_{IS}M_{SV}$	M_{RA} SA_{SV} AX_{PRE}	E_{CS} C_{CST} $C_{PT}C$
Hybrid learning Methods	Logistic Model Trees	COM_{LRDL}	NA	NA
	Bayesian Trees	SMC_{PRO} P_{CRYP}	CL_FA_C	NA
learning	AdaBoost	$P_{EL}ML_{SP}$	S_{THF} V_{AP} SWS_{APP}	TD_{TEZ}
	Random Forest	NA	NA	$OD_{NC/R}$
Hierarchical clustering	AGNES	GN_DD_S	NUM_{CLTR} $S_{PL}E_{IMP}$	N_{UNDO} $N_{OBJ}M_N$ S_{NO}
Centroid (partition) based clustering	*k*-means	$IT_{PDS}C$	U_{FRE} R_{EFF} B_RD_{DW}	$F_{NL}DS$ A_{MN} H_{ND} N_{GC}
	Partition Around Medoids	$DP_{CEN}D_{DP}$	NA	NA
	Fuzzy *c*-means	$A_{MD}P_{DS}$	RES_{ODS} DP_{OCC}	NUM_{IT} $EDUN_{EQ}$
Distribution based clustering	Clustering EM clustering	$F_{MIX}DF$	X_{RD} C_AS	H_C
Density based clustering	DBSCAN	$DIV_{HD}C$	PAC_{DEN}	HUG_{DIS} DIS_{EXP}
Association rules (unsupervised)	Apriori	$F_{TD}AR$	$S_{PL}E_{IMP}$	MUL_{CS} $NM_{DS}L_{IS}$ CG_{MST}
	FP-growth	$M_{FI}T$	RED_{SZ} $ET_{STR}C_{FP}$	CMP_{DS}

BT_{FS}=builds the fastest and shortest tree

4.1.Distribution based Clustering

Distribution based Clustering includes EM algorithm which is also an important algorithm of data mining. Expectation– maximization (EM) algorithm is an iterative method for finding maximum likelihood of parameters in statistical models, where the model depends on unobserved latent variables. Table 1 represents the comparison of various techniques used

maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are used to determine the distribution of the latent variables in the next E step.

4.2.Density Based Clustering

With help of partition and hierarchical methods, spherical-shaped clusters can be found. We cannot find random shaped cluster such as the "S" shape and Oval clusters. With density based method, we can find random shaped clusters, so we can model cluster as dense region in data space, separated by sparse regions. The main idea of density-based clusters is that one can find other than circular shape cluster. DBSCAN (Density-based Spatial Clustering of Application with Noise) was proposed that access density connectivity for handling the random shaped cluster and noise. DENCLUE (Density-based clustering) is a distribution based algorithm, which work effectively on large dataset which contain high level noise, but other way, it works faster compare to DBSCAN, other than this, it contains large number of parameters, so it is good at investing the random shape clusters, but due to non-linear complexity, it can applicable only on small or medium level datasets. The Ordering Points to Identify the Clustering Structure algorithm or OPTICS is procedurally identical to that of the DBSCAN. The OPTICS technique builds upon DBSCAN by introducing values that are stored with each data object, an attempt to overcome the necessity to supply different input parameters.

4.3.Centroid based Clustering

The main idea of centroid based clustering is summarized into below steps:

- Randomly choose k objects from data set as the initial cluster centers.
- Assigns each object to the cluster to which it is associated closely by considering the distance from the given centroids.
- Compute the new position of each centroid by the mean value of object in cluster.
- Repeat step 2 & 3 until the points stop moving, i.e. the mean squared error converges.

K-means, k-medoids and fuzzy c-means are the examples of centroid based clustering. K-means algorithm is very sensitive in terms for selection of initial means. K-means method is applied when the mean of a set of object is defined. Disadvantage of K-means is that, there is no specific answer for find the minimum number of clusters for any given data set. One solution is to compare the results of multiple runs with different clusters and choose best one according to criteria.

4.4.Hierarchical clustering

Hierarchical clustering assign objects into tree like structures, where cluster can have a data point or representation of low level cluster. Fig 2 represents the process of agglomerative and divisive methods of hierarchical clustering.

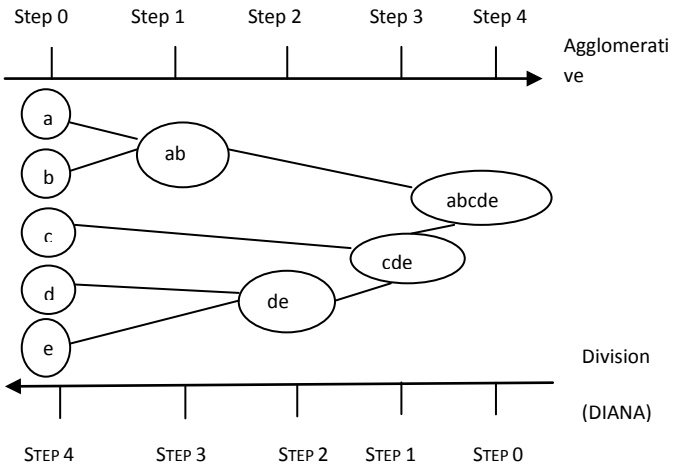

Figure 2. Hierarchical Clustering

Agglomerative: It follows bottom – up approach. It starts for each object letting down in its own clusters and it merges cluster into large clusters, until all objects not merge into one clusters through iteration with specific condition. The single cluster becomes root of hierarchy. For merging, object finds the cluster which is closest to it, and combines the two as one cluster.

Divisive: it works opposite to agglomerative method i.e. top bottom approach. In this method, all objects placed into one cluster i.e. hierarchy root. Then it divides root cluster into several small cluster, and via iteration the root cluster divide into small sub clusters. AGNES is a hierarchical clustering technique and its disadvantage is that as the clusters size grows, the actual patterns of hierarchical clustering become less relevant and also it cannot represent distinct clusters with similar pattern.

5. Research Findings

A hybrid algorithm is proposed using both K-means and AGNES and this section represents the various research findings when applied to Iris dataset.

5.1. Clustering in Iris dataset

Clustering algorithms have been used for Iris dataset i.e. Simple K-Means and Hybrid algorithm to analyze the data. These algorithms are used in Matlab.

5.2. Data Source

To evaluate clustering algorithms, the work has been performed on a data set of Iris database, which is available on UCI Machine learning repository.

5.3. Performance study of Algorithms

Table 2 represents the comparison of the traditional approach with the proposed algorithm on the basis of certain parameters. The accuracy of the proposed algorith is found to be 94.6667 which is very well as compared to the traditional approach. Precision is the fraction of number of positive examples retrieved by the total number of positive and negative numbers. Thus hybrid algorithm retrieved more relevant/positive example i.e 90.9091% than traditional k-means. It is observed that the racall equally performs much better than the traditional approach. Recall is the fraction of number of positive examples retreived by the total number of positive examples in the dataset.

Table 2. Comparison of K-means and proposed algorithm

PARAMETERS	K-MEANS	HYBRID ALGORITHM
ACCURACY	84.2466	94.6667
PRECISION	87.7193	90.9091
RECALL	75.7576	94.3396
FSCORE	81.3008	92.5926
TRUE POSITIVE	50	50
TRUE NEGATIVE	5	7
FALSE POSITIVE	92	73
FALSE NEGATIVE	3	16

F-score or F-measure is a measure of a test's accuracy. It also proved to be better in case of proposed algorithm that is 92.5926 which is observed to be 81.3008 in case of the traditional approach. Similarly the values of confusion matrix i.e, True positive is same in both the cases whereas False positive proved better results 92 which are 73 in k-means.

5.3.1. Results of K-means algorithm

This section represents the values of accuracy, precision, recall and fscore respectively in figure 3 and the values of true positive, true negative, false positive and false negative respectively in figure 4.

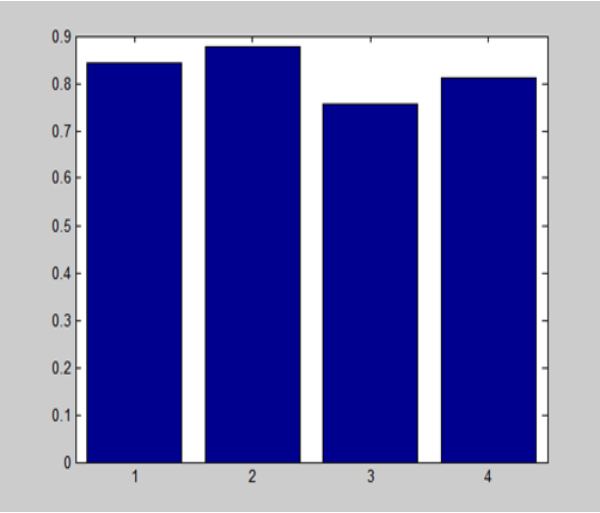

Figure 3. Value of Accuracy, Precision, recall and Fscore

In figure 3 the value of various parameters such as accuracy, precision, recall and fscore has been calculated when applied to the traditional k-means approach. Accuracy depicts the ability of a classifier that how well it can guess the value of the predicted attribute for a given data set. Here, the estimated value for accuracy is 84.2466 and the other parametric values such as precision are calculated to be 87.7193 whereas the recall is 75.7576 and fmeasure or fscore is found to be 81.3008. In figure 4

the confusion matrics parameters such as true positive, true negative, false positive, false negative respectively are estimated in the form of bar graphs.

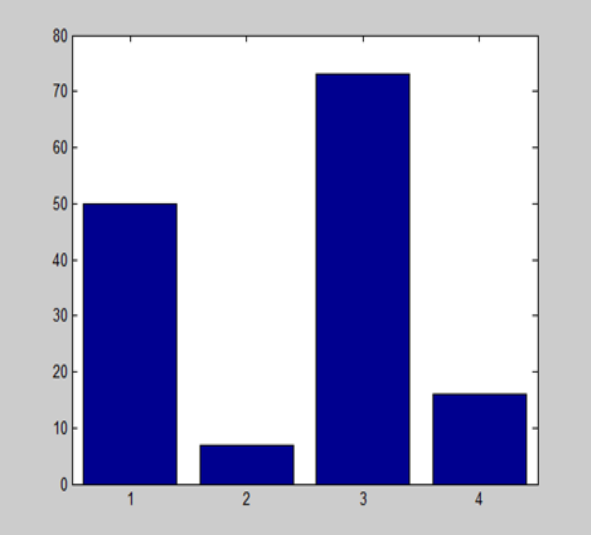

Figure 4. Value of True Positive, true negative, false positive and false negative

5.3.2 Results of Hybrid algorithm

In this section the results of hybrid algorithm have been presented on the basis of several parameters such as accuracy, precision, recall and fscore respectively in figure 5 and the results of confusion Matrix in figure 6 such as true positive, true negative, false positive and false negative respectively have been evaluated.

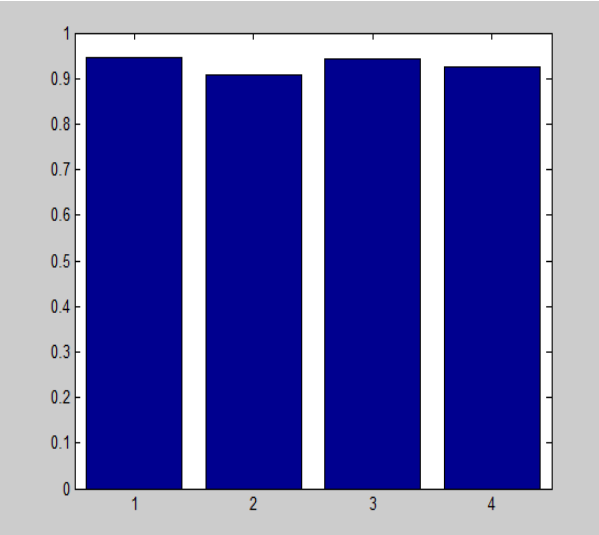

Figure 5. Value of Accuracy, Precision, recall and Fscore

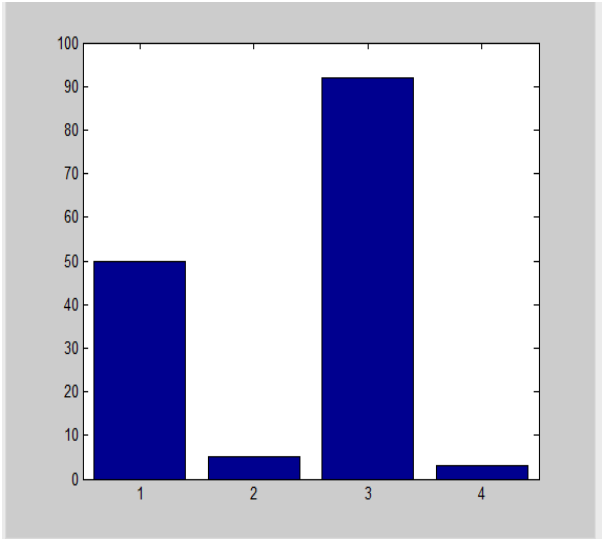

Figure 6. Value of True Positive, true negative, false positive and false negative

6. Conclusion

In this work, the comparison of the various clustering techniques have been presented in detail along with an exhaustive comparative analysis of k-means clustering technique and proposed algorithm with respect to the parameters Accuracy, Precision, recall and fscore as well as the confusion matrix and it showed that the efficiency and output of the proposed algorithm is much better than the traditional approach with the trained data set. The output clearly depicts the performance level has raised up to 94.667% in case of the proposed algorithm. Thus, verifying the performance of proposed algorithm.

7. References

[1] Chintan Shah and Anjali Jivani, “Comparison of Data Mining Clustering algorithms”, IEEE, 2013.

[2] Amit Verma, Iqbaldeep Kaur, Inderjeet Singh, “Comparative analysis of data mining tools and techniques for information retreival”, Indian Journal of Science and Technology,Vol 9(11), DOI:10.17485, March 2016.

[3] Y. Zhang and D. Pi, "A Clustering Algorithm Based on Symmetric Neighborhood of Micro-*clusters*," Computer and Automation Engineering, 2009. ICCAE '09. International Conference on, Bangkok, 2009, pp. 118-122.

[4] Dianfang Gao,Qiang Liu, “An Improved Simulated Annealing Algorithm for Process Mining”, IEEE, pp. 474-479, 2009.

[5] Ren Zai-an, Wang Bin, Zheng Shi-ming , Miao Zhuang, Shao Rong-ming, "A WSRF-enabled Distributed Data Mining Approach to Clustering WEKA4WS -Based", IEEE,pp. 219-226, 2010.

[6] Zhang Xiaodan , "Plain Discussion on Data mining Technology Research", IEEE, pp. 296-298, 2011

[7] Archana Purwar, Sandeep Kumar Singh ,"Issues in Data mining: A comprehensive Survey", IEEE, 2011.

[8] C. M. Velu, K. R. Kashwan, "Visual Data Mining Techniques for Classification of Diabetic Patients", IEEE, pp. 1070-1075, 2012.

[9] A. Lekhal, Dr. C V Srikrishna and Dr. Viji Vinod, "Utility of Association Rule Mining: a Case Study using Weka Tool", IEEE, 2013.

[10] Omer Adel Nassar, Dr. Nedhal A. Al Saiyd, "The Integrating Between Web Usage Mining and Data Mining Techniques", IEEE, pp. 243-247, 2013.

[11] Lei Xu Chunxiao Jiang, Jian Wang, Jian Yuan and Yong Ren , "Information security in Big Data: privacy and Data mining", IEEE, vol.2, pp. 1149-1176, 2014.

[12] Nour E.Owets, Suhail S.Owais, Waseem George, Mona G.Suliman and Vaclav Snasel, "A Survey on Big Data, Mining: (Tools ,Techniques, Applications and Notable uses)", Springer, pp.109-119, 2015.

[13] Hussah A. Al-Odan, Ahmad A. Al-Daraiseh King Saud, "Open Source Data Mining Tools" A Comparitive Study, IEEE, pp. 369-374, 2015.

[14] Abdulmohsen A, Mubarak Al, "Relevance Feature Discovery for Text Mining", IEEE, 2015.

APENDIX

S_Y=Symbols, G_S=Gestures, D_R=during, B_F=Before, F_D=floppy disk, CL_D=cloud, PP_R=paper, CD_{ROM}=CD-ROM, D_{VD} =DVD, U_{SB}=USB,CLF_{IN}= Classifier Instances, ALG_{IMP}=Algorithm Implemented, NUM_{CL}= No. of clusters, CL_{INS}=Cluster instances, NUM_{IT}=No. of iterations, CL_{SUM}=Within clusters sum of squared errors, T_{BM}=Time taken to build models, UCL_{IN}=unclustered instances, CEN_{CL}=Centroid based Clustering, H_{CL}=Hierarchical clustering, DIS_{CL}=Distribution based clustering, DNS_{CL}=Density based clustering, EH_{MN}=Can easily handle missing values and numeric attributes, E_G= Easy to generate, HE_{DT}= highly expressive as decision trees , DIV_{CON}=Combines the divide-and-conquer strategy with separate and conquer strategy of rule learning, E_I= Easy to interpret, P_{DT}=Performance comparable to decision trees, E_{CS}=effect of collusion on security, C_{CST}=communication cost, $C_{PT}C$=computation cost, C_{TM}= 2-class and multi-class problem, ER_{VAR}=estimates relationships among variables, USD_{PF}=Used for prediction and forecasting, UR_{DINP}=used to infer casual relationship between dependent and independent variables, NC_{SE}= not confined to single equation, NL_{VP}=not linearity in variables and parameters, ANN= Artificial Neural Network, SYS_{NEU}= System of interconnected neurons that exchange information among each other, $U_{NSM}P_{LR}$= used to perform nonlinear statistical modeling and provide a new alternative to logistic regression, PO_F= proneness to over fitting, EMP_{DEV}=the empirical nature of model development, $A_{IS}M_{SV}$=Approximating the original decision function by using an infinite series of linear combinations of monomial feature mapped support vectors, M_{RA}= most robust and accurate method, SA_{SV}=Security against attacks on support vectors, COM_{LRDL}=Combines logistic regression and decision tree learning, SMC_{PRO}= SMC-based protocol, P_{CRYP}= Paillier cryptosystem, $P_{EL}ML_{SOLP}$=Provides ensemble learning which deals with methods which employ multiple learners to solve a problem, S_{THF}=Solid theoretical foundation, V_{AP}=very accurate prediction, SWS_{APP}=great simplicity, wide and successful applications, TD_{TEZ}= test error often tends to decrease even after the training error is zero, consists of many decision trees and outputs the class that is the mode of the classes output by individual trees, $OD_{NC/R}$=over fit for some datasets with noisy classification/regression tasks, AGNES=Agglomerative hierarchical clustering algorithm, GN_DD_S= grouping the data one by one on the basis of the nearest distance measure of all the pair wise distance between the data point. Again distance between the data point is recalculated, NUM_{CLTR}=No information about the number of clusters required, N_{UNDO}= can never undo what was previously done, $N_{OBJ}M_N$=No objective function is directly minimized, S_{NO}=Sensitivity to noise and outliers, $IT_{PDS}C$=simple iterative method to partition a dataset into a user specified number of clusters, k, $B_{RST}DS_{DW}$=gives best results when data set are distinct and well separated from each other, R_{EFF}=relatively efficient, UND_{FRE}=Fast, robust and easy to understand, A_{MN}=applicable only when mean is defined, $F_{NL}DS$= fails for non-linear dataset, H_{ND}=unable to handle noisy data, N_{GC}= Does not work well with global clusters, $DP_{CEN}D_{DP}$=chooses datapoints as centers and works with an arbitrary matrix of distances between datapoints, $A_{MD}P_{DS=}$assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point, $RES_{ODS=}$Gives best result for overlapped data set, DP_{OCC}= data point may belong to more than one cluster center, NUM_{IT}=more number of iteration, $EDUN_{EO}$=Euclidean distance measures can unequally weight underlying factors, $F_{MIX}DF$= Focuses on mixture models which can be used to cluster continuous data and to estimate the underlying density function, $DIV_{HD}C$=divides the high-density data into clusters and forms any kind of cluster in the noisy database, PAC_{DEN}=suit for the packing density, HUG_{DIS}=Wispy difference will lead the huge distinction, DIS_{EXP}=disorder selection of parameter can only confirm by experiences, $F_{TD}AR$=To find frequent data items from a transaction data set and derive association rules, MUL_{CS}=does multiple scan over the database to generate candidate set, $NUM_{DS}L_{IS}$=number of database passes are equal to the max length of frequent item set, CG_{MST}=candidate generation process it takes more memory, space and time, $M_{FI}T$=algorithm mines frequent item sets without the time – consuming, $ET_{STR}C_{FP}$=used an extended prefix tree structure for storing compressed and crucial information about frequent pattern, RED_{SZ}= Reduce the overall size of all input data set, CMP_{DS}= Uses complex data structures, NA=Not Acknowleged