International Journal of Engineering Applied Sciences and Technology, 2017
Vol. 2, Issue 3, ISSN No. 2455-2143, Pages 73-78
Published Online January-February 2017 in IJEAST (http://www.ijeast.com)

TWO STEP CLUSTERING APPROACH USING BACK PROPAGATION FOR TUBERCULOSIS DATA

Ravinder Kaur
Dept. of Computer Science and Engineering
CTITR, Jalandhar, India

Anshu Sharma
Assist. Professor
Dept. of Computer Science and Engineering
CTITR, Jalandhar, India

Abstract- **The clustering is the technique which can cluster similar and dissimilar type of data. In the recent times various clustering techniques has been applied which cluster similar and dissimilar type of data. The two step clustering is the efficient clustering algorithm which is based on clustering and classification. To improve performance of two step clustering technique back propagation learning is applied which is based on neural networks. The back propagation learning technique learns from the previous experience and drive new values. The simulation of proposed and existing model is done in matlab and it is been analyzed that accuracy and gini index is improved in the proposed technique.**

Keywords— **Two-step clustering, classification, gini index, accuracy, back propagation neural networks**

I . INTRODUCTION

With the enormous measure of data stored in files, databases, and different repositories, it is progressively important, if a bit much, to develop capable means for examination and may be interpretation of such data and for the extraction of intriguing knowledge that could help in decision-making [1]. Data Mining, additionally famously known as Knowledge Discovery in Databases (KDD), alludes to the nontrivial extraction of implicit, beforehand obscure and potentially helpful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently regarded as synonyms, data mining is very of the knowledge discovery process [2].Clustering is a division of data into groups of comparative objects. Speaking to the data by less clusters essentially loses certain fine points of interest, however accomplishes simplification. It models data by its clusters. Data modeling places clustering in a historical perspective rooted in mathematics, statistics, and numerical examination [3]. From a machine learning perspective clusters compare to hidden patterns, the search for clusters is unsupervised learning, and the subsequent system speaks to a data concept. From a machine learning perspective clusters relate to hidden patterns, the search for clusters is unsupervised learning, and the subsequent system speaks to a data concept. Along these lines, clustering is unsupervised learning of a hidden data concept. Data mining manages large databases that impose on clustering examination extra extreme computational requirements. These challenges prompted the emergence of intense comprehensively applicable data mining clustering methods [4]. There are two sorts of clustering techniques: Partition and Hierarchical. In partition Clustering given a database of n objects, a partition clustering algorithm constructs k partitions of the data, where every cluster optimizes a clustering criterion, for example, the minimization of the sum of squared distance from the mean inside every cluster [5]. Hierarchical algorithms make a hierarchical decomposition of the objects. They are either agglomerative (bottom-up) or divisive (top-down). Agglomerative algorithms begin with every object being a separate cluster itself, and progressively merge groups as indicated by a distance measure. Divisive algorithms take after the opposite strategy. They begin with one group of all objects and progressively split groups into little ones, until every object falls in one cluster, or as fancied [6]. Aside from the two primary classifications of partition and hierarchical clustering algorithms, numerous different methods have emerged in cluster examination, and are predominantly centered on specific issues or specific data sets available. Density-Based Clustering algorithms group objects as per specific density objective functions [7]. Density is typically characterized as the number of objects in a particular neighborhood of a data objects. In these methodologies a given cluster keeps developing the length of the number of objects in the neighborhood surpasses some parameter. Lattice Based Clustering has primary concentrate on spatial data, i.e., data that model the geometric structure of objects in space, their relationships, properties and operations [8]. The objective of these algorithms is to quantize the data set into a number of cells and after that work with objects having a place with these cells. Model-Based Clustering algorithms discover good approximations of model parameters that best fit the data. They can be either partition or hierarchical; contingent upon the structure or model they hypothesize about the data set and the way they refine this model to identify partitioning's. They are nearer to density-based algorithms, in that they develop particular clusters so that the preconceived model is made strides. Categorical Data Clustering algorithms are specifically developed for data where Euclidean, or other numerical-oriented, distance measures can't be connected [9].

A. K-mean clustering

K-means is one of the straightforward unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through some number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different position causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the closest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the last step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As result specified k centers change their location step by step until no more changes are done or in other words centers do not move any more. At last, this algorithm aims at minimizing an objective function knows as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

Where, $\|x_i - v_j\|$ is the Euclidean distance between xi and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

B. Two step clustering algorithm

Two step cluster analysis is technique of the arithmetical software set SPSS used for huge data bases, since ordered and *k* -means clustering do not scale scalable when *n* is very large. Two-step clustering is used to cluster data into different clusters and allocate classes based on variables. The SPSS Two-Step cluster technique is considered as a scalable cluster analysis algorithm that is designed to handle very huge data sets. It is capable to handle both regular and categorical variables and attributes [12]. It needs only one data pass. It is performed in two steps 1) pre-cluster the cases or records into several small sub-clusters 2) assemble the sub-clusters that are the output of pre-cluster step into the preferred number of clusters. It can also spontaneously select the number of groups. This clustering technique is very effective in classification of huge data sets and it has the ability to create clusters by using categorical and continuous variables and it is provided with spontaneous selection of number of clusters.

C. Decision tree

Decision tree is a managed type of learning algorithm which has a pre-defined target variable and this algorithm is mostly used in classification problems. It can work for both regular and categorical output and input variables. According to this algorithm the data sample is isolated into two or more than two homogeneous groups based on most significant differentiator in variables of input data set.

D. Back propagation neural network

The back propagation neural network is an ordinary way of teaching artificial neural networks used in combination with an optimization technique such as gradient descent [18]. This method computes the gradient of a loss function with according to all the masses computed in the network. After that the computed gradient is served to the optimization method which uses it to modernize the weights, in order to minimize the loss function. This algorithm is one of the most popular NN algorithms that is called back propagation algorithm [19]. This back propagation algorithm can be break down to four main stages. After selecting the masses of the network, the back propagation algorithm is used to calculate the necessary alterations.

E. Tuberculosis

Tuberculosis is a very common disease which is caused by mycobacterium and established as severe disease with really serious effects. This disease classically distresses the lungs, but it also can distress any other organ of the body. This disease is typically cured with a schedule of drugs taken for six months to two years depending on the stage of disease [20]. TB is spread by means of air from the infected person to the normal one. The TB bacteria are spread into the air when an infected person with TB disease of the lungs, speaks, sings, or sneezes. The normal people nearby can breathe in these bacteria and infected air and become infected**.** TB is not spread by sharing food, drink and sharing toothbrushes and shaking someone's hand [22].

II. RELATED WORK

Md. Ezaz Ahmed, et.al (2013) proposed in paper [10] that unlabeled document collections are turning out to be progressively common and mining such databases turns into a noteworthy test. As the number of available Web pages grows, it is turned out to be more troublesome for clients finding documents applicable to their interests. Clustering is the classification of a data set into subsets (clusters), so that the data in every subset share some common trait – frequently proximity as per some defined distance measure. By clustering we enhance the quality of websites by grouping comparable websites in groups. This paper addresses the applications of data mining tool, weka by applying k means clustering to discover clusters from huge data sets and discover the characteristics that represent advancement of search engines. Mohnish Patel, et.al (2014) proposed in this paper [11] that Efficient Privacy preserving association rule mining has emerged as a most recent research issue. In this theory work, existing algorithms, Increase Support of Left and Decrease Support of Right are implemented effectively on the real data for Privacy Preserving Association Rule Mining. Keeping in mind the end goal to hide an association rule, a hybrid algorithm is proposed which is based on two previous existing algorithms ISL and DSR. K-Mean, Neural gas Cluster algorithm with number of cluster in this algorithm, first we deplete support of right hand side of the rule in a rule where object to be hidden is in right side. Richa Sharma, et.al (2016) proposed in this paper [12] that one of the applications of data mining is disease diagnosis for this purpose one needs medical dataset to identify hidden patterns lastly extracts valuable knowledge from medical database. As of late, researchers have utilized different classification and clustering algorithms for diagnosing diseases. This paper gives overview on two different complex diseases which incorporates the heart disease and Cancer disease, paper fundamentally watched the existing writing work to discover significant knowledge in this area and summarized different approaches utilized as a part of disease diagnosing, promote examined about the tools available for processing and classification of data. This study reveals the importance of research in area of life debilitating disease diagnosis. G. Anuradha, et.al (2014) proposed in this paper [13] that the popular expression in research is Big Data. Big Data gets described by 5 V's: Volume, Velocity, Variety, Veracity and Value of data. Volume altogether of penta bytes, velocity which alludes to click stream data in various domains, variety containing heterogeneous data, veracity demonstrating the cleanliness of data and value emphasizing on the arrival on investment for companies who invest in Big Data technologies. This Big Data is better modeled not as persistent tables but rather as transient data streams which require different clustering and mining techniques to be effectively processed and managed. In this paper a few suggestions on online learning through clustering and mining of stream data are introduced. Since the volume and velocity of big data keeps expanding consistently, more propelled techniques for clustering and mining such humongous data is the need of the hour.

EdemInang Edem, et.al (2015) proposed in this paper [14] that the proliferation of malware as of late have accounted for the increase in computer crimes and prompted for a more aggressive research into improved investigative strategies, to keep up with the menace. Exploring dynamic examination is unarguably, a positive step to supporting static evidence with malware dynamic conduct logs. In perspective of this, dissecting these huge generated reports raises concerns about speed, accuracy and performance. The implementation results of the sub-components recorded in this study demonstrated a considerable time gain in feature extraction utilizing unique feature approach furthermore yielded an improved data matrix embedding strategy which made data mining clustering of behavioral reports data got from an online sandbox relatively fast utilizing k-means with appropriate distance measure. Cheng-Fa Tsai, et.al (2014) proposed in this paper [15] that this investigation develops another data clustering method. The proposed algorithm's expansion without selecting data points to increase computation cost and it might considerably lower time cost. The experimental results affirm that the displayed approach has genuinely high clustering accuracy and noise filtering rate, and is faster than numerous notable existing density-based data clustering algorithms. Experimental results indicate that the proposed data clustering algorithm surpasses other existing celebrated real approaches, for example, the DB SCAN, IDBSCAN, KIDBSCAN, and FDBSCAN techniques, and its high accuracy and low execution-time cost make it efficient and effective for data clustering in numerous data mining applications.

III. PROPOSED METHODOLOGY

The existing work is based on two step clustering in which is based on k-mean clustering and decision tree classification. The first step of clustering is based on k-mean clustering in which arithmetic mean of the whole dataset is calculated and from the central point Euclidian distance is calculated using distance formula. The clusters are defined on the basis of similarity of the distance. In the second step of clustering decision tree classifier is applied which will classify the data. The two step of clustering is implemented and analyzed that accuracy is less because some points are remained un clustering or wrongly clustered. The proposed work is based on improvement in two step clustering. In the proposed work, Euclidian distance is calculated in the iterative manner using back-propagation learning technique. The back-propagation learning technique is based on unsupervised learning in which network learn from the previous experiences and drive new values as per requirements. The whole dataset is divided into two parts tanning and trained dataset. The dataset values is given as input and formula is applied which will calculate actual value of the Euclidian distance.

Actual Euclidian distance =

$$\sum_{\substack{x=0\\w=o}}^{\substack{x=n\\w=n}} \mathrm{X_n W_n} + \text{bias} \text{-------------------- (1)}$$

The actual Euclidian distance is calculated using equation 1, to calculate Euclidian distance in the iterative manner, old of the first iteration will be input in next iteration unless error is minimized and error is calculated using equation 2

Error =

Actual Euclidian distance – Desired Euclidian distance----- (2)

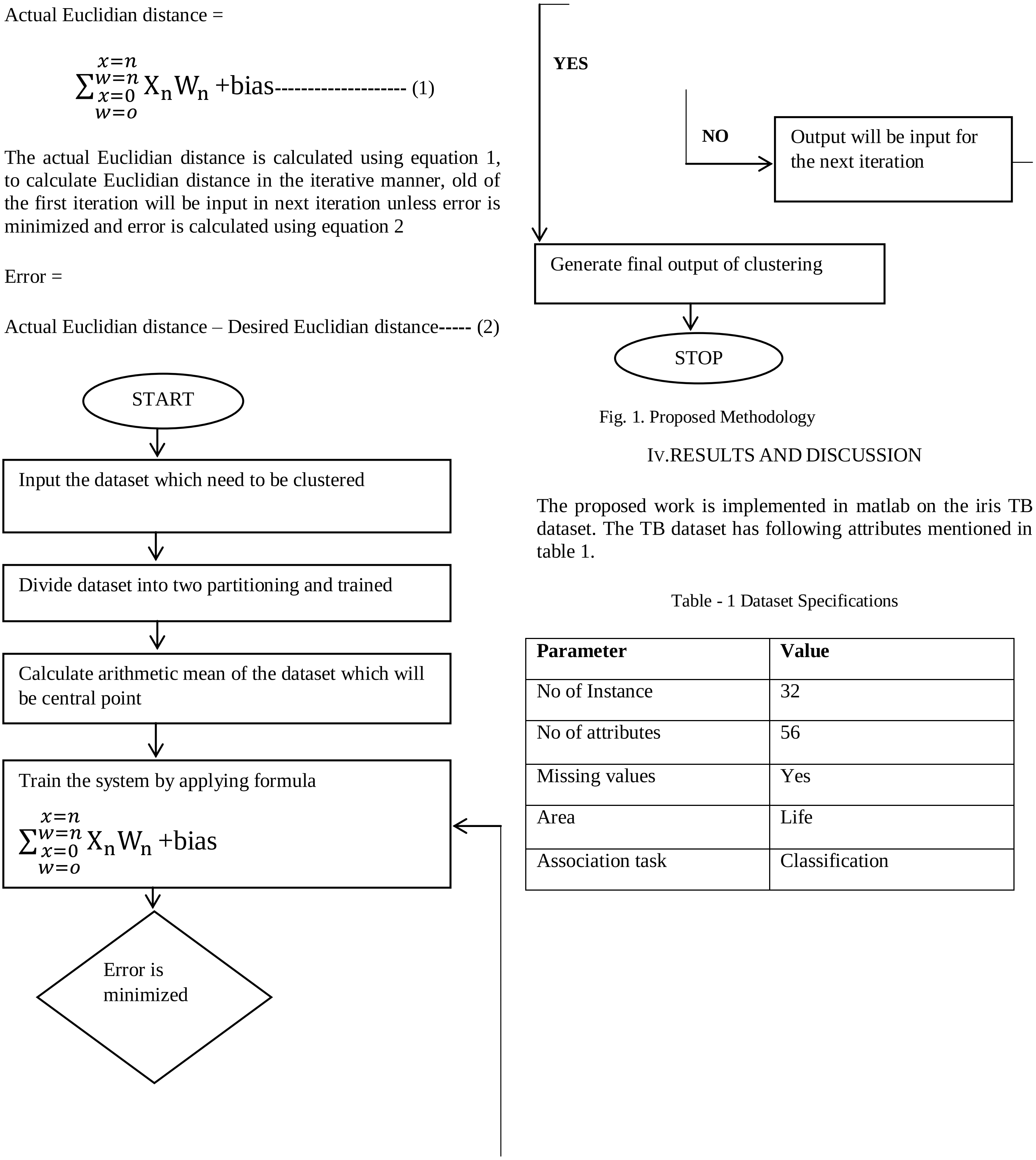
START
Input the dataset which need to be clustered
Divide dataset into two partitioning and trained
Calculate arithmetic mean of the dataset which will be central point
Train the system by applying formula
$\sum_{\substack{x=0\\w=o}}^{\substack{x=n\\w=n}} \mathrm{X_n W_n}$ +bias
Error is minimized
YES
NO
Output will be input for the next iteration
Generate final output of clustering
STOP

Fig. 1. Proposed Methodology

Iv.RESULTS AND DISCUSSION

The proposed work is implemented in matlab on the iris TB dataset. The TB dataset has following attributes mentioned in table 1.

Table - 1 Dataset Specifications

Parameter	**Value**
No of Instance	32
No of attributes	56
Missing values	Yes
Area	Life
Association task	Classification

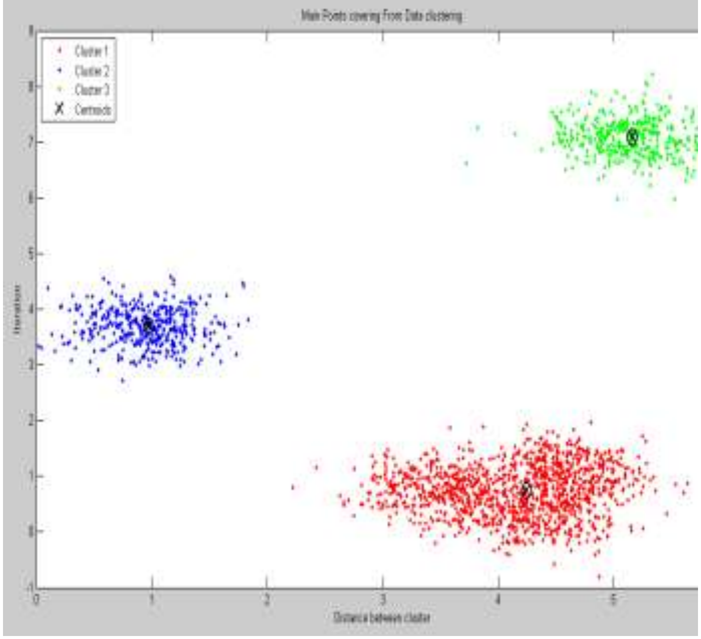

Fig. 2. Final output

As shown in figure 2, the final output of clustering is generated in which three clusters are formed. The values are assigned to each cluster and no points are remained unclusttered or wrongly clustered.

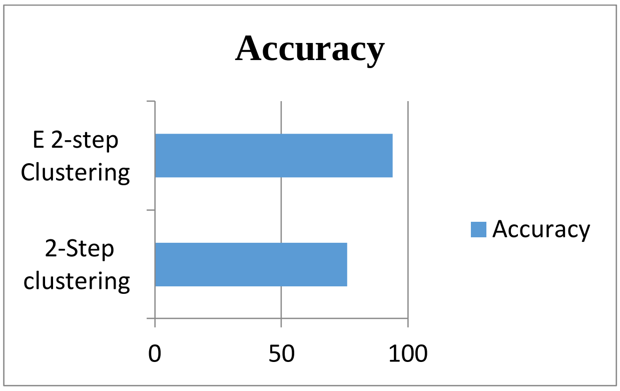

Fig . 3. Accuracy comparison

As shown in figure 3, the comparison is made between existing and proposed technique is terms of accuracy. It is been analyzed that accuracy of proposed two step clustering is increased up to 15-20 percent.

$$\text{Accuracy} = \frac{\text{No of Points clustered}}{\text{Total number of points}} * 100$$

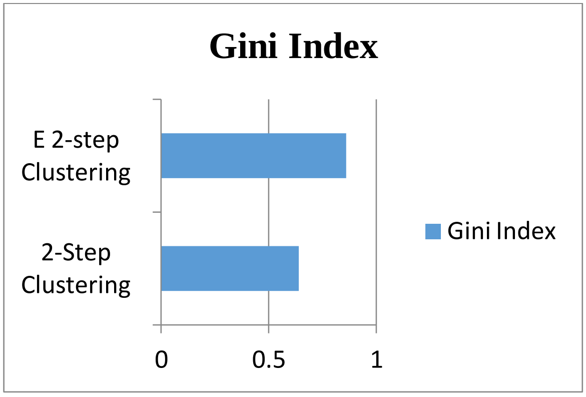

Fig. 4. Gini index comparison

As shown in figure 4, the proposed and existing 2-step clustering is compared in terms of gini index. It is been analyzed that gini index is increased up to 10 percent in the proposed work.

$$\text{Gini index} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n\sum_{i=1}^{n} x_i}$$

V. CONCLUSION AND FUTURE SCOPE

In this work, it is been concluded that clustering is the efficient technique to analyze similar and dis-similar type of data. In this work, two step clustering is improved to increase accuracy of clustering. To improve accuracy if clustering technique of back propagation learning technique is applied which is based on neural networks. The proposed and existing models are implemented in matlab and it is been analyzed that proposed model performed batter in terms of accuracy and gini index. In future the proposed methodology can be used in various real life applications of data mining, web security, medical diagnosis etc. It can be applied for the density based clustering to improve performance. This methodology works as a sequential learning machine taking the input patterns sequentially for recognition but as a future prospect, work should be carried to generate high level networks for recognizing concurrent patterns.

VI. REFERENCES

[1] Geqi Qi, Yiman Du, Jianping Wu, Ming Xu," Leveraging longitudinal driving behavior data with data mining techniques for driving style analysis", 2015, IET Intell. Transp. Syst., Vol. 9, Iss. 8, pp. 792–801

[2] M. Guder, O. Salor, I. Çadirci, B. Ozkan and E. Altintas," Data Mining Framework for Power Quality Event Characterization of Iron and Steel Plants", 2015, IEEE, 0093-9994

[3] Claudia Plant, Andrew Zherdin, Christian Sorg, Anke Meyer-Baese, and Afra M. Wohlschläger," Mining Interaction Patterns among Brain Regions by Clustering", 2014, IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 9

[4] Johannes Grabmeier, Andreas Rudolph," Techniques of Cluster Algorithms in Data Mining", 2002 Kluwer Academic Publishers, 303–360

[5] P. Berkhin," A Survey of Clustering Data Mining Techniques", 2010, Springer, 3485-34-533

[6] G. Sreenivasulu, S. ViswanadhaRaju and N. SambasivaRao," Review of Clustering Techniques", 2016, Springer Science+Business Media Singapore

[7] Lamine M. Aouad, Nhien-An Le-Khac, and Tahar M. Kechadi," Lightweight Clustering Technique for Distributed Data Mining Applications", 2007, ICDM, LNAI 4597, pp. 120–134

[8] Murilo Coelho Naldi," Genetic Clustering for Data Mining", 2001, Springer, 35-343-578, pp-343-967

[9] Francesco Gullo," From Patterns in Data to Knowledge Discovery: What Data Mining Can Do", 2015, Francesco Gullo / Physics Procedia 62, 18 – 22

[10] MD. Ezaz Ahmed, PreetiBansal," Clustering Technique on Search Engine Dataset using Data Mining Tool", 2013, IEEE, 978-0-7695-4941

[11] Mohnish Patel,PrashantRichhariya, AnuragShrivastava," A Novel Approach for Data Mining Clustering Technique using NeuralGas Algorithm", 2014, IEEE, 978-1-4799-4910

[12] Richa Sharma, Dr. Shailendra NarayanSingh, Dr. SujataKhatri," Medical Data Mining Using Different Classification and Clustering Techniques: A Critical Survey", 2016, IEEE, 978-1-5090-0210

[13] G. Anuradha, Bidisha Roy," Suggested Techniques for Clustering and Mining of Data Streams", 2014, IEEE, 978-1-4799-2494

[14] EdemInangEdem, ChafikaBenzaidy, Ameer Al-Nemrat and Paul Watters," Analysis of Malware Behaviour: Using data Mining Clustering Techniques to Support Forensics Investigation", 2015, IEEE, 978-1-4799-8825

[15] Cheng-Fa Tsai, Po-Vi She," A New Efficient Density-Based Data Clustering Technique Using Cross Expansion for Data Mining", 2014, IEEE, 978-1-4799-4215