DOI: http://dx.doi.org/10.26483/ijarcs.v8i9.5186

ISSN No. 0976-5697

Volume 8, No. 9, November-December 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

AN ENHANCED K-MEAN CLUSTERING ALGORITHM

Sapna
Department of Computer Science & Engineering/
Lovely Professional University
Jalandhar, India

***Abstract*—**K-Mean's clustering algorithm is one of the most widely used partitioning algorithm used for grouping the elements. It is the fast, simple and can work with large datasets. But still it has some drawbacks like in the initial stage we have to tell the number of clusters. It can detect only spherical clusters. Number of iterations is more. Here we will propose an enhanced K-Means clustering algorithm which will basically work on the concept of partitioning dataset and reducing the number of iterations. It will abstract some features from two modified K-means algorithms. The benefit of partitioning is that we will be able to deal with larger datasets and the benefit of reducing iterations is that time taken for clusters formation will reduce and in this way the efficiency of the traditional K-means clustering algorithm is increased. The results of the proposed methodology, is applied on Enron dataset to find out spam emails in the spam email dataset.

***Keywords*—** Data Mining, KDD, E-Mail, Spam, Ham, Spam Filter, K-Means Algorithm, K-Means Update Algorithm.

I. INTRODUCTION

The incremental growth of data from last few decades is remarkably high. The reason behind the tremendous increase in the size and the complexity of the data is due to various online commercial sites, Work performed in the engineering field and other social media sites like facebook, you tube etc. The internet contains large amount of raw data, to process the data various tools and techniques are used for the efficient extraction of relevant data. Data Mining is a process established for the possible extraction of unseen information for the sake of gaining knowledge. Facts can vary in dimension, difficulty to formation. Data can be represented in the form of audio, video or simply a text data in alphabetic or numeric form. Data mining is generally desirable, so as to tackle large volume of data and to extract needed properties from the group of the data.

- *Knowledge Discovery Process*

Knowledge or facts from the data can be acquired by undergoing many steps related with each other. Information mining is also categorized as Knowledge detection method, which means an action to extract valuable data from a collection of untreated data. Data mining is a concentric part of knowledge discovery [1], [2], [3]

- *Collection of Raw Data*: Data-group can be gathered from various sources like online and offline, social media sources, public sector banks, retail sector, Insurance companies, Private sector banks etc.
- *Data Selection:* Data can vary in large volume, so it is necessary to extract relevant and essential data that is required for the further processing is selected.
- *Data Pre-Processing:* Raw data can contain false information in the missing values or noise form. So, it is mandatory to pre-process the dataset, so as to remove any kind of vague or false data.
- *Transformation:* The data is transformed into suitable shape so that mining job can be carried out.
- *Data Mining:* Finding the relevance among the data is called as data mining. A variety of data mining approaches can be utilized to carry out the application in the data.
- *Evaluation:* Gained information is evaluated for the exactness of the patterns and its compactness.
- Knowledge: The final required information is called as Knowledge.
- Diverse Methods of Data Mining

The diverse methods relevant in data mining are considered as mentioned underneath [1]. The following steps are performed on raw data to gain and access Knowledge.

- *Anomaly Detection:* Collected information that can be irrelevant or bogus is detected which is termed as an Anomaly or fake. Anomaly detection track the information that contributes with no fact or knowledge.
- *Association Rule Mining (ARM):* It is a procedure of establishing a relationship between the items in the dataset.
- *Clustering:* It is a procedure that labels the similar type of data in one groups called as clusters without knowing any predefined model. Expressive process of grouping the data.
- C*lassification:* It is a procedure that has a predefined known structure which groups the data into known predefined groups. Classification modelling is a predictive model for grouping the data. It helps to target data to different classes.
- *Summarization:* A process of labelling the data in a compact form so that we can visualise and represent it.
- *Electronic mail Spam*

Electronic mails are classified into two broad categories: Spam emails and Ham emails. Spam emails are the unauthenticated emails received from the unknown sources that may contain virus. Spam can originate from any external source like Web, Text messages, etc., depending upon the kind of broadcast, spam can be categorised into a variety of category similar to electronic mail spam, web spam, text spam, social networking spam [4].

The spam emails are scattering at great pace due to the swift and offensive way of contribution data. It was noticed that account holders receive more spam emails than the ham emails. To avoid spam emails, spam filtration is important because spam can lead to time, energy, and bandwidth wastage, along with the misleading information [5].

Email can be labelled as a spam email only depending on these properties:

- *Unsolicited Emails:* E-mails that are received from contacts that are not known to the user.
- *Bulk Mailing:* The kind of emails that are sent in mass or bulk to multiple account holders at the same time.
- *Nameless Mails:* In this type of mails in the identity and the details of the sender are not revealed or demonstrated.

Electronic mail spam is of severe concern and can cause main bandwidth failure and can charge billion of dollars failure to the servicer's. It is necessary to distinguishing between the type of email, a spam email or a ham email. Numbers of algorithms are present that can efficiently distinguish the emails on their characteristics, but because of the rapid change in the technologies, spammers are becoming wiser. So, new and better algorithms are needed which ensure better accuracy rate for successfully labelling the emails. Spam filtration technique is used to group the email as an unsolicited email and prevents it from entering the authorised user's inbox. Filters are grouped into two types [5]:

- *Machine Learning Based Technique***:** machine learning based approach are those that obtain facts and discover by instances, for instance Support Vector Machine, Multi-Layer Perceptron, Naïve Bayes Algorithm, Decision Tree Based etc.
- *Non-Machine Learning Based Technique:* non-machine learning based approach are those that don't obtain facts, instance sandboxing, blacklist or whitelist, heuristic scanning, signature based technique etc.

The likelihood of accomplishment of machine learning algorithms in dissimilarity with non-machine learning algorithms is additional. Both the above mentioned algorithms execute by choosing the finest characteristics from the dataset so as to label the emails, and categorise it into either spam folder or ham email folder. The process of selecting the features can be performed in two possible ways ie. Header based feature selection or content based selection:

- *Header Based Selection:* the selection of the best possible feature from the header of the email is termed as header based selection. It consists of recipient's handle, Blind Carbon Copy, Carbon Copy, To details, From details, Date of the mail and Subject.
- *Content Based Selection:* the process of selection of the finest characteristic from the data and content written in the email is called as content based feature selection. The content can be in any profile like audio, video or text message.

Content Feature Selection technique is considered as a valid feature selection technique in contrast to Header Based feature selection technique because the main problem with Header Based Feature Selection is that it can be simply modify by the hackers as per the requirement.

The research paper is outlined in diverse sections. Section 2 represents the related work introducing various papers in the field of clustering. Section 3 describes the methodology adopted for the identification of ham and spam emails. Section 4 describes a variety of grades conducted on Enron dataset, while section 5 concludes the paper and prescribes the finest algorithm with elevated accuracy for spam discovery.

II. RELATED WORK

Mansoori Eghbal et al. has explained two types of clustering. One is fuzzy clustering and other is non fuzzy clustering. In fuzzy clustering, a data object may fit in to more than one cluster where as in non fuzzy clustering an object belongs exactly to a particular cluster. In case of fuzzy clustering, a membership is provided to each data object. This membership is generally less than one. Fuzzy clustering is better than non fuzzy clustering when the borders between the clusters are not fixed. Fuzzy and non fuzzy clustering both has some drawbacks like both the techniques are sensitive to the number of clusters. In most of the data mining applications the information obtained is not easily understood by the naïve users. In order to remove these drawbacks, fuzzy rule-based clustering algorithm (FRBC) is proposed in this paper. Initially it searches the clusters in the data without human intervention. Then it uses some fuzzy rules to recognize those clusters. It is basically useful when the boundaries of the clusters are not fixed. Different types of datasets are taken and then experiments are performed on them. It is found experimentally that FRBC shows good results as compared to other fuzzy clustering algorithms. The clusters which are obtained in this way are easily understandable with acceptable accuracy. Clustering is the technique of dividing a large number of objects into a number of groups such that the objects which belong to the same group are most similar to one another and the objects of different groups are most dissimilar to one another. So, the process of combining the objects into groups of objects with similar properties is called clustering [6].

Shi Na et al. has first explained the characteristics of k mean algorithm and then a new enhanced k mean algorithm is planned that basically reduces the measure of iterations. The improved algorithm avoids the calculation of the distance of each object to the cluster centre again and again. First it randomly chooses K data points and calculates the first cluster centres on the basis of smallest Euclidean distance. Two arrays are used to store smallest distance of the clusters. The second one is used to store the cluster centre of the object. This information is useful in reducing the number of times the loops are executed. In this way it reduces the efficiency of k-mean algorithm by increasing the execution speed. Two different types of datasets are used. Then both the k mean and enhanced k mean algorithms are run on the dataset. The experiments show that the enhanced k mean algorithm provides superior performance as compared to traditional k mean algorithm [7].

Sourabh Shah *et al.* has taken three algorithms into consideration in this paper. K medoid, k mean and modified k mean algorithms are compared. In PAM algorithm initially K objects are chosen as medoids. Then we calculate the distance of each object with the medoid and in this way we assign the object to the medoid with the smallest distance. In this way, every data item is allocated to the adjoining

medoid. In next step swapping is done. We swap a medoid m with non medoid o. again the same procedure is followed. New cost is calculated. If this cost us lesser than the previous cost, then the newly chosen object becomes the medoid. After this iteration we swap the non medoid with the medoid and the same procedure is repeated. The whole process continues until there is no change in the rate of medoids. The customized k mean algorithm is as well described. It is approximately alike to the k mean algorithm. The only difference is that in modified k mean algorithm instead of implementing k mean on whole of the dataset, the dataset is split into smaller parts or subparts. Then k mean is applied on these subparts. It is found experimentally that the customized algorithm k mean shows enhanced performance as compared to the traditional k mean algorithm and k medoid on the same dataset [8].

M.D. Boomija *et al.* has explained various clustering algorithms in this paper. These include k-mean, medoid, CLARA and CLARAN. In partitioning based algorithms we divide n items into a set of k clusters. So k partitions are obtained. Suppose D is the set of n objects, and k is the amount of clusters which are to be obtained to be formed, the partitioning algorithm will divide all the D objects into k partitions. The objects which belong to a particular cluster are having similar properties or we can say that are similar to one another where as the objects which belong to different clusters are not similar to one another. The main drawback of portioning algorithms is that we need to tell the number of clusters to be obtained initially and the clusters which are obtained are spherical in shape. Steps followed in k-mean are that it arbitrarily selects k points from dataset. Further it assigns every tip to the group with nearby centroid. It again recalculates the centroids. Allocate every tip to nearby centroid. The procedure repeats until there is no transformation in the position of centroids. In k-medoid initially k objects are chosen as medoids. Then we compute the space of every data item with the medoid and in this way we assign the object to the medoid with the smallest distance. In this way, every data item is allocated to the adjoining medoid. In the nest step swapping is done. We swap a medoid m with a non medoid object o. again the same procedure is followed. New cost is calculated. If this cost is lesser than the previous cost, then the newly chosen object becomes the medoid. After this iteration we swap the non medoid with the medoid and the same procedure is repeated. The whole process continues, until no change. In CLARA data is divided into smaller groups. Then PAM is applied to each of these groups. It produces better results as compared to earlier two algorithms. It is not even much affected by outliers. CLARAN Works in the way same as that of CLARA. The only difference is that small samples are chosen randomly from the dataset [9].

Kwai Han et al. clarifies that information concentrated shared (p2p) systems are finding expanding number of uses. Information mining in such P2P surroundings is a typical development. Be that as it may, common solid information mining configuration don't fit well in these sort of surroundings as they more often than not require bringing together the scattered information which is frequently not reasonable in a gigantic P2P arrange. Circulated information mining calculations that avoid huge scale synchronization or information centralization propose a distinctive decision. This paper considers the scattered k-implies grouping exertion where the information and figuring resources are spread over a vast P2P arrange. It offers two calculations which manufacture a gauge of the outcomes made by the standard concentrated k-mean bunching calculation. The essential is intended to work in a dynamic P2P arrange that can make grouping by limited synchronization as it were. The following calculation utilizes reliably inspected peers and gives intelligent certifications concerning the accuracy of bunching on a p2p arrange. Exploratory outcomes represent that both the calculations uncover excellent execution contrasted with their concentrated partners at the unobtrusive correspondence cost [10]

Vinod Kumar Dehariya *et al.* state that data clustering or grouping is an important key step in processing. Current size of database of organizations is increasing exponentially. Databases now a day include huge quantity of text, image. Organizations want to mine these data heavily to get the valuable information which can be used in marketing and other decision making. Now as image it is not easy to mine it easily as they contain pixels and graphics in detail. In the same way text contains more or less unorganized information. So, data clustering proves to be a reliable tool as it can help in mining the proper information so that decision making could be done in every kind of scenario in the organization, which includes image analysis, text and other tools [11].

Artur Abdullin et al. state that in the new age there is a lot on interest in mining of the database. This interest is mainly due to the fact that because an origination need to evaluate the information which is hidden in the database. There are basically three kinds of information one that is apparent, second which we can calculate or deduce, third which is hidden, this is the third type of information for which data mining is done so that hidden pattern of data can be recognized and future forecasting can be done. Different domains have different version of data mining needs. One can achieve these needs by making a distance matrix from the centroid which explains what exactly is the real data which explains what exactly is the real data which is near to the calculated centroids .The clustering approaches can be combined with other approaches to find out an optimum mix which can help the organization to achieve its targets and making the decision for the future contingencies [12].

Shalove Agarwal et al. state that process of grouping of the object is not merely clustering. It also means that grouping of the linked substance. For e.g. grouping on numerical values may be called a clustering in one context but it is not always true in every context. Sometimes number and text can make a group together. For example, all names and pensions of the much company is not paying for the employees which are currently not working. Second fuzzy it means that some values can be in the boundaries of more than one cluster which can help in making the decision n that how changes we need to do in a cluster to find out the valuable information [13].

III. METHODOLOGY

K-Mean is the traditional partitioning algorithm. Till now various researchers have used it in many fields like biology, insurance, banking, marketing etc. it has faced many modifications because it faces various drawbacks like we need to tell the number of clusters initially, how to choose initial points, large number of iterations. Till date many researches have given their solutions for these problems. Some have used the hybrid approach. Some researchers have reduced the calculations by using their own methods to increase the speed. Some researchers have used a different method to choose initial clusters. Others have used their own methods to choose the no of cluster centres where as some researchers have used median, mode or max min distance to find the minimum distance.

K-means deals with many problems like it is hard to assume the significance of K. For different values of K, clusters we get are different. It works only with numerical data. It is not capable of detecting the noise and outliers. It puts all the data into clusters. It cannot deal with irregular shapes. It cannot work with very large datasets. It does not work well with clusters of diverse thickness.

With the analysis of k mean algorithm, we have found that we can try to improve its speed or increase its efficiency by using our approach and moreover the algorithm can be enhanced in order to deal with the very large datasets. We can make it more robust comparatively. So what we have done is that we take a dataset first. Then that dataset is divided into smaller dataset. Then we run an algorithm which is modified form of k-mean clustering algorithm. In this algorithm we have abridged the amount of repetitions in k-mean clustering algorithm which increases its efficiency. But dividing the dataset into smaller datasets we have made the traditional k-mean more robust in the way that now we can deal with comparatively larger datasets as compared to the traditional k-mean algorithm.

In our study, we have merged two approaches basically, one is splitting the dataset into smaller datasets and other is reducing the number of iterations. What we have done is that we take a dataset first. Then that dataset is divided into smaller dataset. Then we run an algorithm which is modified form of k-mean clustering algorithm. In this algorithm we have abridged the amount of repetitions in k-mean clustering algorithm which increases its efficiency. By dividing the dataset into smaller datasets, we have made the traditional k-mean more robust in the way that now we can deal with comparatively larger datasets as compared to the conventional k-mean algorithm.

While doing the research, the methodology we adopted is that first of all we collected the data on data mining which is known as literature survey. Then the second step was to choose the main topic in data mining on which we want to precede our research. Clustering was chosen as the main topic. A number of research papers were studied to find the problem definition. Here we deal basically with k-mean algorithm which is basically partitioning clustering algorithm. The data relevant to the k-mean algorithm was collected and in order to deal with that problem we present here an enhanced k-means clustering algorithm. Mainly k-means is an algorithm to select the early ideals to go after k-means clustering algorithm. If we choose wrong clusters initially it leads to poor clustering. The k-means algorithm initialised with an random set of group centres. We introduced a different way of selecting the centres and then some method to reduce the number of iterations. Basically, we are splitting the data into smaller sets and then implementing am algorithm on these smaller datasets to reduce the number of iterations.

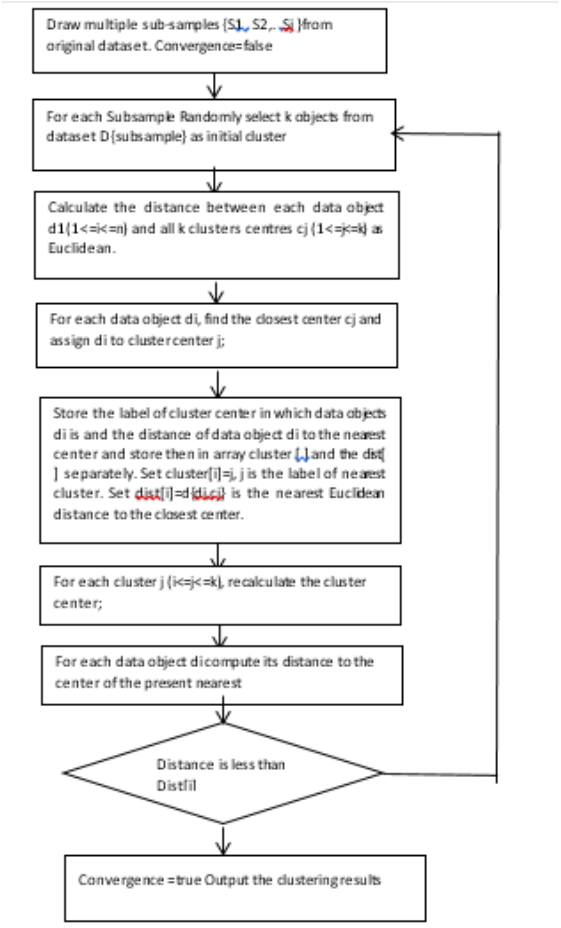

Fig. 1: Flowchart of the proposed methodology.

Steps:

- First of all, draw multiple sub-samples from the original data set.
- From every subsample arbitrarily choose k items from dataset as initial cluster centres.
- Compute the area among every data items and all cluster mid-points as Euclidean area and allocate data items to the adjoining clusters.
- For every data item, locate the nearby centre and set instance to cluster centre.
- Store the tag of cluster middle in which data item is and all the space of data item to the adjoining cluster and accumulate them.
- Recalculate the cluster centre for each cluster.
- For every data item compute its space to the centre of the current adjoining cluster, if this space is fewer than previous distance, the data item stays in the first cluster else for all cluster centre calculate the space of each data item to all the centre, allocate data item to the adjoining middle.
- For every cluster centre recalculate the centres until convergences criteria meets.
- Yield the clustering results.

• IV. RESULTS AND DISCUSSION

What we have done is that first of all we took an inbuilt dataset of weka. Then we ran KMean clustering algorithm which is already defined in weka and then ran KMean updated on the same dataset i.e. email dataset. By running both the algorithms on the same dataset we came to know that our algorithm runs with more efficiency and robustness on the dataset. With efficiency we mean that its processing speed is faster than the traditional KMean algorithm. With robustness we mean that our proposed algorithm can work efficiently with large datasets as compared to traditional K-mean.

Table 1: Time comparison

No. of Clusters	K-Means	K-Means Modified	Hybrid K-Means
clusters=2	10.44	3.54	1.61
clusters=3	18.39	5.94	2.03
clusters=4	20.49	6.55	2.08
clusters=5	21.94	7.1	2.16
clusters=6	22.13	8.5	2.5

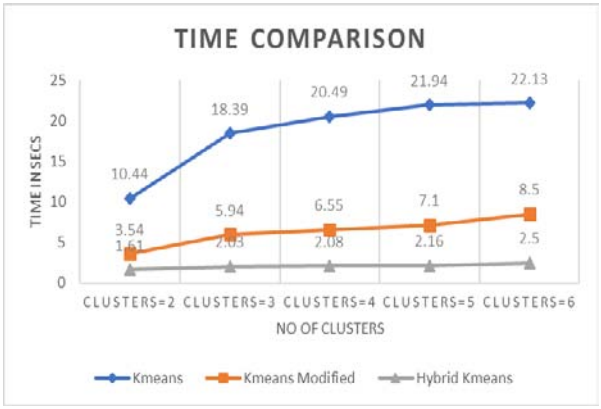

Fig. 2: Time comparison

Fig2, clearly states that the proposed technique executes low time for the processing and evaluation of the results for detecting the type of mail, a spam or ham mail.

Table 2: Various Parameters

Performance	K-Means	K-Means Modified	Hybrid K-Means
Accuracy	59.17	68.6	79.6
Precision	0.76	0.81	0.85
Recall	0.597	0.686	0.78

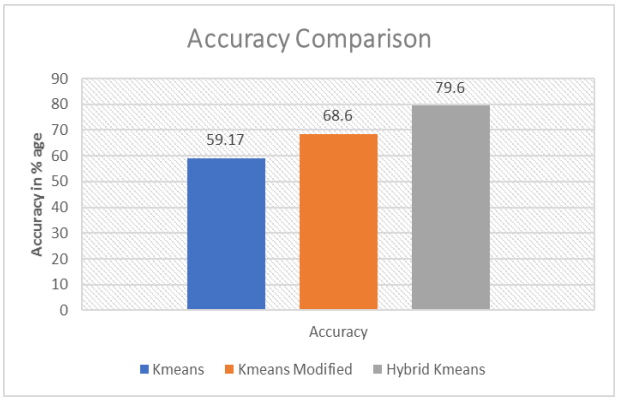

Fig3. Accuracy Comparison

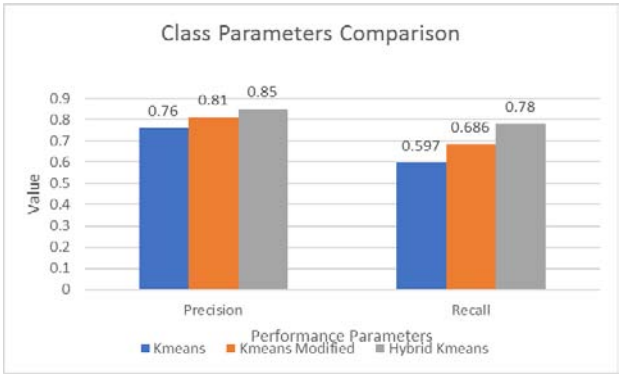

Fig4. Precision and Recall Comparison

Fig 3 and fig 4, demonstrated the various results formulated from our proposed technique. The proposed technique shows 79.6 % accuracy rate. The precision rate is 0.85 and 0.78 rate for recall. This shows that the proposed approach can easily and efficiently detect the spam email or ham email.

IV. CONCLUSION

The proposed algorithm i.e. K-Means Updated emphasizes on the optimum utilization of resources while calculating K-Means. Processing speed increases so processing time reduces. Comparatively large dataset can be processed. With the analysis of K-Mean algorithm, we have found that we can try to improve its speed or increase its efficiency by using our approach and moreover the algorithm can be enhanced in order to deal with very large datasets. We can make it more robust comparatively. So what we have done is that we take a dataset first. Then that dataset is divided into smaller dataset. Then we run an algorithm which is modified form of K-Mean clustering algorithm. In this algorithm we have abridged the amount of repetitions in k-mean clustering algorithm which increases its efficiency. By dividing the dataset into smaller datasets, we have made the traditional K-Mean more robust in the way that now we can deal with comparatively larger datasets as compared to the traditional K-Mean algorithm.

VI. ACKNOWLEDGMENT

The author shows its humble thanks to Lovely Professional University for their appealing contribution and support in the area of research. The author also presents its respect towards Department of Computer Science & Engineering

for the efforts. The author is grateful to mentor Mr. Robin Prakash Mathur for efficient assistance throughout the research work.

VII. REFERENCES

[1] P.Verma, D.Kumar, "Association Rule Mining Algorithm's Variant Analysis", International Journal of Computer Application (IJCA), vol. 78, no. 14, September 2013, pp. 26–34.

[2] Rekha, S. Negi, "A Review on Different Spam Detection Approaches", International Journal of Engineering Trends and Technology (IJETT), vol.11, no.6, May 2014.

[3] Marek Rychly, Pavlina Ticha, "A tool for clustering in data mining", International Federation for Information Processing, 2007.

[4] L.Firte, C.Lemnaru, R.Potolea, "Spam Detection Filter using KNN Algorithm and Resampling", 6th International Conference on Intelligent Computer Communication and Processing- IEEE, 2010, pp.27-33.

[5] G.Kaur, R.K.Gurm, "A Survey on Classification Techniques in Internet Environment", International Journal of Advance Research in Computer and Communication Engineering, vol. 5, no. 3, March 2016, pp. 589–593.

[6] Mansoori Eghal, "A fuzzy rule-based clustering algorithm", IEEE transactions on Fuzzy systems, 2011.

[7] Na shi, "Research on k-means clustering algorithm", 3rd international symposium on intelligent information technology and security informatics, 2011.

[8] Shah Sourabh, Singh Manmohan, "comparison of a time efficient modified k-mean algorithm with k-mean and k-medoid algorithm" international conference on communication systems and network technologies, 2012.

[9] Boomjia M.D, "Comparison of partitioning based clustering algorithms".

[10] Han kwai, "Approximate distributed k-means clustering over a peer-to-peer network", IEEE transactions on knowledge and data engineering, 2009.

[11] Vinod Kumar Dehariya, Shailendra Kumar Shrivastava, R. C. Jain, "Clustering Of Image Data Set Using K-Means And Fuzzy K-Means Algorithms", IEEE 2010 International Conference on Computational Intelligence and Communication Networks, pp. 386-391.

[12] Artur Abdullin, Olfa Nasraoui, "Clustering Heterogeneous Data Sets", IEEE Eighth Latin American Web Congress,2012, pp. 1-8.

[13] Shalove Agarwal, Shashank Yadav, Kanchan Singh, "K-means versus k-means ++ clustering technique", Research Gate, 2012.