International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 06 | June -2020 ISSN: 2582-3930

# Data Clustering Approach for Sentiment Analysis

Kavitha S[1], Usha J[2]

P.G. Student, Dept of Master of Computer Applications, RV College of Engineering®, Bengaluru, Karnataka, India[1]

Professor, Dept of Master of Computer Applications, RV College of Engineering®, Bengaluru, Karnataka, India[2]

**ABSTRACT-** Data mining is defined as the exploration and analysis process for discovering meaningful patterns and rules on a large dataset. In this paper, we present the clustering techniques and the impact on clustering techniques. Clustering is very important for data processing applications and for data mining applications. It is considered one of the Data-Mining Techniques used to find meaning in unlabeled data, and is the most widely used technique of exploratory data analysis used to obtain an understanding of the data meaning. Various ranges of algorithmic techniques such as hierarchical, partitioning, grid-based and density-based algorithms can be used to implement clustering. The technology embraced contributes to the field of Information Retrieval, Data Warehouse and many other fields by providing the required space for data clustering. Sentiment Analysis Technique provides an efficient way to predict future trends and behaviours on large datasets, enabling businesses to make proactive, knowledge-driven, and decision making. The K-means algorithm is one of the simplest, practical and efficient clustering Algorithms that are widely used.

**KEYWORDS -** Data mining, Dataset, K-Mean, Clustering, Analysis.

## I. INTRODUCTION

Mining of data is described as a process by which a large scale amount of data is collected, searched and analyzed in a database as patterns or relationships are discovered. It is a method of evaluating and summing data into valuable and relevant knowledge from various perspectives. Data mining consists of collecting, converting and loading data from transactions into the data warehouse network, storing and handling data in a multidimensional database system, providing data access to business analysts and IT practitioners, analyzing data through application software, presenting data in a useful format such as a graph or chart. It mainly requires identification of anomalies, learning the law of association, classification, regression, description, and clustering.

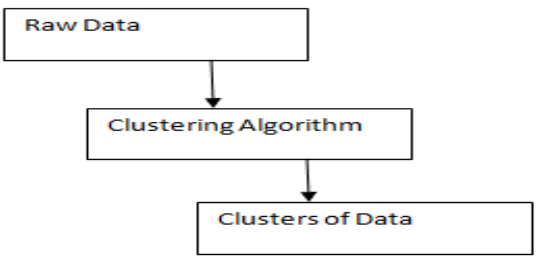


Fig1. Stages of Clustering

The clustering task is to group a set of data objects, so that data objects are more similar in the same group or cluster than data objects in other groups and clusters. It is a major data mining task and a common technique for statistical data analysis in many fields, including recognition of patterns, analysis of image, data recovery, bioinformatics, and compression of data, computer graphics and machine learning. To extract the data, data mining requires two methods of learning, i.e. supervised learning or unsupervised clustering.

**Supervised learning**- The Training data includes the input and expected results. These methods are swift and precise. The appropriate results are known during the learning process and are given in the inputs to the model. Neural network, Multilayer Preceptor, Decision trees are supervised models.

**Unsupervised learning**- The underlying structure of the data can be discovered and it is implemented using two main methods such as principal component and cluster analysis. The unsupervised model does not get the accurate results during the training of machine. This method of learning can be used only for clustering the input data into classes based on their statistical properties. The major task is to group the unsorted data based on their similarity and patterns. It is a machine learning technique and this method is used to find unknown patterns in dataset.

The data mining process is carried out in four steps: data assembly, application of data mining tools on datasets, analysis and outcome assessment, application of results.

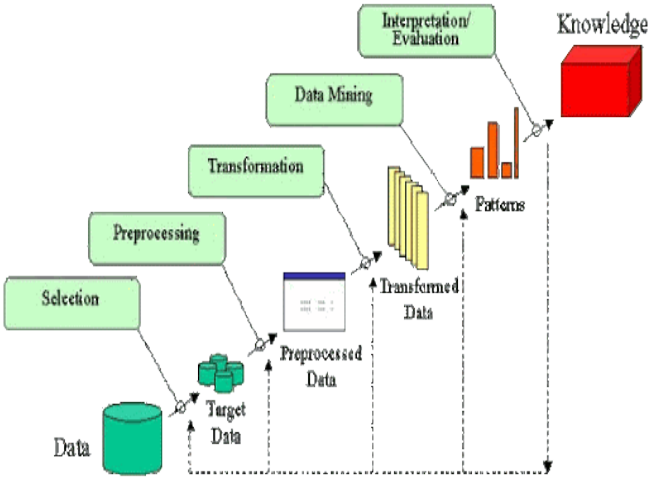


Fig2. Steps of Data Mining Process

Knowledge Discovery in Databases is also known as data mining. As a process, it is carried out using various techniques such as clustering, correlation, and the study of sequential patterns & decision tree. The raw data is processed to build a model that describes the information and brings the report. It can be regarded as a standard measurement of information technology. Data planning and data mining activities complete the data mining process as a method of knowledge-discovery.
It allows more effective usage and resource distribution. It is very much required in processing Big Data for extracting information and also for detection. The data should be given with privacy and security to prevent leakage of the information and it is also used to retrieve the information.

## I. APPLICATIONS

Data Mining is very important because it helps in analyzing significant facts, trends, patterns for the analyst. In Business field, it is used for discovering patterns and relationships in data. By implementing market basket analysis, an appropriate arrangement can be made in a store which helps the customers for selecting frequent buying products. In Bank sector, models can be built from historical customer's data of loans. In manufacturing industry, data mining is used in detecting fault equipments and also to determine the range of perfect product production. It is also used in government agency in analyzing the records of financial transactions.

Various Techniques like classification is used for mapping data in groups, Regression is used for mapping data item to a real valued predicted variable, Clustering is used to form similar group of data items together, Summarization process is used in mapping a data in a subsets and the link analysis is used for defining a relationships among various types of data. The mining task should be scalable and highly efficient to perform on large scale data.

## II. OBJECTIVE

The paper's principal objective is to propose the methodology to cluster the large amount of dataset that helps for prediction and analysis. It reduces the manual work and much of errors and gives the best outcome based on the analysis. The methodology of data mining is embraced to overcome the difficulties faced by the manual work that may eradicate the need for manual entries, and it will help in data prediction and also improves the system performance with increased functionality. Sentiment Analysis is carried out to give an insight to determine the future scope and possibilities and it is also used for decision making.

## III. RELATED WORK

H. Rehioui and A. Idrissi described the large multidimensional data in a huge dataset and obtained the result using clustering approach and also explained that it is a fast technique for data mining [1]. X. Liu attempted to explain the clustering method and pattern analysis, by utilizing algorithm such as k-means clustering algorithm [2]. N. Yambem proposed an approach for processing Bigdata and explained the characteristics issues of Bigdata with various computing techniques [3]. E. Schubert and P.J Rousseeuw in their paper proposed the methodology to find medoids in each cluster using mathematical formula [4]. T. Guptaet al in their paper discussed about the importance of clustering and also gave the comparison of K-means algorithm with other clustering algorithms. They also explained about the advantages of k-means algorithm over other algorithms [5]. L. Matioli explained about the new methodology based on the density estimation by introducing a clustering algorithm [6]. S. Riaz and M. Fatima in their paper brings about the technique for undergoing Sentiment Analysis by implementing a k-means algorithm on large scale data and also explains about the importance of mining method [7]. P. Nithya and M. Kalpana brings about the strategies of Big Data and also discussed about the various clustering algorithms [8]. A. C. Pandey discussed about the analysis technique on Social media based on the people's emotion towards an issue and also gave an insight for decision making using Sentiment Analysis [9]. X. Jin and J. Han in their paper, explained about the clustering mechanism using K-means algorithm and also gave the importance of clustering a huge dataset [10]. B, L. Wang explained about the pattern analysis along with the applications based on the similarity and dissimilarity criterion. He also discussed about the uses of clustering methodology in various fields [11]. G. Pitolli et al in their paper explained the concept of clustering huge dataset. Also explained the identification of clusters based on the similarity and the centroids. [12]. N. D. Dat et al in their paper explained about the Sentiment classification using English words. He discussed about the analysis process which is helpful prediction and decision making [13]. M. Gribaudo et al in his paper brought about the method of improving reliability and also performance of the system in a large scale distributed applications. This gave an insight for data organization and fast performance of the system [14]. H. Suresh et al discussed about an unsupervised clustering technique on Twitter data. He also explained the Sentiment analysis process using the people's comment for further decision making [15]. S. S. Kumari and G. A. Babu in their paper they focused on linear and non-linear clustering using social media comments. They made use of clustering methodology in data mining for sentiment analysis. They explained a methodology for prediction and further analysis process [16]. A. Babu and R. V. Pattani in their paper discussed about one of the data mining technologies known as Segmentation that gave an idea for further implementation of clustering process and sentiment analysis [17]. J. Ovelade et al in their paper discussed about the applications of data mining and also about the advantages of clustering methodology on large scale of data. They also explained about various clustering algorithms that are useful in various application domains [18]. Gelbard et al in their paper proposed an approach for classification by implementing clustering algorithm. They also explained about measures used to apply the algorithm [19]. V. Ajin and L. D Kumar in their paper explained about the big data and also the characteristics associated with the Big data. They also explained about various clustering algorithms used on big data for analysis [20]. F. Jiang et al in their paper mentioned about the methodology by applying formula to calculate mean values and to find centroids in each cluster. They explained about the initialization of k-modes and also about the outlier detection techniques [21]. A. B. Pawar et al explained about the fundamentals and concepts of Sentiment Analysis. They also explained about the methodology used for Sentiment Analysis [22]. Z. Yin et al explained about the methodology for

prediction of data. He also explained about various techniques used for predicting by using various models. This technique is useful for further analysis of data [23]. Y. Yang et al explained about the intertask correlation of data by using clustering methodology on the dataset [24]. H. Zhu et al explained about the sequential classification approach of data for mobile apps. They also explained about the popularity modeling of data [25]. S, R. Saha et al discussed about the features and objective for clustering method based on selection and symmetry using multiple framework [26]. V, M. S. Elyasigomari et al in their paper, they made use of classification technique by means of shuffling by optimizing the data clustering approach based on the selection of data. He explained clustering has more advantage than classification [27]. W, Hu, Q and Pan in their paper, they discussed about the various analysing techniques using supervised clustering algorithm. They explained about the Hierarchical algorithm used for clustering huge amount of data [28]. B. Zerhari et al discussed about the big data clustering and various algorithms for dividing data into similar groups based on closest relation and also discussed about the various challenges faced during clustering process [29]. Aliza et al explained about the Sentiment analysis on Twitter data by implementing data mining techniques used in the application. They explained about the methodology used for processing and analysing twitter comments [30]. The next section is discussed about the materials and methods that describe the hassle.

## IV. MATERIALS AND PROCESS

Various Data Mining tools that are existing for mining data are WEKA, Rapid miner, Tanagra, Orange and Anaconda Navigator. This paper gives the information of data mining tool Anaconda Navigator. For Mining duties, Anaconda Navigator is a desktop graphical user interface (GUI) that is included in the distribution of anaconda that enables to run applications and navigate conda packages, environments and channels quickly without the use command line commands. The tool is free and is an open-source python distribution language that aims to manage the package efficiently after deployed. It can be used in operating systems like Windows, macOS, and Linux. Navigator helps to work in an easy way with conda packages and environments, point-and-click, without having to type any of the condo commands in a terminal window. Various Applications can be accessed using Navigator such as JupyterLab, Jupyter Notebook, Spyder, PyCharm etc. Once the installation has been executed, the window consists of tabs, menus and buttons. The Navigator installs packages in online mode and home tab displays all the available applications, the environment tab manages the installed channels, environments and packages. When the huge dataset is loaded into the application, the various data mining techniques like clustering can be implemented. The unsupervised clustering algorithm like K-means clusters the data based on the similarity and forms the number of clusters by dividing the data. This helps to gain information and is used for further analysis to extract knowledge. As k-means algorithm is an unsupervised algorithm, the dataset need not be trained and tested before loading it to the application. Any type of data can be clustered based on the similarity and the closest mean value. This paper explains the clustering process used on social media such as Face book by extracting the comments of the users and clustering them based on people's emotion towards an issue. After the formation of clusters, analysis process is carried out to extract knowledge that is useful for further decision making process. In the next section, the proposed methodology is discussed with the flow of the process.

## V. PROPOSED METHODOLOGY

The purpose of the Data clustering is to process the huge amount of data to extract knowledge. It is also helpful for fast retrieval of the data. By implementing various data mining techniques, such as Classification, Segmentation, Regression, Prediction, Clustering etc, data is divided into the groups or clusters based on the similarity and organized to extract knowledge and information.

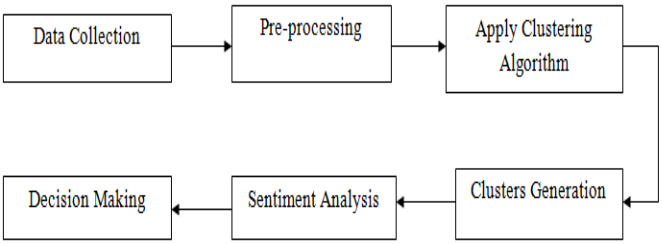


Fig3. Flow Diagram that shows the Clustering and Analysis

### A. Data Collection

The data is collected from Social Media such as Face book which contains the profile details of the users along with the comments and posts. The input is obtained with multiple formats like csv file, excel, etc. The raw data is collected from various sources and further processed which helps in predicting and analysing by applying various clustering algorithms for predictions and comparisons along with decision making. The data collected can be of any data type and also of large scale in size. By the implementation of data mining technologies, data organization is done easily also increases the system performance efficiently.

### B. Pre-processing

The raw data collected from various social media consists of different types of datasets like posts, comments, messages, feedback etc and also with many file formats. In this paper, comments from social media are extracted and preprocessed related to a particular issue. This process is done to minimize the huge amount of data to get a better solution with some amount of accuracy and also the pre-processing is a way of organizing the huge amount of data. This data organization acts as a benefit for better system performance and also in the fast retrieval of information.

For the Clustering Analysis, the data that is collected is provided as an input into Anaconda Navigator in comma separated file format (csv). The file contains comments and feedback given by the users that is used for further clustering

process. This is process is done by applying a clustering algorithm on the dataset that is explained in the following step.

### C. Clustering Algorithm

After pre-processing, the K-means clustering algorithm is applied and implemented on each dataset. The algorithm is the most popular method of cluster partitioning. It is an unsupervised, numerical, non-deterministic, iterative clustering method. Each cluster is represented in k-mean by the mean value of objects within the cluster. Here we partition a set of n objects into a cluster of k so that similarity between inter-clusters is low and similarity between intra-clusters is high. Similarity in terms of mean value of objects in a cluster is calculated. It takes the number of desired clusters and the initial means as inputs, and generates the final means of output. The listed initial and final means are cluster media. If the algorithm is needed to produce K clusters then the initial means will be the value K and the final means will also be the value K. After the implementation of k-means algorithm, each object in a dataset becomes a member of one cluster based on similarity. Shortest distanced mean is considered the mean of the cluster to which the object under examination belongs. K-means algorithm groups the items in the data set into the desired number of clusters. It does some iteration to perform this task until it converges. Calculated means are updated after performing each iteration so that they are closer to the final means. The algorithm eventually converges and ends iterations. In K-means Clustering, various techniques can be used to measure the distance between objects and means. Manhattan Distance and Euclidean Distance are the most popularly used distant metrics. The developer of the clustering system randomly selects the initial means.

In general, we have the xi n data point, i=1... N this must be divided into clusters k. The aim is to assign every data point to a cluster. K-means is a method of clustering which aims to find cluster positions μi, i=1...K which minimizes the distance between cluster data points. This is K-means Remedies to Cluster.

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

Where $S_i$ is represented as the number of points belonging to the cluster i. The clustering of the K-means makes use of the Euclidean distance square.

K-means clustering Algorithmic steps for generating clusters:

Let us consider X = $\{x_1, x_2, x_3, \ldots\ldots\}$ represent a set of data points used to perform clustering and V = $\{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centres used to represent the centres of data points.

1) The initial step is to select 'c' cluster centres randomly.

2) Then calculate the distance from each data point to each cluster centre

3) Assign the data point to the cluster centre at a minimum distance from the cluster centre.

4) In the next step, recalculate using new cluster centre:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, 'ci' stands for the number of cluster data points.

5) Recalculate then, the distance between newly formed cluster centres and every data point, by doing this similar data objects are grouped together forming a loop

6) If there is no reassigned data point then stop, otherwise repeat step 3)

A loop is formed and similar data points are together clustered in one cluster based on their relationships. The procedure to cluster is simple and easy to classify dataset with certain number of clusters. The principal idea for each cluster is to identify k-centroids. K-Means is the simplest algorithm that has been adapted to various problem domains and it is a good technique to work on a randomly generated data points. The following step describes about the clusters generation.

### D. Clusters Generation

After the implementation of clustering algorithm, various clusters are generated from the huge amount of dataset based on the similarity of comments. Clusters are formed based on the mean value by identifying the centroids of the cluster. Each cluster consists of a group of data objects. These data objects in the output clusters are further analyzed and predicted in the following step.

### E. Sentiment Analysis

In Sentiment Analysis, the resulted data clusters are analyzed and calculated with majority comments of the users by the analysis technique that will be a way for further decision making. It also increases the application sustainability. The majority outcome of the system represents the people's opinion towards an issue.

### F. Decision Making

The resulted output is beneficial and used in many cases especially in business matters when the service provided can be improvised and more customers can be attracted. This process has a major role in various fields like Data mining, Information Retrieval, Advertisements, Solving many business problems etc. The outcome and result of clustering methodology is discussed the following section.

## VI. RESULTS AND DISCUSSIONS

Data Mining is an iterative method in which the method of mining can be optimized and new data can be implemented to produce more productive performance. It meets the requirement for effective, scalable and versatile analysis of the data. After the implementation of the k-means clustering algorithm, the data mining tool Anaconda Navigator produces the result. It is one of the best algorithms in the field of data mining and Information Retrieval. As it is polymorphic, it is capable of clustering any data type according to the user's requirements. As it is an unsupervised clustering methodology, the dataset need not be trained and tested before using it to the application. This makes the system work more efficiently when compared to the other clustering algorithms. The result obtained from the clustering technique gives an insight for prediction. It is also used for organizing huge amount of data based on the requirement. The resulted clusters gives the accuracy for

prediction and further the Sentiment Analysis implementation gives a major scope for decision making especially in business matters. In the next section, conclusion about the work and future implementations that can be enhanced for better performance of the system is discussed.

## VII CONCLUSION AND FUTURE WORK

It is shown in the paper that there are several methods for dividing and analyzing huge amount of data. Noise in a real world datasets can degrade the quality of data clustering. Large amount of dataset is also a complex task for analysis. The main advantage of k-means algorithm is its favorable execution time as it is an unsupervised clustering algorithm, class labels need not be created and the dataset is not trained. The system produces the more accurate clustering results and provides the fast time execution with reliable clusters. With the implementation of Sentiment Analysis, people's emotions towards the problem and issue can be identified and determined if the people's overall feedback is happy, unhappy or emotionless. The results of the analysis are very much beneficial especially in business matters for decision making. The clustering methodology also provides a way for fast retrieval of data and reduces much of errors. By considering the input and the outcome, the best approach can be considered. Further research is required to verify the capability of this method when applied to datasets with more complex object distributions. It can also be extended on both real and non-real forms of data. Once the clusters are defined, we can use the categorization algorithms for allotting clusters to the remaining documents. The similarity measure used for clustering can be changed according to the application to improve the results and better system functionality. The following section contains the list of references made during the work carried.

## REFERENSES

[1]. H. Rehioui and A. Idrissi, "A fast clustering approach for large multidimensional data", *Int. J. Bus. Intell. Data Mining*, vol. 15, no. 3, 2019.

[2]. X. Liu, X. Zhu, M. Li, L. Wang, E. Zhu, T. Liu, et al., "Multiple kernel k-means with incomplete kernels", *IEEE transactions on pattern analysis and machine intelligence*, 2019

[3]. N. Yambem and A. Nandakumar, "Big data: Characteristics issues and clustering techniques", *3rd National Conference on Image Processing Computing Communication Networking and Data Analytics*, pp. 348, 2018

[4]. E. Schubert and P. J. Rousseeuw, "Faster k-medoids clustering: Improving the pam clara and clarans algorithms", *arXiv preprint*, 2018.

[5]. T. Gupta and S. P. Panda, "A comparison of k-means clustering algorithm and clara clustering algorithm on iris dataset", *International Journal of Engineering & Technology*, vol. 7, no. 4, pp. 4766-4768, 2018.

[6]. L. Matioli, S. Santos, M. Kleina and E. Leite, "A new algorithm for clustering based on kernel density estimation", *Journal of Applied Statistics*, vol. 45, no. 2, pp. 347-366, 2018.

[7]. S. Riaz, M. Fatima, M. Kamran and M. W. Nisar, "Opinion mining on large scale data using sentiment analysis and k-means clustering", *Cluster Comput.*, pp. 1-16, 2017, [online] Available: https://link.springer.com/journal/10586/onlineFirst/page/75.

[8]. P. Nithya and M. Kalpana, A, "Big data clustering algorithm and strategies", *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 5, no. 6, pp. 1387-1391, 2017.

[9]. A. C. Pandey, D. S. Rajpoot and M. Saraswat, "Twitter sentiment analysis using hybrid cuckoo search method", *Inf. Process. Manage.*, vol. 53, no. 4, pp. 764-779, 2017.

[10]. X. Jin and J. Han, "K-medoids clustering", *Encyclopedia of Machine Learning and Data Mining*, pp. 697-700, 2017.

[11]. B, L. Wang, C. Zhang, F.-z. Wu, Li, Z, et al., "Spectral clustering based on similarity and dissimilarity criterion", *Pattern Analysis and Applications*, vol. 20, no. 2, pp. 495-506, 2017.

[12]. G. Pitolli, L. Aniello, G. Laurenza, L. Querzoni and R. Baldoni, "Malware family identification with birch clustering", *2017 International Carnahan Conference on Security Technology (ICCST)*, pp. 1-6, 2017.

[13]. N. D. Dat, V. N. Phu, V. T. N. Tran, V. T. N. Chau and T. A. Nguyen, "Sting algorithm used english sentiment classification in a parallel environment", *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 31, no. 07, pp. 1750021, 2017.

[14]. M. Gribaudo, M. Iacono, and D. Manini, "Improving reliability and performances in large scale distributed applications with erasure codes and replication," Future Generation Computer Systems, vol. 56, pp. 773– 782, March 2016

[15]. H. Suresh et al., "An unsupervised fuzzy clustering method for twitter sentiment analysis", *Proc. Int. Conf. Comput. Syst. Int. Technol. Sustain. Solutions*, pp. 80-85, 2016.

[16]. S. S. Kumari and G. A. Babu, "Sentiment on social interactions using linear and non-linear clustering", *Proc. 2nd Int. Conf. Adv. Elect. Electron. Inform. Commun. Bio-Inform.*, pp. 177-181, 2016.

[17]. A. Babu and R. V. Pattani, "Efficient density based clustering of tweets and sentimental analysis based on segmentation", *Int. J. Comput. Techn.*, vol. 3, no. 3, pp. 53-57, 2016.

[18]. J. Oyelade et al., "Clustering Algorithms: Their Application to Gene Expression Data", *Bioinform. Biol. Insights*, vol. 10, pp. 237-253, 2016.

[19]. A, Barak, R and Gelbard, "Classification by clustering using an extended saliency measure", *Expert Systems*, vol. 33, no. 1, pp. 46-59, 2016.

[20]. V. Ajin and L. D. Kumar, "Big data and clustering algorithms", *International Conference on Research Advances in Integrated Navigation Systems (RAINS)*, pp. 1-5, 2016.

[21]. F. Jiang, G. Liu, J. Du and Y. Sui, "Initialization of k-modes clustering using outlier detection techniques", *Information Sciences*, vol. 332, pp. 167-183, 2016.

[22]. A. B. Pawar, M. A. Jawale and D. N. Kyatanavar, "Fundamentals of sentiment analysis: concepts and methodology", *Sentiment Analysis and Ontology Engineering*, 2016.

[23] Z. Yin, D. Yin, Z. Chen and Q. Li, "A new combination model for short-term wind power prediction", *Proc. IEEE DRPT*, pp. 1869-1873, Nov. 2015.

[24] Y. Yang, Z. Ma, Y. Yang, F. Nie and H. T. Shen, "Multitask spectral clustering by exploring intertask correlation", *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 1069-1080, May 2015.

[25] H. Zhu, C. Liu, Y. Ge, H. Xiong and E. Chen, "Popularity modeling for mobile Apps: A sequential approach", *IEEE Trans. Cybern.*, vol. 45, no. 7, pp. 1303-1314, Jul. 2015

[26]. S, R. Saha, A. Spandana, Ekbal, S and Bandyopadhyay, "Simultaneous feature selection and symmetry based clustering using multiobjective framework", *Applied Soft Computing Journal*, vol. 29, pp. 479-486, 2015.

[27]. V, M. S. Elyasigomari, H. R. C. Mirjafari, Screen, M. H and Shaheed, "Cancer classification using a novel gene selection approach by means of shuffling based on data clustering with optimization", *Applied Soft Computing*, vol. 35, pp. 43-51, 2015.

[28]. W, Hu, Q and Pan, "Data clustering and analyzing techniques using hierarchical clustering method", *An International Journal*, vol. 74, no. 19, pp. 8495-8504, 2015.

[29]. B. Zerhari, A. A. Lahcen and S. Mouline, "Big data clustering: Algorithms and challenges", *Proc. of Int. Conf. on Big Data Cloud and Applications (BDCA '15)*, 2015.

[30]. Aliza Sarlan, Chayanit Nadam and Shuib Basri, "Twitter sentiment analysis", *Proceedings of the 6th International conference on Information Technology and Multimedia*, 2014