International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Performance Evaluation with K-Mean and K-Mediod in Data Mining

Isha Sharma[1], Kirti Joshi[2]

RIMT-IET, India

Abstract: ***Data mining is the process of extraction of various types of information from different types of dataset that contains various types of attributes. Clustering is an approach that divides the whole information into different clusters. After processing of division of data values into different clusters centeroid have been computed. Cluster centeroid has been done on the basis of distance from other cluster members available in the particular clusters. The main problem in the clustering for data mining process is that text mining contains different problem for division of the text dataset into different cluster. Sometimes in the process of clustering by default empty cluster has been developed. We removed this problem by using K-mean clustering with hybridization of K-mediod algorithm.***

Keywords: Data Mining, K-Method, Clustering, K-mediod

1. Introduction

1.1 Data Mining: Introduction

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is also known as Knowledge Discovery in Data (KDD) [9]. Data mining uses mathematical algorithms to part the data and evaluate the probability of future events. It automatically searches large volume of data to discover pattern and trend. Data mining software is one of a number of analytical tools for analyzing data. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

1.2 Different Forms of Data Mining

a) **Spatiotemporal Data Mining**: Spatiotemporal data are data that relate to both space and time. It refers to the process of discovering patterns and knowledge from spatiotemporal data.
b) **Multimedia Data Mining**: It is discovery of interesting patterns from multimedia databases that store and manage large collections of multimedia objects, including image data, video data, and audio data.
c) **Web Mining:** It is the application of data mining techniques to discover patterns, structures and knowledge from web.
d) **Spatial data mining:** Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases [13]. Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationship and spatial auto correlation [24]. Spatial data are the data related to objects that occupy space. A spatial database stores spatial objects represented by spatial data types and spatial relationship among such objects.

1.3 Data Mining: A KDD Process

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line.

1.4 Clustering

Clustering is the process of partitioning a set of data objects into subsets such that the data elements in a cluster are similar to one another and different from the element of other cluster [9]. The set of cluster resulting from a cluster analysis can be referred to as a clustering. In this context, different clustering methods may generate different clustering's on the same data set. The partitioning is not performed by humans but by the clustering algorithm. Cluster analysis has wide range of application in business intelligence, image pattern recognition, web search, biology, and security.

1.5 Spatial Clustering

Spatial clustering is a process of grouping a set of spatial objects into clusters so that objects within a cluster have high similarity in comparison to one another, but are dissimilar to objects in other clusters [23]. For example, clustering is used to determine the "hot spots" in crime analysis and disease tracking. Hot spot analysis is the

Licensed Under Creative Commons Attribution CC BY

process of finding unusually dense event clusters across time and space. Many criminal justice agencies are exploring the benefits provided by computer technologies to identify crime hot spots in order to take preventive strategies such as deploying saturation patrols in hot spot areas. Spatial clustering can be applied to group similar spatial objects together; the implicit assumption is that patterns in space tend to be grouped rather than randomly located.

2. Review of Literature

Li-Yeh (2012) et al. in the paper "An Improved Particle Swarm Optimization for Data Clustering" proposed an improved particles warm optimization based on Gauss chaotic map for clustering. Gauss chaotic map adopts a random sequence with a random starting point as a parameter, and relies on this parameter to update the positions and velocities of the particles. It provides the significant chaos distribution to balance the exploration and exploitation capability for search process. This easy and fast function generates a random seed processes, and further improve the performance of PSO due to their unpredictability. In the experiments, the eight different clustering algorithms were extensively compared on six test data. The results indicate that the performance of their proposed method is significantly better than the performance of other algorithms for data clustering problem.

Thangamani (2010) et al. in the paper "Integrated Clustering and Feature Selection Scheme for Text Documents" proposed the semantic clustering and feature selection method to improve the clustering and feature selection mechanism with semantic relations of the text documents. The proposed system was designed to identify the semantic relations using the ontology. The ontology was used to represent the term and concept relationship. Results: The synonym, meronym and hypernym relationships were represented in the ontology. The concept weights were estimated with reference to the ontology. The concept weight was used for the clustering process. The system was implemented in two methods. They were term clustering with feature selection and semantic clustering with feature selection. Conclusion: The performance analysis was carried out with the term clustering and semantic clustering methods. The accuracy and efficiency factors were analyzed in the performance analysis.

Jafar (2010) et al. in the paper "Ant-based Clustering Algorithms: A Brief Survey" describes a brief survey on ant-based clustering algorithms. They also present some applications of ant-based clustering algorithms. Ant-based clustering is a biologically inspired data clustering technique. Clustering task aims at the unsupervised classification of patterns in different groups. Clustering problem has been approached from different disciplines during last year's. In recent years, many algorithms have been developed for solving numerical and combinatorial optimization problems. Most promising among them are swarm intelligence algorithms. Clustering with swarm-based algorithms is emerging as an alternative to more conventional clustering techniques. These algorithms have recently been shown to produce good results in a wide variety of real-world applications. During the last five years, research on and with the ant-based clustering algorithms has reached a very promising state.

Sang Jun Lee[2001] et al. in the paper "A review of data mining techniques" Terabytes of data are generated everyday in many organizations. To extract hidden predictive information from large volumes of data, data mining (DM) techniques are needed. Organizations are starting to realize the importance of data mining in their strategic planning and successful application of DM techniques can be an enormous payoff for the organizations. This paper discusses the requirements and challenges of DM, and describes major DM techniques such as statistics, artificial intelligence, decision tree approach, genetic algorithm, and visualization.

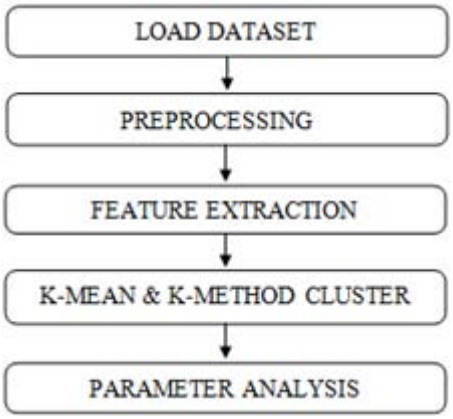
Abbas[2008]et al. in the paper "Comparison between data clustering algorithm" is intended to study and compare different clustering algorithm. The algorithms which are investigated are k-means, hierarchal clustering algorithm, self-organizing algorithm and expectation maximization clustering algorithm. All these algorithms are comparing on the factors: set of dataset, number of cluster, types of dataset and types of software used.

EngYeowCheu[2009] et al. in this paper "On the Two-level Hybrid Clustering Algorithm" present a design of the hybrid clustering algorithms which involve two level clustering. At each of the levels, users can select the k-means, hierarchical or SOM clustering techniques. Unlike the existing cluster analysis techniques, the hybrid clustering approach developed here represents the original data set using a smaller set of prototype vectors (cluster means), which allows efficient use of a clustering algorithm to divide the prototype into groups at the first level. Since the clustering at the first level provides data abstraction first, it reduces the number of samples for the second level clustering. The reduction of the number of samples, hence, the reduction of computational cost is especially important when hierarchical clustering is used in the second stage. The prototypes clustered at the first level are local averages of the data and therefore less sensitive to random variations than the original data..The empirical evaluation of the two-level hybrid clustering algorithms is made at four data sets.

Nedaabdelhamid et al [2015] In this paper "Emerging trends in associative classification data mining" studied emerging trends in associative classification in data mining. Utilising association rule discovery to learn classifiers in data mining is known as associative classification. In the last decade AC algorithms proved to be effective in devising high accurate classification system from various types of supervised datasets. Yet, there are new emerging trends and that can further enhance the performance of current ac method or necessitate the development of new methods. This paper sheds the light on four possible new research trends within AC that could enhance the predictive performance of the classifier or their quality in terms of rules. These possible research directions are considered starting research points for other scholar in rule based classification in data mining.

Licensed Under Creative Commons Attribution CC BY

3. Proposed Work



This figure represents flow diagram of the purposed work that have to be carried out for process of classification. In this process data has been loaded to the system that has been in unstructured format. This data has been undergoes the process of pre processing that is text pre processing and feature extraction from dataset. In the process of pre processing text pre processing has been done using text tokenization and stop word removal filter. This filter has been utilized for division of the dataset into different tokens of the words. Lexical and semantic analysis has been done to divide dataset into different segments. After this process feature selection has been implemented so that best features from the dataset on the basis of class prediction ability can be extracted that can be used for classification or clustering. In the purposed work after feature extraction hybrid algorithm of K-means and medioid has been implemented that has been used for clustering of the dataset on the basis of mean and median distance and density between different centriods of the dataset. On the basis of min distance and maximum density dataset instance has been clustered into different clusters that has been used for extraction of various types of information.

4. Results and Discussions

In the process of data mining various datasets and tools has been used for mining of raw information. In the purposed work WEKA tool has been used for data mining process. This tool provides various functionaries that can be used for data processing. In this too, various supervised and unsupervised filters are available that can be used for preprocessing of the dataset. Classification, clustering and selection algorithm are available that can be used for dataset classification process.

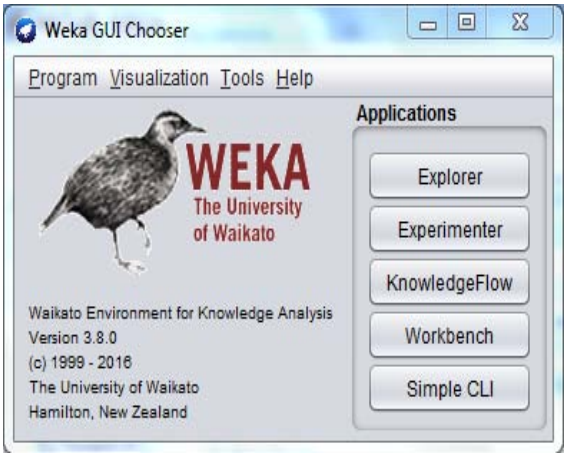


Figure 4.1: WEKA Initialization

This figure is use to represent the initialization of the WEKA. In this window different classes of WEKA have been defined that can be used for data mining process. Explorer is used for data classification, clustering and selection. Experimental class can be used for developing a model that can be used for modalities or generation of various aspects that can be evaluated for data mining.

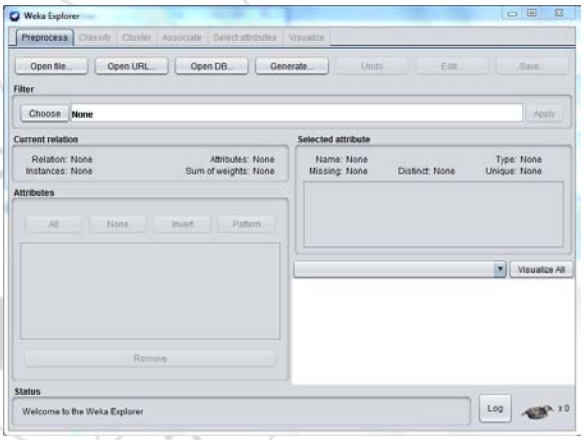


Figure 4.2: Explorer Window

This figure represents explorer window represented in data mining tool. Explorer contains various objects that can be used for loading of the dataset. Dataset can be loaded by using various files in different formats, from URL or DB can also be used for datasets loading. In this process preprocessing has been done to assign class label to dataset, filtering, removal of redundancy, various converters are available that can be used for conversion of the data from numeric to binary, binary to nominal, string to words. After preprocessing filters can be implemented that can be used for data classification or clustering. Selection approaches can be implemented on the dataset to evaluate best attributes for data classification process.

Licensed Under Creative Commons Attribution CC BY

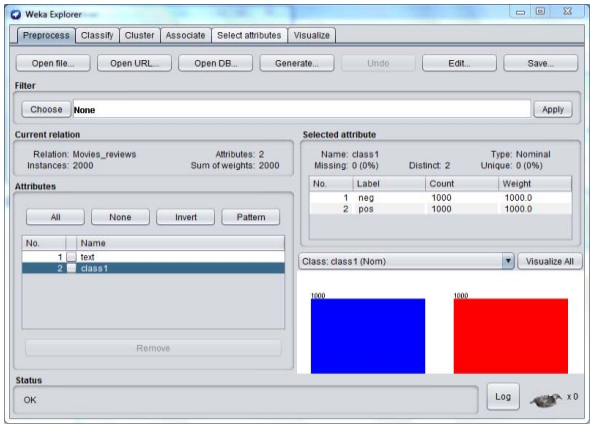


Figure 4.3: Loaded Text Dataset

This figure represents text directory that has been loaded to the system that has been used for text mining. In this two attributes are available one is text that contain review from different users that has been provided by various users, second attribute is class that consists of labels on the basis of reviews of the users. These two attributes having the relation between them to provide description of dataset.

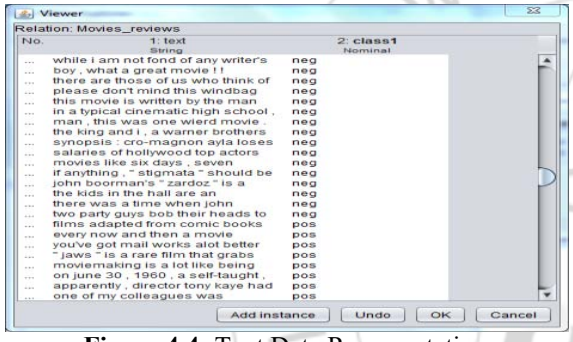


Figure 4.4: Text Data Representation

This Figure represents text data that has been loaded to the system the text directory contains different text reviews from different users that has been used for classification process. These reviews contain a review class that is positive or negative review. In the process of text data stings of different reviews have been represented that has been used for clustering of the data. These strings have been converted into the word that has been done through by implementing string to word convertor filter that use term frequency, inverse document frequency, stop word removal, lemmatization and stemming to convert various unimportant characters and words for clustering of datasets.

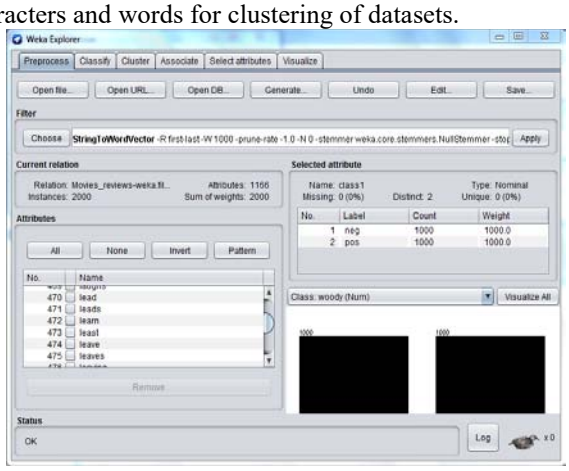


Figure 4.5: Extracted attributes from text

This figure is use to represent the attribute that has been extracted from text, all the numerical symbol characters & verbs have been extracted from text using generic filter that decide the text in to objectives different segments of numerical character.

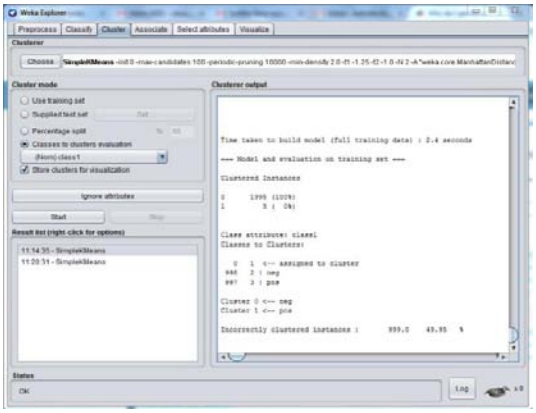


Figure 4.7: Simple K-Mean Clustering

This figure represents clustering of the datasets that has been done through simple K-means approach that has been used for clustering process on the basis of mean values and distance function. This clustering is done to divide datasets into different cluster of positive and negative class that has been used for classification process. Dataset instances have been divided into two different clusters and these have been done using all the attributes available in the dataset.

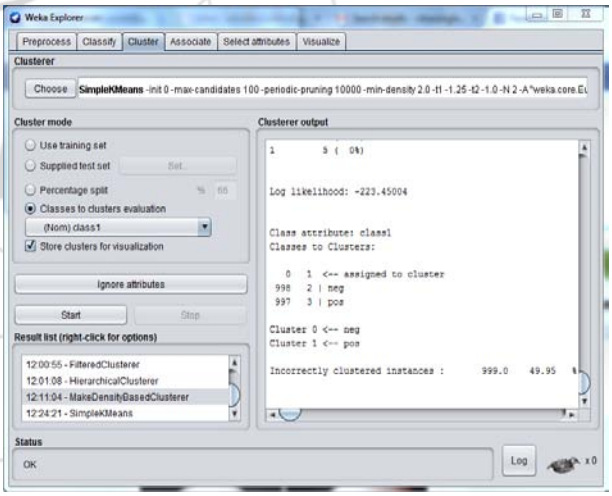


Figure 4.8: K-mediod and K-means based clustering

This figure represents clustering that has been done on the basis of k-means and K-mediod approach that has been used for selection of best center point on the basis of mean and minimum dissimilarity from all data points available in the dataset. On the basis of this clustering using 1166 attributes available in the datasets clustering provide near by 50 percent accuracy.

Licensed Under Creative Commons Attribution CC BY

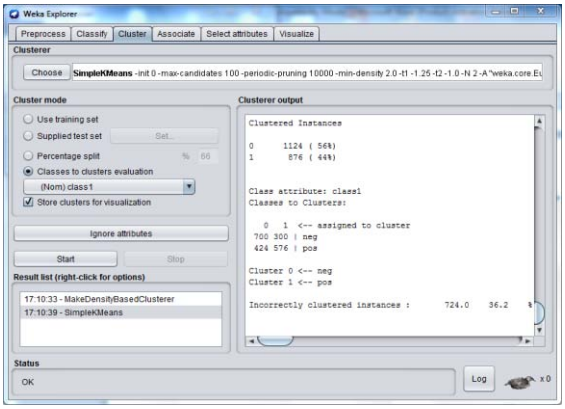


Figure 4.9: Simple K-Mean Clustering using attribute selection

This figure represents clustering of the datasets that has been done through simple K-means approach that has been used for clustering process on the basis of mean values and distance function. This clustering is done to divide datasets into different cluster of positive and negative class that has been used for classification process. Dataset instances have been divided into two different clusters and these have been done using selected attributes in the basis of best first search algorithm.

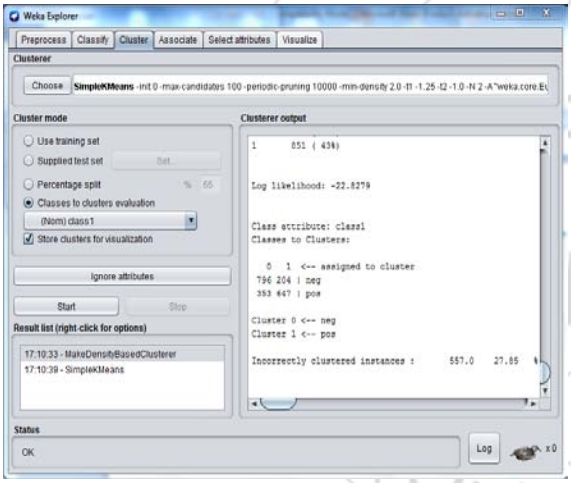


Figure 4.10: K-mediod and K-means based clustering using attribute selection

This figure represents clustering that has been done on the basis of k-means and K-mediod approach that has been used for selection of best center point on the basis of mean and minimum dissimilarity from all data points available in the dataset. On the basis of this clustering using 53 selected attributes available in the datasets clustering provide near by 50 percent accuracy.

Table 5.1: Clustering Accuracy

Clustering approach	Previous	Purposed
Canopy	50.3	63.75
Cobweb	50	50
Farthest first	50.05	52.85
Filtered clusters	50.05	63.80
K-means	50.05	63.80
Hybrid	50.05	72.15

This table represents clustering accuracy provided by different clustering approaches. These different approaches have been implemented on the full dataset and selected attributes dataset and accuracy has been measured that has been given in tabular form.

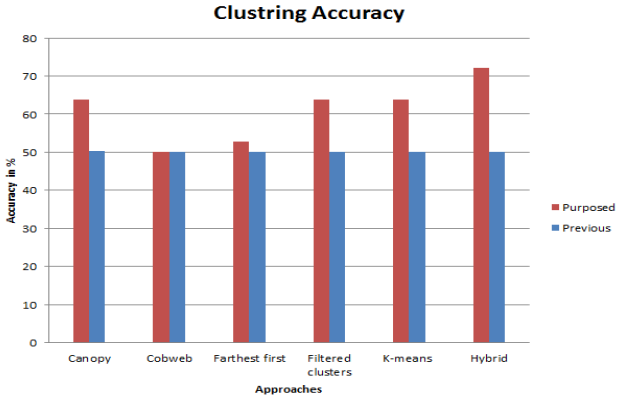


Figure 4.11: Clustering accuracy w.r.t. clustering approaches

5. Conclusion

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. In the processing of data mining various approaches like classification, clustering has been used to divide information into different blocks for extraction of meaning full information. In the process of data mining huge datasets have been used for extraction of knowledge based information. After processing of division of data values into different clusters centeroid have been computed. Cluster centeroid has been done on the basis of distance from other cluster members available in the particular clusters. The main problem in the clustering for data mining process is that text mining contains different problem for division of the text dataset into different cluster. Sometimes in the process of clustering by default empty cluster has been developed. We removed this issue by using K-mean clustering with hybridization of K-mediod algorithm. Purposed approach provides better clustering accuracy.

6. Future Scope

In the future reference this purposed work can be used in real world applications and used for classification after clustering so that prediction of class label can be easy. In this approach in future artificial intelligence approaches can be used that can provide better clustering and selection of attributes.

References

[1] AsmaaBenghabrit, BrahimOuhbi, HichamBehja, BouchraFrikh, "Text Clustering Using Statistical and Semantic Data" World Congress onComputer and Information Technology (WCCIT), pp. 1-6, 22-24 June 2013.

[2] HarpreetKaur,GaganpreetKaur, "A Survey on Comparison between Biogeography Based Optimization and Other Optimization Method" International Journal of Advanced Research in Computer Science and

Licensed Under Creative Commons Attribution CC BY

Software Engineering,Volume 3, Issue 2, February 2013.
[3] J.Durga, D.Sunitha, S.P.Narasimha, B.TejeswiniSunand, "A Survey on Concept Based Mining Model using Various Clustering Techniques" International Journal of Advanced Research in Computer Science and Software Engineering,Volume 2, Issue 4, April 2012.
[4] Mrs. SayantaniGhosh, Mr. Sudipta Roy, and Prof. Samir K. Bandyopadhyay, "A tutorial review on Text Mining Algorithms" International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 4, June 2012.
[5] Li-Yeh Chuang, Yu-Da Lin, and Cheng-Hong Yang, "An Improved Particle Swarm Optimization for Data Clustering" proceedings of the International multiconference on engineers and Computer Scientists 2012, Vol I, March 14-16, 2012.
[6] M. Thangamani and P. Thangaraj, "Integrated Clustering and Feature Selection Scheme for Text Documents" Journal of Computer Science 6 (5): 536-541, 2010.
[7] O.A. Mohamed Jafar and R. Sivakumar, "Ant-based Clustering Algorithms: A Brief Survey" International Journal of Computer Theory and Engineering, Vol. 2, No. 5, October, 2010.
[8] KilianSto_el and AbdelkaderBelkoniene, "Parallel k/h-Means Clustering for Large Data Sets" *gbif.ch/files/content/sites/imi/files/shared/documents/.../Parallel.pdf*
[9] Marcelo N. Ribeiro, Manoel J. R. Neto, and Ricardo B. C. Prudˆencio, "Local feature selection in text clustering" *www.cin.ufpe.br/~rbcp/papers/ICONIP08.pdf*.
[10] Xin-She Yang, Xingshi He, "Firefly Algorithm: Recent Advances and Applications" Int. J. of Swarm Intelligence, 2013 Vol.1, No.1, pp.36 – 50.
[11] A. Abraham, He Guo and Hongbo Liu, "Swarm Intelligence: Foundations, Perspectives and Applications", Swarm Intelligence in Data Mining, A. Abraham, C. Crosan, V. Ramos (Eds.), Studies in Computational Intelligence (series), Springer, Germany, 2006.
[12] A. Abraham and V. Ramos, "Web usage mining using artificial ant colony clustering and linear genetic programming", Proc. Congress on Evolutionary Computation (IEEE Press), Australia, 2003, pp.1384-1391.
[13] Ashish Ghosh, Anindya Halder, Megha Kothari and Susmita Ghosh, "Aggregation pheromone density based data clustering", Information Sciences, Vol. 178, Issue 13, 1 July 2008, pp. 2816-2831.
[14] H. Azzag, N. Monmarche, M. Slimane and G. Venturini, "AntTree: a new model for clustering with artificial ants", Evolutionary Computation, CEC'03, Vol. 4, 2003, 2642-2647.
[15] P. Berkhin, "Survey clustering Data Mining Techniques", Technical Report, Accrue Software, San Jose, California, 2002.
[16] W. Bin and S. Zhongzhi, "A clustering algorithm based on swarm intelligence", Proc. of the Int. Conf. on Info-tech. and Info-net, Beijing,= China, 2001, pp. 58-66.
[17] Nedaabdelhamid, Aladdin Ayesh and FadiThabtah "Emerging trends in associative classification data mining" International journal of electronics and electrical engineering Volume 3, Issue 1, Feb 2015.
[18] E.W.T. Ngai "Application of data mining techniques in customer relationship management: A literature review and classification" Expert Systems with Applications , ELSEVIER, Volume 36, Issue 2, Part 2, March 2009, Pages 2592–2602.
[19] KaustubhS.Chaturbhjet. al.[2016] "Parallel clustering of large data set on Hadoop using data mining techniques"IEEE international conference Futuristic Trends in Research and Innovation for Social Welfare,pp-432-439.
[20] HamzaErolet. al.[2016] "Logical circuit design using orientations of clusters in multivariate data for decision making predictions: A data mining and artificial intelligence algorithm approach" IEEE International Symposium on INnovations in Intelligent SysTems and Applications,pp-2-6.

Licensed Under Creative Commons Attribution CC BY