IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

Performance Analysis on Clustering Approaches for Gene Expression Data

D. Asir Antony Gnana Singh[1], A. Escalin Fernando[2], E. Jebamalar Leavline[3]

Department of CSE, Anna University, BIT Campus, Tiruchirappalli, India[1, 2]

Department of ECE, Anna University, BIT Campus, Tiruchirappalli, India[3]

Abstract: Clustering is a way of finding the structures from a collection of unlabeled gene expression data. A number of algorithms are developed to tackle the problem of clustering the gene expression data. It is important for solving the problems that originate due to unsupervised learning. This paper presents a performance analysis on various clustering algorithm namely K-means, expectation maximization, and density based clustering in order to identify the best clustering algorithm for microarray data. Sum of squared error, log likelihood measures are used to evaluate the performance of these clustering methods.

Keywords: Clustering analysis on microarray data, comparison of clustering algorithms, clustering analysis on gene expression data, literature review on clustering methods, survey on clustering techniques.

I. INTRODUCTION

Clustering is a process of organizing objects into groups whose members are similar in any of the ways. Therefore, a cluster contains similar objects and dissimilar objects are present in different clusters. The ultimate aim of the clustering algorithm is to form the perfect clusters that means grouping objects based on their similarity. There are many similarity measures such as density; distance, etc. have been used. In general, the clustering algorithms learn the unlabeled data therefore it is also called unsupervised learning algorithm. The unsupervised learning algorithm learns the unlabeled data and develops the clustering model. Then, the developed clustering model can be employed to predict the group or cluster of the ungrouped or un-clustered data. The clustering algorithms can be used for various applications such as gene expression data analysis, outlier detection, features selection [1-3], etc. The clustering algorithms can be classified into various types based on the fashion with which the objects are clustered namely model based clustering, density based clustering, connectivity based clustering, centroid based clustering, etc. In this paper, the performance of the density based clustering, expectation maximization (EM) clustering, and K-means clustering are analysed in terms of the sum of squared error (SSE), and log likelihood on various gene expression data.

The rest of the paper is organised as follows. In Section II the clustering techniques are discussed. In Section III experimental setup and the experimental procedures are explained. The results are discussed in Section IV. Finally, Section V concludes this paper.

II. CLUSTERING TECHNIQUES

This section presents the various clustering techniques and their merits and demerits.

A. Model Based Clustering

Cobweb is one of the model based clustering method. It is basically an incremental system intended for hierarchical conceptual clustering. It can set the acuity value which is the minimum standard deviation of the numerical data. The categorical utility threshold value is set by using cut-off to prune the data. Arifovic et al [4] explained that the genetic algorithm would have a better convergence for wider range of parameters. Hommes et al [5] suggested that cobweb model can show consistent rational behaviour for non-linear dynamic models. The similarity measure used in the cobweb is the distance measure. Alejos et al [6] presented a technique to calculate the magnetization of the simulated system with improved accuracy by means of the Preisach model. Zhechong Zhao et al [7] presented a cobweb plot which is used to illustrate graphically the iterative procedure and to analyse stability.

It investigates the quantitative behaviour of one dimensional iterated function using a fixed point known as invariant point. Yuni Xia et al [8] proposed a conceptual clustering algorithm which can explicitly handle the uncertainties in the values of the dataset. Total utility (TU) index is introduced to measure the quality of the clustering. This finally increases the internal probabilistic information of the clustering performance. The advantage of cobweb is that it allows a bidirectional search. It allows merging and splitting of classes using the category utility. The major disadvantage is that it is purely based on the assumption of probability distributions were updating and storing of clusters is quite expensive.

Expectation maximization (EM) clustering is another type under model based clustering, which is an iterative method and is capable of finding the maximum likelihood in statistical methods. The expectation of the likelihood is computed by performing the alternate iteration between the performances of the expectation step. Tood et al [9] stated that this algorithm is suitable for outcomes that are clumped together. Brankov et al [10] proposed the normalized cross correlation that has better performance than the traditionally used Euclidean distance which is

used as the similarity measure for the expectation maximization. The EM algorithm is mainly suitable for the analysis of the image data. Lagendijk et al [11] applied the maximum likelihood approach to identify and restore noisy data in the blurred images. This EM method can facilitate maximizing likelihood functions that arise in statistically estimating problems. Figueiredo et al [12] presented the algorithm for the restoration of the images using the penalized likelihood. Fessler et al [13] presented a new update of sequentially alternating the parameters between the several small hidden data spaces which are defined by the algorithm designer. EM is most suitable for the real world dataset and is best suited for performing cluster analysis for a small scene and when not satisfactory with the results of simple K-means algorithm. The drawback of EM algorithm is its inherent complexity.

B. Farthest First Clustering

The farthest first cluster places the centre of each cluster at a point farthest from the existing cluster centres. It is a variant of simple K-means. Manoj et al [14] suggested that the farthest first algorithm is suitable for the large dataset and the clusters produced are non-uniform. So they developed an optimized farthest first clustering algorithm to produce uniform clusters. Chung-Ming et al [15] proposed a farthest first forwarding algorithm to reduce the transmission delay in the vehicular adhoc networks (VANETs). H. K. Yogish et al [16] proposed a strategy of farthest first traversal for finding the frequent traversal path in the navigation and reorganization of the website structure. This clustering algorithm can eventually speed up the clustering since there are only few adjustments in the data. The constraint based methods and distance-function learning methods according to Bilenko et al [17] are the similarity metric used in the algorithm. The major advantage is that it is a heuristic based method that is fast, scalable and appropriate for large datasets. But it is difficult to compare the quality of the cluster produced. It does not hold good for non-globular clusters and is very sensitive to outliers.

C. Filter Based Clustering

This type of clustering method is for filtering the information or any pattern which are essentially needed. The filtration is carried out based on the keywords that are supplied or some relevant information. Jiang-She Zhang et al [18] proposed a clustering algorithm for the processing of the images. They are computationally stable and insensitive to initialization. They also produce consistent clusters. Thomas et al [19] proposed a collaborative filtering which is a combination of the correlation and singular value decomposition (SVD) to improve accuracy. A weighted co-clustering algorithm is designed in incremental and parallel versions and the results are empirically evaluated. Lagendijk et al [20] proposed two different methods to estimate the performances of individual classifiers and then combine them based on the weight of the individual classifiers. The advantage is that it compares the new arriving keywords with the existing profile and information is provided to the user. It also checks the information instantly rather than waiting for the other information from the user. The major drawback is that the user cannot get the information about the filtering algorithm that is being used. It also depends on the feedback of the retrieved information.

D. Connectivity Based Clustering

Hierarchical clustering is a type of connectivity based clustering and it is the way of relating the objects having core idea of the objects closer than the objects that are farther. The main categories of the hierarchical clustering are “Agglomerative” and “Divisive” methods. Edward J. Coyle et al [21] proposed a randomized algorithm which is mainly applicable for the sensors for the generation of the cluster heads in a hierarchical manner. Michael Dittenbach et al [22] presented a growing hierarchical self-organizing map that evolves on the input data during the unsupervised training process. Guangyu et al [23] developed a comparative analysis and suggested that hierarchical clustering is better when compared with the conventional clustering. It produces an extensive hierarchy of clusters that merge with other ideas that are present at a certain distance. The disadvantage of using hierarchical cluster is that it cannot provide single partitioning of the dataset.

E. Density Based Clustering

The density based clustering (DBC) groups the objects mainly based on the density of the objects that are reachable and connective. Li Tu et al [24] proposed a framework called D-stream for clustering using the density based approach. Mitra et al [25] suggested a nonparametric data reduction scheme. The procedure followed here is separating the dense area objects from less dense area with the aid of an arbitrary object. The density based clusters (DBC) are robust to noise but the datasets are problematic and requires high densely connected data.

F. Centroid Based clustering

K-Means is a centroid based clustering method. It partitions the dataset into various clusters based on the mean distance. It is one of the simplest forms of unsupervised algorithm. The main objective of this algorithm is to reduce the squared error. Tapas et al [26] identified that the algorithm works faster as the separation between the cluster increases. This algorithm is applicable for the segmentation of images and data compression. Kanungo et al [27] proposed that the K-means algorithm runs faster as the separation between the cluster increases. Jakob J. Verbeek et al [28] suggested a solution to reduce the computational load without affecting the quality of the solution significantly. The algorithm is robust, fast and easy to understand. It also yields better results when the dataset are well separated or distinct from each other. It does not work efficiently for non-linear and categorical data. Further, it is unable to handle outliers and noisy data if the cluster centres are randomly chosen.

III.EXPERIMENTAL SETUP

The performances of the clustering methods are compared by considering the gene expression datasets such as SRBCT, Lymphoma and three different Leukaemia

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

datasets namely Leukaemia, Leukaemia3C, and Leukaemia-1. The medical datasets that were considered are SRBCT dataset with 2309 attributes and 83 instances. The Lymphoma dataset consists of 4027 attributes and 66 instances. Three different datasets of Leukaemia with 7130 attributes and 72 instances each have also been used. The performance is evaluated by performing the operation of clustering in each of the datasets. The numbers of the clusters are varied from two to ten and the resulting sum of squared error (SSE) and the log likelihood (LL) are noted for each of the methods. The comparison is carried out among the clustering methods namely, K-means, density based clustering (DBC) and expectation maximization (EM) clustering. The experiment is carried out to obtain the better clustering method for the gene expression data.
The experiments were performed using the WEKA data mining tool. It is developed with the Java programming language and it contains the GUI that is capable of interacting with the various data files and even produces visual results. WEKA tool provides various other options on pre-processing, classification, clustering, association, selection of attributes, and visualization.

2.1 Experimental Procedure
The experiment is carried out using the experimental procedure with the following steps:

Step 1: Read the dataset.
Step 2: Set the number of clusters to be formed for clustering the instances of the dataset.
Step 3: The sum of squared error is noted for the K-means and density based clustering method.
Step 4: The log likelihood is noted for the density based clustering and expectation maximization clustering method.

Initially, the data set is read. Then, the number of clusters to formed is set (from 2 to 10) for clustering the instances of the dataset. Then, sum of squared error is noted for the K- means and density based clustering method and the log likelihood is noted for the density based clustering and expectation maximization clustering method.

IV. RESULTS AND DISCUSSION

This section illustrates the results obtained from the conducted experiments. Table I shows the sum of squared errors for the K-means and density based clustering, Table II shows the log likelihood values of the density based clustering, Table III shows the log likelihood values of the expectation maximization clustering, Figure 1 depicts the sum of squared error for the K-means and density based clustering for five different datasets.
Figure 2 illustrates the log likelihood of the density based clustering and expectation maximization clustering for the SRBCT dataset. Figure 3 depicts log likelihood of the density based clustering and expectation maximization clustering for the Lymphoma dataset. Figure 4 illustrates the log likelihood of the density based clustering and expectation maximization clustering for the Leukaemia dataset. Figure 5 depicts log likelihood of the density based clustering and expectation maximization clustering for the Leukaemia 3C dataset. Figure 6 depicts the log likelihood of the density based clustering and expectation maximization clustering for the Leukaemia-1 dataset.

TABLE 1 SUM OF SQUARED ERRORS FOR THE K-MEANS AND DENSITY BASED CLUSTERING

Datasets	Number of clusters								
	2	3	4	5	6	7	8	9	10
SRBCT	07131.72	06743.21	06440.08	06123.40	05719.74	05467.57	05062.60	05022.87	04817.73
Lymphoma	08970.75	08546.91	07960.22	07684.71	07516.37	07139.95	06880.25	06791.88	06669.48
Leukaemia	16376.73	15409.02	15137.53	14803.38	14006.20	13693.59	13411.96	12966.56	12632.31
Leukaemia3C	16373.73	15407.02	15236.14	14900.66	14028.98	13765.59	13483.96	13011.07	12676.82
Leukaemia-1	16368.73	15400.02	15227.14	14891.66	14020.98	13737.54	13351.06	13005.07	12670.82

TABLE 2 LOG LIKELIHOOD VALUES OF THE DENSITY BASED CLUSTERING

Datasets	Number of Clusters								
	2	3	4	5	6	7	8	9	10
SRBCT	-01252.27	-01093.28	-01016.29	-00865.44	-00701.76	-00596.40	-00418.96	-00334.48	-00286.77
Lymphoma	-03162.03	-02993.16	-02787.76	-02658.80	-02630.60	-02469.81	-02352.89	-02341.43	-02231.36
Leukaemia	-47267.70	-46848.50	-46753.89	-46686.34	-46276.76	-46089.52	-45885.16	-45665.65	-45582.44
Leukaemia3C	-47267.55	-46848.34	-46715.70	-46648.40	-46216.85	-46041.36	-45837.00	-45612.04	-45528.83
Leukaemia-1	-47267.23	-46848.02	-46715.37	-46648.07	-46216.55	-46034.78	-45789.04	-45611.77	-45528.56

TABLE 3 LOG LIKELIHOOD VALUES OF THE EXPECTATION MAXIMIZATION CLUSTERING

Datasets	Number of Clusters								
	2	3	4	5	6	7	8	9	10
SRBCT	-01122.50	-00943.74	-00759.65	-00654.04	-00535.71	-00451.53	-00340.92	-00259.80	-00168.69
Lymphoma	-03137.02	-02916.35	-02759.67	-02683.67	-02413.22	-02260.96	-02355.60	-01979.22	-01910.53
Leukaemia	-47351.64	-46882.29	-46578.92	-46291.84	-46070.29	-45886.79	-45770.37	-45456.07	-45410.69
Leukaemia3C	-47351.49	-46882.13	-46578.76	-46291.69	-46070.18	-45886.67	-45770.27	-45455.95	-45410.59
Leukaemia -1	-47324.03	-46881.82	-46560.46	-46291.44	-46069.86	-45886.44	-45770.03	-45455.67	-45410.34

DOI 10.17148/IJARCCE.2016.5242

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

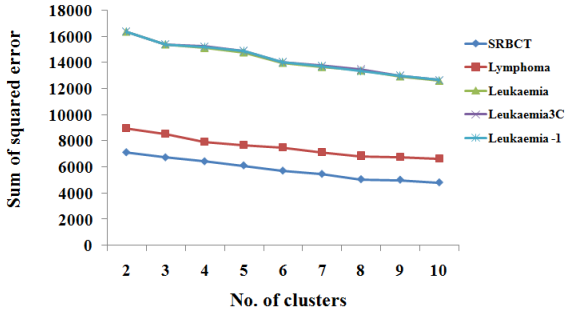

Fig1. Sum of squared error for the K-means and density based clustering for five different datasets.

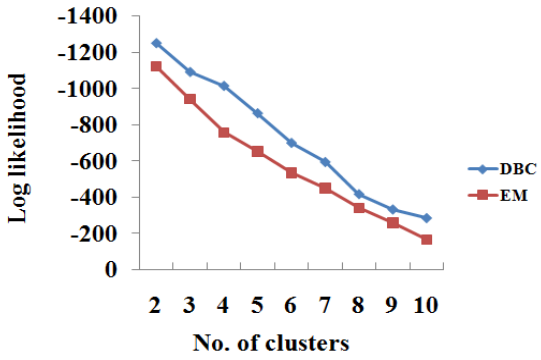

Fig2. Log likelihood of the density based clustering and expectation maximization clustering for the SRBCT dataset.

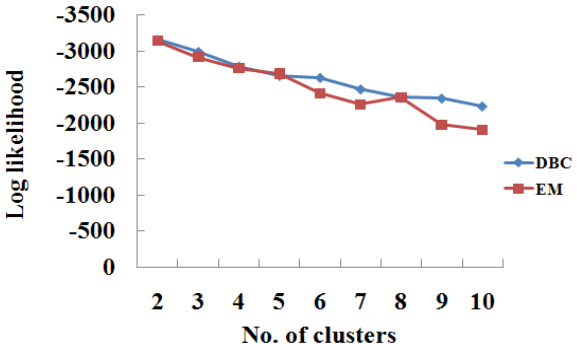

Fig3. Log likelihood of the density based clustering and expectation maximization clustering for the Lymphoma dataset.

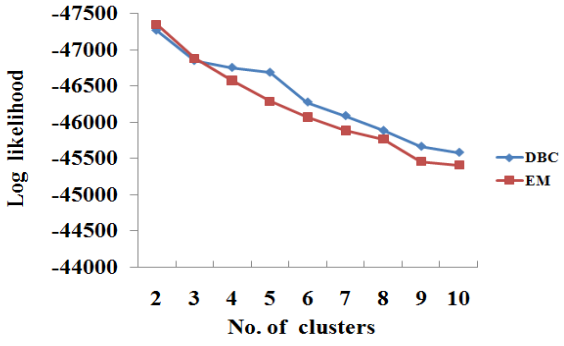

Fig4. Log likelihood of the density based clustering and expectation maximization clustering for the Leukaemia dataset.

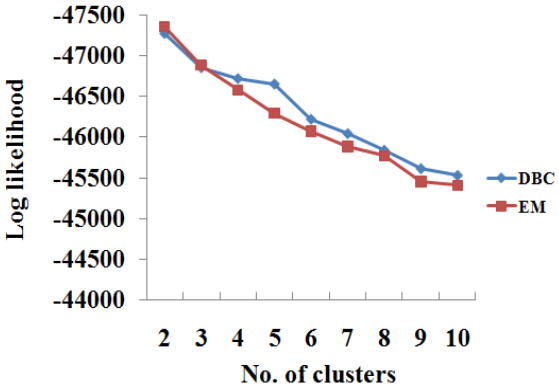

Fig5. Log likelihood of the density based clustering and expectation maximization clustering for the Leukaemia 3C dataset.

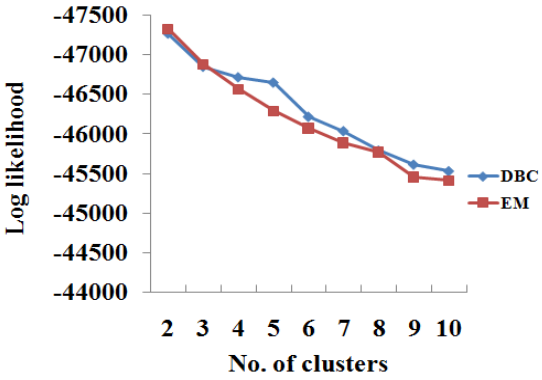

Fig6. Log likelihood of the density based clustering and expectation maximization clustering for the Leukaemia-1 dataset.

From Table 1 and Figure 1, it observed that the K-means clustering and density based clustering methods perform similarly in terms of SSE for different number of clusters on all the datasets and also it is observed that the clustering methods produce lesser SSE with SRBCT dataset compared to other datasets. From the Table 2, Table 3 and Figure 2 to 6, it is observed that the expectation maximization clustering method performs better than the density based clustering method in terms of log likelihood.

V. CONCLUSION

This paper conducted an empirical study on various clustering algorithms in order to observe their performance on gene expression data in terms of sum of squared error and log likelihood. In this empirical study, the performance of the clustering algorithms namely density based clustering, expectation maximization clustering and K-means clustering are evaluated on various gene expression data. From this evaluation, it is observed that the performance of expectation maximization clustering algorithm is comparatively better than the density based clustering algorithm in terms of log likelihood. The clustering algorithms namely K-means and density based clustering method perform similarly in terms of sum of squared error. The SRBCT dataset possesses less sum of

DOI 10.17148/IJARCCE.2016.5242

squared error for the clustering algorithms K-means and density based clustering than all other datasets compared.

REFERENCES

[1] Danasingh Asir Antony Gnana Singh, Subramanian Appavu Alias Balamurugan, Epiphany Jebamalar Leavline, 'A novel feature selection method for image classification', Optoelectronics and Advanced Materials, Rapid Communications, 9.11-12 (2015) :1362- 1368

[2] Danasingh Asir Antony Gnana Singh, Subramanian Appavu Alias Balamurugan, Epiphany Jebamalar Leavline. "An unsupervised feature selection algorithm with feature ranking for maximizing performance of the classifiers." International Journal of Automation and Computing 12.5 (2015): 511-517.

[3] Danasingh Asir Antony Gnana Singh, Subramanian Appavu Alias Balamurugan, Epiphany Jebamalar Leavline, 'Improving the Accuracy of the Supervised Learners using Unsupervised based Variable Selection', Asian Journal of Information Technology, 13.9 (2014): 530-537.

[4] Arifovic, Jasmina. "Genetic algorithm learning and the cobweb model." Journal of Economic dynamics and Control 18.1 (1994): 3-28.

[5] Hommes, Cars H. "On the consistency of backward-looking expectations: The case of the cobweb." Journal of Economic Behavior & Organization 33.3 (1998): 333-362.

[6] Alejos, Óscar, and Edward Della Torre. "The generalized cobweb method." Magnetics, IEEE Transactions on 41.5 (2005): 1552-1555.

[7] Zhao, Zhechong, and Lei Wu. "Stability analysis for power systems with pricebased demand response via Cobweb Plot." Proc. IEEE PES General Meeting. 2013.

[8] Yuni Xia,Bowei Xi "Conceptual Clustering Categorical Data with Uncertainty" 19th IEEE International Conference on Tools with Artificial Intelligence

[9] Moon, Tood K. "The expectation-maximization algorithm." Signal processing magazine, IEEE 13.6 (1996): 47-60.

[10] Brankov, Jovan G., et al. "Similarity based clustering using the expectation maximization algorithm." Image Processing. 2002. Proceedings. 2002 International Conference on. Vol. 1. IEEE, 2002.

[11] Lagendijk, Reginald L., Jan Biemond, and Dick E. Boekee. "Identification and restoration of noisy blurred images using the expectation-maximization algorithm." IEEE Transactions on Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], 38 (7) (1990).

[12] Figueiredo, Mário AT, and Robert D. Nowak. "An EM algorithm for wavelet-based image restoration." Image Processing, IEEE Transactions on 12.8 (2003): 906-916.

[13] Fessler, Jeffrey, and Alfred O. Hero. "Space-alternating generalized expectation-maximization algorithm." Signal Processing, IEEE Transactions on 42.10 (1994): 2664-2677.

[14] Kumar, Manoj. "An optimized farthest first clustering algorithm." Engineering (NUiCONE), 2013 Nirma University International Conference on. IEEE, 2013.

[15] Huang, Chung-Ming, et al. "A farthest-first forwarding algorithm in VANETs." ITS Telecommunications (ITST), 2012 12th International Conference on. IEEE, 2012.

[16] Vadeyar, Deepshree A., and H. K. Yogish. "Farthest First Clustering in Links Reorganization." International Journal of Web & Semantic Technology 5.3 (2014): 17.

[17] Bilenko, Mikhail, Sugato Basu, and Raymond J. Mooney. "Integrating constraints and metric learning in semi-supervised clustering." Proceedings of the twenty-first international conference on Machine learning. ACM, 2004.

[18] Leung, Yee, Jiang-She Zhang, and Zong-Ben Xu. "Clustering by scale-space filtering." Pattern Analysis and Machine Intelligence, IEEE Transactions on 22.12 (2000): 1396-1410.

[19] George, Thomas, and Srujana Merugu. "A scalable collaborative filtering framework based on co-clustering." Data Mining, Fifth IEEE International Conference on. IEEE, 2005.

[20] Lagendijk, Reginald L., Jan Biemond, and Dick E. Boekee. "Identification and restoration of noisy blurred images using the expectation-maximization algorithm." IEEE Transactions on Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], 38 (7) (1990).

[21] Bandyopadhyay, Seema, and Edward J. Coyle. "An energy efficient hierarchical clustering algorithm for wireless sensor networks." INFOCOM 2003. Twenty-Second Annual Joint Conferences of the IEEE Computer and Communications. IEEE Societies. Vol. 3. IEEE, 2003.

[22] Dittenbach, Michael, Dieter Merkl, and Andreas Rauber. "The growing hierarchical self-organizing map." ijcnn. IEEE, 2000.

[23] Pei, Guangyu, et al. "A wireless hierarchical routing protocol with group mobility." Wireless Communications and Networking Conference, 1999. WCNC. 1999 IEEE. IEEE, 1999.

[24] Chen, Yixin, and Li Tu. "Density-based clustering for real-time stream data." Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007.

[25] Mitra, Pabitra, C. A. Murthy, and Sankar K. Pal. "Density-based multiscale data condensation." Pattern Analysis and Machine Intelligence, IEEE Transactions on 24.6 (2002): 734-747.

[26] Kanungo, Tapas, et al. "An efficient k-means clustering algorithm: Analysis and implementation." Pattern Analysis and Machine Intelligence, IEEE Transactions on 24.7 (2002): 881-892.

[27] Kanungo, Tapas, et al. "An efficient k-means clustering algorithm: Analysis and implementation." Pattern Analysis and Machine Intelligence, IEEE Transactions on 24.7 (2002): 881-892.

[28] Likas, Aristidis, Nikos Vlassis, and Jakob J. Verbeek. "The global k-means clustering algorithm." Pattern recognition 36.2 (2003): 451-461.