Clustering Techniques and the Similarity Measures used in Clustering: A Survey

Jasmine Irani
Department of Computer Engineering
MIT Pune
Pune, India

Nitin Pise
Department of Computer Engineering
MIT Pune
Pune, India

Madhura Phatak
Department of Computer Engineering
MIT Pune
Pune, India

ABSTRACT
Clustering is an unsupervised learning technique which aims at grouping a set of objects into clusters so that objects in the same clusters should be similar as possible, whereas objects in one cluster should be as dissimilar as possible from objects in other clusters. Cluster analysis aims to group a collection of patterns into clusters based on similarity. A typical clustering technique uses a similarity function for comparing various data items. This paper covers the survey of various clustering techniques, the current similarity measures based on distance based clustering, explains the limitations associated with the existing clustering techniques and propose that the combination of the advantages of the existing systems can help overcome the limitations of the existing systems.

General Terms
Data Mining, Machine Learning, Clustering, Pattern based Similarity, Negative Data, et. al.

Keywords
pattern based similarity, negative data clustering, similarity measures.

1. INTRODUCTION

1.1 Clustering
Clustering using distance functions, called distance based clustering, is a very popular technique to cluster the objects and has given good results. The clusters are formed in such a way that any two data objects within a cluster have a minimum distance value and any two data objects across different clusters have a maximum distance value

1.2 Similarity of data
Similarity is an amount that reflects the strength of relationship between two data items, it represents how similar 2 data patterns are. Clustering is done based on a similarity measure to group similar data objects together. This similarity measure is most commonly and in most applications based on distance functions such as Euclidean distance, Manhattan distance, Minkowski distance, Cosine similarity, etc. to group objects in clusters. The clusters are formed in such a way that any two data objects within a cluster have a minimum distance value and any two data objects across different clusters have a maximum distance value. Clustering using distance functions, called distance based clustering, is a very popular technique to cluster the objects and has given good results.

There are however, many limitations associated with distance measures based clustering which have been addressed in this paper and are aiming to overcome in our research.

1.3 Categories of Clustering
Clustering algorithms can be categorized broadly into the following categories:

1. Partitional Clustering
2. Density based Clustering
3. Hierarchical clustering

1.3.1 Partitional Clustering
Partitional clustering is considered to be the most popular category of clustering algorithm.

Partition clustering algorithm divides the data points into "k" partitions, where each partition represents a cluster. The partition is done based on a certain objective function. The clusters are formed such that the data objects within a cluster are "similar", and the data objects in different clusters are "dissimilar".

Partitional clustering methods are useful in applications where the number of clusters required are static.

K-means, PAM (Partition around mediods) and CLARA are a few of the partitioning clustering algorithms.

1.3.2 Density Based Clustering
Density-based clustering algorithms create arbitrary-shaped clusters. In this kind of clustering approach, a cluster is considered as a region in which the density of data objects exceeds a particular threshold value.

DBSCAN algorithm is a famous example of Density based clustering approach.

1.3.3 Hierarchical Clustering
Hierarchical clustering algorithms work to divide or merge a particular dataset into a sequence of nested partitions. The hierarchy of these nested partitions can be of two types, viz., agglomerative, i.e., bottom-up or divisive, i.e., top-down.

In the agglomerative method, clustering begins with a single data object in a single cluster and continues to cluster the closest pairs of clusters until all the data objects are grouped together in just one cluster.

Divisive hierarchical clustering, on the other hand, starts with all data objects in a single cluster and keeps splitting larger clusters into smaller ones until all the data objects are split into unit clusters.

BIRCH, (Balance Iterative Reducing and Clustering using Hierarchies), CURE (Cluster Using REpresentatives) are examples of Hierarchical clustering approach.

2. OVERWIEW OF DISTANCE METRICS USED FOR CLUSTERING

Distance metrics play a very important role in order to measure the similarity among the data objects. The main requirement of metric calculation in a specific problem is to obtain an appropriate distance /similarity function. A metric function or distance function is a function that defines a distance between elements/objects of a set [34,35]. A set with a metric is called metric space[7]. This distance metric plays a crucial role in clustering techniques. In this paper, an example of the k-means clustering algorithm using Euclidean distance metric is given. Normally, the job is to define a function Similarity(X,Y), where X and Y are two objects or sets of a certain class, and the value of the function represents the degree of "similarity" between the two[7]. Fig 1 shows the example of a generalized clustering process using distance measures.

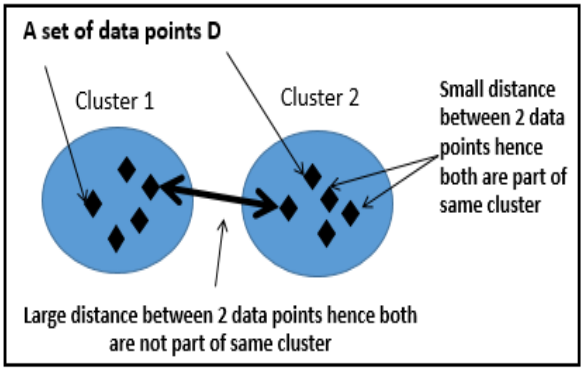

Fig 1: Example of the generalized clustering process using distance measures

2.1 Similarity Measures

A similarity measure can be defined as the distance between various data points. While, similarity is an amount that reflects the strength of relationship between two data items, dissimilarity deals with the measurement of divergence between two data items[9]. In fact, the performance of many algorithms depends upon selecting a good distance function over the input data set[9].

Here, a brief overview of similarity measure functions commonly used for clustering in literature is shown in the following subsections:

2.1.1 Euclidean distance

Euclidean distance is considered as the standard metric for geometrical problems. It is simply the ordinary distance between two points. Euclidean distance is extensively used in clustering problems, including clustering text. The default distance measure used with the K-means algorithm is also the Euclidean distance. The Euclidean distance determines the root of square differences between the coordinates of a pair of objects as shown in equation (1) given below [7].

$$Dist_{XY} = max_k \left| X_{ik} - X_{jk} \right| \quad (1)$$

2.1.2 Cosine distance

Cosine distance measure for clustering determines the cosine of the angle between two vectors given by the following formula[36]. Here θ gives the angle between two vectors and A, B are n-dimensional vectors.

$$\theta = arccos \frac{A.B}{\|A\|\|B\|} \quad (2)$$

2.1.3 Jaccard distance

The Jaccard distance measures the similarity of the two data items as the intersection divided by the union of the data items as shown in equation (3) given below [36]. The Jaccard similarity measure was also used for clustering ecological species[1].

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

2.1.4 Manhattan distance

Manhattan distance is a distance metric that calculates the absolute differences between coordinates of pair of data objects as shown in equation (4) given below[7]:

$$\boldsymbol{Dist_{XY} = \left| X_{ik} - X_{jk} \right|} \quad (4)$$

2.1.5 Chebyshev distance

Chebyshev distance is also called the maximum value distance. This distance metric calculates the absolute magnitude of the differences between coordinate of a pair of data objects as given in equation (5) given below[7]:

$$Dist_{XY} = max_k \left| X_{ik} - X_{jk} \right| \quad (5)$$

2.1.6 Minkowski distance

Minkowski Distance is also known as the generalized distance metric. In equation (6) given below[7], note that when p=2, the distance becomes the Euclidean distance. Chebyshev distance metric is a variant of Minkowski distance metric where p=∞ (taking a limit). This distance can be used for variables that are both ordinal and quantitative in nature.

$$Dist_{XY} = \left(\sum_{k=1}^{d} \left| X_{ik} - X_{jk} \right|^{\frac{1}{p}} \right)^{p} \quad (6)$$

2.1.7 Example Of Using A Distance Metric In Clustering

The K-means clustering algorithm makes use of the Euclidean distance as default distance metric to measure the similarities between the data objects:

Algorithm K-means using basic Euclidean distance metric

Let X = {x1,x2,x3, … ,xn} be the set of data objects and

Let V= {v1,v2, … ,vc} be the set of centers.

1. Randomly choose 'c' cluster centers.
2. Using the Euclidean distance metric, calculate the distance between each data object and cluster centers using equation (7) given below[7]:

$$Dist_{XY} = \sqrt{\sum_{k=1}^{m} \left(X_{ik} - X_{jk} \right)^2} \quad (7)$$

3. Assign data object to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4. Calculate new cluster center using equation (8) given below[7]:

$$v_i = \left(\frac{1}{c_i}\right)\sum_{1}^{c_i} x_i$$

5. The distance between each data object and new obtained cluster centers is recalculated.
6. Stop if no data object was reassigned, else repeat steps 3 to 5.

2.1.8 Limitations Of Distance Metrics In Clustering

Distance metrics are not always good enough when it comes to capturing correlations among the data objects. There is a high probability of the existence of similar data patterns among a set of data objects even if they are far apart from each other as measured by the distance metrics[2]. In order to find similarity between two data points, distance based metrics calculate only the physical distance between two data points and hence, are inadequate when it comes to capturing the behaviour of the data series. The behaviour of data series can be captured by association and disassociation between patterns of data points[2]. This can bring out the closeness between them.[2,3]

3. NEED OF CLUSTERING BASED ON SIMILARITY OF DATA PATTERNS

Sample data patterns are shown in Figures 2 and 3. In Figure 2, bottom two data objects are very close to each other distance-wise and all the data objects form a shifting pattern. In Figure 3, the data patterns form a scaling relationship.

Hence, the goal is to group not only data objects that are physically close together, but also all such kinds of data objects shown below in Fig 2 and Fig 3 having similar patterns[2,3]. This is the direction of this research.

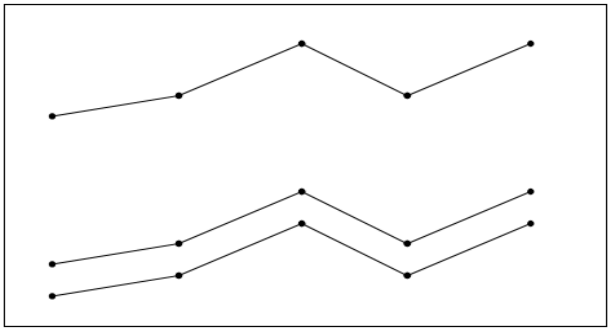

Fig 2: Data objects form a Shifting pattern

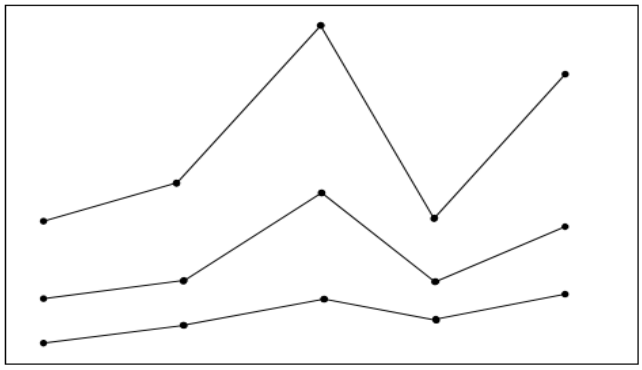

Fig 3: Data objects form a Scaling pattern

4. NEED OF AN EFFICIENT CLUSTERING TECHNIQUE

Along with the problem of incorporating distance based measures for clustering, many clustering techniques exhibit various problems such as having to specify the number of clusters at the start of clustering which increases overheads, and this decision of number of clusters is static, being incapable of handling outliers or noisy data, and many more.

Hence it is not only important to group data based on similar patterns but also data must be clustered in efficient manner eliminating above mentioned problems. This is one of the directions of this research.

5. "CONTEXT" OF DATA:[2]

In text data, terms in a document give meaning or context to a document. In other words, they establish context of the document. Many researchers tend to define 'context' based on the application. Context can also be the relevant terms associated with a document[2].

This concept of finding the "context" of data can be applied to non-textual data as well by using the attributes of the data to generate the context[2]. After the data has been clustered based on behaviour and not on physical closeness, these contexts can be generated.

Every generated cluster has a context associated with it[2]. These contexts can be generated for radar data like Ionosphere data from the UCI (University of California Irvine) Repository, and medical diagnosis data like Pima Indian Diabetes data also from UCI Repository and these contexts can be of use in analysis and decision making. This is one of the directions of this research.

6. CLUSTERING NEGATIVE DATA

Clustering negative data is a challenge. Less research is done in this area. Most of the distance functions and pattern similarity concepts for clustering are not robust enough to handle clustering negative data. They are only equipped to work for positive data like Pima Indian Diabetes data.

The Ionosphere data consists of a series of radar readings which have many negative values. This radar readings are categorized as good or bad depending upon the structures they form in the ionosphere.

Negative data can be:

1. Radar data such as Ionosphere
2. On requirement gathering from doctors, we found that Medical Diagnosis applications can work with negative and positive weights on a scale that can be assigned to observed readings such as Diabetes readings
3. Negative Term-Weighting schemes in text mining consist of assigning negative weights to a term that is absent instead of assigning value zero to that term.

7. LITERATURE SURVEY OF EXISTING CLUSTERING TECHNIQUES

The literature survey in this paper has identified the following major points such as the similarity measure being distance based measures or data pattern based similarity measures, the various inputs to the clustering algorithms such as number of

clusters, similarity matrix, the cluster threshold generation methods.

[2] proposes that the behaviour of data series can be captured by association and disassociation between patterns of data points which can reflect closeness between them which can be applied to find association between text documents. This work proposes a novel approach of document association based on probabilistic approach using context similarity coefficient (CSC). CSC has been used to find the context of textual data. However, CSC calculation fails to work for clustering negative data such as Ionosphere as CSC is not capable of handling negative data. CSC calculation needs modification in that case if context of Ionosphere data is to be found. pCluster[3] captures the similarity of the patterns exhibited by a cluster of objects in a subset of dimensions. pCluster uses a user defined clustering threshold.

[5] uses a distance metric for clustering high dimensional data based on the hitting time of two Minimal Spanning Trees (MST) grown sequentially from a pair of points by Prim's algorithm[5].

[6] proposed the similarity measurement method between

words by deploying Jaccard Coefficient which is a distance based measure.

[7] implemented the k-means algorithm using Euclidean,

Manhattan and Minkowski distance metrics and observed that the selection of distance metric plays an important role in clustering.

[21] identifies population structure from genetic data, using a similarity matrix to cluster which helps in efficient clustering.

[23] proposes modeling the entries of a given similarity matrix as the inner products of the cluster probabilities that are unknown.

[9] analyses the impact of the different distance similarity measures on Shared Nearest Neighbour (SNN) Approach. It is observed that Euclidean function works best with SNN clustering approach in contrast to Cosine, Jaccard distance measures function.

[32] gives a behavioral topic analysis approach to measure similarities between patient traces. A probabilistic graphical

model, i.e., latent Dirichlet allocation (LDA), is used to help discover treatment behaviours of patient traces[32].

[10] performs the clustering process in data streams and detects the outliers in data streams. Analysis of the clustering and outlier performance of BIRCH with CLARANS and BIRCH with K¬Means clustering algorithm for detecting outliers is done. The results show that BIRCH with CLARANS outperforms BIRCH with K-Means.

[30] Suggests that threshold can initially be set to a high value and then reduced till the whole dataset is covered.

[31] suggests a step-wise sequential threshold generation method.

7.1 Limitations And Findings From Existing Systems

After surveying and studying the existing systems, the following limitations and findings have been identified:

The similarity coefficients used in clustering are mostly distance based similarity metrics. Distance metrics are not always good enough when it comes to capturing correlations among the data objects. There is a high probability of the existence of similar data patterns among a set of data objects even if they are far apart from each other as measured by the distance metrics[2].

1. There is a need to decide certain clustering parameters such as number of clusters prior to the start of clustering, and this decision of number of clusters is generally static, i.e., clusters are not formed at runtime.
2. Most clustering techniques have the inherent problem of clusters being affected by outliers.
3. In pattern similarity based clustering, probabilistic based approaches are commonly used and efficiently capture behaviour of data.[2,32]
4. In pattern similarity based clustering, the data belonging to a particular class label can also have different clusters within the same class label.
5. Context Similarity Coefficient[2], an effective probabilistic based data pattern similarity approach for clustering fails to work for clustering negative data such as Ionosphere and therefore, contexts of those clusters cannot be generated.

8. CHALLENGES

The major challenge identified is clustering negative data. Very less research exists in this area. Most of the distance metrics and pattern similarity concepts for clustering are not robust enough to handle clustering negative data. They are only equipped to work for positive data like Pima Indian Diabetes data.

The Ionosphere data consists of a series of radar readings which have many negative values. This radar readings are categorized as good or bad depending upon the structures they form in the ionosphere.

9. CONCLUSION

After a thorough survey of existing systems, we aim to solve the above mentioned limitations in our system. This system proposes to combine the advantages of the reviewed systems. The improvements proposed in our system are as follows:

1. To work with pattern based similarity clustering using an efficient data association method called Context Similarity Coefficient because the distance based measures are not good enough when it comes to capturing the behaviour of data series.
2. To improve Context Similarity Coefficient based clustering by making it work for negative data so that radar data- Ionosphere can be clustered and
3. The context of this clustered Ionosphere data can be generated.
4. To implement Similarity Matrix Input based clustering.
5. To implement an efficient threshold generation mechanism that will help exclude outliers from clusters.
6. Combining above 2 points to help eliminate the need to decide number of clusters prior to clustering and also helps this number to not be static. The clusters are formed at runtime.

The challenges mainly in this regard are related to negative data clustering.

10. ACKNOWLEDGEMENTS

I would like to thank my guide, Professor Nitin Pise who has been very supportive and encouraging and has helped me at every stage of the research work.

I would also like to thank my co-guide, Professor Madhura Phatak for her timely guidance.

11. REFERENCES

[1] S.S. Choi, S.-H. Cha, C. Tappert, A survey of binary similarity and distance measures, Journal of Systematics, Cybernetics and Informatics 8 (1), 2010, 43-48.

[2] Kulkarni, A., Tokekar, V., Kulkarni, P.: Discovering context of labelled text documents using context similarity coefficient. Procedia Computer Science 49C(9),118-127 , Elsevier, 2015.

[3] Haixun Wang , Wei Wang , Jiong Yang , Philip S. Yu , Clustering by Pattern Similarity in Large Data Sets, Proceeding SIGMOD '02 Proceedings of the 2002 ACM SIGMOD international conference on Management of data, Pages 394-405, ACM.

[4] Reinforcement and systemic machine learning for decision making; vol. 1. John Wiley and Sons; 2012., IEEE Press.

[5] Laurent Galluccioa , Olivier Michelb, Pierre Comonb, Mark Kligerc, Alfred O. Herod, Clustering with a new distance measure based on a dual-rooted tree, Information Sciences Volume 251, 1 December 2013, Pages 96-113, Elsevier.

[6] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, Supacha-nun Wanapu, Using of Jaccard Coefficient for Keywords Similarity, Proceedings of the International MultiConference of Engineers and Computer Scientists 2013, Vol I, IMECS 2013, March 13 - 15, 2013, Hong Kong.

[7] Archana Singh, Avantika Yadav, Ajay Rana, K-means with Three different Distance Metrics, International Journal of Computer Applications, Volume 67, No.10, April 2013.

[8] Jian Pei , Xiaoling Zhang , Moonjung Cho , Haixun Wang , Yu, P.S. , MaPle:a fast algorithm for maximal pattern-based clustering, Data Mining, 2003. ICDM 2003. Third IEEE International Conference, Pages 259 - 266.

[9] Anil Kumar Patidar , Jitendra Agrawal , Nishchol Mishra, Analysis of Different Similarity Measure Functions and their Impacts on Shared Nearest Neighbor Clustering Approach, International Journal of Computer Applications, Volume 40, No.16, February 2012.

[10] S. Vijayarani and P. Jothi, "An Efficient Clustering Algorithm for Outlier Detection in Data Streams", International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, Issue 9, (2013) September, pp.3657-3665.

[11] Yadav, A.K. , Tomar, D. , Agarwal, S. , Clustering of lung cancer data using Foggy K-means, Recent Trends in Information Technology (ICRTIT), 2013 International Conference, Pages 13 - 18, IEEE.

[12] Yung-Shen Lin , Jung-Yi Jiang , Shie-Jue Lee , A Similarity Measure for Text Classification and Clustering, Knowledge and Data Engineering, IEEE Transactions (Volume:26 , Issue: 7) , Pages 1575 - 1590.

[13] Bollegala D. , Matsuo, Y. , Ishizuka, M. , A Web Search Engine-Based Approach to Measure Semantic Similarity between Words, Knowledge and Data Engineering, IEEE Transactions on (Volume:23 , Issue: 7) , Pages 977 - 990.

[14] Botsis T. , Scott, J. , Woo, E.J. , Ball, R. , Identifying Similar Cases in Document Networks Using Cross-Reference Structures, Biomedical and Health Informatics, IEEE Journal of (Volume:19 , Issue: 6), Pages 1906 - 1917.

[15] Fuyuan Cao , Jiye Liang , Deyu Li , Liang Baia , Chuangyin Dang , A dissimilarity measure for the k-Modes clustering algorithm, Knowledge-Based Systems, Volume 26, February 2012, Pages 120-127, Elsevier.

[16] Na Chen , Zeshui Xu , Meimei Xia , Correlation coefficients of hesitant fuzzy sets and their applications to clustering analysis, Applied Mathematical Modelling, Volume 37, Issue 4, 15 February 2013, Pages 2197-2211, Elsevier.

[17] Xianchao Zhang , Xiaotong Zhang , Han Liu , Multi-Task Multi-View Clustering for Non-Negative Data, Proceedings of the Twenty-Fourth International Joint Conference on Articial Intelligence, IJCAI 2015.

[18] Gabriella Casalino , Nicoletta Del Buono , Corrado Mencar , Subtractive clustering for seeding non-negative matrix factorizations, Information Sciences, Volume 257, 1 February 2014, Pages 369-387, Elsevier.

[19] Prachi Joshi , Mousami Munot , Parag Kulkarni , Madhuri Joshi , Efficient karyotyping of metaphase chromosomes using incremental learning, IET Science, Measurement and Technology, Volume 7, Issue 5, September 2013, p. 287-295.

[20] Abhishek Kumar , Hal Daume , A Co-training Approach for Multi-view Spectral Clustering, Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 2011.

[21] Daniel John Lawson , Daniel Falush , Similarity matrices and clustering algorithms for population identification using genetic data, Department of Mathematics, University of Bristol, Bristol, BS8 1TW, UK, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig Germany, March, 2012.

[22] Wen-Yen Chen , Yangqiu Song , Hongjie Bai , Chih-Jen Lin , Edward Y.Chang , Parallel Spectral Clustering in Distributed Systems, Pattern Analysis and Machine Intelligence, IEEE Transactions on (Volume:33 , Issue: 3), 2011, Pages 568-586.

[23] Raman Arora , Maya R. Gupta , Amol Kapila , Maryam Fazel , Similarity-based Clustering by Left-Stochastic Matrix Factorization, Journal of Machine Learning Research 14 (2013) 1715-1746.

[24] D. Kuang, C. Ding, and H. Park. Symmetric nonnegative matrix factorization for graph clustering. In Proc. SIAM Data Mining Conf, 2012.

[25] Cluster analysis: a survey by BS Duran, PL Odell - 2013.

[26] Brian Eriksson , Gautam Dasarathy , Aarti Singh , Robert Nowak , Active Clustering: Robust and Efficient Hierarchical Clustering using Adaptively Selected Similarities, Arxiv preprint arXiv:1102.3887, 2011.

[27] Alina Ene , Sungjin Im , Benjamin Moseley , Fast clustering using MapReduce, Proceeding KDD 2011 Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 681-689, ACM.

[28] Nir Ailon , Yudong Chen , Huan Xu , Iterative and Active Graph Clustering Using Trace Norm Minimization Without Cluster Size Constraints, Journal of Machine Learning Research 16, 2015, Pages 455-490.

[29] HwanjoCheng, Yizong, Church, George M., 2000. Biclustering of expression data. In: Proc. Eighth Internat. Conf. on Intelligent Systems for Molecular Biology, AAAI Press, pp. 93-103.

[30] Hwanjo Yu , Duane Searsmith , Xiaolei Li , Jiawei Han , Scalable Construction of Topic Directory with Nonparametric Closed Termset Mining, Data Mining, 2004. ICDM '04. Fourth IEEE International Conference, Pages 563-566.

[31] Stutz WE, Bolnick DI. (2014). Stepwise threshold clustering: a new method for genotyping MHC Loci using next-generation sequencing technology, PLoS One 9:e100587.

[32] Zhengxing Huang , Zhejiang Univ. , Hangzhou China , Wei Dong , Hui-long Duan , Haomin Li , Similarity Measure Between Patient Traces for Clinical Pathway Analysis: Problem, Method, and Applications, Biomedical and Health Informatics, IEEE Journal of (Volume:18 , Issue: 1), Pages 4-14.

[33] Agrawal R., Faloutsos C., Swami A. Efficient similarity search in sequence databases. Proc. 4 The Int. Conf. On Foundations of Data Organizations and Algorithms, 1993. – Chicago. pp. 69-84.

[34] Fast Distance Metric Based Data Mining Techniques Using P-trees: k-Nearest-Neighbor classification and k-Clustering : A Thesis Submitted to the Graduate Faculty Of the North Dakota State University.

[35] Joaquin Perez Ortega, Ma. Del Rocio Boone Rojas and Maria J. Somodevilla Garcia. Research issues on K-means Algorithm: An Experimental Trial Using Matlab.

[36] Shraddha Pandit, Suchita Gupta, A Comparative Study On Distance Measuring Approaches For Clustering, International Journal of Research in Computer Science eISSN 2249-8265, Volume 2 Issue 1 2011, pp. 29-31.