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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 11, November 2014

Analysis of Inter Cluster Migration Estimation (ICME) model in multidimensional clusters

A.M.Rajee[1], F.Sagayaraj Francis[2]

Research Scholar, Department of CSE, Pondicherry Engineering College, Puducherry, India [1]

Professor, Department of CSE, Pondicherry Engineering College, Puducherry, India [2]

Abstract: Clustering evolves as an indigenous unsupervised data mining problem. This paper presents an estimation model, when new unclustered information is fed to the clustered system. The soul of this paper is to test the accuracy of the built Inter Cluster Movement Estimation (ICME) model with multi-dimensional clusters. Clusters of varying sizes and dimensions were constructed from synthetic and real data sets taken from UCI repository. On experimental analysis, the accuracy of the approximation model is found to increase with increased cluster sizes of multiple dimensions.

Keywords: Cluster parameters, Estimation model, Inter cluster movement, Multi-dimensional data set

I. INTRODUCTION

Clustering is an important data mining technique which groups objects with similar characteristics. A cluster is, therefore, a group of data objects similar between them and dissimilar to the objects belonging to other clusters. Cluster analysis is widely used in many areas in the current era especially in market research, data analysis, pattern recognition and image processing [1,2].

A clustering system is a collection of similar objects (or synonymously data points, instances, observations, elements) which is grouped to K clusters N: $\{C_1, C_2,\ldots, C_K\}$. The Euclidean distance between objects say xi and xj is denoted as d (x_i,x_j). The cluster Ci is closest (or synonymously nearest) to cluster C_j if the Euclidean distance between their centers is smallest among all combinations of the cluster pairs in the clustering system. We therefore refer to C_i as closest to C_j.

The scope of this paper relies on the new data point which is induced into already built clustering system. Feeding the system with a new point makes a re-arrangement of the individual clusters, causing other data points to move between the clusters. This process is known as inter cluster movement (or synonymously migration) of data points [3, 4]. This paper gives a brief introduction of Inter Cluster Migration Estimation (ICME) model followed by the experimental setup and analysis of ICME model in multidimensional clusters of varying sizes.

II. RELATED WORKS

Very few research works were made on handling the incoming data point and its impact on the clustering system [5, 6]. A.M. Sowjanya and M. Shashi developed a method for clustering the incoming data point by finding the farthest neighbor point that is nearer to the incoming data point. Mixed data sets were used for testing the cluster efficiency [7]. O. Georgieva, F. Klawonn proposed an approach to assign the new input data stream to the already known data structure or discovers new interesting groups of the data set that currently appeared. Two algorithm variants – hard and fuzzy – are presented in parallel [8]. Charu C. Aggarwal, Philip S. Yu presented an online approach for clustering massive text and categorical data streams. A time-sensitive weightage was assigned to each data point. An entry was maintained to keep track of the last time; a point was added to the cluster [9]. Angie King suggested a discounted center updating rule as a modification of the updating rule proposed by Lloyd. This proposed exponential smoothing heuristic algorithm works for the naturally cluster- able data and for the cluster centers, which are moving over time [10].

Seokkyung Chung and Dennis McLeod performed incremental clustering from web articles (documents) that change over time. The proposed algorithm incrementally clusters documents based on neighborhood search and computes their similarity. The re-clustering was effected by merging the documents to a singleton cluster [11]. Si-Bao Chen, Hai-Xian Wang, Bin Luo introduced dynamic weighting of data by interactively updating centers. Alpha-mean operator is used in the proposed K Alpha Means algorithm and the performance is found to be increased [4]. On dynamic data clustering and visualization using swarm intelligence proposed Dynamic-FClust algorithm which adds new data records, resulting in the change of clustering over time, leading to the need to mine dynamic clusters[12].

III.INTER CLUSTER MOVEMENT ESTIMATION (ICME) MODEL

New unclustered information is fed to the existing clustering system [7, 11]. The new element will become member of its nearest cluster and will either facilitate the movement of other data points between clusters, upsetting the clustering setup or becoming a member of the nearest cluster, without influencing inter cluster movement [3]. This paper is inclined to study the behavior of the clustering system considering the former case.

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 11, November 2014

To reduce the repeated re-clustering procedure, an Inter Cluster Movement Estimation (ICME) model was built, which will predict the occurrence of inter cluster movement, thereby letting the user to decide on the re-execution of the clustering system. Cluster parameters like cluster size and shape, SSE, cluster separation measures get altered with the induction of new data object [10, 13, 14]. The distance of the incoming data object to its closest cluster center is interdependent on various cluster parameters. The correlation relationship of the new point with the cluster parameters is shown in Table I. The ICME model was built by thoroughly studying the dependence of cluster parameters on the new entrée [3].

TABLE I
CORRELATION VALUES

Cluster Parameters	Correlation values
Cluster Separation	Positive
SSE	Negative
Distance between two closest cluster centers	Positive
Number of Clusters	Negative

Studies have shown that there exists a linear dependency of all the cluster parameters with the position of the new point from its nearest cluster center. Some parameters like cluster separation and distance between two cluster centers are positively correlated with the new point, while parameters like SSE and number of clusters seems to acquire a negative dependency. This correlation study plays a major impact for the construction of Inter Cluster Movement Estimation (ICME) model. The ICME model was thus expressed as a function of Cluster Separation (CS), Distance between the center of two closest clusters d (Ci,Cj) and Sum of Squared Error (SSE).

$$D \propto \frac{CS * d(C_i, C_j)}{SSE} \quad (1)$$

Introducing a proportionality constant 'migrator'

$$D = \text{migrator} * \frac{CS * d(C_i, C_j)}{SSE} \quad (2)$$

Assigning a value to the 'migrator' constant depends on the inter dependence of cluster separation, SSE, distance between two closest clusters d (C_i,C_j) and the cluster size. The ICME model will estimate the minimum distance D, that is, whether a new point, when positioned at the distance D will facilitate the movement of data between the clusters or not. This distance D is the Euclidean distance between the new point and the center of its nearest cluster.

IV. EXPERIMENTAL SETUP

A. Synthetic Data sets

Synthetic data sets of varying instances and multiple dimensions were generated to test the effectiveness of the estimation model [8, 15]. A new point is fed to the existing clustering system. Table II gives the distance of the new point causing inter cluster movement for varying instances of two dimensional datasets built from two clusters.

TABLE II
COMPARISON OF ICME VALUES WITH OBSERVED RESULTS FOR SYNTHETIC DATA SETS

Number of Instances	Distance causing inter cluster movement-experimental values	Inter Cluster Movement Estimator D	Relative Error (in %)
1024	191.2	186.7009	2.35
2000	182.5	178.2933	2.31
573	174.4	170.3847	2.30
976	167	163.0461	2.37
3918	153.5	150.3995	2.02
2347	148.1	145.2641	1.91
132	143.3	141.0378	1.58

The values in column 2 show the observed experimental values which caused movement of data points between clusters. Column 3 presents the proposed estimation values. The lower relative error in column 4 depicts the efficiency of this model, explaining the adequacy for its implementation in the clustered system. The ICME model was found to portray improved accuracy, with increasing cluster size and dimension. Fig.1 shows the results with different number of three dimensional clusters. It is obvious from figure 1 that the relative error of the ICME model dips down with the raise in the cluster size.

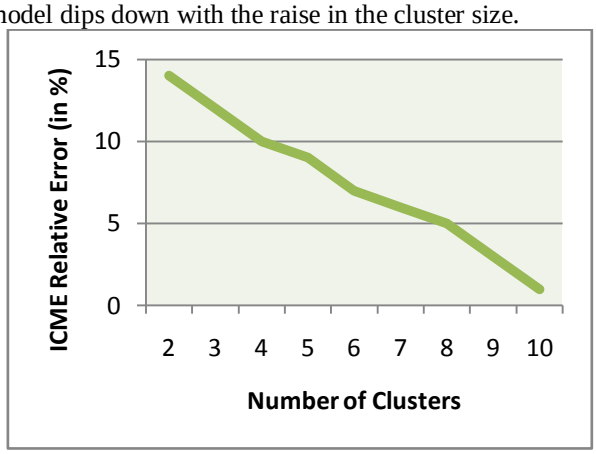


Fig. 1. Relative error of ICME with multiple clusters

The ability of ICME model was tested with multi-dimensional data sets built from four clusters varying the dimensions from two to ten. Fig.2 presents the results.

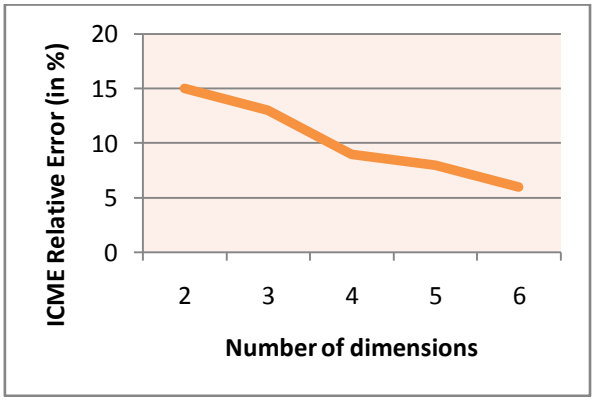


Fig. 2. Relative error of ICME with multiple dimensions

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 11, November 2014

From fig.2, it is obvious that the accuracy of the model increases (decreased error %) with increased dimensionality of the clusters.

B. Synthetic Data sets

Real data sets from UCI Machine Learning repository [16] was taken to check the reliability of ICME model on multiple attributes of the data sets. Iris and Breast Cancer data sets were used for this purpose.

- Iris

Iris data set consists of 150 instances grouped into three equal sized clusters. Each cluster contains 50 data points. The iris data set is represented by three attributes Setosa, Versicolor and Virginica.

- Breast Cancer

Wisconsin Breast Cancer data set consisting of 684 data points was used. The number of clusters is 2 with 9 attributes.

- Newthyroid

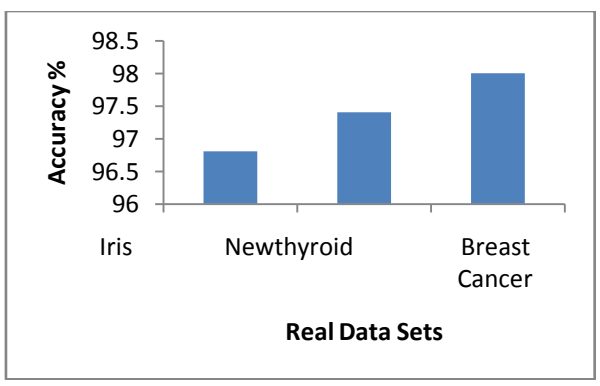
Three clusters were built from a total of 215 data points. The total number of attributes is 5.

Experiments were conducted with Iris and Breast cancer data sets. A new data point was introduced and the d distance of the new point effecting inter cluster movement was estimated. Observed values from the experiments were also recorded as shown in Table III. It is evident that ICME values tend to achieve relatively higher accuracy (lower relative error %) with Iris and Breast cancer and Newthyroid data sets.

TABLE II
COMPARISON OF ICME VALUES WITH REAL DATA SETS

Data set	No. of Instances	No. of dimensions	Accuracy (in %)
Iris	150	3	96.8
Breast Cancer	684	9	98
Newthyroid	215	5	97.4

It is also obvious from table III that even in real data sets, as the number of instances and the attributes increases, there is a linear increase in the accuracy of the ICME model. Fig.3 reflects the findings in Table III.



V. CONCLUSION

This paper presented an approximation model for predicting the inter cluster movement, when a new data point was nurtured to the system. Successful prediction will result in avoiding unnecessary repeated execution of the system. Both synthetic and real data sets from UCI repository was taken to construct clusters with varying size and dimensions. On experimental analysis, it is found that the estimation model is in concurrence with the experimental setup with relatively lower error rate. Also the ICME model is more accurate with increasing cluster size and works well on multidimensional synthetic and real data clusters.

REFERENCES

[1] S.Lloyd, "Least squares quantization in PCM", IEEE Transactions on Information Theory, pp.129-136, 1982.
[2] J. Han and M.Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, 2001.
[3] A.M.Rajee and F.Sagayaraj Francis, "Inter Cluster Movement Estimation model based on cluster parameters", in Proc. IEEE International Conference on Computational Intelligence and Computing Research", 2013, pp.369-372.
[4] Si-Bao Chen, Hai-Xian Wang, Bin Luo, "On Dynamic Weighting of Data in Clustering with K-Alpha Means", International Conference on Pattern Recognition, 2010
[5] A. Campan and G. Serban, "Adaptive Clustering algorithms", Advances in Artificial Intelligence, 2006, Springer.
[6] G.Serban and A.Campan, "Adaptive Clustering using a Core-based Approach", Informatica, Volume L, Number 2, 2005.
[7] A.M. Sowjanya and M. Shashi, "A Cluster Feature-Based Incremental Clustering Approach to Mixed Data", Journal of Computer Science, 2011
[8] O. Georgieva, F. Klawonn, "Dynamic data assigning assessment clustering of streaming data", Journal of Applied Soft Computing, Volume 8 Issue 4, September, 2008
[9] Charu C. Aggarwal, Philip S. Yu, "A Framework for Clustering Massive Text and Categorical Data Streams", In Proc. ACM SIAM Data Mining Conference , 2006
[10] Angie King, "Online k-Means Clustering of Non-stationary Data", Prediction Project Report, 2012
[11] Seokkyung Chung and Dennis McLeod, "Dynamic Pattern Mining: An Incremental Data Clustering Approach", Journal on Data Semantics, Volume 2, 2005
[12] E.Saka, "On dynamic data clustering and visualization using swarm intelligence", In Proc. 26th International conference on Data Engineering Workshops, pp.337-340, 2010.
[13] Olatz Arbelaitz, Ibai Gurrutxaga et al, "An extensive comparative study of cluster validity indices", Pattern Recognition, Elsevier, vol.46, pp. 243–256, 2013.
[14] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn, vol. 2, pp. 740–741, August 1987
[15] M.Young, "The Technical Writer's Handbook", Mill Valley, CA: University Science, 1989
[16] Asuncion, A., Newman, D. J.,"UCI Machine Learning Repository", University of California, School of Information and Computer Science, 2007.