Optimal Number of Clusters

E. Gökçay[1], M. Karakaya[2], G. Şengül[3]

[1]Software Engineering Department, Atilim University,

Street, Incek, Ankara, TURKEY

[2,3]Computer Engineering Department, Atilim University,

Street, Incek, Ankara, TURKEY

[1]erhan.gokcay@atilim.edu.tr, [2]murat.karakaya@atilim.edu.tr, [3]gokhan.sengul@atilim.edu.tr

ABSTRACT

Clustering is one of the important analysis methods in Data Analytics and Pattern Recognition. Clustering divides the data into groups without any supervision or external labels and it is a subjective analysis as the definition of a cluster is context dependent. Because of this reason many algorithms, like kNN, require the number of clusters to be fixed apriori. Each clustering algorithm depends on a distance metric to identify different groups in the data. Once the number of centers are fixed, each algorithm tries to find the best separation according of its distance measure by using an optimization algorithm. The distance metric determines the shape of the clusters generated. There are algorithms, like Ward, to determine the number of clusters in a data set and these algorithms also depend on the same distance metrics where many metrics, like Euclidean and its derivatives, generate hyper ellipsoidal clusters and fail in nonlinearly clustered data. Another computationally expensive approach is to run a specific algorithm for different number of cluster centers and try to choose the best number. In this paper, we attempt to analyze the number of clusters using a previously developed Information Theoretical metric called CEF which; in its original use; can separate nonlinear clusters. Data points that are close to each other are incrementally joined together using a distance measure to create subclusters like joined data points against the rest of the data. The operation continues until all data elements are consumed. The CEF metric is used to measure the distance between obtained clusters where peaks in the measure indicates strong cluster separation. The method is tested in several artificial and real data sets and its advantages and disadvantages are discussed.

Keywords – clustering, distance metric, information theory

1. INTRODUCTION

Clustering is an unsupervised pattern classification method that partitions the input space into clusters. The goal of a clustering algorithm is to perform a partition where objects within a cluster are similar and objects in different clusters are dissimilar. Therefore, the purpose of clustering is to identify natural structures in a dataset [1–5] and it is widely used in many fields such as psychology [6], biology [4], pattern recognition [3], image processing [7] and computer security [8].

Estimation of the number of clusters in a data set is an important theoretical and practical problem in cluster analysis. With too few clusters, one cannot preserve the most relevant information about the structure of the dataset X. On the other hand, with too many clusters, resources are wasted by processing non-relevant data. Many clustering algorithms are not able to determine the number of natural clusters in the data, and therefore they must initially be supplied with this information—frequently known as the k parameter. Since this information is rarely previously known, the usual approach is to run the algorithm several times with a different k value for each run. Then, all the partitions are evaluated and the partition that best fits the data is selected. The process of estimating how well a partition fits the structure underlying the data is known as cluster validation [1].

The algorithm is run over the dataset with a set of m different values for the k parameter, K = {k1, k2, . . . , km}. In this way, a set of m partitions is obtained, S = {P1, P2, . . . , Pm}, but just one of them has partitioned the data with the correct number of clusters.

Cluster validity indices are usually based on external criteria, internal criteria and relative criteria [9]. For example, some indices are based on considering the compactness within each cluster or the separation between clusters [10–13], some are based on the information entropy or based on measuring in-group proportion of objects [14,20,21]. Yu and Cheng [15] study the range where the optimal number of clusters between the minimal number and the maximal number in FCM clustering algorithms. Some of CVI methods are combined with FCM (fuzzy C-mean) algorithms [16–18]. Chen et al. [19] propose a hierarchical method, COPS (clusters optimization on preprocessing stage) method, to determine the number of clusters by the two-steps processing.

Generally speaking, most of the discussed cluster validity checking methods are based on the following process to obtain the optimal number of clusters. A typical process for determining the number of clusters in a data set is: firstly, different partitions, also be called as cluster results or clustering schemes, are obtained through a certain clustering algorithm in which the different parameters used usually such as the number of clusters; then the criteria count on the different partitions; the parameter meets the predetermine conditions is selected to be the optimal number of clusters after analyzing the size or the variability of the different criteria.

In Chapter 2, the distance function used in the calculation will be described. Chapter 3 describes the new method to analyze the number of clusters. Chapter 4 gives results using artificial and real datasets. The discussion of results are given in Chapter 5 and conclusion is given in Chapter 6.

2. INFORMATION THEORY AND CLUSTERING

Clustering is an unsupervised method to separate data into groups by using a certain metric or distance function in such a way that the groups or clusters are more similar within, compared to other grouping choices. Similar to edge detection, clustering is also a subjective division depending on the distance measure. Some distance measures can only divide the data using a linear boundary where others can perform a nonlinear separation. Even the human eye will cluster the same data differently depending on other conditions and context.

Information theory has been used as a distance measure successfully in one of the author's previous paper [5]. Using an information theoretic measure, nonlinear regions can be separated without any supervision. The derivation of the distance measure or CEF (Cluster Evaluation Function) was given elsewhere [5] and it will not be repeated here. Only the final formulation will be used. In (1), p and q are clusters of size N_p and N_q, respectively where $x_i \in p$ and $x_j \in q$. The Gaussian kernel needs a parameter σ for the kernel size. The proper value of this parameter is important in clustering which needs to be determined.

$$CEF(p,q,\sigma) = \frac{1}{N_p N_q} \sum_{i=1}^{N_p} \sum_{j=1}^{N_q} G\left(x_i - x_j, 2\sigma^2\right) \quad (1)$$

A feature vector needs to be extracted from the data to be clustered. The feature set can be any original and/or transformed subset of the data which provides a good description of data. Using the feature set as an input to the distance measure, labels are assigned to each cluster such that the distance between different labels (clusters) has the maximum value. Searching the optimal labeling may need an exhaustive search and some heuristic algorithms are applied to limit the search space. As explained in [5], the final result is very promising and a nonlinear separation of clusters is possible with this method without any supervision as shown in Fig 1. It should be noted that the distance measured by CEF is inversely proportional to the distance between clusters.

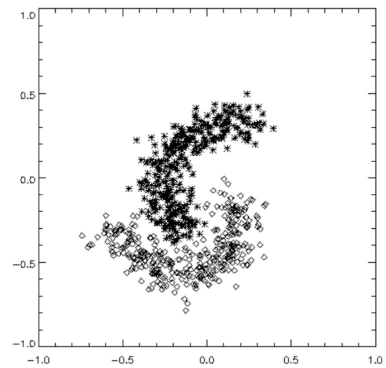

Figure 1: Dataset labels after an unsupervised clustering using CEF function [5]

3. PROPOSED CLUSTER SEPARATION INDEX

In this section, the proposed information theoretic cluster separation method is presented. The motivation of the proposed method is to process the data once and obtain information about possible number of clusters in the data. The distance function (CEF) which is described earlier will be used in the calculation.

3.1 Subclusters and Grouping

The calculations start by forming two cluster sets, i.e. *p* and *q*. One cluster is formed initially by finding two closest point to each other in the dataset using a Euclidean distance measure. Since we are not measuring any cluster separation during this procedure, any distance metric can be used. The second cluster is formed by the rest of the data except the points in the first cluster. At any time, the points belong either to the first cluster or to the second cluster. After the initialization of two clusters, the procedure continues by finding the closest point *x* in cluster *q* to any point in cluster *p* one at a time as shown in Fig. 2. The points x_1 and x_2 are the closest two points in the data set. The first group is formed by putting these points together as cluster *p*. The rest of the data belongs to cluster *q*. During the next step, the closest point is found in cluster *q* to any member of cluster *p*. The point x_3 is the closest point to the group formed by x_1 and x_2. The next stage is to move the point x_3 from cluster *q* to the group (i.e. to cluster *p*). The cluster p grows one at a time by finding the closest point in cluster *q*.

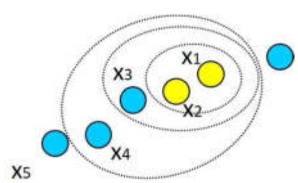

Figure 2: Group forming by finding the closest point to any group members

3.2 Cluster separation

During the iteration, cluster *p* will increase and cluster *q* will decrease one point at a time. After a new point is removed from cluster q and added to cluster p, the distance between these two clusters are measured using CEF distance function. One instance is shown in Fig. 3.

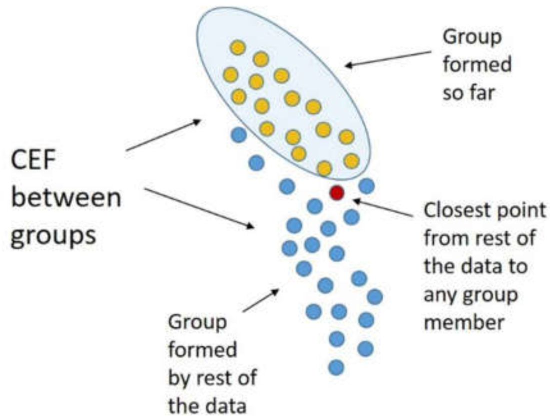

Figure 3: Two cluster groups are formed and CEF between these groups is calculated

Once each point is processed, the array obtained by calculating CEF distance between clusters at every step is analyzed to determine the possible clusters in the dataset. Low peaks in the plot indicate a strong indication of cluster separation. The algorithm is given in Fig. 4.

Number_of_Clusters Algorithm

INPUT:dataset
OUTPUT:CEFarray

```
Cluster_p = []
Cluster_q = [dataset]
[x1,x2] = find_closest_2_points(dataset)
Cluster_p = Cluster_p + [x1,x2]
Cluster_q = Cluster_q – [x1,x2]
CEFarray = size(dataset)
CEFarr(1) = = CEF ( Cluster_p, Cluster_q)
for n=2 to size(dataset)
  x = find_closest_point(Cluster_p, Cluster_q)
  Cluster_p = Cluster_p + x
  Cluster_q = Cluster_q – x
  CEFarray(n) = CEF ( Cluster_p, Cluster_q)
endfor
plot CEFarray
find peaks(CEFarray)
```

Figure 4: Number of Clusters Algorithm

4. EXPERIMENTAL RESULTS

The algorithm is tested using several artificial and real datasets and the results are discussed.

4.1 Artificial datasets

An example dataset is shown in Fig. 5.

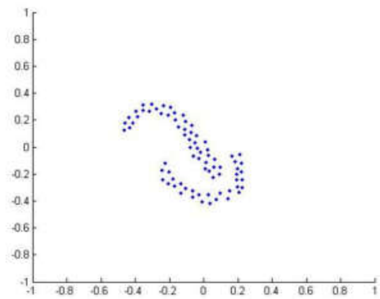

Figure 5: 2-dimensional dataset with 2 main clusters

The plot of CEF distance function is shown in Fig. 6. The horizontal axis consists of all points in the dataset and vertical axis is the log(CEF) value between clusters formed at every step during the iteration. The logarithm is displayed to enhance the small values in the calculation.

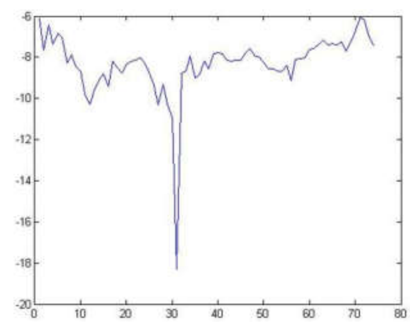

Figure 6: log(CEF) plot of the dataset

The plot indicates a very strong low value at point 31 which indicates that the data has 2 main clusters in it. The other less small peaks indicate other sub clusters and the importance of these clusters is context dependent. The cluster p and cluster q at point 31 is shown in Fig. 7.

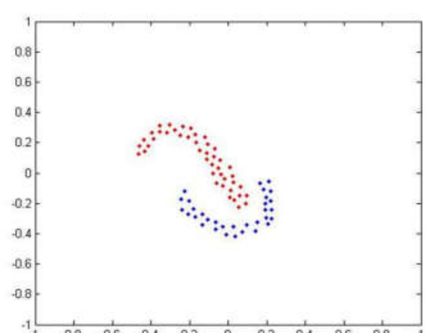

Figure 7: Cluster separation at lowest peak of CEF plot

When a different distance metric is used, the separation point is wrong unlike the point in CEF calculation. When we use a Euclidean distance the plot is given in Fig 8.

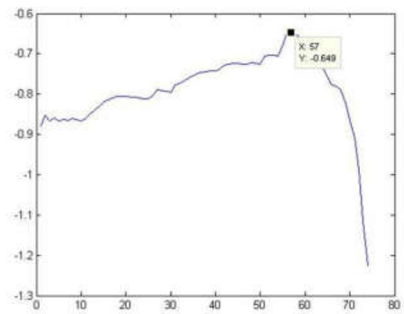

Figure 8: Euclidean Distance plot

The results for other artificial datasets are shown in Fig. 8. to Fig. 12.

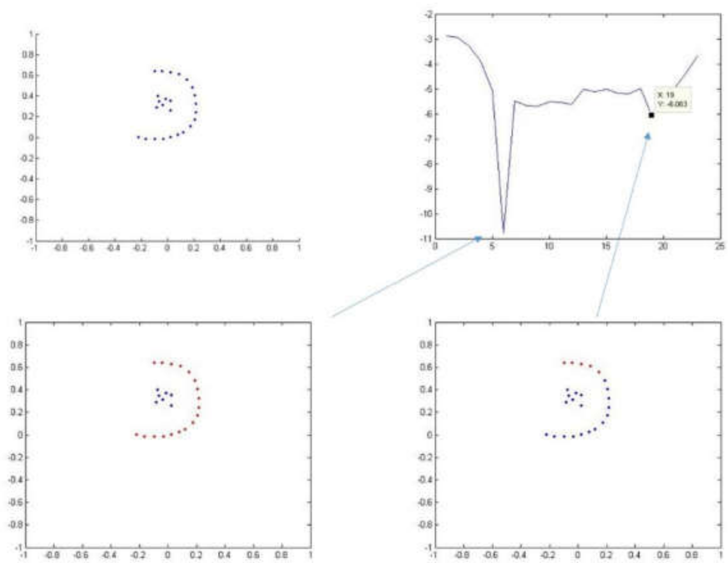

Figure 9: Results using a dataset with 2 clusters

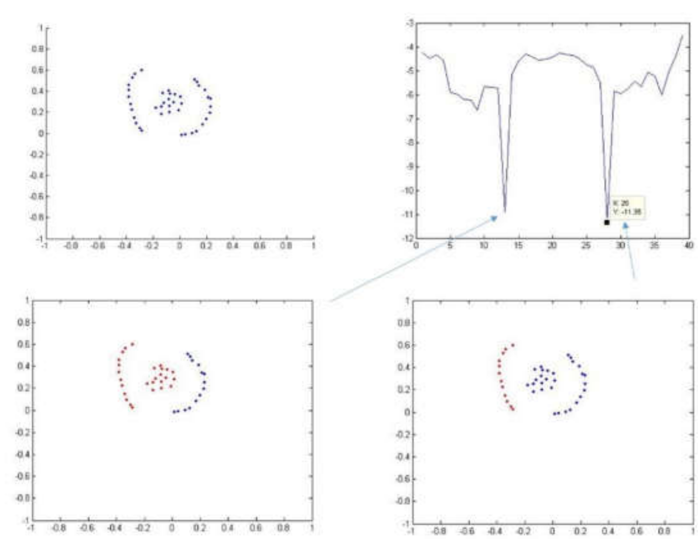

Figure 10: A dataset with 3 different main clusters

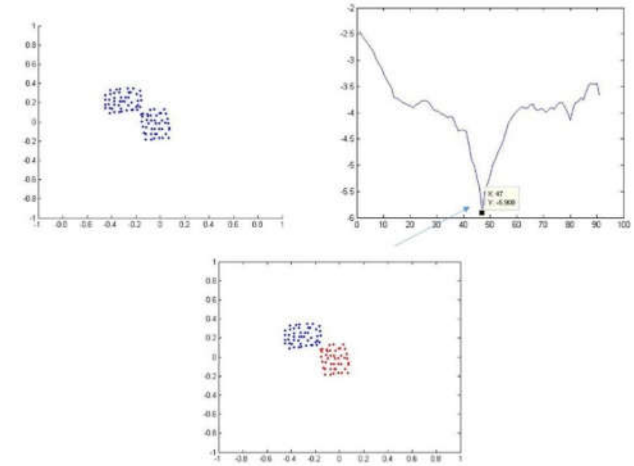

Figure 11: A different dataset with 2 main clusters

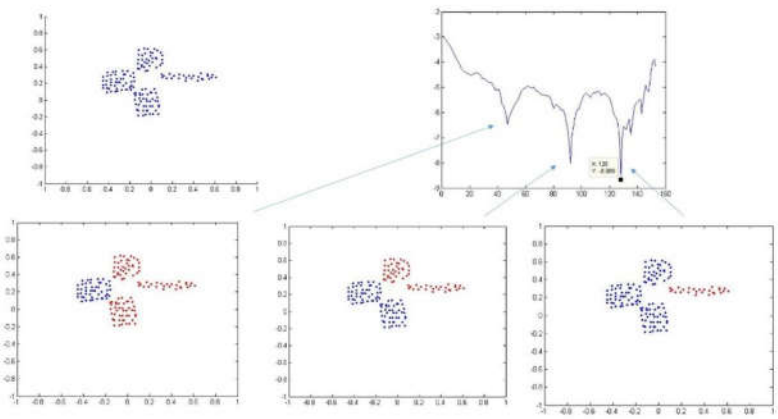

Figure 12: A dataset with 4 different main clusters

4.2 Real datasets

The algorithm is tested using Fisher's Iris data set. The Iris data set is 4 dimensional and it is not possible to plot all dimensions together. It is known that two species have overlapping features. Features 1 and 2 and the CEF plot are shown in Fig. 13. using different color and shapes for different species where the overlap can be seen very clearly. The CEF plot shows the first cluster but fails to show the second one.

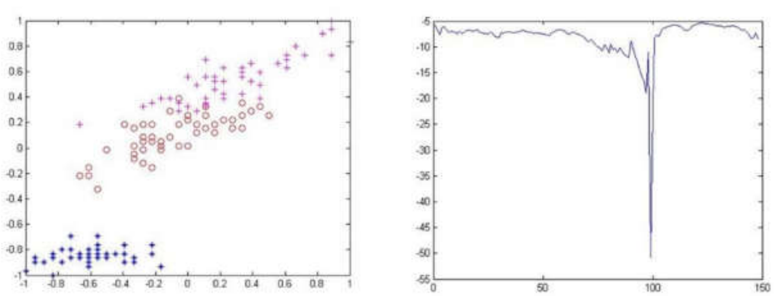

Figure 13: Iris data set and CEF plot

The second dataset is from UCI Machine Learning Repository [22]. The Banknote Authentication dataset has 4 dimensional feature set with two classes. When we plot different combinations of features, we can see that the classes overlap heavily. Also sub clusters are visible in the dataset as in Fig. 14.

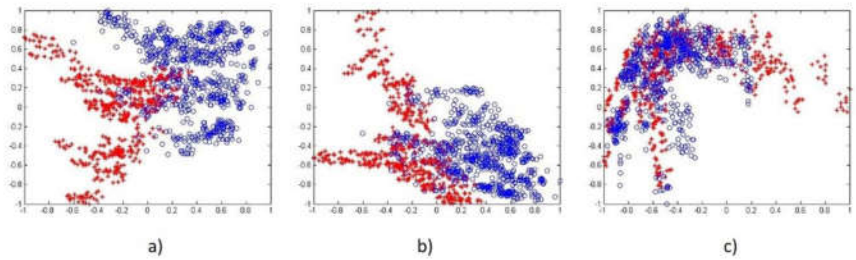

Figure 14: Banknote dataset a) features 1,2 b) features 1,3 c) features 3,4

The CEF plot shows noisy output at borders where one cluster has most of the points and the other cluster is almost empty. Ignoring the end points, the plot shows a strong indication for 2 clusters as shown in Fig 15. But it also shows other major cluster separations as well which can be seen in Fig. 13.

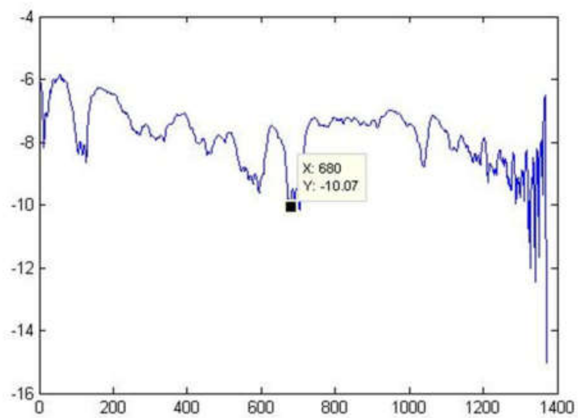

Figure 15: Banknote CEF plot showing 2 main clusters and smaller sub clusters

5. DISCUSSION

When the clusters are well separated even highly nonlinearly (nonconvex), the CEF Index has not trouble identifying strong cluster separation points. The calculation requires only one pass through the data points and it is possible to obtain a good indication about the number of clusters. On the other hand, overlapping features are reducing the separation since no assumptions are done about any model or distribution. Another concern is that the initial and ending conditions of the plot shows noisy output because of the reduction of points in one cluster. It is wise to assume that outliers will create a similar problem. Extra calculations are needed to compensate the reduction.

6. CONCLUSIONS

In this paper an algorithm is proposed to find information about the possible number of clusters in a dataset. The present algorithms are trying to find the number of clusters by running a given clustering algorithm by changing the cluster number from 2 up to the maximum number of points in the dataset. During the calculation different metrics are used to measure the validity of the clusters generated. Running the algorithms repeatedly by changing the assumed number of clusters is a time consuming task. The proposed algorithm is going through the dataset only once without assuming any number of clusters. The nonlinearly weighted CEF distance functions shows a strong separation at cluster boundaries and by counting the minimum peaks, it is possible to get a very reliable information about the possible number of clusters in any dataset.

6. REFERENCES

[1] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "On clustering validation techniques", Journal of Intelligent Information Systems 17 (2001) 107–145.

[2] A.K. Jain, R.C. Dubes, "Algorithms for Clustering Data", Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[3] B. Mirkin, "Clustering for Data Mining: A Data Recovery Approach", Chapman & Hall/CRC, Boca Raton, Florida, 2005.

[4] P.H.A. Sneath, R.R. Sokal, "Numerical Taxonomy, Books in Biology", W.H. Freeman and Company, San Francisco, 1973.

[5] Gokcay, E.; Principe, J.C., "Information theoretic clustering," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.24, no.2, pp.158,171, Feb 2002.

[6] K.J. Holzinger, H.H. Harman," Factor Analysis", University of Chicago Press, Chicago, 1941.

[7] C.-H. Chou, M.-C. Su, E. Lai, "A new cluster validity measure and its application to image compression", Pattern Analysis and Applications 7 (2004) 205–220.

[8] D. Barbara´ , S. Jajodia (Eds.), "Applications of Data Mining in Computer Security", Kluwer Academic Publishers, Norwell, Massachusetts, 2002.

[9] M. Halkidi, Y. Batistakls, M. Vazirgiannis, Clustering validity checking methods: part ii, in: ACM SIGMOD Record, vol. 31, 2002, pp. 19–27.

[10] M. Bouguessa, S.Wang, H. Sun, An objective approach to cluster validation, Pattern Recognition Letters 27 (13) (2006) 1419–1430.

[11] X.L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Trans. On Pattern Analysis and Machine Intelligence 13 (8) (1991) 841–847.

[12] J.C. Dunn,Well separated clusters and optimal fuzzy partitions, Journal of Cybernetics 4 (1974) 95–104.

[13] P. Lingras, M. Chen, D. Miao, Rough cluster quality index based on decision theory, IEEE Transactions on Knowledge and Data Engineering 21 (7) (2009) 1014–1026.

[14] J.Y. Liang, X.W. Zhao, D.Y. Li, F.Y. Cao, C.Y. Dang, Determining the number of clusters using information entropy for mixed data, Pattern Recognition 45 (2012) 2251–2265.

[15] J. Yu, Q.S. Cheng, The range of the optimal number of clusters for the fuzzy clustering algorithms, Science in China (Series E) 32 (2) (2002) 274–280.

[16] S.F. Bahght, S. Aljahdali, E.A. Zanaty, A.S. Ghiduk, A. Afifi, A new validity index for fuzzy c-means for automatic medical image clustering, International Journal of Computer Applications 38 (12) (2012) 1–8.

[17] Y.J. Zhang,W.N.Wang, X.N. Zhang, Y. Li, A cluster validity index for fuzzy clustering, Information Sciences 178 (4) (2008) 1205–1218.

[18] H.J. Sun, S.R.Wang, Q.S. Jiang, Fcm-basedmodel selection algorithms for determining the number of cluster, Pattern Recognition 37 (10) (2004) 2027–2037.

[19] L.F. Chen, Q.S. Jiang, S.R.Wang, A hierarchical method for determining the number of clusters, Journal of Software 19 (2008) 62–72.

[20] A.V. Kapp, R. Tibshirani, Are clusters found in one dataset present in another dataset?, Biostatistics 8 (1) (2007) 9–31.

[21] S. Still,W. Bialek, How many clusters? An information-theoretic perspective, Neural Computation 16 (12) (2004) 2483–2506.

[22] Lichman, M.UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2013