Kuwait J. Sci. **44** (**4**) pp 57-67, 2017

A novel clustering method suitable for clustering of biological signal datasets containing batched outliers

Selahaddinb B. Akben

Osmaniye Korkut Ata University, Bahce Vocational School, Turkey

batuhanakben@osmaniye.edu.tr

Abstract

During clustering analyses, instances of batched outliers of one class falling close to another class can be a significant problem. Such outliers might be incorporated into a false class or lead to the false identification of unreal classes,which can lead to false localization of the cluster centers. Here we propose a novel method for accurate classification of outliers in batched clustering analyses, aimed specifically at the type of outliers most often encountered in biological signals. The recommended divisive hierarchical clustering method is based on how much each element in the dataset is unwanted by other elements. In this method, the reluctance vectors applied to each element by the other elements are first determined. According to the common features of the reluctance vectors (horizontal and vertical components), two initial classes are obtained from some elements. All remaining elements are then included into classes according to their proximity to these classes. Then, using the reluctance vectors developed between the two established classes, class that might be re-divided are identified and further classes are constituted using the same splitting method. To validate this approach, which we named the selfish data clustering (SDC) method, areal dataset was analyzed using the SDC method and other commonly applied clustering methods. We found that our clustering method outperformed the conventional approaches by up to 12% (average is 6%) in datasets with low silhouette values.

Keywords: Batched outliers; clustering; data mining; force fields; sparse data.

1. Introduction

The goal of clustering is to place data into appropriate groups. In clustering analyses, many of the commonly encountered issues (e.g., accurate selection of the clusters and accurate placement of outliers) can emerge, complicating interpretation of the data (Figure 1) (Popat & Emmanuel, 2014; Sarumathi *et al*., 2013).

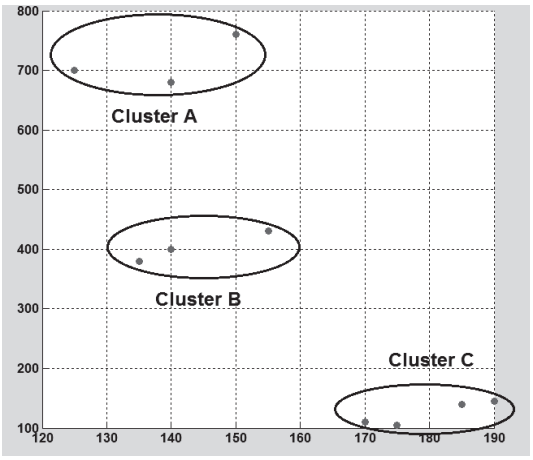

Fig. 1. A probable cluster problem related to the batched outliers.Given the available data, it is difficult to determine whether cluster–B a separate class or whether Clusters A and C are made up of outliers.

As seen in Figure 1, in data clusters in which the outliers belonging to the classes are situated in batches, it can be difficult to determine, which elements make up the clusters and into which cluster the outliers could be incorporated (Kumar *et al*., 2014).

Since biological signals depend on many criteria (e.g., movement, momentary situation, source of signal and external factors), biological signals data often include a great deal of outliers (Wolson & Clarke, 2011; Chrominski & Tkacz, 2010; Nallamhut & Palanichamy, 2015). Therefore, the clustering problems represented in Figure 1 are more likely to be encountered, when dealing with biological signals.

In the datasets containing batched outliers, there are a greater number of elements that are remote to their class and close to another class (Tong & Barfoot, 2011). However, the mean silhouette coefficient of the dataset is actually lower (Rousseeuw, 1987). Then, the solution to be suggested for the datasets comprising outliers in the form of a group is the solution, which shall be proposed for the datasets, whose mean silhouette coefficient is lower.The most frequently employed cluster methods are the centroid-based, divisive-based and density-based methods (Karaboga & Ozturk, 2011). The K-Means and Fuzzy C-Means approaches are the most frequently used centroid-based methods. In both

of these methods, the centers of classes are randomly chosen and the classes are configured according to their proximity to the centroids of the elements. New centroids are then determined and the same processes repeated until the centroids become invariant (Ghosh & Dubey, 2013). Nevertheless, the batched outliers may sometimes be close to an incorrect class center (Figure 2).

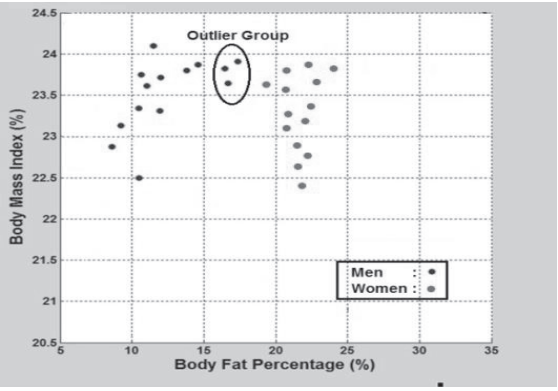

Fig. 2. An exemplary dataset in which the batched outliers are close to another class center.

In situations such as shown in Figure 2, instead of only considering the proximity of the outliers to theclass center, it is also necessary to consider their proximity to the elements of other classes. Figure 2 demonstrates that there are intercommunication elements between the outliers andclass center they belong to.

Support vector clustering is an alternative clustering method that gives attention to the linkage between each one of the same class elements for the issues,similar to the case shown in Figure 2. In this method, the gap where there is the lowest connection among the elements in the dataset is selected and a hyper plane is constructed, the classes are formed from the elements according their position relative to this hyper plane(Ben-Hur *et al*., 2002). However, batched outliers can lead to faulty hyperplane creation (Figure 3).

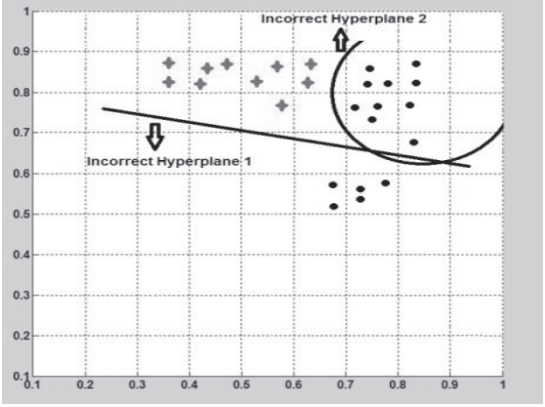

Fig. 3. Possibilities due to faulty hyper. plane formation due to batched outliers inside

As seen from Figure 3, the hyperplane formed according to inter-elements largest gap may develop due to excessive distance of the outliers. If it were information on the actual classes, this error at Figure 3 would not occur. Namely, the problem with Figure 3 may be eliminated, if the members of some classes are expressly known at the beginning.

Therefore, a new clustering method is required in which the initial classes are shapedaccording to those elements, whose class memberships are certain and the outliers are evaluated according to these initial classes. For this purpose,in this study, the silhouette value (silhouette coefficient) was composed of the initial classes, because the potential for the elements with higher silhouette coefficients to become a class member is higher (Zhou & Gao, 2014). The remaining elements were then placed into these initial classes according to their distance from the initial classes. Finally, the dataset was divided into two distinctive classes and further classes obtained using the same dividing method.

For comparison, frequently employed clustering methods and our proposed novel method were applied to a real biological dataset for clustering analyses.

2. Materials

The dataset used for detailed analysis was lower limb EMG dataset. Also, three datasets (Iris, Liver, EEG) were used to approve the result of first dataset. All datasets have been taken from the UCI database. This dataset consists ofEMG signals recorded from 5 channels. The purpose of the data was to be able to detect each of 11 different abnormalities associated with 5 different muscle movements, namely, each of 5 attributes each has 11 classes. This dataset was chosen, because the classes within some channel data (attributes) contain batched outliers. Therefore, the dataset is composed of the attributes possessing mean silhouette coefficients in distinct values. In this way, the silhouette variance-based contribution of the proposed method towards efficient clustering can be assessed, namely, the amount of batched outlier-based method's contribution to the success shall be determined. The mean silhouette coefficients pertaining to the attributes of the dataset are shown in Table 1.

Table 1. Mean silhouette coefficients of the datasets.

	Mean silhouette coefficient
Attribution – 1 Datasets created with other attributes	0.60
Attribution – 2 Datasets created with other attributes	0.51
Attribution – 3 Datasets created with other attributes	0.46
Attribution – Datasets created with other attributes	0.26
Attribution – 5 Datasets created with other attributes	0.43
OVERALL AVERAGE	0.47

The average silhouette coefficients at Table 1, have been produced according to the known class labels. In each line of Table 1, the average of the silhouette coefficient pertaining to the data pair generated by an attribution together with other attributes.

3. Method

Centroid, distribution, density and separation based methods are all currently applied in clustering analyses. Amongst these approaches, the K-means, Fuzzy C-Means, GMM, DBSC and SVC methods are most frequently used; all of which have a distinctive logic. Here we compared the performance of our novel clustering method, against these most frequently used clustering methods using real biological data (Mirkin, 2012; Moses & Deisy, 2015). Furthermore, silhouette coefficient was used to assess the dataset patterns (structure) in this study. Silhouette coefficient is used to measure the consistency within clusters of data. Therefore, the silhouette coefficient was used to evaluate the relation between success rate and cluster structure, in this study. Also, the silhouette coefficient calculation is given below:

Assume the data have been clustered. For each datum i, let $a(i)$ be the average dissimilarity of i with all other data within the same cluster and let $b(i)$ be the lowest average dissimilarity of i to any other cluster, of which i is not a member (Rousseeuw, 1987; Amorim & Hanning, 2015). In this case, the silhouette coefficient of datum i is:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i),\ b(i)\}} \quad (1)$$

Then mean silhouette coefficient of data set containing the datum i, is:

$$ms(i) = \frac{1}{N}\sum_{i=1}^{N} s(i) \quad (2)$$

While N is the element number of data set

3.1. New clustering method: selfish data clustering

To enable the creation of an initial class from those elements with higher silhouette coefficients in a clustering analysis, those elements that might generate a group (batch) first need to be determined. Because the in-group proximity of those elements that might make up the batches and the distance of such groups from each other can enhance the silhouette coefficients of the elements of a group (Piernik *et al*., 2015), if the distance of the elements from each other is grouped with a common peculiarity, those elements that form a group can be determined (Kaufman & Rousseeuw, 2009). The distances of the elements from each other should, therefore, be determined first. The elements can then be subsequently separated into groups using the common peculiarities of these distances.

The distance of an element from other elements may be defined as the reluctance vector (the repulsive force exerted by an element to another element). Also, by virtue of the direction of vectors, common properties can be identified. Because in doing so, to what direction of any element is not wanted by other elements is set down. Consequently, initial batches can be created according to the directions of the reluctance vectors applied to the elements.

To generate such initial batches, we first determined the extent to which each element was unwanted by the other elements (in both the horizontal and vertical direction). Figure 4 illustrates the components of the reluctance vectors applied by the various elements to each other.

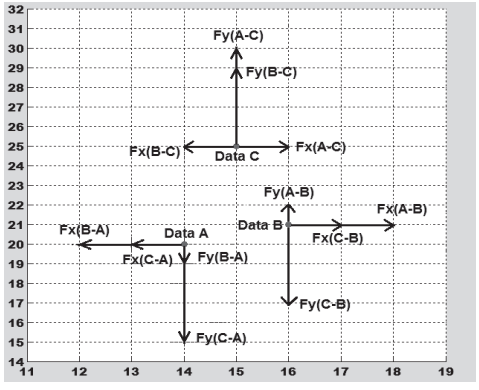

Fig. 4. The components of reluctance vectors over the horizontal and vertical axes. Fx(A-B) is the horizontal component of the reluctance vector applied by element A to element B. Fy(A-B) is the vertical component of the reluctance vector applied by element A to element B.

The components shown in Figure 4 were computed as follows:

For the dataset,

$$DATA\_SET = [A_1, A_2, \cdot \quad \cdot \quad \cdot, A_n] \tag{3}$$

The sums of the horizontal and vertical reluctance vectors applied by each elementto every other element may be calculated as follows:

$$FA_{i_x} = \sum_{z=1}^{z=n} \left(A_{i_x} - A_{z_x}\right)$$
$$FA_{i_y} = \sum_{z=1}^{z=n} \left(A_{i_y} - A_{z_y}\right) \tag{4}$$

Where A_{i_x} is the sum of the reluctance vectors and A_{i_y} is the sum of the vertical components of the reluctance vectors, applied to A_i. Therefore, the total reluctance vector applied to every element can be calculated using:

$$FA_i = \sqrt{\left({FA_{i_x}}^2 + {FA_{i_y}}^2\right)} \tag{5}$$

According to this formulation, it can be said that the sums of sums of reluctance vectors applied to all elements have also 4 each direction according to horizontal and vertical components. The directions of the reluctance vector sums applied to the elements are shown in Figure 5.

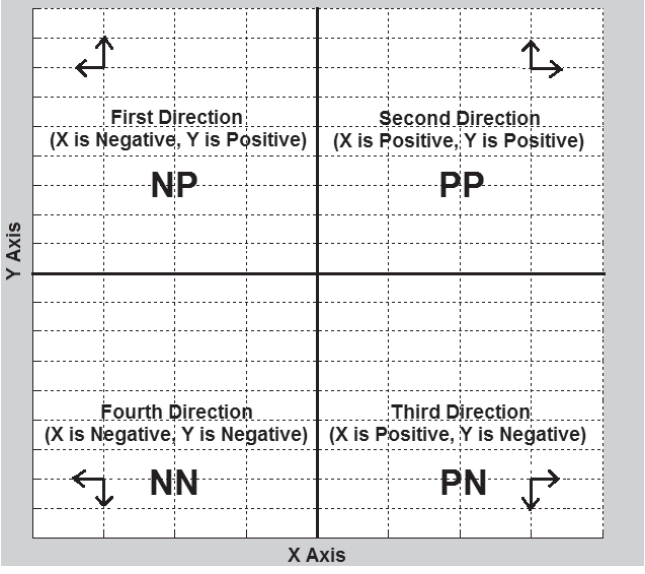

Fig. 5. The directions applied to the sums of reluctance vectors of the elements

As seen in Figure 5, there are four possible directions, named according to their vertical and horizontal components: PP (positive horizontal and vertical components); PN (positive horizontal component, negative vertical component); NN (negative horizontal and vertical components) and NP (negative horizontal and positive components). In this respect, the sum of reluctance vectors with components towards the same directionmay be employed as the common properties of those elements. In this way, the elements can be divided into four different groups according to their common properties.

Nevertheless, the initial purpose of the suggested method is to divide the dataset into two batches. Therefore, it would be necessary to identify, which two of the four groups obtained want most to break with the dataset. This calculation may be accomplished by computing the sum of the sums of the reluctance vectors within every batch. Namely, upon summing the parallel direction reluctance vectors, four feature vectors are obtained. Four such feature vectors are shown in Figure 6.

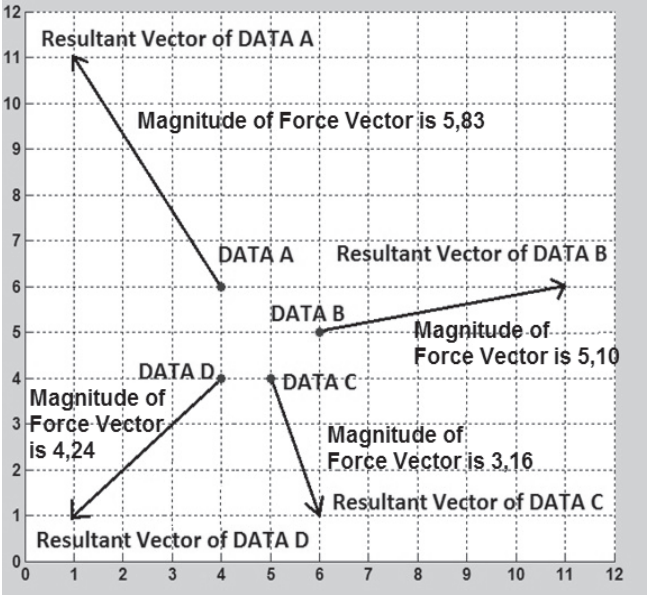

Fig. 6. Feature vectors

As shown in Figure 6, there are four potential directions. Therefore, the sum of reluctance vectors applied to the elements is the same as feature vector too.

It can be said that two of the four biggest feature vectors obtained are composed of the most appropriate elements from which to form a group within the dataset, because greater reluctance forces have been applied to the elements that created these two feature vectors (i.e., that the elements of two feature vectors chosen are composed of the elements with higher separation coefficient, which ensure the silhouette coefficient to be higher). Therefore, the silhouette coefficients belonging to two each of the chosen feature vectors would be higher compared to those elements forming the other two feature vectors. Finally, those elements making up each of the two chosen feature

vectors are called the elements of probable two classes. Figure 7 shows the selection of the feature vector size-based probable class members for an exemplary dataset.

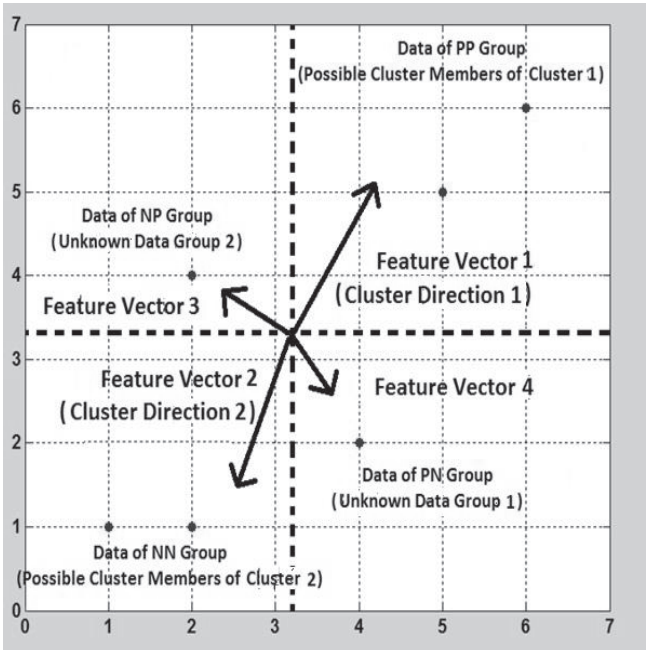

Fig. 7. Probable class member determination by the feature vectors

As seen in Figure 7, the first and second feature vectors are bigger than the other feature vectors. Therefore, the elements making up the first and second feature vectors are the probable members of two distinctive classes.

However, those data points that fall outside of these probable class members, which are closer to the center of the dataset, might have been erroneously classified. To avoid such mistakes, the center of the four separate groups is determined according to the four feature vectors. Subsequently, those points outside of the two previously determined distinctive probable classes, which are closer to their own class center versus the other centers, are determined to certain class members. All remaining elements are considered unknown elements, which are then incorporated into the classes as follows: (1) The unknown elements closest to certain class members are incorporated into that class; (2) In doing so, the number of classes to which that element is added increases and the area of those classes are enlarged and the number of unknown elements decreases; (3) This operation continues until all unknown elements are classified (Figure 8).

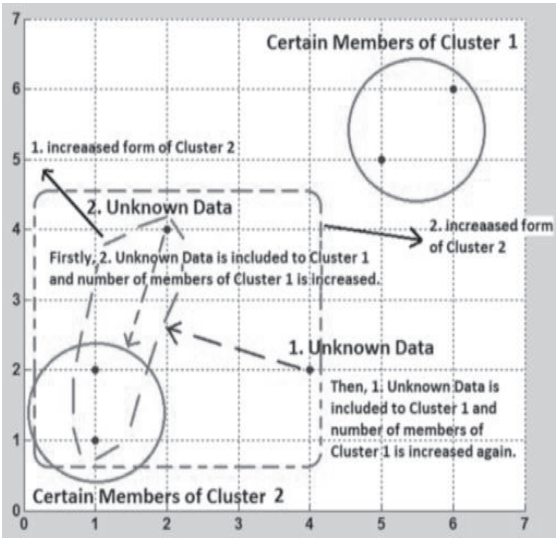

Fig. 8. Incorporation of unknown elements to the classes

As seen in Figure 8, first, the second unknown element has been added to the second class. This is because this element is the element that is closest to the previously ascertained elements of the second class. Upon addition of the second unknown element to the second class, the number of second class elements increases from two to three. The first remaining element has been evaluated according to the elements of this new class.

Because our proposed method is based on the refusal of elements by every other element, we have named this the sel fish data clustering (SDC) method. A flowchart of the SDC method is shown in Figure 9. Also mathematical definitions of SDC algorithm can be seen below Figure 9.

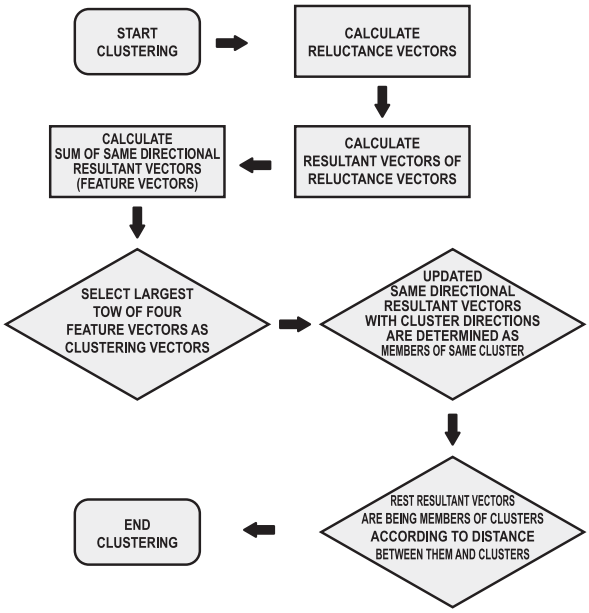

Fig. 9. Flowchart of the proposed selfish data clustering method

As described in Equation 5, $FA_{i_{xy}}$ is the reluctance force applied to datum i which is the element of data set A. Also, assume that $F_{N_{xy}}$ is the feature vector of dataset A and $F_{U_{xy}}$ is the other vectors. In addition, x and y are the vector components. Note that the N and U will be "1" or "2". Based on these variables, temporary membership of $FA_i(MT(N))$ is determinedas in Equation 6.

Then, locations of $mean(Group\ F_N)$ and $mean(Group\ F_U)$ are determined and Equation 7 is applied.

$$MT(N) = \begin{cases} If\ xy\ of\ FA_i\ and\ F_N\ are\ in\ same\ directions & ,\ A_i \in Group\ F_N \\ Otherwise & ,\ A_i \in Group\ F_U \end{cases} \quad (6)$$

$$MT(N) = \begin{cases} If\ A_i\ is\ close\ to\ mean(Group\ F_N) & ,\ A_i \in Group\ F_N \\ If\ A_i\ is\ close\ to\ mean(Group\ F_U) & ,\ A_i \in Group\ F_U \end{cases} \quad (7)$$

At the end of Equation (7), elements of $Group\ F_N$ are the determined members of these 2 groups. Finally, the closest element of $group\ F_U$ to $group\ F_N$ is determined as the membership of the $group\ F_N$ and the element number of $group\ F_N$ decreases. Iteration is ended when all elements of $group\ F_U$ is wither.

3.2. Some exceptional circumstances of elements

If the sum of the reluctance vectors of an element is created by a single component, as in Figure 4 (when there is only a horizontal and vertical axis), at the beginning, such an element is considered to be an unknown element. This is because, in such a case, it would be unclear, what feature vector this element shared common features with.

3.3. Some exceptional circumstances of feature vectors

Where a feature vector is composed of a single component, there are three feature vectors in total. In such a case, the feature vector is created by a single component and the elements having the same directional components are chosen as the elements for one of the initial classes.This is because a single-component feature vector is made up of the sum of two equal feature vectors and would be bigger than the other two feature vectors.

3.4. Creating more than two classes

If creation of more than two classes is required, firstly, two classes are generated. Then, the same method is applied to each of these newly developed classes.The sum of the feature vectors for each of these classes is then compared.

The class with the greatest value for itssum of its feature vectors isdivided into two.The previous method is then applied to the class that must be divided into two, resulting in three distinct classes. These processes are repeated for the four or more classes. Put differently, firstly two classes are created. Then, the class that is most suitable for division is split and the number of classes is augmented by repeating these operations.

4. Experimental results

Our novel method was applied, together with other methods, to Lower Limb EMG Dataset. The outcomes of these analyses are shown in Table 2.

Table 2. Accuracy ratios of the applied clustering methods for Lower Limb EMG Dataset

	SDC	K-MEANS	FCM	GMM	SVC	DBSCAN
Attribution 1: Data sets created with other attributes	87.0	85.3	84.9	81.7	84.0	83.2
Attribution 2: Data sets created with other attributes	84.3	82.2	79.1	79.3	81.1	81.3
Attribution 3: Data sets created with other attributes	81.1	77.0	76.4	75.6	77.7	77.1
Attribution 4: Data sets created with other attributes	79.3	66.0	66.7	65.5	67.8	67.3
Attribution 5: Data sets created with other attributes	80.4	72.6	75.4	72.1	73.2	71.2
Average	82.4	76.6	76.5	74.8	76.8	76.0

As shown in Table 2, the SDC method is approximately 6% more successful than the other clustering methods. To be able to better appreciate the significance of this result, it is necessary to also consider the mean silhouette coefficients of the attributions (Table 1), which has been done in Figure 10. Furthermore, in order to be better able to appreciate the analysis, the mean silhouette coefficients shown in Figure 10 have been multiplied by 100.

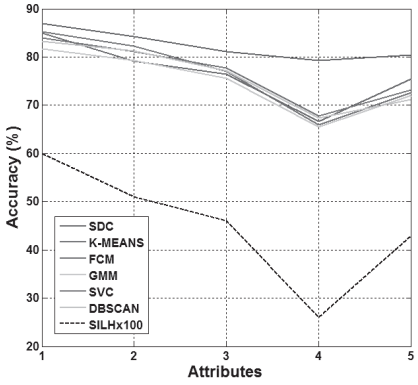

Fig. 10. Joint evaluations of mean silhouette(SILH) coefficients and ratios of achievement for Lower Limb EMG Dataset

By this approach, we found that the SDC is more resistant than those datasets where the mean silhouette coefficient is relatively lower. This resistance appears to be particularly clear for the fourth attribute. Therefore, in addition to being more successful than other methods, the SDC method is also less affected by batched discrete elements (outliers).

Then, the processing time of the methods were also tested. We found that the K-MEANS was fastest, while the SDC method was faster than average (Table 3).

Table 3. Processing times (sec) of the applied clustering methods

	SDC	K-MEANS	FCM	GMM	SVC	DBSCAN
Attribution 1: Data sets created with other attributes	0.071	0.049	0.189	0.292	0.121	0.188
Attribution 2: Data sets created with other attributes	0.884	0.282	1.088	1.680	0.696	1.079
Attribution 3: Data sets created with other attributes	0.584	0.196	0.756	1.168	0.484	0.750
Attribution 4: Data sets created with other attributes	3.330	0.983	3.792	5.858	2.427	3.761
Attribution 5: Data sets created with other attributes	0.368	0.134	0.517	0.799	0.331	0.513
Average	0.819	0.263	1.016	1.570	0.651	1.008

In order to better see the superiority of proposed method over the other methods, three datasets were subjected to comparison again. Comparison results can be seen in Figure 11, Figure 12, Figure 13 and Table 4.

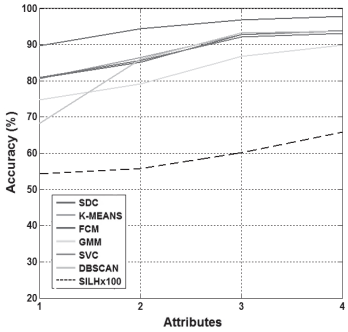

Fig. 11. Joint evaluations of mean silhouette (SILH) coefficients and ratios of achievement for Iris Dataset

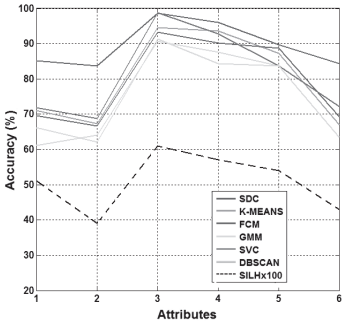

Fig. 12. Joint evaluations of mean silhouette (SILH) coefficients and ratios of achievement for Liver Dataset

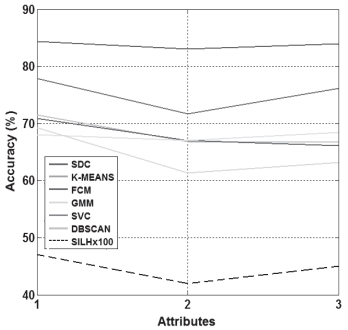

Fig. 13. Joint evaluations of mean silhouette (SILH) coefficients and ratios of achievement for EEG Dataset

Table 4. Accuracy comparison of SDC and others in various datasets

Noise Removed Datasets Except SDC	SDC	K-MEANS	FCM	GMM	SVC	DBSCAN
Lower Limb EMG Dataset	82.4	76.6	76.5	74.8	76.8	76.0
Iris Dataset	94.3	88.1	88.3	83.7	88.6	84.2
Liver Dataset	89.6	80.9	79.6	75.1	81.4	75.2
EEG Dataset	83.81	68.37	68.02	64.59	75.25	67.86
Average	87.52	78.49	78.10	74.54	80.51	75.81

As shown in Figure 11, Figure 12, Figure 13 and Table 4, SDC is superior to others. Also SDC is more resistant than those datasets, where the mean silhouette coefficient is relatively lower. These results are similar with result seen in Figure 10. These similarities confirm the SDC›s superiority defined in conclusion of Figure 10.

At the last stage of study, noise removal process was applied to traditional methods. Then preprocessed traditional methods were compared with SDC again. Distance based noise removal method was applied to data sets (Xiong & Steinbach, 2006). Then noise removed data sets were used as inputs again. The new comparison results can be seen in Table 5.

Table 5. Accuracy comparison of SDC and others in noise removed datasets

Noise Removed Datasets Except SDC	SDC	K-MEANS	FCM	GMM	SVC	DBSCAN
Lower Limb EMG Dataset	82.40	79.05	78.80	76.82	78.87	78.20
Iris Dataset	94.30	90.92	90.95	85.96	90.99	86.64
Liver Dataset	89.60	83.49	81.99	77.13	83.60	77.38
EEG Dataset	83.81	70.56	70.06	66.33	77.28	69.83
Average	87.53	81.00	80.45	76.56	82.69	78.01

As shown in Table 5, noise removal process increases the success of traditional methods. But this increase is poor, because the results in Table 5 are similar to results in Table 4. So, it can be said that SDC is superior to other methods, again in noise removed datasets.

5. Discussion

Here we aimed to develop a novel method, which is better able to assign batched discrete values (outliers during clustering analyses. Erroneous assignment of discrete values is often a result of lower silhouette coefficients (Agresti, 2013). It is therefore necessary that the elements with higher silhouette coefficients are classified first, with the remaining values being incorporated into their classes according to their similarity to the previously classified elements. Silhouette coefficients are higher, when the elements are closer to their own centers. Nevertheless, since the centers of classes are unknown, the two elements with greatest separation factor may be utilized to establish an initial class. Because the separation coefficient is a coefficient, which determines the coefficient of silhouette. The separation and silhouette coefficients are positively correlated (Berkhin, 2006). When applying the SDC method, two initial classesare created from the uttermost periphery of the dataset elements (i.e., those elements that are farthest from each other). In doing so, the exact membership of some elements pertaining to two different classes can be fixed and information about the position (center) of these classes can be obtained. Later, the remaining elements can be assigned to these two classes according their resemblances. In this way, the SDC method produces fewer positioning mistakes than other methods. This is because either the classes are being created at the beginning or the initial classes are being formed randomly. Moreover, owing to this feature, proposed method is being likely to be distinguished at the first glance from the class members with exact discrete values closer to other classes. In stage two, the remaining elements are placed into classes according to their proximity to the predetermined class members. In this way, the remaining elements are distributed into the classes, so as to have the highest silhouette factor.

Because proximity of an element to the class elements results in a small cohesion coefficient, and subsequently a higher silhouette coefficient (Corral *et al*., 2006), the SDC method is aimedat ensuring that the silhouettecoefficientis also higher in the distribution of outliers. In addition, as the otherelement with predetermined membership would be closer to their own class centers, but farther than other centers and places the remaining according to their distances from class centers as well.

However, other methods distribute outliers according to either their proximity to the closest class member or according to the class center. The SDC method is therefore advantageous for the placement of outliers.

In addition to these advantages, the placement stage of the remaining elements by the method suggested may also be used to increase the performance of other methods by being combinedwith other methods. Similarly, the initial stage of the SDC method could be employed as a commencement algorithmin other methods. Although the processing time of the SDC method is slower than some alternative methods, its processing time is faster than the mean processing time of the methods considered here. Therefore, the SDC method is suitable for use with moderately large real-life applications (e.g., biological datasets).

6. Conclusion

Biological datasets often contain many outliers that can form small groups. To date there has been on clear consensus on the best clustering method to address this issue (Fielding, 2007). Here we propose a novel clustering method based upon inter-element resemblances and uniqueness, which we name as the SDC method.

We found that the SDC method outperformed other commonly applied clustering methods by materializing the clustering through predetermination of some class members and acquisition of details on the classes.

By this approach, faulty determination of cluster position within batched outliers is avoided. When using datasets with a low mean silhouette factor, the SDC method outperformed other clustering methods by up to 12%. The reason of this success is that attention was paid to the placement of elements into classes in a way to ensure the elements would have higher silhouette coefficient.

References

Agresti, A. (2013). Categorical data analysis, Wiley.

Amorim, R. C., Hennig, C. (2015). Recovering the number of clusters in data sets with noise features using feature rescaling factors, Information Sciences, **324**:126–145.

Berkhin P. (2006). A survey of clustering data mining techniques, Springer.

Ben-Hur, A., Horn, D., Siegelmann, H.T. & Vapnik.V. (2002). Support vector clustering. The Journal of Machine Learning Research, **2**:125-137.

Fielding, A.H. (2007). Cluster and classification techniques for the biosciences, Cambridge University Press.

Chrominski, K. & Tkacz, M. (2010). Comparision of outlier detection methods in biomedical data. Journal of Medical Informatics & Technologies, **16**:89-94.

Corral, G., Fornells, A., Golobardes, E. & Abella, J. (2006). Cohesion factors: Improving the Clustering Capabilities of Consensus, Springer.

Ghosh, S. & Dubey, S.K. (2013). Comparative analysis of K-Means and Fuzzy C-means algorithms. International Journal of Advanced Computer Science and Applications, **4**:35-39.

Karaboga, D. & Ozturk, C. (2011). A novel clustering approach: Artificial bee colony (ABC) algorithm. Applied Soft Computing, **11**:652–657.

Kaufman, L. & Rousseeuw, P. J. (2009). Finding groups in data an introduction to cluster analysis, Wiley.

Kumar, N.S., Rao, K.N., Govardhan, A. & Reddy, K.S. (2014). An updated literature review on the problem of class imbalanced learning in clustering.International Journal of Engineering and Technical Research, **2**:123-128.

Mirkin,. B. (2012). Clustering: a data recovery approach. Chapman and Hall/CRC.

Mosec, D. & Deisy, C. (2015). A survey of data mining algorithms used in cardiovasculardisease diagnosis from multi-lead ECG data. Kuwait Journal of Science, **42**(2):206-235.

Nallamhut, R. & Palanichamy, J. (2015). Optimized construction of various classification models forthe diagnosis of thyroid problems in human beings. Kuwait Journal of Science, **42**(2):189-205.

Popat, S.K. & Emmanuel, M. (2014). Review and comparative study of clustering techniques. International Journal of Computer Science and Information Technologies, **5**:805-812.

Piernik, M., Brzezinski, D., Morzy, T.& Lesniewska, A. (2015). XML clustering: a review of structural approaches. The Knowledge Engineering Review, **30**:297-323.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, **20**:53–65.

Sarumathi, S., Shanthi, N. & Sharmila, M. (2013). A review: Comparative analysis of different categorical data clustering ensemble methods. International Journal of Computer, Control, Quantum and Information Engineering, **7**:974-984.

Tong, C.H. & Barfoot, T.D. (2011). Batch heterogeneous outlier rejection for feature-poor SLAM.IEEE International Conference on Robotics and Automation (ICRA).Shanghai, China.

Xiong, H. & Steinbach, M. (2006). Enhancing data analysis with noise removal.IEEE Transactions on Knowledge and Data Engineering, **18**(3):304-319.

Woolson, R.F. & Clarke, W.R. (2011). Statistical methods for the analysis of biomedical data. John Wiley & Sons. Inc., New York.

Zhou, H.B. & Gao, J.T. (2014). Automatic method for determining cluster number based on silhouette coefficient.Advanced Research on Intelligent System, **951**:227-230.

Submitted *:* 27/11/2015
Revised *:* 23/02/2016
Accepted *:* 02/03/2016

طريقة تجميع جديدة مناسبة لتجميع مجموعات بيانات الإشارات الحيوية التي تحتوي على قيم متطرفة مجمعة

صلاح الدين أكبين

جامعة كوركوت اطا في العثمانية ، مدرسة بهس المهنية، تركيا

batuhanakben@osmaniye.edu.tr

خـلاصـة

أثناء تحليل المجموعات، يمكن أن تشكل نماذج من قيم متطرفة مجمعة في فئة واحدة تقع بالقرب من فئة أخرى مشكلة كبيرة. فمن الممكن أن تندمج مثل هذه القيم المتطرفة في فئة وهمية أو تؤدي إلى تعريف وهمي لفئات غير حقيقية، والذي قد يؤدي إلى تمركز وهمي لمراكز المجموعات. ونقترح هنا طريقة جديدة للتصنيف الدقيق للقيم المتطرفة في تحليل المجموعات المجمعة، والتي تستهدف على وجه التحديد نوع القيم المتطرفة التي غالباً ما تعترض الإشارات الحيوية. وتستند طريقة التجمعات الهرمية المقسمة الموصى بها إلى مقدار رفض كل عنصر في مجموعة البيانات لعناصر أخرى غير مرغوب فيها. في هذه الطريقة، تم أولاً تحديد ناقلات الممانعة المُستخدمة على كل عنصر بواسطة العناصر الأخرى. ووفقاً للسمات المشتركة لناقلات الممانعة (المكونات الأفقية والعمودية)، تم الحصول على فئتين أوليتين من بعض العناصر. ثم بعد ذلك تم إدراج جميع العناصر المتبقية في فئات حسب قربها من هذه الفئات. بعد ذلك، وباستخدام ناقلات الممانعة التي تم تطويرها بين الفئتين الراسختين، تم تحديد الفئة التي يمكن إعادة تقسيمها، وتم تشكيل فئات أخرى باستخدام نفس طريقة التقسيم. ومن أجل إثبات هذا النهج، والذي أطلقنا عليه اسم طريقة تجميع البيانات الأنانية (SDC selfish data clustering)، تم تحليل مجموعة بيانات حقيقية باستخدام طريقة SDC وطرق التجميع الأخرى المُطبقة بشكل شائع . ووجدنا أن طريقة التجميع الخاصة بنا تفوقت على النهج التقليدية بنسبة تصل إلى 12% (المتوسط 6%) في مجموعات بيانات ذات قيم ظلية منخفضة.