New Cluster Detection using Semi-supervised Clustering Ensemble Method

Huaying Li and Aleksandar Jeremic[*]

Department of Electrical and Computer Engineering, McMaster University Hamilton, Ontario, Canada

Keywords: Clustering, Information Fusion, Cluster Ensemble and Semi-supervised Learning.

Abstract: In the recent years there has been tremendous development of data acquisition system resulting in a whole new set of so called big data problems. Since these data structures are inherently dynamic and constantly changing the number of clusters is usually unknown. Furthermore the ”true” number of clusters can depend on the constraints and/or perception (biases) set by experts, users, customers, etc., which can also change. In this paper we propose a new cluster detection algorithm based on a semi-supervised clustering ensemble method. Information fusion techniques have been widely applied in many applications including clustering, classification, detection, etc. Although clustering is unsupervised and it does not require any training data, in many applications, expert opinions are usually available to label a portion of data observations. These labels can be viewed as the guidance information to combine the cluster labels that are generated by different local clusters. It consists of two major steps: the base clustering generation and the fusion. Since the step of generating base clusterings is unsupervised and the step of combining base clusterings is supervised, in the context of this paper, we name the algorithm as the semi-supervised clustering ensemble algorithm. We then propose to detect a new cluster utilizing the average association vector computed for each data point by the semi-supervised method.

1 INTRODUCTION

Although many clustering algorithms exist in the literature, in practice no single algorithm can correctly identify the underlying structure of all data sets (Jain et al., 1999)(Xu et al., 2005). Furthermore, it is usually difficult to select a suitable clustering algorithm for a given data set when prior information about cluster shape and size are not available. In addition, for a particular clustering algorithm, it usually generates different clusterings for a given data set by starting from different initiations or using different parameter settings. Consequently, we expect to improve the quality of the cluster analysis by combining multiple clusterings into a consensus clustering. The problem involving combination of multiple clusterings is often referred to as clustering ensemble problem in the literature (Strehl and Ghosh, 2003)(Wang et al., 2011)(Ghaemi et al., 2009)(Vega-Pons and Ruiz-Shulcloper, 2011).

In many applications it may be of interest to detect a so called new cluster i.e. an event in which new data (e.g. new type of cells, new type of customers etc.) with different statistical properties appears and cannot be classified to any of the existing clusters. Obviously one possible interpretation of such event could be a need to redefine existing clusters by defining new, larger data groups to accommodate these changes. However in certain instances, e.g. in biochemistry, by applying particular chemical treatments a desired outcome is creation of new cell types and in these cases such an algorithm would be beneficial in order to identify successful treatments. To this purpose in this paper we propose an algorithm for new cluster detection using clustering ensemble method. Usually, clustering algorithms do not require training data to generate cluster labels. However, in many applications, opinions from experts are available to label at least a portion of the data points. These labels can be utilized as supervision and guidance information for the fusion process of the cluster labels. Therefore, in this paper we utilize our previously proposed (Li and Jeremić, 2017) clustering ensemble algorithms for the scenario of presence of training data points (labelled). The proposed method consists of two major steps: the generation and fusion of multiple base clusterings. The first step is to generate a set of base clusterings by applying unsupervised clustering algorithms. The second step is to fuse multiple clusterings into a consensus clustering.

[*]This work was supported by Natural Sciences and Engineering Research Council of Canada.

Li, H. and Jeremic, A.
New Cluster Detection using Semi-supervised Clustering Ensemble Method.
DOI: 10.5220/0006653802210226
In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - Volume 4: BIOSIGNALS*, pages 221-226
ISBN: 978-989-758-279-0

This step is considered as supervised since it utilizes the labels of training data points as the guidance for the points which appears in the confusion region of two or more clusters. In the context of this paper, we name it as the semi-supervised clustering ensemble algorithm (SEA). It computes association vectors for each data point according to different base clusterings. Using these results we then propose a new cluster detection algorithm based on the average association vector generated for each data point by SEA. This proposed algorithm has the ability to automatically determine whether additional data points to a given data set come from existing classes or from a new class. The rest of this paper is outlined as follows. In Section 2, we propose the semi-supervised clustering ensemble algorithm. In Section 3, we propose the new cluster detection algorithm. In Section 4, we provide numerical examples to show the performance of our proposed algorithms using real data sets.

2 SEMI-SUPERVISED CLUSTERING ENSEMBLE

The semi-supervised clustering ensemble algorithms consists of two major steps: the generation and fusion of base clusterings, as shown in Fig. 1. A data set containing N data points is denoted as $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$, where each data point $\mathbf{x_i} \in \Re^F$, for all $i = 1, \ldots, N$, comes from F-dimensional feature space. A clusterer $\mathcal{F}_j$ is a local unit that produces a base clustering for the given data set. The base clustering is usually represented by a label vector $\lambda^{(j)}$. Cluster labels derived by different local clusterers form a set of base clusterings $\{\lambda^{(1)}, \ldots, \lambda^{(j)}, \ldots, \lambda^{(D)}\}$ and it can be viewed as an $N \times D$ label matrix $\mathcal{F}$, the entry of which on the i-th row and j-th column is the cluster label of data point $\mathbf{x}_i$ according to the j-th clustering. All the base clusterings are sent to the fusion center, which produces a consensus clustering λ^0, a better clustering of the given data set in some sense compared with each individual clustering.

In general, combining multiple clusterings is more difficult than combining local decisions (such as detection and classification results) due to many reasons. One of the obvious reasons is that the number and shape of clusters depend on the clustering algorithms that generate them as well as their parameter settings. Another reason is that the desired number of clusters is often unknown due to the lack of prior information about the data set. Furthermore, the most important reason comes from the correspondence problem of multiple clusterings due to the fact that cluster labels are symbolic. Since the clusters are not pre-defined, it is possible that the same cluster label from different clusterings represents two distinct clusters. For example, $\lambda^{(1)} = [1,2,2,1,3,2,3]^T$ and $\lambda^{(2)} = [2,1,1,3,2,3,2]^T$ represents two clusterings of a data set with seven data points. Although the two vectors are distinct, they actually represent the same partitioning of the given data set. This is the so-called correspondence problem and it makes the clustering ensemble problem more difficult to solve.

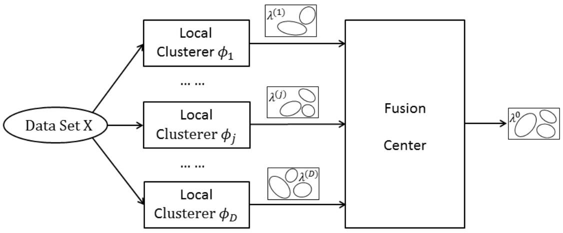

Figure 1: Two major steps of an ensemble method.

As mentioned earlier, labelled data is usually available in many applications. In order to utilize the known labels as the guidance information, in this section we propose a semi-supervised clustering ensemble algorithm. It calculates the association between each data point and the training clusters and relabels the cluster labels in $\mathcal{F}$ according to the training clusters. The fusion idea is stated as follow: (1) for a particular data point count the number of agreements between its label and the labels of training points in each training cluster according to an individual base clustering, (2) calculate the association vector between this data point and the particular base clustering, (3) compute the average association vector for this data point based on all base clusterings, (4) repeat for all data points and derive the soft consensus clustering for all the data points using the association vectors and (5) assign each data point its most associated cluster id according to the average association vector. The summary of the algorithm is provided in Table 1.

For a given data set, we name the subset containing training data points as $\mathbf{X}_r$ and the subset containing testing points, whose labels are unknown, as $\mathbf{X}_u$. The corresponding numbers of data points in these two sets are denoted by N_r and N_u. Suppose the training data points come from K_0 categories and N_r^k is the number of training points from the k-th category, i.e., $\sum_{k=1}^{K_0} N_r^k = N_r$. According to the j-th clustering $\lambda^{(j)}$, we compute the association vector $\mathbf{a}_i^{(j)}$ for the i-th unlabelled data point $\mathbf{x}_i$, where $i = 1, \ldots, N_u$ and $j = 1, \ldots, D$. Since there are K_0 training clusters, the association vector $\mathbf{a}_i^{(j)}$ has K_0 entries. Each entry describes the association between data point $\mathbf{x}_i$ and the corresponding training cluster. The k-th entry of the

Table 1: Semi-supervised clustering ensemble algorithm (SEA).

- * Input: Base clusterings $\mathcal{F}$
- * Output: Soft clustering λ_u

(a) According to label vector λ_r, rearrange base clusterings $\mathcal{F}$ into K_0+1 sub-matrices $\{\mathcal{F}_r^1,\ldots,\mathcal{F}_r^k,\ldots,\mathcal{F}_r^{K_0},\mathcal{F}_u\}$

(b) For data point $\mathbf{x}_i$, calculate the k-th element of the association vector $\mathbf{a}_i^{(j)}$ by

$$\mathbf{a}_i^{(j)}(k) = \frac{\text{occurrence of } \mathcal{F}_u(i,j) \text{ in } \mathcal{F}_r{}^k(:,j)}{N_r^k}$$

and repeat for $k=1,\ldots,K_0$ to form the association vector $\mathbf{a}_i^{(j)}$

(c) Compute the overall association vector $\mathbf{a}_i$ of data point $\mathbf{x}_i$ by $\mathbf{a}_i = \frac{1}{D}\sum_{j=1}^{D}\mathbf{a}_i^{(j)}$.

(d) Compute the association level γ_i of data point $\mathbf{x}_i$ to all training clusters by $\gamma_i = \sum_{k=1}^{K_0}\mathbf{a}_i(k)$.

(e) Compute the membership information of data point $\mathbf{x}_i$ to every cluster by normalizing $\mathbf{a}_i$

(f) Repeat step (b) to (d) to generate the association level vector γ_u and repeat step (b) to (e) to generate the soft clustering λ_u

(g) (Optional) Assign data point $\mathbf{x}_i$ its most associated cluster id, which corresponds to the highest entry in the overall association vector and repeat for all $i = 1,\ldots,N_u$.

association vector $\mathbf{a}_i^{(j)}$ is calculated by the ratio of occurrence of $\mathcal{F}_u(i,j)$ in $\mathcal{F}_r{}^k(:,j)$ to the number of data points in the k-th training cluster, i.e.,

$$\mathbf{a}_i^{(j)}(k) = \frac{\text{occurrence of } \mathcal{F}_u(i,j) \text{ in } \mathcal{F}_r{}^k(:,j)}{N_r^k}, \tag{1}$$

where $\mathcal{F}_u(i,j)$ is the cluster label of data point $\mathbf{x}_i$ according to the j-th base clustering and $\mathcal{F}_r{}^k(:,j)$ represents the labels of all data points in the k-th training category generated by the j-th local clusterer. For each data point $\mathbf{x}_i$, different association vectors $\mathbf{a}_i^{(1)},\ldots,\mathbf{a}_i^{(D)}$ are calculated since there are D base clusterings generated for the given data set. In order to fuse the information, the average association vector $\mathbf{a}_i$ for data point $\mathbf{x}_i$ is computed by averaging all the association vectors $\mathbf{a}_i^{(j)}$, i.e.,

$$\mathbf{a}_i = \frac{1}{D}\sum_{j=1}^{D}\mathbf{a}_i^{(j)}. \tag{2}$$

Each entry of $\mathbf{a}_i$ describes the consolidated association between data point $\mathbf{x}_i$ and one of the training clusters. As a consequence, the summation of all the entries of $\mathbf{a}_i$ could be used to describe the association between data point $\mathbf{x}_i$ and all the training clusters quantitatively. We define it as the association level of data point $\mathbf{x}_i$ to all the training clusters and denote it as γ_i, i.e.,

$$\gamma_i = \sum_{k=1}^{K_0}\mathbf{a}_i(k). \tag{3}$$

By computing the association levels for all the data observations, the association level vector γ_u for the testing set $\mathbf{X}_u$ is made up by stacking all the association levels, i.e., $\gamma_u = [\gamma_1,\gamma_2,\ldots,\gamma_{N_u}]^T$. The soft consensus clustering of testing set $\mathbf{X}_u$ is derived by normalizing the association vector of each data point $\mathbf{a}_i$. Let us denote the soft consensus clustering of test set $\mathbf{X}_u$ by a label matrix λ_u. The i-th row of λ_u is computed by normalizing the association vector $\mathbf{a}_i$, i.e.,

$$\lambda_u(i,:) = \mathbf{a}_i^T/\gamma_i. \tag{4}$$

If a hard consensus clustering is required, the labels for data point $\mathbf{x}_i$ can be obtained by assigning its most associated cluster id, which corresponds to the highest entry in the average association vector. The consensus clustering λ^0 is obtained by repeating this step for all $i = 1,\ldots,N_u$.

3 NEW CLUSTER DETECTION

As mentioned earlier, the lack of prior information about the data set, such as the size, shape and number of the clusters, is one of the reasons that makes the clustering ensemble problem difficult to solve. Although training data usually provides prior information about the given data set, such as the size and the shape of the clusters, since these data points can be viewed as scatter points that clearly outline the shape of each cluster, we may always question on whether the amount of training clusters is enough or whether a new cluster is necessary to describe the nature of the given data set especially when there are new observations available.

Since the known labels of the training data points provide information about the given data set such as the number of the clusters, data points with unknown labels are expected to form a set of clusters, similar to the clusters formed by training data points in size, shape and quantity. When additional data observations become available or the environment of making such observations changes all the time, the prior information derived from the training data may be not sufficient to improve the accuracy of cluster analysis. Therefore, we propose the new cluster detection algorithm in this section, which is based on computing and comparing the association levels of additional data points to all training clusters and the association levels of existing data points to all training clusters. The objective of the proposed algorithm is to determine whether the additional data observations belong to existing classes or a new class.

Suppose additional data observations form a set of data points, denoted by $\mathbf{X}_a$. The summary of the new cluster detection algorithm is listed in Table 2. The input of the proposed algorithm is the combination of the original data set and the additional data set, i.e., $\mathbf{X} = \{\mathbf{X}_r, \mathbf{X}_u, \mathbf{X}_a\}$. Recall in the SEA, in order to obtain the consolidated clustering, the average association vector $\mathbf{a}_i$ is calculated by averaging the association vectors $\mathbf{a}_i^{(j)}$ for $j = 1, \ldots, D$, where D is the total number of base clusterings. The association level γ_i is defined to evaluate the association between data point $\mathbf{x}_i$ and all the training clusters by summing all entries of the average association vector $\mathbf{a}_i$. Intuitively, a data point belonging to the existing classes of the training data points should locate inside the contours outlined by all the training points and it should highly associate with one of the training cluster. Therefore, the association level of this data point to all the training clusters is relatively high. In contrast, a data point from a new class other than the existing training classes should locate outside from the contours outlined by all the training points. The association level of this data point to all the training clusterers should be low compared with that of a data point from existing classes. As a consequence, we could make decisions about the category information of the additional data sets by comparing the distribution of the association level of original data points and the distribution of the association level of the additional data points. If the distribution of the association level of the original data points is consistent with that of the additional data points, the additional data points are expected to come from the existing classes. Otherwise, they are expected to come from a new class.

Table 2: New Cluster Detection Algorithm.

```
* Input: Data set $\mathbf{X}$; Percentage $\eta$
* Output: New cluster indicator $i_{new}$
(a) Apply SSEA on $\mathbf{X}_b = \{\mathbf{X}_r, \mathbf{X}_u\}$ and obtain the association level vector $\gamma_b$
(b) Set the threshold $\gamma_{th} = \max(\gamma_b) * \eta$
(c) Count the number of original data points $N_b$ satisfying $\gamma_b < \gamma_{th}$
(d) Apply SSEA on $\mathbf{X} = \{\mathbf{X}_r, \mathbf{X}_u, \mathbf{X}_a\}$ and obtain the association level vector $\gamma_a$
(e) Count the number of additional data points $N_{new}$ satisfying $\gamma_a < \gamma_{th}$
(f) Set the threshold for the hypothesis testing as $N_{th} = N_a * \frac{N_b}{N_u}$ and determine $i_{new}$ by
    $i_{new} = \begin{cases} 0 & \text{if } N_{new} < N_{th} \\ 1 & \text{if } N_{new} \geq N_{th} \end{cases}$
```

In the proposed new cluster detection algorithm, another input is the pre-defined percentage η (determined by the users) and is used to determine the threshold γ_{th} for the association level, i.e., $\gamma_{th} = \max(\gamma_b) * \eta$. Suppose the association level of the original data points and additional data points to all the training clusters are denoted by γ_b and γ_a respectively. The sizes of testing set $\mathbf{X}_u$ and additional set $\mathbf{X}_a$ are denoted by N_u and N_a respectively. The numbers of original and additional data points whose association levels are less than the threshold γ_{th} are denoted by N_b and N_{new} respectively. To determine whether a new cluster is necessary or not for the given data set, we perform a hypothesis testing with two hypotheses:

H_0 : No data observations come from a new class

H_1 : Some data observations come from a new class.

The threshold N_{th} for the hypothesis testing is calculated by

$$N_{th} = N_a * \frac{N_b}{N_u}. \tag{5}$$

When $N_{new} < N_{th}$, the hypothesis H_0 is favoured and the new cluster indicator is set to be 0. When $N_{new} \leq N_{th}$, the hypothesis H_1 is favoured and the new cluster indicator is set to be 1.

4 NUMERICAL EXAMPLES

In this section, we provide numerical examples to show the performance of our proposed semi-supervised clustering ensemble algorithm and the new cluster detection algorithm. Since the expected cluster labels for each data set are available in the experiments, we use micro-precision as our metric to measure the accuracy of a clustering result with respect to the expected labelling. Suppose there are k_t classes for a given data set $\mathbf{X}$ containing N data points and N_k is the number of data points in the k-th cluster that are correctly assigned to the corresponding class. Corresponding class here represents the true class that has the largest overlap with the k-cluster. The micro-precision is defined by $mp = \sum_{k=1}^{k_t} N_k / N$ (Wang et al., 2011).

We use two types of data to evaluate the proposed clustering ensemble method. One type of data comes from the UCI machine learning repository website

(Bache and Lichman, 2013), which provides hundreds of data sets for the study of classification and clustering. In the literature, many researchers evaluate their clustering algorithms and clustering ensemble methods using data sets from this website (Wang et al., 2011)(Likas et al., 2003)(Zhou and Tang, 2006)(Yan et al., 2009)(Zhang and Gu, 2014). The other type of data comes from a biomedical laboratory and they are used to study human breast cancer cells undergoing treatment of different drugs.

To evaluate our proposed algorithm, we start from applying different clustering algorithms (K-means, Hierarchical agglomerative and Affinity propagation) to the UCI data sets "Ionoshpere" and "Balance" and the biomedical laboratory data sets "3ClassesTest1", "4ClassesTest1" and "5ClassesTest1". The micro-precisions of these algorithms are listed in Table 3. For the comparison purpose, we also list the average micro-precisions of existing clustering ensemble methods reported in (Wang et al., 2011). We only list the ensemble method that performs best on each data set. We apply the clustering ensemble method (MCLA) proposed in (Strehl and Ghosh, 2003) to the biomedical laboratory data sets.

Suppose p% represents the ratio of the number of training points (N_r) to the number of testing points (N_u). To study the effect of the amount of training data to the semi-supervised method, we vary the values of p from $\{3,5,10,15,20,25,30\}$. The performance of our propose semi-supervised method is listed in Table 4. Compared with individual clustering algorithms (K-means, HAC and AP), our proposed algorithm outperforms on the data sets listed in Table 3. Compared with existing ensemble methods, our proposed algorithm also outperforms these data sets. The micro-precisions increase dramatically when p is relatively small and become steady when $p > 15\%$. Therefore, due to the fact that it is expensive and time-consuming to obtain labels from field experts, there is no need to make effort on increasing the amount of training data because the improvement of the accuracy of the semi-supervised method is not always increased by increasing the number of training points.

The biomedical data sets are obtained from the study of human breast cancer cells undergoing treatment of different drugs. When a certain type of drug is injected into cancer cells, the cells usually react differently: a portion of the cells may slightly react to the injected drug (such as slightly enlarged); another portion of the cells may react strongly (such as loss of nucleus); and the rest may not react to the injected drug at all. For those cells that strongly react to the injected drug, it is very likely that their statistical properties vary significantly and they can form a new cluster. Therefore, in the study of the effect of a certain drug to cancer cells, we could apply our proposed new cluster detection algorithm to automatically detect the existence of cancer cells that strongly react to the injected drug.

We provide numerical examples to show the performance of the proposed new cluster detection algorithm. The original test files contains data observations from different classes. Each original test file has a fixed amount of training data. To evaluate the new cluster detection algorithm, we insert additional data points to the original test files and vary the number of additional data points. To evaluate the probability of successful detection of a new cluster, we insert a mixture of data points from a new class and from existing classes and vary the proportion of the data points from a new class. For each original test file and a particular number of additional points, we randomly generate 20 versions of additional data set $\mathbf{X}_a$ using one of the mixture proportions listed in Table 6. The number of total successful detections of a new cluster are provided in Table 5. As expected, the probability of successful detection of a new cluster using the proposed algorithm goes higher when the number of data points from a new class increases.

5 CONCLUSIONS

Since clustering is a more general problem such that no categories/clusters are pre-defined for the clustering algorithms, the fusion of multiple clusterings is more difficult due to the so-called correspondence problem. In this paper, we have proposed the semi-supervised clustering ensemble algorithms to combine multiple clusterings by relabelling the cluster labels according to the training clusters. We presented numerical examples to demonstrate the capability of the proposed algorithms on improving the quality of cluster analysis. The improvement in terms of accuracy of the clustering results depends on the statistical properties of the data set and also depends on the amount of available reference labels. When additional observations become available, we need to determine whether the training data is sufficient for the new observations. Therefore, we have proposed the new cluster detection algorithm to detect the event that new observations come from a new class other than existing training classes. We provided numerical examples to show that the proposed algorithm is capable to detect a new cluster when the number of new observations, not from existing classes, is accumulated to a certain level.

Table 3: Average micro-precisions of different clustering algorithms and existing ensemble methods.

Data Sets	No. of			Clustering Algorithms			Ensemble methods	
	Points	Classes	Features	Kmeans	HAC	AP	Average	Method
Ionosphere	351	2	34	0.7123	0.7182	0.7107	0.7141	BCE
Balance	625	3	4	0.5221	0.5074	0.4834	0.5552	MM
3ClassesTest1	542	3	705	0.4469	0.4299	0.4871	0.4989	MCLA
4ClassesTest1	717	4	705	0.4547	0.3501	0.4923	0.4505	MCLA
5ClassesTest1	916	5	705	0.4004	0.4323	0.4116	0.3895	MCLA

Table 4: Average micro-precisions of the proposed semi-supervised clustering ensemble methods.

Data sets	3%	5%	10%	15%	20%	25%	30%
Ionoshpere	0.7586	0.8047	0.8716	0.8594	0.8662	0.8760	0.8644
Balance	0.6048	0.6251	0.6770	0.6772	0.6692	0.6849	0.6964
3ClassesTest1	0.6351	0.6123	0.6530	0.6825	0.6900	0.7032	0.6868
4ClassesTest1	0.5424	0.5547	0.5869	0.6096	0.6334	0.6150	0.6302
5ClassesTest1	0.4277	0.4205	0.4619	0.4933	0.4902	0.4815	0.4812

Table 5: Number of successful detections of a new cluster when different amount of additional data points are added to the original data sets.

P = 15%	No. of data points		No. of detections of a new cluster					Total Success
	Original	Added	Type 1 (/20)	Type 2 (/20)	Type 3 (/20)	Type 4 (/20)	Type 5 (/20)	(/100)
2ClassesAdd1	257	50	20	17	17	12	3	83
		100	20	20	17	15	1	91
		150	20	20	19	15	2	92
3ClassesAdd1	518	50	20	15	12	9	2	74
		100	20	18	18	8	1	83
		150	20	20	19	9	3	85

Table 6: Mixing proportion of data points from a new class and existing classes.

Mixing Proportion	Proportion of data points from	
	A new class	Existing classes
Type 1	1	0
Type 2	2/3	1/3
Type 3	1/2	1/2
Type 4	1/3	2/3
Type 5	0	1

REFERENCES

Bache, K. and Lichman, M. (2013). UCI machine learning repository.

Ghaemi, R., Sulaiman, M. N., Ibrahim, H., and Mustapha, N. (2009). A survey: clustering ensembles techniques. *World Academy of Science, Engineering and Technology*, 50:636–645.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Li, H. and Jeremić, A. (2017). Semi-supervised distributed clustering for bioinformatics - comparison study. In *BIOSIGNALS 2017*, pages 649–652.

Likas, A., Vlassis, N., and Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461.

Strehl, A. and Ghosh, J. (2003). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617.

Vega-Pons, S. and Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03):337–372.

Wang, H., Shan, H., and Banerjee, A. (2011). Bayesian cluster ensembles. *Statistical Analysis and Data Mining*, 4(1):54–70.

Xu, R., Wunsch, D., et al. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678.

Yan, D., Huang, L., and Jordan, M. I. (2009). Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 907–916. ACM.

Zhang, K. and Gu, X. (2014). An affinity propagation clustering algorithm for mixed numeric and categorical datasets. *Mathematical Problems in Engineering*, 2014.

Zhou, Z.-H. and Tang, W. (2006). Clusterer ensemble. *Knowledge-Based Systems*, 19(1):77–83.