Clustering Ensemble Method for Heterogeneous Partitions

Sandro Vega-Pons and José Ruiz-Shulcloper

Advanced Technologies Application Center (CENATAV), Havana, Cuba
{svega,jshulcloper}@cenatav.co.cu

Abstract. Cluster ensemble is a promising technique for improving the clustering results. An alternative to generate the cluster ensemble is to use different representations of the data and different similarity measures between objects. This way, it is produced a cluster ensemble conformed by heterogeneous partitions obtained with different point of views of the faced problem. This diversity enhances the cluster ensemble but, it restricts the combination process since it makes difficult the use of the original data. In this paper, in order to solve these limitations, we propose a unified representation of the objects taking into account the whole information in the cluster ensemble. This representation allows working with the original data of the problem regardless of the used generation mechanism. Also, this new representation is embedded in the WKF [1] algorithm making a more robust cluster ensemble method. Experimental results with numerical, categorical and mixed datasets show the accuracy of the proposed method.

Keywords: Cluster ensemble, object representation, similarity measure, co-association matrix.

1 Introduction

Cluster ensemble has emerged as a good alternative to improve the quality of clustering results. Traditionally, given a set of objects, a cluster ensemble method consists in two principal steps: Generation, which is about the conformation of a set of partitions of these objects, and Consensus Function, where a new partition which is the *integration* of all partitions obtained in the generation step, is computed.

In the generation step, different representations of the objects can be used or the same representation with different similarity (dissimilarity) measures to obtain each partition in the cluster ensemble. Then, if it is necessary to work with the original data after the generation step, we have to deal with the following questions: Which representation of the objects should be used? Which similarity (dissimilarity) measure should be applied?

To the best of the authors knowledge, these questions have not been boarded before. Taking into account these situations, and giving them an adequate treatment, we can improve the quality of the clustering ensemble algorithms and

E. Bayro-Corrochano and J.-O. Eklundh (Eds.): CIARP 2009, LNCS 5856, pp. 481–488, 2009.

enlarge their scope. Then, the main goal of this paper is to give an answer to these questions. In order to do that, we present a new way of representing the objects unifying the information in all partitions in the cluster ensemble. This representation is based on a new similarity matrix, which is obtained from the co-association of the objects in the clusters of the set of partitions, but taking into account more information than the traditional co-association matrix [2] and therefore expressing better the relationship between objects. By using this new representation of the objects can be applied, for example, any distance function and compute centroids in this new representation space, even when the original data are mixed (composed by numerical and non-numerical attributes).

This paper is organized as follows: In section 2 some related works are presented, highlighting the method proposed in [1] called WKF, as well as the motivation and problem description are outlined. Section 3 introduces our proposal and presents a modification of the WKF method with our new object representation embedded in the steps of this method. Experimental results are discussed in Section 4 and finally in Section 5 are the conclusions of our research.

2 Problem Discussion

We will denote $X = \{x_1, x_2, \ldots, x_n\}$ the set of objects, where each object is a tuple of some $d-$dimensional feature space Ω^d. $\mathbb{P}=\{P_1, P_2, \ldots, P_m\}$ is a set of partitions, where each $P_i = \left\{C_i^1,\ C_i^2, \ldots,\ C_i^{k_i}\ \right\}$ is a partition of the set of objects X, and C_i^j is the j^{th} cluster of the i^{th} partition, for all $i = 1, \ldots, m$. The goal of clustering ensemble methods is to find a partition P^*, which better represents the properties of each partition in $\mathbb{P}$.

Several clustering ensemble methods have been proposed in the literature, e.g. Co-association methods [2] and Hyper-Graph methods [3]. In these methods, it is not necessary to work with the original data after the generation step, i.e., once the set of partitions $\mathbb{P}$ is obtained, all the operations to obtain the consensus partition P^* are achieved taking into account only the partitions in $\mathbb{P}$.

For example, the co-association methods firstly compute the co-association matrix $\mathcal{C}$, where each cell has the following value:

$$\mathcal{C}_{ij} = \frac{1}{m} \sum_{k=1}^{m} \delta\left(P_k\left(x_i\right), P_k(x_j)\right) \tag{1}$$

$P_k\left(x_i\right)$ represents the associate label of the object x_i in the partition P_k, and $\delta\left(a, b\right)$ is 1, if $a = b$, and 0 otherwise. Then, the value in each position (i, j) of this matrix is a measure about how many times the objects x_i and x_j are in the same cluster for all partitions in $\mathbb{P}$. Using the co-association matrix $\mathcal{C}$ as the similarity measure between objects, the consensus partition is obtained by applying a clustering algorithm.

The Hyper-graph methods start by representing each partition in the cluster ensemble with a hyper-edge. Then, the problem is reduced into a hyper-graph

partitioning problem. Three efficient heuristics CSPA, HGPA and MLCA are presented in [3].

However, the set of objects X and their similarity (dissimilarity) values are additional information that, if it is well-used, the combination results can be improved. In other words, more complex methods that make use of this information in order to achieve better results can be developed. This is the case of the WKF method [1].

The WKF method uses the set of objects X and their similarities in an intermediate step, between the generation and the consensus function called Qualitative Analysis of the Cluster Ensemble (QACE). In this new step, it is estimated the relevance of each partition for the cluster ensemble.

The idea is to assign a weight to each partition that represents how relevant it is in the cluster ensemble. In this step, partitions which represent noise for the cluster ensemble are detected, and its effect in the consensus partition is minimized. The impact of this step in the final consensus partition is meaningful, since by using the information obtained in this step, a fairer combination process is achieved. In [1] the QACE step is performed by applying different cluster validity indexes, where each one of them measures the fulfillment of a particular property, e.g., compactness, separability, connectivity. Thus, to a partition that behaves very different to the rest of the partitions with respect to this properties, it is assigned a small weight, because it is probably a noisy partition obtained by a not appropriate generation mechanism. Otherwise, if a partition has an average behavior, it will have a higher weight assigned.

The consensus partition in the WKF method is computed by using a consensus function able to work with the weights computed in the QACE step. For this, each partition is represented by a graph. Furthermore, to obtain an exact representation of the consensus into a Hilbert Space, a kernel function is used as the similarity measure. Finally, an efficient stochastic search strategy is applied to obtain the final consensus partition.

Despite of the advantages that the QACE step offers, it has the limitation that the similarity between the objects must be computed on the original data X. This may cause the appearance of some problems such as:

1. The partitions could be created by using different representations of the data, either by completely different representations given by different modelings of the problem, or the same representation, but using different subset of features to obtain the partitions. Then, which representation of the data should be used in the QACE step?
2. Besides, the partitions in the cluster ensemble could be obtained by using different similarity (dissimilarity) measures but, which one should be used in the QACE step? We would possibly also need to compute a distance between objects in this step but, what can be done if we have not a distance defined for our set of objects?

These problems are more complicated when the original data is mixed, because it is difficult to apply cluster validity indexes to the partitions since it is difficult to

embed the set of objects into a metric space. In this paper, we propose a solution for these questions and it is incorporated in the steps of the WKF algorithm.

3 Generalized WKF

Firstly, we say that the fact that two objects belong to the same cluster in a partition does not contribute with the same information for every partitions in the cluster ensemble. For that reason, we will define the *co-association significance* of two objects x_i and x_j that belong to the same cluster in some partition $P \in \mathbb{P}$. To compute this significance, we will take three factors into account: the number of elements in the cluster to which x_i and x_j belong, the number of clusters in the partition analyzed and the similarity (dissimilarity) of this objects by using the same proximity measure used to conform the partition P. Then, we say that two objects x_i and x_j, grouped in a cluster C of some partition $P \in \mathbb{P}$, which was obtained using the similarity measure Γ_P, have a high co-association significance if the following conditions are satisfied:

1. $|C|$ is small ($|C|$ is the number of elements in the cluster C)
2. $|P|$ is large ($|P|$ is the number of clusters in P)
3. $\Gamma_P(x_i, x_j)$ is large.

If the partition P was obtained by using a dissimilarity measure d_P, we can easily obtain an equivalent similarity measure Γ_P by $\Gamma_P = \frac{1}{d_P+1}$. Then from now on, we assume that the clustering algorithm applied to generate the partition P used the similarity measure Γ_P. Formally, we can define the co-association significance as:

Definition 1. *Given two objects x_i, x_j and a partition P, we define the* co-association significance *of these objects in the partition P as:*

$$CS^P(x_i, x_j) = \begin{cases} \frac{|P|}{|C|} \cdot \Gamma_P(x_i, x_j), & if\ \exists C \in P, such\ that\ x_i \in C, x_j \in C; \\ 0, & otherwise \end{cases}$$

which represents the significance that two objects x_i and x_j had been grouped together in the same cluster or not, in the partition P.

Taking into account the co-association significance of each pair of objects in X, in all partition in $\mathbb{P}$, it is conformed the *Weighted Co-Association Matrix* denoted by WC, where the (i, j) entry of the matrix has the following value:

$$WC_{i,j} = \sum_{k=1}^{m} CS^{P_k}(x_i, x_j) \tag{2}$$

The WC matrix is more expressive than the traditional co-association matrix (1), because in the co-association matrix it is only taken into account if the objects are together or not in the same cluster but, the rest of the information given by the partition is not considered. Let us see an illustrative example:

Example 1. Let $\mathbb{P}_X$ be the set of all possible partitions of the set X. We can define the order relationship *nested in* denoted by $\preceq$, where $P \preceq P'$ if and only if, for all cluster $C' \in P'$ there are clusters $C_{i_1}, C_{i_2}, \ldots, C_{i_k} \in P$ such that $C' = \bigcup_{j=1}^{k} C_{i_j}$. The hierarchical clustering algorithms, like the Single Link and Average Link, produce sequences of nested partitions where, if $P \preceq P'$ means that the criterion used to obtain P is stronger than the used to obtain P'. Then, the fact that a pair of objects belong to the same cluster in P, gives more information about the likeness of these objects than the fact that they had been grouped in the same cluster in the partition P'. Using the traditional co-association matrix (1) this information can not be extracted. However, by using the weighted co-association (2) more significance is given to the partition P since, if $P \preceq P'$ implies that $|P| \geq |P'|$ and $|C| \leq |C'|$. Then $\frac{|P|}{|C|} \cdot \Gamma_P(x_i, x_j) \geq \frac{|P'|}{|C'|} \cdot \Gamma_{P'}(x_i, x_j)$ because in this case $\Gamma_P = \Gamma_{P'}$.

Once we have the matrix WC, it can be considered as a new representation of the objects, where each object $x_i \in X$ is represented by a vector of $\mathbb{R}^n$, $x_i = \{WC_{i,1}, WC_{i,2}, \ldots, WC_{i,n}\}$. The representation by dis(similarities) is extensively studied in [4]. This way, all the information about the possible different representations and proximity measures, used in the generation step, are summarized in the new representation of the objects. Even, when only one representation of the set of objects, and only one similarity measure in the generation step are used, this new representation can give advantages, e.g. when the original data is mixed. In this case, the new representation as a vector of $\mathbb{R}^n$ allows the use of the mathematical tools for vectorial spaces that have not to be available for the original data representation.

In [5], a comparison of different cluster ensemble techniques is achieved. Among other techniques, the co-association matrix (1) is used as a new representation of the objects, obtaining the best results. Then, as an alternative, the new weighted co-association matrix can be used instead of the traditional co-association matrix in the co-association cluster ensemble methods [2]. The results should be better since the WC matrix has more information about the real relationship between the objects in the set of partitions $\mathbb{P}$.

However, the main goal of the construction of the matrix WC is for obtaining a new representation of the objects that allows to use the WKF method without any constraint in the generation step.

3.1 Steps of the Generalized WKF Algorithm

In order to embed our object representation into the WKF methodology, it is necessary to incorporate an intermediate step where the matrix WC is computed and used as a new representation of the objects. We call the algorithm with this modification as GWKF and its principal steps are:

1. Given a set of objects X, generate a set of partitions $\mathbb{P}$ of these objects, by using different clustering algorithms, different initialization of the parameters, even using different representation of the objects, and different similarity (dissimilarity) measures to obtain each partition.

2. With the information in $\mathbb{P}$, compute the *weighted co-association matrix* (2), and use it as a new representation of the set of objects X.
3. Apply the QACE, where any kind of indexes can be used without taking into account the original type of data of the problem or the way that the partitions in the cluster ensemble were generated. The indexes are applied by using the new data representation and can be used any distance function (e.g., the Euclidean distance).
4. Compute an associated weight to each partition by using the indexes defined previously.
5. Apply the consensus function as in [1]. This consensus function automatically selects the appropriate number of clusters in the consensus partition.

4 Experimental Results

In the experiments, we used 7 datasets from the UCI Machine Learning Repository [8]. The characteristics of all datasets are given in Table 1. We denote our method (given by the steps in the previous section) by GWKF. Also, in order to apply the QACE step, we use three simple indexes [6]: *Variance*, *Connectivity* and *Dunn Index*. Each one of them measures the fulfillment of a specific property. The *Variance* is a way to measure the compactness of the clusters in the partition. The *Connectivity* evaluates the degree of connectedness of clusters in the partition, by measuring how many neighbors of one object belong to the same cluster that the object. The *Dunn Index* is a measure of the ratio between the smallest inter-cluster distance and the largest intra-cluster distance.

The three indexes use a distance function defined over the set of objects X. Then, if it is used the WKF algorithm, these indexes can not be applied to a dataset for which there is not defined an appropriate distance function. However, by using the GWKF, we can apply this indexes to any dataset without taking into account the type of data because, after the generation step the objects are represented as vectors of $\mathbb{R}^n$ by using the Weighted Co-Association Matrix (2). After that, any distance function can be applied, we use the Euclidean distance.

Table 1. Overview of datasets

Name	**Type**	**#Inst.**	**#Attrib.**	**#Classes**	**Inst-per-classes**
Iris	Numerical	150	4	3	50-50-50
Digits	Numerical	100	64	10	10-11-11-11-12-5-8-12-9-11
Breast-Cancer	Numerical	683	9	2	444-239
Zoo	Mixed	101	18	7	41-20-5-13-4-8-10
Auto	Mixed	205	26	7	0-3-22-67-54-32-27
Soybeans	Categorical	47	21	4	10-10-10-17
Votes	Categorical	435	16	2	267-168

4.1 Analysis of the Experimental Results

Firstly, we show the behavior of our method (GWKF) in numerical datasets and we compare the results with several clustering ensemble methods (see Table 2). EA-SL and EA-CL are the co-association methods [2] using the Single-Link and Complete-Link algorithm, CSPA, HPGA and MCLA are the three hypergraph methods proposed in [3]. In this experiment, it is generated the cluster ensemble always using the same representation of the objects, and by applying the k-Means algorithm 20 times with different parameters initialization. This experiment shows the good performance of the GWKF method in comparison with the other cluster ensemble methods and how the final results of the GWKF method are very close to the results of the WKF method.

Table 2. Clustering error rate of different clustering ensemble methods

Dataset	Ens(Ave)	EA-SL	EA-AL	CSPA	HPGA	MCLA	WKF	GWKF
Iris	18.1	11.1	11.1	13.3	37.3	11.2	**10.6**	10.8
Breast-Cancer	3.9	4.0	4.0	17.3	49.9	3.8	**3.7**	**3.7**
Digits	27.4	56.6	23.2	**18.1**	40.7	18.5	22.1	20.6

The WKF method can not be applied if the data is mixed or in the generation step were used different representations of the objects. In these cases, the GWKF method gains more importance, because it is able to deal with any type of data and any kind of generation mechanism keeping the good performance of the WKF method.

On the other hand, in Table 3, we compare the GWKF with simple clustering algorithms (in this case the k-Means) in two different mixed datasets with different ensemble sizes (H). In this experiment, the partitions are obtained by using random subset of features and applying the k-Means algorithm with a fixed number of clusters (k). The results show that our algorithm obtains lower errors rate than the average error rate of the k-Means algorithm.

Table 3. Cluster ensemble average error rate and GWKF error rate in mixed datasets

Dataset	H	k	Ensemble(Ave.)	GWKF
Zoo	10	7	26.8	**20.7**
Zoo	20	7	24.3	**19.1**
Auto	10	7	62.0	**57.3**
Auto	20	7	61.3	**54.5**

Finally, we compare our method with 4 algorithms proposed in [7] (CSPA-METIS, CSPA-SPEC, CSBA-METIS and CSBA-SPEC), all of them designed to work with categorical data. In this experiment, we use 2 categorical datasets, and the partitions were generated by using different sets of random features (see Table 4). As in the previous experiment, the original WKF method is not able to

Table 4. Clustering error rate using categorical data

Dataset	k	Ens(Ave)	CSPA-ME	CSPA-SP	CBPA-ME	CBPA-SP	GWKF
Votes	2	13.7	14.0	13.5	14.0	14.2	**11.4**
Soybeans	4	24.4	10.6	12.3	12.8	15.3	**6.3**

work with this generation mechanism, this dataset and the indexes defined for the experiments. However, the GWKF method extends the good performance of the WFF method to this kind of situations.

5 Conclusions

In this paper, the problem about what representation of the objects and what similarity measure should be used in the consensus step, in the cases that many representations and similarity measures are used in the generation step is formulated and solved. A new object representation is proposed, which summarizes the whole information in the set of partitions. This is possible thanks to the definition of a new way of measuring the co-association between objects, which is more expressive about the real similarity between objects in the cluster ensemble than the classical co-association. The new representation is embedded in the WKF method enlarging its scope and obtaining the Generalized WKF method. The experiments with numerical, categorical and mixed datasets respectively and by using different generation mechanisms, show the good performance of the proposed method.

References

[1] Vega-Pons, S., Correa-Morris, J., Ruiz-Shulcloper, J.: Weighted cluster ensemble using a kernel consensus function. In: Ruiz-Shulcloper, J., Kropatsch, W.G. (eds.) CIARP 2008. LNCS, vol. 5197, pp. 195–202. Springer, Heidelberg (2008)
[2] Fred, A.L.N., Jain, A.K.: Combining multiple clustering using evidence accumulation. IEEE Trans. on Pat. Analysis and Machine Intelligence 27, 835–850 (2005)
[3] Strehl, A., Ghosh, J.: Cluster ensembles: a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. 3, 583–617 (2002)
[4] Pekalska, E., Duin, R.P.W.: The Dissimilarity Representation for Pattern Recognition: Foundations And Applications. In: Machine Perception and Artificial Intelligence. World Scientific Publishing Co., Singapore (2005)
[5] Kuncheva, L., Hadjitodorov, S., Todorova, L.: Experimental comparison of cluster ensemble methods. In: Int. Conference on Information Fusion, pp. 1–7 (2006)
[6] Handl, J., Knowles, J., Kell, D.: Computational cluster validation in post- genomic data analysis. Bioinformatics 21, 3201–3212 (2005)
[7] Al-Razgan, M., Domeniconi, C.: Random subspace ensembles for clustering categorical data. Studies in Computational Intelligence (SCI) 126, 31–48 (2008)
[8] UCI machine learning repository, `http://archive.ics.uci.edu/ml/datasets.html`