Learning Similarities by Accumulating Evidence in a Probabilistic Way

Helena Aidos and Ana Fred

Instituto de Telecomunicações, Instituto Superior Técnico,
Universidade de Lisboa, Lisbon, Portugal
{haidos,afred}@lx.it.pt

Abstract. Clustering ensembles take advantage of the diversity produced by multiple clustering algorithms to produce a consensual partition. Evidence accumulation clustering (EAC) combines the output of a clustering ensemble into a co-association similarity matrix, which contains the co-occurrences between pairs of objects in a cluster. A consensus partition is then obtained by applying a clustering technique over this matrix. We propose a new combination matrix, where the co-occurrences between objects are modeled in a probabilistic way. We evaluate the proposed methodology using the dissimilarity increments distribution model. This distribution is based on a high-order dissimilarity measure, which uses triplets of nearest neighbors to identify sparse and odd shaped clusters. Experimental results show that the new proposed algorithm produces better and more robust results than EAC in both synthetic and real datasets.

Keywords: Clustering ensembles, co-association matrix, voting scheme, probablistic learning of similarities, dissimilarity increments distribution.

1 Introduction

Many clustering algorithms have been developed, each producing a different partition for a given dataset, and typically relying on a similarity measure between objects, which can be difficult to choose when no prior knowledge about cluster shapes and structure is available. Furthermore, one single clustering algorithm, with a given similarity measure, can also produce different solutions for the same dataset, depending on the initialization or parameters values, *e.g.*, k-means.

To exploit that diversity, an approach called *clustering ensemble* (CE) has been developed [13,10,3], producing a set of data partitions. These methods combine information given by the set of data partitions produced, and propose a consensus partition. Moreover, it has been shown that CE methods uncover a more robust and stable cluster structure than a single clustering algorithm [6,13]. To combine information from the set of data partitions, different paradigms were followed: (i) similarity between objects, induced by the clustering ensemble [6,13,7]; (ii) similarity between partitions [4,3]; (iii) combining similarity between objects and partitions [5]; (iv) probabilistic approaches to CEs [14,15].

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 596–603, 2014.

This paper focuses on a clustering ensemble technique, namely the *evidence accumulation clustering* (EAC) [7], which combines the results of multiple clusterings into a co-association matrix, corresponding to co-occurrences of pairs of objects in a cluster. This co-occurrence between pairs of objects means that two objects that are very similar, more likely will be grouped together. Therefore, the co-association matrix can be seen as a similarity measure and the final partition of data is obtained by applying a clustering algorithm over this matrix.

We propose a new combination matrix, where each element of the matrix corresponds to the probability of observing a pair of objects co-occurring in a cluster. That co-occurrence can be modeled by any probabilistic model, however we use the dissimilarity increments distribution (DID) [2]. The DID is a probabilistic model for the dissimilarity increments measure [9], which uses the information from triplets of nearest neighbors. This measure identifies the structure of a cluster, in terms of sparsity and shape of clusters.

2 Related Work

The **Evidence Accumulation Clustering** (EAC) [7] is a three step methodology, consisting of: (i) building the clustering ensemble (CE); (ii) learning pairwise similarities by accumulating evidence in a matrix; and (iii) extracting the consensus partition using a clustering algorithm.

A CE can be obtained by applying different clustering algorithms over data, or one algorithm with different initializations or parameters of the same algorithm. Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be a set of N objects and $\mathcal{P}^i = \{C_1^i, C_2^i, \ldots, C_{k_i}^i\}$ a data partitioning into k_i clusters obtained by applying a given clustering algorithm i. A *clustering ensemble*, $\mathbb{P} = \{\mathcal{P}^1, \mathcal{P}^2, \ldots, \mathcal{P}^M\}$, is a set of M different partitions of the data X.

The *evidence accumulation* approach consists of learning similarities between pairs of objects induced by the clustering ensemble, since objects that are similar and should be grouped together are probably going to be assigned to the same cluster in different data partitions. Equivalently, we count the co-occurrences of pairs of objects in the same cluster, n_{ij}, among the M partitions, yielding a $N \times N$ co-association matrix:

$$\mathbb{C}(i, j) = \frac{n_{ij}}{M}, \tag{1}$$

Finally, the consensus partition is found by applying a clustering algorithm to the co-association matrix.

On the other hand, each element of the co-association matrix in eq. (1) can be viewed as an independent realization of binomial random variables counting the number of times two objects occur in the same cluster [11].

3 Combination of Evidence: A Probabilistic Point of View

We can interpret each element of the co-association matrix as the probability of observing a pair of objects, $\mathbf{x}_i$ and $\mathbf{x}_j$, in the same cluster for a given partition $\mathcal{P}^l$.

Therefore, in this paper, we formulate the co-association matrix in a probabilistic way. Formally, an element of matrix $\mathbb{C}$ is given by the probability of observing $\mathbf{x}_i$ and $\mathbf{x}_j$ in the same cluster among the M partitions of the ensemble, *i.e.*,

$$\mathbb{C}(i,j) \equiv P(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{M} P(\mathbf{x}_i, \mathbf{x}_j, \mathcal{P}^l) = \sum_{l=1}^{M} P(\mathbf{x}_i, \mathbf{x}_j | \mathcal{P}^l) P(\mathcal{P}^l). \tag{2}$$

The term $P(\mathcal{P}^l)$ allows giving different weights to each partition, depending on how well a partition fits the data, creating a weighted version of this approach. Here it is assumed uniform, *i.e.*, each partition contributes equally to the matrix $\mathbb{C}$ ($P(\mathcal{P}^l) = {}^1\!/_M$). The term $P(\mathbf{x}_i, \mathbf{x}_j | \mathcal{P}^l)$ can be described by any probabilistic model of observing $\mathbf{x}_i$ and $\mathbf{x}_j$ belonging to the same cluster in partition $\mathcal{P}^l$.

Depending on the model assumed for the observed variables, one can obtain an asymmetric matrix or a symmetric matrix. Typical hierarchical clustering algorithms can be used to extract the consensus partition, however usually they can only be applied if the matrix is symmetric. If matrix $\mathbb{C}$ is asymmetric one must use clustering algorithms appropriate for asymmetric measures, like spectral clustering [12], or one can transform $\mathbb{C}$ into a symmetric matrix, *e.g.*, by computing the average ($\mathbb{C} = (\mathbb{C} + \mathbb{C}^T)/2$).

Here we consider the dissimilarity increments distribution (DID) [2] as the model of observing two objects in the same cluster, leading to a matrix called *DID-based consensus matrix*. This new matrix combines the probability of a triplet of objects being in the same cluster. This approach is quite similar to the voting scheme proposed by Fred and Jain [7], but the voting scheme combines in a matrix evidence of pairs of objects belonging to the same cluster, here we combine in a matrix evidence of triplets of objects belonging to the same cluster.

3.1 Dissimilarity Increments Distribution

Let X be a dataset and $d(\cdot,\cdot)$ some dissimilarity measure between objects. Consider $\mathbf{x}_i$, an object from X. A triplet of nearest neighbors, $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$, is obtained by searching $\mathbf{x}_j$ and $\mathbf{x}_k$, corresponding to the nearest neighbor of $\mathbf{x}_i$, and the nearest neighbor of $\mathbf{x}_j$ different from $\mathbf{x}_i$, respectively. Thus, the **dissimilarity increment** [9] associated with the triplet is defined as

$$d_{\mathrm{inc}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) = \left| d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{x}_j, \mathbf{x}_k) \right|. \tag{3}$$

This measure can be applied to characterize the structure of data, in terms of sparse clusters and different shapes and densities, since abrupt changes of dissimilarity increments should not occur inside a cluster, and higher values of this measure occur between well separated clusters. Therefore, the increment between objects belonging to different clusters is positioned on the tail of the dissimilarity increments distribution of a cluster.

The **dissimilarity increments distribution** (DID) was derived in [2], assuming a Gaussian distribution of the data, and that $d(\cdot,\cdot)$ is the Euclidean distance. Accordingly, let $w = d_{\text{inc}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$ be a dissimilarity increment, defined in $[0, \infty)$. Thus, the probability density function (pdf) of w is given by

$$\begin{aligned} p_{d_{\text{inc}}}(w; \lambda) = & \frac{\pi\beta^2}{4\lambda^2} w \exp\left(-\frac{\pi\beta^2}{4\lambda^2} w^2\right) \\ & + \frac{\pi^2\beta^3}{8\sqrt{2}\lambda^3}\left(\frac{4\lambda^2}{\pi\beta^2} - w^2\right) \exp\left(-\frac{\pi\beta^2}{8\lambda^2} w^2\right) \operatorname{erfc}\left(\frac{\sqrt{\pi}\beta}{2\sqrt{2}\lambda} w\right), \end{aligned} \quad (4)$$

where erfc(·) is the complementary error function, $\beta = 2 - \sqrt{2}$, and λ is the parameter of the distribution corresponding to the mean of the set of increments inside a cluster, $S_{d_{\text{inc}}}(C)$, given by $\mathbb{E}[S_{d_{\text{inc}}}(C)] \equiv \lambda$. This parameter has influence in the distribution, since the smaller the values of λ, the narrower is the pdf, indicating that the data is more dense. On the other hand, the higher the values of λ, the wider is the pdf, corresponding to sparser data.

3.2 DID-Based Probability Matrix

We propose using the DID to model the probability of observing $\mathbf{x}_i$ and $\mathbf{x}_j$ belonging to the same cluster in the following way: consider a partition of data, $\mathcal{P}^l = \{C_1^l, \ldots, C_{k_l}^l\}$ from the clustering ensemble. Firstly, we obtain the set of 1-nearest neighbor of each object inside a cluster and the set of increments $S_{d_{\text{inc}}}(C_m^l)$, with $m = 1, 2, \ldots, k_l$. This set of increments is obtained by constructing a minimum spanning tree inside the cluster C_m^l and applying eq. (3).

Now, for each object $\mathbf{x}_i$ inside the cluster C_m^l, we find the dissimilarity increment $w = d_{\text{inc}}(\mathbf{x}_i, \mathbf{x}_j, NN(\mathbf{x}_j))$, where $NN(\mathbf{x}_j)$ is the nearest neighbor of $\mathbf{x}_j$. Thus, the probability of observing $\mathbf{x}_i$ and $\mathbf{x}_j$ in the same cluster is given by

$$P(\mathbf{x}_i, \mathbf{x}_j | \mathcal{P}^l) = \frac{p_{d_{\text{inc}}}(w; \lambda_{C_m^l})}{p_{d_{\text{inc}}}(0; \lambda_{C_m^l})}, \quad (5)$$

where $\lambda_{C_m^l}$ is the mean of increments of cluster C_m^l, and $p_{d_{\text{inc}}}(w; \lambda_{C_m^l})$ is the pdf for the dissimilarity increments given by eq. (4). The term $p_{d_{\text{inc}}}(0; \lambda_{C_m^l})$ is used to normalize the pdf of clusters with different shapes and densities. The overall procedure is summarized in Algorithm 1. Notice that the new consensus matrix as defined by eq. (2) is asymmetric and can be seen as the similarity between objects i and j.

3.3 DID-Based Probabilistic Consensus Clustering Algorithm

Similar to EAC, the proposed algorithm is a three step methodology, where the difference is in the construction of the consensus matrix. While EAC is based on a voting scheme, the proposed algorithm, called DID-based Probabilistic Consensus Clustering (PCC$^{\text{DID}}$), uses a probabilistic model based on DID of observing pairs of objects as belonging to the same cluster in a partition of the ensemble. The overall procedure is summarized in Algorithm 2.

Algorithm 1. DID-based consensus matrix (DIDCM)

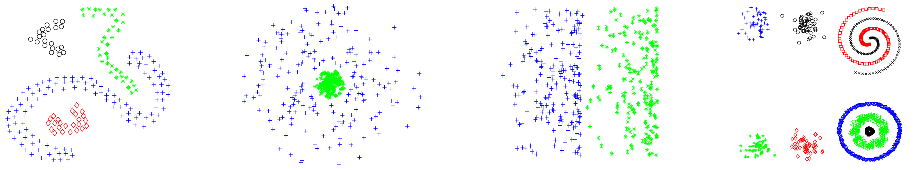
```
Require: data with $N$ samples
Require: partition of data $\mathcal{P} = \{C_1, \ldots, C_K\}$
Require: consensus matrix $\mathbb{C}$
for each cluster $C_k$ do
    if $|C_k| >= 3$ then
        For all $\mathbf{x}_j$ inside $C_k$, find the 1-nearest neighbor (also inside $C_k$), $NN(\mathbf{x}_j)$
        Get $S_{d_{\mathrm{inc}}}(C_k)$ using a minimum spanning tree
        Compute $\lambda_{C_k}$ and $p_{d_{\mathrm{inc}}}(0; \lambda_{C_k})$, using (4)
        for each $\mathbf{x}_i \in C_k$ do
            Find $w = d_{\mathrm{inc}}(\mathbf{x}_i, \mathbf{x}_j, NN(\mathbf{x}_j))$, for all $\mathbf{x}_j \in C_k$
            Compute $p_{d_{\mathrm{inc}}}(w; \lambda_{C_k})$, using (4)
            /* Update consensus matrix $\mathbb{C}$. */
            $\mathbb{C}(i, j) \leftarrow \mathbb{C}(i, j) + p_{d_{\mathrm{inc}}}(w; \lambda_{C_k}) / p_{d_{\mathrm{inc}}}(0; \lambda_{C_k})$
        end for
    end if
end for
```

Algorithm 2. DID-based Probabilistic Consensus Clustering ($\mathrm{PCC}^{\mathrm{DID}}$)

```
Require: data with $N$ samples
$\mathbb{C} = \mathrm{zeros}(N, N)$
for $i = 1$ to $M$ do
    /* Step 1: Build the clustering ensemble. */
    Obtain partition $\mathcal{P}^i$ by applying clustering algorithm(s)
    /* Step 2: Compute the consensus matrix. */
    $\mathbb{C} \leftarrow \mathbb{C} + \mathrm{DIDCM}(\mathbb{C}, \mathcal{P}^i) \times {}^1/_M$, using (2)
end for
/* Step 3: Extract the consensus partition */
Apply a clustering algorithm to the consensus matrix, converting $\mathbb{C}$ to a symmetric
    matrix if needed
return Consensus partition
```

4 Experiments

We built the clustering ensemble by performing $M = 100$ runs of k-means, with k randomly chosen between $k_{min} = \max\{\sqrt{N}/2, N/50\}$ and $k_{max} = k_{min} + 20$, where N is the number of samples of the dataset. Moreover, we extracted the consensus partition by applying the single-link (SL) and the average-link (AL) algorithms to the consensus matrix, assuming the true number of clusters is known. The proposed methodology produces an asymmetric matrix, and in order to be able to apply SL and AL, one needs to convert it into a symmetric matrix. We used two different approaches: (i) average, *i.e.*, $\mathbb{C}(i, j) = \mathbb{C}(j, i) = (\mathbb{C}(i, j) + \mathbb{C}(j, i))/2$; and (ii) maximum, *i.e.*, $\mathbb{C}(i, j) = \mathbb{C}(j, i) = \max\{\mathbb{C}(i, j), \mathbb{C}(j, i)\}$, for $i, j = 1, \ldots, N$. The proposed method using the first symmetric approach is designated by $\mathrm{PCC}^{\mathrm{DID}}_{\mathrm{mean}}$, and the other one by $\mathrm{PCC}^{\mathrm{DID}}_{\mathrm{max}}$.



(a) D2 (b) Circs (c) Bars (d) Mixed Image

Fig. 1. Synthetic datasets

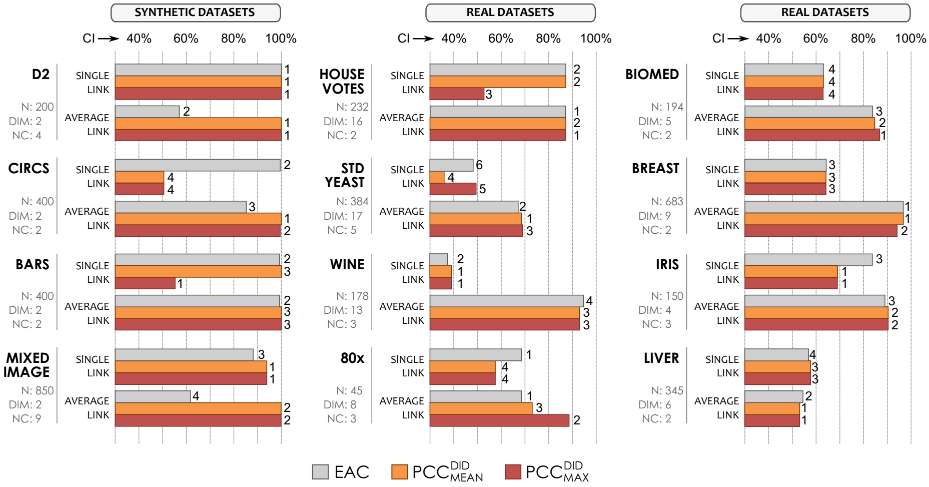


Fig. 2. Consistency index (CI, in %) for EAC, and the proposed $\text{PCC}^{\text{DID}}_{\text{mean}}$ and $\text{PCC}^{\text{DID}}_{\text{max}}$, where the final partition is extracted using SL and AL. The number on top of each bar corresponds to the rank using G-DID to sort the six methodologies. N is the number of samples, dim the dimension of the feature space, and Nc the number of clusters.

We test the performance of the proposed method in 12 datasets: 4 synthetic datasets and 8 real datasets from the UCI Machine Learning Repository[1]. The synthetic datasets are presented in fig. 1. We assessed the quality of the consensus partitions through the consistency index (CI) [6], which is the percentage of agreement between the given partition and the true labeling. Moreover, we use the Graph-based Dissimilarity Increments Distribution (G-DID) [8] to automatically choose the best methodology among all six possibilities (EAC, $\text{PCC}^{\text{DID}}_{\text{max}}$ and $\text{PCC}^{\text{DID}}_{\text{mean}}$, using SL and AL to extract the final partition), and rank, from better to worse, the six methodologies. Fig. 2 presents the consistency index and the numbers on top of the bars corresponding to the ranking given by G-DID.

Note that the probabilistic approach (PCC^{DID}) is always better or equal to the voting scheme (EAC), and the best results are achieved when the consensus partition is extracted using AL. It seems that there is no difference between

[1] http://archive.ics.uci.edu/ml

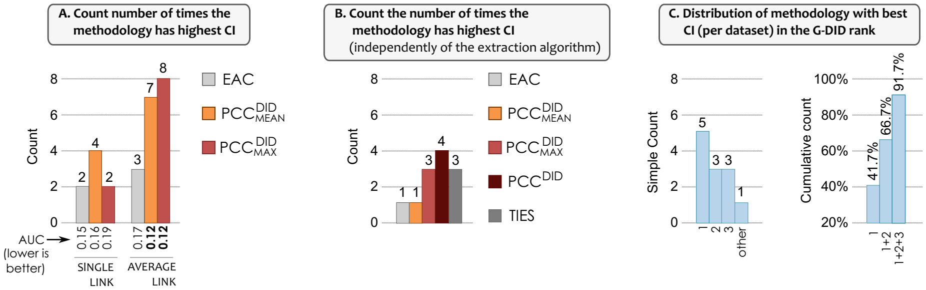


Fig. 3. Analysis of fig. 2. The AUC corresponds to the area under an error ROC curve [1], used as a measure of robustness (lower values indicate more robust methods).

PCC$^{\mathrm{DID}}_{\mathrm{mean}}$ and PCC$^{\mathrm{DID}}_{\mathrm{max}}$, however PCC$^{\mathrm{DID}}_{\mathrm{max}}$ has better results. Fig. 3 presents a summary of the results presented in fig. 2.

The first two plots in fig. 3 present the number of datasets with highest CI for: (i) the six methodologies; and (ii) the combination strategy (voting or probabilistic), independent of the extraction method. Thus, PCC$^{\mathrm{DID}}$ is clearly the best strategy, with 8 datasets against 1 dataset for EAC. Those 8 datasets are distributed as follows: 1 for PCC$^{\mathrm{DID}}_{\mathrm{mean}}$, 3 for PCC$^{\mathrm{DID}}_{\mathrm{max}}$ and 4 for both (max and mean are tied). Moreover, AL is the best extracting algorithm for PCC$^{\mathrm{DID}}$, with 7 and 8 datasets having the highest CI in PCC$^{\mathrm{DID}}_{\mathrm{mean}}$ and PCC$^{\mathrm{DID}}_{\mathrm{max}}$, respectively. Furthermore, the AUC values correspond to the area under a ROC curve [1], which is used to test the robustness of each methodology. Therefore, PCC$^{\mathrm{DID}}$ with AL used as an extraction algorithm are more robust than EAC.

The two plots on the right of fig. 3, present an analysis of the G-DID rankings. Looking at those plots, we notice that G-DID chooses, as the best partition fitted to data, the one with the highest CI for 5 datasets, and it ranks in second and third place the best CI in 3 datasets each. In 91.7% of datasets, G-DID puts the partition with highest CI in the first three places of the rank, and in 41.7% of datasets it founds the best partition, the one with the highest CI value.

5 Conclusions

This paper presents a new methodology for learning similarities from clustering ensembles, which models the co-occurrence of pairs of objects being in the same cluster using some probabilistic model. In this paper we model that co-occurrence using the dissimilarity increments distribution, but any other model of observing pairs of objects can be used. Moreover, we assumed that each partition of the clustering ensemble contributes equally to the combination matrix, but some partitions may have a better fit to the data than others. In future work, we will study more models for observing pairs of objects in the same cluster and, also, find a criterion to measure how well a partition of the ensemble fits the data, constructing a weighted version of the proposed method.

The proposed methodology performs better than the voting scheme (EAC), which consists in counting the co-occurrences of pairs of objects in a cluster. Moreover, the proposed method is more robust than EAC.

Acknowledgments. This work was supported by the Portuguese Foundation for Science and Technology grant PTDC/EEI-SII/2312/2012.

References

1. Aidos, H., Duin, R., Fred, A.: The area under the ROC curve as a criterion for clustering evaluation. In: Proc. of the 2nd Int. Conf. on Pattern Recognition Applications and Methods (ICPRAM 2013), Barcelona, pp. 276–280 (2013)
2. Aidos, H., Fred, A.: Statistical modeling of dissimilarity increments for *d*-dimensional data: Application in partitional clustering. Pattern Recognition 45(9), 3061–3071 (2012)
3. Ayad, H.G., Kamel, M.S.: Cumulative voting consensus method for partitions with variable number of clusters. IEEE Trans. Pattern Anal. Mach. Intell. 30(1), 160–173 (2008)
4. Dimitriadou, E., Weingessel, A., Hornik, K.: A combination scheme for fuzzy clustering. In: Pal, N.R., Sugeno, M. (eds.) AFSS 2002. LNCS (LNAI), vol. 2275, pp. 332–338. Springer, Heidelberg (2002)
5. Fern, X.Z., Brodley, C.E.: Solving cluster ensemble problems by bipartite graph partitioning. In: Machine Learning - Proc. of the 21st Int. Conf. (ICML 2004), Banff, Alberta, Canada (2004)
6. Fred, A.: Finding consistent clusters in data partitions. In: Kittler, J., Roli, F. (eds.) MCS 2001. LNCS, vol. 2096, pp. 309–318. Springer, Heidelberg (2001)
7. Fred, A., Jain, A.: Combining multiple clusterings using evidence accumulation. IEEE Trans. Pattern Anal. Mach. Intell. 27(6), 835–850 (2005)
8. Fred, A., Jain, A.: Cluster validation using a probabilistic attributed graph. In: 19th Int. Conf. on Pattern Recognition (ICPR 2008), Florida, USA, pp. 1–4 (2008)
9. Fred, A., Leitão, J.: A new cluster isolation criterion based on dissimilarity increments. IEEE Trans. Pattern Anal. Mach. Intell. 25(8), 944–958 (2003)
10. Kuncheva, L.I., Hadjitodorov, S.T.: Using diversity in cluster ensembles. In: Proc. of the IEEE Int. Conf. on Systems, Man & Cybernetics, The Hague, Netherlands, pp. 1214–1219 (2004)
11. Lourenço, A., Bulò, S.R., Rebagliati, N., Fred, A., Figueiredo, M., Pelillo, M.: Probabilistic evidence accumulation for clustering ensembles. In: Proc. of the 2nd Int. Conf. on Pattern Recognition Applications and Methods (ICPRAM 2013), pp. 58–67. Barcelona (2013)
12. Meila, M., Pentney, W.: Clustering by weighted cuts in directed graphs. In: Proc. of the SIAM Int. Conf. on Data Mining (SDM 2007), pp. 135–144. Minnesota (2007)
13. Strehl, A., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research 3, 583–617 (2002)
14. Topchy, A., Jain, A., Punch, W.: Combining multiple weak clusterings. In: Proc. of the 3rd IEEE Int. Conf. on Data Mining (ICDM 2003), Melbourne, Florida, USA, pp. 331–338 (2003)
15. Wang, H., Shan, H., Banerjee, A.: Bayesian cluster ensembles. In: Proc. of the SIAM Int. Conf. on Data Mining (SDM 2009), Nevada, USA, pp. 211–222 (2009)