

COMBINING MULTIPLE PARTITIONS CREATED WITH A GRAPH-BASED CONSTRUCTION FOR DATA CLUSTERING

*L. Galluccio*¹, *O. Michel*², *P. Comon*¹, *A. O. Hero*³ and *M. Klinger*⁴

¹ Laboratory I3S - UMR 6070 - University of Nice Sophia Antipolis - France

² Gipsa - Lab - UMR 5216 - Grenoble INP - France

³ Dept. of EECS - University of Michigan - Ann Arbor - USA

⁴ Medasense Biometrics Ltd. - Ofakim - Israel

ABSTRACT

This paper focusses on a new clustering method called evidence accumulation clustering with dual rooted prim tree cuts (EAC-DC), based on the principle of cluster ensembles also known as “combining multiple clustering methods”. A simple weak clustering algorithm is introduced based upon the properties of dual rooted minimal spanning trees and it is extended to multiple rooted trees. Co-association measures are proposed that account for the cluster sets obtained by these methods. These are exploited in order to obtain new ensemble consensus clustering algorithms. The EAC-DC methodology applied to both real and synthetic data sets demonstrates the superiority of the proposed methods.

1. INTRODUCTION

Data clustering (or unsupervised learning) is the task consisting of partitioning a set of data into non-overlapping subsets such that the inter-cluster variance is maximized and the intra-cluster variance is minimized ; patterns belonging to a same cluster share more similarity with each other than with patterns belonging to different cluster. Notion of similarity does not have a widely accepted definition and remains context dependent. This problem has been addressed in many fields of research: data mining, pattern recognition, machine learning. Many approaches have been developed to tackle this problem, among them hierarchical methods and partitioning algorithms are the most popular, see e.g. [1] or [2]. In this paper, we focus our attention on a more recent concept derived from studies on multi-learner systems. This new trend has recently emerged from the idea of combining multiple clustering results, called evidence accumulation clustering or consensus clustering. Since different clustering methods or initializations can produce different partitions on the same data, many approaches have been de-

veloped to extract significant clusters from a ensemble of partitions.

In [3], Fred and Jain propose to combine different data partitions by using a co-association measure (see next section for a definition). In [4], the authors explore the same idea but consider a set of clustering partitions obtained by varying the initialization parameters of a K -means algorithm [5]. Such approaches were extended to the problem of clustering different subsets of objects and to combining clusters from a set of projections onto subspaces [6], [7]. All these methods require the definition of consensus functions based on either co-association matrix [8], majority voting [9], hypergraph partitioning [10] or mixture of models [6]. A common appealing feature of these approaches is that combination of weak clustering methods can significantly enhance the quality of the clustering algorithms [6].

In this paper, we present a method for data clustering that combines the concepts of cluster ensembles and minimal spanning tree (MST) -based clustering, called evidence accumulating clustering with dual rooted prim tree cuts (EAC-DC). As in other work [11] basis for the method is the evolution of Prim’s algorithm for constructing the minimal spanning tree. An attractive feature of Prim-based clustering comes from its ability to account for both local and global geometrical information about the distribution of the data that is partitioned. Starting from the work of Grikschat et al. [11], we present modified dual rooted MST distances that account for higher order neighborhood information in the distance matrix. We propose a clustering method which extracts sets of clusters from dual rooted and multi rooted diffusion trees. A K -rooted tree defines K clusters within the initial dataset without the need to define a partition of the dataset because not all vertices need to be connected. However, like most other clustering methods, the K -rooted tree exhibits variability with respect to the set of initializations of parameters. Ensemble clustering methods are applied to the K -rooted tree clusters to reduce variability and enhance clustering performance. We define new co-association measures, which compute the percentage of time a pair of points

⁰This research was partially supported by PPF-ISSO, Université de Nice Sophia Antipolis and by grants from the US National Science Foundation under grants CCR-0325571 and CCF 0830490.

is classified in the same group. The final cluster set is computed by applying a spectral clustering algorithm [12] on co-association matrices and applying ensemble consensus approaches.

In Section 2, Prim’s algorithm for constructing a MST is reviewed. Dual rooted and multi rooted MST-based clustering approaches are developed and discussed, then co-association measures that make use of the obtained MST-based cluster sets are proposed. The consensus clustering is presented to combine the cluster partitions. Different cluster validity indices are described in Section 3. Section 4 presents experimental results on both synthetic and real data sets.

2. EVIDENCE ACCUMULATION BASED ON MULTIPLE GRAPH CUTS

In this section, we consider the set V of N data points $\in \mathbf{R}^L$, which we want to partition into K clusters. Let $P_i = \{C_1, \dots, C_K\}$ stands for a set of clusters, as obtained from the data by applying P_i . Notice that P_i and P_j may be identical algorithms with different initialization parameters, or different clustering algorithms. M different partitions of the data are available ; \mathbf{P} represents the cluster ensemble: $\mathbf{P} = (P_1, \dots, P_M)$.

Let $G = (V, E)$ be an undirected graph where $V = (v_1, \dots, v_N)$ is the set of N vertices and $E = (e_1, \dots, e_{N-1})$ denotes the set of edges. The weight of an edge measures the dissimilarity or separation between two vertices. Here, only Euclidean distances will be examined although other metrics can be used [13].

There exist different algorithms for constructing the MST [14]; in this paper, the proposed method is the Prim construction algorithm [15]. The MST is fully connected, acyclic and the full MST does not depend upon the initial vertex on to which it is rooted. At iteration i , the Prim’s algorithm connects the closest non-connected vertex to the tree constructed at iteration $i - 1$. Subtrees obtained after $N - 1$ iterations of the Prim’s algorithm are equivalent MSTs defined over the set of N connected vertices. If Prim’s algorithm is initialized at a vertex x and is stopped prematurely after k iterations, $k < N - 1$, one obtains a subtree that we call a “Prim subtree rooted at x ”. The weight of a MST is minimal, where the weight is defined as the sum of the weights of edges connecting the vertices. There exist two principal ways to obtain clusters from the MST: 1) forward algorithms that successively add edges to a MST subtree until a cluster criterion is reached; 2) pruning algorithms that successively remove edges from the MST [2]. For example in a pruning method one segments the MST graph into k clusters by removing the $k - 1$ largest edges. This method, similar to single-linkage clustering [2], is known to be unstable, especially when the data contain outliers or are

corrupted by noise. On the other hand, the proposed method of evolving several rooted Prim subtrees over the data set is an example of a forward algorithm. Forward algorithms are often less sensitive to outliers, as shown in the K-MST work of Hero and Michel [16].

2.1. Dual-rooted tree

In this paper, we propose to construct a cluster ensemble \mathbf{P} by applying Prim construction with different initializations. In [11], Grikschat et al. propose a graph-based distance measure between two vertices derived from the hitting-time of the two Prim subtrees rooted at each pair of distinct vertices. In Grikschat paper, each Prim subtree is grown simultaneously at each iteration. A slight modification of the distance measure is proposed here where at each step of the tree growing procedure, only one of the two Prim subtrees is grown at each iteration, namely the one for which the new edge has minimal weight. As [11], this process continues until the two subtrees collide. However, unlike in [11] numbers of vertices connected within each subtree are no longer always identical.

The tree obtained by the union of the two Prim subtrees is referred to as **Dual Rooted Prim (Droopi)**¹. This Droopi tree enjoys many interesting properties, among which one will be extensively used in the rest of the paper : for a given couple of vertices $\{v_1, v_2\}$ serving as roots of two subtrees, the last constructed edge, of weight noted w_{last} connects the two subtrees together, is always the largest (with maximum weight) among the set of all edges from both subtrees. This maximum weight edge defines a distance - $d(v_1, v_2) = w_{last}$ - that is symmetric, positive and satisfies the triangular inequality. Consequently, each pair of root vertices is associated with a distance defined by the Droopi tree that connects a subset of vertices. This subset is easily partitioned into two clusters by applying a single graph cut to remove the largest, and last constructed edge. For each pair of root vertices considered, the members of a given cluster receive the same label, and these labels are recorded. In Fig. 1, the Droopi tree is illustrated. We observe two labeled classes corresponding to vertices belonging to each subtree and a set of unlabeled vertices that are the unconnected vertices.

For each run of the graph-partitioning algorithm with different initializations, let $P_i = \{C_{1i}, C_{2i}\}$ be the resulting partition of the set of connected vertices for the couple of roots $\{v_{1i}, v_{2i}\}$ (where i indexes this particular choice of roots). It is important to emphasize that $C_{1i} \cup C_{2i} \neq V$ and therefore, the algorithm exhibits two labeled classes and a set of unlabeled vertices. This will lead to some refinements in the definition of co-association measure (see Section 2.3).

¹Details and properties of this new measure will be published separately.

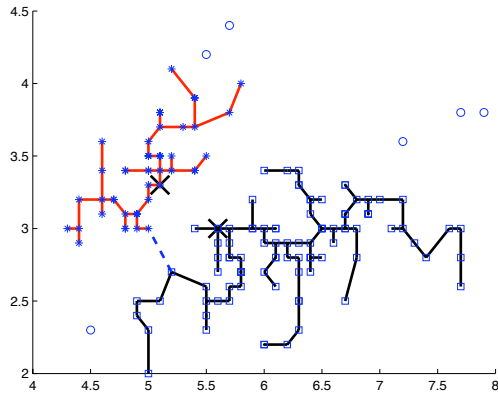


Fig. 1. Dual rooted tree built on a data set. Symbol X marks the rooted vertices. The dashed edge is the last connected edge.

2.2. Multi-rooted tree

Dual rooted tree approaches as introduced in the previous section, are straightforwardly extended to multi-rooted tree constructions. Starting with K initial vertices, K subtrees are constructed until all subtrees collide. At each step of construction, only one subtree is grown: among the K candidates (K new vertices associated with K subtrees grown by Prim's algorithm) one connects the vertex that requires the shortest (smallest weight) edge.

The algorithm could be stopped as soon as the K subtrees grown from the K roots are all connected. The $K - 1$ cuts that will be considered are given by the longest edge that connects a subtree to the other subtrees. However, to circumvent the problem that some vertices are not included in any of the K subtrees (therefore not labeled), a slight modification is suggested. Once the $K - 1$ candidate cuts are identified, the Prim's algorithm is iterated until all points are connected. The full MST is thus constructed and each vertex will receive a label according to a subtree to which it is ultimately connected. In this manner, for each run of the graph-partitioning algorithm with different initializations set of K roots, a partition $P_i = \{C_{1i}, \dots, C_{Ki}\}$ is stored. Note that $C_{1i} \cup \dots \cup C_{Ki} = V$. In Fig. 2, a multi-rooted tree with $K = 4$ rooted vertices is shown.

2.3. Co-association measures

In this section, we propose a new way of combining clustering methods, based upon an evidence accumulation approach. Our approach relies on the assumption that two vertices belonging to the same cluster tend to receive the same cluster label every time that a different clustering algorithm is applied.

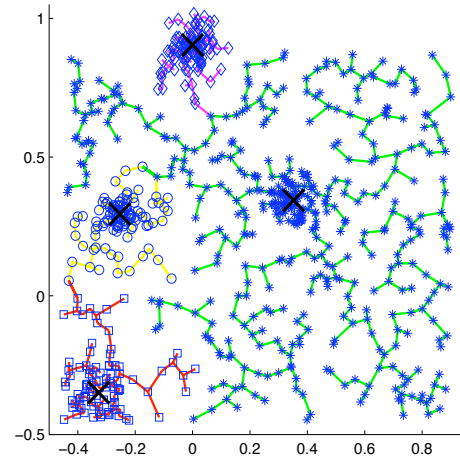


Fig. 2. Multi-rooted tree built on a data set. Symbol X marks the rooted vertices.

From the cluster ensemble $P = \{P_1, \dots, P_M\}$, obtained by applying M different clustering processes, and following the ideas introduced in [10], [8], we compute the following co-association measure

$$co_assoc1(x_i, x_j) = \frac{n(x_i, x_j)}{M}, \quad (1)$$

where $n(x_i, x_j)$ stands for the number of times the two vertices received the same label, and therefore belong to the same cluster.

When Droopi tree based clustering is implemented, there may be some instances when not all vertices receive a label (see Section 2.1). In order to account for this behaviour, a modified co-association measure is proposed

$$co_assoc2(x_i, x_j) = \frac{n(x_i, x_j)}{m(x_i, x_j)}, \quad (2)$$

where $n(x_i, x_j)$ is defined as above and $m(x_i, x_j)$ stands for the number of times (among M) both vertices x_i and x_j receive a label. By construction, $0 \leq m(x_i, x_j) \leq M$.

Although these definitions of co-association measures rely upon some reasonable heuristics, we propose to modify them. We consider all unlabeled vertices as being the elements of a "rejection cluster". If $nn(x_i, x_j)$ stands for the number of times both x_i and x_j belong to this rejection cluster, a natural modification of the definition (1) can thus be $co_assoc_{mod} = \frac{n+nn}{M}$. As this measure gives an equal importance to the actual detected clusters and to the rejection cluster, we propose here to consider instead

$$co_assoc3(x_i, x_j) = \frac{n(x_i, x_j)}{m(x_i, x_j)} + \frac{nn(x_i, x_j)}{M}, \quad (3)$$

As $m \leq M$, this will apply a larger weight to the detected clusters, and a lower weight to the rejection cluster.

2.4. Evidence accumulation clustering

For a given co-association criterion consider the set of all pairwise co-association measures $co_assoc(x_i, x_j)$. We propose to apply spectral clustering with consensus to extract the final clusters. This differs from the approach of Fred and Jain [8], who apply hierarchical clustering algorithms on the co-association measures, and from Strehl and Gosh use a graph partitioning algorithm [10].

Here, we propose to construct an affinity matrix A defined as:

$$A(i, j) \propto \beta \exp \left\{ + \frac{co_assoc(x_i, x_j)}{\sigma} \right\} \quad (i, j) \in [1, N]^2,$$

where σ and β are constants to be adjusted. Note that σ defines the size of the neighborhood of significant pairwise influence between x_i and x_j .

Following the work of Ng et al. [12], we propose to identify the clusters by using a spectral clustering algorithm, which extracts the eigenstructure of A .

In the remainder of the paper, we will denote the application of spectral clustering to an affinity matrix of dual rooted Prim distances as EAC-DC (Evidence Accumulation Clustering by Droopi tree Cut). The co-association measures used to construct the affinity matrix will be denoted by adding an index: EAC-DC i means that co_assoc_i was used. Multi rooted tree algorithm (Multi EAC-DC) leads to label all vertices. Therefore, only co_assoc_1 measure will be considered for such multi-tree algorithms..

Before we present the experimental results, the cluster validity indices used to evaluate the performances of the proposed approaches are introduced in the next section.

3. CLUSTER VALIDITY INDICES

Validity indices are used to compare two cluster partitions P and P^* . If P^* is some known ground truth reference partition, these indices can be used to benchmark the performances of the methods. For sake of completeness, we will also estimate the overall accuracy index which computes the percentage of points correctly labelled.

The Rand (R) index [17] is defined as

$$R(P, P^*) = \frac{a + b}{a + b + c + d}$$

where a is the number of pairs of points in V which have been classified as belonging to in the same set in P and in the same set in P^* , b is the number of pairs of points in V which have been classified in different sets in P and in different sets in P^* , c is the number of pairs of points in V which have been classified in the same set in P and in different sets in P^* and d is the number of pairs of points in V which have been classified in different sets in P and in the

same set in P^* .

By construction, $0 < R(P, P^*) < 1$. 1 is obtained when the partitions are exactly the same, up to an arbitrary permutation of partition labels.

The adjusted Rand (AR) index is defined as

$$AR(P, P^*) = \frac{a - \frac{t_1 t_2}{a+b+c+d}}{\frac{1}{2}(t_1 + t_2) - \frac{t_1 t_2}{a+b+c+d}},$$

where $t_1 = \sum_{i=1}^K \binom{N_i}{2}$ and $t_2 = \sum_{j=1}^K \binom{N_j}{2}$. N_i is the number of instances of cluster i in partition P and N_j is the number of instances of cluster j in partition P^* . A discussion of the motivation and the properties of this index may be found in [18].

The Jaccard (J) index [17] is expressed as

$$J(P, P^*) = \frac{a}{a + c + d}$$

Strehl and Gosh [10] defined an index based on some notions of information theory: the normalized mutual information (NMI). The latter aims at defining information shared partitions P and P^* and reads

$$NMI(P, P^*) = \frac{2.I(P, P^*)}{H(P^*) + H(P)}$$

$H(P) = -\sum_{i=1}^k \frac{n_i}{N} \log\left(\frac{n_i}{N}\right)$, where n_i is the number of points classified in cluster C_i , $C_i \in P$.

$I(P^*, P) = \sum_{i=1}^K \sum_{j=1}^K \frac{n_{ij}^*}{N} \log\left(\frac{n_{ij}^*/N}{n_i/n_i \cdot n_j/n_j}\right)$, n_{ij}^* is the number of shared points in $C_i \in P$ and $C_j^* \in P^*$. $H(P)$ represents the Shannon entropy of partition P ; $I(P, P^*) = I(P^*, P)$ -mutual information- measures the agreement between P and P^* .

4. EMPIRICAL STUDY

In order to validate and evaluate the performances of the proposed approaches, tests on both synthetic data and real data sets from the UCI Machine Learning Repository [19] were conducted. In all our experiments, for each tested algorithm, a set of $M = 100$ random different initialization parameters was considered. We compare our method to several of the most commonly used algorithms in the literature: co-association measure, voting, graph representation:

- evidence accumulation clustering with average-link algorithm (EAC - AL) [8], [3],

- graph and hypergraph representation as described in [10]: cluster-based similarity partitioning algorithm (CSPA), hypergraph partitioning algorithm (HPGA) and meta-clustering algorithm (MCLA),

- cumulative voting, see [9] for definitions: unnormalized reference based cumulative voting (URCV), reference based cumulative voting (RCV), adaptive cumulative voting (ACV),

- median partition: quadrature mutual information (QMI) [6].

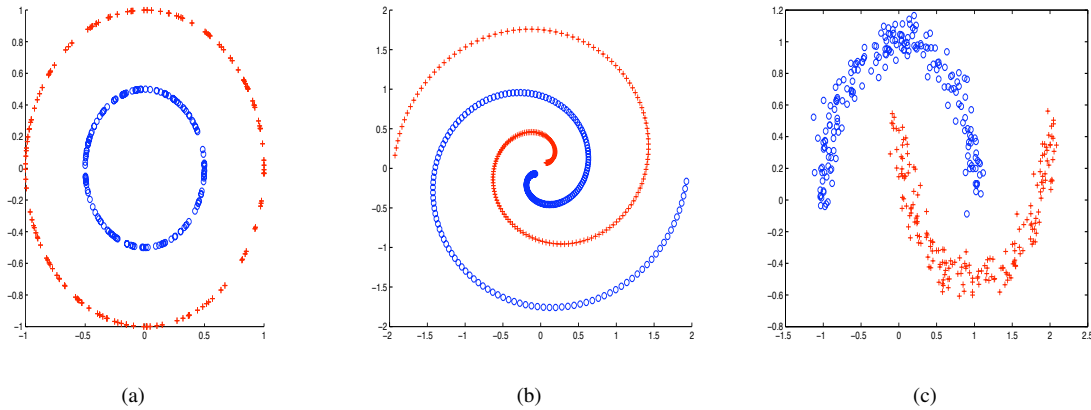


Fig. 3. Results obtained with our methods on synthetic data sets: (a) two rings, (b) two spirals, (c) two half-moons. Clusters labels are indicated by symbols.

4.1. Numerical comparisons of synthetic data sets

Table 1 reports the results obtained with the methods introduced in the previous section. Performances are measured with the simple accuracy criterion (percentage of correctly labeled points). Results of our proposed methods on the synthetic data are shown in Fig. 3. Our method clearly outperforms the previous methods proposed in the literature: this is due to the fact that the clusters in Fig. 3 are not convex. The multi rooted tree based approaches allow to solve this difficulty in the sense that the local geometry of distribution is explored through the Prim construction, whereas most other implemented methods rely on the constructions of partitions into convex subsets. Note that all these methods are unsupervised, except that the correct number of clusters is assumed to be known. Grikschat's method gives the same results as the EAC-DC but subtrees are computed with each pair of vertices, whereas in our approach only few subtrees are computed to obtain a distance matrix. Therefore, our method is less consuming in computing time. Further studies will be realized in a larger version of this paper.

Methods	Half moons	Two spirals	Two circles
EAC-DC (1, 2, 3)	1	1	1
Multi EAC-DC	1	1	1
EAC-AL	0.7675	0.5550	0.5200
QMI	0.7550	0.5050	0.5150
URCV	0.7600	0.5000	0.5200
RCV	0.7650	0.5300	0.5200
ACV	0.5000	0.5000	0.5000
CSPA	0.7300	0.5025	0.5550
HGPA	0.5000	0.5000	0.5000
MCLA	0.7275	0.5050	0.5400

Table 1. Results obtained on synthetic data

4.2. Results on real data sets

The first data set considered is Iris (150 points in a 4-dimensional feature space) [19]. It contains three classes of flower features with one class well separated from the others and two close classes. The obtained results are presented in Table 2 for different algorithms and validity indices. It appears that for these data set, the proposed multi EAC-DC significantly outperforms the other methods for any validity index. A second data set, breast cancer Wisconsin data [19], has been analyzed. It consists of 683 9-dimensional feature vectors separated into two classes. For these data, the EAC-DC dominates the other methods (Table 3), the voting approaches also give good results and notice that multi EAC-DC seems to be better in the case of more than two clusters. It will be valuable to investigate this point further on.

Methods	Accuracy	Rand	Adjusted Rand	Jaccard	NMI
EAC-DC 1	0.9067	0.8893	0.7509	0.7577	0.7637
EAC-DC 2	0.9267	0.9124	0.8032	0.8010	0.8315
EAC-DC 3	0.9200	0.9045	0.7853	0.7863	0.8022
Multi EAC-DC	0.9600	0.9495	0.8858	0.8876	0.8705
EAC - AL	0.8733	0.8580	0.6809	0.6944	0.6914
CSPA	0.9267	0.9124	0.8015	0.8100	0.7900
HGPA	0.9200	0.9055	0.7859	0.7956	0.7773
MCLA	0.8933	0.8797	0.7294	0.7360	0.7496
URCV	0.8800	0.8664	0.6997	0.7096	0.7130
RCV	0.9000	0.8859	0.7455	0.7414	0.7980
ACV	0.9067	0.8923	0.7592	0.7557	0.8057
QMI	0.8800	0.8664	0.6989	0.7114	0.7065

Table 2. Results obtained on the Iris data set

5. DISCUSSIONS

We have introduced a new clustering algorithm called EAC-DC, that is developed in two steps. First, Prim-based clustering algorithms are used to compute co-association measures. Secondly, the co-association measures are analysed

Methods	Accuracy	Rand	Adjusted Rand	Jaccard	NMI
EAC-DC 1	0.9678	0.9376	0.8743	0.9184	0.7889
EAC-DC 2	0.9605	0.9240	0.8466	0.8991	0.7458
EAC-DC 3	0.9606	0.9240	0.8466	0.8969	0.7458
Multi EAC-DC	0.9385	0.7884	0.5744	0.7268	0.4502
EAC - AL	0.9429	0.8922	0.7816	0.8488	0.6827
CSPA	0.8448	0.7374	0.4749	0.6523	0.4809
HGPA	0.6501	0.5444	0	0.4856	0
MCLA	0.9575	0.9186	0.8355	0.8869	0.7363
URCV	0.9590	0.9223	0.8409	0.8907	0.7427
RCV	0.9663	0.9348	0.8685	0.9106	0.7755
ACV	0.9649	0.9321	0.8630	0.8067	0.7684
QMI	0.9356	0.8793	0.7561	0.8366	0.6376

Table 3. Results obtained on Breast Cancer Wisconsin data

in order to provide a final consensus partition. For this purpose, we have used spectral clustering approaches. Although it seems clear that Prim-based methods handle problem of non-convex subsets, the efficiency of the proposed methods may also be due to the non-convex adaptive properties of spectral clustering algorithms. We think that the primary strength of the EAC-DC approach is its simultaneous use of Prim-based co-association measures and spectral clustering approaches. However further study is needed to confirm the relative contribution of the two steps in EAC-DC.

6. CONCLUSION

In this paper, we have presented new clustering methods, based on the concept of evidence accumulation and on the combination of multiple clustering results. We proposed to exploit the properties of Prim-based co-association measures to form the cluster ensemble. We also proposed to extend this method to a multi-rooted approach. This graph construction better captures both the local and the global intrinsic geometry of the data set. Multiple realizations of MST-cut clustering are performed for a set of random initialization parameters and allow to construct a cluster ensemble from which new co-association measures are proposed. The performances of these methods have been evaluated over a set of both synthetic and real data, thus highlighting very interesting and promising features of the proposed algorithms.

7. REFERENCES

- [1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.
- [2] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, third edition, 2006.
- [3] A. L. N. Fred and A. K. Jain, "Learning pairwise similarity for data clustering," in *Proceedings of 18th International Conference on Pattern Recognition*, Hong Kong, China, 2006.
- [4] P. S. Bradley and U. M. Fayyad, "Refining initial points for k-means clustering," in *Proceedings of 15th International Conference on Machine Learning*, Madison, Wisconsin, USA, 1998, pp. 91–99.
- [5] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Ed., Berkeley, 1967, vol. 1, pp. 281–287.
- [6] A. Topchy, A. K. Jain, and W. Punch, "Clustering ensembles: Models of consensus and weak partitions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1866–1881, December 2005.
- [7] X. Z. Fern and C. E. Brodley, "Random projection for high dimensional data clustering: A cluster ensemble approach," in *Proceedings of 20th International Conference on Machine Learning*, Washington DC, USA, 2003.
- [8] A. L. N. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 835–850, June 2005.
- [9] H. G. Ayad and M. S. Kamel, "Cumulative voting consensus method for partitions with a variable number of clusters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 1, pp. 160–173, January 2008.
- [10] A. Strehl and J. Ghosh, "Cluster ensemble - a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2002.
- [11] S. Grikschat, J. A. Costa, A. O. Hero, and O. Michel, "Dual rooted-diffusions for clustering and classification on manifolds," in *Proceedings of 31th IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, 2006.
- [12] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances on Neural Information Processing Systems 15*, 2001.
- [13] O. J. J. Michel, P. Bendjoya, and P. RojoGuer, "Unsupervised clustering with MST: Application to asteroid data," in *Proceedings of 4th Physics in Signal and Images Processing*, Toulouse, France, 2005.
- [14] D. J. Marchette, *Random graphs for statistical pattern recognition*, Wiley, New York, 2004.
- [15] R. Prim, "Shortest connection networks and some generalizations," *Bell System Technical Journal*, vol. 36, pp. 1389–1401, 1957.
- [16] A. O. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k -point random graphs," *IEEE Transactions on Information Theory*, vol. 45, pp. 1921–1939, 1999.
- [17] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, pp. 107–145, 2001.
- [18] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, 1985.
- [19] A. Asuncion and D. J. Newman, "UCI machine learning repository," 2007.