Cluster Ensembles

Kurt Hornik

Institut für Statistik und Mathematik,
Wirtschaftsuniversität Wien,
Augasse 2–6, A-1090 Wien, Austria

Abstract. Cluster ensembles are collections of individual solutions to a given clustering problem which are useful or necessary to consider in a wide range of applications. Aggregating these to a "common" solution amounts to finding a consensus clustering, which can be characterized in a general optimization framework. We discuss recent conceptual and computational advances in this area, and indicate how these can be used for analyzing the structure in cluster ensembles by clustering its elements.

1 Introduction

Ensemble methods create solutions to learning problems by constructing a set of individual (different) solutions ("base learners"), and subsequently suitably aggregating these, e.g., by weighted averaging of the predictions in regression, or by taking a weighted vote on the predictions in classification. Such methods, which include Bayesian model averaging (Hoeting et al. (1999)), bagging (Breiman (1996)) and boosting (Friedman et al. (2000)) have already become very popular for supervised learning problems (Dietterich (2002)).

In general, aggregation yields algorithms with "low variance" in the statistical learning sense so that the results obtained by aggregation are more "structurally stable". Based on the success and popularity of ensemble methods, the statistical and machine learning communities have recently also become interested in employing these in unsupervised learning tasks, such as clustering. (Note that in these communities, the term "classification" is used for discriminant analysis. To avoid ambiguities, we will use "supervised classification" to refer to these learning problems.) For example, a promising idea is to obtain more stable partitions of a given data set using bagging (Bootstrap Aggregating), i.e., by training the same base clusterer on bootstrap samples from the data set and then finding a "majority decision" from the labelings thus obtained. But obviously, aggregation is not as straightforward as in the supervised classification framework, as these labelings are only unique up to permutations and therefore not necessarily matched. In the classification community, such aggregation problems have been studied for quite some time now. A special issue of the Journal of Classification was devoted to "Comparison and Consensus of Classifications" (Day (1986)) almost two decades ago. By building on the readily available optimization framework for obtaining

consensus clusterings it is possible to exploit the full potential of the cluster ensemble approach.

Employing cluster ensembles can be attractive or even necessary for several reasons, the main ones being as follows (see e.g. Strehl and Ghosh (2002)):

- To improve quality and robustness of the results. Bagging is one approach to reduce variability via resampling or reweighting of the data, and is used in Leisch (1999) and Dudoit and Fridlyand (2002). In addition, many clustering algorithms are sensitive to random initializations, choice of hyper-parameters, or the order of data presentation in on-line learning scenarios. An obvious idea for possibly eliminating such *algorithmic* variability is to construct an ensemble with (randomly) varied characteristics of the base algorithm. This idea of "sampling from the algorithm" is used in Dimitriadou et al. (2001, 2002). Aggregation can also *leverage* performance in the sense of turning weak into strong learners; both Leisch (1999) and Dimitriadou et al. (2002) illustrate how e.g. suitable aggregation of base k-means results can reveal underlying non-convex structure which cannot be found by the base algorithm. Other possible strategies include varying the "features" used for clustering (e.g., using various preprocessing schemes), and constructing "meta-clusterers" which combine the results of the application of *different* base algorithms as an attempt to reduce dependency of results on specific methods, and take advantage of today's overwhelming method pluralism.
- To aggregate results over conditioning/grouping variables in situations where repeated measurements of features on objects are available for several levels of a grouping variable, such as the 3-way layout in Vichi (1999) where the grouping levels correspond to different time points at which observations are made.
- To reuse existing knowledge. In applications, it may be desired to reuse legacy clusterings in order to improve or combine these. Typically, in such situations only the cluster labels are available, but not the original features or algorithms.
- To accommodate the needs of distributed computing. In many applications, it is not possible to use all data simultaneously. Data may not necessarily be available in a single location, or computational resources may be insufficient to use a base clusterer on the whole data set. More generally, clusterers can have access to either a subset of the objects ("object-distributed clustering") or the features ("feature-distributed clustering"), or both.

In all these situations, aggregating (subsets of) the cluster ensemble by finding "good" consensus clusterings is fundamental. In Section 2, we consider a general optimization framework for finding consensus partitions. Extensions are discussed in Section 3.

2 Consensus partitions

There are three main approaches to obtaining consensus clusterings (Gordon and Vichi (2001)): in the *constructive* approach, a way of constructing a consensus clustering is specified: for example, a strict consensus clustering is defined to be one such that objects can only be in the same group in the consensus partition if they were in the same group in all base partitions. In the *axiomatic* approach, emphasis is on the investigation of existence and uniqueness of consensus clusterings characterized axiomatically. The *optimization* approach formalizes the natural idea of describing consensus clusterings as the ones which "optimally represent the ensemble" by providing a criterion to be optimized over a suitable set $\mathcal{C}$ of possible consensus clusterings. Given a function d which measures dissimilarity (or distance) between two clusterings, one can e.g. look for clusterings which minimize average dissimilarity, i.e., which solve

$$C^* = \text{argmin}_{C \in \mathcal{C}} \sum_{b=1}^{B} d(C, C_b)$$

over $\mathcal{C}$. Analogously, given a measure of similarity (or agreement), one can look for clusterings maximizing average similarity. Following Gordon and Vichi (1998), one could refer to the above C^* as the *median* or *medoid* clustering if the optimum is sought over the set of all possible base clusterings, or the set $\{C_1, \ldots, C_B\}$ of the base clusterings, respectively.

When finding consensus *partitions*, it seems natural to look for optimal *soft* partitions which make it possible to assign objects to several groups with varying degrees of "membership" (Gordon and Vichi (2001), Dimitriadou et al. (2002)). One can then assess the amount of belongingness of objects to groups via standard impurity measures, or the so-called classification margin (the difference between the two largest memberships). Note that "soft" partitioning includes fuzzy partitioning methods such as the popular fuzzy c-means algorithm (Bezdek (1974)) as well as probabilistic methods such as the model-based approach of Fraley and Raftery (2002). In addition, one can compute global measures Φ of the softness of partitions, and use these to extend the above optimization problem to minimizing

$$\sum_{b=1}^{B} \omega_b d(C, C_b) + \lambda \Phi(C)$$

over all soft partitions, where the ω indicate the importance of the base clusterings (e.g., by assigning importance according to softness of the base partitions), and λ controls the amount of "regularization". This extension also allows for a soft-constrained approach to the "simple" problem of optimizing over all hard partitions. Of course, one could consider criterion functions resulting in yet more robust consensus solutions, such as the median or trimmed mean of the distances $d(C, C_b)$.

One should note that the above optimization problems are typically computationally very hard. Finding an optimal hard partition with K labels in

general makes it necessary to search all possible hard partitions (the number of which is of the order $(K+1)^n$ (Jain and Dubes (1988)) for the optimum. Such exhaustive search is clearly impossible for most applications. Local strategies, e.g. by repeating random reassigning until no further improvement is obtained, or Boltzmann-machine type extensions (Strehl and Ghosh (2002)) are still expensive and not guaranteed to find the global optimum.

Perhaps the most popular similarity measure for partitions of the same data set is the Rand index (Rand (1971)) used in e.g. Gordon and Vichi (1998), or the Rand index corrected for agreement by chance (Hubert and Arabie (1985)) employed by Krieger and Green (1999). Finding (hard) consensus partitions by maximizing average similarity is NP-hard in both cases. Hence, Krieger and Green (1999) propose an algorithm (SEGWAY) based on the combination of local search by relabeling single objects together with "smart" initialization using random assignment, latent class analysis (LCA), multiple correspondence analysis (MCA), or a greedy heuristic. Note also that using (dis)similarity measures adjusted for agreement by chance works best if the partitions are stochastically independent, which is not necessarily the case in all cluster ensemble frameworks described in Section 1.

In what follows, the following terminology will be useful. Given a data set $\mathcal{X}$ with the measurements of the same features (variables) on n objects, a *K-clustering* of $\mathcal{X}$ assigns to each x_i in $\mathcal{X}$ a (sub-)probability K-vector $C(x_i) = (\mu_{i1}, \ldots, \mu_{iK})$ (the "membership vector" of the object) with $\mu_{i1}, \ldots, \mu_{iK} \geq 0$, $\sum_k \mu_{ik} \leq 1$. Formally,

$$C : \mathcal{X} \to M \in \mathcal{M}_K; \qquad \mathcal{M}_K = \{M \in \mathbb{R}^{n \times K} : M \geq 0, M1_K \leq 1_K\},$$

where 1_K is a lenght K column vector of ones, and $M1_K$ is the matrix product of M and 1_K. This framework includes hard partitions (where each $C(x_i)$ is a unit Cartesian unit vector) and soft ones, as well as incomplete (e.g., completely missing, for example if a sample from $\mathcal{X}$ was used) results where $\sum_k \mu_{ik} < 1$. Permuting the labels (which correspond to the columns of the membership matrix M) amounts to replacing M by $M\Pi$, where Π is a suitable permutation matrix.

The dissimilarity measure used in Models I and II of Gordon and Vichi (2001) and in Dimitriadou et al. (2002) use the Euclidean dissimilarity of the membership matrices, adjusted for optimal matching of the labels. If both partitions use the same number of labels, this is given by

$$d_F(M, \tilde{M}) = \min_\Pi \|M - \tilde{M}\Pi\|^2$$

where the minimum is taken over all permutation matrices Π and $\|\cdot\|$ is the Frobenius norm (so that $\|Y\|^2 = \mathrm{tr}(Y'Y)$, where $'$ denotes transposition). As $\|M - \tilde{M}\Pi\|^2 = \mathrm{tr}(M'M) - 2\,\mathrm{tr}(M'\tilde{M}\Pi) + \mathrm{tr}(\Pi'\tilde{M}'\tilde{M}\Pi) = \mathrm{tr}(M'M) - 2\,\mathrm{tr}(M'\tilde{M}\Pi) + \mathrm{tr}(\tilde{M}'\tilde{M})$, we see that minimizing $\|M - \tilde{M}\Pi\|^2$ is equivalent to maximizing $\mathrm{tr}(M'\tilde{M}\Pi) = \sum_{i,k} \mu_{ik}\tilde{\mu}_{i,\pi(k)}$, which for hard partitions is

the number of objects with the same label in the partitions given by M and $\tilde{M}\Pi$. Finding the optimal Π is thus recognized as an instance of the assignment problem (or weighted bipartite graph matching problem), which can be solved by a linear program using the so-called Hungarian method in time $O(K^3)$ (e.g., Papadimitriou and Steiglitz (1982)). If the partitions have different numbers of labels, matching also includes suitably collapsing the labels of the finer partition, see Gordon and Vichi (2001) for details.

Finding the consensus K-clustering of given base K-clusterings with membership matrices $M_1, \ldots, M_B$ amounts to minimizing $\sum_{b=1}^{B} d_F(M, M_b)$ over $\mathcal{M}_K$, and is equivalent to minimizing $\sum_{b=1}^{B} \|M - M_b\Pi_b\|^2$ over $M \in \mathcal{M}_K$ and all permutation matrices $\Pi_1, \ldots, \Pi_K$. Dimitriadou et al. (2002) show that the optimal M is of the form

$$M = \frac{1}{B} \sum\nolimits_{b=1}^{B} M_b \Pi_b$$

for suitable permutation matrices $\Pi_1, \ldots, \Pi_B$. A hard partition obtained from this consensus partition by assigning objects to the label with maximal membership thus performs simple majority voting after relabeling, which motivates the name "voting" for the proposed framework. The $\Pi_1, \ldots, \Pi_B$ in the above representation are obtained by simultaneously maximizing the profile criterion function

$$\sum\nolimits_{1 \le \beta, b \le B} \mathrm{tr}(\Pi_\beta' M_\beta' M_b \Pi_b)$$

over all possible permutation matrices (of course, one of these can be taken as the identity matrix). This is a special case (but not an instance) of the multiple assignment problem, which is known to be NP-complete, and can e.g. be approached using randomized parallel algorithms (Oliveira and Pardalos (2004)). However, we note that unlike in the general case, the above criterion function only contains second-order interaction terms of the permutations. Whether the determination of the optimal permutations and hence of the consensus clustering is possible in time polynomial in both B and K is currently not known.

Based on the characterization of the consensus solution, Dimitriadou et al. (2002) suggest a greedy forward aggregation strategy for determining approximate solutions. One starts with $\tilde{M}_0 = M_1$ and then, for all b from 1 to B, first determines a locally optimal relabeling $\tilde{\Pi}_b$ of M_b to $\tilde{M}_{b-1}$ (i.e., solves the assignment problem $\mathrm{argmin}_\Pi \|\tilde{M}_{b-1} - M_b\Pi\|^2$ using the Hungarian method), and determines the optimal $M = \tilde{M}_b = (1/b) \sum_{\beta=1}^{b} \tilde{M}_\beta \tilde{\Pi}_\beta$ for fixed $\tilde{\Pi}_1, \ldots, \tilde{\Pi}_b$ by on-line averaging as $\tilde{M}_b = (1 - 1/b)\tilde{M}_{b-1} + (1/b)M_b\tilde{\Pi}_b$. The final $\tilde{M}_B$ is then taken as the approximate consensus clustering. One could extend this approach into a fixed-point algorithm which repeats the forward aggregation, with the order of membership matrices possibly changed, until convergence. Gordon and Vichi (2001) propose a different approach

which iterates between simultaneously determining the optimal relabelings $\Pi_1, \ldots, \Pi_B$ for fixed M by solving the corresponding assignment problems, and then optimizing for M for fixed $\Pi_1, \ldots, \Pi_B$ by computing the average $(1/B)\sum_{b=1}^{B} M_b \Pi_b$.

In the aggregation strategy Bag1 of Dudoit and Fridlyand (2002), the same base clusterer is applied to both the original data set and B bootstrap samples thereof, giving membership matrices M_{ref} and $M_1, \ldots, M_B$. Optimal relabelings Π_b are obtained by matching the M_b to M_{ref}, and (a hard version of) the consensus partition is then obtained by averaging the $M_b \Pi_b$. There seems to be no optimization criterion underlying this constructive approach.

According to Messatfa (1992), historically the first index of agreement between partitions is due to Katz and Powell (1953), and based on the Pearson product moment correlation coefficient of the off-diagonal entries of the co-incidence matrices MM' of the partitions. (Note that the (i, j)-th element of MM' is given by $\sum_{k=1}^{K} \mu_{ik}\mu_{jk}$, which in the case of hard partitions is one if objects i and j are in the same group, and zero otherwise, and that relabeling does not change MM'.) A related dissimilarity measure (using covariance rather than correlation) is

$$d_C(M, \tilde{M}) = \|MM' - \tilde{M}\tilde{M}'\|^2$$

The corresponding consensus problem is the minimization of $\sum_b \|MM' - M_b M_b'\|^2$, or equivalently of

$$\left\| MM' - \frac{1}{B} \sum\nolimits_{b=1}^{B} M_b M_b' \right\|^2$$

over $\mathcal{M}_K$. This is Model III of Gordon and Vichi (2001), who suggest to use a sequential quadratic programming algorithm (which can only be guaranteed to find local minima) for obtaining the optimal $M \in \mathcal{M}_K$. The average co-incidence matrix $(1/B)\sum_{b=1}^{B} M_b M_b'$ also forms the basis of the constructive consensus approaches in Fred and Jain (2002) and Strehl and Ghosh (2002).

3 Extensions

The optimization approach to finding consensus clusterings is also applicable to the case of hierarchical clusterings (Vichi (1999)). If these are represented by the corresponding ultra-metric matrices $U_1, \ldots, U_B$, a consensus clustering can be obtained e.g. by minimizing $\sum_b \|U - U_b\|^2$ over all possible ultra-metric matrices U.

In many applications of cluster ensembles, interest is not primarily in obtaining a global consensus clustering, but to analyze (dis)similarity patterns in the base clusterings in more detail—i.e., to cluster the clusterings. Gordon and Vichi (1998) present a framework in which all clusterings considered

are hard partitions. Obviously, the underlying concept of "clustering clusterings", based on suitable (dis)similarity measures between clusterings, such as the ones discussed in detail in Section 2, is much more general. In particular, it is straightforward to look for hard prototype-based partitions of a cluster ensemble characterized by the minimization of

$$\sum_{k=1}^{K} \sum_{C(M_b)=e_k} d(M_b, P_k),$$

where e_k is the k-th Cartesian unit vector over all possible hard assignments C of membership matrices to labels and all suitable prototypes $P_1, \ldots, P_K$. If the usual algorithm which alternates between finding optimal prototypes for fixed assignments and reassigning the M_b to their least dissimilar prototype is employed, we see that finding the prototypes amounts to finding the appropriate consensus partitions in the groups. Similarly, soft partitions can be characterized as the minima of the fuzzy c-means style criterion function $\sum_{k,b} u_{kb}^q d(M_b, P_k)$.

References

BEZDEK, J. C. (1974): Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biology, 1, 57–71.*

BREIMAN, L. (1996): Bagging predictors. *Machine Learning, 24(2), 123–140.*

DAY, W. H. E. (1986): Foreword: Comparison and consensus of classifications. *Journal of Classification, 3, 183–185.*

DIETTERICH, T. G. (2002): Ensemble learning. In: M. A. Arbib (Ed.): *The Handbook of Brain Theory and Neural Networks.* The MIT Press, Cambridge, MA, 405–408.

DIMITRIADOU, E., WEINGESSEL, A. and HORNIK, K. (2001): Voting-merging: An ensemble method for clustering. In: G. Dorffner, H. Bischof and K. Hornik (Eds.): *Artificial Neural Networks – ICANN 2001*, volume 2130 of *LNCS.* Springer Verlag, 217–224.

DIMITRIADOU, E., WEINGESSEL, A. and HORNIK, K. (2002): A combination scheme for fuzzy clustering. *International Journal of Pattern Recognition and Artificial Intelligence, 16(7), 901–912.*

DUDOIT, S. and FRIDLYAND, J. (2002): A prediction-based resampling method to estimate the number of clusters in a dataset. *Genome Biology, 3(7), 0036.1–0036.21.*

FRALEY, C. and RAFTERY, A. E. (2002): Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97, 611–631.* URL `http://www.stat.washington.edu/mclust`.

FRED, A. L. N. and JAIN, A. K. (2002): Data clustering using evidence accumulation. In: *Proceedings of the 16th International Conference on Pattern Recognition (ICPR 2002)*, 276–280.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000): Additive logistic regression: A statistical view of boosting. *The Annals of Statistics, 28(2), 337–407.*

GORDON, A. D. and VICHI, M. (1998): Partitions of partitions. *Journal of Classification, 15, 265–285*.

GORDON, A. D. and VICHI, M. (2001): Fuzzy partition models for fitting a set of partitions. *Psychometrika, 66(2), 229–248*.

HOETING, J., MADIGAN, D., RAFTERY, A. and VOLINSKY, C. (1999): Bayesian model averaging: A tutorial. *Statistical Science, 14, 382–401*.

HUBERT, L. and ARABIE, P. (1985): Comparing partitions. *Journal of Classification, 2, 193–218*.

JAIN, A. K. and DUBES, R. C. (1988): *Algorithms for Clustering Data*. Prentice Hall, New Jersey.

KATZ, L. and POWELL, J. H. (1953): A proposed index of the conformity of one sociometric measurement to another. *Psychometrika, 18, 149–256*.

KRIEGER, A. M. and GREEN, P. E. (1999): A generalized Rand-index method for consensus clustering of separate partitions of the same data base. *Journal of Classification, 16, 63–89*.

LEISCH, F. (1999): *Bagged clustering*. Working Paper 51, SFB "Adaptive Information Systems and Modeling in Economics and Management Science". URL `http://www.ci.tuwien.ac.at/~leisch/papers/wp51.ps`.

MESSATFA, H. (1992): An algorithm to maximize the agreement between partitions. *Journal of Classification, 9, 5–15*.

OLIVEIRA, C. A. S. and PARDALOS, P. M. (2004): Randomized parallel algorithms for the multidimensional assignment problem. *Applied Numerical Mathematics, 49(1), 117–133*.

PAPADIMITRIOU, C. and STEIGLITZ, K. (1982): *Combinatorial Optimization: Algorithms and Complexity*. Prentice Hall, Englewood Cliffs.

RAND, W. M. (1971): Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66(336)*.

STREHL, A. and GHOSH, J. (2002): Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research, 3, 583–617*.

VICHI, M. (1999): One-mode classification of a three-way data matrix. *Journal of Classification, 16, 27–44*.