Ensemble Clustering in Medical Diagnostics

Derek Greene, Alexey Tsymbal, Nadia Bolshakova, Pádraig Cunningham

Department of Computer Science, Trinity College Dublin, Ireland

Abstract. Ensemble techniques have been successfully applied in the context of supervised learning to increase the accuracy and stability of classification. Recently, analogous techniques for cluster analysis have been suggested. Research has demonstrated that, by combining a collection of dissimilar clusterings, an improved solution can be obtained. In this paper, we examine the potential of applying ensemble clustering techniques with a focus on the area of medical diagnostics. We present several ensemble generation and integration strategies, and evaluate each approach on a number of synthetic and real-world datasets. In addition, we show that diversity among ensemble members is necessary, but not sufficient to yield an improved solution without the selection of an appropriate integration method.

1 Introduction

Current electronic repositories, especially in medical domains, can contain vast amounts of information. Knowledge discovery and data mining methods have been applied to discover patterns and relations in these complex datasets. Of these, cluster analysis is one of the most important approaches. Such unsupervised learning procedures may be distinguished from other data mining tasks by the unavailability of predefined class labels that partition data. The goal of a clustering algorithm is to expose the underlying structure of the data by uncovering the "natural" groupings of samples.

In the past, cluster analysis in areas such as medical diagnostics has often involved the repeated execution of a clustering procedure, followed by the manual selection of an individual solution that maximises a user-defined criterion. However, rather than merely selecting a "winning" partition, recent work has shown that combining the strengths of an ensemble of clusterings can often yield better results. Ensemble techniques have been successfully applied in supervised learning to improve the accuracy and stability of classification algorithms [2, 13]. However, only recently have attempts been made to apply analogous techniques to domains where class information is unavailable. This research has focused on exploiting the additional information provided by a collection of diverse clusterings to generate a superior partition of the data [8, 11].

In this paper, we evaluate the potential of applying ensemble techniques to several problems of medical diagnostics. We discuss a variety of ensemble generation strategies and integration schemes, and suggest an optimal set of

parameters for each of the datasets under consideration. In addition, we examine the role that diversity plays in producing a successful ensemble.

In Section 2 we introduce the design issues that must be addressed when employing ensemble clustering in practise. Empirical results, based on the application of these techniques to both medical and synthetic datasets, are provided in Section 3. Finally, in Section 4 we conclude and suggest possible directions for future research.

2 Ensemble Design

Ensemble techniques require three key issues to be addressed. Firstly, how does one generate a collection of base clusterings from which the ensemble is composed? Secondly, how many clusterings are required to give a stable accurate solution? Thirdly, how does one combine the ensemble members to produce the final partition? In this section we present an overview of ensemble generation techniques that have been proposed in the recent literature, and discuss suitable integration schemes.

2.1 Ensemble Generation

The first phase of ensemble clustering involves constructing a collection of τ base clustering solutions, denoted $\mathbb{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_\tau\}$, which represent the members of the ensemble. This is typically done by repeatedly applying a chosen clustering algorithm in a manner that leads to diversity among the members. It has been demonstrated [9] that classifier ensembles are most successful when constructed from a set of predictors whose errors occur in different parts of the data space. Diversity is often introduced artificially to improve the output of an ensemble. For many clustering algorithms, the solutions produced over many trials will typically be highly similar. Clearly, if all ensemble members agree on how a dataset should be partitioned, aggregating the clusterings will show no improvement over any of the constituent members.

Several approaches comparable to those used in supervised learning have been proposed to introduce artificial instabilities in clustering algorithms. These generation strategies yield different clusterings of the same data, which can potentially improve the quality and robustness of ensemble output. In this paper, we empirically examine the following ensemble generation strategies:

- *Plain.* A simple approach to producing a collection of ensemble members is to rely solely on some stochastic element in the base clustering algorithm to provide diversity, such as the selection of random initial clusters in k-means.
- *Random-k.* The output of clustering algorithms such as standard k-means is dependent on the initial choice of the number of clusters k. This has been exploited as a source of ensemble diversity by generating clusterings using randomly selected values of k from a user-specified interval [8]. In our experiments, we used the range $[2, k + 10]$, where k is the natural number of clusters for a given dataset.

- *Random-k+.* It has been shown that a collection of clusterings generated at a much higher resolution than the value of k used for the final partition can provide better results [4]. This generation method is the same as *random-k*, but with the interval $[k, k + 30]$.
- *Bagging.* A common solution to the lack of diversity in classifier ensembles is to train individual predictors on random subsamples of the data, as in bagging. An analogous method for ensemble clustering was suggested in [10], where subsets of the original data are produced by randomly selecting objects with replacement.
- *Random subspacing.* Instability can also be introduced to an ensemble by ensuring that individual members have only a partial view of each data point. A simple approach that has been used in classifier ensembles to accomplish this task is random subspacing [5]. This method is also suitable for ensembles in unsupervised learning, where each base clustering is generated on a randomly selected subset of the original dimensions.
- *Random projection.* Another effective ensemble creation method was proposed in [12], involving the generation of a set of dissimilar clusterings by randomly projecting the data onto a lower dimensional subspace. Each ensemble member is produced by transforming the original $n \times m$ dataset to a reduced set of m' new dimensions, based on a randomly generated transformation matrix. In the experiments discussed in this paper, the value of m' for each ensemble member was randomly selected from the interval $[1, m]$.
- *Heterogeneous ensembles.* In *homogeneous* ensembles, members are created using repeated runs of a single base clustering algorithm. As an alternative, *heterogeneous* ensembles may be employed, where diversity is induced by allowing each base clustering to be generated using a different algorithm.

Another design consideration in this context is the choice of one or more base clustering algorithms that will be used to produce each ensemble member. In our experiments, we employed standard k-means, k-medoids and a fast "weak clustering" technique, where k centroids are chosen at random and the remaining objects are assigned to the cluster with the nearest centroid. Unlike k-means and k-medoids, no subsequent attempt is made here to improve the partition, resulting in highly diverse solutions.

2.2 Ensemble Integration

Once a collection of diverse base clusterings has been generated, these clusterings should be aggregated to produce a single solution. An intuitive ensemble integration method is to use the information provided by the different clusterings to determine the level of association between each pair of objects in the dataset [8]. The fundamental assumption here is that objects occurring in the same "natural" cluster will be frequently assigned to the same cluster across the base clusterings.

This *co-association* approach resembles the majority voting schemes commonly used in classifier ensembles. For each base clustering in $\mathbb{C}$, a pair of objects

occurring in the same cluster signifies a "vote" for the pair being co-located in the final partition. The collection of base clusterings can effectively be mapped to a symmetric $n \times n$ *co-association matrix* $\mathbf{M}$, where each entry M_{ij} represents the fraction of times that the pair of objects (x_i, x_j) has been assigned to the same cluster. Once this intermediate representation has been constructed, a standard similarity-based clustering algorithm may be applied to $\mathbf{M}$ to produce a consensus solution. Algorithms that have been employed for this purpose include agglomerative clustering [7] and multi-level graph partitioning [11]. In our experiments, we apply the following hierarchical clustering algorithms: single-linkage, complete-linkage and average-linkage.

3 Experimental Results and Discussion

In order to evaluate the ensemble strategies described in Section 2, experiments were conducted on two synthetic datasets and six benchmark medical datasets from the UCI repository [1]. We examined all possible combinations of base clustering algorithms, generation strategies and integration approaches discussed previously. For each dataset, the co-association approach was used as an integration function, where the final number of clusters k was set to the known number of clusters for the dataset.

3.1 Evaluation of Ensemble Performance

In unsupervised learning, there is no definitive measure of accuracy, making the task of evaluating any ensemble clustering technique non-trivial. Many clustering validation measures are parametric and tend to favour bell-shaped distributions, making them inappropriate for the task of ensemble evaluation. An alternative strategy for cluster validation is to apply the algorithm to a dataset for which a reference partition or "ground truth" is available, typically in the form of predefined class labels. External validation indices make use of this information, unavailable to the clustering algorithm itself, to quantify the level of agreement between the algorithm's output and the set of k' natural classes $\mathcal{C}' = \{C'_1, \ldots, C'_{k'}\}$ in a reference partition. To evaluate the ensemble strategies describe previously, we use two such criteria.

One approach to external validation is to count the pairs of objects for which the clusters and natural classes agree on their co-assignment. A representative index, the *Jaccard coefficient* [6], has been commonly applied to assess the similarity between binary sets. It is also possible for this measure to be used in the context of external validation, where the level of agreement of between the disjoint partitions $\mathcal{C}'$ and $\mathcal{C}$ is given by normalising the number of positive agreements

$$J(\mathcal{C}', \mathcal{C}) = \frac{a}{a+b+c} \tag{1}$$

where a denotes the number of pairs of objects with the same label in $\mathcal{C}'$ and assigned to the same cluster in $\mathcal{C}$, b denotes the number of pairs with the same

Dataset	n	m	k	τ	Base	Generator	Linkage	Jac.	Acc.
2spirals	212	2	2	1000	k-medoids	random-k+	single	1.00	1.00
halfrings	600	3	2	500	weak	plain	single	1.00	1.00
breast	277	51	2	1000	k-means	random-k+	complete	0.59	0.76
diabetes	768	8	2	2000	heterogen.	plain	average	0.55	0.68
heart	270	13	2	2000	weak	plain	single	0.50	0.60
iris	150	4	3	3000	k-medoids	random-k	single	0.78	0.89
liver	345	7	2	2000	weak	plain	single	0.51	0.59
lymph	148	18	4	2000	k-medoids	subspacing	average	0.48	0.62
thyroid	215	5	3	2000	k-means	bagging	single	0.63	0.79

Table 1. Details, optimal ensemble parameters, and corresponding external validation scores for experimental datasets.

label, but in different clusters and c denotes the number of pairs in the same cluster, but with different class labels. This index produces a result in the range $[0, 1]$, where a value of 1 indicates that $\mathcal{C}'$ and $\mathcal{C}$ are identical.

Another external validation approach is to identify a match between each cluster and a corresponding natural class in the reference partition. Therefore, as our second validation index, we consider a simple *accuracy* score that uses external class information in a manner similar to the mis-assignment rate described in [12]. By finding the optimal correspondence between a set of annotated class labels and the clusters in an ensemble partition, a performance measure may be derived that reflects the proportion of objects that were assigned to the correct cluster.

A summary of the datasets under consideration is provided in Table 1, together with optimal ensemble parameters and evaluation results. A minimum number of ensemble members τ required to give a stable solution is also suggested for each dataset. In experiments performed over 30 trials, ensembles with the specified parameters consistently yielded results superior to those produced by single independent runs of individual clustering algorithms. When compared with the output of a single k-means algorithm, the increase in accuracy values ranged from 0.002 on the *thyroid* dataset to 0.453 on the *2spirals* dataset. Similar improvements were also evident in the Jaccard values, with increases ranging from 0.068 on the *thyroid* dataset to 0.667 on the *2spirals* dataset.

3.2 Evaluation of Ensemble Diversity

As noted previously, it has been observed [3, 13] that the success of a supervised ensemble depends not only on the presence of a diverse set of base classifiers, but also on the ability of the meta-level classifier to exploit the resulting diversity. To assess the relationship between methods for creating ensemble members and diversity, we examined the amount of disagreement resulting from each of the generation techniques and base clustering algorithms considered previously. To quantify diversity, we use a non-pairwise entropy measure based on that proposed

for classifier ensembles in [13]. Formally, the diversity of a collection of base clusterings $\mathbb{C}$ constructed on a dataset of n objects is given by the expression

$$div\_ent(\mathbb{C}) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} -\left(M_{ij} \log_2 M_{ij} + (1 - M_{ij}) \log_2 (1 - M_{ij})\right) \tag{2}$$

where M_{ij} represents the fraction of times the objects x_i and x_j are co-located in the same cluster.

In each case, we examined the amount of disagreement between 1000 base clusterings, averaged across the datasets. Figure 1 provides a comparison of the level of diversity produced by each ensemble generator and the average performance of ensembles utilising the technique. In Figure 2 we show a comparison of the diversity afforded by the three base clustering algorithms. Note that, while weak-clustering only shows small improvements when averaging across the datasets, it proved to be the optimal choice for the base algorithm on several

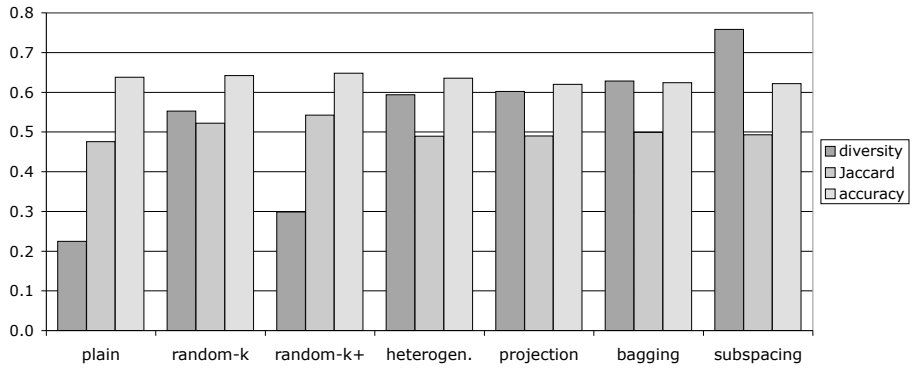

Fig. 1. Comparison of accuracy and diversity scores for ensemble generation strategies.

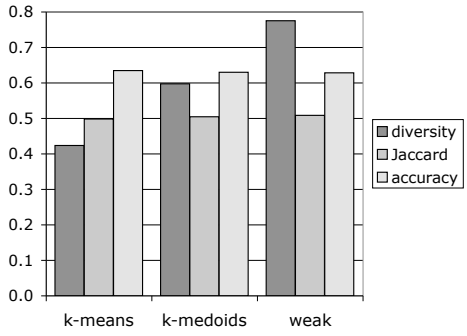

Fig. 2. Comparison of accuracy and diversity scores for base clustering algorithms.

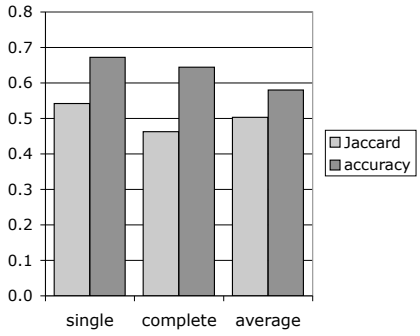

Fig. 3. Comparison of accuracy scores for integration strategies.

datasets. Since it provides a high level of diversity, it may be used in conjunction with the "plain" generation strategy, making it less computationally expensive than the other methods.

Both Figures 1 and 2 indicate that, while combining the output of multiple clusterers is useful only if there is disagreement between the partitions they produce, diversity alone is not sufficient to yield an improved solution. Rather, the choice of a suitable integration strategy appears to greatly dictate the success of an ensemble. Figure 3 shows that, in our studies, single-linkage gave the best average performance across the datasets. For practical applications, each of the integration algorithms could be applied, with the best partition selected by combining the output of multiple consensus functions [11].

4 Conclusion

In this paper, we discussed ensemble clustering and conducted a series of experiments on synthetic and real-world datasets, examining a range strategies for generating and integrating the ensembles. We also suggested an optimal configuration for each dataset that resulted in consistent improvements over single independent runs of individual clustering algorithms.

We have demonstrated that ensemble clustering offers considerable potential to improve our ability to identify the underlying structure of both artificial and real datasets in unsupervised scenarios. However, it is apparent from our results that the ability to exploit this potential relies to a great extent on making several important design decisions relating to the choice of base clustering algorithm, generation technique, number of ensemble members and final integration algorithm. In addition, we have observed that diversity among ensemble members is necessary, but not sufficient to yield an improved solution without the selection of an effective integration scheme.

Future research could consider other combination strategies that may be more successful in exploiting diversity. Alternative methods of quantifying diversity could also be investigated, such as pairwise or variance-based measures. An important aspect of ensemble clustering that remains to be explored is the relationship between the various measures of ensemble performance and the accuracy of its constituent members, which could provide further insight into the process of selecting appropriate ensemble parameters.

Acknowledgements: This material is based upon works supported by the Science Foundation Ireland under Grant No. SFI-02IN.1I111. We would like to thank the UCI machine learning repository of databases, domain theories and data generators for the datasets used in this study.

References

1. C. Blake and C. Merz. UCI repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Science, 1998.

2. L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
3. C. Brodley and T. Lane. Creating and exploiting coverage and diversity. In *Proc. AAAI Workshop on Integrating Multiple Learned Models*, pages 8–14, Portland, Oregon, 1996.
4. J. Ghosh, A. Strehl, and S. Merugu. A consensus framework for integrating distributed clusterings under limited knowledge sharing. In *Proc. NSF Workshop on Next Generation Data Mining*, pages 99–108, November 2002.
5. T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.
6. P. Jaccard. The distribution of flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.
7. A. K. Jain and A. Fred. Data clustering using evidence accumulation. In *Proc. 16th International Conference on Pattern Recognition (ICPR'02)*, volume 4, pages 276–280, December 2002.
8. A. K. Jain and A. Fred. Evidence accumulation clustering based on the k-means algorithm. *Structural, Syntactic, and Statistical Pattern Recognition*, LNCS 2396:442–451, 2002.
9. A. Krogh and J. Vedelsby. Neural network ensembles, cross validation and active learning. *Advances in Neural Information Processing Systems*, pages 231–238, 1995.
10. F. Leisch. Bagged clustering. Working Paper 51, SFB "Adaptive Information Systems and Modeling in Economics and Management Science", August 1999.
11. A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining partitionings. In *Proc. Conference on Artificial Intelligence (AAAI'02)*, pages 93–98. AAAI/MIT Press, July 2002.
12. A. Topchy, A. Jain, and W. Punch. Combining multiple weak clusterings. In *Proc. Third IEEE International Conference on Data Mining (ICDM'03)*, pages 331–338, Melbourne, Florida, November 2003.
13. A. Tsymbal, M. Pechenizkiy, and P. Cunningham. Diversity in ensemble feature selection. Technical report, Trinity College Dublin, 2003. http://www.cs.tcd.ie/publications/tech-reports/reports.03/TCD-CS-2003-44.pdf.