Constraint Selection by Committee: An Ensemble Approach to Identifying Informative Constraints for Semi-Supervised Clustering

Derek Greene, Pádraig Cunningham

University College Dublin, Ireland
{derek.greene,padraig.cunningham}@ucd.ie

Abstract. A number of clustering algorithms have been proposed for use in tasks where a limited degree of supervision is available. This prior knowledge is frequently provided in the form of pairwise must-link and cannot-link constraints. While the incorporation of pairwise supervision has the potential to improve clustering accuracy, the composition and cardinality of the constraint sets can significantly impact upon the level of improvement. We demonstrate that it is often possible to correctly "guess" a large number of constraints without supervision from the co-associations between pairs of objects in an ensemble of clusterings. Along the same lines, we establish that constraints based on pairs with uncertain co-associations are particularly informative, if known. An evaluation on text data shows that this provides an effective criterion for identifying constraints, leading to a reduction in the level of supervision required to direct a clustering algorithm to an accurate solution.

1 Introduction

Recently, a considerable amount of attention has been paid to the application of machine learning algorithms in problems that do not perfectly correspond to the standard distinction between supervised and unsupervised learning [1]. In many domains, a limited degree of background knowledge will be available when performing exploratory data analysis. While this may take the form of labelled training data, in other situations a simpler type of supervision will be available that describes the relations between pairs of data objects. The latter is commonly represented as a set of pairwise constraints, where each constraint indicates that two objects should either always be assigned to the same cluster (*must-link*) or never be assigned together (*cannot-link*). A number of popular clustering algorithms, such as standard k-means, have been adapted to incorporate this type of information. While the addition of pairwise supervision has the potential to improve clustering accuracy, the choice of constraints will often dictate the level of improvement attained [2]. Many semi-supervised clustering tasks will be active in nature, where the constraint oracle takes the form of a human expert. In such applications, the number of queries for constraints that can be made will be strictly limited.

In this paper, we tackle the problem of identifying constraints that are "informative" in the context of semi-supervised clustering. That is, we seek constraints that will be most effective in guiding a clustering algorithm to produce more accurate solutions. We differentiate these from constraints whose presence does not lead to any noticeable improvement in clustering accuracy. To make this distinction, we firstly establish a connection between hard pairwise constraints and the frequency of co-assignment, or *co-association*, between pairs of objects in an ensemble of clusterings. Specifically, Section 3.1 describes a process by which it is often possible to "guess" or impute a large number of constraints without supervision by examining these co-association values. Following from this, in Section 3.2 we propose a new approach for selecting informative constraints by identifying objects whose cluster assignments are ambiguous. In Section 4 we evaluate this approach on text data, where it is shown to lead to a reduction in the number of actual oracle queries required to produce a significant improvement in clustering accuracy.

2 Related Work

2.1 Semi-Supervised Clustering

Given a set of n data objects $\mathcal{X} = \{x_1, \ldots, x_n\}$, a common representation for background information pertaining to $\mathcal{X}$ is in the form of pairwise constraint sets: must-link constraints $\mathcal{M}$ and cannot-link constraints $\mathcal{C}$. This information can be incorporated into traditional partitional clustering algorithms by adapting the objective function to include penalties for violated constraints. For instance, the Pairwise Constrained k-means (PCKM) algorithm [2] modifies the standard sum of squared errors function to take into account both object-centroid distortions in a clustering $\mathcal{P} = \{\pi_1, \ldots, \pi_k\}$ and any associated constraint violations

$$J_{pckm}(\mathcal{P}) = \sum_{c=1}^{k} \sum_{x_i \in \pi_c} ||x_i - \mu_c||^2 + \sum_{(x_i,x_j)\in\mathcal{M}, li \neq lj} w_{ij} + \sum_{(x_i,x_j)\in\mathcal{C}, li=lj} \bar{w}_{ij} \quad (1)$$

where μ_c is the centroid of the cluster π_c, and l_i denotes the cluster label of the object x_i in $\mathcal{P}$. The weight w_{ij} signifies the size of the penalty incurred when a must-link constraint between a pair (x_i, x_j) is violated, while $\bar{w}_{ij}$ is the penalty for violating a cannot-link constraint between the pair. These weights control the influence given to external information during the assignment phase of the algorithm. The objective (1) has been shown to have a probabilistic basis related to the assignment of labels in Hidden Markov Random Fields (HMRFs).

As with standard partitional algorithms, the choice of initialisation strategy for semi-supervised methods such as PCKM can greatly affect clustering accuracy. An effective strategy in this context involves computing the transitive closure of the graph formed by the constraints in $\mathcal{M}$, and using the centroids of the resulting λ neighbourhoods. If $\lambda > k$, where k is the desired number of clusters, then a weighted variant of farthest-first initialisation may be employed to select a subset of k well-separated centroids [3].

While research in the area of semi-supervised clustering has largely focused on the development of new clustering algorithms, relatively little emphasis has been placed on the important issue of selecting useful constraints. An initial foray into this area was made with the two-stage *Explore and Consolidate* (E&C) approach proposed by Basu *et al.* [2]. In the exploration stage, a set of k initial well-separated neighbourhoods is identified, each of which belongs to be a different natural class. Once the neighbourhoods have been formed, the consolidation stage proceeds by randomly selecting unlabelled objects and assigning them to correct neighbourhoods in a manner that requires as few constraints as possible. The resulting centroids and constraint sets were used to provide supervision for the PCKM algorithm.

2.2 Ensemble Clustering

It has been shown that combining the strengths of a diverse set of clusterings can often yield more accurate and robust solutions [4]. Unsupervised ensemble approaches typically involves two phases: a *generation* phase where a collection of base clusterings is produced, and an *integration* phase where an aggregation function is applied to the ensemble members to produce a consensus solution. The most frequently employed integration strategy has been to use the information provided by an ensemble to determine the level of association between pairs of objects in a dataset [4, 5]. The fundamental assumption underlying this strategy is that pairs belonging to the same natural class will frequently be co-assigned during repeated executions of a clustering algorithm. In practice, these pairwise co-associations are represented using a symmetric co-association matrix. A consensus solution is recovered by applying a similarity-based algorithm to the matrix, such as single-linkage agglomerative clustering.

Pairwise co-association values have also been used to gather information from unlabelled data in order to improve the performance of kernel-based classification algorithms. The *bagged cluster kernel* technique proposed by Weston *et al.* [6] involves modifying a base kernel to include co-association information aggregated from multiple k-means clusterings, which are generated on bootstrap samples.

2.3 Uncertainty Sampling

For many learning problems, large numbers of training examples will not be available due to the expense of providing class labels. In these cases, active learning techniques can be employed to identify and label informative data objects that will serve to maximise classification accuracy. One approach to active learning that has widely been used is *uncertainty sampling* [7], where unlabelled objects are prioritised based upon the level of uncertainty regarding their class membership. An intuitive basis for measuring uncertainty is to consider the disagreement between the predictions made by a committee of classifiers [8]. For instance, Melville & Mooney [9] suggested measuring the uncertainty for an unlabelled object based on the margin between its maximum class probability and the probability of the next best competing class.

3 Constraint Identification

The composition and cardinality of the sets $\mathcal{M}$ and $\mathcal{C}$ can significantly impact upon the improvements achieved by semi-supervised algorithms. In addition, as the number of data objects n increases, the number of possible constraints also significantly increases. If constraints are selected at random, many oracle queries may be required before any noticeable improvement in clustering accuracy is achieved. To illustrate this, Figure 1 shows the effect of adding constraints for randomly chosen pairs on the normalised mutual information (NMI) [4] scores produced when the PCKM algorithm is applied to the *3-news-similar* dataset. Even after the addition of 1000 constraints, little significant increase in accuracy is evident. For many semi-supervised tasks, it will be the case that the oracle is a human expert. Since it is unrealistic to expect a human to respond to so many queries, an intelligent strategy for choosing constraints is desirable.

3.1 Imputing Constraints from Pairwise Co-associations

When seeking to choose a small set of highly informative constraints, it may be helpful to eliminate those "easy" constraints that can be found without the aid of a supervisor. In this section, we show that it is possible to identify such constraints by examining the relationship between pairs of objects over a large collection of base clusterings, denoted $\mathbb{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_\tau\}$.

Like their supervised counterparts, it has been demonstrated that unsupervised ensembles are most effective when constructed from solutions that are both accurate and diverse [4]. To encourage diversity, a commonly employed strategy has been to apply a partitional clustering algorithm, such as standard k-means, to different subsamples of the same dataset. In practice, typically 60-80% of the data is included when generating each base clustering. After each sample is clustered, membership assignments for the out-of-sample objects are determined by applying a suitable classification scheme, such as a nearest centroid classifier.

Once an ensemble $\mathbb{P}$ has been generated, it is customary to represent the co-assignments between objects across all clusterings in the form of a symmetric

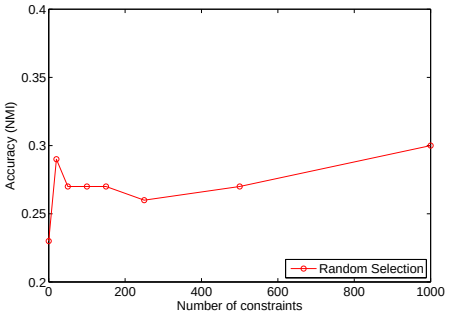


Fig. 1. Effect of randomly selected constraints on *3-news-similar* dataset.

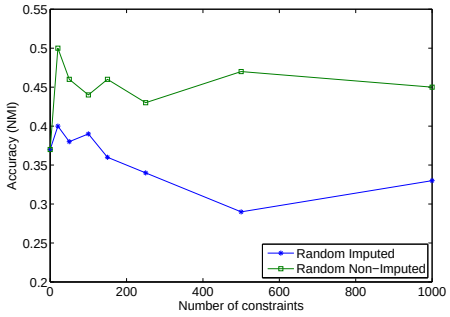


Fig. 2. Effect of randomly selected imputed constraints on *3-news-related* dataset.

1. Initialise $\mathbf{A}$ as the $n \times n$ empty co-association matrix for the dataset $\mathbf{X}$.
2. For $t = 1$ to τ:
   1. Draw a sample of objects $\mathcal{X}_t$ by random sampling without replacement.
   2. Generate a base clustering $\mathcal{P}_t$ by clustering the sample $\mathcal{X}_t$.
   3. Classify each out-of-sample object based on the clusters in $\mathcal{P}_t$.
   4. For each pair (x_i, x_j) assigned to the same cluster in $\mathcal{P}_t$, update $\mathbf{A}$:

$$A_{ij} = A_{ij} + 1/\tau$$

3. Construct imputed constraint sets $(\mathcal{M}', \mathcal{C}')$ based on the co-association of each unique pair (x_i, x_j), according to the rule given in Eqn. 2.

Fig. 3. Pairwise constraint imputation procedure.

$n \times n$ co-association matrix $\mathbf{A}$. In this matrix, an entry $A_{ij} \in [0, 1]$ denotes the fraction of clusterings in $\mathbb{P}$ in which the objects x_i and x_j were assigned to the same cluster. Both the ensemble and cluster kernel techniques discussed in Section 2.2 are motivated by the assumption that the matrix $\mathbf{A}$ encodes information describing the probability or confidence with which a pair of objects will be grouped together in the natural classes of the data. For a sufficient number of base clusterings, a value $A_{ij} \approx 1$ is a strong indicator that the pair (x_i, x_j) belong to the same class, while $A_{ij} \approx 0$ indicates that they belong to different classes. On the other hand, a value $A_{ij} \approx 0.5$ implies that we are highly unsure about whether the objects are actually conceptually related to one another.

When performing ensemble clustering, we are typically interested in producing a complete disjoint partition of the data. This can result in "uncertain" pairs being grouped together in the consensus clustering. For instance, the integration approach described in [5] only requires a value $A_{ij} > 0.5$ for a pair to be placed in the same consensus cluster. However, given the goal of accurately deducing constraints, we focus on pairs with unambiguous associations. By thresholding the values in $\mathbf{A}$, we can eliminate uncertain pairs from consideration. Specifically, we choose a threshold value κ_m for must-link constraints, which represents the minimum level of confidence required for the co-assignment of two objects, and a threshold κ_c for cannot-link constraints, which represents the maximum level of uncertainty allowed when concluding that two objects are unrelated. Both values lie in the range $[0, 1]$, with the natural requirement that $\kappa_m >> \kappa_c$. Formally, we construct imputed constraint sets $(\mathcal{M}', \mathcal{C}')$ by using the rule:

$$\begin{array}{ll} A_{ij} \geq \kappa_m & \text{add } (x_i, x_j) \text{ to } \mathcal{M}' \\ A_{ij} \leq \kappa_c & \text{add } (x_i, x_j) \text{ to } \mathcal{C}' \\ \kappa_c < A_{ij} < \kappa_m & \text{ignore.} \end{array} \tag{2}$$

The application of the method outlined in Figure 3 frequently produces constrained pairs that correspond to those generated from natural class labels.

While ensemble clustering can often provide a more comprehensive picture of the natural structures present in a dataset, it is interesting to note that constraints imputed in this way, even if correct, will rarely prove directly useful in semi-supervised clustering. Surprisingly, in some situations these constraints can actually prove harmful. We suggest that this phenomenon is due to the fact that these easily imputed constraints leave regions in certain underlying classes under-represented, so that initial clusters resulting from the imputed must-link constraints are skewed. As an example, Figure 2 shows the effect of randomly adding constraints from the set of pairs that were correctly imputed on the *3-news-related* dataset. Here, the addition of a large number of imputed constraints actually results in less accurate solutions. In contrast, when randomly choosing from among non-imputed pairs, the quality of the resulting clusterings increases.

3.2 Selecting Informative Constraints

Motivated by the observations made in the previous section, we now describe a new ensemble-based selection procedure that makes use of pairwise co-associations to focus on informative constraints. This procedure consists of two phases: firstly we use the imputed set $\mathcal{M}'$ to identify a set of representative objects $\{r_1, \ldots, r_k\}$ which correspond to distinct classes in the data; subsequently we construct clusters around these representatives by adding constraints relating to objects whose cluster assignments are difficult to determine.

While imputed constraints may not be directly useful for semi-supervised clustering, they do provide a starting point for finding representative objects. This can be achieved by examining the set of neighbourhoods produced by computing transitive closure of the imputed must-link constraints in $\mathcal{M}'$. We frequently observe that the largest neighbourhoods produced in this way will correspond to distinct natural classes in the data. This provides a basis for selecting representatives for k different classes using only a small number of oracle queries. Firstly, the neighbourhoods are arranged in descending order by size, and the median object of each neighbourhood is identified (*i.e.* the object nearest the neighbourhood centroid). The median of the largest neighbourhood is

---

1. Identify imputed constraint sets $(\mathcal{M}', \mathcal{C}')$ from a co-association matrix $\mathbf{A}$.
2. Compute the transitive closure of $\mathcal{M}'$, and identify the median objects of each neighbourhood based on the values in $\mathbf{A}$.
3. Choose the first representative r_1 to be the median of the largest neighbourhood.
4. For $c = 2$ to k:
   - Select r_c as the median of the next largest neighbourhood, such that a cannot-link constraint exists between r_c and each of $\{r_1, \ldots, r_{c-1}\}$.
5. Output a clustering $\mathcal{P} = \{\pi_1, \ldots, \pi_k\}$, where $r_c \in \pi_c$, together with any other object with a must-link constraint to r_c.

Fig. 4. Constraint set initialisation phase.

elected to be the first representative r_1. Each of the $(k-1)$ other representatives is chosen to be the median object of the largest remaining neighbourhood, such that a cannot-link constraint exists between that median and all previously selected representatives (*i.e.* it belongs to a new class). The application of this initialisation scheme leads to an initial clustering $\mathcal{P} = \{\pi_1, \ldots, \pi_k\}$, where $r_c \in \pi_c$. Any objects involved in must-link constraints are also assigned to the appropriate cluster in $\mathcal{P}$. The complete initialisation procedure is outlined in Figure 4. A particular advantage of this approach is that, even if constraints are only available for a subset of objects, good representatives can be identified using imputed neighbourhoods derived from clusterings of the entire dataset.

In the second phase of the proposed constraint selection procedure, we expand the clustering $\mathcal{P}$ by incrementally assigning objects using pairwise supervision. Objects are processed using an ordering based upon the level of uncertainty regarding their association to the existing clusters, thereby prioritising those objects for which queries to an external oracle are particularly necessary. Formally, let $\mathbf{S} \in \mathbb{R}^{n \times k}$ denote the object-cluster association matrix, such that S_{ic} is the mean co-association between the object x_i and the members of the cluster π_c:

$$S_{ic} = \frac{1}{|\pi_c|} \sum_{x_j \in \pi_c} A_{ij} \tag{3}$$

To evaluate the degree of uncertainty in assigning an object to a cluster in $\mathcal{P}$, we use a criterion based on the well-known *silhouette index* [10], which is often employed in internal cluster validation. Rather than using distance values computed on the raw data, we consider the margin between competing clusters based on object-cluster associations. Specifically, for a candidate query object x_i, let π_a denote the cluster with which it has the highest level of association, and let π_b denote the next best alternative cluster. The certainty of the assignment of x_i can be measured using the expression:

$$u(x_i) = \frac{2 \cdot S_{ia}}{S_{ia} + S_{ib}} - 1 \tag{4}$$

Since it is always the case that $S_{ia} \geq S_{ib}$, Eqn. 4 produces an evaluation in the range $[0, 1]$, where a smaller value is indicative of a greater degree of uncertainty.

Unfortunately, if objects are chosen based on an ordering of the uncertainty scores $u(x_i)$, this can potentially result in the generation of a large succession of constraints for a single natural class. The use of such unbalanced constraint sets can reduce the performance gain achieved by semi-supervised algorithms when using small constraint sets. To address this problem, we introduce a bias in favour of under-represented classes. This is accomplished by weighting object-cluster association values with respect to cluster size, leading to an adjusted certainty criterion

$$w(x_i) = \frac{2 \cdot T_{ia}}{T_{ia} + T_{ib}} - 1 \quad \text{such that} \quad T_{ic} = \frac{|\pi_c|}{\sum_j |\pi_j|} \cdot S_{ic} \tag{5}$$

where T_{ia} denotes the maximum weighted object-cluster association value, and T_{ib} is the next highest value. This weighting has the effect of producing higher

1. Update the object-cluster association matrix **S**.
2. Select the next most uncertain object x_i with the minimum value for $w(x_i)$, as calculated using Eqn. 5.
3. Arrange the clusters in descending order using the values in the i-th row of **S**.
4. For each cluster π_c:
   - Query the oracle for the pair (x_i, r_c) until a must-link constraint is found.
5. Assign x_i to the cluster containing the correct representative.
6. Repeat from Step 1 until no further oracle queries are possible.

Fig. 5. Constraint set expansion phase.

scores for objects that have strong associations with large clusters. Since objects with lower scores are prioritised, this encourages the selection of constraints for objects that are likely to be assigned to smaller clusters.

Once an unassigned object x_i has been selected based on the minimal value for Eqn. 5, its correct cluster in $\mathcal{P}$ is found by querying the oracle for constraints between x_i and each of the k representatives. Following the observations made in [2], it is apparent that the correct cluster can be located using at most $(k-1)$ queries. We can potentially further reduce the number of queries required by sorting the values in the i-th row of **S** in descending order. Candidate clusters are processed in this order until a must-link pair (x_i, r_c) is generated. If such a constraint is not found after $(k-1)$ queries, it can be assumed that the object belongs to the final cluster without the requirement for an additional query. After assigning x_i to the correct cluster, uncertainty scores for the remaining objects are recalculated. An outline of the expansion phase is provided in Figure 5.

4 Evaluation

In this section, we describe the results of two sets of experimental evaluations conducted on text data. Firstly, we assess the veracity of constraints imputed using the approach discussed in Section 3.1. In the second set of experiments, we evaluate the performance of semi-supervised clustering when constraints are selected using the ensemble-based procedure proposed in Section 3.2.

Both sets of experiments were performed on six text corpora, which present different degrees of difficulty when performing document clustering. The *bbc* corpus contains news articles pertaining to five topical areas: business, entertainment, politics, sport and technology. The *bbcsport* corpus consists of a smaller set of sports news articles from the same source[1]. The *cstr* dataset[2] represents a small collection of technical abstracts. The *3-news-related* dataset (also referred to as *ng17-19*) is a commonly used subset of the *20-newsgroups* collection[3], con-

[1] Both available from http://mlg.ucd.ie/datasets/
[2] Original abstracts available from http://www.cs.rochester.edu/trs
[3] Available from http://people.csail.mit.edu/jrennie/20Newsgroups/

sisting of three groups pertaining to politics that exhibit some overlap. Another benchmark subset, the *3-news-similar* dataset, consists of three IT-related newsgroups that overlap significantly. The *reuters5* dataset is a subset of the widely-used *Reuters-21578* corpus of news articles, containing documents from the five largest categories. To pre-process the datasets, we applied standard stop-word removal and stemming techniques. We subsequently removed terms occurring in less than three documents and applied log-based *tf-idf* normalisation.

4.1 Validation of Imputed Constraints

To investigate the effectiveness of the constraint imputation technique, we generated an ensemble consisting of 2000 members for each dataset. These clusterings were formed by applying standard k-means with cosine similarity and random initialisation to samples of documents, using a subsampling rate of 80%. We subsequently constructed a co-association matrix and a corresponding set of imputed constraints for each corpus by following the procedure outlined in Figure 3. In practice, we found that conservative thresholds of $\kappa_m = 0.98$ and $\kappa_c = 0$ were suitable for use with a variety of text datasets.

Table 1 presents details of the imputed must-link and cannot-link constraint sets generated for each dataset. Note that the numbers reported do not take into account any additional cannot-link constraints that can be inferred from the imputed must-link constraints. We compare the imputed sets to the correct pairwise relations defined by the natural classification of the datasets, using measures of *pairwise precision* (PP) and *pairwise recall* (PR). Given an imputed set $\mathcal{Y}'$, the former refers to the fraction of imputed pairs that are correctly constrained, while the latter represents the fraction of the complete set $\mathcal{Y}$ recovered:

$$PP(\mathcal{Y}', \mathcal{Y}) = \frac{|\mathcal{Y}' \cap \mathcal{Y}|}{|\mathcal{Y}'|} \qquad PR(\mathcal{Y}', \mathcal{Y}) = \frac{|\mathcal{Y}' \cap \mathcal{Y}|}{|\mathcal{Y}|} \tag{6}$$

On each of the datasets considered, a large number of must-link and cannot-link constraints are correctly imputed. In all but one case, pairwise precision scores of 0.9 or higher were achieved for both constraint types. Table 1 also lists the mean NMI scores of the base clusterings in each ensemble. It is interesting to

Table 1. Details of imputed constraint sets for text datasets.

Dataset	n	Base	Must-Link			Cannot-Link		
		NMI	Selected	PP	PR	Selected	PP	PR
bbc	2225	0.80	191619	0.98	0.38	1021257	1.00	0.52
bbcsport	737	0.71	4842	1.00	0.08	19516	1.00	0.09
cstr	505	0.64	4389	0.99	0.12	40874	0.99	0.44
reuters5	2317	0.46	145336	0.94	0.15	1202021	0.91	0.61
3-news-related	2625	0.41	245886	0.90	0.19	12620	1.00	0.01
3-news-similar	2938	0.22	17761	0.67	0.01	3025	0.95	0.01

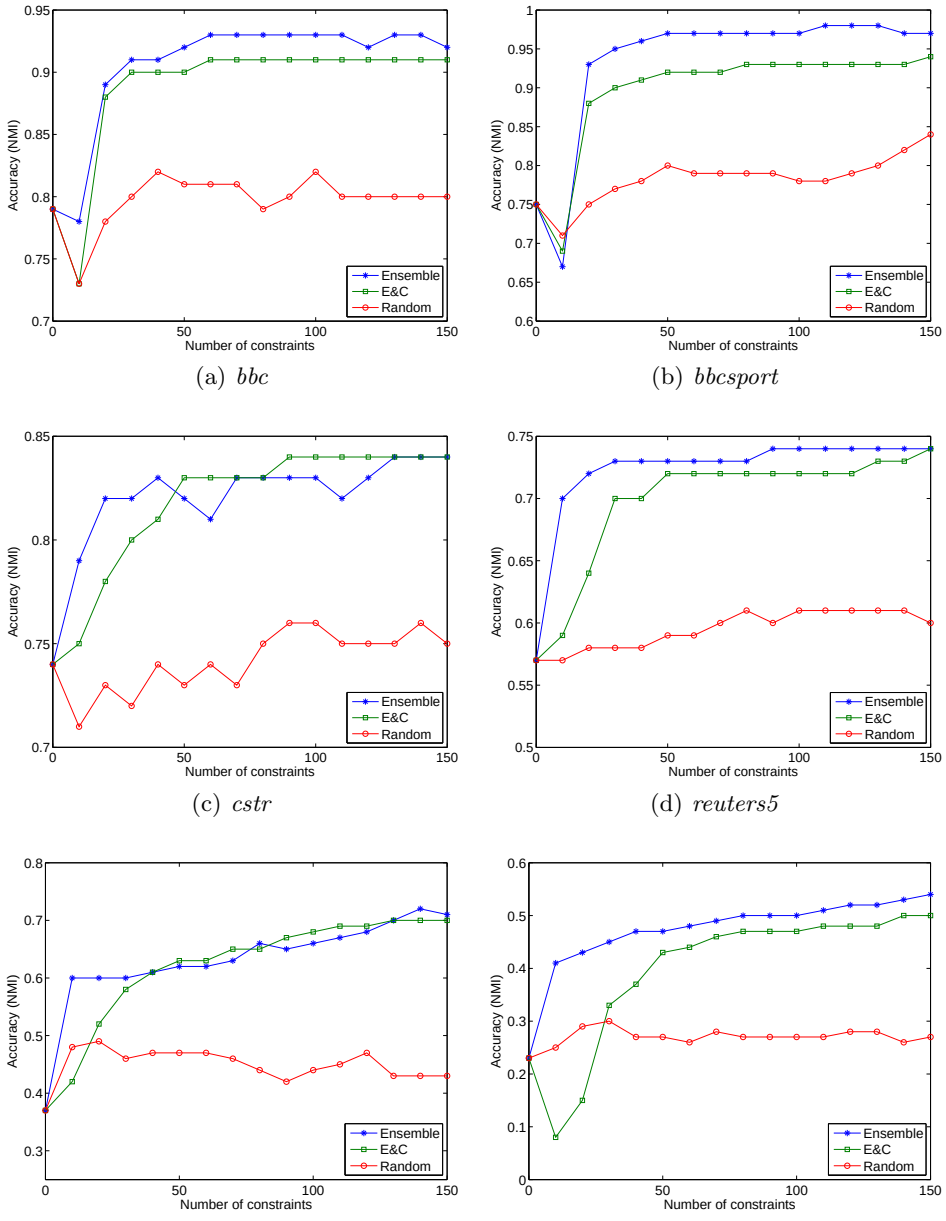
observe that, even when the quality of the base clusterings used to construct a co-association matrix is poor, it is still possible to produce an accurate set of imputed constraints. Due to the use of conservative threshold values in the imputation process, the level of recall is significantly lower than the level of precision. However, for all the datasets under consideration, the number of correctly imputed pairs is significantly higher than the number of constraints we could expect to be provided by a human resource.

4.2 Constraint Selection Evaluation

We now compare the performance of the constraint selection approach proposed in Section 3.2 with that of two alternative strategies. The first is the *Explore and Consolidate* (E&C) approach described in [2]. As a baseline, we also consider the random selection of constraints from all the available pairs in the data. As our choice of semi-supervised clustering algorithm, we employ PCKM with cosine similarity, and set the value of k to correspond to the number of natural classes in the data. When evaluating the case where no constraints are present, initial centroids are selected for the clustering algorithm using farthest-first initialisation. The constraint selection approaches were evaluated over 50 two-fold cross validation trials. As an oracle, we use the natural classification supplied for each dataset. Each oracle query results in either a must-link or a cannot-link constraint. In each trial, constraints are available for 90% of the data, while the remaining 10% of the data constitutes the test set. In the assignment phase of PCKM, all constraints are given an equal weighting of 0.001.

Figure 6 shows a comparison of the mean NMI scores achieved by the three constraint selection strategies when applied to the six datasets under consideration. Note that the reported scores are calculated solely based on the assignment of objects in the test set. We focus on the performance of the three selection strategies for the first 150 queries, since the selection of a larger number of constraints by a human oracle in this context is unrealistic. For all three methods, the points on the validation plots indicate the mean NMI score achieved using the first p selected constraints. This ensures that each method has the same level of supervision.

Firstly, it is clear that both the E&C and ensemble strategies represent significantly better options than simply choosing constrained pairs at random. For data with poorly separated clusters, such as the *3-news-related* and *3-news-similar* datasets, little improvement in clustering accuracy is evident after 150 random queries. In contrast, the ensemble strategy leads to a significant increase in accuracy, even after the addition of only 10 constraints. In general, we observed that ensemble-based selection led to greater increases in accuracy after the first 10–30 constraints than afforded by the E&C technique. This may be attributed to the selection of good representatives based on imputed must-link constraints, and, in particular, the use of the weighted uncertainty criterion (5) to encourage the selection of constraints from under-represented classes. For the *bbc* and *bbcsport* datasets, both intelligent selection methods did result in an initial drop in accuracy when using a very small number of constraints. However,



(a) *bbc* (b) *bbcsport*

(c) *cstr* (d) *reuters5*

(e) *3-news-related* (f) *3-news-similar*

Fig. 6. Comparison of mean accuracy (NMI) scores for constraint selection strategies when applied to text datasets.

the subsequent increases in accuracy were substantial. It is interesting to note that the recall of the imputed constraints did not have a direct impact on the choice of suitable representatives for the first phase of ensemble-based selection. Also, in the case of the *3-news-similar* dataset, which achieved a relatively low level of pairwise precision as shown in Table 1, both the imputed constraints and the related co-association values still proved useful when selecting real constraints. While initialising the proposed ensemble approach does require more time than the E&C strategy, the running times were not prohibitive in practice. We suggest that, for many applications, the cost of additional machine cycles will be less than the expense of making additional queries to a human oracle.

5 Conclusion

In this paper, we demonstrated that it is often possible to correctly impute sets of pairwise constraints for data by examining the co-associations in an ensemble of clusterings. Furthermore, we proposed a new approach for selecting informative constraints for use in semi-supervised clustering tasks, based upon the uncertainty of object-cluster associations. Evaluations on text data have shown this approach to be effective in improving clustering accuracy, particularly when working with a small number of constraints. We suggest that the notion of imputed constraints may also be relevant in other contexts, such as when integrating information from different feature spaces, or where prior knowledge is available in the form of one or more existing clusterings of the data.

References

1. Chapelle, O., Schölkopf, B., Zien, A., eds.: Semi-Supervised Learning. MIT Press, Cambridge (2006)
2. Basu, S., Banerjee, A., Mooney, R.: Active semi-supervision for pairwise constrained clustering. In: Proc. 4th SIAM Int. Conf. Data Mining. (2004) 333–344
3. Basu, S., Bilenko, M., Mooney, R.J.: A probabilistic framework for semi-supervised clustering. In: Proc. 10th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining. (2004) 59–68
4. Strehl, A., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. J. Machine Learning Research **3** (2002) 583–617
5. Fred, A.: Finding consistent clusters in data partitions. In: Proc. 2nd Int. Workshop on Multiple Classifier Systems (MCS'01). Volume 2096. (2001) 309–318
6. Weston, J., Leslie, C., Ie, E., Zhou, D., Elisseeff, A., Noble, W.: Semi-supervised protein classification using cluster kernels. Bioinformatics **21**(15) (2005) 3241–3247
7. Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Proc. 11th Int. Conf. Machine Learning. (1994) 148–156
8. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proc. 5th Workshop on Computational Learning Theory, Morgan Kaufmann (1992) 287–294
9. Melville, P., Mooney, R.: Diverse ensembles for active learning. In: Proc. 21st Int. Conf. Machine Learning. (2004) 584–591
10. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Computational and Applied Mathematics **20**(1) (1987) 53–65