Efficient Prediction-Based Validation for Document Clustering

Derek Greene, Pádraig Cunningham

University of Dublin, Trinity College,
Dublin 2, Ireland
{derek.greene,padraig.cunningham}@cs.tcd.ie

Abstract. Recently, stability-based techniques have emerged as a very promising solution to the problem of cluster validation. An inherent drawback of these approaches is the computational cost of generating and assessing multiple clusterings of the data. In this paper we present an efficient *prediction-based* validation approach suitable for application to large, high-dimensional datasets such as text corpora. We use kernel clustering to isolate the validation procedure from the original data. Furthermore, we employ a *prototype reduction* strategy that allows us to work on a reduced kernel matrix, leading to significant computational savings. To ensure that this condensed representation accurately reflects the cluster structures in the data, we propose a *density-biased* selection strategy. This novel validation process is evaluated on a large number of real and artificial datasets, where it is shown to consistently produce good estimates for the optimal number of clusters.

1 Introduction

The task of evaluating the output of a clustering algorithm, referred to as cluster validation, is a fundamental problem in unsupervised learning. One common application of validation is in the identification of suitable values for algorithm parameters such as the optimal number of clusters $\hat{k}$. Internal validation indices, which make assessments based on intrinsic properties of the raw data, have frequently been used for this task in the past [1]. However, many of these indices make assumptions about the distribution of clusters and are only useful when used in conjunction with certain distances measures. On the other hand, *external* validation techniques, which make use of *a priori* information to evaluate clustering accuracy, are not directly applicable here as external knowledge will typically be unavailable during the clustering process.

Recently, methods based on stability analysis have proved popular for the task of model selection. The *stability* of a clustering model refers to its ability to consistently replicate similar solutions on data originating from the same source [2]. Since there is often only a single set of data available in unsupervised learning scenarios, solutions are typically obtained by clustering subsamples of the original dataset. If the solutions on different samples agree, we may conclude that the model is appropriate for the data. A related approach for estimating

the optimal number of clusters was proposed by Tibshirani *et al.* [3], which is motivated by the concept of prediction accuracy in supervised learning. This *prediction-based* validation scheme estimates $\hat{k}$ by assessing, for each candidate value k, the degree to which we can consistently construct a classifier on a training set that will accurately predict the assignment of objects in a clustering of a corresponding test set.

A key advantage of these methods lies in their ability to evaluate clustering solutions without making assumptions about the true cluster structures in the data. However, from a computational perspective, the use of stability analysis in cluster validation has significant drawbacks. Due to the time required to generate and compare multiple clusterings of the data, such methods have rarely been applied to high-dimension, large-scale datasets such as text corpora.

In this paper, we tackle the computational issues of stability analysis by proposing an efficient prediction-based validation scheme. Our approach makes use of kernel clustering methods so that we no longer require multiple partitions to be generated in the original high-dimensional space. Furthermore, we propose a novel unsupervised *prototype reduction* strategy that allows us to construct a condensed kernel matrix, leading to substantial efficiency improvements in the subsequent validation procedure without significantly impacting upon its ability to correctly identify $\hat{k}$. Rather than explicitly computing a new set of reduced prototypes in the original feature space, we rely on the "kernel trick" [4] to produce implicit representations of the new objects in the kernel-induced space. Prototype reduction is a delicate process as the reduced dataset must be a good proxy for the full dataset in the validation process. To achieve this, we present a *density-biased* selection strategy that allows us to consistently produce good estimates for the number of clusters in text corpora. On text data, our evaluation shows that the proposed method results in a 16-20 fold speed-up without any loss in acuity as a validation score.

The remainder of this paper is organised as follows. The next section provides a summary of relevant work pertaining to cluster validation and prototype reduction. In Section 3 we discuss our proposed validation scheme, with a particular focus on its application to document clustering. To demonstrate the effectiveness of the scheme, in Section 4 we compare it to existing methods on a large number of real and artificial datasets. Finally, Section 5 presents concluding remarks and suggestions for future work.

2 Related Work

2.1 Cluster Validation

The task of identifying the optimal number of clusters presents a significant challenge when clustering documents. Popular partitional algorithms such as k-means require the *a priori* selection of a value for k. In practice, users will often generate multiple clusterings over a range of k values and select the best partition of the data according to some objective function. Alternatively, when

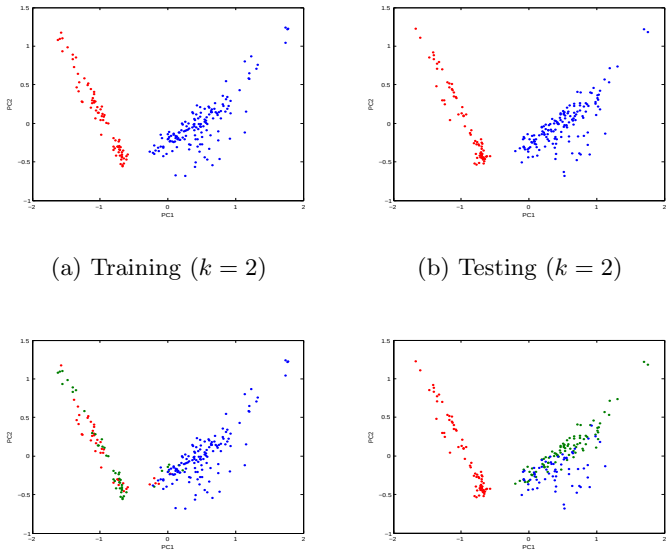
hierarchical clustering algorithms are employed, a termination criterion is often used to identify a suitable point at which agglomeration or sub-division ceases. In either case, some form of *internal* validation criterion is required to evaluate partition quality. In the past, measures such as the *gap statistic* [5] and the *Bayesian information criterion* [6] have been applied in certain domains to select a value for the number of clusters. However, these tend to be model dependent in the sense that they make assumptions about the structure of clusters in data [2]. In addition, many internal criteria are tied to a specific distance function or clustering technique. As a result, their ability to detect arbitrarily-shaped clusters in complex text datasets is generally limited.

2.2 Stability-Based Validation

Validation techniques based on stability analysis have recently been shown to be particularly effective in determining the optimal number of clusters in data [2]. These methods seek to infer $\hat{k}$ based on a clustering model's ability to consistently generate similar partitions on data originating from the same source. If the chosen number of clusters is too large, repeated clusterings will lead to arbitrary partitions of the data, resulting in a low level of stability. On the other hand, if the number of clusters is too small, the clustering algorithm will be forced to produce partitions that merge subsets of objects which should remain separate, also resulting in poor stability. In contrast, repeated clusterings generated using the optimal number of clusters $\hat{k}$ should be robust with respect to perturbations of the data produced by subsampling or the addition of noise, resulting in high stability.

One commonly-used approach to stability analysis involves applying sampling to the original data to generate τ non-disjoint subsets [7, 8]. For each candidate value of k, all subsets are clustered and the agreement between the resulting partitions is computed. To determine the level of agreement, partition similarity measures such as those used in external validation are employed. An evaluation of the stability achieved using k clusters is obtained by taking the mean or median agreement over all pairs of partitions.

Tibshirani *et al.* [3] proposed a novel method for stability analysis which is motivated by the concept of prediction accuracy in supervised learning. To illustrate the relevance of this idea in the context of validation, we consider a small subset of the well-known 20 newsgroups collection (20NG), consisting of 300 documents from the 'cryptography' group and 150 documents from the 'hockey' group. Figures 1(a) and 1(b) show training and testing partitions of this data generated for $k = 2$. Clearly, a suitable classifier built upon the clusters in the former is likely to be successful at predicting the assignment of documents in the latter partition. In contrast, the partitions of the same sets for $k = 3$ shown in figures 1(c) and 1(d) are significantly different, making it unlikely that a classifier constructed on the former will accurately predict the latter. If these patterns are frequently replicated over many different splits of the data, it is reasonable to conclude that $k = 2$ represents a more appropriate choice for the number of clusters than $k = 3$.



(a) Training ($k = 2$) (b) Testing ($k = 2$)

(c) Training ($k = 3$) (d) Testing ($k = 3$)

Fig. 1. Plot of first two Principal Components for partitions generated on subset of 450 documents from 20NG collection.

In practice, each run of the validation process involves applying two-fold cross-validation to randomly split the dataset $\mathcal{X} = \{x_1, \ldots, x_n\}$ into disjoint training and test sets, denoted $\mathcal{X}_a$ and $\mathcal{X}_b$ respectively. Both sets are then clustered to produce partitions $\mathcal{C}_a$ and $\mathcal{C}_b$, using an algorithm such as k-means. Subsequently, a prediction $\mathcal{P}_b$ for the assignment of objects in the test set is produced by assigning each $x_i \in \mathcal{X}_b$ to the nearest centroid in $\mathcal{C}_a$. Prediction accuracy is measured by evaluating the degree to which the class memberships in $\mathcal{P}_b$ correspond to the cluster assignments in $\mathcal{C}_b$. To formally produce an evaluation, the authors in [3] proposed a new measure for comparing partitions, referred to as *prediction strength*. For each cluster in the test clustering $\mathcal{C}_b = \{C_1, \ldots, C_k\}$, we identify the number of pairs of objects assigned to the same cluster that are also assigned to the same class in the predicted partition $\mathcal{P}_b = \{P_1, \ldots, P_k\}$. These associations can be represented as a $\frac{n}{2} \times \frac{n}{2}$ binary matrix $\mathbf{M}$, where $M_{ij} = 1$ only if the objects x_i and x_j are co-assigned in both $\mathcal{C}_b$ and $\mathcal{P}_b$. From this matrix, an evaluation is computed based on the cluster containing the minimum fraction of correctly predicted pairs:

$$S(\mathcal{C}_b, \mathcal{P}_b) = \min_{1 \leq h \leq k} \left[\frac{1}{|C_h|\,(|C_h| - 1)} \sum_{x_i \neq x_j \in C_h} M_{ij} \right] \tag{1}$$

This prediction process is repeated for τ runs for each candidate value k in a fixed range $[k_{min}, k_{max}]$. A final estimate for $\hat{k}$ is made using a heuristic approach that

involves identifying the largest k such that the corresponding mean prediction strength is above a user-defined threshold.

2.3 Supervised Prototype Reduction

Prototype reduction techniques have been extensively used in supervised learning for tasks involving large datasets, typically in conjunction with a nearest-neighbour classifier. These techniques are concerned with producing a minimal set of objects or prototypes to represent a dataset, while ensuring that a classifier applied to this set will perform approximately as well as on the original dataset. In the literature, these techniques are generally divided into two categories: *prototype selection* is the process of identifying a subset of representative objects from the original data, while *prototype extraction* involves the creation of an entirely new set of objects. A comprehensive overview of supervised reduction schemes has been provided by Bezdek and Kuncheva [9].

Many reduction techniques are computationally intensive, often involving clustering-like procedures to identify relevant prototypes. In contrast, Hamamoto *et al.* [10] proposed a simple, fast, stochastic technique (BTS), based on bootstrap editing. Initially, a random sample of n' seed objects is drawn from the dataset. Each seed object is then replaced by a new prototype constructed from the mean of its p-nearest neighbours and the seed itself. A 1-NN classifier is then applied to the new set of n' prototypes. The entire process may be repeated multiple times to give improved results. In [11] a novel framework was proposed which involves using a chosen reduction scheme, such as BTS, to produce a smaller set of prototypes, on which a reduced-kernel matrix is constructed. Ensemble classifier methods are then employed on this kernel to compensate for any loss in accuracy resulting from the reduction in dataset size.

2.4 Unsupervised Prototype Reduction

While most work in prototype reduction has focused on supervised learning tasks, the concept has been used implicitly as part of many clustering algorithms. It has particular relevance for clustering applications in limited resource scenarios, such as interactive information retrieval. Cutting *et al.* proposed a technique, referred to as *fractionation*, to improve the efficiency of hierarchical clustering methods for large text corpora, which can be viewed as a form of prototype reduction. The procedure involves randomly splitting the corpus into fractions. The documents in each fraction are then clustered separately so that, by treating the generated clusters as if they were individual "meta-documents", the number of data objects is reduced. The algorithm is repeated using these meta-documents, with the process terminating when only k clusters remain. It should be noted that the use of prototype selection in clustering is closely related to both the problem of outlier removal [12] and the choice of seeds in cluster initialisation [13].

3 Proposed Method

For small datasets, stability-based validation techniques offer an attractive option for inferring a value for the optimal number of clusters. However, for larger, high-dimensional data such as text corpora, the cost of generating and evaluating multiple clustering solutions is often prohibitive. When using the standard vector space model, each document in a corpus will be represented by an m-dimensional feature vector. As the dimensionality m increases, the cost of repeatedly running a clustering algorithm such as k-means will greatly increase. The number of documents n will also be a limiting factor, as an increase in n will greatly affect the running time of the clustering and the stability assessment procedures, which will typically run in $O(n^2)$ time or slower. To tackle the computational issues of stability analysis, we now introduce an efficient prediction-based validation method suitable for use in document clustering tasks.

3.1 Kernel-Based Stability Analysis

To avoid having to work in the m-dimensional vector space, we make use of recently proposed kernel clustering methods. A kernel function is usually represented by an $n \times n$ kernel matrix $\mathbf{K}$, where K_{ij} indicates the affinity between objects x_i and x_j. The advantage of using kernel methods in the context of stability analysis stems from the fact that, having constructed a single kernel matrix, we may subsequently generate multiple partitions of the data without referring back to the original high-dimensional feature space. A variety of popular clustering techniques have been re-formulated for use in a kernel-induced space. As the standard k-means algorithm has commonly been used in both stability analysis and document clustering, we focus here on the use of the corresponding kernelised k-means algorithm [14].

To form the basis for our validation scheme, we choose the prediction-based method proposed in [3] due to its sound theoretical foundation, computational advantage over other stability-based methods and empirical success. We construct disjoint training and test sets by employing two-fold cross-validation. Experimental observations support the authors' assertion that using a larger number of folds is unlikely to be beneficial. Consequently, each run of the kernel k-means algorithm involves only a sample of $\frac{n}{2}$ objects. We apply this process τ times for each candidate k in a fixed range, assessing the stability achieved by the clustering model at each run.

The stability assessment phase involves measuring the level of agreement between a clustering of the test set and the assignments as predicted by a given classifier. To perform this comparison, some authors have suggested the use of set matching measures, including normalised hamming distance [2] and partition similarity [15]. However, we make use of an adjusted version of the *prediction strength* measure (1) because of its strong theoretical foundation and superior empirical performance. Rather than using a heuristic method to choose among the candidate values of k, we select the value k that leads to the maximum average score over τ runs. Since Eqn. 1 exhibits a natural bias toward smaller

values of k, we employ the widely-used adjustment technique described in [16] to correct for chance agreement:

$$S'(\mathcal{C}_b, \mathcal{P}_b) = \frac{S(\mathcal{C}_b, \mathcal{P}_b) - \bar{S}(\mathcal{C}_b, \mathcal{P}_b)}{1.0 - \bar{S}(\mathcal{C}_b, \mathcal{P}_b)} \tag{2}$$

where $\bar{S}(\mathcal{C}_b, \mathcal{P}_b)$ is the expected prediction strength on the split $(\mathcal{X}_a, \mathcal{X}_b)$ for a given k, which may be approximated by calculating the mean value of Eqn. 1 over a large number of pairs of random partitions.

As noted in [2], the choice of classifier used to make predictions should complement the clustering algorithm. For k-means, a nearest centroid classifier is appropriate, where a prediction is made by associating each object in the test set $\mathcal{X}_b$ with the closest training centroid. To "mimic" the assignment behaviour of the kernel k-means algorithm, we employ a kernel nearest centroid classifier, such that each object in $\mathcal{X}_b$ is classified as being a member of the class represented by the nearest pseudo-centroid in the training clustering. Subsequently, we use Eqn. 2 to evaluate the degree to which the predicted classification agrees with the clustering of $\mathcal{X}_b$ as produced by kernel k-means.

3.2 Kernel Reduction

In Section 3.1 we described a method for stability-based validation that is suitable for application to high-dimensional data. However, the validation process still requires τ runs consisting of clustering and prediction assessment phases, which both run in $O((\frac{n}{2})^2)$ time. Additionally, for a given value of n, to produce robust results we will need a sufficiently large number of runs to compensate for the variance introduced by subsampling. Clearly, decreasing n will make the validation process significantly less computationally expensive. Motivated by existing techniques such as fractionation [17], it is apparent that a natural approach for accomplishing this is to create a reduced set of $n' < n$ objects, upon which the validation procedure may be subsequently applied. However, any such reduction must be performed in a way that preserves the structure of the underlying "natural classes" in the data so as not to overly impact upon our ability to robustly estimate $\hat{k}$. Specifically, we wish to ensure that the expected number of prototypes representing each class is proportional to the size of that class. In addition, we wish to sample uniformly from within that class, so that we cover both core and outlying regions.

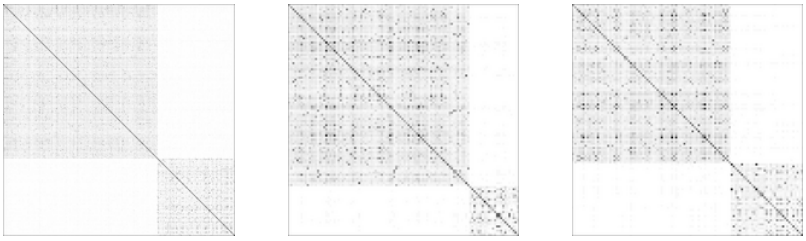
Meeting these requirements without any form of supervision is not a trivial task. In [9] it was noted that "pre-supervised" reduction approaches, which use both the raw data and class information to produce new prototypes, tend to be far more successful than their purely unsupervised counterparts. Since the former generally involve processing each class separately, the resulting reduced prototypes will be "meaningful" in the sense that they will represent regions from a single class only. In document clustering we will generally not have access to any form of external knowledge, so we must rely instead upon intrinsic properties of the data to ensure that all cluster structures are adequately represented.

Unfortunately, text corpora often contain unbalanced cluster sizes, which may also differ in their relative densities, making the task particularly problematic.

To address these issues, we propose a reduction scheme consisting of two phases. In the first phase, prototype extraction is used to generate a set of candidate prototypes representing small homogeneous regions of the data. The second phase selects from among these a subset of n' prototypes to form a reduced kernel matrix $\mathbf{K}'$. This selection phase involves the application of a deterministic density-biased strategy to select from the set of possible prototypes.

To create the extracted prototypes, we employ a bootstrap method similar to the supervised BTS reduction scheme [10], where new prototypes are formed by locally combining subsets of objects from the original data. Firstly, we define a *neighbourhood* $\mathcal{N}_a$ as a subset of $\mathcal{X}$ consisting of a seed object x_a together with its set of p nearest neighbours. A new prototype s_a may be constructed from the mean of these $p+1$ objects. Since we wish to work in the kernel-induced space only, we consider s_a to be the pseudo-centroid of the subset $\mathcal{N}_a$ as calculated from the values in $\mathbf{K}$. Motivated by the need to construct meaningful prototypes, we observe that, by the well-known statistical concept of *neighbourhood consistency*, if objects assigned to the same natural class are highly similar, then the nearest neighbour of any given object in a class are also likely to belong that class. This principle has recently been used in both clustering [18] and internal validation [19]. We assert that, as regions forming cluster structures will be locally homogeneous, the majority of the set of neighbours of each object should belong to the same natural class as that object. Therefore, prototypes constructed from the centroid of sufficiently small neighbourhoods will generally be representative of a single natural class.

However, the problem remains of selecting a subset $\mathcal{S}'$ of n' optimal prototypes from the set of n candidates, denoted by $\mathcal{S} = \{s_1, \ldots, s_n\}$. One possible solution is to apply unbiased random sampling to choose $\mathcal{S}'$ in the same manner as employed in BTS, where each seed object x_i (and corresponding reduced prototype s_i) has an equal probability of being selected. However, this approach has several notable drawbacks in the context of validation. As stated previously, we wish to select a fraction of prototypes from each class that is proportional to the size of that class in the original dataset. A single random subsampling of $\mathcal{S}$ is not guaranteed to achieve this. As an example, we consider the case of the 20NG subset described in Section 2.2. Figures 2(a) and 2(b) respectively show the block-ordered matrices corresponding to the full kernel matrix and a reduced matrix produced by randomly selecting seeds. From the latter, it is evident that the smaller 'hockey' class is not adequately represented after the random reduction process. We observed in our evaluation that subsets of reduced prototypes randomly chosen in this way frequently fail to produce a true proxy for the dataset, resulting in poor estimations for $\hat{k}$ in the subsequent validation process. In these cases, the failure is often due to the neglection of smaller clusters or important sub-regions within clusters. While we could run the process multiple times and aggregate the results, in practice the resulting computational cost would typically negate the benefits of performing prototype reduction.



(a) (b) (c)

Fig. 2. Gram matrix for (a) full kernel; (b) kernel reduced by random sampling; (c) kernel reduced by density selection.

As an alternative, the second phase of our reduction procedure employs a deterministic density-biased strategy to select the subset $\mathcal{S}'$. This procedure has similar goals to existing density-biased sampling techniques (*e.g.* [20]), but is stochastic and does not require that we partition the original high-dimensional feature space. Firstly, we define the *compactness* of a neighbourhood $\mathcal{N}_a$ to be the average of the pair-wise affinities between its constituent members:

$$C(\mathcal{N}_a) = \frac{\sum_{x_i, x_j \in \mathcal{N}_a} K_{ij}}{|\mathcal{N}_a|^2} \tag{3}$$

where $|\mathcal{N}_a| = p + 1$. This is equivalent to the "self-similarity" of the pseudo-centroid formed from $\mathcal{N}_a$. As stated previously, we wish to select a fraction of prototypes from each class that is approximately proportional to the size of that class in the original dataset. To achieve this, the prototypes in $\mathcal{S}$ are ranked in descending order according to their compactness. We now uniformly choose $n' = \frac{n}{\rho}$ prototypes from the ranked list, where ρ is the *reduction rate* that determines the degree to which the number of objects should be reduced. Specifically, we select every ρ-th prototype from the ranked list, thereby ensuring that we represent all density patterns in the data. The n' selected prototypes are then used to build the reduced kernel matrix $\mathbf{K}'$. Rather than computing explicit representations of the new prototypes in the original feature space, we can make use of the affinity values in the original kernel matrix to directly construct $\mathbf{K}'$. Formally, the affinity between a pair of reduced prototypes s_i and s_j is calculated as:

$$K'_{ij} = \frac{\sum_{x_a \in S_i, x_b \in S_j} K_{ab}}{(p+1)^2} \tag{4}$$

Referring back to our previous example, we can see that, unlike in the case of random sampling, the reduced kernel matrix in Figure 2(c) is clearly representative of the two classes in the original dataset, despite their differing sizes and densities. In practice, we consistently observe that this density-biased selection strategy produces a set of extracted prototypes that accurately summarise the underlying structures in the data. We contend that this success is due to the in-

clusion of regions of all densities in the data, ensuring good coverage of clusters of varying densities and all sub-regions within those clusters.

Once we have constructed the reduced kernel matrix, the prediction phase of the validation procedure proceeds as described in Section 3.1. While it is possible that a matrix defined by Eqn. 4 will not represent a valid Mercer kernel in the sense that it may not be positive semi-definite, it has previously been shown in [21] that this does not pose a significant problem for the kernel k-means algorithm.

The application of the proposed reduction procedure results in a significant decrease in the computational cost of the validation process. Our approach does require a once-off initialisation phase, with complexity $O(n \log n)$ for the prototype extraction phase and $O(n'^2 p^2)$ for the construction of $\mathbf{K}'$. However, the computational gains in the subsequent validation process are substantial. For each of the τ runs of the validation process, the costs associated with clustering and prediction assessment both become $O((\frac{n}{2\rho})^2)$. In practice, particular benefits may be derived from the greatly reduced clustering time, which represents a bottleneck in traditional stability-based validation procedures.

3.3 Application to Document Clustering

While our proposed method may be used in conjunction with any valid kernel function, for the purpose of document clustering we make use of a linear kernel that has been normalised according to the approach described by Schölkopf and Smola in [4], yielding values in the range $[0, 1]$. The matrix of this normalised kernel is equivalent to that produced by the widely used cosine similarity measure, so that the affinity between a pair of documents x_i and x_j is given by

$$K_{ij} = \frac{\langle x_i, x_j \rangle}{\sqrt{\langle x_i, x_i \rangle \langle x_j, x_j \rangle}} \tag{5}$$

Typically, the cost of producing this matrix can be significantly reduced by using a sparse matrix representation and computing pair-wise similarities across non-zero features values only.

While a kernel defined by Eqn. 5 represents an intuitive choice for document clustering, its matrix will typically suffer from the problem of *diagonal dominance*. This phenomenon occurs when, for a given kernel function, self-similarity values are large relative to between-object similarities. It has been shown in [21] that this can negatively impact upon the accuracy and stability of centroid-based kernel clustering algorithms. To reduce the dominance effect, we apply a negative shift to the diagonal of the kernel matrix so as to minimise its trace, as described in [21]. In Section 4, we see that this frequently has the effect of increasing clustering accuracy during the validation process, leading to a noticeable improvement in validation performance. A summary of the complete validation process is provided in Figure 3.

As mentioned previously, our proposed method is based on the assumption that regions will be locally homogeneous, which should generally be the case

Initialisation Phase

- Construct full $n \times n$ kernel matrix $\mathbf{K}$ from the original dataset using Eqn. 5.
- Extract set of n candidate prototypes $\mathcal{S}$, consisting of neighbourhood centroid vectors.
- Evaluate the compactness of each candidate and rank them in descending order.
- Select the set of reduced prototypes $\mathcal{S}'$, such that $|\mathcal{S}'| = n/\rho = n'$, based on the compactness ranking.
- Construct the $n' \times n'$ reduced kernel matrix $\mathbf{K}'$ from $\mathbf{K}$ using prototypes in $\mathcal{S}$.
- Apply zero-trace diagonal shift to $\mathbf{K}'$.

Validation Phase

- Produce τ splits of $\mathcal{S}'$ into training and test sets.
- For each value of $k \in [k_{min}, k_{max}]$:
  1. For each split $(\mathcal{X}_a, \mathcal{X}_b)$:
     (a) Apply kernel k-means to training set $\mathcal{X}_a$ using kernel $\mathbf{K}'$.
     (b) Apply kernel nearest centroid classifier to predict the assignment of documents in $\mathcal{X}_b$.
     (c) Apply kernel k-means to test set $\mathcal{X}_b$ using kernel $\mathbf{K}'$.
     (d) Evaluate prediction strength and correct for chance as in Eqn. 2.
  2. Compute mean corrected prediction strength for k.
- Select $\hat{k}$ to be the candidate k with the highest mean prediction strength.

Fig. 3. Complete kernel prediction-based validation scheme, with prototype reduction.

when an appropriate kernel function is chosen. To maximise neighbourhood homogeneity, we select a low value for the number of neighbours, with $p = 5$ being used for our experiments in the following section. We note that the use of a small value for p also has the effect of reducing the time required to construct the reduced kernel matrix $\mathbf{K}'$.

Bezdek and Kuncheva [9] concluded that the goals of supervised prototype reduction tasks, the identification of a minimal number of prototypes while maximising accuracy, will naturally conflict. Similarly, it is unsurprising that the extent to which we reduce our kernel matrix will impact upon our ability to correctly identify $\hat{k}$. In our experimental evaluations, we have observed that a value of $\rho = 4$ for the reduction rate substantially reduces the time required for the validation process, without significantly affecting its accuracy. The selection of ρ will also influence the maximum number of runs τ, where the computational gains resulting from prototype reduction allows us to select a larger value (*e.g.* $\tau > 100$) to guarantee the robustness of the overall validation procedure.

It must be stressed that, in our experiments on text data, the use of these "general purpose" parameter values proved to be effective on a diverse range of datasets, indicating that the proposed validation method is quite robust to the choice of values for these parameters. This allows us to focus on the more immediate task of selecting the number of clusters.

4 Empirical Evaluation

In this section we compare the newly proposed validation scheme with prediction-based techniques operating on the full data. Specifically, we consider four validation methods. The first involves applying k-means in conjunction with the prediction strength criterion (KM-S). Assessments are performed using a version of Eqn. 1 corrected for chance agreement, so that we do not require a final value for k to be manually selected by inspecting the plot of results. The second method also uses k-means, with assessments made using the *partition similarity* criterion described in [15] (KM-P). Note that this technique is also essentially equivalent to that described in [2]. The final two methods are those proposed in this paper: kernel k-means with prediction strength (KKM-S), and kernel k-means with prediction strength after prototype reduction (RED-S). Both kernel-based techniques employ the diagonal shift technique prior to validation to address the issue of diagonal dominance. For comparison, when applying k-means, we make use of the standard cosine similarity measure, which is equivalent to the normalised kernel defined in Eqn. 5. All clustering algorithms are initialised by randomly assigning documents to clusters.

The experimental process involved applying the schemes to each dataset across a reasonable range of values for k (for the data in this paper, we chose $[2, 10]$) and comparing their output with the "true" number of natural classes. It should be noted that, while we make use of this external information in the evaluation, it is possible that the validation schemes may identify other values of k as being potentially valid, serving to highlight alternative interesting groupings in the data. In all cases, we used $\tau = 200$ to minimise any variance introduced by subsampling.

4.1 Evaluation on Artificial Data

For our initial experimental evaluation, we required a large number of datasets to illustrate significant differences between the validation strategies. While many authors examining stability-based validation techniques have made use of synthetic datasets, generating data that realistically models the distribution of term frequency values in text data is difficult. As an alternative, we used the 20NG collection as a source for artificially constructed datasets because it contains a range of topics that overlap to varying degrees. From the collection we derived a large number of smaller datasets for which the correct value of $\hat{k}$ is known. We constructed 84 sets in total, 12 for each value of $\hat{k} \in [2, 10]$. Half of these datasets consist of newsgroups that are reasonably compact and well-separated (*e.g.* 'graphics', 'hockey', 'mideast'). The remaining sets consist of newsgroups that overlap considerablely (*e.g.* 'mac', 'windows'). These two groups are further divided into sub-groups of datasets containing clusters of different proportions, in a manner similar to that suggested for producing artificial data in [15]: balanced clusters containing 500 documents each, unbalanced clusters where one cluster contains 10% of the documents in the dataset, and unbalanced clusters

where one cluster contains 60% of the documents. In all cases the documents were randomly drawn from each class[1].

Table 1. Percentage of correct and top-3 estimations for $\hat{k}$ on artificial data.

Datasets	#	KM-S		KM-P		KKM-S		RED-S	
		First	Top 3	First	Top 3	First	Top 3	First	Top 3
Balanced	28	54%	68%	61%	89%	71%	86%	79%	89%
Unbalanced	56	21%	61%	25%	70%	30%	71%	36%	66%
Non-overlapping	42	45%	76%	43%	81%	62%	90%	67%	88%
Overlapping	42	19%	50%	31%	71%	26%	62%	33%	60%
Overall	84	32%	63%	37%	76%	44%	76%	50%	74%

Table 1 summarises the relative performance of the four methods under consideration in terms of the the percentage of datasets on which each method was successful in identifying $\hat{k}$. These results indicate that both kernel-based techniques consistently outperformed those employing the standard k-means algorithm. In these cases, the application of the diagonal shift frequently lead to significantly higher prediction accuracy. Furthermore, we see that, across the 84 artificial datasets, the reduced validation process (RED-S) generally lead to more instances where the true number of clusters was correctly identified. This is particularly apparent for datasets with non-overlapping clusters. The difference was less pronounced on datasets with overlapping clusters, where object neighbourhoods were generally less homogeneous. When performing the evaluation on such a large number of datasets, we observed that the speed-up achieved by working on $\frac{n}{4}$ reduced prototypes was dramatic.

4.2 Evaluation on Real Data

In our second evaluation, we compare the four validation schemes on real-world corpora that have previously been used in document clustering. The *classic3* and *classic* datasets are collections of technical abstracts taken from Cornell's SMART repository[2], which have been widely used in information retrieval. The *bbc* corpus contains news articles from the BBC corresponding to stories in five topical areas: business, entertainment, politics, sport and technology. The *bbc-sport* corpus consists of a smaller set of sports news articles from the same source[3]. The *ng17-19* dataset is a commonly used subset of the 20NG collection, consisting of three groups relating to politics that exhibit considerable overlap. The *ng3* dataset is another subset derived from the same collection, composed of three relatively well-separated groups pertaining to astronomy, politics and

[1] See http://www.cs.tcd.ie/Derek.Greene/research/datasets.html for full description of artificial datasets

[2] Available from ftp://ftp.cs.cornell.edu/pub/smart

[3] Both available from http://www.cs.tcd.ie/Derek.Greene/research/datasets.html

computer graphics. The *reviews* dataset contains articles from the TREC collections relating to food, movies, music, radio and restaurants[4]. Further details for these datasets are given in Table 2.

Table 2. Details of real datasets.

Dataset	Description	Documents	Terms	$\hat{k}$
bbc	News articles from BBC	2225	9635	5
bbcsport	Sports news articles from BBC	737	4613	5
classic3	CISI/CRAN/MED	3893	6733	3
classic	CACM/CISI/CRAN/MED	7097	8276	4
cstr	Technical abstracts	505	2117	4
ng17-19	Overlapping newsgroups	2625	11841	3
ng3	Approximately disjoint newsgroups	2900	12875	3
reviews	Entertainment news articles (TREC)	4069	18391	5

Table 3 shows the results of the comparison between the four schemes, indicating the top three estimated values for $\hat{k}$ on the eight real datasets, together with the corresponding criterion scores. In almost all cases, the reduced clustering method (RED-S) recommended the same value of k as that chosen when validation was performed on the full kernel matrix (KKM-S). Only in the case of the *reviews* dataset, which contains significantly overlapping clusters, did it fail to rate $\hat{k}$ among its top three choices. However, the methods based on k-means also performed poorly on this corpus. In the case of the *cstr* dataset, all four techniques demonstrated a preference for three clusters, which is explained by the considerable overlap between the 'ai' and 'vision' classes. Specifically, the mean affinity between documents across the two classes is greater than the mean affinity between documents in the 'vision' class, suggesting that $k = 3$ does in fact represent a suitable choice. It is interesting to note that, as with the artificial data, the kernel-based methods generally outperformed those relying on the standard k-means algorithm. Once again, we observed that employing a diagonal dominance reduction technique prior to validation results in higher prediction scores and better estimates of $\hat{k}$ when using kernel k-means. In general, we observed that using prototype reduction with $\rho = 4$ consistently afforded a 16-20 fold decrease in the time required for the validation process.

5 Conclusion

We have proposed a practical approach to stability-based validation suitable for the task of estimating the number of clusters in large, high-dimensional datasets such as text corpora. The use of kernel clustering methods allows us to work on a single kernel matrix rather than repeatedly computing distances in the original feature space. Moreover, we have demonstrated that we can significantly decrease

[4] Available from http://www.cs.umn.edu/∼karypis/cluto

Table 3. Summary of results on real datasets.

Dataset	Method	1st	2nd	3rd
bbc ($\hat{k} = 5$)	KM-S	**5** (0.51)	4 (0.38)	6 (0.36)
	KM-P	**5** (0.75)	6 (0.71)	7 (0.66)
	KKM-S	**5** (0.62)	6 (0.46)	4 (0.42)
	RED-S	**5** (0.51)	6 (0.43)	4 (0.42)
bbcsport ($\hat{k} = 5$)	KM-S	4 (0.29)	**5** (0.27)	3 (0.27)
	KM-P	**5** (0.55)	6 (0.53)	4 (0.50)
	KKM-S	**5** (0.45)	6 (0.43)	4 (0.41)
	RED-S	**5** (0.47)	4 (0.43)	6 (0.41)
classic3 ($\hat{k} = 3$)	KM-S	**3** (0.85)	2 (0.56)	4 (0.44)
	KM-P	**3** (0.90)	2 (0.73)	4 (0.72)
	KKM-S	**3** (0.94)	4 (0.49)	5 (0.45)
	RED-S	**3** (0.95)	4 (0.48)	5 (0.44)
classic ($\hat{k} = 4$)	KM-S	3 (0.58)	5 (0.57)	2 (0.50)
	KM-P	3 (0.83)	5 (0.78)	2 (0.76)
	KKM-S	5 (0.78)	**4** (0.74)	2 (0.63)
	RED-S	5 (0.65)	**4** (0.57)	2 (0.52)
cstr ($\hat{k} = 4$)	KM-S	3 (0.58)	2 (0.29)	**4** (0.28)
	KM-P	3 (0.78)	**4** (0.57)	2 (0.53)
	KKM-S	3 (0.82)	**4** (0.50)	5 (0.38)
	RED-S	3 (0.75)	**4** (0.43)	5 (0.36)
ng3 ($\hat{k} = 3$)	KM-S	**3** (0.61)	4 (0.45)	2 (0.40)
	KM-P	**3** (0.86)	4 (0.77)	5 (0.67)
	KKM-S	**3** (0.69)	4 (0.46)	2 (0.38)
	RED-S	**3** (0.63)	2 (0.56)	4 (0.50)
ng17-19 ($\hat{k} = 3$)	KM-S	5 (0.37)	4 (0.33)	6 (0.29)
	KM-P	5 (0.62)	4 (0.56)	6 (0.56)
	KKM-S	5 (0.40)	4 (0.39)	**3** (0.34)
	RED-S	4 (0.33)	5 (0.33)	**3** (0.32)
reviews ($\hat{k} = 5$)	KM-S	2 (0.87)	3 (0.42)	6 (0.40)
	KM-P	2 (0.97)	8 (0.71)	9 (0.70)
	KKM-S	2 (0.91)	**5** (0.53)	4 (0.51)
	RED-S	2 (0.95)	6 (0.42)	3 (0.42)

the computational demands of the validation process by employing a form of prototype reduction to construct a reduced kernel matrix. To ensure that the use of a condensed representation does not adversely impact upon the accuracy of the validation process, we have proposed a density-biased strategy for selecting a set of reduced prototypes that adequately represent the underlying classes in the data, regardless of their relative sizes or densities. Notably, the reduction process does not require that we explicitly represent these new prototypes as feature vectors. Extensive experimental evaluations have shown this validation process to be effective on a large number of real and artificial datasets, where it consistently produced good estimates for the optimal number of clusters, often outperforming existing methods that are significantly more computationally expensive.

While we have particularly focused on validation in the area of document clustering, we believe that our approach is applicable for a wide variety of other domains and kernel functions, where large datasets would otherwise make stability analysis unfeasible. We also expect that, while the new prototype reduction technique has been used in conjunction with prediction-based validation, the underlying principles may also be useful in improving the efficiency of other computationally costly learning methods, such as ensemble clustering.

References

1. Jain, A.K., Dubes, R.C.: Algorithms for clustering data. Prentice-Hall, Inc. (1988)
2. Lange, T., Roth, V., Braun, M.L., Buhmann, J.M.: Stability-based validation of clustering solutions. Neural Comput. **16** (2004) 1299–1323
3. Tibshirani, R., Walther, G., Botstein, D., Brown, P.: Cluster validation by prediction strength. Technical report, Statistics Department, Stanford University (2001)
4. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA (2001)
5. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a dataset via the gap statistic. Technical Report 208, Dept. of Statistics, Stanford University (2000)
6. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics **6** (1978) 461–464
7. Levine, E., Domany, E.: Resampling method for unsupervised estimation of cluster validity. Neural Computation **13** (2001) 2573–2593
8. Ben-Hur, A., Elisseeff, A., Guyon, I.: A stability based method for discovering structure in clustered data. In: Proceedings of the 7th Pacific Symposium on Biocomputing (PSB 2002), Lihue, Hawaii (2002) 6–17
9. Bezdek, J.C., Kuncheva, L.: Nearest prototype classifier designs: An experimental study. International Journal of Intelligent Systems **16** (2001) 1445–1473
10. Hamamoto, Y., Uchimura, S., Tomita, S.: A bootstrap technique for nearest neighbor classifier design. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997) 73–79
11. Kim, S.W., Oommen, B.J.: On using prototype reduction schemes and classifier fusion strategies to optimize kernel-based nonlinear subspace methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 455–460

12. Kollios, G., Gunopulos, D., Koudas, N., Berchtold, S.: Efficient biased sampling for approximate clustering and outlier detection in large datasets. IEEE Transactions on Knowledge and Data Engineering **15** (2003)
13. Juan, A., Vidal, E.: Comparison of four initialization techniques for the k-medians clustering algorithm. In: Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition, Springer-Verlag (2000) 842–852
14. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. **10** (1998) 1299–1319
15. Giurcaneanu, C., Tabus, I.: Cluster structure inference based on clustering stability with applications to microarray data analysis. EURASIP Journal on Applied Signal Processing **1** (2004) 64–80
16. Hubert, L.J., Arabie, P.: Comparing partitions. Journal of Classification **2** (1985) 193–218
17. Cutting, D.R., Pedersen, J.O., Karger, D., Tukey, J.W.: Scatter/gather: A cluster-based approach to browsing large document collections. In: Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (1992) 318–329
18. Ding, C., He, X.: K-nearest-neighbor consistency in data clustering: incorporating local information into global optimization. In: SAC '04: Proceedings of the 2004 ACM symposium on Applied computing, ACM Press (2004) 584–589
19. Frederix, G., Pauwels, E.J.: Shape-invariant cluster validity indices. Lecture Notes in Computer Science **3275** (2004)
20. Palmer, C.R., Faloutsos, C.: Density biased sampling: an improved method for data mining and clustering. In: ACM SIGMOD International Conference on Management of Data. (2000) 82–92
21. Greene, D., Cunningham, P.: Practical solutions to the problem of diagonal dominance in kernel document clustering,. In: Proceedings of the 23rd International Conference on Machine Learning. (2006)