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 regarding the structure of clusters. On the other hand, *external* validation techniques, which assess the degree to which a clustering corresponds to the "natural classes" in the data, are not directly applicable for parameter selection since 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 $\hat{k}$

was proposed in [3], which is motivated by the concept of prediction accuracy in supervised learning. This *prediction-based* validation scheme involves assessing, for a given number of clusters 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 need to generate clusterings 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 an implicit representation of the new objects in the kernel-induced space. To ensure that this reduced representation is a good proxy for the full dataset, we present a *density-biased* prototype 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 scheme 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. Note that an extended version of this paper is available as a technical report with the same title [5][1].

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 values for k and select the best partition of the data according to some objective function. Alternatively, when

[1] TODO: URL here?

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* [6] and the *Bayesian information criterion* [7] 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. In practice, a clustering algorithm employing $\hat{k}$ should be robust with respect to perturbations of the data produced by subsampling, resulting in a high level of stability across many clusterings.

Tibshirani *et al.* [3] proposed a novel method for stability analysis which is motivated by the concept of prediction accuracy in supervised learning. 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:

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

This prediction process is repeated for τ runs for each candidate value k in a reasonable range $[k_{min}, k_{max}]$. A final estimate for $\hat{k}$ is made by identifying the largest k such that the corresponding mean prediction strength is above a user-defined threshold.

2.3 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 the data, 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* techniques seek to identify a subset of representative objects from the original data, while *prototype extraction* techniques involve the creation of an entirely new set of objects. A comprehensive overview of supervised reduction schemes has been provided by Bezdek and Kuncheva [8].

Many reduction techniques are computationally intensive, often involving clustering-like procedures to identify relevant prototypes. In contrast, Hamamoto *et al.* [9] 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 [10] 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 smaller 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.

While most work in prototype reduction has focused on supervised learning tasks, the concept has been used implicitly as part of many clustering algorithms. Notably, 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 each cluster as a single "meta-document", the number of data objects is subsequently reduced. It is interesting to note that the application of prototype selection in clustering is closely related to both the problem of outlier removal [11] and the choice of seeds in cluster initialisation [12].

3 Proposed Method

For small datasets, stability-based validation techniques offer an attractive option for inferring a value for $\hat{k}$. However, for larger, high-dimensional data, their use is often unfeasible. As the number of dimensions increases, the time required to repeatedly apply an algorithm such as k-means will greatly increase. The number of objects n will also be a limiting factor, as a larger value for n will substantially increase the computational cost of the clustering and the stability assessment procedures, which typically run in $O(n^2)$ time or slower. To tackle these issues, we now introduce an efficient prediction-based validation method suitable for use on text corpora.

3.1 Kernel-Based Stability Analysis

To avoid having to work in the original feature space 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 without referring back to the original data. 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.

To form the basis for our validation scheme, we choose the prediction-based method proposed in [3] due to its empirical success and computational advantage over other stability-based methods. The latter derives from the fact that, since we employ two-fold cross validation to produce disjoint training and test sets, each run of the kernel k-means algorithm involves only a sample of $\frac{n}{2}$ objects. To assess prediction accuracy, some authors have suggested the use of set matching measures, such as *partition similarity* [13]. 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 [14] to correct for chance agreement:

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

Note that $\bar{S}(\mathcal{C}_b, \mathcal{P}_b)$ 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 discussed in [2], the choice of classifier used to make predictions should complement the clustering algorithm. 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.

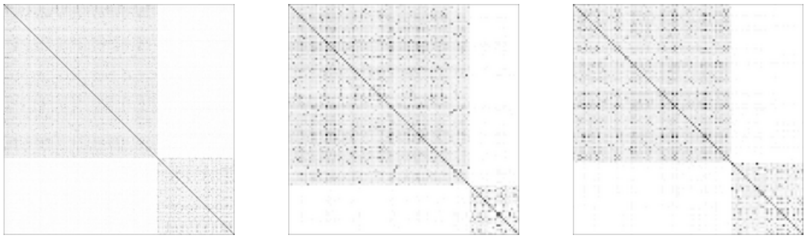
3.2 Kernel Reduction

In the previous section we described a method for stability-based validation that is suitable for use on 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. Clearly, decreasing n will make the validation process significantly less computationally expensive. Motivated by existing techniques such as fractionation [15], it is apparent that an intuitive solution 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 true classes in the data. Specifically, we wish to ensure that the expected number of prototypes representing each class is approximately proportional to the size of that class. In addition, we wish to cover both core and outlying regions within these classes.

Meeting these requirements without any form of supervision is not a trivial task. In [8] it was noted that reduction approaches utilising class information 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 the absence of class labels we must rely 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 formed from small homogeneous regions of the data. The second phase selects from among these a subset of n' prototypes to build a reduced kernel matrix $\mathbf{K}'$.

Firstly, we create a set of extracted prototypes $\mathcal{S} = \{s_1, \ldots, s_n\}$ in a manner similar to that employed by the supervised BTS reduction scheme [9], where new prototypes are formed by locally combining subsets of the original dataset $\mathcal{X}$. Formally, we define a *neighbourhood* $\mathcal{N}_i$ as a subset of $\mathcal{X}$ consisting of a seed object x_i together with its set of p nearest neighbours. A new prototype s_i 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_i to be the pseudo-centroid of the subset $\mathcal{N}_i$ as calculated from the values in $\mathbf{K}$. Motivated by the need to construct meaningful prototypes, it is apparent that, as regions forming cluster structures will normally be locally homogeneous, the majority of the set of neighbours of each object are likely belong to the same cluster as that object [16]. 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 n possible candidates. A possible solution is to apply unbiased random sampling to choose $\mathcal{S}'$. However, this approach has several 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 sample from $\mathcal{S}$ is not guaranteed to achieve this. As an example, we consider the case of the 20NG subset described in Section 2.2. Figures 1(a) and 1(b) respectively show the block-ordered matrices corre-



(a) (b) (c)

Fig. 1. Gram matrix for (a) full kernel; (b) kernel reduced by random sampling; (c) kernel reduced by density selection.

sponding 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 evaluations that subsets of reduced prototypes 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, the resulting computational cost would typically negate the benefits of performing prototype reduction.

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.* [17]), but is stochastic and does not require that we partition the original high-dimensional feature space. Firstly, we define the *compactness* of a neighbourhood $\mathcal{N}_a$ as 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 $\mathcal{N}_a$. In the selection process, the prototypes in $\mathcal{S}$ are ranked in descending order according to their compactness. We now uniformly choose $n' = \frac{n}{\rho}$ prototypes, 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 ordered list, thereby ensuring that we represent all density patterns in the data. We then build the reduced kernel matrix $\mathbf{K}'$ based on these n' prototypes. 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 defined as:

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

While it is possible that a matrix constructed in this way may not always be positive semi-definite, it has previously been shown in [18] that this does not pose a significant problem for the kernel k-means algorithm.

Referring back to our previous example, we can see that, unlike in the case of random sampling, the reduced kernel matrix in Figure 1(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 inclusion 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 validation scheme proceeds as described in Section 3.1. The application of the proposed reduction strategy results in a significant decrease in the computational cost of the validation process. Our approach does involve a once-off initialisation step, requiring time $O(n \log n)$ 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, the costs associated with clustering and prediction assessment are both reduced to $O((\frac{n}{2\rho})^2)$.

3.3 Application to Document Clustering

While our proposed method may be used in conjunction with any valid kernel function, for document clustering we make use of a linear kernel that has been normalised according to the approach described in [4], resulting in a kernel matrix that is equivalent to that produced by the standard cosine similarity measure. Although this kernel 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 [18] 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. This frequently leads to a non-trivial improvement in validation performance. A summary of the complete validation process is provided in Figure 2.

As mentioned previously, our proposed method is based on the assumption that regions will be locally homogeneous, which should generally be the case when an appropriate kernel function is chosen. To maximise homogeneity, we select a low value for the number of neighbours, with $p = 5$ being used for our experiments in the following section. Empirical evidence suggests 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 is also related to the maximum number of runs τ, where the computational gains resulting from prototype reduction allows the use of a larger value to guarantee the robustness of the overall validation procedure. It must be stressed that, in our experiments, the use of these "general purpose" parameter values proved to be

```
Initialisation Phase
  • Extract candidate prototypes $\mathcal{S}$, consisting of $n$ neighbourhood centroid vectors.
  • Evaluate compactness of candidates in $\mathcal{S}$ and sort accordingly in descending order.
  • Uniformly select set of $n'$ reduced prototypes $\mathcal{S}'$ from the ordered list.
  • 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 $\tau$ 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) Predict the assignment of documents in $\mathcal{X}_b$ based on centroids from clustering of $\mathcal{X}_a$.
          (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. 2. Complete kernel prediction-based validation scheme, with prototype reduction.

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 [13] (KM-P). 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. 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. 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 of "artificial" data. We created 84 datasets in total, containing clusters of different proportions which vary in their degree of overlap. A full discussion on the construction of these datasets is provided in [5].

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 eight real-world corpora that have previously been used in document clustering. For further details of these datasets, consult [5]. Table 2 shows the results of the comparison, indicating the top three estimated values for $\hat{k}$ on the real corpora. 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. It is interesting to note

Table 2. Summary of top-3 estimations for $\hat{k}$ on real datasets.

Dataset	$\hat{k}$	KM-S	KM-P	KKM-S	RED-S
bbc	5	**5**, 4, 6	**5**, 6, 7	**5**, 6, 4	**5**, 6, 4
bbcsport	5	4, **5**, 3	**5**, 6, 4	**5**, 6, 4	**5**, 4, 6
classic3	3	**3**, 2, 4	**3**, 2, 4	**3**, 4, 5	**3**, 4, 5
classic	4	3, 5, 2	3, 5, 2	5, **4**, 2	5, **4**, 2
cstr	4	3, 2, **4**	3, **4**, 2	3, **4**, 5	3, **4**, 5
ng3	3	**3**, 4, 2	**3**, 4, 5	**3**, 4, 2	**3**, 2, 4
ng17-19	3	5, 4, 6	5, 4, 6	5, 4, **3**	4, 5, **3**
reviews	5	2, 3, 6	2, 8, 9	2, **5**, 4	2, 6, 3

that, as with the artificial data, the kernel-based methods generally outperformed those relying on the standard k-means algorithm. 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 the computational demands of the validation process by employing a form of prototype reduction to construct a reduced kernel matrix. To ensure that this 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 2001-21, Stanford University (2001)
4. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press (2001)
5. Greene, D., Cunningham, P.: Efficient prediction-based validation for document clustering. Technical Report CS-2006-??, Trinity College Dublin (2006)
6. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a dataset via the gap statistic. Technical report, Stanford University (2000)
7. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics **6** (1978) 461–464
8. Bezdek, J.C., Kuncheva, L.: Nearest prototype classifier designs: An experimental study. International Journal of Intelligent Systems **16** (2001) 1445–1473
9. Hamamoto, Y., Uchimura, S., Tomita, S.: A bootstrap technique for nearest neighbor classifier design. IEEE Tran. Pattern Anal. Machine Intell. **19** (1997) 73–79
10. Kim, S.W., Oommen, B.J.: On using prototype reduction schemes and classifier fusion strategies to optimize kernel-based nonlinear subspace methods. IEEE Tran. Pattern Anal. Machine Intell. **27** (2005) 455–460
11. Kollios, G., Gunopulos, D., Koudas, N., Berchtold, S.: Efficient biased sampling for approximate clustering and outlier detection in large datasets. IEEE Tran. Knowledge and Data Engineering **15** (2003)
12. Juan, A., Vidal, E.: Comparison of four initialization techniques for the k-medians clustering algorithm. In: Advances in Pattern Recognition: Joint IAPR International Workshops, Springer-Verlag (2000) 842–852
13. 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
14. Hubert, L.J., Arabie, P.: Comparing partitions. Journal of Classification **2** (1985) 193–218
15. Cutting, D.R., Pedersen, J.O., Karger, D., Tukey, J.W.: Scatter/gather: A cluster-based approach to browsing large document collections. In: Proc. 15th Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval. (1992) 318–329
16. Ding, C., He, X.: K-nearest-neighbor consistency in data clustering: incorporating local information into global optimization. In: Proc. 2004 ACM symposium on Applied computing, ACM Press (2004) 584–589
17. 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
18. Greene, D., Cunningham, P.: Practical solutions to the problem of diagonal dominance in kernel document clustering,. In: Proc. 23rd International Conference on Machine Learning. (2006)