Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering

Derek Greene, Pádraig Cunningham

University of Dublin, Trinity College,
Dublin 2, Ireland

Abstract. In supervised kernel methods, it has been observed that the performance of the SVM classifier is poor in cases where the diagonal entries of the Gram matrix are large relative to the off-diagonal entries. This problem, referred to as *diagonal dominance*, often occurs when certain kernel functions are applied to sparse high-dimensional data, such as text corpora. In this paper we investigate the implications of diagonal dominance for unsupervised kernel methods, specifically in the task of document clustering. We discuss a selection of strategies for addressing this issue, and evaluate their effectiveness in producing more accurate and stable clusterings.

1 Introduction

In many domains it will often be the case that the average similarity of one object to another will be small when compared to the "self-similarity" of the object to itself. This characteristic of many popular similarity measures does not constitute a problem for some similarity-based machine learning techniques. However, it does pose a problem for supervised kernel methods. If, for a given kernel function, self-similarity values are large relative to between-object similarities, the Gram matrix of this kernel will exhibit *diagonal dominance*. This will result in poor generalisation accuracy when using Support Vector Machines (Smola & Bartlett, 2000; Cancedda *et al.*, 2003; Schölkopf *et al.*, 2002). Recently Dhillon *et al.* (2004) suggested that this issue might also impact upon the performance of centroid-based kernel clustering algorithms, as the presence of large self-similarity values can limit the extent to which the solution space is explored beyond the initial state.

An unfortunate characteristic of this problem is that matrices which are strongly diagonally dominanted will be positive semi-definite and measures to reduce this dominance run the risk of rendering the matrix indefinite so that it no longer represents a valid Mercer kernel. Consequently there is a tension between diagonal dominance on the one hand and the requirement that the matrix be positive semi-definite on the other.

In this paper we are concerned with the implications of diagonal dominance for clustering documents using the kernel k-means algorithm. As such, we compare several practical techniques for addressing the problem. We examine the

use of subpolynomial kernels, which have the effect of "flattening" the range of values in the kernel matrix. We also explore the use of a diagonal shift to reduce the trace of the kernel matrix. Since both techniques can render the matrix indefinite, we evaluate the use of the empirical kernel map (Schölkopf *et al.*, 1999) to overcome this. Finally, we consider an algorithmic approach that involves adjusting the kernel k-Means algorithm to remove the influence of self-similarity values.

The evaluation presented in Section 4 demonstrates that all these reduction approaches have merit. An interesting point arising from the experiments is that the techniques employing indefinite kernel matrices still produce good clusterings and can be terminated after a tractable number of iterations without a significant decrease in clustering accuracy. This suggests that kernel k-means may not be as susceptible to this issue as supervised kernel-based techniques, when considering text data. Before presenting our results, the issue of diagonal dominance in kernel clustering is discussed in Section 2 and the details of the techniques under evaluation are described in Section 3.

2 Dominant Diagonals in Kernel Clustering

Kernel methods involve the transformation of a dataset to a new, possibly high-dimensional, space where non-linear relationships between objects may be more easily identified. Rather than explicitly computing the representation $\phi(x)$ of each object x, the application of the "kernel trick" allows us to consider the affinity between a pair of objects x_i and x_j using a given kernel function κ, which is defined in terms of the dot product

$$\kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \tag{1}$$

By re-formulating algorithms using only dot products and subsequently replacing these with kernel affinity values, we can efficiently apply learning algorithms in the new non-linear space. The kernel function κ is usually represented by an $n \times n$ kernel matrix (or Gram matrix) $\mathbf{K}$, where $K_{ij} = \kappa(x_i, x_j)$. Following this notation, the squared Euclidean distance between a pair of objects in the transformed space can be expressed as

$$||\phi(x_i) - \phi(x_j)||^2 = K_{ii} + K_{jj} - 2K_{ij} \tag{2}$$

This may be used as a starting point for the identification of structures in the new space.

2.1 Kernel K-means

A variety of popular clustering techniques have been re-formulated for use in a kernel-induced space, including the standard k-means algorithm. Given a set of

objects $\{x_1, \ldots, x_n\}$, the kernel k-means algorithm (Schölkopf *et al.*, 1998) seeks to minimise the objective function

$$\sum_{a=1}^{k} \sum_{x_i \in C_a} ||\phi(x_i) - \mu_a||^2 \tag{3}$$

for clusters $\{C_1, \ldots, C_k\}$, where μ_c represents the centroid of cluster C_c. Rather than explicitly constructing centroid vectors in the transformed feature space, distances are computed using dot products only. From Eqn. (2), we can formulate the squared object-centroid distance $||\phi(x_i) - \mu_c||^2$ as the expression

$$K_{ii} + \frac{\sum_{x_j, x_l \in C_c} K_{jl}}{|C_c|^2} - \frac{2 \sum_{x_j \in C_c} K_{ij}}{|C_c|} \tag{4}$$

The first term above may be excluded as it remains constant; the second is a common term representing the self-similarity of the centroid, which need only be calculated once for each cluster; the third term represents the affinity between x_i and the centroid μ_c.

The kernelised algorithm proceeds in the same manner as standard batch k-means, alternating between reassigning objects to clusters and updating the centroids until convergence. In this paper we follow the standard convention of regarding the clustering procedure as having converged only when the assignment of objects to centroids no longer changes from one iteration to another.

2.2 Diagonal Dominance

It has been observed (Cancedda *et al.*, 2003; Schölkopf *et al.*, 2002) that the performance of the SVM classifier can be poor in cases where the diagonal values of the Gram matrix are large relative to the off-diagonal values. This problem, sometimes referred to as *diagonal dominance* in machine learning literature, frequently occurs when certain kernel functions are applied to data that is sparse and high-dimensional in its explicit representation. It is particularly problematic in text mining tasks, where linear or string kernels can often produce diagonally dominated Gram matrices. However, this phenomenon can also arise with other kernel functions, such as when employing the Gaussian kernel with a small smoothing parameter, or when using domain-specific kernels for learning tasks in image retrieval (Tao *et al.*, 2004) and bioinformatics (Saigo *et al.*, 2004). These cases are all characterised by the tendency of the average of the diagonal entries of the kernel matrix $\mathbf{K}$ to be significantly larger than the average of the off-diagonal entries, resulting in a *dominance ratio*

$$\frac{\frac{1}{n} \sum_i K_{ii}}{\frac{1}{n(n-1)} \sum_{i,j,i \neq j} K_{ij}} \gg 1 \tag{5}$$

We can interpret this to mean that the objects are approximately orthogonal to one another in this representation. In many cases a classifier applied to such a matrix will effectively memorise the training data, resulting in severe overfitting.

The phenomenon of diagonal dominance also has implications for centroid-based kernel clustering methods. Observe that, when calculating the dissimilarity between a centroid μ_a and a document $x_i \in C_a$, the expression (4) can be separated as follows:

$$K_{ii} + \frac{\sum_{x_j, x_l \in C_a} K_{jl}}{|C_a|^2} - \frac{2\sum_{x_j \in C_a - \{x_i\}} K_{ij}}{|C_a|} - \frac{2K_{ii}}{|C_a|} \tag{6}$$

If $\mathbf{K}$ is diagonally dominated, the last term in Eqn. (6) will often result in x_i being close to the centroid of C_a and distant from the remaining clusters, regardless of the affinity between x_i and the other documents assigned to C_a. Consequently, even with random cluster initialisation, few subsequent reassignments will be made and the algorithm will converge to a poor local solution.

The problem of dominant self-similarity has previously been shown to adversely affect centroid-based clustering algorithms in high-dimensional feature spaces (Dhillon *et al.*, 2002). Therefore, it is unsurprising that similar problems should arise when applying their kernel-based counterparts using kernel functions that preserve this sparseness. Dhillon *et al.* (2004) observed that the accuracy of kernel k-means can decrease significantly when document-cluster distances are dominated by self-similarity values.

For the remainder of the paper, we make use of a linear kernel that has been normalised according to the approach described by Schölkopf & Smola (2001), yielding values in the range $[0, 1]$. The matrix of this normalised kernel, denoted here as $\mathbf{S}$, corresponds to the similarity matrix of the widely used cosine similarity measure, so that

$$S_{ij} = \frac{\langle x_i, x_j \rangle}{\sqrt{\langle x_i, x_i \rangle \langle x_j, x_j \rangle}} \tag{7}$$

While a kernel formulated in this way represents an intuitive choice for document clustering, its matrix will typically suffer from diagonal dominance. Thus, although we will always have $S_{ii} = 1 \, \forall i$, it will often be the case for sparse text data that $S_{ij} \ll 1$ for $i \neq j$.

As an example, we consider the *cstr* dataset[1], which consists of 505 technical abstracts relating to four fields of research. For a matrix $\mathbf{S}$ constructed from this data, the dominance ratio (5) is 16.54, indicating that the matrix is significantly diagonally dominated. This can be seen clearly in the graphical representation of the matrix in Figure 1. When applying kernel k-means using this matrix, the large diagonal entries may prevent the identification of coherent clusters. Often incorrect assignments in an initial partition will fail to be subsequently rectified as the large self-similarity may obscure similarities between pairs of documents belonging to the same natural grouping.

[1] http://www.cs.rochester.edu/trs

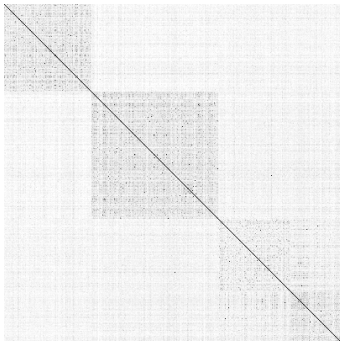

Fig. 1. Linear kernel matrix for *cstr* dataset with dominant diagonal.

3 Reducing Diagonal Dominance

In this section we describe a number of practical strategies for reducing the effects of diagonal dominance.

3.1 Diagonal Shift (DS)

To reduce the influence of large diagonal values, Dhillon *et al.* (2004) proposed the application of a negative shift to the diagonal of the Gram matrix. Specifically, a multiple σ of the identity matrix is added to produce

$$\mathbf{K_{DS}} = \sigma\mathbf{I} + \mathbf{S} \tag{8}$$

The parameter σ is a negative constant, typically selected so that the trace of the kernel matrix is approximately zero. For a normalised linear kernel matrix with trace equal to n, this will be equivalent to subtracting 1 from each diagonal value, thereby eliminating the first and last terms from the document-centroid distance calculation (6).

However, the shift technique is equivalent to the addition of a negative constant to the eigenvalues of $\mathbf{S}$. As a result, $\mathbf{K_{DS}}$ will no longer be positive semi-definite and the kernel k-means algorithm is not guaranteed to converge when applied to this matrix. Figure 2 compares the trailing eigenvalues for the matrix shown in Figure 1, before and after applying a diagonal shift of $\sigma = -1$. Notice that the modification of the diagonal entries has the effect of shifting a large number of eigenvalues below zero, signifying that the modified matrix is indefinite.

The application of diagonal shifts to Gram matrices has previously proved useful in supervised kernel methods. However, rather than seeking to reduce diagonal dominance, authors have most frequently used the technique to ensure that a kernel matrix is positive semi-definite. Both Saigo *et al.* (2004) and Wu

et al. (2005) proposed the addition of a non-negative constant to transform indefinite symmetric matrices into valid kernels. Unfortunately, this will have the side effect of increasing the dominance ratio. Specifically, Saigo *et al.* (2004) suggested adding a shift $\sigma = |\lambda_n|$, where λ_n is the negative eigenvalue of the kernel matrix with largest absolute value. However, as evident from Figure 2, such a shift will often negate the benefits of the diagonal shift, resulting in a matrix that is once again diagonally dominated. In addition, computing a full spectral decomposition for a large term-document matrix will often be impractical, although Wu *et al.* (2005) did suggest an approach for estimating λ_n.

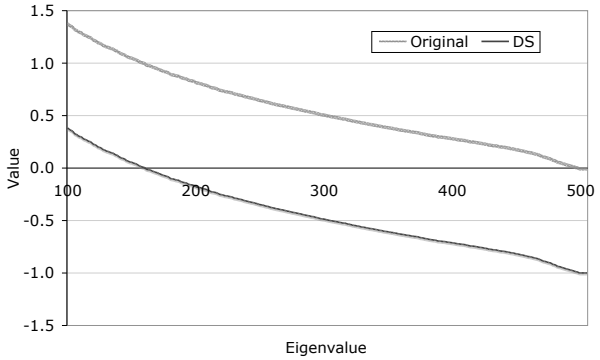

Fig. 2. Eigenvalues in range $[100, 505]$ for normalised linear kernel matrix of *cstr* dataset.

3.2 Subpolynomial Kernel With Empirical Kernel Map (SPM)

To address the problems introduced by large diagonals in SVMs, Schölkopf *et al.* (2002) proposed the use of *subpolynomial* kernels. Given a positive kernel based on the function ϕ, a subpolynomial kernel function is defined as

$$\kappa_{SP}(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle^p \tag{9}$$

where $0 < p < 1$ is a user-defined parameter. As the value of the degree p decreases, the ratio of diagonal entries to off-diagonal entries in the matrix $\mathbf{K}_{SP}$ also decreases. Unlike the diagonal shift technique, the non-linear transformation directly modifies the pair-wise affinities between the off-diagonal entries in $\mathbf{S}$, which may potentially distort the underlying cluster structure.

Since κ_{SP} may no longer be a valid kernel, the authors suggest the use of the *empirical kernel map* method (Schölkopf *et al.*, 1999) to render the matrix positive definite. This involves mapping each document x_i to an n-dimensional feature vector

$$\phi_m(x_i) = (\kappa(x_i, x_1), \ldots, \kappa(x_i, x_n))^T \tag{10}$$

By using this feature representation, we can derive a positive definite kernel matrix by simply computing the dot products

$$\mathbf{K_{SPM}} = \mathbf{K_{SP}}\mathbf{K_{SP}}^T \tag{11}$$

In practice, normalising all rows of $\mathbf{K_{SP}}$ to unit length prior to computing the dot product leads to significantly superior results.

An important issue that must be addressed when using a subpolynomial kernel is the selection of the parameter p. If the value is too large the Gram matrix will remain diagonally dominated, while a value of p that is too small will obscure cluster structure as all documents will become approximately equally similar. Schölkopf *et al.* (2002) suggest the use of standard cross-validation techniques for selecting p. However, this may not be feasible in cases where other key parameters such as the number of clusters k must also be determined by repeatedly clustering the data.

3.3 Diagonal Shift With Empirical Kernel Map (DSM)

While the empirical map technique was used by Schölkopf *et al.* (2002) to produce a valid kernel from the matrix of a subpolynomial kernel, this approach can be applied in combination with other reduction methods. Thus, even if we alter the diagonal of the kernel matrix in an arbitrary manner so that it becomes indefinite, we may still recover a positive definite matrix that will guarantee convergence for the kernel k-means algorithm.

Here we consider the possibility of applying a negative shift so as to minimise the trace of the matrix as described previously. This is followed by the construction of the empirical map $\mathbf{K_{DSM}} = \mathbf{K_{DS}}\mathbf{K_{DS}}^T$, after normalising the rows of $\mathbf{K_{DS}}$ to unit length. While this approach does reduce the dominance ratio 5, it should be noted that the application of the dot product will produce a kernel matrix with trace greater than zero.

3.4 Algorithm Adjustment (AA)

When attempting to apply supervised kernel methods to matrices that are not positive semi-definite, Wu *et al.* (2005) distinguished between two fundamental strategies: *spectrum transformation* approaches that perturb the original matrix to produce a valid Gram matrix, and *algorithmic* approaches that involve altering the formulation of the learning algorithm. A similar distinction may be made between diagonal dominance reduction techniques. We now discuss an algorithmic approach that involves adjusting the kernel k-means algorithm described by Schölkopf *et al.* (1998) to eliminate the influence of self-similarity values.

If one considers the distance between a document x_i and the cluster C_a to which it has been initially assigned, a dominant diagonal will lead to a large value in the third term of Eqn. (4). As noted in Section 2.2, this will often cause x_i to remain in C_a during the reassignment phase, regardless of the affinity between x_i and the other documents in C_a. A potential method for alleviating this problem

is to reformulate the reassignment step as a "split-and-merge" process, where self-similarity values are not considered. Rather, we seek to assign each document to the nearest centroid, where the document itself is excluded during centroid calculation.

Formally, each document x_i is initially removed from its cluster C_a, leaving a cluster $C_a - \{x_i\}$ with centroid denoted $\mu_{a'}$. For each alternative candidate cluster C_b, $b \neq a$, we consider the gain achieved by reassigning x_i to C_b rather than returning it back to C_a. This gain is quantified by the expression

$$\Delta_{ab} = ||\phi(x_i) - \mu_{a'}||^2 - ||\phi(x_i) - \mu_b||^2 \tag{12}$$

Note that from Eqn. (4), the diagonal value K_{ii} is not considered in the computation of Δ_{ab}. If $\arg\max_b \Delta_{ab} > 0$, then x_i is reassigned to that cluster C_b which results in the maximal gain. Otherwise, x_i remains in cluster C_a. As with the standard formulation of kernel k-means, centroids are only updated after all n documents have been examined.

This strategy could potentially be applied to improve the performance of the standard k-means algorithm in sparse spaces where self-similarity values have undue influence. However, the repeated adjustment of centroids in a high-dimensional space is likely to be impractical. Fortunately, in the case of kernel k-means we can efficiently compute Δ_{ab} by caching the contribution of each document to the common term in Eqn. (4), making it unnecessary to recalculate the term in its entirety when evaluating each document for reassignment.

3.5 Comparison of Reassignment Behaviour

Dhillon *et al.* (2002) observed that spherical k-means often becomes trapped at an initial clustering, where the similarity of any document to its own centroid is much greater than its similarity to any other centroid. As discussed previously, a diagonally dominated kernel matrix frequently elicits similar behaviour from the kernel k-means algorithm. Consequently, the algorithm will converge after relatively few reassignments have been made to a local solution that is close to the initial partition. If the initial clusters are randomly selected, it is possible that the final clustering will be little better than random. In addition, multiple runs may produce significantly different partitions of the same data.

To gain a clearer insight into this problem, we examine the reassignment behaviour resulting from the application of each of the reduction strategies. Figure 3 illustrates the expected number of reassignments occurring during the first 10 iterations of the kernel k-means algorithm when applied to the *cstr* dataset. It is evident that applying the algorithm to the original dominated matrix results in significantly fewer reassignments, which can be viewed as a cursory search of the solution space. It is interesting to note that these reassignment patterns are replicated to varying degrees across all the datasets considered in our evaluation in Section 4.

Clearly the number of reassignments may not necessarily be a good predictor of clustering accuracy. However, the experimental results presented in the next

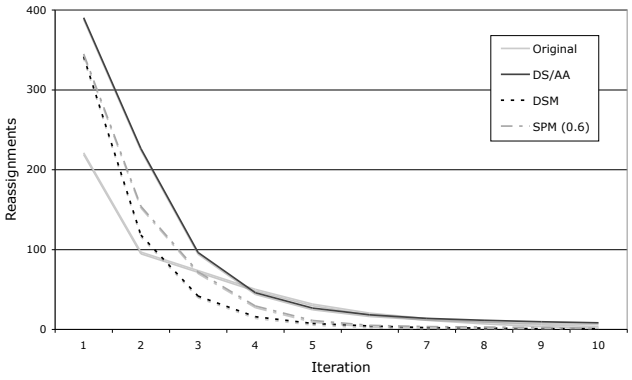

Fig. 3. Average number of assignments per iteration for *cstr* dataset over first 10 iterations.

section do suggest a significant correlation between depth of search and clustering accuracy. In particular, the methods DS and AA, which show equivalent reassignment behaviour, frequently perform well.

4 Empirical Evaluation

4.1 Experimental Setup

In order to assess the reduction methods described in Section 3, we conducted a comparison on eight datasets that have previously been used in the evaluation of document clustering algorithms (see Table 1). For further information regarding these document collections, consult Greene & Cunningham (2005). To pre-process the datasets we applied stop-word removal and stemming techniques. We subsequently removed terms occurring in less than three documents and applied standard TF-IDF normalisation to the feature vectors.

Dataset	Description	Documents	Terms	k	Ratio
bbc	News articles from BBC	2225	9635	5	24.18
bbcsport	Sports news articles	737	4613	5	16.19
classic3	CISI/CRAN/MED	3893	6733	3	39.47
classic	CACM/CISI/CRAN/MED	7097	8276	4	46.47
cstr	Computer science technical abstracts	505	2117	4	16.54
ng17-19	Overlapping newsgroups	2625	11841	3	28.70
ng3	Approximately disjoint newsgroup	2900	12875	3	30.13
reviews	Entertainment news articles (TREC)	4069	18391	5	27.18

Table 1. Details of experimental datasets, including original dominance ratios.

We initialised the clustering procedure with random clusters and averaged the results over 250 trials. For each trial, a maximum of 100 assignment iterations

was permitted. We set the number of clusters k to correspond to the number of natural classes in the data. For experiments using a subpolynomial kernel, we tested values for the degree parameter from the range $[0.4, 0.9]$. Values for $p < 0.4$ invariably resulted in excessive flattening of the range of values in the kernel matrix, producing partitions that were significantly inferior to those generated on the original matrix.

When comparing the accuracy of document clustering techniques, external class information is generally used to assess cluster quality. We employ the *normalised mutual information* (NMI) measure proposed by Strehl & Ghosh (2002), which provides a robust indication of the level of agreement between a given clustering and the target set of natural classes.

An alternative approach to cluster validation is based on the notion of *cluster stability* (Roth *et al.*, 2002), which refers to the ability of a clustering algorithm to consistently produce similar solutions across multiple trials on data originating from the same source. It is well documented that the k-means algorithm and its variations are particularly sensitive to initial starting conditions. This makes them prone to converging to different local minima when using a stochastic initialisation strategy. Therefore, when selecting a diagonal reduction method, we seek to identify a robust approach that will allow us to consistently produce accurate, reproducible clusterings. In our experiments we assessed the stability of each candidate method by calculating the *average normalised mutual information* (ANMI) (Strehl & Ghosh, 2002) between the set of all partitions generated on each dataset.

4.2 Analysis of Results

Our experiments indicate that all of the reduction approaches under consideration have merit. In particular, Table 2 shows that the AA and DS methods yield improved clustering accuracy in all but one case. Generally, we observed that diagonal dominance reduction has a greater effect on some datasets than on others. While the difference in reassignment behaviour after reduction is less pronounced on datasets such as *classic3*, there is no strong correlation between the distribution of the affinity values in the kernel matrix and the increase in accuracy. However, it is apparent from Table 3 that applying kernel k-means to a dominated kernel matrix consistently results in poor stability. It is clear that the restriction placed upon the number of reassignments made in these cases frequently results in less deviation from the initial random partition, thereby increasing the overall disagreement between solutions.

Diagonal Shift (DS). Table 2 shows that the negative diagonal shift approach frequently produced clusterings that were more accurate than those generated on the original dominated kernel matrices. As noted in Section 3.1, this method provides no guarantee of convergence. However, our results support the assertion made by Dhillon *et al.* (2004) that, in practice, lack of convergence may not always be a problem. Frequently we observed that a comparatively stable

Dataset	Original	DS	DSM	AA	p=0.4	p=0.5	p=0.6	p=0.7	p=0.8	p=0.9
bbc	0.81	**0.83**	0.81	**0.83**	0.81	0.82	0.81	0.81	0.81	0.81
bbcsport	0.72	**0.80**	0.78	**0.80**	0.69	0.75	0.76	0.76	0.76	0.78
classic3	**0.94**	**0.94**	0.90	**0.94**	0.88	0.89	0.89	0.89	0.89	0.90
classic	0.74	**0.75**	**0.75**	**0.75**	0.71	0.73	**0.75**	0.74	0.74	0.74
cstr	0.64	**0.74**	**0.74**	**0.74**	0.57	0.69	0.71	0.72	0.73	0.73
ng17-19	0.38	0.40	0.46	0.40	0.45	0.46	0.46	0.45	**0.47**	0.45
ng3	0.82	0.83	0.84	0.84	0.84	0.85	**0.86**	0.85	0.85	0.83
reviews	0.58	0.59	0.60	0.58	0.61	**0.62**	**0.62**	0.61	0.61	0.60

Table 2. Accuracy (NMI) scores for reduction methods, with linear kernels and subpolynomial kernels for various values of p.

Dataset	Original	DS	DSM	AA	p=0.4	p=0.5	p=0.6	p=0.7	p=0.8	p=0.9
bbc	0.82	0.86	0.90	0.87	**0.94**	**0.94**	0.92	0.92	0.90	0.89
bbcsport	0.64	0.79	**0.82**	0.79	0.77	0.80	0.80	0.80	0.79	0.81
classic3	0.98	0.98	0.97	0.98	0.99	**1.00**	0.99	0.99	**1.00**	0.99
classic	0.86	0.89	0.81	**0.90**	0.79	0.79	0.80	0.80	0.80	0.79
cstr	0.60	0.78	**0.83**	0.79	0.82	0.82	0.82	0.81	0.81	0.81
ng17-19	0.45	0.48	0.60	0.47	0.62	**0.64**	0.63	0.61	0.62	0.60
ng3	0.81	0.83	0.92	0.85	**1.00**	**1.00**	0.99	0.97	0.94	0.91
reviews	0.77	0.81	0.84	0.80	0.92	**0.98**	0.95	0.90	0.87	0.84

Table 3. Stability (ANMI) scores for reduction methods, with linear kernels and subpolynomial kernels for various values of p.

partition is identified after a relatively few number of iterations. At this stage the algorithm proceeds to oscillate indefinitely between two nearly identical solutions without ever attaining convergence. As a solution to this problem, we chose to terminate the reassignment procedure after five consecutive oscillations were detected. This resulted in no significant adverse effect on clustering accuracy. However, the lack of complete convergence did impact upon the stability of the partitions generated using the DS method, as apparent by the relatively low ANMI scores reported in Table 3.

Diagonal Shift With Empirical Kernel Map (DSM). While the application of the empirical kernel map technique subsequent to a diagonal shift does guarantee convergence after relatively few iterations, the map also has the effect of increasing the dominance ratio, resulting in accuracy gains that are not so significant as those achieved by the DS approach. The higher level of consistency between partitions generated using this method does suggest that it represents a good trade-off between accuracy and stability. However, there remains the additional computational expense of constructing the matrix $\mathbf{K}_{DSM}$, which requires $O(n^3)$ time.

Subpolynomial Kernel With Empirical Kernel Map (SPM). For subpolynomial kernel reduction method, our experimental findings underline the

difficulty of setting the degree parameter p. The gains in accuracy resulting from this approach were significant, though less consistent than those achieved by the other methods. On certain datasets, such as the *3ng* and *reviews* collections, specific values of p lead to a large improvement in both accuracy and stability, while in other cases there was little or no improvement. This suggests that the alteration of cluster structure induced by the subpolynomial function may prove beneficial in some cases, but not in others. Therefore, while a value of $p = 0.6$ was found to perform best on average, we conclude that the selection of a value for p is largely dataset dependent. Once again, the expense of calculating the empirical map must be taken into consideration when making use of this reduction method.

It is interesting to note that employing a subpolynomial kernel without subsequently rendering the kernel matrix positive definite still resulted in complete convergence in all experiments. However, the accuracy and stability scores returned in these cases were generally lower. As with the DSM approach, the application of the empirical map resulted in a marked increase in cluster stability.

Algorithm Adjustment (AA). The adjusted kernel clustering algorithm, as described in Section 3.4, yielded improvements in accuracy that were marginally better than those produced by the diagonal shift method (DS), while also achieving slightly higher cluster stability scores. The correlation between the two methods is understandable given their similar reassignment behaviour. This stems from the fact that applying a negative diagonal shift of $\sigma = -1$ to a matrix with trace equal to n effectively eliminates the dominant last term in Eqn. (6), leading to document-centroid distances that are approximately the same as those achieved using the "split-and-merge" adjustment. It should be noted that, while the AA reduction method frequently failed to achieve complete convergence, the oscillation detection technique described previously resolved this problem satisfactorily on all datasets.

5 Conclusion

We have considered a range of practical solutions to the issues introduced by diagonally dominated kernel matrices in unsupervised kernel methods. Furthermore, we have demonstrated the effectiveness of the solutions when performing the task of document clustering. From our evaluation it is apparent that the presence of disproportionately large self-similarity values precipitates a reduction in the number of reassignments made by the kernel k-means algorithm. This may limit the extent to which the solution space is explored, causing the algorithm to become stuck close to its initial state. In cases where the initialisation strategy is stochastic or unsuitable, this can result in a appreciable decrease accuracy and cluster stability.

For practical purposes, the diagonal shift and adjusted k-means techniques both represent efficient strategies for reducing diagonal dominance. However, it

is possible that the tendency of these methods to become trapped in a cycle of oscillating reassignments may prove problematic under certain circumstances. Applying the empirical kernel map technique subsequent to a negative diagonal shift leads to a good trade-off between accuracy and stability, although the cost of computing the empirical map may be prohibitive for large datasets. This factor is also relevant when employing subpolynomial kernels to reduce diagonal dominance. In addition, for this latter reduction method we conclude that the choice of the degree p is largely dataset dependent. The requirement of an additional user-selected parameter in the clustering process makes this approach less attractive than the other methods we have discussed.

An interesting direction for future research would be to investigate the factors that determine the extent to which diagonal dominance reduction affects clustering accuracy. In addition, we believe that the techniques described in this paper will also have merit in the application of unsupervised kernel methods to other domains such as bioinformatics and image retrieval, where the ratio of diagonal to off-diagonal entries in the kernel matrix will often be significantly higher.

Bibliography

Cancedda, N., Gaussier, E., Goutte, C. & Renders, J.M. (2003). Word sequence kernels. *J. Mach. Learn. Res.*, **3**, 1059–1082.

Dhillon, I., Guan, Y. & Kulis, B. (2004). A unified view of kernel k-means, spectral clustering and graph cuts. Tech. Rep. TR-04-25, UTCS.

Dhillon, I.S., Guan, Y. & Kogan, J. (2002). Iterative clustering of high dimensional text data augmented by local search. In *Proceedings of The 2002 IEEE International Conference on Data Mining*.

Greene, D. & Cunningham, P. (2005). Producing accurate interpretable clusters from high-dimensional data. Tech. Rep. TCD-CS-2005-42, Department of Computer Science, Trinity College Dublin.

Roth, V., Braun, M., Lange, T. & Buhmann, J. (2002). A resampling approach to cluster validation. In *Proceedings of the 15th Symposium in Computational Statistics*.

Saigo, H., Vert, J.P., Ueda, N. & Akutsu, T. (2004). Protein homology detection using string alignment kernels. *Bioinformatics*, **20**, 1682–1689.

Schölkopf, B. & Smola, A.J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

Schölkopf, B., Smola, A. & Müller, K.R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, **10**, 1299–1319.

Schölkopf, B., Mika, S., Burges, C., Knirsch, P., Müller, K.R., Rätsch, G. & Smola, A. (1999). Input space versus feature space in kernel-based methods. *IEEE Transactions on Neural Networks*, **5**, 1000–1017.

Schölkopf, B., Weston, J., Eskin, E., Leslie, C. & Noble, W.S. (2002). A kernel approach for learning from almost orthogonal patterns. In *ECML '02: Proceedings of the 13th European Conference on Machine Learning*, 511–528, Springer-Verlag, London, UK.

Smola, A.J. & Bartlett, P.J., eds. (2000). *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, USA.

Strehl, A. & Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *JMLR*, **3**, 583–617.

Tao, Q., Scott, S., Vinodchandran, N.V., Osugi, T.T. & Mueller, B. (2004). An extended kernel for generalized multiple-instance learning. In *16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004)*, 272–277.

Wu, G., Chang, E. & Zhang, Z. (2005). An analysis of transformation on non-positive semidefinite similarity matrix for kernel machines. Tech. rep., UCSB.