An Efficient Alternative to SVM Based Recursive Feature Elimination with Applications in Natural Language Processing and Bioinformatics

Justin Bedo[1,2], Conrad Sanderson[1,2], and Adam Kowalczyk[1,2,3]

[1] Australian National University, ACT 0200, Australia
[2] National ICT Australia (NICTA), Locked Bag 8001, ACT 2601, Australia
[3] Dept. Electrical & Electronic Eng., University of Melbourne, VIC 3010, Australia

Abstract. The SVM based Recursive Feature Elimination (RFE-SVM) algorithm is a popular technique for feature selection, used in natural language processing and bioinformatics. Recently it was demonstrated that a small regularisation constant C can considerably improve the performance of RFE-SVM on microarray datasets. In this paper we show that further improvements are possible if the explicitly computable limit $C \to 0$ is used. We prove that in this limit most forms of SVM and ridge regression classifiers scaled by the factor $\frac{1}{C}$ converge to a centroid classifier. As this classifier can be used directly for feature ranking, in the limit we can avoid the computationally demanding recursion and convex optimisation in RFE-SVM. Comparisons on two text based author verification tasks and on three genomic microarray classification tasks indicate that this straightforward method can surprisingly obtain comparable (at times superior) performance and is about an order of magnitude faster.

1 Introduction

The *Support Vector Machine* based *Recursive Feature Elimination* (RFE-SVM) approach [1] is a popular technique for feature selection and subsequent classification, especially in the bioinformatics area. At each iteration a linear SVM is trained, followed by removing one or more "bad" features from further consideration. The goodness of the features is determined by the absolute value of the corresponding weights used in the SVM. The features remaining after a number of iterations are deemed to be the most useful for discrimination, and can be used to provide insights into the given data. A similar feature selection strategy was used in the *author unmasking* approach, proposed for the task of authorship verification [2] (a sub-area within the natural language processing field). However, rather than removing the worst features, the best features were iteratively dropped.

Recently it has been observed experimentally on two microarray datasets that using very low values for the regularisation constant C can improve the performance of RFE-SVM [3]. In this paper we take the next step and rather than pursuing the elaborate scheme of recursively generating SVMs, we use the limit $C \to 0$. We show that this limit can be explicitly calculated and results in a centroid based classifier. Furthermore, unlike RFE-SVM, in this limit the removal of one or more

A. Sattar and B.H. Kang (Eds.): AI 2006, LNAI 4304, pp. 170–180, 2006.

features does not affect the solution of the classifier for the remaining features. The need for multiple recursion is hence obviated, resulting in considerable computational savings.

This paper is structured as follows. In Section 2 we calculate the limit $C \to 0$. In Sections 3 and 4 we provide empirical evaluations on two authorship attribution tasks and on three bioinformatics tasks, respectively, showing that the centroid based approach can obtain comparable (at times superior) performance. A concluding discussion is given in Section 5.

2 Theory

Empirical risk minimisation is a popular machine learning technique in which one optimises the performance of an estimator (often called a predictor or hypothesis) on a "training set" subject to constraints on the hypothesis class. The predictor f, which is constrained to belong to a Hilbert space $\mathcal{H}$, is selected as the minimiser of a regularised risk which is the sum of two terms: a regulariser that is typically the squared norm of the hypothesis, and a loss function capturing its error on the training set, weighted by a factor $C > 0$.

The well studied limit of $C \to \infty$ results in a hard margin Support Vector Machine (SVM). In this paper, we focus on the other extreme, the limit when $C \to 0$. This is a heavy regularisation scheme as the regularised risk is dominated by the regulariser. We formally prove that if the classifier is scaled by the factor $1/C$, then it converges in most cases to a well defined and explicitly calculable limit: the centroid classifier.

We start with an introduction of the notation necessary to precisely state our formal results. Given a training set $\mathbb{Z} = \{(\mathbf{x}_i, y_i)\}_{i=1,\dots,m} \in \mathbb{R}^n \times \{\pm 1\}$ a Mercer kernel $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with the associated reproducing kernel Hilbert space $\mathcal{H}$, and with a feature map $\Phi : \mathbb{R}^n \to \mathcal{H}$, $\mathbf{x} \mapsto k(\mathbf{x}, .)$ [4,5], the regularised risk on the training set is defined as

$$R^{reg}[f, b] := \|f\|^2_{\mathcal{H}} + \sum\nolimits_{y=\pm 1} C_y \sum\nolimits_{i, y_i = y} l(1 - y_i(f(\mathbf{x}_i) + b)) \qquad (1)$$

for any $(f, b) \in \mathcal{H} \times \mathbb{R}$. Here $C_{+1}, C_{-1} \geq 0$ are two class dependent regularisation constants. These constants can be always written in the form

$$C_y = C \frac{1 + yB}{2m_y}, \qquad (2)$$

where $C > 0$, $0 \leq B \leq 1$ and $m_y := |\{i \,;\, y_i = y\}|$ is the number of instances with label $y = \pm 1$. Further, $l(\xi)$ is the *loss function* of one of the following three forms: $l(\xi) := \xi^2$ for the *ridge regression* (RRegr), the *hinge* loss $l(\xi) := \max(0, \xi)$ for the *support vector machine* (SVM) with the *linear* loss, and $l(\xi) := \max(0, \xi)^2$ for the SVM^2 with *quadratic* loss [4,5,6].

We are concerned with *kernel machines* defined as

$$f := f_H := \arg \min_{f \in \mathcal{H}} R^{reg}[f, 0], \qquad (3)$$

in the *homogeneous case* (no bias b), or, in general, as the sum

$$f = f_H + b, \qquad (f_H, b) := \arg \min_{(f,b) \in \mathcal{H} \times \mathbb{R}} R^{reg}[f, b]. \qquad (4)$$

2.1 Centroid Based Classification and Feature Selection

The *centroid classifier*, is defined for every $-1 \leq B \leq 1$ as

$$g_B(x) = \frac{1+B}{2} \sum\nolimits_{i,y_i=+1} \frac{k(x,x_i)}{m_{+1}} - \frac{1-B}{2} \sum\nolimits_{i,y_i=-1} \frac{k(x,x_i)}{m_{-1}} + b, \tag{5}$$

where b is a bias term. In terms of the feature space, it is the projection onto the vector connecting (weighted) arithmetic means of samples from both class labels, degenerating to the mean of a single class for $B = \pm 1$, respectively. For a suitable kernel k it implements the difference between Parzen window estimates of probability densities for both labels. For the linear kernel it simplifies to

$$g_B(\mathbf{x}) = \mathbf{w}_B \cdot \mathbf{x} + b := \left(\frac{1+B}{2} \bar{\mathbf{x}}_{+1} - \frac{1-B}{2} \bar{\mathbf{x}}_{-1} \right) \cdot \mathbf{x} + b \tag{6}$$

where $\bar{\mathbf{x}}_y$ is the centroid of the samples with label $y = \pm 1$ and $\mathbf{w}_B \in \mathbb{R}^n$ is the *centroid weight vector*. In this work we have used $B = 0$ and $b = \mathbf{w}_B \cdot (\bar{\mathbf{x}}_{+1} + \bar{\mathbf{x}}_{-1})/2$, such that the decision hyperplane is exactly halfway between $\bar{\mathbf{x}}_{+1}$ and $\bar{\mathbf{x}}_{-1}$. The weight vector can be used for feature selection in the same manner as in RFE-SVM, i.e. the features are selected based on a ranking given by $r_i = |\mathbf{w}_B^{(i)}|$, where $\mathbf{w}_B^{(i)}$ is i-th dimension of $\mathbf{w}_B$. However, unlike RFE-SVM, the removal of a feature does not affect the solution for the remaining features, thus the need for recursion is obviated. This form of combined feature selection and subsequent classification will be referred to as the *centroid approach.*

For the case of the linear kernel, we note that the centroid classifier can be interpreted as a form of *Linear Discriminant Analysis (LDA)* [7], where the two classes share a diagonal covariance matrix with equal diagonal entries.

2.2 Formal Results

In the following theorem we prove that using either form of *quadratic* loss specified above, the machine converges to the centroid classifier as $C \to 0$; the case of linear loss is covered by Theorem 2.

Theorem 1. *Let $|B| < 1$, the loss be either $l(\xi) = \xi^2$ or $l(\xi) = \max(0,\xi)^2$ and*

$$f_C := f_{H,C} + b_C \quad \text{and} \quad (f_{H,C}, b_C) := \arg\min_{f,b} R^{reg}[f,b]$$

for every $C > 0$. Then $\lim_{C \to 0^+} |\frac{f_C}{C}| = \infty$ if $B \neq 0$ but $\lim_{C \to 0^+} \frac{f_{H,C}}{C} = \frac{1-B^2}{2} g_0$.

Thus although the whole predictor f_C scaled by C^{-1} diverges with $C \to 0$, its homogeneous part $f_{H,C}$ converges to an explicitly calculable limit.

Proof outline. The kernel trick reduces the proof to the linear kernel $k(\mathbf{x}, \mathbf{x}') := \mathbf{x} \cdot \mathbf{x}'$ on $\mathbb{R}^n$ and $\mathcal{H}$ isomorphic to $\mathbb{R}^n$. In such a case $f(\mathbf{x}) = f_H(\mathbf{x}) + b = \mathbf{x} \cdot \mathbf{w} + b$, where $\mathbf{w} \in \mathbb{R}^n$ and $\|f_H\|^2_{\mathcal{H}} = \|\mathbf{w}\|^2$. The regularised risk (1-2) takes the form

$$R^{reg}[\mathbf{w}, b] = \|\mathbf{w}\|^2 + C \sum\nolimits_i \frac{y_i + B}{2m_{y_i}} l(1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + b))$$

and $f_C(\mathbf{x}) = \mathbf{w}_C \cdot \mathbf{x} + b_C$ for $\mathbf{x} \in \mathbb{R}^n$, where $(\mathbf{w}_C, b_C) = \arg\min_{\mathbf{w},b} R^{reg}[\mathbf{w}, b]$. This implies

$$\|\mathbf{w}_C\|^2 \leq \min_{\mathbf{w},b} R^{reg}[\mathbf{w}, b] \leq R^{reg}[0, 0] = C \tag{7}$$

and due to continuous differentiability of the loss functions in our case:

$$\frac{\partial R^{reg}}{\partial b}[\mathbf{w}_C, b_C] = 0 \quad \text{and} \quad \frac{\partial R^{reg}}{\partial \mathbf{w}}[\mathbf{w}_C, b_C] = 0. \tag{8}$$

Solving the first of these equations we obtain

$$b_C = B - \sum\nolimits_{i \in SV} \frac{1 + y_i B}{2m_{y_i}} \mathbf{w}_C \cdot \mathbf{x}_i$$

where $SV := \{i : 1 - y_i \mathbf{x}_i \mathbf{w}_C - y_i b_C > 0\}$ is the set of the *support vectors* in the case of the hinge loss and the whole training set in the case of regression. Furthermore,

$$\left| \sum\nolimits_{i \in SV} \frac{1 + y_i B}{2m_{y_i}} \mathbf{w}_C \cdot \mathbf{x}_i \right| \leq \|\mathbf{w}_C\| \max_i \|\mathbf{x}_i\| \leq \sqrt{C} \max_i \|\mathbf{x}_i\|,$$

by (7), hence, denoting by $O(\xi)$ a term such that $|O(\xi)/\xi|$ is bounded on for $\xi > 0$, we get

$$b_C = B + O(\sqrt{C}). \tag{9}$$

Case $l(\xi) = \xi^2$. The first equation in (8) takes the form

$$2\mathbf{w}_C - 2C \sum\nolimits_i \frac{y_i + B}{2m_{y_i}} (1 - y_i(\mathbf{w}_C \cdot \mathbf{x}_i + b_C))\mathbf{x}_i = 0$$

which upon consideration of (7) and (9) implies

$$\frac{f_{H,C}(\mathbf{x})}{C} = \sum\nolimits_i y_i \frac{(1 + y_i B)(1 - y_i B + O(\sqrt{C}))}{2m_{y_i}} \mathbf{x}_i \cdot \mathbf{x} = \frac{1 - B^2}{2} g_0(\mathbf{x}) + O(\sqrt{C}),$$

$$\frac{f_C(\mathbf{x})}{C} = \frac{f_{H,C}(\mathbf{x}) + b_C}{C} = \frac{1 - B^2}{2} g_0(\mathbf{x}) + O(\sqrt{C}) + \frac{B + O(\sqrt{C})}{C}.$$

This immediately proves the current case of the theorem.

Case $l(\xi) = \max(0, \xi)^2$. This case reduces to the previous one, once we observe that every training sample becomes a support vector for sufficiently small C. Indeed, Eqns. (7) and (9) imply

$$1 - y_i \mathbf{w}_C \cdot \mathbf{x}_i - y_i b_C = 1 - y_i B + O(\sqrt{C}) > 0$$

if C is sufficiently small, since $|B| < 1$. □

Note that the above result does not cover the most popular case of SVM, namely the linear loss $l(\xi) := \max(0, \xi)$. The example in Figure 1 shows that in such a case the limit depends on the particular data configuration in a complicated way and cannot be simply expressed in terms of the centroid classifier.

The extension of Theorem 1 to the case of homogeneous machines follows. A proof, similar to the above one, is not included.

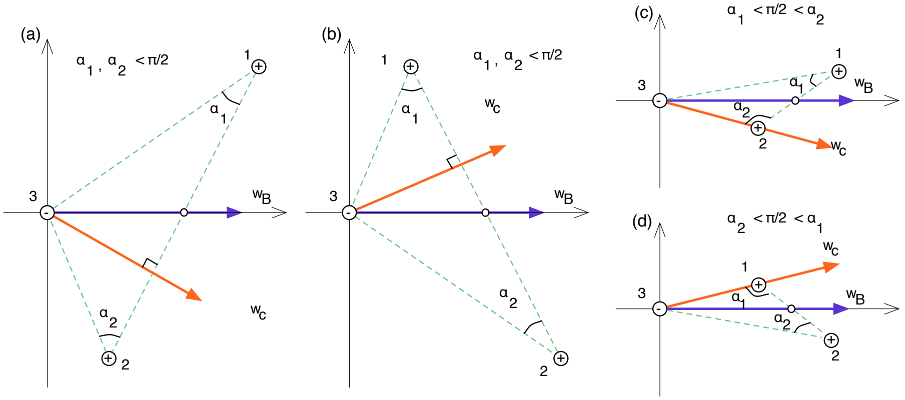

Fig. 1. Four subsets of $\mathbb{R}^2$ that share the same centroid classifier (the blue line), but which have different vectors $\mathbf{w}_C$ (the red line) for the SVM[1] solutions $\mathbf{x} \rightarrow \mathbf{w}_C \cdot \mathbf{x} + b_C$

Theorem 2. *Let $|B| \leq 1$, the loss be either $l(\xi) = \xi^2$ or $l(\xi) = \max(0, \xi)^p$ for $p = 1, 2$ and $f_C := \arg\min_f R^{reg}[f, 0]$ for every $C > 0$. Then $\lim_{C \to 0^+} f_C/C = g_B(\boldsymbol{x})$.*

Using the kernel $k(\mathbf{x}, \mathbf{x}') + 1$ rather than $k(\mathbf{x}, \mathbf{x}')$ in the last theorem we obtain:

Corollary 1. *Let $f_C := f_{H,C} + b_C$ for every $C > 0$, where $(f_{H,C}, b_C) := \arg\min_{f,b} R^{reg}[f, 0] + b^2$. Then $\lim_{C \to 0^+} f_C/C = g_B(\boldsymbol{x}) + B$.*

Note the differences between the above of results. Theorem 2 and Corollary 1 hold for $B = \pm 1$, which is virtually the case of one-class-SVM, while in Theorem 1 this is excluded. The limit $\lim_{C \to 0} f_C/C$ exists in the case of Corollary 1 but does not exist in the case of Theorem 1. Finally, in Theorem 1 the convergence is to a scaled version of the balanced centroid g_0 for all $|B| < 1$, while in the latter two results to g_B or $g_B + B$, respectively.

3 Authorship Verification Experiments

Recently an RFE-SVM based technique was proposed for the task of authorship verification [2], a particular problem that spans the fields of natural language processing and humanities [8,9]. Given a reference text from a purported author and a text of unknown authorship, an *author unmasking* curve is built as follows. Both texts are divided into chunks, with each chunk represented by counts of preselected words. The chunks are partitioned into training and test sets. Each point in the author unmasking curve is the accuracy of discriminating (using a linear SVM) between the test chunks from the two texts. At the end of each iteration several of the most discriminative words are removed from further consideration. A vector representing the essential features of the curve is then classified as representing either the “same author” or a “different author” [2].

The underlying hypothesis is that if the two given texts have been written by the same author, the differences between them will be reflected in a relatively

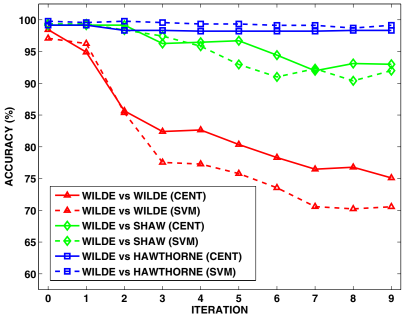

Fig. 2. Demonstration of the unmasking effect for Wilde's *An Ideal Husband* using Wilde's *Woman of No Importance* as well as the works of other authors as reference texts.

Table 1. Classification accuracies for same-author (SA) and different-author (DA) curves on the Gutenberg dataset, using centroid and RFE-SVM (R/S) based unmasking. The second term in the method indicates the secondary classifier type.

Method	SA Acc.	DA Acc.	Time
Cent. + Cent.	92.3%	95.4%	14 mins
Cent. + SVM	92.3%	95.2%	14 mins
R/S + Cent.	69.2%	97.2%	124 mins
R/S + SVM	92.3%	94.9%	124 mins

Table 2. As per Table 1, but for the Columnists dataset

Method	SA Acc.	DA Acc.	Time
Cent. + Cent.	82%	89.6%	99 mins
Cent. + SVM	86%	90.3%	104 mins
R/S + Cent.	84%	88%	709 mins
R/S + SVM	84%	88.5%	714 mins

small number of features. In [2] it was observed that for texts authored by the same person, the extent of the accuracy degradation is much larger than for texts written by different authors.

We used two datasets in our experiments: Gutenberg and Columnists. The former is comprised of 21 books[1] from 10 authors (as used in [2]), while the latter (used in [10]) is comprised of 50 journalists (each covering several topics) with two texts per journalist; each text has approximately 5000 words.

The setup of experiments was similar to that in [2], with the main difference being that books by the same author were not combined into one large text. For the Gutenberg dataset we used chunks with a size of 500 words, 10 iterations of removing 6 features, and using 250 words with the highest average frequency in both texts as the set of pre-selected words. For each pair of texts, 10 fold cross-validation was used to obtain 10 curves, which were then represented by a mean curve prior to further processing. For the Columnist dataset we used a chunk size of 200 words and 100 pre-selected words (based on preliminary experiments).

A leave-one-text-out approach was used, where the remaining texts were used for generation of same-author (SA) and different-author (DA) curves, which in turn were used for training a secondary classifier. The left-out text was then unmasked against the remaining texts and the resulting curves classified. For the Gutenberg dataset, there was a total of 24 SA and 394 DA classifications, while for Columnists there were 100 SA and 9800 DA classifications. Both the SVM and the centroid were also used as secondary classifiers.

[1] Available from Project Gutenberg (www.gutenberg.org).

As can be observed in Figure 2 (which closely resembles Fig. 2 in [2]) the unmasking effect is also well present for the centroid based approach. Tables 1 and 2 indicate that the RFE-SVM and the centroid based unmasking approaches are comparable in terms of performance, but differ greatly in terms of wall clock time[2]. The centroid based approach is about seven to nine times faster even if we neglect the significant time required for searching for the optimal C for SVM.

4 Experiments with Classification of Microarray Datasets

In this section we perform experiments on three publicly available microarray datasets: *colon* cancer [1,3,11], *lymphoma*[3] cancer [3,12,13], and *Cancer of Unknown Primary (CUP)* [14]. All these datasets are characterised by a relatively small number of high dimensional samples. The colon dataset contains 62 samples (22 normal and 40 cancerous), with each sample containing the expression values for 2000 genes. The lymphoma dataset contains three subtypes, with a total of 62 samples. Each sample contains measurements of 4026 expressed genes. We used a subset of the CUP dataset, containing samples for 226 primary and metastatic tumours, representing 12 tumour types[4], with each sample containing expressions for approximately 10500 genes measured on a cDNA microarray platform. The number of samples per class widely varied, form 8 to 50.

In addition to comparing the performance of the RFE-SVM and centroid approaches, we also evaluated the multi-class *Shrunken Centroid (SC)* approach [15], proposed specifically for processing microarray datasets. Briefly, in SC the class centroids are shrunk towards the overall centroid, involving a form of t-statistic which standardises each feature for each class by its within-class standard deviation. The degree of shrinking for all features is controlled by single tunable parameter. A feature is not used for prediction of a class when it has been shrunk "past" the overall centroid. We note that one of the main differences between SC and the centroid approach (described in Section 2.1) is that in the latter standard deviation estimates are not used.

For classification of multiclass datasets (lymphoma and CUP) with the centroid and SVM approaches we have used the popular one-vs-all architecture [16]. Feature selection for these datasets was done follows. Each class had its own ranking of features, of which the top T were selected. The final set of features was the union of the selected features from all classes. Note that since the class-specific feature subsets can intersect, the final set can have less than KT features, where K is the number of classes.

The experiment setup for the colon and lymphoma datasets followed [3]. Testing was based on 50 random splits of data into disjoint training and test sets. In each split approximately 66% of the samples from each class were assigned to

[2] Using LibSVM 2.71 (www.csie.ntu.edu.tw/~cjlin/libsvm) on a Pentium 4, 3.2 GHz.

[3] Available from http://llmpp.nih.gov/lymphoma/data.shtml

[4] The full CUP dataset contains 13 tumour types, however we have omitted the smallest class ('testicular') which was represented by only three samples.

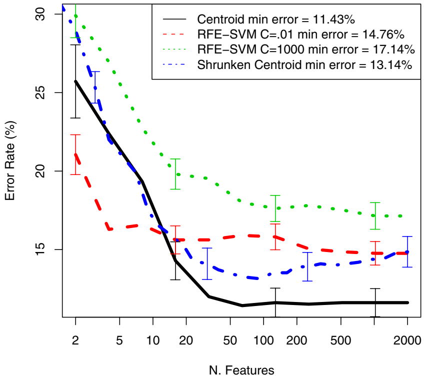

Fig. 3. Results for the colon dataset

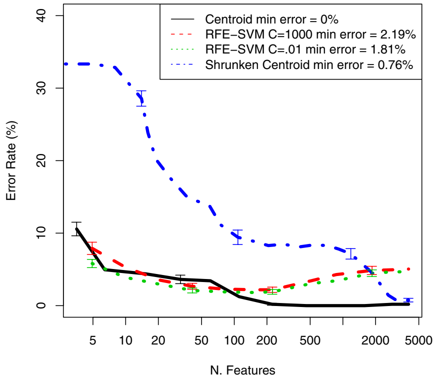

Fig. 4. Results for the lymphoma dataset

the training set, with the remainder assigned to the test set. Results are shown[5] in Figures 3 and 4.

As the value of C decreases, the performance of RFE-SVM improves, achieving a minimum error rate of 14.76% and 1.82% with $C = 0.01$ on the colon and lymphoma datasets, respectively. This agrees with results in [3]. When using the $C \to 0$ limit (i.e. the centroid approach), the minimum error rate drops to 11.43% and 0% on the colon and lymphoma datasets, respectively. The SC approach obtained a comparable minimum of 13.14% and 0.76% on the respective datasets. However, on the lymphoma dataset its error rate increased rapidly when fewer features were used. Specifically, the centroid approach selects about 200 genes that achieve near perfect discrimination, while the SC approach requires the full set of 4026 genes to achieve its minimum error rate.

The CUP dataset was compiled for building a clinical test for determination of the origin of a cancer from the tissue of a secondary tumour. In [14] it was demonstrated (for a subset of five cancers), that in order for CUP classification to be practical for use in pathology tests, it is necessary to select a relatively small subset of marker genes which are then used to build a considerably cheaper and also more robust test, i.e., via a Polymerase Chain Reaction based microfluidic card, where the maximum number of gene probes is 384 [14]. Results in Figure 5 indicate that the centroid method outperforms both RFE-SVM and SC in the practically important region of a few hundred genes.

Finally, wall clock timing results shown in Table 3 indicate that the centroid approach is between four to 17 times faster than RFE-SVM. For for the two-class dataset (colon), the SC and centroid approaches have comparable time requirements. For multi-class datasets (lymphoma and CUP), the SC approach is roughly twice as fast as the centroid approach, due to the use of the one-vs-all architecture for the latter.

[5] We have evaluated a large set of C values for RFE-SVM, however for purposes of clarity only a subset of the results is shown in Figures 3, 4 and 5.

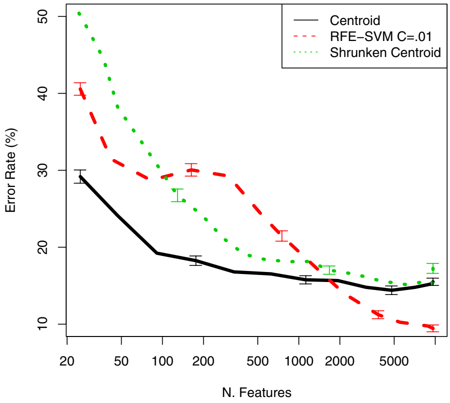

Fig. 5. Results for the CUP dataset

Table 3. Timing results for the three bioinformatics datasets for a single 50-permutation test (1.83 GHz Intel Core Duo machine)

Dataset	Algorithm	Time
Colon	RFE-SVM	146 sec.
	Centroid	20 sec.
	Shrunken Cent.	28 sec.
Lymp.	RFE-SVM	505 sec.
	Centroid	113 sec.
	Shrunken Cent.	66 sec.
CUP	RFE-SVM	19844 sec.
	Centroid	1135 sec.
	Shrunken Cent.	478 sec.

5 Discussion and Conclusions

In this paper we have shown that the limit $C \rightarrow 0$ of most forms of SVMs is explicitly computable and converges to a centroid based classifier. As this classifier can be used directly for feature ranking, in the limit we can avoid the computationally demanding recursion and convex optimisation in RFE-SVM. We have also shown that on some real life datasets the use of the centroid approach can be advantageous.

The results may at first be surprising, raising the question: *Why is the centroid approach a competitor to its more sophisticated counterparts?* One explanation is that its relative simplicity makes it well matched for classification and feature selection on datasets comprised of a small number of samples from a noisy distribution. On such datasets, even the simple estimates of variance which are employed by, say, the Shrunken Centroid approach [15], are so unreliable that their usage brings more harm than good. Indeed, the results presented in Section 4 support this view.

Another independent argument can be built on the theoretical link established in Section 2, relating the centroid classifier to the extreme case of regularisation of SVMs. Once we accept the well supported perspective of the SVM community that regularisation is the way to generate robust empirical estimators and classifiers from noisy data [4,6], it naturally follows that we should get robust classifiers in the limit of extreme regularisation. It is worth noting that the same argument can be extended to the theoretical and empirical comparison with LDA [7], which due to space restrictions is not included here.

We note that the centroid approach has been applied and used sporadically in the past: we have already linked it to the "ancient" Parzen window; its kernel form can be found in [4] and its application to classification of breast cancer microarrays is given in a well cited paper [17]. However, what this paper brings is a formal link to SVMs and RFE-SVM in particular. Furthermore, to our knowledge, this type of centroid based feature selection and subsequent classification has not been explored previously and its application to genomic microarray classification as well as natural language processing (author unmasking) is novel.

To conclude, we have shown that the centroid approach is a good baseline alternative to RFE-SVM, especially when dealing with datasets comprised of a low number of samples from a very high dimensional space. It it very fast and straightforward to implement, especially in the linear case. It has a nice theoretical link to SVMs and regularisation networks, and given the right circumstances it can deliver surprisingly good performance. We also recommend it as a baseline classifier, facilitating quick sanity cross checks of more involved supervised learning techniques.

Acknowledgements

We thank Simon Guenter for useful discussions. National ICT Australia (NICTA) is funded by the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council.

References

1. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning **46** (2002) 389–422
2. Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. In: Proc. 21st Int. Conf. Machine Learning (ICML), Banff, Canada (2004)
3. Huang, T.M., Kecman, V.: Gene extraction for cancer diagnosis by support vector machines - an improvement. Artificial Intelligence in Medicine **35** (2005) 185–194
4. Schölkopf, B., Smola, A.: Learning with Kernels. MIT Press, Cambridge (2002)
5. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK (2000)
6. Vapnik, V.: Statistical Learning Theory. John Wiley and Sons, New York (1998)
7. Duda, R., Hart, P., Stork, D.: Pattern Classification. John Wiley & Sons (2001)
8. Gamon, M.: Linguistic correlates of style: authorship classification with deep linguistic analysis features. In: Proc. 20th Int. Conf. Computational Linguistics (COLING), Geneva (2004) 611–617
9. Love, H.: Attributing Authorship: An Introduction. Cambridge University Press, U.K. (2002)
10. Sanderson, C., Guenter, S.: Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. In: Proc. 2006 Conf. Empirical Methods in Natural Language Processing (EMNLP), Sydney (2006) 482–491
11. Ambroise, C., McLachlan, G.: Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. National Acad. Sci. **99** (2002) 6562–6566
12. Alizadeh, A., Eisen, M., Davis, R., et al.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature **403** (2000) 503–511
13. Chu, F., Wang, L.: Gene expression data analysis using support vector machines. In: Proc. Intl. Joint Conf. Neural Networks. (2003) 2268–2271

14. Tothill, R., Kowalczyk, A., Rischin, D., Bousioutas, A., Haviv, I., et al.: An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. Cancer Research **65** (2005) 4031–4040
15. Tibshirani, R., Hastie, T., et al.: Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Statistical Science **18** (2003) 104–117
16. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. Journal of Machine Learning Research **5** (2004) 101–141
17. van't Veer, L., Dai, H., van de Vijver, M., He, Y., Hart, A., et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature **415** (2002) 530–536