Feature Selection using One Class SVM: A New Perspective

Yamuna Prasad *
Department of CSE
I.I.T. Delhi
New Delhi, India

K. K. Biswas †
Department of CSE
I.I.T. Delhi
New Delhi, India

Parag Singla ‡
Department of CSE
I.I.T. Delhi
New Delhi, India

Abstract

Feature selection is an important task for mining useful information from datasets in high dimensions, a typical characteristic of biology domains such as microarray datasets. In this paper, we present an altogether new perspective on feature selection. We pose feature selection as a one class SVM problem of modeling the space in which features can be represented. We show that finding the support vectors in our one class formulation is tantamount to performing feature selection. Further, we show that our formulation reduces to the standard QPFS formulation in the dual problem space. Not only our formulation gives new insights into the task of feature selection, solving it directly in the primal space can give significant computational gains when the number of the samples is much smaller than the number of features. We validate our thesis by experimenting on three different microarray datasets.

1 Introduction

DNA microarray technology measures the mRNA level of thousands of genes simultaneously under certain conditions in a cell (sample). These mRNA levels represent the gene expression values and are known as gene expression dataset. The analysis of this data is typically carried out through classification/regression [1]. These datasets are characterized by very large number (thousands) of features(genes) while the number of samples are very small (hundreds) [2]. It is well known that the presence of large number of features in a dataset leads to poor generalization accuracy and high execution time [3].

Several methods have been proposed in the literature to reduce the dimensionality of microarray data in classification (or regression) tasks. There are mainly two ways of dealing with this curse of dimensionality. In the first approach, the data is transformed into an entirely new low dimensional sub-space. PCA, IDA, LDA are examples of this approach [4]. One key problem with this approach is that the dimensions in the transformed sub-space may not correspond to any physical interpretation to the domain expert. In the second approach, dimensionality reduction is achieved by selecting a subset of original features using Feature Selection methods [1, 2, 5, 6, 7, 8]. The objective of feature selection is to find a minimal subset of non-redundant and relevant features from the data which maximizes classification/regression accuracy. The features obtained by this approach can generally be directly interpreted by the domain expert. The quadratic programming feature selection (QPFS) [7] proposed recently has been shown to outperform existing feature selection methods such as mRMR, MaxRel and reliefF [6].

*http://www.cse.iitd.ernet.in/~yprasad and e-mail:yprasad@cse.iitd.ac.in
†http://www.cse.iitd.ernet.in/~kkb and e-mail:kkb@cse.iitd.ac.in
‡http://www.cse.iitd.ernet.in/~parags and e-mail:parags@cse.iitd.ac.in

In this paper, we pose the problem of feature selection as a one class SVM problem. One class SVMs represent the underlying data by finding a hyperplane which maximally separates the data points from the origin [9, 10]. They have been widely used for outlier detection. Our formulation strives to separate the set of features from the set of non-features (outliers) in the space where each feature represents a data point and each example represents a dimension. This can be done by finding a hyperplane which maximally separates the given set of features from the origin. The support vectors describing the hyperplane boundary are exactly the set of features which are required for representing the underlying set of features. Hence, the task of feature selection essentially corresponds to finding these support vectors. Under the assumption of unit norm (which is typically the case for feature selection), the one class SVM formulation using a hyperplane boundary becomes equivalent to the Support Vector Data Description (SVDD) formulation [10], which tries to find a hypersphere which most compactly encloses the given set of points.

We show that our formulation corresponds to the QPFS formulation [7] in the dual problem space. This helps in presenting a principled perspective of feature selection in both the primal as well as the dual space. When the number of samples is significantly less than the number of features, as is the case with problems in biology domains such as microarray datasets, it can be computationally much more efficient to solve the problem directly in the primal space. Experiments on three different microarray datasets corroborate our claim.

The rest of the paper is organized as follows. We describe our proposed formulation in Section 2. Experimental results are presented in Section 3. We conclude our work in Section 4.

2 Proposed Framework for Feature Selection

The main goal in feature selection is to select a subset of features which jointly minimize redundancy and maximize relevance. One way to achieve this goal is to select the subset of features which can describe the boundary (hyperplane) separating the set of features from the set of non-features (outliers). This framework is inspired by the one class Support Vector Machine (SVM) [9] formulation where we are looking for a hyperplane which separates the given set of points from the outliers. Typically, there exist a subset of points which is sufficient to describe the separating hyperplane. The points in this subset are called support vectors [9]. In our formulation, the features ($f_i, i = 1, ..., M$) represent the data points and examples ($x_i, i = 1, \ldots, N$) correspond to the dimensions. The support vectors correspond to the set of support (informative) features. One class SVM formulation [9] for this can be written as:

$$\begin{aligned} &\min_{w,b} \frac{1}{2} w^T w + b \\ &\text{subject to} \\ &\quad w^T \phi(f_i) + b \geq 1, \forall i = 1, ..., M; \end{aligned} \tag{1}$$

where ϕ is a transformation in the dot product space and can be computed via a kernel $k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$, w is a normal to the separating hyperplane $w^T \phi(x_i) + b = 0$ and b is the bias term.

In any given feature selection task, the goal is to maximize the relevance and minimize the redundancy [1, 6, 7]. Typically, an explicit relevance vector r ($r_i, i = 1, \ldots, M$) is computed based on correlation or mutual information with class labels [7]. In our formulation, we allow each feature to have a separate margin boundary based on its relevance(r_i). Greater the relevance, larger the margin. Redundancy (similarity) is captured implicitly in our framework. The features lying on the respective margin boundaries can be considered as non-redundant while those lying beyond the respective margin boundaries can be considered as redundant. The choice of ϕ (transformed space) determines the kind of correlation among the features (more details on this later).

Based on the above artifacts, we present the following primal formulation for feature selection :

$$\begin{aligned} &\min_{w,b} \frac{1}{2} w^T w + b \\ &\text{subject to} \\ &\quad w^T \phi(f_i) + b \geq r_i, \forall i = 1, ..., M; \end{aligned} \tag{2}$$

Figure 1 illustrates the intuition behind our proposed framework in the linear dot product space. In the figure, $w^T f + b = 0$ represents the separating hyprplane. The distance of this hyperplane from the origin is given by $-b/||w||$. The first term in the objective of (2) tries to minimize $w^T w$ i.e. maximize $1/||w||$. The second term in the objective tries to minimize b i.e. maximize $-b$. Hence, the overall objective tries to push the plane away from the origin.

The ith dashed plane represents the margin boundary for the ith feature. The distance of this marginal hyperplane from the separating hyperplane is given by $r_i/||w||$ where r_i is the pre-computed relevance of the ith feature. Therefore, minimizing $w^T w$ in the objective also amounts to maximizing this marginal distance ($r_i/||w||$). Hence, the objective has the dual goal of pushing the hyperplane away from the origin while also maximizing the margin for each feature (weighted by its relevance). The features which lie on the respective marginal planes are the support features (encircled points).

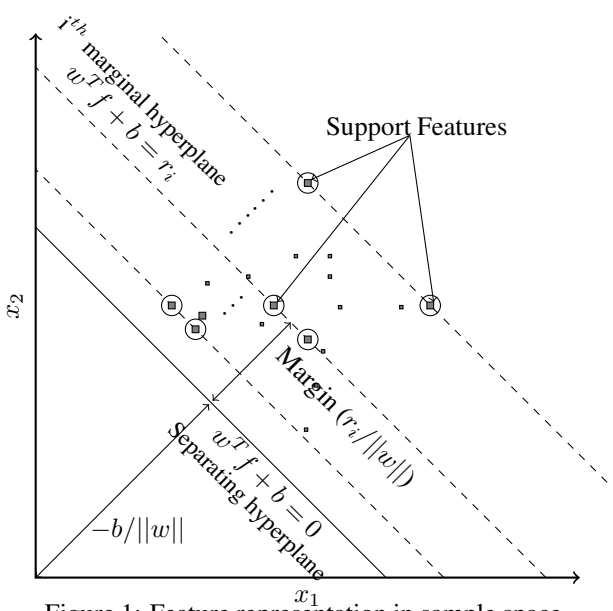

Figure 1: Feature representation in sample space

The redundancy is usually captured using correlation or mutual information in feature selection tasks [7]. In our framework, the dot product space (kernel) captures the similarity (redundancy) among the features [11]. The required similarity metric can be captured by selecting the appropriate dot product space. The linear kernel ($f_i^T f_j$) represents the correlation among the features when the features are normalized to zero mean and unit variance[1]. Since the value of the correlation ranges between -1 and 1, a degree two homogeneous polynomial kernel defined over normalized data represents the squared correlation (i.e. $\phi(f_i)^T \phi(f_j) = (f_i^T f_j)^2$). The choice of this kernel is quite intuitive for feature selection as it gives equal importance to the positive and negative correlations. Gaussian kernel can also be used to approximate the mutual information (MI) [12] which is the key metric for non-linear redundancy measure in feature selection problems [6, 7]. The dual formulation of (2) using the method of Lagrangians can be described as follows:

$$\max_{\alpha} \sum_{i=1}^{M} \alpha_i r_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \alpha_i \alpha_j k(f_i, f_j)$$

$$\text{subject to} \tag{3}$$

$$\alpha_i \geq 0, \forall i = 1, ..., M; \text{and } \sum_{i=1}^{M} \alpha_i = 1$$

where $k(,)$ is a suitable kernel function corresponding to the transformation ϕ such that $k(f_i, f_j) = \phi(f_i)^T \phi(f_j)$. Note that the dual precisely represents the QPFS formulation [7] where $k(f_i, f_j)$ represents the entries of the similarity matrix Q and r_i represents the relevance of ith feature f_i. In (3), the relevance and redundancy have been given equal importance. In order to incorporate a relative importance (weight) for relevance and redundancy a scalar parameter $\theta \in [0, 1)$ is introduced and the first term of the dual objective ($\sum_{i=1}^{M} \alpha_i r_i$) is scaled by $\theta/(1-\theta)$ [7]. This is equivalent to scaling the r_i's by $\theta/(1-\theta)$ in the primal.

Let D be the dimensions of the transformed space ϕ. Then the number of variables and the number of constraints in the primal formulation (2) are $D + 1$ and M, respectively. The corresponding numbers for the dual are M and $M + 1$, respectively. Typically in microarray datasets, the number

[1]It is typical to normalize the data to zero mean and unit variance for feature selection.

of samples is much less compared to the number of features [1]. When the number of samples in the mapped space ϕ is very small compared to the number of features ($D<<M$), the primal optimization will be faster than the dual. Using linear and second degree polynomial transformations the size of the transformed space (D) will still be very small (assuming $D<<M$) compared to the number of features and therefore, it would be beneficial to solve the primal. For the large values of D (possibly infinite as in the case of Gaussian kernel) solving the dual will be beneficial.

3 Experiments

3.1 Algorithms and Datasets

In our experiments, we compared the performance of our proposed feature selection approach in the primal as well as the dual. We used both the linear correlation (LinC) and the squared correlation (SqC) to compute the feature similarity and the feature relevance. We also compared our methods with QPFS [7] using mutual information. For our experimental study, we used three publicly available benchmark microarray datasets, namely, Lymphoma, Leukemia and Type 2 Diabetes (T2D) datasets with 4026(45), 7129(72) and 22283(34) features(samples), respectively [2]. These datasets are binary classification datasets and have been used for feature selection by many researchers [1, 2, 5, 6, 8]. We report the leave-one-out cross-validation (LOOCV) [2] accuracy.

The value of the scale parameter θ was fixed to 0.5 in each of the methods for the purpose of comparison. After feature selection was done, linear SVM (L2-regularized L2-loss support vector classification in primal) [13] was used to train a classifier using the optimal set of features output by each algorithm. The code was implemented in Matlab. All the experiments were run on a Intel Core™ i7 3.10GHz with 8GB RAM.

3.2 Results

Table 1 compares the average execution times (in seconds) of solving the feature selection in the primal and the dual using LinC and SqC. For Lymphoma and Leukemia LinC in the primal is three orders of magnitude faster than the dual. Using SqC the primal is an order of magnitude faster than the dual. For the T2D dataset the execution in the dual ran out of memory and could not be completed whereas the primal had no issues. We also compared the accuracies for both LinC and

Table 1: Average execution time (in seconds)

Dataset	**LinC**		**SqC**	
	Primal	**Dual**	**Primal**	**Dual**
Lymphoma	**1.1**	**1039.1**	**25.1**	1035.3
Leukemia	**1.1**	**5606.2**	**375.83**	5609.4
T2D	**1.3**	–	**8.4**	–

SqC[2]. As expected SqC performed better than LinC because of flexibility in modeling the decision boundaries. The best set of accuracies (varying the number of top-K features selected) were as follows: Lymphoma(LinC:100, SqC:100), Leukemia(LinC:90.27, SqC:97.22) and T2D(LinC:100, SqC:100). Running QPFS [7] with mutual information as similarity measure gave the same set of accuracies as that of SqC.

4 Conclusion

In this paper, we have presented a novel framework for feature selection using one class SVM. Our proposed formulation in the dual space corresponds to the QPFS formulation. Our experiments show that solving the problem in the primal can be significantly more efficient than solving it in the dual specially when dealing with the datasets with large number of features and small number of samples, a characteristic of many problems in the biology domains.

[2]Note that both the primal and the dual will produce the same set of accuracies.

References

[1] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46(1-3):389–422, March 2002.

[2] P. Ganesh Kumar, T. Aruldoss Albert Victoire, P. Renukadevi, and D. Devaraj. Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm. *Expert Syst. Appl.*, 39(2):1811–1821, February 2012.

[3] Manik Varma and Bodla Rakesh Babu. More generality in efficient multiple kernel learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1065–1072, New York, NY, USA, 2009. ACM.

[4] Jian J. Dai, Linh Lieu, and David Rocke. Dimension reduction for classification with gene expression microarray data. *Statistical Applications in Genetics and Molecular Biology*, 5(1):Article 6, 2006.

[5] Yu Wang, Igor V. Tetko, Mark A. Hall, Eibe Frank, Axel Facius, Klaus F. X. Mayer, and Hans W. Mewes. Gene selection from microarray data for cancer classification-a machine learning approach. *Comput. Biol. Chem.*, 29(1):37–46, February 2005.

[6] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1226–1238, 2005.

[7] Irene Rodriguez-Lujan, Ramon Huerta, Charles Elkan, and Carlos Santa Cruz. Quadratic programming feature selection. *J. Mach. Learn. Res.*, 11:1491–1516, August 2010.

[8] Hao Jiang and Wai-Ki Ching. Correlation kernels for support vector machines classification with applications in cancer data. *Comp. Math. Methods in Medicine*, 2012:7, 2012.

[9] Bernhard Schölkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. *Advances in neural information processing systems*, 12(4):582–588, 2000.

[10] David M. J. Tax and Robert P. W. Duin. Support vector data description. *Mach. Learn.*, 54(1):45–66, January 2004.

[11] Heng Lian. On feature selection with principal component analysis for one-class svm. *Pattern Recogn. Lett.*, 33(9):1027–1031, July 2012.

[12] Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *J. Mach. Learn. Res.*, 6:2075–2129, December 2005.

[13] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.