Class Separability in Spaces Reduced By Feature Selection

Erinija Pranckeviciene[1], Tin Kam Ho[2], Ray Somorjai[1]
[1]Institute for Biodiagnostics, National Research Council Canada
{erinija.pranckevie, ray.somorjai}@nrc-cnrc.gc.ca
[2]Bell laboratories, Lucent Technologies
tkh@research.bell-labs.com

Abstract

We investigated the geometrical complexity of several high-dimensional, small sample classification problems and its changes due to two popular feature selection procedures, forward feature selection (FFS) and Linear Programming Support Vector Machine (LPSVM). We found that both procedures are able to transform the problems to spaces of very low dimensionality where class separability is improved over that in the original space. The study shows that geometrical complexities have good potentials for comparing different feature selection methods in aspects relevant to classification accuracy, yet independent of particular classifier choices.

1. Introduction

Practical classification problems often contain substantial geometrical structure in the class distributions that manifests as patterns learnable by human or machine classifiers. Yet with sparse samples in a high-dimensional space, such structure could be difficult to uncover due to an excessive degree of freedom in parameter choices and interferences from noninformative or redundant features. Feature selection is a popular treatment for such data that aims at enhancing the differences between classes that impact the data geometry and subsequently classification accuracy.

Feature selection has to deal with the challenges of overfitting and selection bias, both occurring because the search procedure aggressively adapts to the given limited data. Sparse data with many redundant features may also yield alternatives among several equally well-performing feature subsets to be chosen by ease of interpretation[8]. Such scenarios are often encountered in biomedical domains, where sophisticated instruments yield high-resolution measurements on very few available specimens.

Typical feature selection procedures are guided by a particular merit criterion that is related to class separability, or follow a wrapper approach to optimize performance of a particular classifier. To evaluate the final outcome, i.e., the classification problem projected to the selected subspace, claims are sometimes made on improved accuracy of a predesignated classifier, but more often on higher efficiency of automatic learning with fewer dimensions to consider. It is often unclear to what extent the feature selection transformation has changed the difficulty of the classification problem, in widening the gap between classes, compressing the discriminatory information, and removing irrelevant dimensions. In this study we propose a way to quantitatively describe such changes, using a set of data complexity measures previously shown useful for characterizing classification problems [4][5]. We intend to examine whether a feature selection procedure transforms the initial problem into spaces where classification difficulty is reduced, in the sense that it pushes the problem to a position in the complexity space farther away from a “pattern-less” random class labeling than the distance maintained by the full-dimensional data from a similar “pattern-less” labeling. Using this method, we investigate two popular feature selection procedures: Forward Feature Selection (FFS) and Linear Programming Support Vector Machine (LPSVM) [2], named LIKNON by [1]. LPSVM has yielded promising classifiers in microarray analysis [1], face recognition [3], and classification of biomedical spectra [7]. We will describe these procedures briefly and then present details and results of our proposed evaluation methodology.

2. Feature Selection

2.1. Forward Feature Selection

Forward feature selection is to incrementally add features to optimize a criterion that is either a class distance measure or the accuracy of a classifier. In our study we optimize the ratio of between and within class scatter, using an implementation from PRTools [9].

2.2. LPSVM Feature Selection

LPSVM is a variant of SVM where an L_1 norm is used in the regularization term. The dual constraints and the optimality conditions suggest how the values of the regulariza-

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

tion parameter affect the selected feature subspace. In this work the regularization parameter is used to control the size of the feature subset, as described below.

The primal minimization problem. LPSVM implements a linear rule for two-class classification: $y_i = sign(\mathbf{x}_i \mathbf{w}^T + w_0)$, where $\mathbf{x}_i = [x_i^1, \dots, x_i^d]$ are d-dimensional samples, y_i is the class label of sample i, and $N = N_1 + N_2$, the total number of samples in the two classes. The important features in LPSVM are given by the large weights w_j of the vector **w**. A component of the weight vector w_j and its absolute value are modeled through two non-negative variables: $w_j = u_j - v_j$, $|w_j| = u_j + v_j$. Define $g_i^j = y_i x_i^j$ to merge vectors from both classes. The weights **w** of the separating hyperplane are found by solving the following optimization problem [1][3]:

$$\begin{aligned} \frac{\arg\min}{(u_1,,u_d,v_1,,v_d,\xi_1,,\xi_N)} \mathbf{J}_{min} &= \textstyle\sum_{j=1}^{d}(u_j + v_j) + C \sum_{i=1}^{N} \xi_i \\ s.t. \ \textstyle\sum_{j=1}^{d} u_j g_i^j - \sum_{j=1}^{d} v_j g_i^j + u_0 - v_0 + \xi_i &\geq 1, \\ \xi_i \geq 0,\ u_j \geq 0,\ v_j &\geq 0, \\ j = 1, \dots, d,\ i = 1, \dots, N. & \qquad (1) \end{aligned}$$

The dual maximization problem. The dual of LPSVM is:

$$\begin{aligned} & \frac{\arg\max}{(z_1,\dots,z_N)} \mathbf{J}_{max} = \textstyle\sum_{i=1}^{N} z_i \\ s.t. \quad & g_1^j z_1 + \dots + g_N^j z_N \leq 1, \\ & -g_1^j z_1 - \dots - g_N^j z_N \leq 1, \\ & y_1 z_1 + \dots + y_N z_N = 0, \\ & 0 \leq z_i \leq C, \\ & i = 1, \dots, N,\ j = 1, \dots, d. \qquad (2) \end{aligned}$$

The optimal solution lies on a vertex of the feasible region given by constraints in (2) and $z_i \leq C$.

Optimality conditions of LPSVM. The optimal solutions(*) of primal and dual LPSVM satisfy the optimality conditions for every feature j and sample i:

$$\begin{gathered} u_j^* \left(\textstyle\sum_{i=1}^{N} g_i^j z_i^* - 1 \right) = 0, v_j^* \left(\textstyle\sum_{i=1}^{N} g_i^j z_i^* + 1 \right) = 0, \\ \xi_i^* \left(z_i^* - C \right) = 0, \qquad (3) \\ z_i^* \left(\textstyle\sum_{j=1}^{d} g_i^j (u_j^* - v_j^*) + y_i (u_0^* - v_0^*) + \xi_i^* - 1 \right) = 0. \end{gathered}$$

Binding constraints determine the nonzero components w_j of **w**, which correspond to the selected features. We use hyperplane H^j to denote any constraint $\pm g_1^j z_1 \pm \dots \pm g_N^j z_N = 1$, which becomes binding for feature j.

Subset selection. For some fixed C, the optimal solution satisfies $\sum_{i=1}^{N1} z_i^* x_i^{j_k} - \sum_{i=1}^{N2} z_i^* x_i^{j_k} = \pm 1$ for a set of features $\mathbf{j} = (j_1, \dots, j_m), m \leq d$ that correspond to the binding constraints. When the bounding box $z_i \leq C$ expands, the constraints H^j are encountered sequentially, depending on how far they are from the origin. At some C_{max} the bounding box will fully contain the feasible region formed by all hyper-planes H^j, including the farthest. Increasing the C beyond C_{max} likely would not alter the ***identities*** of the selected features. This analysis suggests that the initial and final values of C should be:

$$C_{min,max} = \frac{1}{\sum_{i=1}^{N} y_i x_i^{j_{max,min}}} + \varepsilon, \qquad (4)$$

where ε is a small number. The $j_{max,min}$ are the indices of the individual features corresponding to the maximum and minimum values of $y_i x_i^j$ out of all $j = 1, \dots, d$. The subspaces of the original data space are determined traversing the interval of C values $[C_{min} : C_{step} : C_{max}]$ non-uniformly, with C_{step} matching the ascending distances of H^j's from the origin. We collect the feature subset when its size changes.

3. Data Complexity In Reduced Spaces

Feature selection procedures like FFS and LPSVM are intended to transform a problem to a new, reduced space, where only important discriminatory information is retained. Ideally, this should mean an increase in class separability from that of the "mother" problem in the original space. Our study attempts to provide a *quantitative* description of these changes, using several measures of classification complexity proposed in [4] to describe each transformed problem. These measures describe the data geometry without reference to the performance of a particular classifier, yet they have been found useful in characterizing a classifier's domain of dominant competence [5]. Three of these measures are especially useful for high dimensional, sparse data:

1. `boundary` – percentage of points on class boundary, estimated by a minimum spanning tree method,
2. `intra-inter` – ratio of averaged intra-class nearest-neighbor (NN) distance to averaged inter-class NN distance,
3. `ballcenter` – percentage of points needed as centers of maximal balls to cover the class, also known as the `pretopology` measure.

These measures do not involve a linear separability assumption, are not dependent on a classifier, and are insensitive to the orientation of the class boundary w.r.t. the axes. Though, previous observations with these measures show that they may be strongly influenced by the sampling density, i.e., number of points per feature dimension. This calls for special caution in our study, because we are comparing problem formulations in spaces of drastically different dimensionalities, e.g., 1D or 2D versus 1500D, while the number of samples remains the same. An important concern is how to isolate the effects inherent to high-dimensional geometry from those due to the interleaving of the two classes.

As a way to alleviate this problem, we propose to compare the complexity values ($C(p_s)$ for measure $C(.)$, problem p, and feature subset s) computed for the transformed point set (p_s) with values computed for "pattern-less" labeling of the same point set (r_s), in the space of the same dimensionality. A large difference would mean that the transformed problem is farther away from the case where the two classes are thoroughly mixed, i.e., the more class-discriminatory information is retained. The pattern-less complexity values are obtained by randomly shuffling the class labels among the feature vectors and then applying the complexity measures, as in a *permutation test* well known in statistics. We used the averaged value ($\overline{C(r_s)}$) over 10 realizations of the random shuffling. We note that with sparse samples, one may afford to have far more realizations and even compute the exact distribution of the complexity measures for all labelings.

To show the change (or enhancement in class separability) due to the feature subset projection, we normalize the complexity difference obtained for the feature subset by the difference computed for the mother problem (p_m) and its pattern-less labeling (r_m) in the original feature space, i.e., $C(p_m) - \overline{C(r_m)}$. The differences are comparable because each complexity measure spans a known, fixed range or scale of values [4]. For each feature subset s, a ratio $t(s) = \frac{C(p_s)-\overline{C(r_s)}}{C(p_m)-\overline{C(r_m)}}$ gives the normalized deviation from pattern-less mean due to the projection of the data set to s. Systematic methods for generating s would produce a *trajectory* of $t(s)$ as a function of the controlling parameter (e.g. the subset size), showing the changes in class separability. In the experiment described below, we test the utility of this way of evaluating the FFS and LPSVM feature selection procedures.

4. Experimental Setup

We experimented with two-class problems with five datasets from real-world biomedical applications. Typical of this domain, they are sparse and high-dimensional. The public data sets Ovarian and Colon are gene expression microarrays of cancer classification problems [6]. The sets Spectra1[7] and Spectra2 (problems of discrimination of pathogenic fungi) and Spectra3 (cancerous vs. normal tissues) were magnetic resonance spectra provided by the Institute for Biodiagnostics, NRC Canada. The size of these datasets and their partition into training (Tr) and testing (Te) sets are as follows.

Property	Colon	Ovarian	Spectra1	Spectra2	Spectra3
Dim.	2000	1536	1500	1500	1500
N1+N2	40+22	30+24	104+75	141+114	77+51
Tr1+Tr2	15+15	16+16	50+50	76+76	34+34
Te1+Te2	25+7	14+8	54+25	65+38	43+17

FFS and LPSVM are applied to each of 10 random splits, and the complexity trajectory is computed for each data set, each split and each selection method. The test errors of 9 classifiers from PRTools[9] are computed for the mother problem and for every subset: nearest neighbor (1,3 5) *knnc*, linear fisher discriminant *fisherc*, decision tree *treec*, linear svm *svc*, logistic linear classifier *loglc*, nearest mean classifier *nmc* and linear discriminant using the data pca expansion *pcldc*. As it occurs, feature selection may improve or degrade classification accuracy. We attempt to understand when improvements happen by showing the changes in class separability due to the transformation.

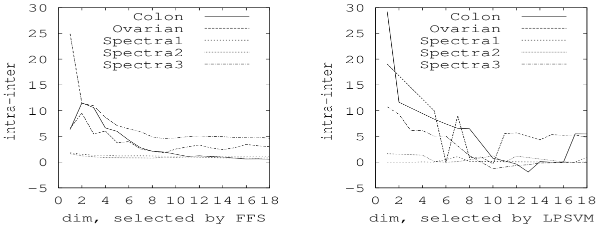

Figure 1. Trajectories of changes $t(s)$ of `intra-inter` **on the five datasets due to FFS and LPSVM.**

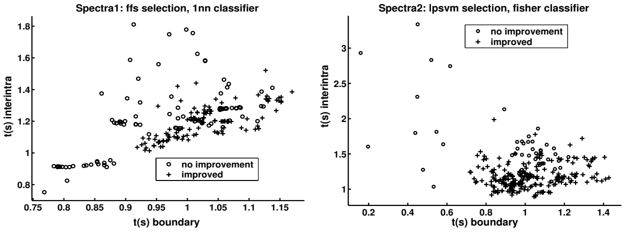

Figure 2. Comparison of complexity changes in feature subsets with or without classification accuracy improvement over mother problem.

5. Results and Observations

Figure 1 shows, as an example, trajectories of changes of the `intra-inter` measure for the five datasets using the two feature selection methods. Figure 2 shows typical distributions of complexity changes for feature subsets with or without accuracy improvement. Table 1 summarizes the results from FFS for subsets of dimensionality 1,3,...,15. The entries are the complexity values of the transformed training set in the first random split, and (after the "/") the mean and stddev of the pattern-less training set projected to the same subspace. Row "m" describes the mother problem with full dimensionality, and "s" shows the mean and stddev of the complexity values over the 10 random splits of the mother. Table 2 shows, for the Colon data, average test errors of the mother problem (over 10 random splits) with each classifier, and those of the feature subsets yielding an improvement.

For all datasets and both feature selection methods, the values of the complexity measures show some variability over the data splits. This highlights the need for further research on quantifying the uncertainty in complexity estimates due to small samples. Yet some discernible trends can be observed. These are summarized as follows.

Dim.	Colon	Ovarian	Spectra1	Spectra2	Spectra3
	boundary				
1	33.3/72.0,13.4	40.6/75.9,9.6	26.0/74.7,6.2	57.2/77.3,3.5	61.8/75.7,4.5
3	26.7/69.0,8.8	34.4/71.9,9.2	15.0/70.8,5.5	38.2/70.9,3.7	51.5/71.0,5.9
5	36.7/66.3,9.6	40.6/71.3,10.9	16.0/73.0,5.2	40.1/66.6,3.5	52.9/68.8,6.5
10	63.3/68.3,8.9	40.6/69.1,11.0	17.0/72.2,3.8	38.8/69.0,2.9	54.4/68.2,7.9
15	60.0/65.3,10.9	56.3/70.6,8.0	18.0/70.6,5.7	39.5/69.3,3.1	50.0/67.4,7.0
m	43.3/72.3,11.3	62.5/67.5,9.1	16.2/69.5,3.4	35.5/70.1,4.3	70.6/71.2,7.1
s	42.33,7.0	51.87,7.7	15.10,2.6	38.15,3.5	67.50,5.0
	intra-inter				
1	0.61/0.92,0.3	0.73/1.06,0.2	0.15/1.12,0.1	0.56/0.96,0.2	0.63/1.39,0.5
3	0.55/1.06,0.1	0.73/1.00,0.1	0.30/1.02,0.1	0.73/0.98,0.1	0.73/1.06,0.1
5	0.72/1.01,0.0	0.84/1.03,0.1	0.37/1.01,0.1	0.77/0.98,0.0	0.79/1.00,0.1
10	0.93/1.00,0.1	0.88/1.01,0.1	0.38/1.01,0.1	0.77/1.00,0.0	0.87/1.01,0.1
15	0.97/1.00,0.1	0.88/1.02,0.0	0.39/1.01,0.1	0.79/0.99,0.0	0.86/1.01,0.1
m	0.97/1.01,0.03	0.95/1.00,0.02	0.46/0.99,0.04	0.76/1.01,0.04	0.97/1.00,0.04
s	0.92,0.02	0.95,0.01	0.42,0.03	0.78,0.02	0.97,0.02
	ballcenter				
1	60.0/76.7,12.2	59.4/82.5,6.1	40.0/80.7,4.8	67.7/82.7,3.2	73.5/80.3,4.9
3	86.7/99.3,1.4	87.5/97.8,2.6	86.0/98.8,1.4	99.3/98.9,0.9	100.0/98.5,1.2
5	100.0/98.0,1.7	100.0/99.4,1.3	92.0/99.9,0.3	99.3/99.9,0.3	100.0/99.6,0.7
10	100.0/100.0,0.0	96.9/98.4,1.6	96.0/99.5,0.5	98.7/100.0,0.0	100.0/100.0,0.0
15	100.0/100.0,0.0	93.8/99.4,2.0	92.0/99.7,0.5	99.3/100.0,0.0	100.0/100.0,0.0
m	100.0/100.0,0.0	100.0/100.0,0.0	96.8/99.7,0.5	98.0/100.0,0.0	100.0/100.0,0.0
s	100.00,0.0	100.00,0.0	95.27,1.4	99.41,0.8	100.00,0.0

Table 1. Complexity values of problems transformed by forward feature selection.

Selection	knnc1	knnc3	knnc5	fisherc	treec	svc	loglc	nmc	pcldc
mother	0.262	0.250	0.232	0.154	0.333	0.157	0.157	0.175	0.161
ffs	0.218	0.193	0.197	0.118	0.235	0.120	0.120	0.114	0.148
lpsvm	0.164	0.150	0.149	0.112	0.213	0.114	0.113	0.132	0.140
% ffs	26.1	20.7	23.4	5.4	41.9	1.8	3.2	8.1	19.4
% lpsvm	86.3	89.3	96.2	18.3	77.3	26.8	7.7	74.9	44.5

Table 2. Average test errors of mother problem and improved FFS and LPSVM subsets, and percentage of subsets with reduced errors (Colon data).

1. The intrinsic difficulty of the problem is reflected in the complexity values. e.g. Spectra3, for which all classifiers are at their worst, has generally much higher boundary values than Spectra1.
2. Low dimensional (< 5) projections have significantly different complexity characteristics from the rest. With dim.> 5, complexity quickly approaches that of the mother problem.
3. Classes are severely compressed in 1-dim. projections, causing low values of ballcenter. Even the pattern-less sets show substantial compactness, which may become false classes that can be "learned" by an automatic classifier.
4. Feature selection appears to be best for the microarray data (Colon,Ovarian): for all splits, all low dim. projections have lower complexity than the mother ($t(s) > 1$), in terms of all three measures. Relative merits degrade gracefully as dimensionality increases. Feature selection is useful for MR spectra only in the very low dim. projections (≤ 3). Other than that, very often the selected feature subsets are no better than the mother ($t(s) < 1$).
5. Complexity changes due to FFS and LPSVM are similar; FFS produces monotonic trends more often.
6. Feature subsets that improve accuracy for several classifiers simultaneously almost always have high values in the ratio $t(s) > 1$ for both boundary and intra-inter. For the microarray datasets, the subsets selected by LPSVM improve accuracies more often. For Spectra1 FFS is obviously better.

Though we observe some separation (e.g. in Figure 2) of the $t(s)$ values of the feature subsets that lead to more accurate classifiers from those of the less accurate ones, a definitive and quantitative relationship between complexity changes measured in this way and accuracy improvements remains to be an open question. Future efforts towards this need to consider the small sample effects as well as the sensitivity of each type of classifiers to the specific aspect of data complexity represented by each measure.

6. Conclusions

We describe a method for using geometrical complexity measures to characterize changes in class separability due to feature selection. We report an early experiment where we applied this to evaluate two feature selection procedures, and found interesting evidences of their merits on five high-dimensional biomedical problems with extremely sparse samples. From the results, LPSVM appears to be useful for these data, and it produces accuracies comparable to FFS while being much faster.

References

[1] C. Bhattachariya, L. Grate, A. Rizki, and et al. Simultaneous relevant feature identification and classification in high-dimensional spaces: application to molecular profiling data. *Signal Processing*, 84(4):729–743, 2003.

[2] F. Glen and O. Mangasarian. A feature selection newton method for support vector machine classification. *Comp. Optim. & Appl.*, 28:185–202, 2004.

[3] G. Guo and C. Dyer. Learning from examples in the small sample case: face expression recognition. *IEEE Trans. SMC-Part B*, 35(3):477–488, June 2005.

[4] T. K. Ho and M. Basu. Complexity measures of supervised classification problems. *IEEE Trans. PAMI*, 24(3):289–300, March 2002.

[5] E. B. Mansilla and T. K. Ho. On classifier domains of competence. *Proc. of the 17th ICPR*, pages 1051–4651, August 22-25, Cambridge, UK 2004.

[6] S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T. Golub, and J. Mesirov. Estimating dataset size requirements for classifying dna microarray data. *J. of Comp. Biology*, 10(2):119–142, 2003.

[7] E. Pranckeviciene, R. Somorjai, R. Baumgartner, and M. Jeon. Identification of signatures in biomedical spectra using domain knowledge. *AI in Medicine*, 35(3):215–226, November 2005.

[8] R. Somorjai, B. Dolenko, and R. Baumgartner. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, cavets, cautions. *Bioinformatics*, 19(12):1484–1491, 2003.

[9] F. van der Heijden, R. Duin, D. de Ridder, and D. Tax. *Classification, parameter estimation and state estimation: an engineering approach using MATLAB*. John Willey and Sons Ltd., West Sussex, England, 2004.