Control of Sparseness for Feature Selection

Erinija Pranckeviciene[1,2], Richard Baumgartner[1], Ray Somorjai[1], and Christopher Bowman[1]

[1] Institute for Biodiagnostics, National Research Council Canada, 435 Ellice Avenue Winnipeg ,Canada
{Richard.Baumgartner,Ray.Somorjai, Christopher.Bowman}@nrc-cnrc.gc.ca
[2] Kaunas University of Tecnology, Studentu 50, LT 3031, Kaunas, Lithuania
Erinija.Pranckeviciene@ktu.lt

Abstract. In linear discriminant (LD) analysis high sample size/feature ratio is desirable. The linear programming procedure (LP) for LD identification handles the curse of dimensionality through simultaneous minimization of the L1 norm of the classification errors and the LD weights. The sparseness of the solution – the fraction of features retained - can be controlled by a parameter in the objective function. By qualitatively analyzing the objective function and the constraints of the problem, we show why sparseness arises. In a sparse solution, large values of the LD weight vector reveal those individual features most important for the decision boundary.

1 Introduction

In a high-dimensionality / small sample size scenario, many linear classification rules are possible. When the sample to feature ratio (SFR) is low, we face the problem of overfitting - many perfect classification rules for the training data and poor generalization on the test data. Achieving the proper ratio between number of features and available sample size is of great interest [1],[17]. Conventional dimensionality reduction techniques [2], [3] are not very useful if retaining the original feature positions is important. For high-dimensional situations, methods producing sparse solutions are in demand. Sparse means that only a few solution coefficients have large values. The linear programming (LP) technique of identifying a linear discriminant function belongs to the category of methods producing sparse solutions. Its usefulness in feature selection has been demonstrated [4]. There are case studies showing the potential of the technique in microarray analysis [5] and in face recognition [6]. This LP technique is a variant of linear support vector machine (SVM), the only difference being in the objective function. Selecting the value of a parameter in the objective function will force sparseness on the linear discriminant solutions of LP. The sparseness of the solution depends on the geometrical configuration of the data points. Although there exist studies on SVM via linear programming [8] and [7] there is a lack of systematic analysis on how the sparse solution is obtained, and what factors govern the sparseness. A deeper insight is also missing concerning the characteristics properties of the features the sparse solution identifies. Our analysis concerns the objective function of the LP formulation for linear discriminant and constraints imposed by the dataset. In

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 707–715, 2004.

the following, variables denoting vectors will be bold. We consider a 2-class classification problem. Our dataset consists of the vectors $\mathbf{x}_i \in X$ having components $\mathbf{x}_i = [x_i^1, x_i^2, ..., x_i^p]$, labeled by $y_i \in \{+1, -1\}$, where $i = 1, ..., N_1 + N_2$ are the number of samples in the classes and p is data dimensionality. Our problem is to find a linear discriminant function classifying the samples into one of the two classes ω_1, ω_2:

$$g(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i + w_0, \quad \begin{matrix} g(\mathbf{x}_i) \geq 0 \Rightarrow \mathbf{x}_i \in \omega_1 \\ g(\mathbf{x}_i) < 0 \Rightarrow \mathbf{x}_i \in \omega_2 \end{matrix}. \tag{1}$$

We may formulate this problem as a system of linear equations:

$$\mathbf{X}\mathbf{w} = \mathbf{y}. \tag{2}$$

If there are more equations than unknowns, then (2) represents a system of over determined equations. We may obtain the solution for the weight vector by least squares or by minimizing the total absolute error [10]. When there are more unknowns (features) than equations, the system (2) is underdetermined and has many solutions.

During the last decade, SVM or maximal margin classifiers were used extensively. This learning algorithm is not parametric and implements an approximation of the unknown functional relationship between training data and class label/target. The unknown discriminant weight vector is found by optimizing a functional derived in learning theory [11], [12]:

$$J(\mathbf{w}) = G * \left\| \mathbf{w} \right\|_{Lp} + C * \left\| \boldsymbol{\xi} \right\|_{Lp}, \tag{3}$$

where L_p denotes the p-norm, $\mathbf{w}$ is the vector of the weights of the linear discriminant to be found, $\boldsymbol{\xi}$ is the vector of errors between the actual output and the desired output: $\boldsymbol{\xi} = \mathbf{X}\mathbf{w} - \mathbf{y}$. The choice of norm and the values of the constants in (3) span a range of criteria for regression and classification problems [12], [13] and [14]. Different criteria implement different methods to get the solution. Some instances of the criterion function with different choices of norm and parameter values are summarized in Table 1. The criterion for LD identification by the linear programming technique producing sparse solution is:

$$J(\mathbf{w}) = \left\| \mathbf{w} \right\|_{L1} + C * \left\| \boldsymbol{\xi} \right\|_{L1}. \tag{4}$$

Two terms are minimized: the total absolute error and the sum of the components of the linear discriminant. The constraints are the same as for SVM in the linearly not separable case, and are given by (5). For some values $C > C_{max}$, we get the maximal margin classifier, identical to linear SVM. If the value of C is in the interval $0 < C < C_{max}$, then we get a sparse solution for the weight vector. For different datasets, the value of C_{max} is different. Our goal is to show how C influences the objective function and why small values of C lead to sparse solutions.

Table 1. Criteria and solution methods spanned by different norms and values of the parameters in the objective function (3). When G>0 the function (3) is studied in [16].

Norm	Values of the parameters		
	G = 1, C > C_{max}	**G= 0, C = 1**	**G = 1, 0 < C < C_{max}**
L_1	SVM, Linear programming method	Minimization of $\lvert Xw - y \rvert$. Least absolute error problem, usually solved by LP method.	Sparse solutions for linear discriminant function. Implements SVM by Linear programming method. The value of C, controlling sparseness, depends on dataset configuration. C is the upper bound on variables in the dual problem.
L_2	SVM, Quadratic programming method	Minimization of $(Xw - y)^T (Xw - y)$. Least Squares problem.	Sparse solutions for linear discriminant function. Implements SVM by Quadratic programming method. C is the upper bound of Lagrange multipliers.

2 Analysis of the Constraints and Objective Function

2.1 Formulation of the Problem for the General Linear Program Solver

The optimal solution minimizing (4) is usually obtained by using general linear program solvers [8]. SVM imposes the constraints onto the separating hyperplane. It has to be at the desired distance from the training points and have maximal margin with respect to vectors of the opposite classes [9]:

$$y_i \left(w_1 x_i^1 + ... + w_p x_i^p + w_0 \right) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1,..., N_1 + N_2 . \tag{5}$$

The inequalities in (5) define the region of feasible solutions for (4). For an arbitrary hyper plane, the actual distance of the points x_i from it is:

$$d_i = \frac{\mathbf{x}_i \mathbf{w}^T + w_0}{\|\mathbf{w}\|_2} . \tag{6}$$

The desired distance imposed by the constraints is not less than Δ:

$$\Delta = \frac{1}{\|\mathbf{w}\|_2} . \tag{7}$$

The quantities ξ_i in (5) are proportional to the differences between the desired and actual distances:

$$\frac{\xi_i}{\|\mathbf{w}\|_2} = \Delta - y_i d_i . \tag{8}$$

The value of ξ_i shows the position of the data point with respect to the separating hyperplane: $\xi_i = 0$ means that the data point is exactly at distance Δ from the sepa-

rating hyperplane, $\xi_i < 0$ means that the data point is correctly classified, $\xi_i > 0$ means that the data point is misclassified or closer to the hyper plane than Δ. Minimizing the second term in (4) means minimizing the empirical risk/classification errors. The constraint $\xi_i \geq 0$ in (5) restricts the feasible region of the weights of linear discriminant vector $\mathbf{w}$ to the region where classification errors occur or where the data point is closer than the desired distance. In order to present the minimization problem in a form suitable for a general linear program solver, the variables in the objective function should be positive. Thus, each weight component variable is modeled as a difference of two non-negative variables, as is common in linear programming [15], page 32:

$$w_j = u_j - v_j, \tag{9}$$

and the absolute value of the weight is:

$$|w_j| = u_j + v_j. \tag{10}$$

The pair u_j, v_j satisfying (9) and (10) is unique. Only three choices are possible simultaneously satisfying (9) and (10): 1) $u_j = 0 \quad v_j = 0$, 2) $u_j = 0 \quad v_j \neq 0$ and 3) $u_j \neq 0 \quad v_j = 0$. The constraints (5) now are:

$$\begin{aligned} y_1\left(x_1^1 u_1 + \ldots + x_1^p u_p - x_1^1 v_1 - \ldots - x_1^p v_p + u_0 - v_0\right) + 1*\xi_1 + \ldots + 0*\xi_N \geq 1 \\ \vdots \\ y_N\left(x_N^1 u_1 + \ldots + x_N^p u_p - x_N^1 v_1 - \ldots - x_N^p v_p + u_0 - v_0\right) + 0*\xi_1 + \ldots + 1*\xi_N \geq 1 \end{aligned}, \tag{11}$$

and

$$u_j \geq 0, v_j \geq 0, \xi_i \geq 0, \quad j = 0,\ldots,P, \quad i = 1,\ldots,N, N = N_1 + N_2. \tag{12}$$

The objective function (4) is transformed to:

$$J(\mathbf{u},\mathbf{v},\boldsymbol{\xi}) = \sum_{j=1}^{P}(u_j + v_j) + C\sum_{i=1}^{N}\xi_i. \tag{13}$$

We need to find:

$$(\mathbf{u}^*, \mathbf{v}^*, \boldsymbol{\xi}^*) = \arg\min J(\mathbf{u},\mathbf{v},\boldsymbol{\xi}). \tag{14}$$

One basic feasible solution of (13) subject to (11) and (12) is $\mathbf{u} = \mathbf{0}, \mathbf{v} = \mathbf{0}, \boldsymbol{\xi} = \mathbf{1}$ not useful for classification. In (13) the empirical risk is minimized, thus only non-negative ξ_i are considered and the modulus of ξ_i in (4) is equivalent to a positive ξ_i in (13). If the exact objective function (4) is minimized, then each ξ_i should be modeled by two positive variables as were the components of linear discriminant $\mathbf{w}$ in (9) and (10). More details on the SVM formulations with different norms can be found in [7]. The slacks in (11) are decoupled from the weights of the linear discriminant, although, strictly speaking, slacks and weights depend on each other.

2.2 What Is the Origin of the Sparseness?

The sparseness of the optimal solutions (13) subject to (11) and (12) depends on the value of C. In available SVM software, this parameter is set to the default value, or is determined by cross validation [16]. To discover the origin of sparseness, we analyze qualitatively the dependence of the shape of the objective function (4) on the parameter C. With real instances of the weight vector and a given set of data points, the slack variables equal to the deviations from the target:

$$\xi_i = 1 - y_i \left(w_1 x_i^1 + \ldots + w_p x_i^p + w_0 \right). \tag{15}$$

Substituting (15) directly into (4), we express (4) as a function:

$$J(w_1,\ldots,w_p,w_0) = \sum_{j=1}^{p} \left| w_j \right| + C \sum_{i=1}^{N_1+N_2} \left| 1 - y_i \left(w_1 x_i^1 + \ldots + w_p x_i^p + w_0 \right) \right|. \tag{16}$$

The function (16) has two parts:

$$J(\mathbf{w}, w_0) = R(\mathbf{w}) + C * A(\mathbf{w}, w_0), \tag{17}$$

R is called a regularizer and A is the empirical risk or penalty and loss in [16]. The objective function (16) is piecewise linear. Convex piecewise linear functions of the type (16) are analyzed in depth in [15]. In a constrained optimization problem, the optimal solutions for the objective function lie in the feasible region defined by the constraints. Here this region is fixed and determined by the data points as defined by (5). The objective function is controlled by the values of C leading to the different optimal solutions. When C is large, the term A dominates in the objective function. When C is small, the term R dominates in the objective function. When C is approaching zero, the objective function becomes flat and balanced. The minimum point of function (16) is forced to approach the zero origin point by small C. This narrows the set of possible optimal solutions to the points of feasible region lying near the origin. We illustrate this statement graphically by using a one-dimensional example. It is depicted in Figure 1. In higher dimensions, visualization of the concepts becomes intractable. Let the data consist of three points: (x1=0.5, y1=1), (x2= -2, y2=-1) and (x3=5, y3=-1). Let $w_0 = 0$ in the example. The linear discriminant w in the example is a scalar. The function (16) with these values is:

$$J(w) = |w| + C(|1 - 0.5w| + |1 - 2w| + |1 + 5w|). \tag{18}$$

It is a sum of convex functions and is convex itself. The coefficients 0.5, 2 and 5 can be interpreted as the influence of the data on the objective function. Higher values of the data-dependent coefficients increase the slopes of the components of the objective function and dominate the total sum. The constraints are:

$$\begin{aligned} &\xi_1 \geq 1 - 0.5w, \quad \xi_1 \geq 0, \\ &\xi_2 \geq 1 - 2w, \quad \xi_2 \geq 0, \\ &\xi_3 \geq 1 + 5w, \quad \xi_3 \geq 0. \end{aligned} \tag{19}$$

The functions e1(w)=1-0.5w, e2(w)=1-2w and e3(w)=1+5w for a given dataset represent the functional relationship (8) for all values of linear discriminant w. The interval of w values, satisfying all (19) constraints (feasible region, which does not

change) is $w \in [-0.2 \quad 0.5]$. However the shape of the objective function is determined by C. In Fig.1 we illustrate the difference of the objective functions J1(w), J2(w) and J3(w) given in (18) corresponding to different values of C: C=0.2, C=1.5 and C=5. Function J1(w) attains its minimum at the point $w = 0$ (sparse solution), which is forced by the C=0.2.

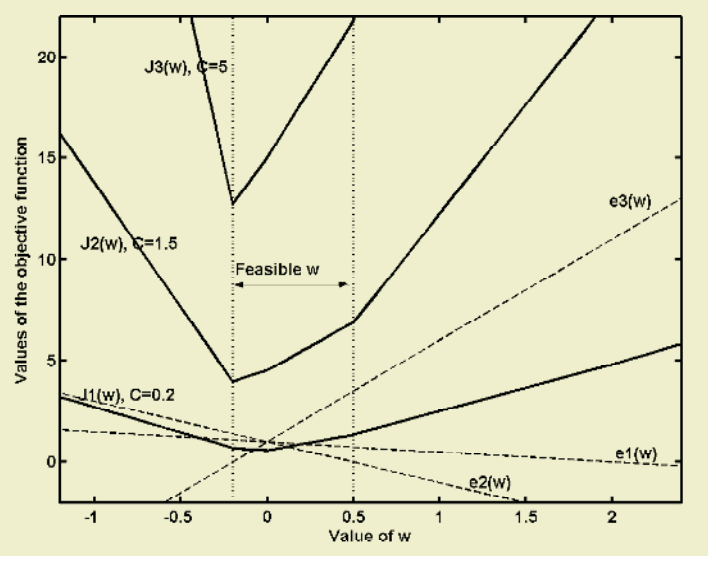

Fig. 1. The influence of the parameter C on the objective function. Solid lines represent the objective functions J1(w), J2(w) and J3(w) of (18) for different values of C: C=0.2, C=1.5 and C=5. The feasible region is shown by dotted lines. The functions e1(w), e2(w) and e3(w) are represented by dashed lines.

The simple one-dimensional example illustrates the effect of small values of C on the objective function of the form (4). In the high-dimensional case, many data points form a complicated convex surface for the feasible region. The objective function of form (4) is a superposition of hyperplanes defined by constraints plus a regularization term. When C approaches zero, the objective function is flattened. The minimum value of this function is forced to lie near the coordinate origin. Since all variables in the minimization problem (13) subject to (11) and (12) are constrained to be non-negative, the feasible region is restricted to the positive half of the high-dimensional space where minimization takes place. For C approaching zero, the optimal solutions of (13) will be at the points where the hyperplane of the objective function encounters the borders of the positive half of coordinate space. The level of sparseness depends on the dataset, determining the orientations of the constraints.

If we take the expression of ξ_i in (15) and substitute it into (13), then express the components of the weight vector $\mathbf{w}$ through $\mathbf{u}$ and $\mathbf{v}$ using (9) and (10) and rearrange the terms, we express (13) as a linear combination of the components of the vectors $\mathbf{u}$ and $\mathbf{v}$:

$$J_1(\mathbf{u},\mathbf{v},u_0,v_0)=\sum_{j=1}^{p}u_j(1-Ck_j)+\sum_{j=1}^{p}v_j(1+Ck_j)-(u_0-v_0)C\sum_{i=1}^{N}y_i+CN\,. \quad (20)$$

The term

$$k_j=\sum_{i=1}^{N}y_i x_i^j \quad (21)$$

is the data-dependent term. In expression (20) coefficients $(1\text{-}Ck_j)$, $(1+Ck_j)$ $j=1\ldots p$, represent the coordinates of the normal vector of the hyper plane of the objective function (20) which is equivalent to (13). They determine the direction in which the function decreases. For the positive coordinates of a normal vector, the decreasing direction of the objective function hyperplane is towards the origin. As C vanishes, more coefficients become positive, depending on the term (21). C should be $C<1/|k_j|$ in order to set the corresponding normal coefficient positive. If all normal coefficients are positive, the optimal minimum value of (20) is zero $\mathbf{u}=\mathbf{0},\mathbf{v}=\mathbf{0}$. The analysis of (20) reveals how sparse solutions evolve and the type of influence the data has on the solutions. Noting that the empirical mean of the observations is:

$$\hat{m}_j=\frac{1}{N}\sum_{i=1}^{N}x_i^j\,, \quad (22)$$

For (21) we have:

$$k_j=N_1\hat{m}_j^{\omega_1}-N_2\hat{m}_j^{\omega_2}\,. \quad (23)$$

The data term k_j is the difference between the weighted centroids of the two classes. (20) and (23) show that the last retained non-zero component of the sparse solution corresponds to the feature that has the largest distance between the centroids of the two classes.

3 Classification Example

We illustrate the geometrical property of the sparse solution induced by small C on a simple artificial example of linearly separable data. In order to compare with other methods, we present several decision boundaries obtained by LP with different values of C and linear SVM, least squares presented in Fig 2. Sparse solutions of the weight vector have zero components. The interpretation of zero components is that they identify unimportant features. "Unimportance" means that individual features, corresponding to zero components of the linear discriminant, do not contribute to the decision boundary. The geometrical property of **unimportant features** is that their **centroids for the two classes are closer** than those of the important features.

4 Conclusions

We presented the analysis of a particular example of the objective function used in the LP method for identification of a linear discriminant. Our analysis is qualitative, aim-

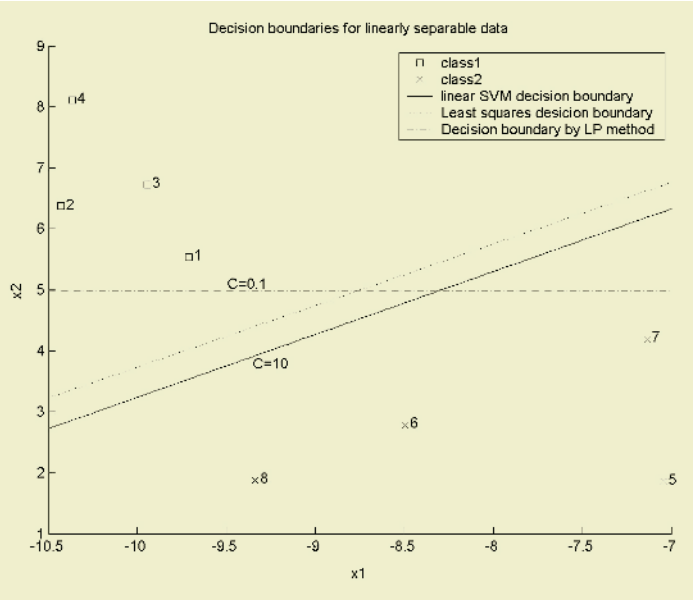

Fig. 2. Decision boundaries for the linearly separable problem for different values of C. C=0.1 gives sparse decision, where x1 is an unimportant feature. C=10 gives identical separation boundary to that of a linear SVM.

ing at a better understanding of the relationships between data, constraints and shape of the objective function. We show that we can control the sparseness of the solution by the parameter C. Small values of C induce sparseness, making the objective function flat and moving its extreme points towards zero. The solutions of the weight vector are affected by the changes in the objective function. The *practical effect* is that for individual features, with centroids for the two classes close (in the Euclidean sense), the corresponding components of the weight vector are very small. In the high dimension/small sample scenario, the method is useful for finding subsets of individual features that contribute to the class separation. However, the value of the sparseness-controlling parameter C for different sets must be identified experimentally. We are currently investigating the impact of the parameter C on the solution of the L1 norm classification problem in real life applications of high-dimensional biomedical spectra.

References

1. Raudys, S.: Statistical and neural classifiers, Springer Verlag, (2001)
2. Chen, L., Liao, H.M., Ko, M., Lin, J. and Yu, G.: A new lda-based face recognition system which can solve the small sample size problem. Pattern recognition, Vol. 33.(2000) 1713-1726

3. Howland, P., Jeon, M. and Park, H.: Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. SIAM Journal on Matrix Analysis and Applications, 25-1,(2003) 165 - 179
4. Bradley, P., Mangasarian, O. and Street, W.: Feature selection via mathematical programming. INFORMS Journal on Computing, 10(2), (1998) 209 - 217
5. Bhattacharyya, C., Grate, L.R., Rizki, A. et al. : Simultaneous relevant feature identification and classification in high-dimensional spaces: application to molecular profiling data. Signal Processing, Vol. 83, Issue 4, (2003) 729 - 743
6. Guo, G.D and Dyer, C.: Simultaneous Feature Selection and Classifier Training via Linear Programming: A Case Study for Face Expression Recognition. Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, Madison, Wisconsin, 18-20 June ,(2003) 346 - 352
7. Pedroso J.P., Murata N. Support vector machines with different norms: motivation, formulations and results, Pattern recognition letters, Vol. 12,Issue 2, (2001) 1263-1272
8. Kecman, V. and Hadzic, I.: Support vectors selection by linear programming. Proc. IJCNN 2000, Vol. 5, (2000) 193 - 198
9. Vapnik, V.: Introduction to statistical learning theory, Springer, (2001)
10. Rosen, J.B., Park, H., Glick, J. and Zhang, L.: Accurate solutions to overdetermined linear equations with errors using L1 norm minimization. Computational optimization and applications, Vol. 17, (2000) 329 - 341
11. Szeliski, R.: Regularization in neural nets. In: Mathematical perspectives on neural networks, Eds: P.Smolensky at.al, Lawrence Erlbaum Associates, (1996) 497 - 532
12. Poggio, T. and Smale, S.: The mathematics of learning:dealing with data. Notices of the American Mathematical Society(AMS), Vol. 50,No. 5, (2003) 537 - 544
13. Saunders, C., Gammerman, A.and Vovk, V.: Ridge regression learning algorithm in dual variables. Proceedings of the 15th International Conference on Machine Learning, (1998)
14. Mika, S., Ratsch, G., Weston, J., Schoelkopf, B., Smola, A. and Muller, K.R.: Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces. IEEE PAMI, Vol. 25, No. 5, (2003) 623 - 628
15. Arthanari, T.S and Dodge, Y.: Mathematical programming in statistics. John Willey and sons, (1981)
16. Hastie T., Rosset S., Tibshirani R. and Zhu J.: The entire regularization path for the support vector machine (2004)
17. Figureido M.: Adaptive sparseness for supervised learning. IEEE PAMI, Vol. 25, No. 9, (2003) 1150 - 1159