Large Margin Classifier Based on Affine Hulls

Hakan Cevikalp and Hasan Serhan Yavuz
Electrical and Electronics Engineering, Eskisehir Osmangazi University, Eskisehir, Turkey
hakan.cevikalp@gmail.com,hsyavuz@ogu.edu.tr

Abstract

This paper introduces a geometrically inspired large-margin classifier that can be a better alternative to the Support Vector Machines (SVMs) for the classification problems with limited number of training samples. In contrast to the SVM classifier, we approximate classes with affine hulls of their class samples rather than convex hulls, which may be unrealistically tight in high-dimensional spaces. To find the best separating hyperplane between any pair of classes approximated with the affine hulls, we first compute the closest points on the affine hulls and connect these two points with a line segment. The optimal separating hyperplane is chosen to be the hyperplane that is orthogonal to the line segment and bisects the line. To allow soft margin solutions, we first reduce affine hulls in order to alleviate the effects of outliers and then search for the best separating hyperplane between these reduced models. Multi-class classification problems are dealt with constructing and combining several binary classifiers as in SVM. The experiments on several databases show that the proposed method compares favorably with the SVM classifier.

1. Introduction

The Support Vector Machine (SVM) classifier is a successful binary classification method that simultaneously minimizes the empirical classification error and maximizes the geometric margin, which is defined as the distance between the separating hyperplane and closest samples from the classes [5,2]. To do so, SVM first approximates each class with a convex hull and finds the closest points in these convex hulls [1]. Then, these two points are connected with a line segment. The hyperplane, orthogonal to the line segment that bisects the line, is chosen to be the separating hyperplane [1]. From this geometrical point of view, in the separable case, the two closest points on the convex hulls determine the separating hyperplane, and the SVM margin is merely equivalent to the minimum distance between the convex hulls that represent classes. However, convex hull approximations tend to be unrealistically tight in high-dimensional spaces since the classes typically extend beyond the convex hulls of their training samples. For example, a convex hull constructed by randomly sampled points from a high-dimensional hypersphere can include only a negligible fraction of the volume of the sphere even if the chosen samples are well spaced and close to the surface of the sphere [4]. This situation may also be observed when the low-dimensional data samples are mapped to a higher-dimensional feature space through kernel mapping during estimation of the nonlinear decision boundaries between classes.

As opposed to the convex hulls, affine hulls (i.e., spanning linear subspaces that have been shifted to pass through the centroids of the classes) give rather loose approximations to the class regions, because they do not constrain the positions of the training points within the affine subspaces. Therefore, they may be better alternatives to convex hulls for some pattern classification problems especially when the data samples lie in high-dimensional spaces [3]. This paper introduces a new large margin classifier that is based on explicitly building maximum margin separators between pairs of affine hulls.

2 Method

Consider a binary classification problem with the training data given in the form $\{\mathbf{x}_i, y_i\}$, $i = 1, ..., n$, $y_i \in \{-1, +1\}$, $\mathbf{x}_i \in \mathbb{R}^d$. The proposed method begins by approximating each class (positive and negative classes) with an affine hull of its training samples. An affine hull of a class is the smallest affine subspace containing them. This is an unbounded, and hence typically rather loose model for each class. The affine hull of samples $\{\mathbf{x}_i\}_{i=1,...,n}$ contains all points of the form $\sum_{i=1}^{n} \alpha_i \mathbf{x}_i$ with $\sum_{i=1}^{n} \alpha_i = 1$. More formally affine hull of a class with samples $\{\mathbf{x}_i\}_{i=1,...,n}$ can be written as

$$H^{\mathit{aff}} = \left\{ \mathbf{x} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i \;\middle|\; \sum_{i=1}^{n} \alpha_i = 1 \right\}. \qquad (1)$$

Our goal is to find the maximum margin linear separating hyperplane between affine hulls of classes. The points $\mathbf{x}$ which lie on the separating hyperplane satisfy $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$, where $\mathbf{w}$ is the normal of the separating hyperplane, $|b|/||\mathbf{w}||$ is the perpendicular distance from the hyperplane to the origin, and $||\mathbf{w}||$ is the Euclidean norm of $\mathbf{w}$. For any separating hyperplane, all points $\mathbf{x}_i$ in the positive class satisfy $\langle \mathbf{w}, \mathbf{x}_i \rangle + b > 0$ and all points $\mathbf{x}_i$ in the negative class satisfy $\langle \mathbf{w}, \mathbf{x}_i \rangle + b < 0$ so that $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 0$ for all training data points. Finding the best separating hyperplane problem can be solved by computing the closest points on the affine hulls. The optimal separating hyperplane will be the one that bisects perpendicularly the line segment connecting the closest points. The offset (also called threshold), b, can be chosen as the distance from the origin to the point halfway between the closest points along the normal $\mathbf{w}$. Once the best separating hyperplane is determined, a new sample $\mathbf{x}$ is classified based on the sign of the decision function, $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$.

Next, we will first show how to find the best separating hyperlane for linearly separable affine hulls and then extend the idea for inseparable case. Then, we explain kernelization process.

2.1 Linearly Separable Case

Suppose that affine hulls belonging to the positive and negative classes are linearly separable. The affine hulls of two classes do not intersect, i.e., they are linearly separable, if the affine combinations of their samples satisfy the rule

$$\sum_{i:y_i=+1} \alpha_i \mathbf{x}_i \neq \sum_{j:y_j=-1} \alpha_j \mathbf{x}_j \text{for} \sum_{i:y_i=+1} \alpha_i = \sum_{j:y_j=-1} \alpha_j = 1. \qquad (2)$$

It should be noted that linear separability of data points does not necessarily guarantee the separability of corresponding affine hulls of classes. For linearly separable case, it is more convenient to write an affine hulls as

$$H^{\mathit{aff}} = \{ \mathbf{x} = \mathbf{U}\mathbf{v} + \mu \,|\, \mathbf{v} \in \mathbb{R}^l \}, \qquad (3)$$

where $\boldsymbol{\mu} = (1/n) \sum_i \mathbf{x}_i$ is the mean of the samples (or any other reference point in the hull) and $\mathbf{U}$ is an orthonormal basis for the directions spanned by the affine subspace. The vector $\mathbf{v}$ contains the reduced coordinates of the point within the subspace, expressed with respect to the basis $\mathbf{U}$. Numerically, $\mathbf{U}$ can be found as the U-matrix of the 'thin' Singular Value Decomposition (SVD) of $[\mathbf{x}_1 - \boldsymbol{\mu}, ..., \mathbf{x}_n - \boldsymbol{\mu}]$. Here, 'thin' indicates that we take only the columns of U corresponding to "significantly non-zero" singular values λ_k; l is the number of such non-zero singular values. This subspace estimation process is essentially orthogonal least squares fitting. Discarding near-zero singular values corresponds to discarding directions that appear to be predominantly "noise".

Now suppose that we have two affine hulls with point sets $\{\mathbf{U}_+ \mathbf{v}_+ + \boldsymbol{\mu}_+\}$ and $\{\mathbf{U}_- \mathbf{v}_- + \boldsymbol{\mu}_-\}$. (They may have different numbers of dimensions l). A closest pair of points between the two hulls can be found by solving

$$\min_{\mathbf{v}_+, \mathbf{v}_-} ||(\mathbf{U}_+ \mathbf{v}_+ + \boldsymbol{\mu}_+) - (\mathbf{U}_- \mathbf{v}_- + \boldsymbol{\mu}_-)||^2. \qquad (4)$$

Defining $\mathbf{U} \equiv \begin{pmatrix} \mathbf{U}_+ & -\mathbf{U}_- \end{pmatrix}$ and $\mathbf{v} \equiv \begin{pmatrix} \mathbf{v}_+ \\ \mathbf{v}_- \end{pmatrix}$, this can be written as the standard least squares problem

$$\min_{\mathbf{v}} ||\mathbf{U}\,\mathbf{v} - (\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)||^2 \qquad (5)$$

whose solution is $\mathbf{v} = (\mathbf{U}^\top \mathbf{U})^{-1} \mathbf{U}^\top (\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)$. Taking the decision boundary $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$,

$$\mathbf{w} = \frac{1}{2}(\mathbf{x}_+ - \mathbf{x}_-) = \frac{1}{2}(\mathbf{I} - \mathbf{P})(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-) \qquad (6)$$

where $\mathbf{P} = \mathbf{U}\,(\mathbf{U}^\top \mathbf{U})^{-1} \mathbf{U}^\top$ is the orthogonal projection onto the joint span of the directions contained in the two subspaces, $\mathbf{I} - \mathbf{P}$ is the corresponding projection onto the orthogonal complement of this span[1], and $\mathbf{x}_+$ and $\mathbf{x}_-$ denote the closest points on the affine hulls of positive and negative classes, respectively. Note that $\mathbf{w}$ lies along the line segment joining the two closest points and it is half the line segment's size. The offset b of the separating hyperplane is given by

$$b = -\mathbf{w}^\top (\mathbf{x}_+ + \mathbf{x}_-)/2. \qquad (7)$$

2.2 Inseparable Case

A problem arises if the affine hulls of classes intersect, i.e., affine hulls are not linearly separable. If the affine hulls of classes are close to being linearly separable and they overlap because of a few outliers, we can restrict the influence of outlying points by reducing affine hulls. Note that ignoring directions corresponding to the overly small singular values during affine

[1] If the two subspaces share common directions, $\mathbf{U}^\top \mathbf{U}$ is not invertible and the solution for $(\mathbf{v}_+, \mathbf{v}_-)$ and $(\mathbf{x}_+, \mathbf{x}_-)$ is non-unique, but the orthogonal complement remains well defined, giving a unique minimum norm separator $\mathbf{w}$. Numerically all cases can be handled by finding $\tilde{\mathbf{U}}$, the U matrix of the thin SVD of $\mathbf{U}$, and taking $\mathbf{P} = \tilde{\mathbf{U}}^\top \tilde{\mathbf{U}}$.

hulls constructions reduces the effects of noise and outliers up to the some point. But, we will use a different approach here in order to cope with the outliers. To this end, we use the initial affine hull formulation (1) and introduce upper and lower bounds on coefficients α_i to reduce affine hulls inspired by the idea that is introduced to reduce convex hulls in [1]. It should be noted that the reduced affine hulls are not simply uniformly scaled versions of the initial complete affine hulls. One may go further and choose different lower and uper bounds, or define a different interval for every sample in the training set if a-priori information is available. For instance, if the lower bound is set to zero, then the method will be equivalent to the SVM classifier. Finding the closest points on the reduced affine hulls can be written as a quadratic optimization problem

$$\begin{aligned}&\min_{\alpha} \; || \sum_{i:y_i=+1} \alpha_i \mathbf{x}_i - \sum_{i:y_i=-1} \alpha_i \mathbf{x}_i ||^2 \\ &s.t. \;\; \sum_{i:y_i=+1} \alpha_i = 1, \;\; \sum_{i:y_i=-1} \alpha_i = 1, \;\; -\tau \leq \alpha_i \leq \tau,\end{aligned} \tag{8}$$

where τ is the user-chosen bound. This optimization problem (8) can be written in a more compact form as

$$\begin{aligned}&\min_{\alpha} \; \sum_{ij} \alpha_i \alpha_j y_i y_j \left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle \\ &s.t. \;\; \sum_{i} \alpha_i y_i = 0, \;\; \sum_{i} \alpha_i = 2, \;\; -\tau \leq \alpha_i \leq \tau,\end{aligned} \tag{9}$$

where $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ denotes the inner product of $\mathbf{x}_i$ and $\mathbf{x}_j$. This is a quadratic programming problem that can be solved using standard optimization techniques. Note that the Hessian matrix, $\mathbf{G} = [G_{ij}] = y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$, is a positive semi-definite matrix, thus the objective function is convex and a global minimum exists as in SVM classifier. Moreover, if the Hessian matrix is strictly positive definite, the solution is unique and it is guaranteed to be the global minimum.

Since the coefficients are bounded between $-\tau$ and $+\tau$, the solution is determined by more points and no extreme point or noisy point can excessively influence the solution for well-chosen τ. Once we compute the optimal values of coefficients α_i, the normal and the offset of the separating hyperplane can be computed as in the linearly separable case

$$\mathbf{w} \;=\; \frac{1}{2}(\sum_{i:y_i=+1} \alpha_i \mathbf{x}_i - \sum_{i:y_i=-1} \alpha_i \mathbf{x}_i), \tag{10}$$

$$b \;=\; -\frac{1}{2} \mathbf{w}^{\top} (\sum_{i:y_i=+1} \alpha_i \mathbf{x}_i + \sum_{i:y_i=-1} \alpha_i \mathbf{x}_i), \tag{11}$$

We call this method Large Margin Classifier of Affine Hulls (LMC-AH) since it uses affine hulls to approximate class regions and finds the optimal separating hyperplane yielding the largest margin between the affine hulls.

If the underlying geometry of the classes is highly complex and nonlinear, and approximating classes with linear affine hulls is not appropriate, we can map the data into a higher-dimensional space where the classes can be approximated with linear affine hulls. Note that the objective function of (9) is written in terms of the dot products of samples, which allows the use of the kernel trick. Thus, by using kernel trick, – i.e., replacing $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ with the kernel function $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ where $\phi : \mathbb{R}^d \rightarrow \Im$ is the mapping function from the input space to the mapped space $\Im$ – we can find the best seperating hyperplane parameters in the mapped space. As a result, more complex nonlinear decision boundaries between classes can be approximated by using this trick.

3 Experiments

We tested the linear and kernelized versions of the proposed methods on a number of data sets and compared them to the SVM classifier. For the linearly separable case, linear separator is determined by using affine subspace estimation formulation, and subspace dimensions are set by retaining enough leading eigenvectors to account for 95-98% of the total energy in the eigen-decomposition. For the inseparable and nonlinear cases, we used quadratic programming formulations. Both one-against-rest (OAR) and one-against-one (OAO) approaches [6] are used for multi-class classification problems, and we report the results for the one which performs the best. We first tested the linear LMC-AH method on a face recognition problem to demonstrate that affine hull approximations are more appropriate than convex hull approximations when the dimensionality of the input space is high. To assess the generalization performances of kernelized versions of the methods, we tested them on seven low-dimensional databases chosen from the UCI repository.

3.1 Experiments on the AR Face Database

The AR Face data set [7] contains 26 frontal images with different facial expressions, illumination conditions and occlusions for each of 126 subjects, recorded in two 13-image sessions spaced by 14 days. For this experiment, we randomly selected 20 male and 20 female subjects. The images were down-scaled (from 768×576), aligned so that centers of the two eyes fell at fixed coordinates, then cropped to size 105×78. Raw

Table 1. Classification Rates (%) on the AR Face Database.

Methods	$n = 7$	$n = 13$	$n = 20$
LMC-AH	**95.19**±0.6	**98.95**±0.3	**99.62**±0.3
SVM	94.54±0.6	98.66±0.2	99.58±0.3

pixel values were used as features. For training we randomly selected $n = 7, 13, 20$ samples for each individual, keeping the remaining $26 - n$ for testing. This process was repeated 10 times, with the final classification rates being obtained by averaging the 10 results.

The results are shown in Table 1. Best results are obtained by OAR strategy for both tested methods. The proposed method gives better classification rates than soft-margin linear SVM classifier in all cases. The performance difference is more apparent for $n = 7$. These results support our claims, suggesting that affine hulls can be better models for representing classes in high-dimensional spaces when the number of samples is limited.

3.2 Experiments on the UCI Databases

In this group of experiments, we tested the kernelized versions of the methods (quadratic programming formulations) on seven lower-dimensional datasets from the UCI repository (http://www.ics.uci.edu/∼mlearn/MLRepository.html): Ionosphere, Iris, Letter Recognition (LR), Multiple Features (MF) - pixel averages, Pima Indian Diabetes (PID), Wine, and Wisconsin Diagnostic Breast Cancer (WDBC). We used the Gaussian kernels, and all design parameters are set based on random partitions of datasets into training and test sets. OAO strategy was used for multi-class problems. Reported classification rates given in Table 2 are computed by 5-fold cross-validation. Although being quite mixed, results indicate that generalization performance of LMC-AH compares favorably with SVM classifier.

Table 2. Classification Rates (%) on the UCI Datasets.

UCI	LMC-AH	SVM
Ionosphere	**93.7**±2.9	92.9±3.2
Iris	94.7±2.9	**95.3**±3.8
LR	**99.98**±0.02	99.64±0.12
MF	**98.4**±0.4	98.0±0.4
PID	**99.9**±0.3	**99.9**±0.3
Wine	**98.8**±1.6	98.2±1.6
WDBC	96.0±2.5	**97.6**±0.7

4 Summary and Conclusion

We investigated the idea of basing large margin classifiers on affine hulls of classes as an alternative to the SVM (convex hull large margin classifier). Given two affine hull models, their corresponding large margin classifier is easily determined by finding a closest pair of points on these two models and bisecting the displacement between them. The experimental results provided useful insights on the potential application areas of the proposed method. The proposed method is much more efficient than SVM classifier in terms of classification accuracy and real-time performance (testing time) when the dimensionality of the sample space is high and affine hulls are linearly separable (in this case solution is easily determined based on subspace estimation which requires simple linear algebra whereas SVM formulation requires solving a quadratic programming). For the low-dimensional databases generalization performance of the proposed method compares favorably with SVM classifier but SVM is more efficient in terms of testing time. This is because of the fact that all training data points contribute to the affine hull models (almost all computed α_i coefficients are nonzero), thus the proposed quadratic optimization solutions lack sparseness, and we need more computations to evaluate decision functions. Nevertheless, some pruning techniques can be employed to overcome this problem.

References

[1] K. P. Bennett and E. J. Bredensteiner. Duality and geometry in svm classifiers. In *International Conference on Machine Learning*, 2000.

[2] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[3] H. Cevikalp and B. Triggs. Large margin classifiers based on convex class models. In *International Conference on Computer Vision Workshops*, 2009.

[4] H. Cevikalp, B. Triggs, and R. Polikar. Nearest hyperdisk methods for high-dimensional classification. In *International Conference on Machine Learning*, 2008.

[5] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[6] C. Hsu and C. Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13:415–425, 2002.

[7] A. M. Martinez and R. Benavente. The AR face database. Technical report, Computer Vision Center, Barcelona, Spain, 1998.