

Classification by Reflective Convex Hulls

Mineichi Kudo Atsuyoshi Nakamura
Division of Computer Science
Graduate School of Information Sci. and Tech.
Hokkaido University, Sapporo 060-0814, JAPAN
E-mail: {mine,atsu}@main.ist.hokudai.ac.jp

Ichigaku Takigawa
Institute for Chemical Research
Bioinformatics Center
Kyoto University
E-mail: takigawa@kuicr.kyoto-u.ac.jp

Abstract

A set of convex bodies including samples of a single class only is used for classification. The convex body is defined by some facets (hyper-planes) that separate the class from the other classes. This paper describes an algorithm to find a set of such convex bodies efficiently and examine the performance of a classifier using them. The relationship to the support vector machines is also discussed.

1. Introduction

The convex hull $\text{conv}(S)$ of a finite set S in m -dimensional Euclidean space is one of central concepts in computational geometry. In pattern recognition, the convex hulls which cover all the training samples of one class allows us to measure the separability among classes. Indeed, the relationship between those convex hulls and support vector machines (SVMs) have been well studied [2, 5, 8]. A typical view of such trials is that the hyper-plane of an SVM is identical to the bisector hyper-plane between the closest points of convex hulls of two classes [8].

When we use convex hulls for classification, the following problems arise: 1) The convex hull of a finite set is hard to be constructed in high dimensions, 2) It costs much to calculate the distance between a point and the convex hull, and 3) In general, we need more than one convex hull for approximating a class region. For 1), there is no efficient algorithm to find the convex hull explicitly in high dimensions. Indeed, the number of facets is often exponential in m . For 2), the problem to calculate the distance $D(x, \text{conv}(S))$ for $x \in \text{conv}(S)$ is known to be NP-hard in the representation size of $\text{conv}(S)$ [7]. For 3), we need more than one convex hull to exclude samples from the classes other than a target class. The authors have already proposed such an

approach using quasi convex hulls with restricted angles [6].

In this paper, to cope with these three problems, we use several randomize techniques. We would obtain efficiency at the expense of loosing the perfection to some extent.

2. Convex Hulls and Support Functions

The simplest definition of the convex hull $\text{conv}(S)$ of a given dataset S , is the intersection of all convex sets containing S . For a finite set S , $C = \text{conv}(S)$ is a polyhedron with at most $|S|$ vertices. Such a polyhedron can be defined in several ways. By ∂C , we denote the boundary of C and divide it into q -faces according to the dimensions. For example, 0-faces are the vertices of C and $(m-1)$ -faces are the facets or hyper-planes. Let $V(C)$ be the set of vertices of C and $F(C)$ be the set of facets of C . The second definition is called \mathcal{V} -representation and is defined as $C = \{y = \sum c_x x \mid \sum c_x = 1, c_x \geq 0, x \in V(C)\}$. The third one is called \mathcal{H} -representation and is defined as $C = \{y \mid \langle w, y \rangle \leq c, \forall (w, c) \in F(C)\}$, where $\langle \cdot, \cdot \rangle$ is the inner product and a facet (w, c) is specified by a normal vector w ($\|w\| = 1$) and a constant $c \in R$.

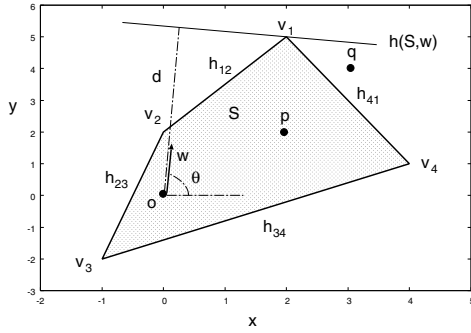
In this paper, as the fourth definition, we use *support functions* to express a convex hull C . A support function with a unit vector w ($\|w\| = 1$) is given by

$$H(S, w) = \sup\{\langle x, w \rangle \mid x \in S\},$$

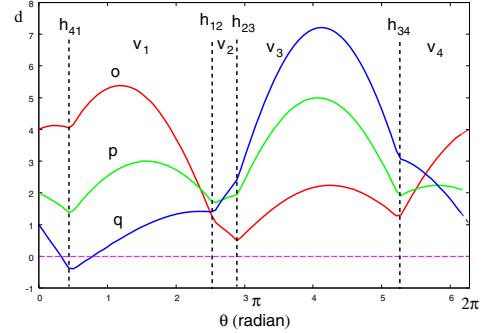
where \sup denotes the supremum. With all possible directions w , we can specify C as

$$C = \bigcap_{w: \|w\|=1} \{x \mid \langle x, w \rangle \leq H(S, w)\}.$$

Of course, it is sufficient to use w of $(w, c) \in F(C)$ instead of all possible w 's. To enhance the role we call a plane $h(S, w) = \{x \mid \langle x, w \rangle = H(S, w)\}$ a *support*



(a) The convex hull



(b) The support functions at three points

Figure 1. A convex hull and the corresponding support functions at three points.

plane. Such a support function representation has been well studied in [4].

In support functions, we may take any point as the starting point of w instead of the origin. As seen in Fig. 1, a vertex is specified by some range of angles of w regardless the starting point of w . Indeed, when we use a point q as the starting point, the support functions change to

$$H(S - \{q\}, w) = \sup\{\langle x - q, w \rangle, x \in S\},$$

but C , $V(C)$ and $F(C)$ are unchanged.

To find all vertices $V(C)$, we do not need to examine all possible w 's. In this paper, we will find $V(C)$ with a small subset of w 's. When a directional vector w is chosen randomly according to a uniform distribution, the probability of a vertex to be found is proportional to the range of angles that specify the vertex (Fig. 1). Note that a vertex with a narrower angle is easier to be found. In this respect, we can expect a good approximation of C even with a small set of randomly chosen w 's. For example, in Fig. 1, vertex v_2 may be missed if we use 5-10 randomly chosen w 's. Even so, $\text{conv}(\{v_1, v_3, v_4\})$ is a good approximation of C .

3. Reflective Convex Hulls

A convex hull $\text{conv}(S)$ is for a single finite set S . Here we consider two finite sets S and T . Then we can consider another convex body for S . The *reflective convex hull* $\text{conv}_r(S)$ of S against T is a convex body specified by *reflective support functions* which are support functions perfectly or partly separating S from T (Fig. 2). When S and T are both in the same side of a closed half-space specified by a support plane, the plane is useless for distinguishing them. Therefore, we only use useful reflective support functions. From the definition, $\text{conv}(S) \subseteq \text{conv}_r(S)$.

To measure the degree of separation between S and T , we introduce a dual function of $H(S, w)$ as

$$G(S, w) = \inf\{\langle x, w \rangle, x \in S\},$$

where \inf is the infimum. The duality is shown as $G(S, w) = H(S, -w)$.

Now we can define the *margin* $M(S, T, w)$ between S and T in direction w as

$$M(S, T, w) = G(T, w) - H(S, w).$$

Note that when S and T are linearly separable, then there exist support planes with a positive margin in both $\text{conv}(S)$ and $\text{conv}(T)$. A reflective support plane with w can be formally defined as a support plane satisfying

$$H(T, w) - H(S, w) > 0.$$

The distance between any point x and a convex hull $\text{conv}(S)$ is calculated by

$$D(x, \text{conv}(S)) = \sup\{M(S, \{x\}, w)\}.$$

The distance takes a negative value when $x \in \text{conv}(S) - \partial\text{conv}(S)$. In this case, the problem to calculate distance $D(x, \text{conv}(S))$ is generally known to be NP-hard in the representation size of $\text{conv}(S)$. However, using a finite set P of w 's, we can calculate $D(x, \text{conv}(S))$ approximately in a linear order of $|P|$.

4. Algorithm

The algorithm to find $\text{conv}_r(S)$ approximately is quite simple. We choose randomly some pairs (x, y) in which x is taken from a target class and y from the other classes. Put them into a pool $P = \{w = y - x / \|y - x\|\}$ as a set of unit directional vectors. Here, we measure the inner product of w and any point z by $\langle z - x, w \rangle$.

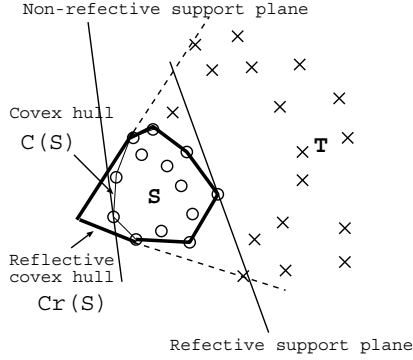


Figure 2. A Reflective Convex Hull.

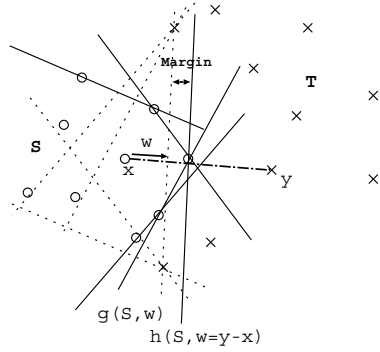


Figure 3. Approximated reflective convex hull by support planes. Only a few support planes are shown.

For obtaining an approximation $\text{conv}_r(S, P)$ of $\text{conv}_r(S)$ for S (the training sample set of the target class), we collect the reflective support planes for all $w \in P$ (Fig. 3). In Fig. 3, the margin in direction w is negative. The number of reflective support planes is $|P|$, although most of them may be connected to a few vertices.

To determine a class region by convex hulls, only one reflective convex hull is sufficient when $\text{conv}_r(S, P) \cap T = \emptyset$. Otherwise, we need a collection $\mathcal{C} = \{\text{conv}_r(U, P) \mid \text{conv}_r(U, P) \cap T = \emptyset, U \subseteq S\}$, where U is taken as a maximal subset with such property. For find \mathcal{C} efficiently, we use a randomized procedure as seen in a *randomized subclass method* [6].

The algorithm is summarized as follows:

Algorithm for finding $\text{conv}_r(S, P)$

1. Let S be the positive sample set of a target class and T be the negative sample set of the other

classes. Let $\mathcal{C} = \emptyset$.

2. For a predetermined number K , find randomly K vectors $w = \frac{y-x}{\|y-x\|}$, where $x \in S, y \in T$, to have a pool $P = \{w\}$ of directional vectors.
3. Repeat L times following Steps 4-5.
4. Let $U = \emptyset$. According to a random presentation order of positive samples, add a positive sample x to U as long as $\text{conv}_r(U \cup \{x\}, P) \cap T = \emptyset$.
5. Add the obtained $\text{conv}_r(U, P)$ into \mathcal{C} .
6. Select a minimal subset of \mathcal{C} in a greedy set cover procedure for positive samples.

In the random choice of the presentation order of positive samples, by weighting, we raise the selection probability of the positive samples that have not been used for any U so far. We carry out this procedure twice for all classes: the first one is for finding irregular samples that are included in only small convex hulls, and the second one is for obtaining the final convex hulls after removal of irregular samples. To judge the *irregularity* of samples, we use a threshold θ as the ratio of the number of samples included in a convex hull to the number of positive samples.

Note that in Step 4, it is guaranteed that we can have one reflective convex hull of some maximal subset U of S . This is because $\text{conv}_r(C', P) \subseteq \text{conv}_r(C, P)$ holds for any $C' \subseteq C$ and every positive sample is necessarily scanned in each iteration. Of course, this algorithm does not always guarantee the perfection in several aspects because of the randomness. Found family $\tilde{\mathcal{C}}$ is a subset of \mathcal{C} , each $\tilde{C} \in \tilde{\mathcal{C}}$ may lose a few vertices, and $\tilde{\mathcal{C}}$ is approximated by a finite subset of support functions. Nevertheless, we can expect the found family $\tilde{\mathcal{C}}$ is sufficiently close to \mathcal{C} even for a moderate pool size $|P| = K$.

The complexity of this algorithm is $O(L|P|n^2m)$ for n samples in m -dimensional space. Note that we can use a subset $T' = \text{conv}_r(S, P) \cap T$ instead of T in such a way that $\text{conv}_r(U \cup \{x\}, P) \cap T' = \emptyset$ is used in Step 4. The complexity for $D(x, \text{conv}_r(U, P))$ is $O(|P|m)$.

5. Experiments

A synthetic data of two classes in 2-dimensional space was firstly examined. The results are shown in Fig. 4. We used $K = 100$ for P (10 from one class and 10 from the opposite class). The number L of iterations was set to $L = 100$. We also set the value of θ to $\theta = 0.01\%$ for irregular sample judgement. As a competitor, we used an SVM with RBF kernel in which

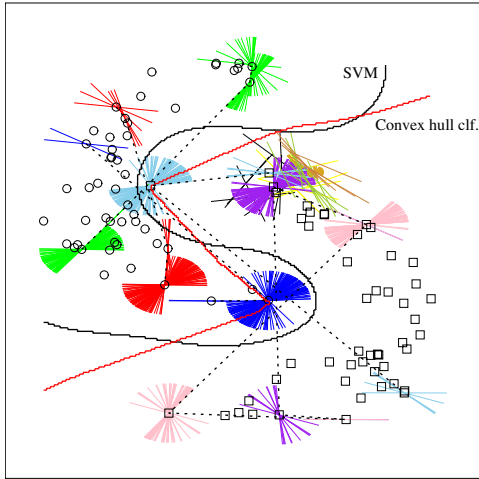


Figure 4. Obtained reflective convex hulls and the decision boundaries of SVM and the proposed classifier. A reflective convex hull is represented by dotted lines and a vertex is by a set of support planes.

the default values of parameters (the standard deviation is $\sigma = 10.0$ and the soft margin parameter $\gamma = 100.0$) were used [3].

Next, we used 17 databases from UCI machine learning repository [1]. To estimate the recognition rates, we used 10-fold cross validation. For P , we chose 10 samples from each class, thus $K = 10 \times 10 \cdot (\# \text{ of classes} - 1)$ and $L = 20$. These results show that our convex-hull-based classifier is comparable to SVM, although SVM was not tuned fully.

6. Relationship to SVM

The proposed classifier using reflective convex hulls is strongly connected to linear SVMs. Indeed, the proposed classifier achieves the largest margin locally. If two sets S and T are linearly separable, that is, $\text{conv}(S) \cap \text{conv}(T) = \emptyset$, support planes of both classes with the largest margin are parallel to the hyper-plane of SVM.

7. Conclusion

We have shown an algorithm to find a family of convex hulls for maximal subsets of training samples. It is a randomized algorithm so that the result is not precise to some extent, but the family is sufficient to find a

Table 1. Recognition rate on 17 datasets.

Database	SVM with RBF	Convex Clf.
balance-scale	93.2	92.1
diabetes	64.1	72.3
ecoli	79.8	84.2
glass	66.3	69.7
heart-statlog	59.3	68.9
ionosphere	94.0	86.9
iris	98.0	96.0
liver-disorders	63.2	63.4
optdigits	96.4	95.2
pendigits	77.2	94.8
page-blocks	92.3	92.9
segment	92.6	92.5
sonar	77.4	77.9
spam-base	86.3	70.8
vehicle	56.0	62.3
waveform-5000	83.7	85.1
wine	72.5	76.5

class region being separated from the other classes. It includes a polynomial-time procedure to find a convex hull approximately even in high dimensions for which no polynomial-time algorithm to find the exact convex hull is known.

References

- [1] A. Asuncion and D. Newman. UCI machine learning repository, 2007.
- [2] K. P. Bennett and E. J. Bredehsteiner. Duality and geometry in SVM classifiers. In *Proc. 17th International Conf. on Machine Learning*, pages 57–64. Morgan Kaufmann, San Francisco, CA, 2000.
- [3] R. Collobert and S. Bengio. Svmtorch: Support vector machines for large-scale regression problems [<http://www.idiap.ch/learning/svmtorch.html>]. *Journal of Machine Learning Research*, 1:143–160, 2001.
- [4] P. K. Ghosh and K. V. Kumar. Support function representation of convex bodies, its application in geometric computing, and some related representations. *Computer Vision and Image Understanding*, 72(3):379–403, 1998.
- [5] K. Ikeda and T. Aoshi. An asymptotic statistical analysis of support vector machines with soft margins. *Neural Networks*, 18(3):251–259, 2005.
- [6] M. Kudo, Y. Torii, Y. Mori, and M. Shimbo. Approximation of class regions by quasi convex hulls. *Pattern Recognition Letters*, 19:777–786, 1998.
- [7] O. L. Mangasarian. Polyhedral boundary projection. *SIAM J. on Optimization*, 9(4):1128–1134, 1999.
- [8] D. Zhou, B. Xiao, and H. Zhou. Global geometry of SVM classifiers. Technical report, AI Lab, Institute of Automation, Chinese Academy of Sciences, 2002.