The Convex Subclass Method: Combinatorial Classifier Based on a Family of Convex Sets

Ichigaku Takigawa[1], Mineichi Kudo[2] and Atsuyoshi Nakamura[2]

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
takigawa@kuicr.kyoto-u.ac.jp
[2] Graduate School of Information Science and Technology, Hokkaido University, Kita 13, Nishi 8, Kita-ku, Sapporo 060-8014, Japan
{mine, atsu}@main.ist.hokudai.ac.jp

Abstract. We propose a new nonparametric classification framework for numerical patterns, which can also be exploitable for exploratory data analysis. The key idea is approximating each class region by a family of convex geometric sets which can cover samples of the target class without containing any samples of other classes. According to this framework, we consider a combinatorial classifier based on a family of spheres, each of which is the minimum covering sphere for a subset of positive samples and does not contain any negative samples. We also present a polynomial-time exact algorithm and an incremental randomized algorithm to compute it. In addition, we discuss the soft-classification version and evaluate these algorithms by some numerical experiments.

1 Introduction

The goal of pattern classification is, given a training set as examples, to develop a *classifier* which can assign the class label to any possible patterns in the feature space and minimizes the probability of error[1–3]. We consider the classification on the feature space $\mathbb{R}^d$ such that all patterns are described as d numerical measurements (features). Thus, an m-class classification involves partitioning the feature space into m disjoint regions corresponding to each class. Such regions should consist of points which are likely to belonging to that class.

In pattern classification, we can use only a *finite* training set although the background probability distribution is often unknown. Moreover, if we assume the i.i.d. property behind data, the training patterns must be very carefully labeled as example patterns, thus it is often a heavy task and requires high cost to obtain a good training set in general. Consequently, the size of training set is often too small to obtain enough result by classic statistical methods.

Hence we focus attention on the nonparametric classification framework motivated by Vapnik's principle[4] "When solving a given problem using a restricted amount of information, try to avoid solving a more general problem as an intermediate step." Even for given training samples, we will need a nonlinear discrimination in general. Thus, we introduce a decomposition of such a complicated discriminative structure of data into smaller, easy-to-handle convex pieces.

This paper proposes such a framework as a generalization of the subclass method based on rectangles by Kudo *et al.*[5]. According to that, we develop the new combinatorial classifier based on spheres.

2 The Convex Subclass Method

2.1 General Methodology

We focus attention on a geometric intuition for problems such as how data are and how the classification will be done. Any learning algorithms encode some a priori knowledge on the given problem. Actually, many conventional classifier uses, explicitly or implicitly, some kind of computational geometric structures to classify incoming patterns. For examples, SVM uses a hyperplane (or a halfspace) and Nearest neighbor method uses a Voronoi diagram. Using a hyperplane is the simplest way to distinguish two classes, but it is still unsure whether it fits tasks for more than 3 classes or not.

In our approach, we consider representing the dispersion of each class data against other classes by covering all samples of each class with some simple convex sets (such as boxes, balls, ellipsoids, halfspaces, convex hulls, or cylinders) which does not contain any samples of other classes. Each convex set $R(Z)$ is defined by a certain subset Z of positive samples[3] (Figure 1). When given such convex sets for each class, we can assign the class label to every point $x \in \mathbb{R}^d$, based on the minimum distance between the convex sets for each class and the point x.

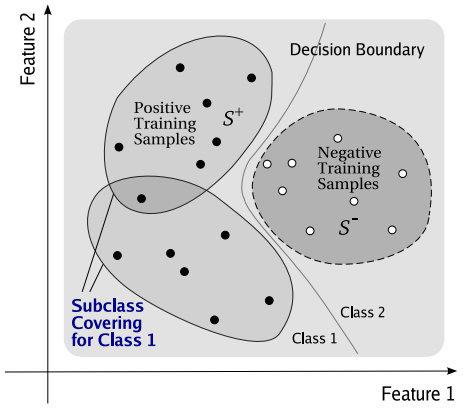

Fig. 1. The idea of the convex subclass method.

[3] Besides Z, we can also use all negative samples in order to define a consistent convex set $R(Z)$, but we do not mention it here.

Convex subclasses are a family of subsets of positive samples (it forms a hypergraph) constrained by all negative samples and the type of used convex set. This idea is motivated by Kudo's subclass method [5, 6] which uses the minimum bounding box (i.e. axis-parallel rectangle) containing the subset Z as the corresponding convex set $R(Z)$. Rectangles are, however, sometimes not suitable for given classification problems because they depend on the choice of the coordinate systems, there may exist too long and thin boxes, or the resultant decision boundary is not smooth enough.

Thus, in this paper, we extend the original subclass method *et al.*[5] to more general framework and develop the method based on spheres according to it. Our method can give a combinatorial classifier which can be exploitable for exploratory data analysis of given classification problem, such as examining the difficulty (or complexity) of problem or the effectiveness of used features. Moreover, this framework can introduce relaxation for the exclusion of negative samples, and also have a potentiality for realizing a parallel computable classifier.

2.2 Subclass Covering for Target Class

Now, we can give more formal definition of our framework. Given two finite point sets $S^+, S^- \subset \mathbb{R}^d$ as a positive set and a negative set for the target class, respectively. In this paper, although we regard $R(Z)$ as the smallest enclosing ball of a point set Z basically, any computable convex set will be available if it can be defined by only $Z \subset S^+$ (and possibly S^-).

Definition 1 (Subclass Cover). Let $R(Z) \subset \mathbb{R}^d$ be a convex set defined by a point set $Z \subset \mathbb{R}^d$. The *subclass family* $\mathcal{F}$ of S^+ (against S^-) is a family of subsets of S^+, which satisfies the following conditions:

1.1 Inclusion of positive samples: $S^+ \subset (\cup_{Z \in \mathcal{F}} R(Z))$,
1.2 Exclusion of negative samples: $S^- \cap (\cup_{Z \in \mathcal{F}} R(Z)) = \varnothing$,
1.3 Maximality of each element: for each $Z \in \mathcal{F}$,
$$\forall W \subset S^+ \setminus Z, S^- \cap R(Z \cup W) \neq \varnothing.$$

We call each subset $Z \in \mathcal{F}$ a *subclass*. If 1.1 and 1.2 are satisfied, the subclass family is said to be *feasible*. We can obtain the *unique* subclass family by collecting subsets which satisfy 1.1-1.3 among all subsets.

In other words, for the union of $R(Z), Z \in \mathcal{F}$, the condition 1.1 means "it contains all positive samples", the condition 1.2 means "it cannot contain any negative samples", and the condition 1.3 means that for any $Z \in \mathcal{F}$, if we add any other positive samples to Z, it must violate the condition 1.2. In addition, from the condition 1.2, each $R(Z), Z \in \mathcal{F}$ cannot also contain any negative samples (Figure 2).

2.3 Weak Subclass and Relaxed Subclass

Computation of subclass is often demanding. We can use a *weak* subclass instead which is approximately maximal. This weak subclass is often sufficient for pattern classification, and it can reduce the computational cost as we see later.

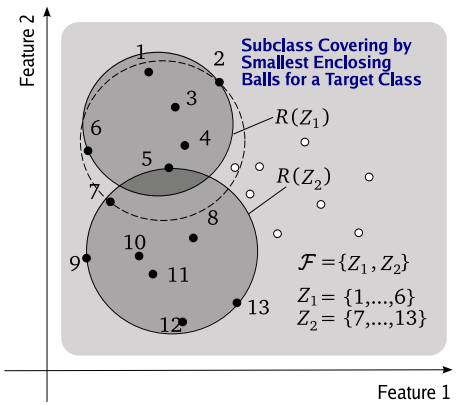

Fig. 2. An example of subclass covering by smallest enclosing balls of subsets. Maximal condition means that we cannot add any other positive samples W to Z_1 (and Z_2). In this example, $W = \{7\}$ violates exclusion of negative samples (*dashed circle*).

For this purpose, we define another condition instead of maximality. A subclass family $\mathcal{F}$ has no elements which becomes a subset of other element. Such a family of subsets is called *Sperner family*[7] (also known as *antichain*), and hence we call this property *Sperner condition*.

Definition 2 (Weak Subclass). We consider the Sperner condition

$$1.4.\ \text{Sperner Condition: } A \in \mathcal{F} \Rightarrow \forall B \in \mathcal{F} \setminus \{A\},\ A \not\subset B,$$

as a weaken condition instead of the maximality condition 1.3 in Definition 1. We call the subclass family satisfying 1.1,1.2 and 1.4 a *weak subclass*.

Strong subclasses always produce perfectly a consistent hypothesis, but in some applications we often need to tolerate training error in order to avoid overfitting. We can relax the condition 1.1 or 1.2. The relaxed condition can depend on the type of $R(\cdot)$ or the type of problem, and we will give later the definition of *relaxed subclass* for spheres in the subsection 3.3.

2.4 Related Previous Approach: Class Cover Problem

In this paper, we will examine the subclass method based on balls. From this viewpoint, we here refer the previously proposed framework called *class cover problem* which has a similar flavor to the subclass cover problem with balls, and discuss the difference between them.

To the best of our knowledge, the class cover problem was introduced by Cannon and Cowen[8] originally as a conference paper in 2000. Subsequently, Priebe *et al.*[9], Marchette[10], and DeVinney[11] studied and developed this

framework, and proposed the graph-theoretic method called *Class Cover Catch Digraph(CCCD)* for computing it.

Let $B(c, r)$ be a ball centered at $c \in \mathbb{R}^d$ with a radius $r \in \mathbb{R}$. The class cover problem is the following problem: Suppose we draw a ball centered at each positive sample with a radius of the distance to the nearest negative sample. Then, find a *minimal* subset of positive samples, such that the corresponding balls can cover all positive samples. This problem is a generalization of the classic set cover problem (see for example [12]).

The subclass cover appears to be an unconstrained and inhomogeneous class cover neglecting that the balls are open or close, but it is not necessarily the minimal family; the class cover problem requires the minimality of a resultant cover while the subclass cover problem requires the maximality of each subclass. Another difference is that the class cover considers a subset of S^+ whereas the subclass cover considers a *family* of subsets.

The previous work mainly focused on the constrained class cover, which can give a method for prototype selection from positive samples, and can provide a prototype-based classifier[10]. Altough the class cover problem of this type becomes NP-hard problem[8, 13] unfortunatelly, the subclass cover problem has a polynomial-time algorithm as presented later.

The property of such a multi-spheres classifier is also discussed by Adam *et al.*[13]. This study extended the classical Vapnik-Chervonenkis learning theory to the data-dependent hypothesis classes. As an example, they discussed the constrained class cover classifier, and showed some interesting properties.

In these contexts, the proposed classification method based on spheres will be also interesting.

3 The Subclass Method Based on a Family of Spheres

We will develop the subclass method based on balls. Hereafter, $R(Z)$ denotes the minimum enclosing ball for the point set Z. It should be noted that the minimum enclosing ball for a set consisting of only one point is defined as a ball centered the point with radius 0.

3.1 Algorithms for Constructing Subclass Family

Exact Algorithm First, we show a polynomial-time exact algorithm which can enumerate all sets in the unique subclass family of target class. This is based on the simple fact that a sphere in $\mathbb{R}^d$ can be determined by at most $d+1$ points[14]. Thus, it is always available when $R(Z)$ is defined by at most d points and the number d does not depend on the size $|Z|$.

Algorithm 1. For two given point sets S^+ and S^-, do the following:

1. Set $\mathcal{H} := \{V \subset S^+ : |V| \leqslant d + 1\}$.
2. Remove sets which cannot exclude negative samples from $\mathcal{H}$.
3. Set $\mathcal{F} := \{S^+ \cap R(V) : V \in \mathcal{H}\}$ and eliminate duplication.

4. For each element in $\mathcal{F}$, if it becomes a subset of other element, remove it.

Roughly speaking, this algorithm first enumerates all subsets of size at most $d+1$ which can exclude all negative samples. Then, we can obtain a subclass family $\mathcal{F}$ as an irreducible set with respect to the Sperner condition 1.4.

Note that the condition 1.3 is satisfied. Suppose $\mathcal{F}$ does not satisfy the maximality condition 1.3, there exists $Z \in \mathcal{F}$ such that $\exists W \subset S^+ \backslash Z, S^- \cap R(Z \cup W) = \varnothing$. Since this $Z \cup W$ can exclude all negative samples, therefore $Z \cup W \in \mathcal{F}$ contradicting our assumption for the existence of Z because $Z \subset Z \cup W$ must be removed in step 4. Hence this algorithm can enumerate all elements of the unique subclass family of S^+.

The number of subsets of size at most $d+1$ is polynomial with respect to the size of inputs. Assuming that the other steps requires only polynomial-time, Algorithm 1 is also polynomial-time computable. This is one of the advantages against the constrained class covers, which are NP-hard[8, 13].

Incremental Algorithm Practically, since the computational cost of Algorithm 1 is still high, improvement of its efficiency should be required. Instead of the enumeration of unique subclasses, we consider the enumeration of elements in any of weak subclass families. In this approach, the maximality and the uniqueness of subclass family are not always satisfied, but such a subclass family is often sufficient in order to construct a pattern classifier.

Algorithm 2. For two given point sets S^+ and S^-, do the following:

1. Let D be a randomly ordered set of S^+.
   Set $C \leftarrow \varnothing$ for the set of tested points, $\mathcal{F} \leftarrow \varnothing$ for the output family, respectively.
2. Repeat the following until $D \setminus C = \varnothing$ is satisfied:
   (a) Select randomly $x \in D \setminus C$, set $Z \leftarrow \{x\}$ and $C \leftarrow C \cup \{x\}$.
   (b) For all $\tilde{x} \in D \setminus \{x\}$, do the following sequentially: If a point set $Z \cup \{\tilde{x}\}$ can exclude S^- then, set $Z \leftarrow Z \cup \{\tilde{x}\}$ and $C \leftarrow C \cup \{\tilde{x}\}$.
   (c) $\mathcal{F} \leftarrow \mathcal{F} \cup Z$.
3. After eliminating duplication, for each element in $\mathcal{F}$, if it becomes a subset of other element, remove it.

3.2 Classification based on Subclass Family

We now turn to the pattern classification problem. To implement the original idea described in section 2.1, we use the directed length of the minimal projection onto spheres for classifying the test samples. The directed length of projection of the point x onto sphere $B(c, r)$ is defines as

$$\tilde{d}(x, B(c, r)) := \|x - c\| - r.$$

It should be noted that if the point x is in $B(c, r)$, the value of $\tilde{d}(x, B(c, r))$ becomes negative.

For given subclass families $\mathcal{F}_1, \ldots, \mathcal{F}_C$ for each class $i = 1, \ldots, C$, the classification is based on

$$f(x) := \arg \min_{i=1,\ldots,C} \min_{Z \in \mathcal{F}_i} \tilde{d}(x, R(Z)).$$

We can have no training error when the exclusion of negative samples are perfect. Under this quasi-distance, each subclass ball acts like prototypes for the corresponding class.

3.3 Relaxed subclass family for balls

As touched in 2.3, the perfect exclusion of negative samples often yields overfitting for practical problems; Thus we often need the relaxed version of exclusion condition to tolerate training error.

In Definition 1 of subclass family, the condition 1.2 can be relaxed. As a benefit from the formalization, we can easily develop the soft-classification version of subclass method by replacing the condition 1.2 by the relaxed condition. For both Algorithm 1 and 2, the required modification is only this replacement when we check the exclusiveness of subclass. In soft-classification version, for a given parameter ξ, "$B(c, r)$ can exclude negative samples" means

$$r = 0 \quad \text{or} \quad \sum_{x \in S^-} \max\left(0, 1 - \frac{\|x - c\|}{r}\right) \leqslant \xi.$$

From the definition, when $\xi = 0$, it is consistent with the perfect exclusion of negative samples (i.e. hard-classification version). In addition, we consider the second additional condition: For a given parameter δ,

$$\delta > \frac{\#\text{ of containing negatives}}{\#\text{ of all negatives}}.$$

This additional condition is sometimes needed for avoiding excessively incorporation of negative samples to the subclass ball when we use the relaxed condition.

3.4 Computational Issues

Monotonicity of Representation We identify $R(\cdot)$ with a function that maps any $Z \subset \mathbb{R}^d$ to $R(Z)$. We call $R(\cdot)$ a *representation*. For a given representation $R(\cdot)$ and any two point sets $U, V \subset \mathbb{R}^d$, if $U \subset V \Rightarrow R(U) \subset R(V)$ holds true, we say that the representation $R(\cdot)$ is *monotonic*.

The axis-parallel rectangles are monotonic. When the representation is monotonic, the incremental algorithm does not violate the maximality condition. However the minimum enclosing balls are non-monotonic, and thus the incremental algorithm will compute just a approximation of maximal subclasses. It should be noted that the exact algorithm can enumerate the unique subclasses in both cases.

Minimum Enclosing Ball Computation For an implementation of Algorithm 1 or 2, the efficient method computing the minimum enclosing ball for a given point set is required. Computation of the minimum enclosing ball has a long history[14] and many algorithms have been developed. Recently, computation in higher-dimensional space or computation for large-scale problem has been studied. Our implementation is based on the simple algorithm [15] which works efficiently for $d < 30$. For more higher-dimensional problems, we can use alternatively the computational geometric method[16] or the aggregation function method and second-order cone programming-based method[17].

4 Examples

In Figure 3, we showed the illustrative example in 2-dimensional classification problem including the result by class cover catch digraph method[9] for comparison. The results were computed by Algorithm 1 and we can see that the original idea of convex subclass method described in section 2.1 was realized well.

In order to examine the behavior for more higher dimensional data, we compared three methods: (1) the subclass method based on balls, (2) the relaxed subclass method, (3) support vector machines [4] with Gaussian kernel $K(x, y) := \exp(-\gamma\|x - y\|^2)$ and a regularization parameter C, and (4) k-nearest neighbor method. The numerical experiments were based on 10 fold cross-validation for 3 numerical datasets from UCI machine learning repository[18]: **iris** (4 features, 3 classes, 150 samples), **glass** (9 features, 6 classes, 214 samples), and **wine** (13 dimensional, 3 classes, 178 samples). The result shown in Table 1 was computed by Algorithm 2. Therefore, it depends on randomness in Algorithm 2 and the obtained subclass is not unique. But the result seems to be good enough compared with the conventional classifiers and the approximated subclasses will work well in higher-dimensional spaces. We can also see the effect of the second additional condition.

Table 1. Estimated classification rate by 10-fold CV (correct number)

	Subclass	Subclass($\xi = 0.5$)		SVM($C = 1$)		SVM($C = 100$)		k-NN	
		δ		γ		γ		k	
		-	0.1	0.01	0.25	0.01	0.25	1	5
iris	96.0	90.0	94.0	88.0	96.0	96.0	94.7	96.0	96.0
(150)	(144)	(135)	(141)	(133)	(144)	(144)	(142)	(144)	(144)
glass	72.0	63.1	65.0	50.5	68.7	67.3	61.9	70.1	65.9
(214)	(154)	(135)	(139)	(108)	(147)	(144)	(136)	(150)	(141)
wine	94.9	89.9	94.9	97.8	96.6	96.1	97.2	94.9	96.1
(178)	(169)	(160)	(169)	(174)	(172)	(171)	(173)	(169)	(171)

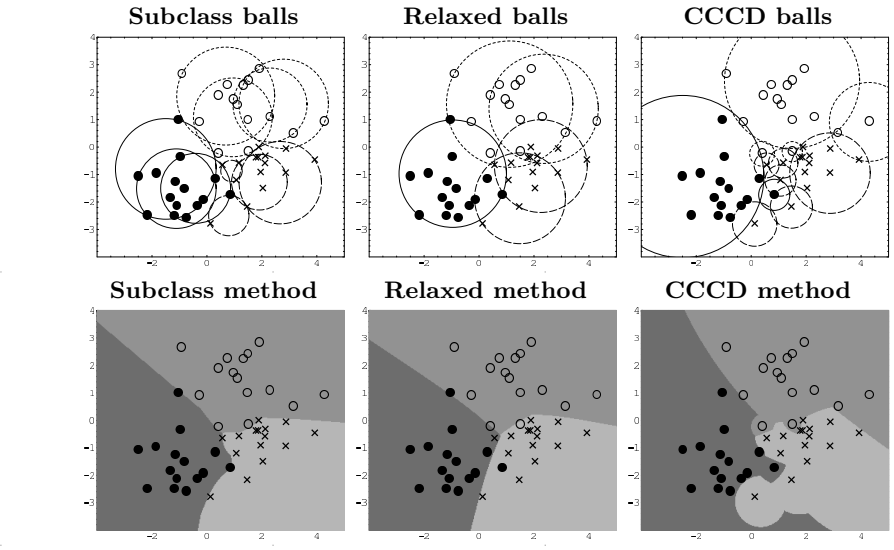

Fig. 3. Balls and decision boundaries of subclass method, relaxed subclass method with $\xi = 1$, and class cover catch digraph method[9]

5 Conclusion

We proposed a new nonparametric classification framework: The (convex) subclass method. According to that, we developed a combinatorial classifier based on a family of spheres, and showed a polynomial-time exact algorithm and an incremental algorithm. Additionally, the relaxed subclasses were considered and through some numerical examples we confirmed its effectiveness. Further researches will include some theoretical analysis on the dependency of randomness and the expected computational cost, developing more efficient computational methods, implementing parallel computing of subclasses, examining the subclasses based on various convex sets, and considering better relaxed conditions.

References

1. Bousquet, O., Boucheron, S., Lugosi, G.: Theory of classification: A survey of recent advances. ESAIM Probability and Statistics, (to appear) (2004)
2. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer-Verlag, New York (1996)
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification 2nd Ed. John Wiley & Sons (2001)
4. Vapnik, V.N.: The Nature of Statistical Learning Theory 2nd Ed. Springer-Verlag, New York (2000)
5. Kudo, M., Yanagi, S., Shimbo, M.: Construction of class regions by a randomized algorithm: A randomized subclass method. Pattern Recognition **29** (1996) 581–588

6. Takigawa, I., Abe, N., Shidara, Y., Kudo, M.: The boosted/bagged subclass method. International Journal of Computing Anticipatory Systems **14** (2004) 311–320
7. Erdős, P., Kleitman, D.: Extremal problems among subsets of a set. Discrete Mathematics **8** (1974) 281–294
8. Cannon, A.H., Cowen, L.J.: Approximation algorithms for the class cover problem. Annals of Mathematics and Artificial Intelligence **40** (2004) 215–223
9. Priebe, C.E., Marchette, D.J., DeVinney, J.G., Socolinsky, D.A.: Classification using class cover catch digraphs. Journal of Classification **20** (2003) 3–23
10. Marchette, D.J.: Random Graphs for Statistical Pattern Recognition. John Wiley & Sons (2004)
11. DeVinney, J.G.: The Class Cover Problem and Its Application in Pattern Recognition. Ph.D. Thesis, The Johns Hopkins University (2003)
12. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms 2nd Ed. MIT Press (2001)
13. Cannon, A.H., Ettinger, J.M., Hush, D., Scovel, C.: Machine learning with data dependent hypothesis classes. Journal of Machine Learning Research **2** (2002) 335–358
14. Welzl, E.: Smallest enclosing disks (balls and ellipsoids). New Results and New Trends in Computer Science, LNCS **555** (1991) 359–370
15. Gärtner, B., Schönher, S.: Fast and robust smallest enclosing balls. Proc. 7th Annual European Symposium on Algorithms (ESA), LNCS **1643** (1999) 325–338
16. Fischer, K., Gärtner, B., Kutz, M.: Fast smallest-enclosing-ball computation in high dimensions. Proc. the 11th Annual European Symposium on Algorithms (ESA) (2003)
17. Zhou, G.L., Toh, K.C., Sun, J.: Efficient algorithms for the smallest enclosing ball problem. Computational Optimization and Applications, (accepted) (2004)
18. Blake, C., Merz, C.: UCI repository of machine learning databases (1998)