Available online at www.sciencedirect.com
SCIENCE DIRECT®
Discrete Applied Mathematics 154 (2006) 1975–1982
DISCRETE APPLIED MATHEMATICS
www.elsevier.com/locate/dam

A new family of proximity graphs: Class cover catch digraphs

Jason DeVinney[a], Carey E. Priebe[b]

[a]*Center for Computing Sciences, Bowie, MD, USA*
[b]*Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, USA*

Received 12 August 2004; received in revised form 6 April 2006; accepted 12 April 2006
Available online 30 May 2006

Abstract

Motivated by issues in machine learning and statistical pattern classification, we investigate a class cover problem (CCP) with an associated family of directed graphs—class cover catch digraphs (CCCDs). CCCDs are a special case of catch digraphs. Solving the underlying CCP is equivalent to finding a smallest cardinality dominating set for the associated CCCD, which in turn provides regularization for statistical pattern classification. Some relevant properties of CCCDs are studied and a characterization of a family of CCCDs is given.

Keywords: Class cover problem; Pattern classification; Machine learning; Digraph

1. Class cover problem

Consider a set $\mathscr{X}$ with a dissimilarity measure d and two finite, non-empty sets $\mathscr{X}_+, \mathscr{X}_- \subseteq \mathscr{X}$, with a distinction of *target* class given to one of the sets. Recall that a dissimilarity measure on $\mathscr{X}$ is a function $d : \mathscr{X} \times \mathscr{X} \to \mathbb{R}_+$ such that $\forall x_1, x_2 \in \mathscr{X}$, $d(x_1, x_1) = 0$ and $d(x_1, x_2) = d(x_2, x_1) \geqslant 0$. We will denote $\mathscr{X}_+$ as the target class unless otherwise specified. In its most general form, the class cover problem (CCP) is to find a minimum cardinality set of open covering balls B_i, with center c_i and radius r_i ($B_i = \{x \in \mathscr{X} : d(x, c_i) < r_i\}$), whose union (the *cover*) contains all of the target class and does not contain any of the non-target class. The CCP is, therefore, a special case of the classic set cover problem [1] with constraints on the type of covering sets that are considered. We write the list $((\mathscr{X}, d), \mathscr{X}_+, \mathscr{X}_-)$ to denote an instance of the CCP.

The above formulation allows several variations of the CCP. The CCP was introduced in [3] by Cannon and Cowen. They considered the constrained (all covering balls must be centered at target class points) homogeneous (all covering balls must have the same radius) CCP. The constrained inhomogeneous CCP (CICCP) was introduced in [17] and is the version we will focus on in this paper.

The CCP was originally motivated by supervised pattern classification, and while the focus of this paper is not classification, we will review the topic and its connection to the CCP to demonstrate this motivation. We will describe a model of two-class supervised pattern classification. Consider a process in which we first draw, at random, a label from the set $\{-1, +1\}$ and then (conditionally on the value of the chosen label) choose a random observation from the set $\mathscr{X}$. Let Y be the random variable whose value is that of the chosen label and X be the random variable whose

E-mail addresses: jgdevin@super.org (J. DeVinney), cep@jhu.edu (C.E. Priebe).

0166-218X/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2006.04.004

value is the observation from $\mathcal{X}$. Assume the joint distribution function for X and Y exists and is denoted $F_{X,Y}$. Then the *class conditional* distribution functions are $F_+ = F_{X|Y=+1}$ and $F_- = F_{X|Y=-1}$, and the *prior probabilities of class membership* are $\pi_+ = P[Y = +1]$ and $\pi_- = P[Y = -1]$. A *training set* is a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from $F_{X,Y}$. The training set will be separated, based on the value of Y, into two sets X_+ and X_-. Note that X_+ and X_- are random sets, while $\mathcal{X}_+$ and $\mathcal{X}_-$ are observed sets. A classifier is a function g that returns a label in $\{-1, 0, +1\}$ for each point in $\mathcal{X}$. (We include the possibility of "no decision" represented by $g(x) = 0$.) The goal of classification in this case is to find a classifier that satisfies

$$\arg\min_{\mathcal{G}} P[g(X) \neq Y], \tag{1}$$

where $\mathcal{G}$ is the class of functions $g : \mathcal{X} \to \{-1, 0, +1\}$. (Use of the "no decision" option $g(x) = 0$ cannot improve the probability of misclassification, of course, but is of value in some practical situations.) Since the classifiers we create will rely on some observed training set $\mathcal{X}_+, \mathcal{X}_-$, we will denote them as $g_{\mathcal{X}_+,\mathcal{X}_-}(\cdot)$. Assuming that the class conditional probability density functions f_+ and f_- exist, the *discriminant region* for the $+$ class is the subset of $\mathcal{X}$ where $f_+ > f_-$. The classifier that satisfies

$$g(x) = \begin{cases} +1 & f_+(x) > f_-(x), \\ -1 & f_-(x) > f_+(x), \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

is optimal and is called the *Bayes optimal classifier*. Since in general we have little or no information regarding the class conditional distributions, the challenge is to find classifiers that closely approximate the Bayes optimal classifier.

Priebe et al. [18] describe a method of constructing a classifier from solutions to a CCP (CCP classifiers). By switching the role of target class between the classes of training observations, two different instances of some variant of the CCP can be solved, resulting in two covers C_+ and C_-. Each cover can be used to provide a simple estimate of the discriminant region for the class it covers. This is achieved by defining a cover-dissimilarity function ρ which gives a dissimilarity between points in $\mathcal{X}$ and a cover. The classifier then labels each point in $\mathcal{X}$ with the label of the cover, C_+ or C_-, to which it is closest. A simple CCP classifier $g_{\mathcal{X}_+,\mathcal{X}_-}(x) : \mathcal{X} \to \{-1, 0, +1\}$ uses the simple cover dissimilarity function

$$\rho_S(x, C) = \begin{cases} 0 & \text{if } x \in C, \\ 1 & \text{otherwise} \end{cases}$$

and is explicitly written as

$$g_{\mathcal{X}_+,\mathcal{X}_-}(x) = \begin{cases} +1 & \rho_S(x, C_+) < \rho_S(x, C_-), \\ -1 & \rho_S(x, C_+) > \rho_S(x, C_-), \\ 0 & \text{otherwise} \end{cases}$$

or equivalently

$$g_{\mathcal{X}_+,\mathcal{X}_-}(x) = \begin{cases} +1 & x \in C_+ \cap C_-^c, \\ -1 & x \in C_- \cap C_+^c, \\ 0 & \text{otherwise}, \end{cases}$$

where C^c is the set of points in $\mathcal{X}$ that are not in C. More elaborate strategies for classifier construction are presented in [18].

The art of classifier construction involves assessing empirical error (classifier performance on the training data) and inferring generalization performance (classifier performance on non-training data); in general, optimizing the empirical error does not imply optimal generalization performance, and some sort of *regularization* or complexity penalty is necessary. In the CCP classifier above, the CCP chooses the centers of the balls in the cover as representatives or prototypes for the entire class [7]. The complexity of the approximation of the discriminant region for the classifier is, in general, increasing in the number of points chosen to make up the covers; therefore, small cardinality covering sets (as chosen by the CCP) should have superior generalization performance over a classifier which chooses more balls in

its cover. That is, finding small cardinality covering sets provides regularization. This is analogous to the motivation behind the reduced nearest neighbor classifier found in [20].

Recent efforts in CCP classification, while demonstrating significant practical value, have been based on heuristics with few theoretical results [15,18,19,8]. This is partly due to the complex nature of the covers formed in the solution of the CCP. In Section 2 we demonstrate how the CICCP can be easily formulated as a problem on a special class of directed graphs called class cover catch digraphs (CCCDs). Translating the CCP to a problem on directed graphs is useful for two reasons. It is convenient to use the pre-established and familiar language of graph theory to describe CCP related concepts. Also, we hope the existence of this equivalent and easily stated problem formulation will increase our theoretical understanding of the problem by providing new approaches to viewing the problem. We also believe that CCDs are an interesting new addition to the family of proximity graphs [21,11,14]. The remainder of the paper introduces some fundamental properties of CCCDs.

2. Class cover catch digraphs

Recall that a *catch digraph* of a collection of sets $\mathcal{S} = S_1, S_2, \ldots, S_n$ and corresponding base points $\mathcal{T} = t_1, t_2, \ldots, t_n$ is the digraph with vertex set $V = v_1, v_2, \ldots, v_n$ with an arc (a directed edge) from v_i to v_j if and only if $t_j \in S_i$ (see [16]). We will say that the catch digraph formed as described above is the catch digraph *induced by* $\mathcal{S}$ and $\mathcal{T}$.

Since the centers of the open balls in the CICCP must be located at target class points and the radii can be as large as possible without covering any non-target class points, we may define a maximal covering ball at each target class point. Given two sets of points $\mathcal{X}_+, \mathcal{X}_- \subseteq \mathcal{X}$ with $\mathcal{X}_+$ as the target class, we define such a ball for each $x_i \in \mathcal{X}_+$ as $B_{x_i} := \{x \in \mathcal{X} : d(x_i, x) < \min_{x_- \in \mathcal{X}_-} d(x_i, x_-)\}$. B_{x_i} is the largest ball centered at x_i not covering any points in $\mathcal{X}_-$. We call the catch digraph D induced by the collection of B_{x_i} and their centers x_i the CCCD for $\mathcal{X}_+, \mathcal{X}_-$ in the dissimilarity space $(\mathcal{X}, d)$. We will define $C((\mathcal{X}, d), n, m)$ to be the family of all possible unlabeled CCCDs induced by n target class points and m non-target class points in the space $(\mathcal{X}, d)$. A digraph on n vertices is a CCCD if there exists some dissimilarity space $(\mathcal{X}, d)$ and m such that the digraph is in $C((\mathcal{X}, d), n, m)$. Note that the property of being a CCCD is hereditary. That is, if $D = (V, A) \in C((\mathcal{X}, d), n, m)$ and $W \subset V$ with $|W| = k$, then the induced digraph $D' = (W, A')$ is a member of $C((\mathcal{X}, d), k, m)$.

In our study of the CCP, we will gain some insight by studying CCCDs. One of the first things we may wish to achieve is a characterization of CCCDs or conditions on $(\mathcal{X}, d)$ such that a given digraph is a CCCD. We begin by giving a necessary condition for a digraph to be a CCCD in Theorem 1. We must first define the notion of a *ball digraph* and a *simple cycle*. The ball digraph of a set of points $z_i \in \mathcal{X}$ and associated radii $r_i \in \mathbb{R}$ is the catch digraph induced by the collection of $B(z_i, r_i)$ and their centers z_i, where $B(z, r) = \{x \in \mathcal{X} : d(x, z) < r\}$. (Notice that any CCCD is also a ball digraph.) A *bidirected arc* between two vertices v and u is the pair of arcs (v, u) and (u, v). A *simple cycle* is a directed cycle that contains no bidirected arcs.

Theorem 1. *If D is a ball digraph then D contains no simple cycles.*

Proof. Let D be a ball digraph induced from points in a general dissimilarity space $(\mathcal{X}, d)$. Suppose for contradiction that D has a simple cycle C consisting of vertices $v_1, \ldots, v_l$. For each $i = 1, \ldots l$, there is an arc from v_i to v_{i+1} (all addition in this proof is assumed to be *modulo* l) but not an arc from v_i to v_{i-1} since C is a simple cycle. Since D is a ball digraph there is a set of points x_i in $(\mathcal{X}, d)$ and associated radii $r_i \in \mathbb{R}$ such that $d(x_i, x_{i+1}) < r_i \leqslant d(x_i, x_{i-1})\ \forall i$. This is so since $B(x_i, r_i)$ must contain x_{i+1} but must not contain x_{i-1}. Such a set of inequalities are impossible since they imply that $d(x_i, x_{i+1}) < d(x_i, x_{i+1})\ \forall i$. Therefore, D cannot contain a simple cycle. □

For a general dissimilarity space, the converse is not true; the lack of simple cycles is not a sufficient condition to be a ball digraph on that dissimilarity space. For example consider the discrete metric on the space of the integers $(\mathbb{Z})$. The discrete metric ρ on $\mathbb{Z}$ is defined as $\rho : \mathbb{Z} \times \mathbb{Z} \to \{0, 1\}$ where for $x, y \in \mathbb{Z}$, $\rho(x, y) = 1$ if and only if $x \neq y$. Using this metric as a dissimilarity measure, the ball digraph $D = (V, A)$ induced by any set of points $x_i \in \mathbb{Z}$ and associated radii $r_i \in \mathbb{R}$ will have the property that all vertices have degree zero or $|V| - 1$. Therefore, a directed path on three vertices is an example of a digraph that does not contain any simple cycles, yet is not a ball digraph on the dissimilarity space $(\mathbb{Z}, \rho)$.

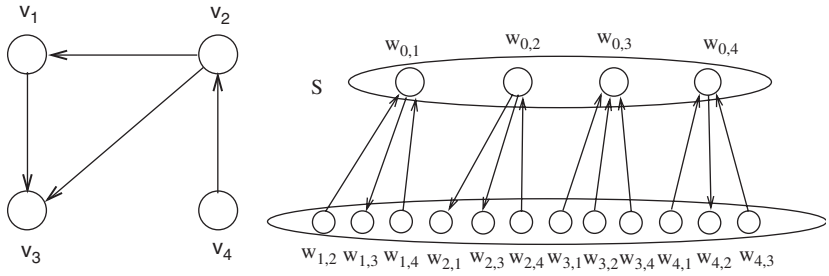

Fig. 1. An example of G and G^*.

2.1. Euclidean CCCDs

Consider the special case where $\mathscr{X} = \mathbb{R}^q$ and $d(x, y) = (\sum_{i=1}^{q}(x_i - y_i)^2)^{1/2}$ is the L_2 metric. We will call a CCCD in $C((\mathbb{R}^q, L_2), n, m)$ a *Euclidean CCCD*. Euclidean CCCDs are a special case of sphere digraphs as introduced by Maehera in [13]. A digraph D on n vertices is a Euclidean CCCD if there exists a set of n target class points and $m > 0$ non-target class points in $\mathbb{R}^q$ for some q which induce (via the Euclidean L_2 metric) a CCCD which is isomorphic to D. In this section, we will demonstrate a characterization of Euclidean CCCDs.

The *dissimilarity* matrix of a set of n points in some dissimilarity space is the $n \times n$ matrix whose i, j entry is the dissimilarity between point i and point j. An $n \times n$ matrix B is said to be *Euclidean embeddable* if there are points in $\mathbb{R}^{n-1}$ with dissimilarity matrix (where the dissimilarity is the Euclidean metric) equal to B. It is well known that for any $n \times n$ symmetric matrix B' with zero along the diagonal and non-negative off diagonal entries, there is a constant c such that $B' + c \cdot ee^{\mathrm{T}} - c \cdot I$ is Euclidean embeddable, where e is the n-dimensional vector of ones and I is the n-dimensional identity matrix (see for instance [5]). Finally, for a digraph $D = (V, A)$, the transitive closure of D is the digraph $D' = (V, A')$ where $A' = \{(x, y) | \exists$ a directed path from x to y in $D\}$.

Theorem 2. *A digraph $G = (V, A)$ on n vertices is in $C((\mathbb{R}^n, L_2), n, 1)$ if and only if it has no simple cycles.*

Proof. ($\Rightarrow$) Since any digraph in $C((\mathbb{R}^n, L_2), n, 1)$ is a ball digraph, this direction is implied by Theorem 1.

($\Leftarrow$) Let $G = (V, A)$ be a digraph on vertices $\{v_1, v_2, \ldots, v_n\}$ with no simple cycles. We will prove that there exists a set of points $\{x_0, x_1, \ldots, x_n\} \in \mathbb{R}^n$ (x_0 will be the lone non-target class point) which induce a CCCD isomorphic to G. Form a new digraph $G^* = (W^*, A^*)$ with $W^* = \{w_{i,j} : i \neq j,\ i, j \in \{0, 1, \ldots, n\}\}$,

$$(v_i, v_j) \in A \iff (w_{0,i}, w_{i,j}) \in A^* \quad \forall i, j \neq 0$$

and

$$(v_i, v_j) \notin A \iff (w_{i,j}, w_{0,i}) \in A^* \quad \forall i, j \neq 0.$$

G^* is a bipartite graph with the partition $R = \{w_{i,j} : i, j \neq 0, i \neq j, 0 \leqslant i, j \leqslant n\}$ and $S = \{w_{0,j} : 0 \leqslant j \leqslant n\}$. (See Fig. 1.) Clearly G^* does not contain any cycles since every vertex in R has degree exactly one. Let $G^{**} = (W^*, A^{**})$ be the transitive closure of G^*. G^{**} is also acyclic (where acyclic refers to *directed* cycles) since the transitive closure of an acyclic digraph is also acyclic. We associate an element $d_{i,j}$ with $w_{i,j}$ for $0 \leqslant i, j \leqslant n$. Since G^{**} is acyclic we may define a partial order on the set $M = \{d_{i,j} : 0 \leqslant i, j \leqslant n\}$ by demanding $d_{i,j} > d_{s,t} \iff (w_{i,j}, w_{s,t}) \in A^{**}$ (see for instance [2]). We now extend the partial order on the elements of M to a strict total order. Finally we assign a value of k to the kth largest element in our total order and then create a dissimilarity matrix $D = [d_{i,j}]$. We create a matrix $D' = [d'_{i,j}]$ (Euclidean embeddable) by adding an appropriate constant to all off-diagonal entries of D. The addition of a constant to each off-diagonal entry in D preserves the ranking in the total order, that is, $d_{i,j} > d_{s,t}$ iff $d'_{i,j} > d'_{s,t}$.

We claim that a set of points (non-target class) $\{x_0\}$ and (target class) $\{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^n$ that satisfy $d_2(x_i, x_j) = d'_{i,j}$ induce a CCCD isomorphic to G. To see why, suppose (v_i, v_j) is an edge in the CCCD induced by $\{x_0, x_1, \ldots x_n\}$. Then it must be the case that $d_2(x_i, x_j) < d_2(x_0, x_i)$. This implies that $(w_{0,i}, w_{i,j}) \in A^*$ and thus $(v_i, v_j) \in A$. Conversely,

suppose $(v_i, v_j) \in A$, then it follows that $d_{0,i} > d_{i,j}$ and so $d_2(x_i, x_j) < d_2(x_0, x_i)$. Thus, $B_{d_2}(x_i, d_2(x_i, x_0))$ contains x_j and so (v_i, v_j) is an edge in the CCCD. □

Corollary 1. *A digraph D is a Euclidean CCCD* $\iff$ *D has no simple cycles.*

2.2. Minkowski metrics

We will let the metric $d_p : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}_+$ be the familiar L_p metric; that is, $d_p(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p}$. A dissimilarity space (I, d) consists of a non-empty set I and dissimilarity measure $d : I \times I \to \mathbb{R}_+$. A dissimilarity space (I, d) is said to be embeddable in a metric space (M, ρ) if there is a function $\phi : I \to M$ such that

$$\rho(\phi(i), \phi(j)) = d(i, j) \quad \forall i, j \in I.$$

Critchley and Fichet [6] showed that a finite subset of $(\mathbb{R}^q, d_2)$ is embeddable in $(\mathbb{R}^q, d_1)$ and $(\mathbb{R}^q, d_\infty)$. This along with Theorem 2 immediately implies the following corollary.

Corollary 2. *A digraph* $G = (V, A)$ *is in* $C((\mathbb{R}^n, L_p), n, 1)$ *if and only if it has no simple cycles* ($p \in \{1, 2, \infty\}$).

Note that among the above Minkowski metrics, there are differences in $C((\mathbb{R}^q, L_p), n, 1)$ for some $q < n$ among different values of p. For example, the empty graph with eight vertices is in $C((\mathbb{R}^2, L_\infty), 8, 1)$ and $C((\mathbb{R}^2, L_1), 8, 1)$, but not $C((\mathbb{R}^2, L_2), 8, 1)$.

2.3. Domination and independence in CCCDs

A set S of vertices of a digraph $D = (V, A)$ is a *dominating set* if for all $v \in V : v \in S$ or $\exists x \in S$ such that $(x, v) \in A$. If J is a collection of indices such that $\{B_j : j \in J\}$ is an optimal solution to an instance of CICCP then the set $\{v_j : j \in J\}$ is a minimum cardinality dominating set in the CCCD induced by the instance and vice versa. The CICCP is, therefore, equivalent to finding a minimum cardinality dominating set in the CCCD induced by $((\mathscr{X}, d), \mathscr{X}_+, \mathscr{X}_-)$. An *independent set* in a digraph $D = (V, A)$ is a set of vertices, $W \subseteq V$ such that for all v, w pairs in W, neither arc (v, w) nor (w, v) is in A. For a digraph $D = (V, A)$, let $\alpha(D)$ be the size of the largest independent set and $\gamma(D)$ be the size of the smallest dominating set.

Theorem 3. *If a digraph D has no simple cycles then* $\alpha(D) \geqslant \gamma(D)$.

Proof. Let $D = (V, A)$ be a digraph with no simple cycles with $|V| = n$. Then by Theorem 2, D is a Euclidean CCCD ($D \in C((\mathbb{R}^n, L_2), n, 1)$), thus there are sets $\mathscr{X}_+, \mathscr{X}_- \in \mathbb{R}^n$ (with $|\mathscr{X}_+| = n$ and $\mathscr{X}_- = \{0\}$) which induce a digraph isomorphic to D. We will find an independent dominating set of size $\hat{\gamma}(G)$ in D. This will show for any such digraph $\alpha(D) \geqslant \hat{\gamma}(D) \geqslant \gamma(D)$. The following *greedy radius algorithm* run on $\mathscr{X}_+$ and $\mathscr{X}_-$ finds an independent dominating set. The greedy radius algorithm is similar to the standard greedy algorithm for the set covering problem [10].

$E = \emptyset, C = V, i = 1$
while $C \neq \emptyset$
  $i^* = \arg\max\{d(x_i, 0) : i \in C\}$
  Let $O_{i*} = \{v \in C : (v_{i^*}, v) \in A\}$
  $C = C - O_{i^*} - v_{i^*}$
  $E = E \cup v_{i^*}$
return E

To see that the set E is independent, consider two points v_i and v_j in E with associated covering balls with radii r_i and r_j. Without loss of generality, suppose the algorithm chose v_i before v_j, implying $r_i \geqslant r_j$. It is obvious that $(v_i, v_j) \notin A$ since the algorithm only chooses points which have not been covered. Also, $(v_j, v_i) \notin A$ since $d(x_i, x_j) \geqslant r_i \geqslant r_j$. The set E is a dominating set since $C = \emptyset$ at the conclusion of the algorithm and points are removed from C only after they are covered by some point in E. □

In $\mathbb{R}^q$, using the Euclidean metric, define a *kissing set* as a set of centers of non-intersecting hyper-spheres with radius one, whose boundaries intersect the boundary of a hyper-sphere of radius one (we imagine this central sphere being centered at the origin). The *kissing number*, $\tau(q)$, is the size of the largest possible kissing set in $\mathbb{R}^q$ [4]. For $x \in \mathbb{R}^q$, we denote $\|x\|$ as the Euclidean distance from x to the origin. For points a, b and c in $\mathbb{R}^n$, we will denote the angle formed by the line segments (a, b) and (b, c) as $\langle a, b, c$.

Lemma 1. *A set K of points in $\{x : \|x\| = 2\}$ is a kissing set in $\mathbb{R}^q$ if and only if for any two points $a, b \in K$, $\theta = \angle a, \{0\}, b \geqslant \pi/3$.*

Proof. ($\Rightarrow$) We know that $\|a\| = \|b\| = 2$ and $\|a - b\| \geqslant 2$ (since the hyper-spheres centered at a and b are non-intersecting). Thus using the Law of Cosines,

$$\begin{aligned}\cos(\theta) &= \frac{\|a\|^2 + \|b\|^2 - \|a - b\|^2}{2\|a\|\|b\|} \\ &\leqslant \frac{1}{2}.\end{aligned}$$

Which implies $\theta \geqslant \pi/3$.

($\Leftarrow$) The above can be reversed to show the converse. □

Lemma 2. *Let $D = (V, A)$ be a digraph in $C((\mathbb{R}^q, L_2), n, 1)$ with a target class point $x_i \in \mathbb{R}^q$ associated with each vertex v_i and a non-target class point located at the origin. If $\{v_i, v_j\}$ are independent vertices, then the angle $\phi = \angle x_i, 0, x_j \geqslant \pi/3$.*

Proof. Let $\phi = \angle x_i, 0, x_j$ and without loss of generality let $\|x_i\| \geqslant \|x_j\|$. Using the Law of Cosines,

$$\|x_i - x_j\|^2 = \|x_i\|^2 + \|x_j\|^2 - 2\|x_i\|\|x_j\|\cos(\phi)$$

which implies,

$$2\|x_i\|\|x_j\|\cos(\phi) \leqslant \|x_j\|^2$$

since $\|x_i - x_j\| \geqslant \|x_i\|$ (by our assumption of independence). Finally we get

$$\begin{aligned}\cos(\phi) &\leqslant \frac{\|x_j\|}{2\|x_i\|} \\ &\leqslant \frac{1}{2}\end{aligned}$$

which implies that $\phi \geqslant \pi/3$. □

Theorem 4. *For a digraph $D \in C((\mathbb{R}^q, L_2), n, 1)$, $\alpha(D) \leqslant \tau(q)$.*

Proof. Given a digraph $D = (V, A)$, in $C((\mathbb{R}^q, L_2), n, 1)$, we will construct a kissing set in $\mathbb{R}^q$ of size $\alpha(D)$. Let $\mathscr{X}_+$ and $\mathscr{X}_-$ be sets of points in $\mathbb{R}^q$ (with $|\mathscr{X}_+| = n$ and $\mathscr{X}_- = \{0\}$) which induce a digraph isomorphic to D. Let $S \subset V$ be an independent set in D and let $\mathscr{S}_+ \subset \mathscr{X}_+$ be corresponding points in $\mathscr{X}_+$. For each $x_i \in \mathscr{S}_+$ define a new point $z_i = 2x_i/\|x_i\|$ (this is the radial projection of each point onto the hyper-sphere of radius two centered at the origin). By Lemmas 1 and 2, the z_i's form a kissing set. □

We show that this bound is tight. Given a kissing set of size $\tau(q)$ in $\mathbb{R}^q$, we will construct an edgeless digraph in $C((\mathbb{R}^q, L_2), \tau(q), 1)$. Let $\mathscr{X}_- = \{0\}$ and let $\mathscr{X}_+$ be the $\tau(q)$ points in the kissing set. For any pair (x_i, x_j) it must be the case that $\|x_i - x_j\| \geqslant 2$ since the open spheres of radius one centered at these points do not intersect. Let $D = (V, A)$ be the CCCD induced by these sets. Since the radius of each B_i is 2 it follows that $x_i \notin B_j \ \forall i \neq j$ which implies that $A = \emptyset$. Therefore, $\alpha(D) = |V| = \tau(q)$.

Given a set of points X in some dissimilarity space $\mathscr{X}$, the *Voronoi region* of $x \in X$ is the set of points in $\mathscr{X}$ which are closer to x than any other point in X.

Corollary 3. *For* $D \in C((\mathbb{R}^q, L_2), n, m)$, $\gamma(D) \leqslant m \cdot \tau(q)$.

Proof. Let $\mathscr{X}_+, \mathscr{X}_- \subset \mathbb{R}^q$ ($|\mathscr{X}_+| = n$, $|\mathscr{X}_-| = m$) be sets which induce a digraph isomorphic to D. We partition $\mathbb{R}^q$ into the Voronoi regions V_i for each point $y_i \in \mathscr{X}_-$. We may now bound the cardinality of the solution to each instance of CCP $((\mathbb{R}^q, L_2), X \cap V_i, \{y_i\})$, corresponding to each point $y_i \in \mathscr{X}_-$ ($i = 1, 2, \ldots, m$), by $\tau(q)$ by Theorems 3 and 4. The result follows. □

In a general digraph or graph, the problem of finding a minimum cardinality dominating set is NP-Hard [9]. But we note that for fixed dimension q^* and fixed m^* the calculation of minimum cardinality dominating set for a digraph in $C((\mathbb{R}^{q^*}, L_2), n, m^*)$ is polynomial-time solvable. This is a consequence of Corollary 3, implying the calculation can be exhaustively done with at most $O(n^{m^* \cdot \tau(q^*)})$ operations.

3. Size of $C((\mathbb{R}^q, L_2), n, m)$

In this section we investigate the size of the family of Euclidean CCCDs. We show that compared to the family of all digraphs, $C((\mathbb{R}^q, L_2), n, m)$ is small, but its growth rate is exponential in n.

To see that $C((\mathbb{R}^q, L_2), n, m)$ is small, we will generate a random digraph on n vertices in a manner similar to the Erdos–Rényi random graph. Between any two vertices i and j, we will add an arc (i, j) with probability $\frac{1}{4}$, arc (j, i) with probability $\frac{1}{4}$, both arcs (i, j) and (j, i) with probability $\frac{1}{4}$, and no arcs with probability $\frac{1}{4}$. Arc additions between any pair of vertices is independent from any other pair of vertices. Let E be the event that D has no simple cycles and F_i be the event that D has no simple cycles of length i. Then $P[E] = P[\cap F_i] \leqslant P[F_3]$. The probability that any particular set of three vertices forms a simple cycle is $\frac{2}{64}$. By considering non-intersecting sets of three vertices from D, we can say $P[F_3] \leqslant \left(\frac{31}{32}\right)^{n/3}$. Therefore, as n gets large, the probability that a random digraph is a Euclidean CCCD goes to zero.

However, the family of labeled Euclidean CCCDs does grow at an exponential rate. Suppose there are $N(n)$ labeled Euclidean CCCDs on n vertices. A lower bound on the number of ways we can add a vertex to each of these $N(n)$ digraphs and avoid introducing any simple cycles is 3^n—consider adding either no arc, a bidirected arc or an arc directed into the new vertex for each vertex in the original digraph. In this manner we are guaranteed to not add any simple cycles, and therefore the resulting digraph will be a Euclidean CCCD on $n + 1$ vertices. Thus, $N(n + 1) \geqslant N(n) \cdot 3^n$.

4. Conclusion

We have shown that for any CCCD $D = (V, A)$, it is possible to find a set of $|V| = n$ target class points and one non-target class point in $\mathbb{R}^n$ which induce (via the L_1, L_2, or L_∞ metric) a digraph isomorphic to D. A natural complement to this result is to prove the *minimum* dimension necessary to embed a given CCCD. Another characterization of interest is to give conditions on the dissimilarity measure such that "no simple cycles" is a necessary and sufficient condition to be a CCCD.

Throughout this paper, we have presented the CCP in a deterministic manner. In the spirit of the application of the CCP to the construction of classifiers, there is an interest in the study of the CCP and CCCDs when the *random* sets X_+ and X_- drawn from some distributions F_+ and F_- are considered directly, as opposed to the observed sets $\mathscr{X}_+$ and $\mathscr{X}_-$ studied herein. In this random case, CCCDs can be seen as vertex random graphs [12]. The theorems presented in this paper represent an effort to gain some understanding of the class of classifiers available based on CCDs and the regularization provided by the CCP. As mentioned in Section 1, a classifier built using the CCP has complexity directly related to the size of the solution to the CCP or, equivalently, the size of the minimum dominating set (γ) in the random CCCD. The exact distribution for γ in a family of one-dimensional CCCDs is calculated in [17]. In the general case, Corollary 3 shows an upper bound on γ. In order to better understand the complexity characteristics of the classifier derived from CCP solutions, additional information about the distribution of γ is being sought.

Acknowledgments

The authors thank David J. Marchette and John C. Wierman for many valuable discussions. In addition, the authors thank anonymous referees for a thorough and thoughtful review of a previous version of this manuscript; this final version is improved thanks to these editorial efforts.

References

[1] D. Bertsimas, J. Tsitsiklis, Linear Otimization, Athena Scientific, 1997.
[2] K. Bogart, Introductory Combinatorics, Harcourt, Brace and Jovanovich, New York, 1990.
[3] A.H. Cannon, L.J. Cowen, Approximation algorithms for the class cover problem, Ann. of Math. Artificial Intelligence 40 (2004) 215–223.
[4] J. Conway, N. Sloane, Sphere Packings, Lattices and Groups, third ed., Springer-Verlag, New York, NY, 1999.
[5] T. Cox, A. Cox, Multidimensional Scaling, Chapman & Hall, CRC, London, Boca Raton, FL, 2001.
[6] F. Critchley, B. Fichet, Lecture Notes in Statistics: Classification and Dissimilarity Analysis, vol. 93, Springer-Verlag, New York, NY, 1994 (Chapter 2).
[7] L. Devroye, L. Gyorfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, Berlin, 1996.
[8] C.K. Eveland, D.A. Socolinsky, C.E. Priebe, D.J. Marchette, A hierarchical methodology for class detection problems with skewed priors, J. Classification 22 (2005) 17–48.
[9] T. Haynes, S. Hedetniemi, P. Slater, Fundamentals of Domination in Graphs, Marcel Dekker, Inc., New York, 1998.
[10] D. Hochbaum, (Ed.), Approximation Algorithms for NP-Hard Problems, PWS Publishing Co., Massachusetts, 1997.
[11] J.W. Jaromczyk, G.T. Toussaint, Relative neighborhood graphs and their relatives, Proc. IEEE 80 (9) (1992) 1502–1517.
[12] M. Karonski, E. Scheinerman, K. Signer-Cohen, On random intersection graphs: the subgraph problem, Combin. Probab. Comput. 8 (1999) 131–159.
[13] H. Maehara, A digraph represented by a family of boxes or spheres, J. Graph Theory 8 (1984) 431–439.
[14] D.J. Marchette, Random Graphs for Statistical Pattern Recognition, Wiley, 2004.
[15] D.J. Marchette, C.E.Priebe. Characterizing the scale dimension of a high dimensional classification problem, Pattern Recognition 36 (2003) 45–60.
[16] T. McKee, F. McMorris, Topics in Intersection Graph Theory, SIAM, Philadelphia, PA, 1999.
[17] C.E. Priebe, J. DeVinney, D. Marchette, On the distribution of the domination number for random class cover catch digraphs, Statist. Probab. Lett. 55 (3) (2001) 239–246.
[18] C.E. Priebe, D. Marchette, J. DeVinney, D. Socolinsky, Classification using class cover catch digraphs, J. Classification 20 (1) (2003) 3–23.
[19] C.E. Priebe, J.L. Solka, D.J. Marchette, B.T. Clark, Class cover catch digraphs for latent class discovery in gene expression monitoring by DNA microarrays, Comput. Statist. Data Anal. 43 (4) (2003) 621–632.
[20] D.B. Skalak, Prototype selection for composite nearest neighbor classifiers, Technical Report, University of Massachusetts, 1995.
[21] G.T. Toussaint, The relative neighborhood graph of a finite planar set, Pattern Recognition 12 (1980) 261–268.