A Hierarchical Methodology for Classification Problems with Skewed Priors

Carey E. Priebe
cep@jhu.edu

“The wealth of your practical experience with sane and interesting problems will give to mathematics a new direction and a new impetus.”
— Leopold Kronecker to Hermann von Helmholtz

March 25, 2004
JHU AM&S – p.1/35
Abstract

We describe an extension of the Class Cover Catch Digraph (CCCD) classifier, specifically designed for detection problems — two-class classification problems in which the natural priors on the classes are severely skewed. The emphasis of our approach is on computationally efficient classification for real-time applications. Our principal contribution is a boosted tree classifier built upon the CCCD structure. This classifier achieves performance comparable to that of the original CCCD classifier, but at drastically reduced computational expense. An analysis of classification performance and computational cost is performed using data from a face detection application. Comparisons are provided with Support Vector Machines (SVMs). These comparisons show that while some SVMs may achieve superior classification performance, their computational burden is so high as to make them unusable in real-time applications. On the other hand, our proposed classifiers combine high detection performance with extremely fast classification.
Abstract

We describe an extension of the Class Cover Catch Digraph (CCCD) classifier, specifically designed for detection problems — two-class classification problems in which the natural priors on the classes are severely skewed. The emphasis of our approach is on computationally efficient classification for real-time applications. Our principal contribution is a boosted tree classifier built upon the CCCD structure. This classifier achieves performance comparable to that of the original CCCD classifier, but at drastically reduced computational expense. An analysis of classification performance and computational cost is performed using data from a face detection application. Comparisons are provided with Support Vector Machines (SVMs). These comparisons show that while some SVMs may achieve superior classification performance, their computational burden is so high as to make them unusable in real-time applications. On the other hand, our proposed classifiers combine high detection performance with extremely fast classification.
Abstract

We describe an extension of the Class Cover Catch Digraph (CCCD) classifier, specifically designed for detection problems — two-class classification problems in which the natural priors on the classes are severely skewed. The emphasis of our approach is on computationally efficient classification for real-time applications. Our principal contribution is a boosted tree classifier built upon the CCCD structure. This classifier achieves performance comparable to that of the original CCCD classifier, but at drastically reduced computational expense. An analysis of classification performance and computational cost is performed using data from a face detection application. Comparisons are provided with Support Vector Machines (SVMs). These comparisons show that while some SVMs may achieve superior classification performance, their computational burden is so high as to make them unusable in real-time applications. On the other hand, our proposed classifiers combine high detection performance with extremely fast classification.
Abstract

We describe an extension of the Class Cover Catch Digraph (CCCD) classifier, specifically designed for detection problems — two-class classification problems in which the natural priors on the classes are severely skewed. The emphasis of our approach is on computationally efficient classification for real-time applications. Our principal contribution is a boosted tree classifier built upon the CCCD structure. This classifier achieves performance comparable to that of the original CCCD classifier, but at drastically reduced computational expense. An analysis of classification performance and computational cost is performed using data from a face detection application. Comparisons are provided with Support Vector Machines (SVMs). These comparisons show that while some SVMs may achieve superior classification performance, their computational burden is so high as to make them unusable in real-time applications. On the other hand, our proposed classifiers combine high detection performance with extremely fast classification.
Collaborators

Lenore Cowen
Adam Cannon
Dave Marchette
Jason DeVinney
Diego Socolinsky
Chris Eveland

Priebe, Marchette, DeVinney, and Socolinsky,
“Classification using class cover catch digraphs,”

Eveland, Socolinsky, Priebe, and Marchette,
“A Hierarchical Methodology for One-Class Problems
with Skewed Priors,” *Submitted for publication*.

<http://www.ams.jhu.edu/~priebe/cccddflassif.html>

Research partially supported by DARPA & ONR.
Collaborators

Lenore Cowen
Adam Cannon
Dave Marchette
Jason DeVinney
Diego Socolinsky
Chris Eveland

Priebe, Marchette, DeVinney, and Socolinsky,
“Classification using class cover catch digraphs,”

Eveland, Socolinsky, Priebe, and Marchette,
“A Hierarchical Methodology for One-Class Problems
with Skewed Priors,” Submitted for publication.

<http://www.ams.jhu.edu/~priebe/cccdclassif.html>

Research partially supported by DARPA & ONR.
Collaborators

Lenore Cowen
Adam Cannon
Dave Marchette
Jason DeVinney
Diego Socolinsky
Chris Eveland

Priebe, Marchette, DeVinney, and Socolinsky,
“Classification using class cover catch digraphs,”

Eveland, Socolinsky, Priebe, and Marchette,
“A Hierarchical Methodology for One-Class Problems
with Skewed Priors,” *Submitted for publication*.

<http://www.ams.jhu.edu/~priebe/cccddclassif.html>

Research partially supported by DARPA & ONR.
Collaborators

Lenore Cowen
Adam Cannon
Dave Marchette
Jason DeVinney
Diego Socolinsky
Chris Eveland

Pribe, Marchette, DeVinney, and Socolinsky,
“Classification using class cover catch digraphs,”

Eveland, Socolinsky, Pribe, and Marchette,
“A Hierarchical Methodology for One-Class Problems
with Skewed Priors,” *Submitted for publication*.

<http://www.ams.jhu.edu/~priebe/cccdclassif.html>

Research partially supported by DARPA & ONR.
# Collaborators

- Lenore Cowen
- Adam Cannon
- Dave Marchette
- **Jason DeVinney**
- Diego Socolinsky
- Chris Eveland

Priebe, Marchette, DeVinney, and Socolinsky,  
“Classification using class cover catch digraphs,”  

Eveland, Socolinsky, Priebe, and Marchette,  
“A Hierarchical Methodology for One-Class Problems with Skewed Priors,” *Submitted for publication*.

<http://www.ams.jhu.edu/~priebe/cccddclassif.html>

Research partially supported by DARPA & ONR.
Collaborators

Lenore Cowen
Adam Cannon
Dave Marchette
Jason DeVinney
Diego Socolinsky
Chris Eveland

Priebe, Marchette, DeVinney, and Socolinsky,
“Classification using class cover catch digraphs,”

Eveland, Socolinsky, Priebe, and Marchette,
“A Hierarchical Methodology for One-Class Problems with Skewed Priors,” *Submitted for publication*.

<http://www.ams.jhu.edu/~priebe/cccdclassif.html>

Research partially supported by DARPA & ONR.
Collaborators

Lenore Cowen
Adam Cannon
Dave Marchette
Jason DeVinney
Diego Socolinsky
Chris Eveland

Priebe, Marchette, DeVinney, and Socolinsky,
“Classification using class cover catch digraphs,”

Eveland, Socolinsky, Priebe, and Marchette,
“A Hierarchical Methodology for One-Class Problems
with Skewed Priors,” *Submitted for publication*.

<http://www.ams.jhu.edu/~priebe/cccdclassif.html>

Research partially supported by DARPA & ONR.
The Class Cover Problem (CCP)

Given training data $\mathcal{X}_0, \mathcal{X}_1 \subset \mathbb{R}^q$, our goal is to find a smallest collection of balls centered at observations in $\mathcal{X}_0$ such that every observation in $\mathcal{X}_0$ is in at least one of the balls and no observation in $\mathcal{X}_1$ is in any ball. That is, we want to find a smallest class cover.
The Class Cover Problem (CCP)

Given training data $X_0, X_1 \subseteq \mathbb{R}^q$, our goal is to find a smallest collection of balls centered at observations in $X_0$ such that every observation in $X_0$ is in at least one of the balls and no observation in $X_1$ is in any ball. That is, we want to find a smallest class cover.
The Class Cover Problem (CCP)

Given training data $\mathcal{X}_0, \mathcal{X}_1 \subset \mathbb{R}^q$, our goal is to find a smallest collection of balls centered at observations in $\mathcal{X}_0$ such that every observation in $\mathcal{X}_0$ is in at least one of the balls and no observation in $\mathcal{X}_1$ is in any ball. That is, we want to find a smallest class cover.
The Class Cover Problem (CCP)

Given training data $\mathcal{X}_0, \mathcal{X}_1 \subset \mathbb{R}^q$, our goal is to find a smallest collection of balls centered at observations in $\mathcal{X}_0$ such that every observation in $\mathcal{X}_0$ is in at least one of the balls and no observation in $\mathcal{X}_1$ is in any ball. That is, we want to find a smallest class cover.
The Class Cover Problem (CCP)

Given training data $\mathcal{X}_0, \mathcal{X}_1 \subset \mathbb{R}^q$, our goal is to find a smallest collection of balls centered at observations in $\mathcal{X}_0$ such that every observation in $\mathcal{X}_0$ is in at least one of the balls and no observation in $\mathcal{X}_1$ is in any ball. That is, we want to find a smallest class cover.
Consider a collection of sets \( \{B_1, B_2, \ldots, B_n\} \) with associated “base” points \( \{t_1, t_2, \ldots, t_n\} \).

We form the *catch digraph* \( D \) with \( V = \{v_1, v_2, \ldots, v_n\} \) and a directed edge from \( v_i \) to \( v_j \) iff \( t_j \in B_i \).

A *sphere digraph* is a catch digraph where the sets are spheres and the base point for each sphere is its center.
Catch Digraphs

Consider a collection of sets \( \{B_1, B_2, \ldots, B_n\} \) with associated “base” points \( \{t_1, t_2, \ldots, t_n\} \).

We form the catch digraph \( D \) with \( V = \{v_1, v_2, \ldots, v_n\} \) and a directed edge from \( v_i \) to \( v_j \) iff \( t_j \in B_i \).

A sphere digraph is a catch digraph where the sets are spheres and the base point for each sphere is its center.
Catch Digraphs

Consider a collection of sets \( \{B_1, B_2, \ldots, B_n\} \) with associated “base” points \( \{t_1, t_2, \ldots, t_n\} \).

We form the catch digraph \( D \) with \( V = \{v_1, v_2, \ldots, v_n\} \) and a directed edge from \( v_i \) to \( v_j \) iff \( t_j \in B_i \).

A sphere digraph is a catch digraph where the sets are spheres and the base point for each sphere is its center.
For any sets $\mathcal{X}_0, \mathcal{X}_1 \subset \mathbb{R}^q$ we can define the class cover catch digraph to be the catch digraph formed from the base points $x_i \in \mathcal{X}_0$ and the associated sets $B_i = \{ z \in \mathbb{R}^q : d(x_i, z) < d(x_i, \mathcal{X}_1) \}$. 
For any sets $\mathcal{X}_0, \mathcal{X}_1 \subset \mathbb{R}^q$ we can define the \textit{class cover catch digraph} to be the catch digraph formed from the base points $x_i \in \mathcal{X}_0$ and the associated sets $B_i = \{z \in \mathbb{R}^q : d(x_i, z) < d(x_i, \mathcal{X}_1)\}$.
Class Cover Catch Digraphs (CCCD)

For any sets $\mathcal{X}_0, \mathcal{X}_1 \subset \mathbb{R}^q$ we can define the \textit{class cover catch digraph} to be the catch digraph formed from the base points $x_i \in \mathcal{X}_0$ and the associated sets $B_i = \{ z \in \mathbb{R}^q : d(x_i, z) < d(x_i, \mathcal{X}_1) \}$. 
For any sets $\mathcal{X}_0, \mathcal{X}_1 \subset \mathbb{R}^q$ we can define the \textit{class cover catch digraph} to be the catch digraph formed from the base points $x_i \in \mathcal{X}_0$ and the associated sets $B_i = \{z \in \mathbb{R}^q : d(x_i, z) < d(x_i, \mathcal{X}_1)\}$.
Define a *dominating set*, $S$, of a digraph $D = (V, A)$ as follows: $S \subset V$ such that $\forall v \in V, \ v \in S$ or $\exists w \in S$ such that $(w, v) \in A$. 
Define a *dominating set*, $S$, of a digraph $D = (V, A)$ as follows: $S \subseteq V$ such that $\forall v \in V$, $v \in S$ or $\exists w \in S$ such that $(w, v) \in A$.

Solution to CCP $\Leftrightarrow$ Minimum Size Dominating Set in CCCD
Class Cover Catch Digraphs

class conditional data

\[ X_i | Y_i = j \sim f_j \]
Class Cover Catch Digraphs

For $X_i$ s.t. $Y_i = j$,

$$B_i = \{ x : d(X_i, x) < r_i := \min_{X_k : Y_k = 1 - j} d(X_i, X_k) \}$$
Class Cover Catch Digraphs

\[ V_j = \{X_i : Y_i = j\} \]

For \( i_1 \neq i_2 \), \((X_{i_1}, X_{i_2}) \in A_j \iff X_{i_2} \in B_{i_1} \).
Class Cover Catch Digraphs

\[ D_j = (V_j, A_j) \]
Class Cover Catch Digraphs

Choose a (minimum) dominating set $\hat{S}_j$ for $D_j$
Consider $\{B_i : X_i \in \hat{S}_j\}$
Class Cover Catch Digraphs

Go and do likewise for class $1 - j$
Class Cover Catch Digraphs

\[ g(z) = \arg \min_j \min \{x_i \in S_j\} \frac{d(z, X_i)}{r_i} \]
Class Cover Catch Digraphs

\[ g(z) = \arg \min_j \min_{\{X_i \in \hat{S}_j\}} d(z, X_i) / r_i \]
Class Cover Catch Digraphs

\( \hat{L}(\text{nearest neighbor}) = 0.123 ; \hat{L}(CCC D) = 0.074 ; L^* = 0.035 \)

\( (L(g) := P[g(X) \neq Y] \) is “probability of misclassification”\)
Theorem 1:

Let $\hat{S}_j$ be dominating sets for CCCDs $D_j$ and $g(z) = \arg \min_j \min_{\{X_i \in \hat{S}_j\}} d(z, X_i)/r_i$.

Then $\hat{L}_n^{(R)}(g) = \sum_{i=1}^{n} I\{g(X_i) \neq Y_i\} = 0$. 
Theorem 2:
Assume, in addition to the conditions of Theorem 1, that $d$ is “well-behaved” (e.g. $L_p$) and the class-conditional distributions $F_j$ are strictly separable. Then $g$ is consistent.
That is, $L_n(g) \to L^* := L(Bayes\ optimal)$. 
Algorithmic extension: robustness
(a) to contamination
(b) to outliers
\[ \alpha = 0; \beta = 0 ; \hat{L} \approx 0.21 \]

\[ \alpha = 10; \beta = 5 ; \hat{L} \approx 0.16 \]
a random walk ...
For each point $x \in \mathcal{X}_j$, we consider the random walk

$$R_x(r) = |\{x' \in \mathcal{X}_j : d(x, x') < r\}| - |\{x' \in \mathcal{X}_{(1-j)} : d(x, x') < r\}|,$$

for $r \in \mathbb{R}_+$, $j \in \{0, 1\}$. This random walk can be thought of as a ball growing around the point $x$. As it reaches an observation, the walk takes a step either up or down depending on the class of that observation.

Then

$$r_x^* = \arg \max_r R_x(r) - P(r),$$

where $P(\cdot)$ is a penalty function that biases the choice of radius.
The choice of prototypes for each class proceeds in a greedy fashion, using

\[ T_x = R_x(r^*_x) - P(r^*_x) \]

as a surrogate criterion for the classifier performance.

For a given class, the first prototype \( c_1 \) is chosen to be that with the highest value of \( T \). All training observations \( x \) for which \( d(x, c_1) < r^*_c \) are then deleted from the training set, all radii \( r^*_x \) are recomputed for the remaining training observations, and the next prototype is chosen using the surrogate criterion \( T \) as before. This process continues until all but a predetermined proportion of the class-\( j \) training data has been deleted.
Consider modifying the radii via

\[ \tilde{r}_i = t r_i \quad \text{for} \quad x_i \in \mathcal{X}_0, \]
\[ \tilde{r}_i = t^{-1} r_i \quad \text{for} \quad x_i \in \mathcal{X}_1, \]

for \( 0 < t < \infty \).

Values of \( t \) in \((0, 1)\) favor lower class-1 error at the expense of higher error rate on class-0.

Geometrically, the value of \( t \) has the effect of shrinking or growing the estimated support of each class.
a CCCD tree

Sequential CCCD structure with three stages of three sub-stages each.
example prototypes

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Example Prototype 1" /></td>
<td><img src="image2.png" alt="Example Prototype 2" /></td>
<td><img src="image3.png" alt="Example Prototype 3" /></td>
</tr>
</tbody>
</table>

Schematic representation of a sample CCCD tree stage. In this case, there is a single target-class prototype and three sub-stages identified by distinct non-target prototypes.
Sample non-face training data before and after the first boosting stage.
**how it works?**

Sequential refinement of the discriminant region through stages and sub-stages. Each row corresponds to a stage and each column to a sub-stage therein.
Example Result!
Clusters found by the CCCD classifier while training a boosted tree classifier. Each image represents the mean face within that cluster. While the clusters are chosen in an effort to maximize classification performance, they do correspond to major image classes that are intuitively expected. Frontal, left and right illumination are the dominant clusters.
ROC curves for a full Gaussian SVM, reduced Gaussian SVM and boosted CCCD tree, all trained on the same data.
“The wealth of your practical experience with sane and interesting problems will give to mathematics a new direction and a new impetus.”

— Leopold Kronecker to Hermann von Helmholtz