Class Cover Digraphs for Latent Class Discovery

J. L. Solka
Code B10
NSWCDD
Dahlgren, VA 22448

C. E. Priebe
Mathematical Sciences Department
JHU
Baltimore, MD 21218

D. J. Marchette
B10
NSWCDD
Dahlgren, VA 22448

Abstract

This paper examines the robustness of a new graph theoretic method of latent class discovery. This new method allows for the discovery of new latent classes within sets of observations residing in a high dimensional space. The robustness of the methodology is studied using a single gene expression data set.

Keywords

latent class discovery, gene expression data, statistical pattern recognition, graph theory, clustering

Introduction

One is often presented with a set of observations from various labeled classes for construction of a classifier. For the discussions within a classifier can be thought of as a mapping from our set of observations, residing in $\mathcal{R}^n$ to a set of class labels $1, 2, \cdots, j$. There are numerous ways that one may construct such a mapping. Some of the ways are based on estimating the unknown distributions of the observations that make up each of the individual classes while other methods merely require one to be able to estimate the discriminant boundary that separates the sets of observations.

Other times one is presented with a set of unlabeled observations and one is interested in identifying structural clusters within the group of observations. In this case one is not provided with any sort of class labels and one is interested in clustering the observations based on some other criteria. As can be imagined there are numerous criteria that one might use in order to cluster the observations or evaluate a clustering after the clustering has been obtained.

A more interesting case, which is the focus of the discussions within, is the case where we are provided labeled observations, for which we wish to create a classifier, but we are also interested in the identification of clusters or latent classes that reside within the labeled observations. In particular we are interested in the identification of said clusters based on the relationship of the observations to the discriminant boundary.

This is the focus of our discussions within. We examine the robustness of a graph theoretic method of latent class discovery. This method is essentially based on a clustering of observations obtained using a graph thoretic surrogate for their relationship to the discriminant boundary.

The paper is laid out as follows. First we provide the reader with some background material on the latent class discovery problem and our previous endeavors in this area. This section will also detail some of the interesting questions, such as the robustness of the discovered latent class structure to the derived classifier. In the next or results section, we will illustrate the application of this methodology to a gene expression dataset. In this case the identified latent classes correspond to previously identified disease types. We will also present some preliminary results in this section that attempt to characterize the robustness of the discovered latent class structure to the particular classifier that was employed. Finally we wrap up the paper with some discusses that illuminate our planned future activities. As with many of the papers that we write, we hope that the reader will be inspired to go forth and investigate some of the new strategies that might be sparked by the discussions outlined below.

Background

Given a set of possibly hyperdimensional observations we are interested in identifying any latent classes, from a discriminant analysis standpoint, that reside within the set of class labeled observations. The first step in the process is the construction of a graph theoretic classifier.This classifier construction methodology is predicated on the existence of a metric or pseudometric that measures the distances between the observations. Our previous work has illustrated how the choice of this distance measure may markedly effect the performance of the classifier but that is not the focus of our discussions within [Priebe et al., 2000].The obtained models are then subjected to the latent class discovery process. The obtained latent classes are then analyzed using multidimensional scaling or nonlinear dimensionality reduction methods. Figure 1 provides a flowchart of our overall research strategy.

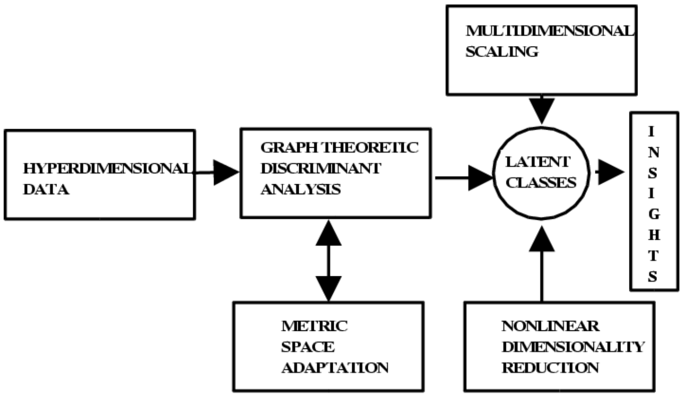

Figure 1: A flowchart detailing our latent class discovery research strategy.

One of the classification methods that can be employed in the latent class discovery process is the class cover catch digraph (CCCD) method of Priebe [Priebe et al., 2003a]. The construction of this classifier begins by placing a ball about each of one of the class's observations. One then obtains a subset of the balls that cover the observations through a greedy based algorithm. If one considers each of the observations as a vertex in a graph and then produces a directed graph by drawing a directed edge from vertex a to vertex b if the ball centered on vertex a covers vertex b then the construction of this reduced cover can be cast as the solution of a dominating set. The reader is referred to [Chartrand and Lesniak, 1996] for a full discussion of domination in graphs.

Given this reduced set of balls that cover one of the class's observations the latent class discovery process proceeds as follows. We cluster the balls based on their radii. The exact manner in which the number of ball clusters is determined is discussed later. Once the balls are clustered then the discovered latent classes consist of those observations residing within a particular cluster of balls. In Figure 2 we illustrate the latent class discovery process using a toy problem. We have colored each ball based on the latent class that it belongs to and hence there are 4 latent classes. The observations are colored according to their original two classes and the CCCD solution that was used during the latent class discovery process originally covered the red observations.

We note that the proposed latent class discovery process is clustering observations based on their relationship to the discriminant boundary. This clustering could of course be accomplished using any sort of classifier that produces a discriminant boundary that allows one to measure distances to the discriminant boundary and subsequently cluster the observations based on this distance. Figure 3 illustrates this idea using a quadratic classifier.

Results

We now illustrate our proposed latent class discovery process on an example gene expression data set. We have chosen to use an ALL/AML data set first proposed by Golub [Golub and Slonin, 1999]. This particular study investigated the ability to distinguish between two forms or leukemia, ALL and AML, and measure the responses of roughly 7000 genes on 72 patients using an Affymetrix microarray system. Figure 4 illustrates our strategy for the application of our latent class discovery process to the gene expression data.

We will now discuss the determination of the number of clusters during the latent class discovery process. We first define the estimated error rate of our CCCD-based classifier as follows. For each candidate nuumber of clusters $k = 1, \cdots, \widehat{\gamma}$ an empirical risk (resubstitution error rate estimate) $\widehat{L}_k$ is calculated as

$$\begin{aligned}\widehat{L}_k \quad := \quad & (1/(n+m))(\sum_{i=1}^{n} I\{x_i \notin \cup_{j=1,\cdots,k} \cup_{v\in\widehat{S}_j} B(v, \min_{w\in\widehat{S}_j} r_w)\} \\ & + \sum_{i=1}^{m} I\{y_i \in \cup_{j=1,\cdots,k} \cup_{v\in\widehat{S}_j} B(v, \min_{w\in\widehat{S}_j} r_w)\})\end{aligned}$$

We proceed by defining the "scale dimension" $\widehat{d^\star}$ to be the cluster map dimension that minimizes a dimensionality-penalized empirical risk; $\widehat{d^\star_\delta} := \min\{\arg\min_k \widehat{L}_k + \delta \cdot k\}$ for some penalty coefficient $\delta \in [0,1]$. This scale dimension determines the number of clusters to be used during the latent class discovery process.

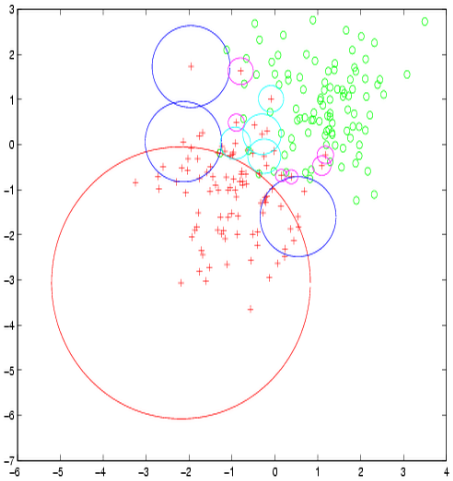

Figure 2: CCCD-based latent class discovery. Ball color indicates ball cluster membership. The CCCD solution was created based on the red observations.

Figure 5 presents performance as a function of scale space dimension for the gene expression data. A visual inspection of the plot indicates a scale space dimension, indicated by the abscissa of the elbow of the plot, of 5.

The latent classes discovered in the gene expression data correspond to the well known B-cell and T-cell ALL subtypes. These latent classes were first discovered using a custom in-house developed visualization framework known as the Interactive Hyperspectral Exploratory Data Analysis Tool (IHEDAT). It turns out that the measured distance from the ALL B-cell to the AML observations differs from the measured distance from ALL T-cell observations to the AML observations. In fact the ALL T-cell observations are in general more distant from the AML observations than the ALL B-cell observations. This distance is apparent in Figure 6 where we plot ALL and AML observations in the MDS projection space. The reader is referred to [Priebe et al., 2003b] for a more detailed description of how the original latent class discovery was performed.The reader is referred to [Solka et al., 2002] for an

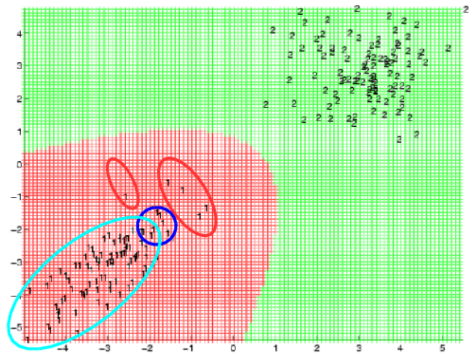

Figure 3: Quadratic classifier based latent class discovery. Circles indicate clusters of observations based on the distance to the quadratic discriminant boundary.

in-depth treatment of IHEDAT.

After the initial success of our latent class discovery methodology, we began to wonder how robust the procedure was to other possible CCCD coverings. In order to answer this question one of us, DJM, developed a method to enumerate all possible coverings of the ALL observations. Some of the these possible solutions correspond to greedy solutions while others do not. There are 180 21 node, ball, solutions. Sixteen of the nodes remain fixed across the solutions. There are 14 greedy solutions. In Figure 7 we plot scale dimension as a function of the various solutions.The red symbols indicate the locations of the greedy solutions. The green symbol indicates the location of the previous solution used to discover the T/B subclass.

In Figure 8 we present a histogram of the scale dimension across all solutions. We note that the value of 5 obtained in our original analysis was a little high as compared with the perceived mean/mode of this distribution. It is an open question as to how the choice of the scale dimension would effect the latent class discovery process.

In Figure 9 we present the dominating sets for each vertex.The triangles at the top of the plot indicate the 16 vertices that appear as part of all 180 solutions. For each of the other vertices we plot the number of times that the vertex appears in one of the covers. We have also used color to indicate whether the ball that corresponds to this vertex is centered on an ALL B-cell, blue, or ALL T-cell, red, observation. We note that only one T-cell vertex is in the set of 16 that does not change.

We may also analyze the variety of coverings through a characterization of the graphs that make up the coverings. In Figure 10 we present the unique induced subgraphs for the 5 changing vertices of the 180 dominating sets (top) and the unique induced subgraphs in the 5 changing vertices of the 14 greedy dominating sets (bottom).We note that the greedy based graphs are a proper subset of those graphs obtained based on the 5 changing vertices of the 180 dominating sets.

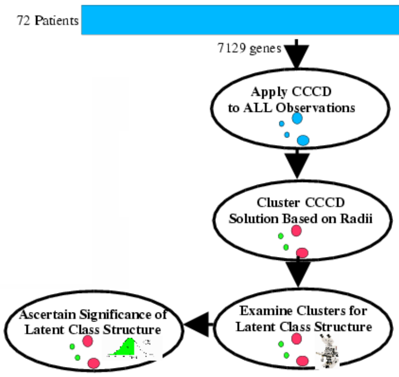

Figure 4: Latent class discovery strategy for the Golub gene expression data.

The analysis, with regard to the robustness of the procedure, presented so far in the paper has been interesting but not particularly compelling or even necessarily relevant to ascertaining whether the latent class structure identified during the original analysis would be identified if one were to use a different covering. The original greedy solution contained 3 clusters of balls that only contained ALL B-cell observations and 1 cluster of balls that contained 8/9 of the ALL T-cell observations. One way to evaluate the other coverings would be through the use of a figure of merit that captured the information contained within this first solution. We have chosen to use a figure of merit consisting of the percentage of B points that are in pure B clusters and the highest percentage of T points in any one cluster. In Figure 11 we present the calculated figures of merit for each solution. We note that all of the greedy folutions contain eight ninths of the T points in one cluster. We also note that .4 or more of the B points are in pure B clusters. It is a little hard to discern the exact nature of the solutions, based on ther plot, due to overplotting but it is reasonable to infer that one may have been able to idnetify the B-cell T-cell distinction utilizing any of a number of the other greedy solutions.

Conclusions

We have discussed a new method of latent class discovery that is appropriate for application to hyperdimensional data sets. This method allows one to discover previously unidentified classes within a class based on the relationship of the observations to the discriminant boundary. We have illustrated the application of this methodology to one gene expression data set. We have also presented some preliminary results that attempt to quantify the robustness of this method with regards to the dominating set that was used to facilitate the latent class discovery process. We have performed some rudimentary exploratory data analysis on the enumerated dominating sets. We have presented a strategy that we have developed to study the robustness of the latent class discovery process to choice of dominating set. This

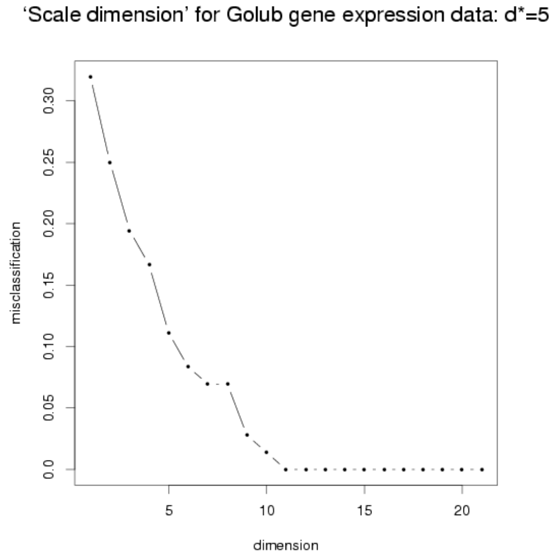

Figure 5: Cross validated performance as a function of scale dimension for the ALL Golub gene expression data with respect to the AML data.

problem is a very difficult one and will continue to be the subject of our ongoing research efforts.

Acknowledgments

The work of the first author, JLS, was sponsored by the NSWCDD ILIR Program. The work of the second author, CEP, was partially supported by the Office of Naval Research Grant N00014-01-1-0011 and the Defense Advanced Research Projects Agency Grant F49620-01-1-0395, and the work of the third author, DJM, was sponsored by the NSWCDD ILIR Program.

References

[Chartrand and Lesniak, 1996] Chartrand, G. and Lesniak, L. (1996). *Graphs and Digraphs*. Chapman and Hall/CRC.

[Golub and Slonin, 1999] Golub, T. R. and Slonin, D. K. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.

[Priebe et al., 2000] Priebe, C. E., J., M. D., and Solka, J. L. (2000). On the selection of distance for a high dimensional cla ssification problem. *ASA Proceedings of the Sections on Statistica and Statistical Graphics*, (58–63).

[Priebe et al., 2003a] Priebe, C. E., Marchette, D. J., DeVinney, J., and Scolinsky, D. (2003a). Classification using catch cover digraphs. *to appear Journal of Classification*.

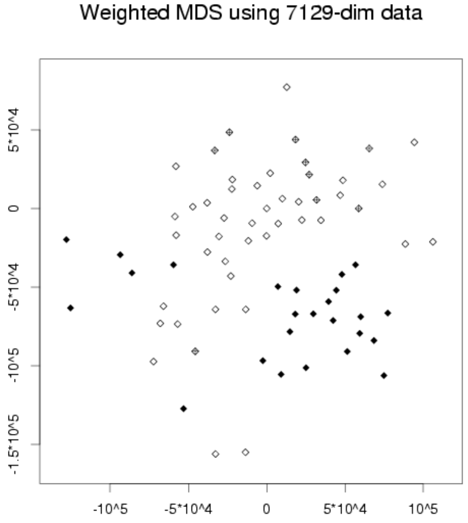

Figure 6: ALL, T and B cell, and AML observations in the MDS derived space. T-cell observations are cross hatched diamonds, B-cell observations are open diamonds, and AML observations are filled diamonds.

[Priebe et al., 2003b] Priebe, C. E., Solka, J. L., Marchette, D. J., and Clark, B. T. (2003+b). Class cover catch digraphs for latent class discovery in gene expression monitoring by dna microarrays. *to appear the Special Issue, of Computational Statistics and Data Analysis on Statistical Visualization.*

[Solka et al., 2002] Solka, J. L., Priebe, C. E., and Clark, B. T. (2002). A visualization framework for the analysis of hyperdimensional data. *International Journal of Image and Graphics*, 2(1):145–161.

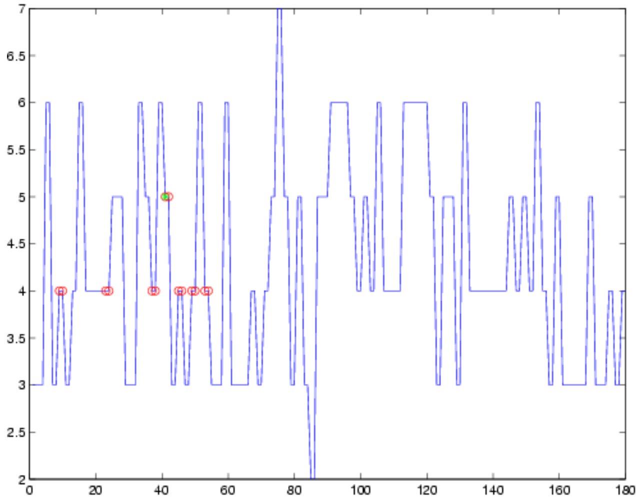

Figure 7: Scale dimension plotted as a function of the various Golub ALL dominating sets with respect to the AML observations. Red symbols indicate greedy solutions. The green symbol indicates our original solution used to make the T/B cell discovery.

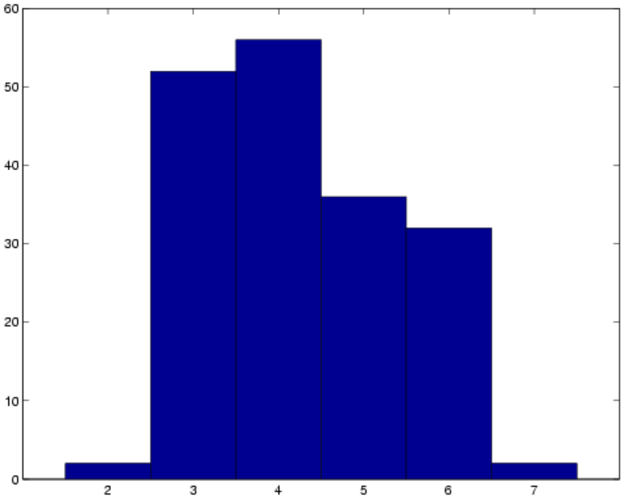

Figure 8: Histogram of the scale dimensions obtained with all 180 Golub ALL dominating sets.

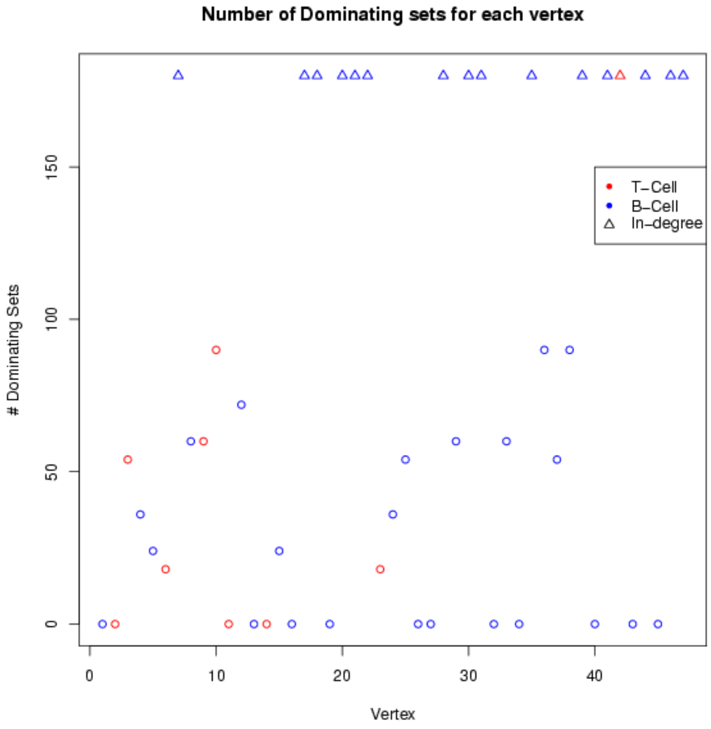

Figure 9: Dominating sets for each vertex in the Golub ALL data. Triangles represent the 16 vertices that are members of all 180 solutions. Color indicates T -cell or B-cell membership.

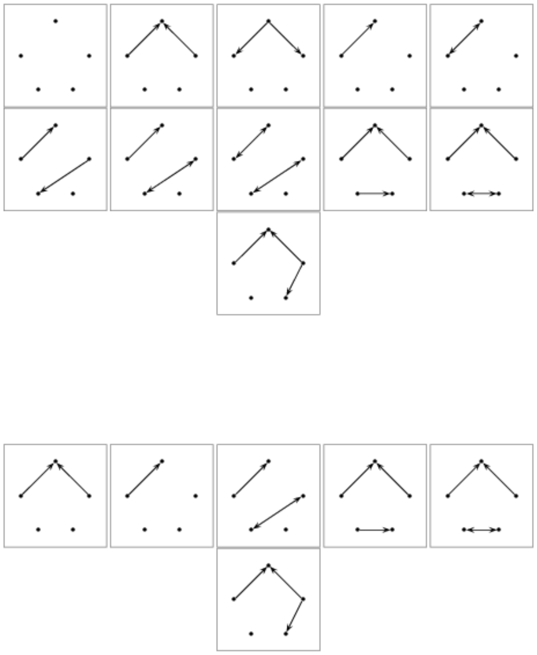

Figure 10: Unique induced subgraphs in the 5 changing vertices of the 180 dominating sets for the Golub ALL data with respect to the Golub AML data (top) and the unique induced subgraphs in the 5 changing vertices of the 14 greedy dominating sets for the Golub ALL data with respect to the Golub AML data (bottom).

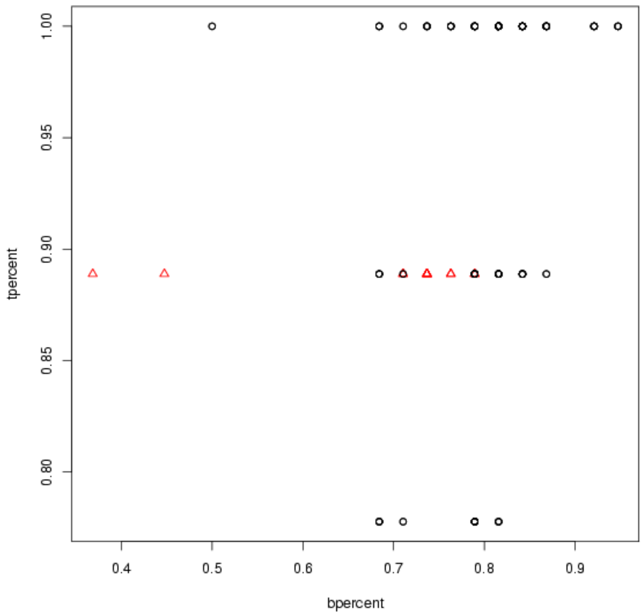

Figure 11: Proportion of B points that are in pure B clusters vs. highest proportion of T points in any one cluster for the 180 dominating sets of the Golub ALL data with respect to the Golub AML data. Red triangles indicate greedy solutions.