

# Inducing Classification Rules for Public Health Data

John W. Sheppard  
Department of Computer Science  
The Johns Hopkins University  
Baltimore, Maryland 21218  
sheppard@cs.jhu.edu

## Abstract

Analyzing data with the intent of inducing classification rules typically proceeds from a set of training data in which classifications are known. In the event classifications are unknown, algorithms exist for performing unsupervised learning to determine concept classes inherent in the data. In this paper, we describe experiments applying multiple learning strategies for classifying unlabeled data. Specifically, three unsupervised learning algorithms were applied to a large set of public health data in order to determine likely concept classes for the data based on the inherent features in the data. After inducing the concept classes, the data were processed by a decision tree algorithm in order to determine more efficient classification rules under the assumption that the concepts induced during unsupervised learning were correct.

**Key words:** Classification, unsupervised learning, clustering, decision trees

## 1. Introduction

The machine learning literature describes several approaches for classifying numerical

data. For example, decision trees (such as those generated by Quinlan's ID3 and C4 algorithms) select attributes as internal test nodes of a tree to determine the class to which a data point belongs, given at the leaves of the tree (Quinlan, 1986). Nearest neighbor algorithms store training examples paired with a classification (Aha *et al.*, 1991). When a new point is presented, the stored point that is *closest* in some sense (such as Euclidean distance or Hamming distance) is selected and the corresponding classification reported.

At times, labels providing classification information are not available with the training set. In these instances, unsupervised learning approaches may be employed to detect *clusters* of the data. These clusters can then be used to develop an initial set of classification labels (albeit non-symbolic) for the data.

In this paper, we will describe applying multiple learning strategies to a large set of psychiatric data (Eaton and Ritter, 1988; Eaton *et al.*, 1989). Specifically, we will compare three clustering algorithms and discuss the results of processing resultant clusters with a decision tree algorithm to provide an efficient classification strategy. The psychiatric data, provided by the Johns Hopkins School of Public Health, consists of over 7,000 data

points describing patients with respect to clinical depression or anxiety. Each data point has 58 fields indicating, for example, whether a patient has various fears, feelings of worthlessness, thoughts of suicide, etc. Our experiments used 20 binary attributes from the 58 provided. According to the School of Public Health, these 20 attributes characterize depression where the others provide demographic information and characterize anxiety. Note that none of the data, as of yet, have been classified (i.e., labels are not known *a priori*), thus motivating the analysis of unsupervised learning techniques.

The three clustering algorithms examined include a non-hierarchical approach, a hierarchical approach (thus resulting in a decision tree), and a connectionist approach. The nonhierarchical approach is based on a variation of MacQueen's *k*-means method (MacQueen, 1967). The standard *k*-means method assumes *k* clusters and fits the data in the clusters with the nearest centroids. The variation of this method used permits *k* to vary so that an estimate of the number of classes in the data may be determined.

The second cluster analysis approach is hierarchical. Hierarchical approaches either divide data or combine data in a tree structure. Divisive approaches begin with one large cluster and divide into smaller clusters based on the attributes. Agglomerative approaches begin with one cluster for each training sample and combine clusters based on similarity. The approach used in this part of the study is a divisive approach called *association analysis*. This approach selects an attribute to divide clusters by computing a matrix of chi-square coefficients for each attribute and selecting the coefficient with the maximum sum of chi-square values (Everitt, 1974).

Finally, Rumelhart and Zipser (1986) describe a connectionist approach to clustering using

competitive learning. The approach proceeds under the assumption that dominant attributes will generally determine the classification, and the network reinforces detection of the dominant attributes by strengthening weights associated with their corresponding input nodes. The output layer then applies a winner-take-all competition strategy to determine the cluster to which a data point belongs.

Since the experimental data used was not provided with classification labels, the second phase of the study consists of generating decision trees based on the classifications derived from the clustering techniques. Quinlan's C4 algorithm is applied to the results of all three clustering techniques, and the resulting trees compared to rules that can be derived from the clustering algorithms themselves.

## 2. Inducing Concept Classes Using Unsupervised Learning

There are many ways to characterize machine learning algorithms. One approach is based upon whether or not an external "teacher" exists. The two resulting types of learning algorithms are referred to as *supervised* learning and *unsupervised* learning. Typically, supervised learning proceeds when the results of some action are analyzed by a critic in comparison with known or expected results. Discrepancies between the two are used to determine ways to modify internal representations of the data so as to improve performance.

Unsupervised learning, on the other hand, does not have the advantage of an external teacher to determine "appropriate" behavior or "correct" classifications. Rather, data are examined and organized in such a way as to identify internal consistency. The class of cluster analysis algorithms generally fall

within the set of unsupervised learning algorithms. In the following sections, we will describe the details of the three unsupervised learning algorithms used in this study.

## 2.1 Clustering by $k$ -means

The first technique for clustering fits within the class of non-hierarchical techniques. Non-hierarchical clustering begins by selecting an initial set of clusters and alters the partitions so as to improve some metric. For example, *nearest centroid* methods attempt to develop partitions such that classification is made by comparing a point to the centroids of the clusters. The class corresponding to the nearest centroid is the one identified for that data point.

One of the most common approaches to non-hierarchical clustering is MacQueen's  $k$ -means algorithm (MacQueen, 1967). The  $k$ -means algorithm attempts to determine the  $k$  best clusters for a set of data such that classification is made by finding the cluster with the nearest Euclidean distance. Recall that the Euclidean distance between two points is computed as follows:

$$\text{dist}(p1, p2) = \sqrt{\sum_{v_i} (x1_i - x2_i)^2}$$

where  $x1_i$  is the  $i$ th attribute of point  $p1$  and  $x2_i$  is the  $i$ th attribute of point  $p2$ . Since all of the attributes in the data set are binary, distance reduces to the square root of the Hamming distance.

The basic  $k$ -means algorithm consists of the following steps:

1. Select the first  $k$  data points as initial clusters with one member in each cluster.

2. Assign the remaining  $m - k$  data points to the cluster with the nearest centroid.
3. After assigning each point, recompute the corresponding centroid of the cluster with the new point.
4. After all of the data points have been assigned, use the  $k$  clusters as seed points and pass through the data one more time for a final classification.

Variations of this algorithm exist in which the clusters converge to improved clusters. These variants require several passes through the data, but the law of diminishing returns may be experienced fairly early in the process.

Unfortunately, for our purposes, the basic  $k$ -means algorithm has a more serious drawback. This algorithm assumes the number of clusters is known and force fits all of the data into exactly  $k$  clusters. For this reason, MacQueen also proposed a variant in which the number of clusters is not known. This algorithm is the one selected for this study and is composed of the following steps:

1. Select values for an initial  $k$  and two additional parameters,  $C$  (coarsening) and  $R$  (refining).
2. As in the basic  $k$ -means algorithm, select the first  $k$  data points as the initial clusters.
3. Compute all of the pairwise distances between each of the cluster centroids. If the smallest distance is less than  $C$ , then merge the two corresponding clusters and recompute the corresponding centroid. Continue merging until no other merges occur.
4. Assign the remaining  $m - k$  data points one at a time to the cluster with the nearest centroid. If the distance to the nearest centroid is greater than  $R$ , then consider the point a new cluster and goto step 3.

5. After all of the data points have been assigned, use the cluster centroids as seed points and pass through the data one last time assigning the points to the clusters with the nearest centroids.

This algorithm can also follow convergent approaches, and as before, it has been found that diminishing returns exhibit themselves early in the process.

## 2.2 An associative clustering algorithm

For the second clustering technique, a hierarchical approach was used. Hierarchical clustering produces a decision tree by which data points can be classified according to the determined clusters. In general, hierarchical clustering is either *divisive* or *agglomerative*. Agglomerative approaches proceed with each data point treated as individual clusters. Clusters are then combined to form higher level clusters. This process continues until a group of high level clusters (or a single cluster) is identified. Divisive approaches begin with a single cluster and divide the cluster into sub-clusters. This process continues recursively until base clusters are determined.

In addition, hierarchical approaches can be classified as *monothetic* or *polythetic*. Monothetic techniques attempt to cluster according to single attributes where polythetic techniques cluster according the values of all of the attributes.

The technique used in this part of the study is a monothetic, divisive cluster analysis algorithm called *association analysis* (Everitt, 1974). Association analysis divides clusters by selecting the single attribute that provides the “best” split. The concept of a best split has been defined in several ways. For example, decision tree algorithms frequently employ concepts from Shannon’s information

theory to select the attribute that provides the most information independent of the actual values of the attributes (Shannon, 1948).

Association analysis selects attributes that maximize the chi-square coefficients of the data. Recall that chi-squared is computed as follows:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

where  $s^2$  is that sample variance,  $\sigma^2$  is the population variance, and  $n$  is the sample size.

For association analysis, we assume all of the attributes are binary. The computation of the chi-square coefficients on binary data is similar to the standard equation. Let  $attrib_{ij}$  be the  $j$ th attribute of the  $i$ th data point and  $attrib_{ik}$  be the  $k$ th attribute of the  $i$ th data point. Then

$$a_{jk} = \sum_{\forall i} attrib_{ij} * attrib_{ik}$$

$$b_{jk} = \sum_{\forall i} (1 - attrib_{ij}) * attrib_{ik}$$

$$c_{jk} = \sum_{\forall i} attrib_{ij} * (1 - attrib_{ik})$$

$$d_{jk} = \sum_{\forall i} (1 - attrib_{ij}) * (1 - attrib_{ik})$$

Then the chi-square coefficients are simply computed as

$$\chi_{jk}^2 = \frac{(ad - bc)^2 n}{(a+b)(a+c)(b+d)(c+d)}$$

and the attribute is selected such that  $\left\{ \sum_{j,k} z_{jk}^2 \right\}$  is maximized.

### 2.3 Clustering by competitive learning

For the final clustering technique, a connectionist algorithm was selected. In particular, the competitive learning neural network described by Rumelhart and Zipser (1986) was implemented. (Note that variants on this network are described by von der Malsburg (1973) and Grossberg (1987)) The idea behind competitive learning is that the network develops a set of “feature detectors.” When data containing a learned feature are submitted to the network, then the activity of the network identifies which feature is present. To identify features, nodes within the network “compete” among themselves to respond to the stimulus pattern. The node that wins the competition has the feature associated with it. Consequently, when that node becomes active, the feature has been identified.

In order to train a competitive learning network, the weight matrix is constructed with  $m$  rows and  $n$  columns, where  $m$  = the number of output nodes and  $n$  = the number of input nodes. The weight matrix is initialized with the following:

$$w_{ij} = \frac{1}{n_o} + \delta; \quad \forall i, j$$

where  $n_o$  is the number of nodes at the input layer and  $\delta$  is a small random number generated for each weight.

The network is trained by processing a set of training data. Then, for each output node in the network, and for each training case, a “winning” node is determined. This winner

is used to determine which node’s weights are to be updated. The winner is determined as follows:

$$winner = \underset{j}{\text{max}} \left\{ \sum_{i=0}^{n_o-1} w_{ji} I_i \right\}$$

where  $w_{ji}$  is the value in the weight matrix corresponding to row  $j$  and column  $i$ ,  $I_i$  is the activation value of input node  $i$ , and  $j$  ranges over the number of outputs.

The competitive learning rule is then applied to the winner for the given training instance. In other words, the weights in the weight matrix are only modified for the connections between the input nodes and the winning output node. The update rule for modifying the weights in the network is as follows:

$$\Delta w_{ji} = \eta \left\{ \frac{I_i}{\left\{ \sum_{k=1}^{n_o} I_k \right\}} - w_{ji} \right\}$$

where  $\Delta w_{ji}$  is the change in the weight matrix and  $\eta$  is a learning factor.

The clusters are identified by winning nodes when a data point is presented to the network. A further analysis of the network can help to identify the attributes that are most significant in clustering the data. In particular, since the update rule “strengthens” the connections between winning nodes and the significant inputs (i.e., attributes), the strong attributes for a given class will have weights greater than  $1/n_o$ .

### 3. Inducing Decision Trees

The three clustering algorithms described in the previous sections provide approaches to labeling data not previously labeled for

classification. Once labels have been assigned, the next step is to determine efficient and effective means for classifying data according to the concepts learned that have not previously been encountered. One approach for such concept learning is the induction of decision trees. Perhaps the most famous decision tree algorithm is ID3 and its successor C4, both developed by Quinlan (1986).

ID3 and C4 allow attributes to be multi-valued (i.e., they do not limit attributes to binary values) and construct classification trees by selecting attributes that provide the best split among the data according to known classifications. The resulting tree is then used to classify data including data not used in training. The rules generated for the decision tree then permit classification to generalize so as to classify new data. Of course, since we do not know what the correct classifications are for our experiments, it is difficult to determine how well the trees generalize. (Note that C4, the program used in these experiments, automatically constructs trees on a subset of the training data using ten-way cross-validation and selects a tree that generalizes the best on the remaining data.)

In order for ID3 and C4 to determine the best attribute at a given point, Quinlan incorporated the information entropy function described by Shannon (1948). The information value of a set of data  $\mathbf{T}$  is

$$I(\mathbf{T}) = - \sum_{i=1}^{|\mathbf{C}|} \frac{\text{freq}(c_i, \mathbf{T})}{|\mathbf{T}|} \log_2 \frac{\text{freq}(c_i, \mathbf{T})}{|\mathbf{T}|}$$

where  $\mathbf{C}$  is the set of classes,  $\mathbf{T}$  is the set of training instances, and  $\text{freq}(c_i, \mathbf{T})$  is the frequency of class  $i$  occurring in  $\mathbf{T}$ . The expected information value of  $\text{attrib}_j$  is

$$E(\text{attrib}_j) = \sum_{i=1}^{v_j} \frac{|T_i|}{|\mathbf{T}|} I(T_i)$$

where  $v_j$  is the number of values  $\text{attrib}_j$  can have and  $T_i$  is the subset of  $\mathbf{T}$  with  $\text{attrib}_j$  having the  $i^{\text{th}}$  value. Then the information gain is simply  $I(\mathbf{T}) - E(\text{attrib}_j)$ . The attribute with the maximum gain is selected for the root of the current subtree. C4 adds several techniques for pruning the trees, thus making the final trees more efficient than the initial ones (Quinlan, 1987). Also, C4 applies a gain ratio criterion for its splitting criterion, but when all attributes are binary, the result is identical to applying information gain.

#### 4. The Public Health Data

For this study, psychiatric data on anxiety and depression were analyzed. This data set consisted of over 7,000 samples with 58 binary attributes. The data set was collected from the East Baltimore Epidemiologic Catchment Area (ECA) Program and was supplied by the Johns Hopkins School of Public Health (Eaton and Ritter, 1988; Eaton *et al.*, 1989). The data was not categorized prior to analysis, so the object of the study was to identify regularities within the data that might suggest natural classifications.

For this study, the data set was reduced in three ways. First, several of the samples had attributes with unknown values. All samples with more than five unknown attributes were eliminated from the data set. Second, since all of the attributes were negative characteristics, all samples in which all of the attributes were zero were also removed. This resulted in a data set of approximately 2,000 points. Finally, 20 binary attributes were identified as specifically relevant to depression. Therefore, all of the clustering algorithms limited consideration to these 20 attributes. The 20

attributes used in the study are as shown in Table I:

## 5. Experiments

As mentioned above, the experiments described in this report followed four major steps. First,  $k$ -means clustering (with the described modification) was applied. Second, the reduced data set was processed by association analysis to generate a decision tree. Third, the competitive learning neural network was applied to data. Finally, classification labels were assigned to the data points based on the results for each of the clustering methods and decision trees were generated by C4. The results of C4 generating decision trees will be discussed at the end of each relevant section. Unfortunately, space limitations prevent us from including all of these trees. The following sections describe the results of the clustering studies.

### 5.1 $K$ -means clustering

$K$ -means clustering provides a technique for determining clusters within the data using a principle based on nearest neighbor. As such, it is not capable of handling overlapping clusters. On the other hand, it is capable of clustering based on all of the attributes rather than limiting its view to single clusters (i.e., it is polythetic). Of course, this makes it more difficult to determine relevant rules for classification, but we attempt to extract rules from the results of the analysis.

Recall that this technique requires an initial value for  $k$  to be provided as well as a coarsening and refining parameter. The latter two parameters were determined empirically, and  $k$  was set initially to 10. In particular, the coarsening parameter was set to 0.5 and the refining parameter was set to 1.95. It was found that coarsening was highly sensitive to values near 1.0 and refining was highly

sensitive near 2.0.  $K$ -means was actually applied last, so the parameters were selected to yield results similar to the other two techniques.

Following  $k$ -means clustering on the public health data, 12 clusters were identified. Attributes of their centroids are listed in Table II. It was found that the two least similar clusters were Cluster 8 and Cluster 11. It is believed that these clusters would represent the extremes on the spectrum of depression. As such, it would be valuable to decipher the centroids to determine the relevant characteristics. Cluster 8 showed very low incidence of depression related attributes with the exception of increased eating. On the other hand, Cluster 11 show high incidence of depression related attributes in all but two attributes— increased eating and moving all the time.

The attributes at the centroids can be considered as weighted “presence” of that attribute in determining whether or not a point belongs to some cluster. These weights spanned 0.1 to 0.9, so a decision tree generated by C4 will not divide cleanly along the attributes (as one might expect from a hierarchical clustering analysis such as the one discussed in the next section). In fact the pruned decision tree generated by C4 has 62 paths and a maximum depth of 13 steps.

It is interesting to note that the top attributes of the C4 tree are feelings of worthlessness, being sad for two weeks, and thinking slowly. The first two were also found to be significant in the study reported in Eaton and Ritter (1988). On the other hand, thoughts of death (considered to be the most significant attribute in the Eaton, *et al.* study) appears fairly deep in the tree.

**Table I.** Attribute for Public Health Data on Depression and Anxiety

1.	CONCENT	Trouble concentrating
2.	CRYING	Crying spells
3.	DEATHT	Thought about death
4.	DEATHW	Wanted to die
5.	EATLESS	Lost appetite
6.	GAIN2LB	Eating increased
7.	HOPELESS	Life hopeless
8.	LOSE2LB	Lost weight
9.	MOVMORE	Moving all of the time
10.	SAD2WK	Sad for two weeks
11.	SAD2YRS	Sad for two years
12.	SEXDIM	Diminished interest in sex
13.	SLPLESS	Trouble falling asleep
14.	SLPMORE	Sleeping too much
15.	SUIDTRY	Attempted suicide
16.	SUITHINK	Thought of suicide
17.	THINKSLO	Thoughts slower
18.	TIRED	Tired out
19.	TMSLOW	Talked more slowly
20.	WSG2WK	Worthless, sinful, guilty

## 5.2 Chi-square clustering

The results of running the chi-square association analysis on the public health data was a decision tree that yielded 16 classifications (Table III). Since the basic goal in classifying this data is to determine whether or not a patient is depressed, it is apparent that subcategories may exist within the data. Unfortunately, we are not in a position to determine the nature of these subcategories without the basic labeling of the data.

Perhaps the most interesting observation to be made from this analysis was determining which of the attributes are considered most significant in separating the data. Since association analysis is a hierarchical technique, attributes used near the root of the tree

differentiate between high level clusters where attributes used near the leaves of the tree differentiate between finer grained clusters. So the first observation is that the attribute DEATHW (i.e., wanting to die) should be highly indicative of whether or not a patient is depressed, assuming only the two classifications exist and a single attribute can distinguish the two clusters. Of course, this assumption may be totally inappropriate. Another plausible interpretation is that the clusters generated by this technique (and by the others) represent “degrees” of depression. As such, wanting to die may suggest more severe depression while the lack of such thoughts may not completely eliminate depression.

**Table II.** Cluster Attributes from *K*-Means Algorithm.

<u>CLUSTER</u>	<u>HIGHEST ATTRIBUTES</u>	<u>LOWEST ATTRIBUTES</u>
1	TIRE	SUIDTRY
2	EATLESS	CONCENT, CRYING, DEATHW, GAIN2LB, HOPELESS, MOVMORE, SAD2WK, SEXDIM, SUIDTRY, SUITHINK, THINKSLO, WSG2WK
3	WSG2WK	EATLESS, GAIN2LB, LOSE2LB, SEXDIM, SUIDTRY, SUITHINK, THINKSLO
4	HOPELESS	EATLESS, GAIN2LB, LOSE2LB, SAD2WK, SAD2YRS, SLPMORE, SUIDTRY, THINKSLO, TMSLOW
5	TIRE	CRYING, DEATHT, DEATHW, GAIN2LB, HOPELESS, MOVMORE, SAD2WK, SAD2YRS, SLPLESS, SUIDTRY, SUITHINK, WSG2WK
6	SLPLESS	CONCENT, DEATHT, DEATHW, EATLESS, GAIN2LB, HOPELESS, LOSE2LB, SAD2WK, SAD2YRS, SEXDIM, SLPMORE, SLPLESS, SUIDTRY, SUITHINK, TMSLOW WSG2WKS
7	DEATHT	CONCENT, CRYING, DEATHW, EATLESS, GAIN2LB, HOPELESS, LOSE2LB, SAD2WK, SAD2YRS, SEXDIM, SLPMORE, SLPLESS, SUIDTRY, SUITHINK, TMSLOW WSG2WKS
8	SAD2WK	DEATHW, GAIN2LB, MOVMORE, SLPMORE, SUIDTRY, SUITHINK, TMSLOW, WSG2WK
9	GAIN2LB	CONCENT, CRYING, DEATHW, EATLESS, HOPELESS, LOSE2LB, SAD2YRS, SUIDTRY, SUITHINK, THINKSLO, TMSLOW, WSG2WK
10	CONCENT, THINKSLO	MOVMORE, SUIDTRY, SUITHINK
11	DEATHT, DEATHW, HOPELESS, LOSE2LB, MOVMORE, SAD2WK, SUITHINK, TIRE, WSG2WK	EATLESS, GAIN2LB, SLPLESS SLPMORE, THINKSLO, TMSLOW
12	CONCENT, DEATHT, DEATHW, HOPELESS, SAD2WK, SUITHINK, THINKSLO	GAIN2LB, LOSE2LB

**Table III.** Decision Rules from Chi-Square Clustering.

<u>CLUSTER</u>	<u>RULE</u>
1	DEATHW=1, CONCENT=1, SUI THINK=1, HOPELESS=1
2	DEATHW=1, CONCENT=1, SUI THINK=1, HOPELESS=0
3	DEATHW=1, CONCENT=1, SUI THINK=0, THINKSLO=1
4	DEATHW=1, CONCENT=1, SUI THINK=0, THINKSLO=0
5	DEATHW=1, CONCENT=0, SUI DTRY=1
6	DEATHW=1, CONCENT=0, SUI DTRY=0, LOSE2LB=1
7	DEATHW=1, CONCENT=0, SUI DTRY=0, LOSE2LB=0, SAD2WK=1
8	DEATHW=1, CONCENT=0, SUI DTRY=0, LOSE2LB=0, SAD2WK=0
9	DEATHW=0, SLP MORE=1
10	DEATHW=0, SLP MORE=0, SUI THINK=1
11	DEATHW=0, SLP MORE=0, SUI THINK=0, CONCENT=1, DEATH T=1
12	DEATHW=0, SLP MORE=0, SUI THINK=0, CONCENT=1, DEATH T=0
13	DEATHW=0, SLP MORE=0, SUI THINK=0, CONCENT=0, SAD2WK=1, HOPELESS=1
14	DEATHW=0, SLP MORE=0, SUI THINK=0, CONCENT=0, SAD2WK=1, HOPELESS=0
15	DEATHW=0, SLP MORE=0, SUI THINK=0, CONCENT=0, SAD2WK=0, THINKSLO=1
16	DEATHW=, SLP MORE=0, SUI THINK=0, CONCENT=0, SAD2WK=0, THINKSLO=0

It is also interesting to note that, in Eaton and Ritter (1988), classification according to dysphoria (i.e., general depression) indicates the highest correlation with thoughts of death.

Further, two of the four classes of depression identified indicated dysphoric symptoms (indicated by thoughts of death) as a leading attribute. The attribute, SAD2WKS was also considered highly indicative of dysphoria. The chi-square analysis performed here also found this attribute to be significant but nearer the leaves of the tree.

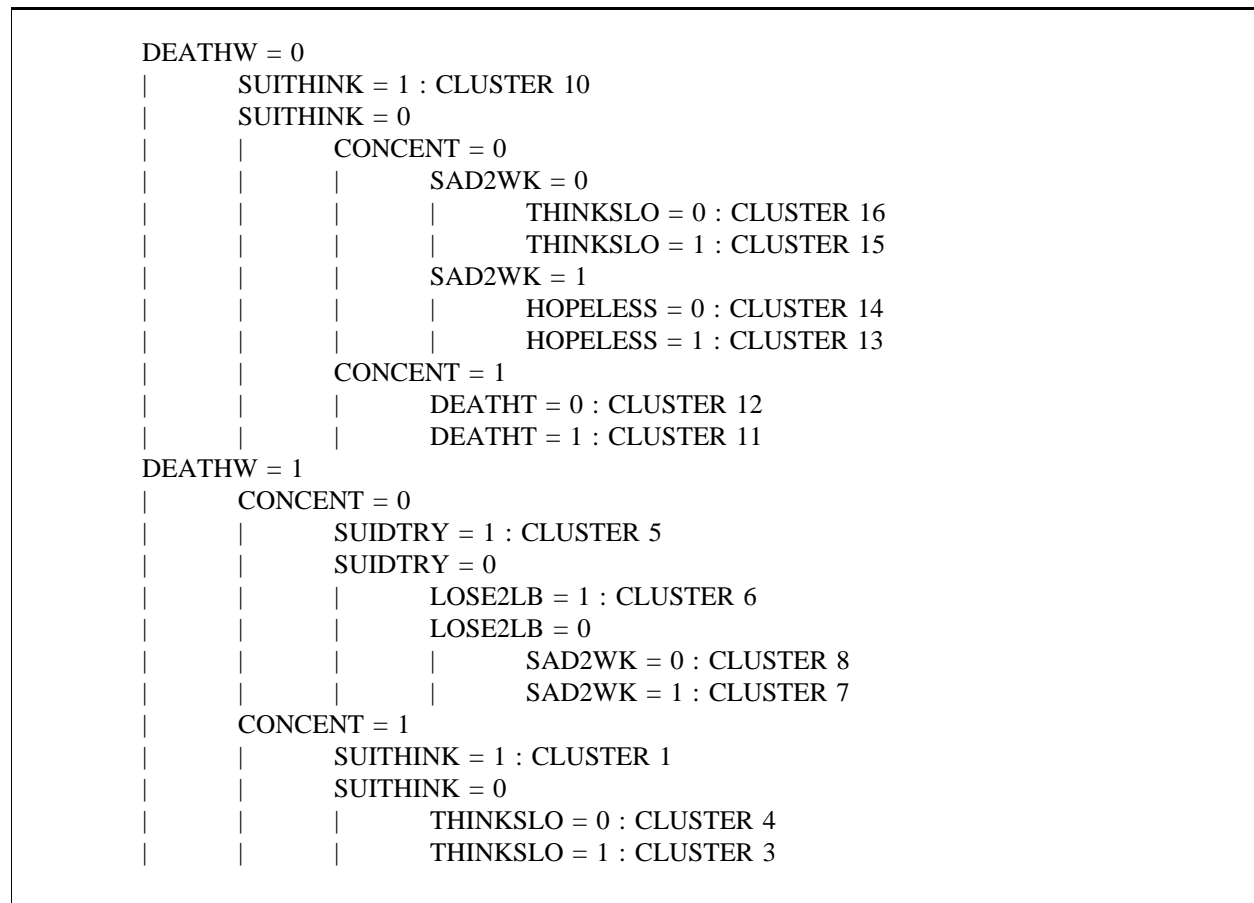
Finally, the results of the chi-square approach were processed by C4 (Table IV). The most important observation that we made from the resulting tree was that the two trees are very similar but not identical. One would expect the trees to be similar since the classifications were initially made with attributes providing “perfect” splits of the data. The reason for the difference in the trees lies in the metric

used to select an attribute. In the chi-square approach, the chi-square metric is used to find high level variation along the lines of the attributes. C4 attempts to select attributes to build the decision tree under a similar motivation, but the metric used is information gain. The information gain metric attempts to evenly split the data into near equal subsets. In fact, we find that the maximum depth of the chi-square tree is six and the maximum depth of the C4 tree is five. For 16 classes, optimal depth of the tree (assuming equal sized clusters) would be four on each branch. No calculations were conducted to determine expected cost to classify based on the size of the data set and the path lengths; however, it is conjectured that C4’s tree will be slightly better.

### 5.3 Competitive clustering

Finally, the competitive learning algorithm was applied to the public health data. This approach assumes there will be fewer clusters

**Table IV.** C4 Decision Tree from Chi-Square Clustering.



than attributes, so since there were 20 attributes, one naturally expects fewer than 20 clusters. Indeed, competitive learning identified twelve clusters as in *k*-means. The results of applying competitive learning are shown in Table V.

One should observe right away that the results are very similar to the *k*-means results. First, the number of clusters is the same. Examining the attributes that are significant (by examining the values of the weight matrix) reveals that there are several similar clusters within the network, and some of these clusters correspond to the *k*-means clusters. However, it also appears that the *k*-means clusters are more distinct. One possible explanation for this is that the competitive learning algorithm

has difficulty due to its sensitivity to the order in which the data are presented.

The corresponding decision tree generated by C4 is also very complex. It has 54 paths and a maximum depth of 17, thus its complexity is analogous to the *k*-means tree. One significant difference, however, is the selection of primary attributes (i.e., attributes near the root). The clusters generated from competitive learning resulted in primary attributes of sleeplessness, crying, and hopelessness. Only the latter is one of the significant attributes in Eaton and Ritter (1988). In fact, the more significant attributes appeared nearer to the leaves in this tree.

**Table V.** Cluster Attributes from Competitive Learning Algorithm.

<u>CLUSTER</u>	<u>HIGHEST ATTRIBUTES</u>	<u>LOWEST ATTRIBUTES</u>
1	CONCENT, THINKSLO, TMSLOW	HOPELESS, SUIDTRY, SUITHINK
2	DEATHW, SUITHINK	CRYING, EATLESS, SLPLESS, SLPMORE
3	DEATHT	SUIDTRY
4	HOPELESS	LOSE2LB
5	GAIN2LB, SLPMORE	CONCENT, CRYING, DEATHW, EATLESS, SAD2YRS, SUIDTRY
6	EATLESS, LOSE2LB	DEATHW, EATLESS, SUIDTRY
7	SEXDIM	DEATHW, HOPELESS, LOSE2LB, SAD2YRS, SUIDTRY, SUITHINK, WSG2WK
8	SAD2WK, WSG2WK	SUIDTRY, SUITHINK
9	CRYING, HOPELESS, SAD2WK	SUIDTRY
10	TIRE	GAIN2LB, SAD2WK, SUIDTRY, SUITHINK
11	MOVMORE	SUIDTRY
12	SLPLESS, TIRE	HOPELESS, LOSE2LB, SUIDTRY, SUITHINK

## 6. Discussion

The results of this study suggest that several degrees of clinical depression may exist. This is evident by the fact that all three clustering algorithms identified on the order of 12 to 16 clusters within the data. Recall that this data was reduced so as to consider attributes most relevant to depression; however, some carryover from anxiety is expected to have occurred. Nevertheless, the number of clusters identified is strong evidence that finer classifications may exist for depression.

In a previous study applying latent class analysis (Grove *et al.*, 1987), a reduced set of clusters was assumed. Specifically, this study assumed two classes. The studies reported in (Eaton and Ritter, 1988; Eaton *et al.*, 1989) also applied latent class analysis and found

four classes. In a more recent study (Furukawa and Sumita, 1992), a hierarchical clustering algorithm was applied to a similar data set and three clusters identified. Unfortunately, the data set used was extremely small (40 subjects) thus making it difficult to compare with our results.

For our experiments, we were able to observe the following. First, both *k*-means and competitive learning found 12 clusters with similar attributes. Unfortunately, the “significance” of the attributes for the two techniques (as evidenced by the C4 decision trees) did not agree. Second, the association analysis generated 16 clusters by considering clean partitions of the data along individual attributes. Now it is unreasonable to assume that all 20 of the attributes are independent, so the idea that such a clean partitioning can occur becomes difficult to accept. In fact,

many of the classification rules have combinations of thoughts of death, wanting to die, thinking about suicide, and attempting suicide. But the other rules seem to suggest grades of depression when combinations of these attributes (and others) have conflicting values (e.g., Cluster 8 included wanting to die but thinking about suicide was notably absent).

Note that both  $k$ -means and competitive learning are polythetic algorithms while chi-square clustering and C4 are monothetic algorithms. From this it should not be surprising that chi-square clustering and C4 yield comparable results as do  $k$ -means and competitive learning. It is also understandable, given this difference, that the C4 trees for  $k$ -means and competitive learning would be much more complex than the C4 tree for chi-square clustering.

Aside from the obvious differences in the trees generated by all three techniques, these trees also had several similarities. First, the principal attributes all tended to agree with Eaton and Ritter (1988) and the trees tended to be highly complex. Further, in post-pruning, all three trees showed minimal rearrangement. Thus the initial trees appeared to be near optimal for the training set.

Several additional analyses could be done on this data. First, if the data were classified, then the classifications could be compared to the clusters identified to determine if, indeed, degrees or hierarchies of depression exist. Second, closer examination of the centroids of the clusters generated by *all three* techniques may be useful in determining how similar the tree results really are. For example, it is possible that the 12 clusters identified by  $k$ -means may closely correlate to the 12 clusters identified by competitive learning (although the decision trees seem to indicate the opposite). Unfortunately, time did not permit such a correlation analysis to be run.

Finally, additional classification algorithms could provide interesting results. For example, AutoClass by Cheeseman *et al.* (1988) is a Bayesian classification tool that attempts to identify the most probable set of clusters within the data. Running AutoClass on the data would provide another valuable data point in determining the character of the depression data.

Another alternative clustering system that we may apply is Fisher's COBWEB (1987). COBWEB is an incremental system for hierarchical conceptual clustering. While our problem does not need to be examined incrementally, COBWEB offers the advantage of applying a different utility measure (i.e., category utility) to evaluate generated clusters. It also constructs the classification tree by using traditional search operators such as merging and splitting (corresponding to generalization and specialization respectively). Finally, since it represents concepts probabilistically, COBWEB should be better suited to the large data set than more rigid clustering algorithms such as  $k$ -means or chi-square clustering.

Traditional conceptual clustering as introduced by Michalski and Stepp (1983) and further developed by Stepp and Michalski (1986) rely on incorporating background knowledge in evaluating the quality of the resulting clusters. In our problem, little to no background knowledge was available, so this traditional approach could not be applied easily. COBWEB's advantage over CLUSTER/2 (Michalski and Stepp, 1983) or CLUSTER/S (Stepp and Michalski, 1986) is that the evaluation function is domain independent. However, we would expect the availability of domain knowledge to improve classification strategies.

## 7. Summary

In this paper we presented the results of three approaches to analyzing and clustering a large set of psychiatric data. As a result of this study, it is apparent that depression cannot be categorized either as simply present or absent. Further, it is unlikely as few as three or four classes of depression are sufficient. The results of this study suggest that there are many degrees of depression ranging from no depression to severe depression. Further, depending on the means by which clusters are identified, it is also apparent that a relatively well defined (although not necessarily small) set of rules can be derived to assist in classifying a patient as fitting in one of the categories. These rules may be expressed either in terms of a decision tree (as in association analysis) or as a linear equation (as in the neural net). And in each of these cases, additional decision trees can be constructed which clearly delineate the rules to be applied for classification.

## Acknowledgements

I would like to thank Joe Gallo of the Johns Hopkins School of Public Health for providing the patient data and for his assistance in interpreting the data. I would also like to thank Steven Salzberg and Simon Kasif for their ideas and support through this research. Finally, I would like to thank David Aha for providing several insightful comments after reading an early version of the paper. Support for this research was provided by ARINC Research Corporation.

## References

Aha, D. W., D. Kibler, and M. K. Albert, "Instance Based Learning Algorithms," *Machine Learning*, Vol. 6, pp. 37-66, 1991.

Cheeseman, Peter, James Kelley, Matthew Self, John Stutz, Will Taylor, and Don Freeman, "AutoClass: A Bayesian Classification System," *Proceedings of the Fifth International Workshop on Machine Learning*, San Mateo, California: Morgan Kaufmann, 1988.

Eaton, W. W., and C. Ritter, "Distinguishing Anxiety and Depression with Field Survey Data," *Psychological Medicine*, Vol. 18, pp. 155-166, 1988.

Eaton, W. W., A. McCutcheon, A. Dryman, and A. Sorenson, "Latent Class Analysis of Anxiety and Depression," *Sociological Methods and Research*, Vol. 18, No. 1, pp. 104-125, 1989

Everitt, B., *Cluster Analysis*, Wiley and Sons, New York, 1974

Fisher, D. H., "Knowledge Acquisition Via Incremental Conceptual Clustering," *Machine Learning*, Vol. 2, pp. 139-172, 1987.

Furukawa, T., and Y. Sumita, "A Cluster-Analytically Derived Subtyping of Chronic Affective Disorders," *Acta Psychiatr Scand*, Vol. 85, pp. 177-182, 1992.

Grossberg, S., "Adaptive Pattern Classification and Universal Recoding: Part I. Parallel Development and Coding of Neural Feature Detectors," *Biological Cybernetics*, Vol. 23, pp. 121-134, 1976.

Grossberg, S., "Competitive Learning: From Interactive Activation to Adaptive Resonance," *Cognitive Science*, The Cognitive Science Society, Vol. 11, pp. 23-63, 1987.

Grove, W. M., N. C. Andreasen, M. Yound, J. Endicott, M. B. Keller, R. M. A. Hirschfeld, and T. Reich, "Isolation and Characterization of a Nuclear Depressive Syndrome,"

*Psychological Medicine*, Vol. 17, pp. 471-484, 1987.

MacQueen, J. B., "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, California, pp. 281-297, 1967.

Michalski, R. S., and R. E. Stepp, "Learning from Observation: Conceptual Clustering," in *Machine Learning: An Artificial Intelligence Approach*, Vol. 1, eds. R. Michalski, J. Carbonnel, and T. Mitchel, Tioga Publishing Company, Palo Alto, California, pp. 331-363, 1983.

Quinlan, J. R., "Induction of Decision Trees," *Machine Learning*, Vol. 1, pp. 81-106, 1986.

Quinlan, J. R., "Simplifying Decision Trees," *Journal of Man-Machine Studies*, Vol. 27, pp. 221-234, 1987.

Rumelhart, D. E. and D. Zipser, "Feature Discovery by Competitive Learning," in

*Parallel Distributed Processing*, David E. Rumelhart and James L. McClelland (eds.), The MIT Press, Cambridge, Massachusetts, pp. 151-193, 1987.

Shannon, C. E., "A Mathematical Theory of Communications," *Bell System Technical Journal*, Vol. 27, pp. 379-423, 1948.

Stepp, R. E., and R. S. Michalski, "Conceptual Clustering: Inventing Goal-Oriented Classifications of Structured Objects," in *Machine Learning: An Artificial Intelligence Approach*, Vol. 2, eds. R. Michalski, J. Carbonnel, and T. Mitchel, Morgan Kaufmann Publishers, Palo Alto, California, pp. 471-498, 1986.

von der Malsburg, C., "Self-organizing of Orientation Sensitive Cells in the Striate Cortex," *Kybernetik*, Vol. 14, pp. 85-100, 1973.