Anomaly Detection, Explanation and Visualization

Ian Davidson
SGI
iand@sgi.com

Abstract: Anomaly detection is the process of identifying unusual behavior. It is widely used in data mining, for example, to identify fraud, customer behavioral change, and manufacturing flaws. We discuss how a probabilistic framework can elegantly support methods to automatically explain why observations are anomalous, assign a degree of anomaliness, visualize the normal and abnormal observations and automatically name the clusters. To our knowledge, interactive visualization of anomalies has not previously been addressed, nor automatic naming of clusters for verification in the anomaly detection field. We specifically discuss anomaly detection using mixture models and the EM algorithm, however our ideas can be generalized to anomaly detection in other probabilistic settings. We implement our ideas in the SGI MineSet product as a mining plug-in re-using the MineSet visualizers.

Keywords: Clustering, Anomaly detection, multivariate outlier detection, mixture model, EM, visualization, explanation, MineSet

Introduction to Anomaly Detection

"What does the data tell us?", is the general question that data mining, machine learning and statistical analysis attempts to answer. More specific questions involve determining what can we *predict* from the data and how can we *summarize* and *generalize* the data. Anomaly detection asks questions with a different aim. Given a set of data we wish to ascertain *what* observations don't "belong" and *which* are interesting and should be investigated. Some researchers have postulated that anomaly detection is a separate class of knowledge discovery task along with dependency detection, class identification and class description [1].

Anomaly detection has been used in many different contexts: detection of unusual images from still surveillance images [2], identifying unusual organic compounds [3], data cleaning [4] and identifying flaws in manufactured materials [5]. In most applications the basic steps remain the same:

1) Identify normality by calculating some "signature" of the data.
2) Determine some metric to calculate an observation's degree of deviation from the signature.
3) Set some criteria/threshold which, if exceeded by an observation's metric measurement means the observation is anomalous.

Various application areas of anomaly detection have different methods of addressing each step.

The signature of the data consists of identifying regularities in the data. The type of data and domain determines the method of identifying the regularities. For example, network intrusion applications might use learning techniques to exploit the sequential nature of the data [6]. Similarly, the criteria to determine if an observation is anomalous is typically application specific. In some domains observations where only one variable deviates from the signature are called anomalous, whilst in others a systematic deviation across all variables is tolerable. However, the metric used to measure the degree of anomalousness is specific to the problem formulation (modeling technique). In a probabilistic problem formulation it may be the maximum likelihood or posterior probability of an observation or in a symbolic formulation some function of the distance between the observations and the remaining observations.

In this discourse we address the questions of explaining *why* an observation is anomalous and to more precisely answer *which* observations are interesting. We argue that using a probabilistic modeling tool and evaluating the anomalies in a probabilistic framework offer flexibility and are naturally conducive to answering the questions that anomaly detection asks. We further illustrate that the innovative methods of anomaly explanation and identifying local anomalies that have been proposed in the distance based outlier detection [1] can be applied to mixture models.

In the next section we introduce anomaly detection using mixture modeling and specify typical criteria and metrics that can be used to determine if an observation is anomalous and the degree of anomalousness. We then describe how a distance measure can be derived from mixture models which enables application of the ideas from the distance based outlier detection field. Visualization of abnormal and normal observations are described next, to our knowledge 3D visualization of clustering-based anomaly detection is unique. We conclude the work by describing our methods of generating explanations of why observations are anomalous and automatically naming clusters for verification purposes. We illustrate that this can be achieved using standard approaches in probability.

We will focus on anomaly detection using mixture modeling, a common probabilistic clustering technique. However, our approach is not limited to mixture modeling or a probabilistic framework. Throughout this paper, we demonstrate our approaches on the UCI Adult data set which consists of 48,842 observations/records of 12 variables/columns. Each record represents an adult, variables are both categorical and continuous measurements of demographic information such age, income and years of education.

Anomaly Detection Using Mixture Modeling

Mixture modeling [7], also called model-based clustering is philosophically different from traditional numerical taxonomy clustering techniques such as K-Means clustering [8]. The aim of K-Means clustering is to find the set partition of observations (using hard assignment) that minimizes the distortion or vector quantization error. Let the k classes partition the observations into the subsets $C_{1...k}$, the cluster centroids be represented by $w_{1...k}$ and the n elements to cluster be $x_{1...n}$. The minimum distortion or vector quantization error that the K-Means algorithm attempts to minimize is shown in equation (1). The mathematical trivial solution which minimizes this expression is to have a cluster for each observation.

$$Distortion = \sum_{j=1}^{k}\sum_{i=1}^{N}\sum_{S_i \in C_j} D(x_i, w_{Class(S_i)})^2 \text{ , where } D \text{ is some distance metric} \qquad (1)$$

This can be viewed as grouping together observations that are most similar, whilst attempting to make the groups as dis-similar as possible to each other. However, model based techniques such as mixture modeling apriori specifies a model and attempts to estimate the parameters of the model. Philosophically the model is considered to be the generation mechanism that produced the observations. To better model the generation mechanism, partial assignment of observations to clusters/components is used.

The parameters we wish to estimate are: K the number of clusters, the relative weight/size of each cluster and the probability distributions for each variable/column for each cluster. From these we can partially assign the observations to the clusters. Formally, we can describe a K component mixture model as follows:

$$p(\mathbf{X}=\mathbf{x} \mid \theta) = \sum_{k=1}^{K} \pi_k \; p(\mathbf{X}=\mathbf{x} \mid C = c_k, \theta_k) \qquad (2)$$

where :

$\mathbf{X}$ is a multivariate random variable

C is a categorical variable with a values from $\{c_1 ... c_k\}$

π_k is the marginal probability (the weight) of the k^{th} class

θ_k is the parameter estimates of the k^{th} class containing the probability distributions for each column/variable

We use the EM algorithm [9] to find the local maximum likelihood estimates for π and θ. Once we have these then we can determine the normalized likelihood probabilities for each observation from equation (3).

$$p(\mathbf{X}=\mathbf{x} \mid \theta_i) = \frac{\pi_i \; p(\mathbf{X}=\mathbf{x} \mid C = c_i, \theta_i)}{\sum_{k=1}^{K} \pi_k \; p(\mathbf{X}=\mathbf{x} \mid C = c_k, \theta_k)} \qquad (3)$$

Mixture modeling is superior for anomaly detection over other clustering techniques such as K-Means because of several properties that maximum likelihood and Bayesian estimators provide. The specific properties include:

1. K-Means cannot model overlapping classes (due to hard assignments) and provides biased class parameter estimates.
2. K-Means is an inconsistent estimator.
3. The estimator is not invariant to scale transformations.
4. The estimator requires the a-priori specification of the number of classes with no theory (or method) for comparing which is the best model space.
5. Euclidean distance measures can unequally weight variables
6. Continuous variables are represented in a non-parametric form.

See [10] for a detailed discussion on each of these limitations.

The cluster parameter estimates are the generation mechanisms that produced the observations. The likelihood of observations produced by one of the generation mechanisms should be high. Conversely if an observation's normalized likelihood is not high for any particular cluster, then one could infer that none of the generation mechanisms produced the observations and hence the observation is an outlier and is anomalous. This is the simplest method of anomaly detection using mixture modeling. One sets a minimum likelihood threshold and those observations that do not belong to any one cluster with a likelihood greater than the threshold are deemed anomalous. We shall call this *minimum likelihood threshold anomaly detection* which we formally define below:

$$\text{if for every } \theta_i : p(\mathbf{X}=\mathbf{x} \mid \theta_i) < t, i = 1...K, \text{then anomalous} = \text{true otherwise anomalous} = \text{false} \qquad (4)$$

$$where:$$

$$t \text{ is the minimum likelihood threshold, t} = [0,1]$$

More complicated forms of anomaly detection can be created in the probabilistic framework by considering the marginal probabilities of the observation likelihood (equation (3)).

Consider a mixture model of *K* classes and *M* columns. In the most simplest situation where each of the columns is independent from each other, then the likelihood is the product of the marginal likelihoods for each variable.

$$\begin{aligned} p(\mathbf{X}=\mathbf{x} \mid \theta_i) &= \frac{\pi_i \, p(\mathbf{X}=\mathbf{x} \mid C = c_i, \theta_i)}{\sum_{k=1}^{K} \pi_k \, p(\mathbf{X}=\mathbf{x} \mid C = c_k, \theta_k)} \\ &= \frac{\pi_i \prod_{j=1}^{M} p(\mathbf{X}_j = \mathbf{x}_j \mid C = c_i, \theta_{ij})}{\sum_{k=1}^{K} \pi_k \prod_{j=1}^{M} p(\mathbf{X}_j = \mathbf{x}_j \mid C = c_k, \theta_{kj})} \end{aligned} \qquad (5)$$

We can now apply a minimum threshold, not to the entire likelihood but to the marginal likelihoods, which we will call *minimum marginal likelihood threshold anomaly detection* and is defined in equation (6)

$$\text{if for every } \theta_{ij} : p(\mathbf{X}_j = \mathbf{x}_j \mid \theta_{ij}) < t, i = 1...K, j = 1...M, \text{then anomalous} = \text{true else anomalous is false} \qquad (6)$$

$$where:$$

$$t \text{ is the minimum likelihood threshold, t} = [0,1]$$

Even more complicated anomaly detection thresholds can be created by considering the consistency of the marginal likelihoods.

Comparison of Mixture Modeling and Distance Based Anomaly Detection

Ng and co-authors have proposed a novel method which they call distance based outlier detection [1] [11]. In this section I will illustrate that one can derive a distance measure from a mixture model and possibly use the methods described in the distance based outlier/anomaly detection literature.

Distance based outlier detection considers an observation in an *m* dimensional space an outlier if at least *p* fraction of the objects in the data base do not fall within a distance *D* from the observation. From this definition the authors have described methods to categorize outliers as being trivial, non-trivial, weak and strongest. Qualitatively speaking an outlier in *m* dimensional space is termed non-trivial if and only if it is not considered an outlier in any of the sub-spaces of the *m* dimensions. An outlier in an *m* dimensional space is termed strongest if no outliers exists in any of the sub-spaces of the *m* dimensions. Conversely an outlier in an *m* dimensional space is termed weak if it is not the strongest outlier. These definitions are instrumental in their method of explaining *why* an observation is an outlier [1]. The distance based approach [11] has been extended to include the notion of degree of anomaliness.

Consider the distance-based criterion for an observation. This measure can be replicated in mixture modeling in a few ways. An observation x_n is an outlier with regard to the remaining observations $x_{1\ldots n-1}$ if the probability of predicting the observation from the remaining observations, that is $P(x_n \mid x_{1\ldots n-1})$ is below the *p* threshold. However, this would be computationally intensive to calculate. A more feasible approach could use the mean KL distance between two observations's likelihood distribution as shown in equation (7). With this distance measure the distance based outlier detection approach and associated methods can be applied to mixture modeling.

$$D(x_i, x_j) = \sum_{k=1}^{K} 0.5P(x_i \mid \theta_k).\log\left[\frac{P(x_i \mid \theta_k)}{P(x_j \mid \theta_k)}\right] + 0.5P(x_j \mid \theta_k).\log\left[\frac{P(x_j \mid \theta_k)}{P(x_i \mid \theta_k)}\right] \quad (7)$$

Visualization of Normal and Abnormal Observations

In this section we describe a visualization approach for *minimum likelihood threshold anomaly detection* (threshold set to 0.8) that can be used for other types of anomaly detection. Our aim is to effectively communicate which observations are anomalous and why. Furthermore, we wish to create a visualization that will enable easier understanding of the normalized likelihood thresholds for each observation.

Figure 1 shows our visualization of anomalies. The cluster centers that are close to each other are more similar than those cluster centers far apart. Furthermore, the position of the observations around the cluster centers reflects the "pull" of the clusters on the observations. The force of the pull is obtained from the likelihoods. Those observations that belong strongly to a cluster are towards its central region. Anomalies (that do not belong strongly to any one cluster) tend to be *between* the cluster centers. The further away an observation is from any of the cluster centers, the more anomalous it is. Observations in the central region are the most anomalous because they are being pulled towards all clusters. Specific details of the visualization will be patented

It can be seen from Figure 1 that there are different types of outliers. Consider the upper left hand region of the first figure. This region actually consists of anomalous observations which are marginally anomalous because they could belong to one other class. These type of outliers are different to those that are contained in the center of the figures, which are essentially those observations that do not belong to any one class particularly strongly. If we were to measure the entropy (degree of disorder) in the likelihoods of the observations we would find that the later type of anomalies has a higher entropy that the first type.

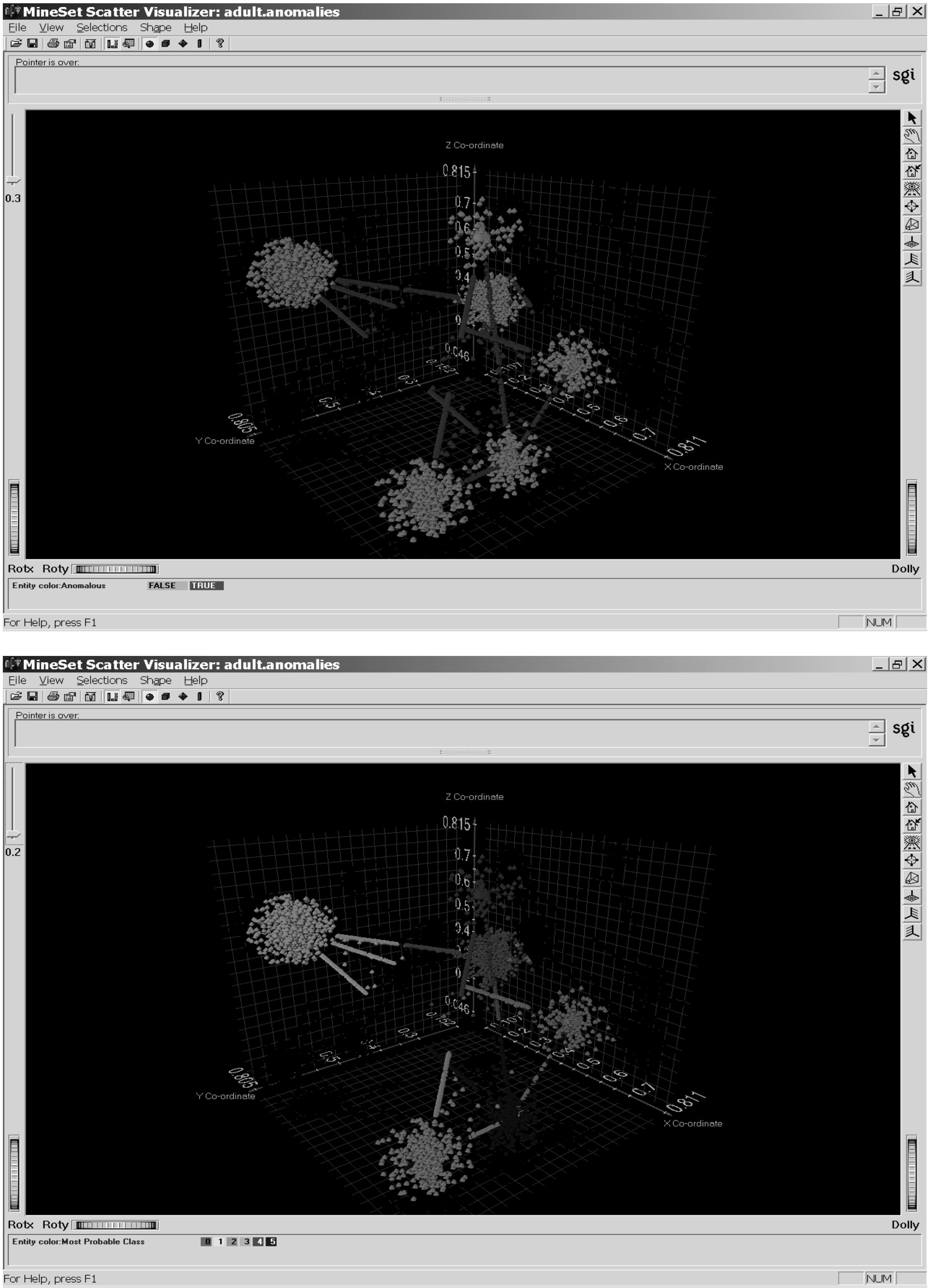


Figure 1): Visualization of anomalies generated from the UCI adult data. The first figure shows anomalies versus normal observations. With anomalies colored as red. The second figure shows the observations assigned to their most probable class with each class a different color.

Automatic Naming of Clusters/Components

The purpose of clustering is to take a population of observations and find distinct sub-populations. Therefore, when naming a cluster we wish to convey how the cluster is different from the entire population and also all other clusters. This enables a domain expert to verify that the classes found are intuitive and accurately represent the domain. This is particularly important for applications of anomaly detection as it is quite feasible that a given cluster, in the opinion of a domain expert, is a cluster of anomalies that is worth investigating.

The probability framework allows us to consider how a different a column is populated for two clusters by using the Kullback-Leibler distance. The difference between two clusters is the sum of the columns differences. We can compare a cluster's and the population's distribution of a column by considering the population to be one larger undivided cluster. We propose two methods of automatically naming clusters.

Naming The Cluster By Columns That Differ From the Population

In this approach to name a cluster, we take each column and measure the mean KL Distance between the probability distribution for the cluster and the population. Those columns not within a distance of 0.5 are deemed to be significant and differentiate the cluster from the population. For continuous variables (for example Gaussians) we can determine how the columns differ by comparing the mean values. For categorical variables we are limited to stating they are different. Figure 2 illustrates the results of automatic naming for 6 clusters identified in the adult data set, whilst Figure 3 illustrates the differences between the population and first component's distribution of the marital status and relationship categorical variables.

Component #1: Age:Very Low, Gross Income:Low, Marital Status:Different, Relationship:Different, Hours Worked:Low,
Cluster Size is :5880.15

Component #2: Education Type:Different, Relationship:Different, ,
Cluster Size is :11814

Component #3: Education Type:Different, Years of Education:High, Marital Status:Different, Relationship:Different,
Cluster Size is :7106.51

Component #4: Education Type:Different, Marital Status:Different, Relationship:Different,
Cluster Size is :10449.2

Component #5 : Gross_income:Very High, Education Type:Different, Years of Education:Very High, Marital Status:Different, Relationship:Different,
Cluster Size is :7645.1

Component #6: Education Type:Different, Years of Education:Very Low,
Cluster Size is :5946.99

Figure 2: Automatic cluster naming of the UCI adult data set.

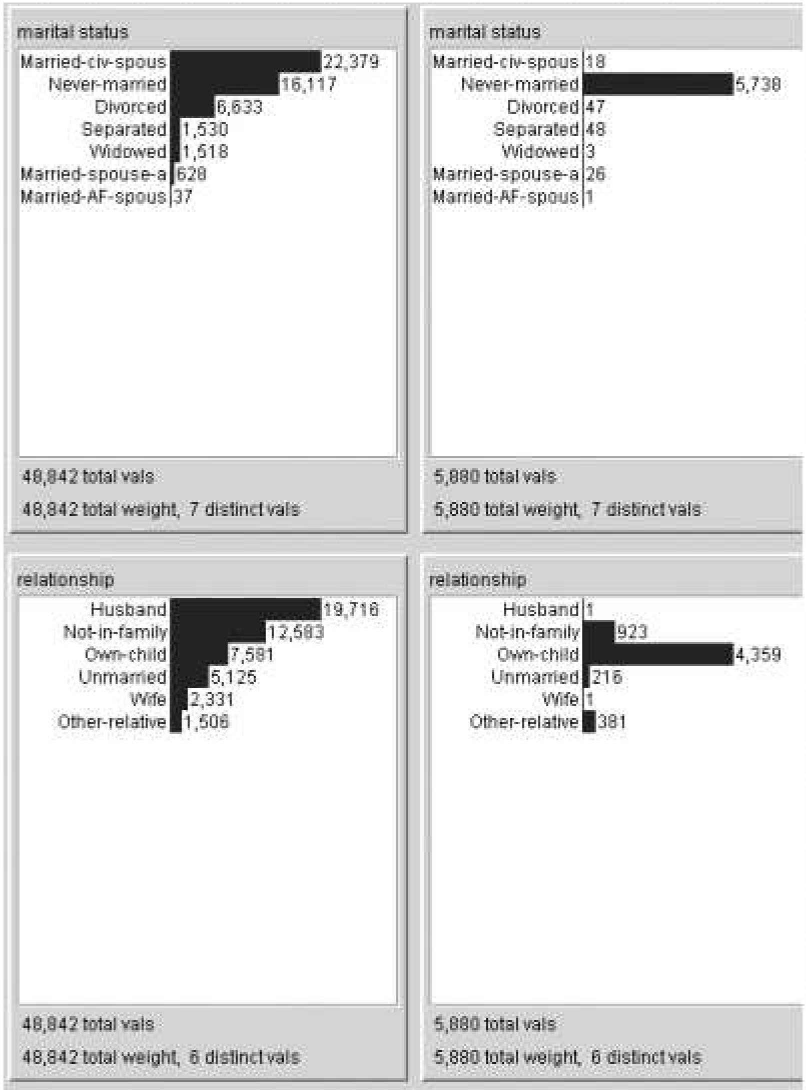


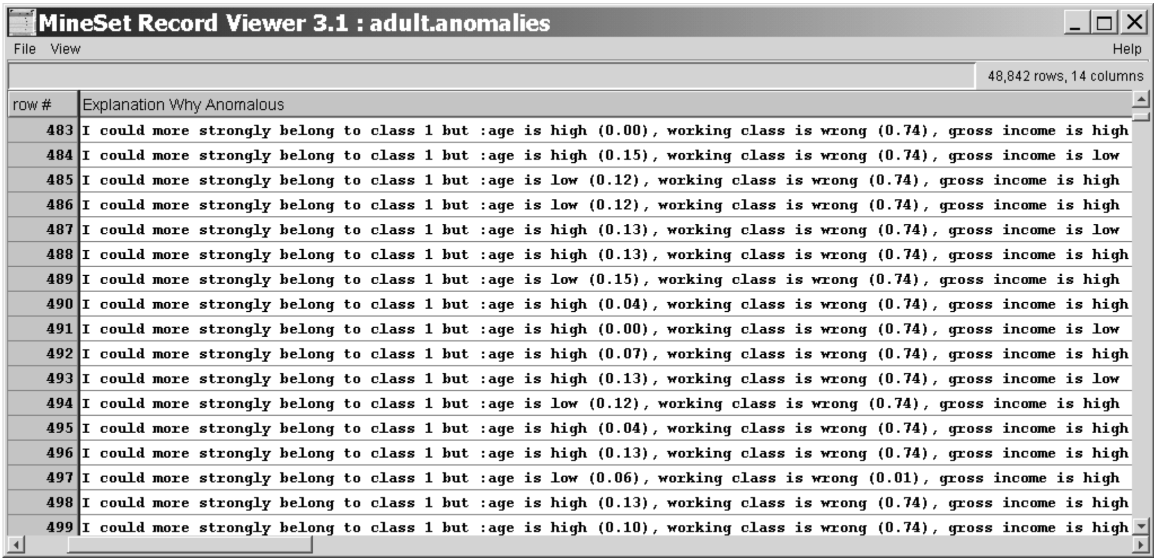
Figure 3: Comparison of population (left) and component #1 (right) distribution of Marital Status and Relationship categorical attributes.

Naming The Cluster By Columns That Are Different From All Clusters

In this approach to name a cluster we can extend the prior approach to naming clusters by considering the KL distance between the probability distribution for a specific cluster and all remaining clusters. Then if every KL distance between a cluster's column and all remaining clusters' columns are larger than 0.5 then the column is deemed significant and differentiates the cluster from all other clusters.

Explanation of Why an Observation is Anomalous

In this section we begin by describing an explanation approach for *minimum likelihood threshold anomaly detection*. An observation is anomalous in this situation if it does not belong to a cluster with a likelihood greater than some threshold, *t*. Using a *t* value of 0.8, in total there were 1867 anomalous observations. For such observations we can easily explain why an observation is anomalous by considering the most probable class and determining what changes to the observation could have increased the class likelihood above *t*. This involves iterating through all columns and determining which marginal likelihoods are below *t* and stating why as is shown in Figure 4. This figure shows the anomalies that could have belonged to component #1 had some of the observation values been different. Component #1 observations are typically very young single individuals who don't work many hours, earn much money, students could comprise a large proportion of this class. Most of the anomalies that could belong to this class more strongly were typically either too old (perhaps part time married workers) to be in the class or earned too much money by working too many hours and did not have as much education (perhaps young manual labourers).



MineSet Record Viewer 3.1 : adult.anomalies

File View Help

48,842 rows, 14 columns

row #	Explanation Why Anomalous
483	I could more strongly belong to class 1 but :age is high (0.00), working class is wrong (0.74), gross income is high
484	I could more strongly belong to class 1 but :age is high (0.15), working class is wrong (0.74), gross income is low
485	I could more strongly belong to class 1 but :age is low (0.12), working class is wrong (0.74), gross income is high
486	I could more strongly belong to class 1 but :age is low (0.12), working class is wrong (0.74), gross income is high
487	I could more strongly belong to class 1 but :age is high (0.13), working class is wrong (0.74), gross income is low
488	I could more strongly belong to class 1 but :age is high (0.13), working class is wrong (0.74), gross income is high
489	I could more strongly belong to class 1 but :age is low (0.15), working class is wrong (0.74), gross income is high
490	I could more strongly belong to class 1 but :age is high (0.04), working class is wrong (0.74), gross income is high
491	I could more strongly belong to class 1 but :age is high (0.00), working class is wrong (0.74), gross income is low
492	I could more strongly belong to class 1 but :age is high (0.07), working class is wrong (0.74), gross income is high
493	I could more strongly belong to class 1 but :age is high (0.13), working class is wrong (0.74), gross income is low
494	I could more strongly belong to class 1 but :age is low (0.12), working class is wrong (0.74), gross income is high
495	I could more strongly belong to class 1 but :age is high (0.04), working class is wrong (0.74), gross income is high
496	I could more strongly belong to class 1 but :age is high (0.13), working class is wrong (0.74), gross income is high
497	I could more strongly belong to class 1 but :age is low (0.06), working class is wrong (0.01), gross income is high
498	I could more strongly belong to class 1 but :age is high (0.13), working class is wrong (0.74), gross income is high
499	I could more strongly belong to class 1 but :age is high (0.10), working class is wrong (0.74), gross income is high

Figure 4: Explanations (partial list) of why observations from UCI Adult data-base whose most probable component is #1 are anomalous. The values in parentheses are marginal likelihoods.

An extension of this approach would involve iterating through all classes and generating a conjunction of the columns that do not pass the threshold to explain the anomaly.

Conclusions and Further Work

We believe that a probabilistic framework allows us to elegantly address questions that anomaly detection asks for three main reasons. Firstly mixture models has many benefits such as the ability to model overlapping classes and produce unbiased parameter estimates that are useful in anomaly detection. Secondly, in the field of probability there are well known and accepted approaches to measure ubiquitous qualities such as distance and degree of belongings. Thirdly, a probabilistic framework allows any number of parametric distributions to be used to model the data and still retain the same framework we have used for explanation, visualization and component naming. One can extend the mixture model to non-vector data such as curves and sequences the later which our tool supports. An advantage mixture modeling based anomaly detection has over distance based approaches is that one does not need to introduce complicated distance functions for types of data (such as sequences) where the notion of distance is non-intuitive.

We have illustrated that a likelihood formulation of mixture modeling can be used to address issues of visualizing and explaining anomalies as well as naming clusters for verification. We believe our visualization of anomalies can graphically convey which observations are anomalous and why. The approach also allows users to form categories of anomalies that will most likely vary between domains.

The work of distance based outlier detection adds the novel notions of non-trivial, strongest and weakest outliers. We have shown how by creating a distance metric from the likelihood probabilities that these notions can be applied to mixture modeling.

Though we have used a likelihood formulation of mixture modeling, we could easily extend our approach to a Bayesian formulation that would allow, amongst other things, encoding any prior knowledge and making *k* become an unknown in the problem [12]. The anomaly detection visualization can be generalized to be a general purpose cluster visualization tool, particularly if a semantic meaning is given to the co-ordinate system.

Acknowledgements

Thanks to Professor Matthew Ward, WPI for his help on specific visualization techniques.

References

[1] E. Knorr, and R.Ng, "Finding Intensional Knowledge of Distance-Based Outliers", *Proceedings of the 25th VLDB Conference, Edinburgh, Scotland, 1999*.

[2] L. Lee, R. Romano, and G. Stein. "Monitoring Activities from Multiple Video Streams: Establishing a Common Coordinate Frame.", to appear in IEEE PAMI Special Section on Video Surveillance and Monitoring (2000).

[3] G. C. Osbourn, J. W. Bartholomew, A. J. Ricco, and G. C., "A New Approach to Sensor Array Analysis Applied to Volatile Organic Compound Detection: Visual Empirical Region of Influence (VERI) Pattern Recognition", http://www.sandia.gov/imrl/XVisionScience/Xchempap.htm

[4] B.D. Wright, S. Rebrik and K.D. Miller, Spike-sorting of Tetrode Recordings in Cat LGN and Visual Cortex: Cross-channel Correlations, Thresholds, and Automatic Methods, Society for Neuroscience 28th Annual Meeting November 7-12, 1998, Los Angeles

[5] J. G. Campbell, C. Fraley, F. Murtagh and A. E. Raftery, *Linear Flaw Detection in Woven Textiles using Model-Based Clustering*, Pattern Recognition Letters: 18(1997):1539-1548

[6] T. Fawcett, and F. Provost, Activity Monitoring: Noticing Interesting Changes in Behavior, in KDD'99, ACM Press, 1999.

[7] B.S Everitt, D.J. Hand, Finite Mixture Distributions. Chapman and Hall, London, UK, London, 1981.

[8] MacQueen, J., Some Methods for classification and analysis of multivariate observations, Proceedings of the Fifty Berkeley Symposium on Mathematics, Statistics and Probability, volume 1, pages 281-296, 1967.

[9] Dempster, A.P et al, Maximum Likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society B*, Vol 39 pages 1-39, 1977.

[10] I. Davidson, "Understanding the Limitations of K-Means Clustering", To Appear

[11] M. Breunig, H. Kriegel, R. Ng, J. Sander, "LOF: Identifying Density-Based Local Outliers", *Proceedings of the ACM SIGMOD 2000, International Conference on Management of Data, Dallas, Texas,* 2000

[12] I. Davidson, "Minimum Message Length Clustering Using Gibbs Sampling", 16th International Conference on Uncertainty in A.I., 2000