Some Statistical Tools for Data Mining Applications

David M. Rocke and
David L. Woodruff
Center for Image Processing and Integrated Computing
University of California, Davis

February 16, 1998

1 Statistics and Data Mining

Statistics is about the analysis of data. Some statistical ideas are designed for problems in which well-formulated prior hypotheses are evaluated by the collection and analysis of data, but other currents of thought in the field are aimed at more exploratory ends. In this sense, data mining (defined as the exploratory analysis of large data sets) should be a branch of statistics. Yet the field of data mining has evolved almost independently of the community of statistical researchers. Why is this?

One reason for the separate development of data mining is that the methods were developed by those who needed to solve the problems, and these rarely included researchers whose primary areas of interest were statistical theory and methodology. Several authors have recently pointed out ways in which statistical ideas are relevant to data mining (see Elder and Pregibon 1996; Friedman 1997; Glymour et al. 1996, 1997; as well as many contributors to the NRC CATS report 1996). If the only reason for the lack of use of standard statistical and data analytic methods was unfamiliarity, this approach would be sufficient. But perhaps there are other problems.

Statistical methods have traditionally been developed to make the greatest use of relatively sparse data. The concept of statistical efficiency, for example, is crucial when data are expensive, and less important when additional data can be obtained for the price of fetching them from the disk. On the other hand, computational efficiency has always played a smaller role, especially since the advent of electronic computers. When analyzing potentially vast data sets, the importance of various considerations is changed or reversed compared to what statisticians are traditionally used to. It is perhaps this that has made the use of standard statistical methods less common in the field of data mining than it might otherwise have been.

2 Computational Complexity of Statistical Methods for Data Mining Applications

We consider this issue in the context of two fundamentally important methods in data mining: data cleaning and data segmentation. Data cleaning here will mean the identification of anomalous data (outliers for short) that may need to be removed or separately addressed in an analysis. Variable-by-variable data cleaning is straightforward, but often anomalies only appear when many attributes of a data point are simultaneously considered. Multivariate data cleaning is more difficult, but is an essential step in a complete analysis. We will avoid technical details of these methods (see Rocke and Woodruff 1996), but the essential nature of the methods is to identify the main "shape" of the data and identify as outliers those data points that lie too far from the main body of data.

Data segmentation is here taken to mean the division of the data into nonoverlapping subsets that are relatively similar within a subset. Some points may be outliers, and so belong to no group. We do not assume that any reliable prior knowledge exists as to which points belong together, or even how many groups there may be. Statistical methods for this problem often go by the name "cluster analysis."

In both cases of data cleaning and data segmentation, consideration will be restricted to measurement data, rather than categorical data. This will simplify the discussion so that many separate sub-cases do not need to be separately described.

It seems like an obvious point that the computational effort cannot rise too fast with the size of the dataset, otherwise processing large datasets would be impossible. If there are n data points, each with p associated measurements, then many statistical methods naively used have a computational complexity of $O(np^3)$, while more complex methods may in principle be high order polynomial, or even exponential in n or p. This is obviously not satisfactory for larger data sets.

In the quest for low computational effort, one limit is that there must of necessity be a piece which is linear in n. This is because we cannot identify outliers unless each point is examined. It will be important that the linear part has a low constant; that is, the effort will rise proportional to n, but the constant of proportionality must be small. This will be the case if we simply plug each point into a quick outlier identification routine. In no case, however, should we tolerate calculations that rise more than linearly with n.

In data segmentation, it might seem difficult to avoid a piece that is proportional to n^2, since pairwise distances are often needed for classification. We avoid this difficulty by employing sampling (in the first of two ways). If n is very large, the basic structure of a data set can be estimated using a much smaller number of points than n, and that basic structure can be used to classify the remaining points. For example, if a complex algorithm for data segmentation is $O(n^3)$, but we instead perform the calculation on a subset of size proportional to $n^{1/4}$ (for large n), the net complexity is a sublinear $O(n^{3/4})$.

An additional advantage of estimating the structure of the data set on a subset of the data is that it can be done more than once and the results compared. This allows for a kind of independent verification of results and avoids making conclusions based on accidental

appearances. If each subcalculation is sublinear, then repeating it any fixed number of times that does not rise with n is also sublinear.

Sampling has an additional role to play in estimating complex structures within a sample. Some estimation methods have a computational effort that rises exponentially with n if done naively. An example of this is a method of data cleaning that depends on finding the most compact half of the data and then evaluating each point against this half. Specifically, the MCD method finds that half of the data for which the determinant of the covariance matrix of the data is smallest. Since the space that must be searched to find even a gross approximation of the MCD rises exponentially with n, there is a danger that the computational effort to find a solution within given quality bounds will also rise rapidly (Woodruff and Rocke 1993). A solution (Woodruff and Rocke 1994) is to conduct most computations only on the cells of a partition of the data, and then use the full sample only for the final stages.

3 The Role of Algorithms in Statistical Science

At one time, it would have been feasible to argue that algorithms were not central to statistical science because all a better algorithm did was somewhat speed up arriving at essentially the same estimate. Things have changed. More and more, we are using estimators that have a substantial stochastic component, in which the "theoretical estimator" is not achievable, and all we get are various approximations of varying degrees of exactness. Due to space limitations, we will describe this phenomenon in terms of an example—S-estimators of multivariate location and scale—but it clearly applies equally to Markov chain Monte Carlo, optimal experimental design, and other important areas of statistical science.

Given a function $\rho()$, nondecreasing on $[0, \infty)$, the S-estimator of multivariate location and scale consists of that choice of a location vector $\boldsymbol{t}$ and PDS matrix $\boldsymbol{C}$ which minimizes $|\boldsymbol{C}|$ subject to

$$n^{-1} \sum \rho \left([(\boldsymbol{x}_i - \boldsymbol{t})^\top \boldsymbol{C}^{-1} (\boldsymbol{x}_i - \boldsymbol{t})]^{1/2} \right) = b_0 \tag{1}$$

which we write as

$$n^{-1} \sum \rho(d_i) = b_0. \tag{2}$$

It has been shown by Lopuhaä (1989) that S-estimators are in the class of M-estimators with standardizing constraints with weight functions $v_1(d) = w(d)$, $v_2(d) = pw(d)$, $v_3(d) = v(d)$, where $\psi(d) = \rho'(d)$, $w(d) = \psi(d)/d$, $v(d) = \psi(d)d$, with constraint (2) (Rocke and Woodruff 1993).

The S-estimator can always be shown to exist in some theoretical sense, but there exists no known algorithm that can always produce it, even in an arbitrarily long computation. The theoretical S-estimator must be a solution to the associated M-estimating equations; however, there is no known method of determining how many such local solutions there are. Given a list of such local solutions, the best of them is a candidate for the theoretical

S-estimator, but there is no known method of determining whether there is another solution with smaller determinant. Furthermore, there can be very large differences between the best solution (the theoretical S-estimator), and the second-best solution (Rocke and Woodruff 1993, 1997).

Under these circumstances, which occur in more and more modern statistical methods, the algorithm used to find local solutions is critical in the quality of the approximation to the theoretical S-estimator. Use of a less effective algorithm can result, not in a small degradation of performance, but in a very large one (Woodruff and Rocke 1994; Rocke and Woodruff, 1996). In a very real sense, "the algorithm is the estimator," and different algorithms result in essentially different estimators, which can have very different finite-sample properties.

Furthermore, if the only available algorithm for an estimator has exponential computational complexity, estimation in large, or even medium sized samples is impossible. For example, the exact computation of the MCD is only possible in samples of size 30 or less. No conceivable improvement in computational speed could raise this level higher than 100. For massive data sets, even polynomial-time algorithms may be far too slow. An $O(n^3)$ algorithm may be feasible in samples of size 1000, but not in samples of size 1,000,000. Linear, or even sub-linear, algorithms are required to deal with massive data sets.

4 Data Cleaning = Outlier Identification

While methods of detection of sporadic outliers in multivariate data have existed for many years (see Hawkins 1980), the problem of detecting outliers can be extremely difficult. This essentially requires robust estimation of multivariate location and shape, and most estimators are known to fail when the fraction of contamination is greater than $1/(p+1)$, where p is the dimension of the data. Thus detecting outliers or a disparate population that compose more than a small fraction of the data has been impractical in high dimension.

In Rocke and Woodruff (1996) we give new insights into why the problem of detecting multivariate outliers is so difficult and why the difficulty increases with the dimension of the data. We then describe significant improvements in methods for detecting outliers and demonstrate using extensive experiments that a hybrid method extends the practical boundaries of outlier detection capabilities. Determination of the exact envelope is complicated by the fact the probability of detecting outliers depends on many things such as the computer time expended, dimension, number of data points, fraction of data contaminated, type of contamination and algorithm parameters. Nonetheless, we are able to specify approximately what levels of contamination can be detected by this algorithm under a variety of conditions. The method we implement is based on extensive theoretical and methodological work, some by us and some by others.

For some statistical procedures, it is relatively straightforward to obtain estimates that are resistant to a reasonable fraction of outliers—for example, one-dimensional location (Andrews et al. 1972) and regression with error-free predictors (Huber 1981). The multivariate location and shape problem is more difficult, since most known methods will break

down if the fraction of outliers is larger than $1/(p+1)$, where p is the dimension of the data (Maronna 1976; Donoho 1982; Stahel 1981). This means that, in high dimension, a very small fraction of outliers can result in very bad estimates.

We are particularly interested in obtaining estimates that are *affine equivariant*. Technically, a location estimator $\boldsymbol{t}_n \in \mathcal{R}^p$ is affine equivariant if and only if for any vector $\boldsymbol{b} \in \mathcal{R}^p$ and any non-singular $p \times p$ matrix $\boldsymbol{A}$

$$\boldsymbol{t}_n(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}) = \boldsymbol{A}\boldsymbol{t}_n(\boldsymbol{X}) + \boldsymbol{b}.$$

A shape estimator $\boldsymbol{C}_n \in \mathrm{PDS}(p)$, the set of $p \times p$ positive-definite symmetric matrices, is affine equivariant if and only if for any vector $\boldsymbol{b} \in \mathcal{R}^p$ and any non-singular $p \times p$ matrix $\boldsymbol{A}$

$$\boldsymbol{C}_n(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}) = \boldsymbol{A}\boldsymbol{C}_n(\boldsymbol{X})\boldsymbol{A}^T$$

Informally, affine equivariance means that nonessential changes in the data, such as changing the measurement scale, do not change the solution. In addition, identification of a structure such as a low-dimensional surface that contains an important fraction of the data should not depend on the orientation or location of that structure.

4.1 Multivariate Outlier Rejection: Current State of the Art

All known affine-equivariant methods for this problem consist of the following two phases:

> PHASE I: Estimate a location and shape.
>
> PHASE II: Scale the shape estimate so that it be used to suggest which points are far enough from the location estimate to be considered possible outliers.

For details of these phases, see Rocke and Woodruff (1996).

Rocke and Woodruff (1996) give some simulation results to support the good behavior of the proposed two-phase method when the data are multivariate normal. In computational experiments, the fraction of data rejected as outliers was always near to the nominal rejection fraction used in the construction of the algorithm.

In Rocke and Woodruff (1996), we compared results using three different strategies for PHASE I: the hybrid algorithm, random search over elemental subsets (Rousseeuw 1985: MINVOL), and the forward algorithm (Atkinson 1992: FORWARD). The steepest descent algorithm (Hawkins 1993b; FSA) was not separately shown since it has been incorporated into the hybrid algorithm.

In these studies, we found that the FORWARD algorithm is greatly superior to random search over elemental subsets at all levels of contamination. The hybrid algorithm in turn is noticeably more effective than FORWARD. Similar results obtain for many choices sample sizes, times, and distances, and dimensions. In higher dimension, limited trials suggest that the superiority of the hybrid algorithm is even greater.

The software described here is publicly available from statlib in the form of a C program (http://lib.stat.cmu.edu/jasasoftware/rocke).

5 Data Segmentation = Cluster Analysis

There is not space here to review the extensive literature on cluster analysis. Since we focus on affine equivariant methods, we restrict out comments to selected previous work with this property. Our goal in this short account is not to be comprehensive, but merely to indicate where our methods could contribute to the efficacy of cluster analysis techniques, especially in massive data sets.

Our perspective is similar theoretically to that of McLachlan and Basford (1988) and Banfield and Raftery (1993). We propose to fit a mixture likelihood to identify clusters and to do so in such a way as to avoid sensitivity to outliers. McLachlan and Basford among many others use a form of the EM algorithm to obtain a mixture likelihood; Raftery uses a classification likelihood in which the data are partitioned into groups instead of being assigned vectors of posterior probabilities.

We make three main innovations in this literature. First, we use robust "pseudo-likelihoods" corresponding to M-estimators with goodness criteria depending on ellipsoidal volume measurements. This allows clustering to be free of outliers without requiring a pre-specification of the form of the outliers. Second, we are developing improved search methods for the combinatorial problem of finding good classification likelihoods. Exchange methods such as those of Späth (1985) can be improved upon using methods of heuristic search. Furthermore, our methods will be selected to be robust to outlying observations. Third, we use a two-stage method, as we have in the one sample problem, in which a combinatorial (classification likelihood) estimate is followed up by a mixture likelihood method chosen to be robust to outliers.

There are many important theoretical (even philosophical) questions that must be resolved before usable robust model-based affine-equivariant clustering methods can be developed. For example, even the definition of robustness requires thought when a cluster of "outliers" is better thought of as another cluster than as outlying observations.

Cluster analysis within the framework we use is an extension of the problem of outlier identification by robust estimation of multivariate location and shape. The problem as formulated for data cleaning was that a data set is given in which at least the majority of the points come from a well-behaved, perhaps multivariate normal, population—the other points are arbitrary. The goal is to estimate the location and shape of the population of well-behaved points as well as possible, and to identify which points come from the main population (cluster) and which are "outliers."

Now we attack a more difficult problem using extensions of the methods we previously employed. Note that the most difficult type of outliers to deal with in the above formulation is when they form a population with a mean far from that of the "good" data, but with a covariance of the same shape, and possibly smaller size (Rocke and Woodruff 1996). Now we generalize this difficult situation and suppose that we are faced with data from K populations, where K may or may not be known ex ante. Each of these populations has its own mean. They may share a covariance matrix, or they may each have a unique shape. In addition, there may be "outliers" defined now as points coming from none of the

populations.

The robust estimation/outlier detection problem can be thought of as the case when $K = 1$ (outliers are spread out) or $K = 2$ (outliers form a second cluster). We now intend to attack the problem for general K.

5.1 Formulation of the Cluster Identification Problem

Suppose that a multivariate population of dimension p is composed of a mixture of subpopulations indexed by $1 \leq i \leq K$, each of which has an elliptical (perhaps multivariate normal) distribution with mean vector $\boldsymbol{\mu}_i$ and shape matrix $\boldsymbol{\Sigma}_i$ with mixing proportions α_i, $1 \leq i \leq K$. K may or may not be known ex-ante, but the mean vectors, shape matrices, and mixing proportions are not known.

We may take as a starting point the normal maximum likelihood mixture model (Everitt 1993; Everitt and Hand 1981; Mclachlan and Basford 1988). We propose to deal with two related problems in the existing approaches to cluster identification via normal mixture models. The first problem is that even in the one dimensional, two-population case, finding the maximum likelihood estimate is a nontrivial problem (Hathaway 1986; Redner and Walker 1984) because numerous infinite and very large maxima in the likelihood exist corresponding to very small clusters. In principle, one can merely constrain the minimum cluster size, but there are still a very large number of local minima and finding a solution that actually separates even well defined clusters is not easy. This can be seen also from the fact that the multivariate outlier problem when the outliers lie in a second cluster is such a problem with $K = 2$, and yet even here it is very difficult.

A second problem we address in our methods is that there may be points that are true outliers in the sense of belonging to no defined cluster. Such points will lead to very poor estimates of multivariate location and shape if they are not discounted in the estimation process. This means that we should not look at normal MLE estimates, but at M-estimate alterations that allow for the down-weighting of aberrant points: we use the t-biweight estimate for its robustness and outlier resistance (Rocke 1996).

Our overall strategy is be a generalization of our strategy for the outlier identification problem. We search for preliminary estimates of a cluster using combinatorial methods, fit an estimate to that cluster using the t-biweight pseudo-likelihood, and decide if we have identified a potential coherent cluster. The outcomes are more complex than the outlier detection problem; we must allow for the possibility that we have found not a cluster, but a group of clusters. Finally, we can fit overall pseudo-MLE models using the t-biweight, and compare different models. Since asymptotics seem to be a poor guide to likelihoods even for the normal model, we will utilize simulation and the bootstrap to set cutoffs and compare models.

5.2 Algorithm for Cluster Identification

We are given a population $\boldsymbol{X}$ of n points in $\Re^p$ that come from an unknown number, K, of distinct populations with unknown parameters each of which contributes in excess of $h > p$

points, plus perhaps additional distinct populations that constitute "outliers" in that they contribute fewer than h points. Our objective is to determine K as well as estimate the location and shape of each of the K populations. We do not assume that we have been given a metric in advance and we require that our methods be affine equivariant. Our methods do not assume a distribution, but scaling is based on multivariate normal distributions.

This is in sharp contrast to the large literature concerning methods that assume that a distance metric is known or that the Euclidean metric is valid (e.g., Ward 1963, Lance and Williams 1967, Johnson 1967, Gower 1967, Mulvey and Crowder 1979, Amorin et al 1992, Gersho and Gray 1992, Dorndorf and Pesch 1994). Certainly, various assumptions can be reasonable or desirable in some settings, but there are other situations where they are not. Our proposed research concerns the case where we are given a data set and asked to determine what elliptical clusters we can find in the data. We are given no other information.

Our model is closer to those that assume only that K is known (some work has been done on the problem of estimating K; see e.g., Windham and Culter 1992) and attempt to find a clustering with $n_i,\ i = 1, \ldots, K$ points in each cluster. One method of finding the members of the K clusters is to assign them so as to minimize the determinant of

$$\boldsymbol{W} = \frac{1}{n-K} \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_j)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)'$$

where $\boldsymbol{x}_{ij}$ is the jth point in the ith cluster and $\bar{\boldsymbol{x}}_i$ is the mean of the ith cluster (Rubin 1967; Mariott 1971, 1982). An alternative approach when K is known is based on maximizing the log likelihood for a normal mixture, which is

$$L = \sum_{i=1}^{n} \ln \left[\sum_{j=1}^{K} \alpha(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j; \boldsymbol{x}_i) \right]$$

where $\alpha(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x})$ is the multivariate normal density with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ evaluated at $\boldsymbol{x}$, and $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are the mean and covariance of the jth cluster.

There are no methods for solving either of these two optimization problems exactly for data sets of realistic size. Combinatorial steepest descent from a given starting set of clusters has been proposed for minimization of W (e.g., Späth 1985). An iterative reweighting descent from a given set of prior p_i is suggested for maximizing the likelihood (e.g., Everitt and Hand 1981; McLachlan and Basford 1988). Both descent methods terminate at local minima that are very sensitive to the starting configuration given (see e.g., Everitt 1992). The algorithm that we propose can be used to get estimates of the clusters, which can be used directly or can be used as starting points for algorithms to minimize $\boldsymbol{W}$ or maximize the log likelihood function.

Note that there are other paradigms for cluster analysis such as hierarchical clustering that do not directly speak to the problem as we have formulated it (Everitt 1993; Hartigan 1975; Kaufman and Rousseeuw 1990). We will not address these methods here.

6 Conclusion

A statistical perspective on data mining can yield important benefits if the data mining perspective on statistical methods is kept in mind during their development. In the data mining perspective, statistical efficiency is not a particularly important goal, whereas computational efficiency is critical. Statistical methods must be used that do not depend on prior knowledge of the exact structure of the data; the questions to be asked as well as the answers to be derived must be free to depend on the outcome of the analysis.

Although data mining has developed mostly independently of statistics as a discipline, a fusion of the ideas from these two fields will lead to better methods for the analysis of massive data sets.

References

Banfield, J. D. and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, **49**, 803–821.

Cabena, Peter, Hadjinian, Pablo, Stadler, Rolf, Verhees, Jaap, Zanasi, Alessandro (1998) *Discovering Data Mining: From Concept to Implementation*, Upper Saddle River, NJ: Prentice Hall.

Elder IV, John and Pregibon, Daryl (1996) "A Statistical Perspective on Knowledge Discovery in Databases," in *Advances in Knowledge Discovery and Data Mining*, Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy (eds.) Cambridge, MA: MIT Press

Fayyad, Usama M., Piatetsky-Shapiro, Gregory, Smyth, Padhraic, and Uthurusamy, Ramasamy (eds.) (1996) *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: MIT Press.

Friedman, Jerome H. (1997) "On Bias, Variance, 0/1-Loss, and the Curse-of-Dimensionality," *Data Mining and Knowledge Discovery*, **1**, 55–78..

Glymour, Clark, Madigan, David, Pregibon, Daryl, and Smyth, Padhraic (1996) "Statistical Inference and Data Mining," *Communications of the ACM*, **39**, 35–41.

Glymour, Clark, Madigan, David, Pregibon, Daryl, and Smyth, Padhraic (1997) "Statistical Themes and Lessons for Data Mining," *Data Mining and Knowledge Discovery*, **1**, 11–28.

Grübel, R. and Rocke, D. M. (1990) "On the Cumulants of Affine Equivariant Estimators in Elliptical Families," *Journal of Multivariate Analysis*, **35**, 203–222.

Hawkins (1993b) 'The Feasible Solution Algorithm for the Minimum Covariance Determinant Estimator in Multivariate Data," Computational Statistics and Data Analysis, **17**, 197-210.

McLachlan, G.J., and K.E. Basford (1988) *Mixture Models: Inference and Applications to Clustering* Marcel Dekker, New York.

National Research Council, Board on Mathematical Sciences, Committee on Applied and Theoretical Statistics (1996) *Massive Data Sets: Proceedings of a Workshop*, Washington, DC: National Acadamy Press.

Rocke, D. M. (1996) "Robustness Properties of S-Estimators of Multivariate Location and Shape in High Dimension," *Annals of Statistics*, **24**, 1327–1345.

Rocke, D. M. and Woodruff, D. L. (1993) "Computation of Robust Estimates of Multivariate Location and Shape," *Statistica Neerlandica*, **47**, 27-42.

Rocke, D. M. and Woodruff, D. L. (1996) "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association*, **91**, 1047–1061.

Rocke, D. M. and Woodruff, D. L. (1997) "Robust Estimation of Multivariate Location and Shape," *Journal of Statistical Planning and Inference*, **57**, 245–255.

Späth, H. (1985) *Cluster Dissection and Analysis*, New York: Halsted Press.

Woodruff, D. L., and Rocke D. M. (1993) "Heuristic Search Algorithms for the Minimum Volume Ellipsoid," *Journal of Computational and Graphical Statistics*, **2**, 69–95.

Woodruff, D. L. and Rocke, D. M. (1994) "Computable robust estimation of multivariate location and shape in high dimension using compound estimators," *Journal of the American Statistical Association*, **89**, 888–896.