Clustering with finite data from semi-parametric mixture distributions

Yong Wang

Computer Science Department
University of Waikato, New Zealand
Email: yongwang@cs.waikato.ac.nz

Ian H. Witten

Computer Science Department
University of Waikato, New Zealand
Email: ihw@cs.waikato.ac.nz

Abstract

Existing clustering methods for the semi-parametric mixture distribution perform well as the volume of data increases. However, they all suffer from a serious drawback in finite-data situations: small outlying groups of data points can be completely ignored in the clusters that are produced, no matter how far away they lie from the major clusters. This can result in unbounded loss if the loss function is sensitive to the distance between clusters.

This paper proposes a new distance-based clustering method that overcomes the problem by avoiding global constraints. Experimental results illustrate its superiority to existing methods when small clusters are present in finite data sets; they also suggest that it is more accurate and stable than other methods even when there are no small clusters.

1 Introduction

A common practical problem is to fit an underlying statistical distribution to a sample. In some applications, this involves estimating the parameters of a single distribution function—e.g. the mean and variance of a normal distribution. In others, an appropriate mixture of elementary distributions must be found—e.g. a set of normal distributions, each with its own mean and variance. Among many kinds of mixture distribution, one in particular is attracting increasing research attention because it has many practical applications: the semi-parametric mixture distribution.

A *semi-parametric mixture distribution* is one whose cumulative distribution function (CDF) has the form

$$F_G(x) = \int_\Theta F(x;\theta)\; dG(\theta), \tag{1}$$

where $\theta \in \Theta$, the parameter space, and $x \in \mathcal{X}$, the sample space. This gives the CDF of the mixture distribution $F_G(x)$ in terms of two more elementary distributions: the *component distribution* $F(x;\theta)$, which is given, and the *mixing distribution* $G(\theta)$, which is unknown. The former has a single unknown parameter θ, while the latter gives a CDF for θ. For example, $F(x;\theta)$ might be the normal distribution with mean θ and unit variance, where θ is a random variable distributed according to $G(\theta)$.

The problem that we will address is the estimation of $G(\theta)$ from sampled data that are independent and identically distributed according to the unknown distribution $F_G(x)$. Once $G(\theta)$ has been obtained, it is a straightforward matter to obtain the mixture distribution.

The CDF $G(\theta)$ can be either continuous or discrete. In the latter case, $G(\theta)$ is composed of a number of mass points, say, $\theta_1, \ldots, \theta_k$ with masses $w_1, \ldots, w_k$ respectively, satisfying $\sum_{i=1}^k w_i = 1$. Then (1) can be re-written as

$$F_G(x) = \sum_{i=1}^{k} w_i F(x;\theta_i), \tag{2}$$

each mass point providing a component, or cluster, in the mixture with the corresponding weight. If the number of components k is finite and known *a priori*, the mixture distribution is called *finite*; otherwise it is treated as countably infinite. The qualifier "countably" is necessary to distinguish this case from the situation with continuous $G(\theta)$, which is also infinite.

We will focus on the estimation of arbitrary mixing distributions, i.e., $G(\theta)$ is any general probability distribution—finite, countably infinite or continuous. A few methods for tackling this problem can be found in the literature. However, as we shall see, they all suffer from a serious drawback in finite-data situations: small outlying groups of data points can be completely ignored in the clusters that are produced.

This phenomenon seems to have been overlooked, presumably for three reasons: small amounts of data may be assumed to represent a small loss; a few data points can easily be dismissed as outliers; and in the limit the problem evaporates because most estimators possess the property of *strong consistency*—which means that, almost surely, they converge weakly to any given $G(\theta)$ as the sample size approaches infinity. However, often these reasons are inappropriate: the loss function may be sensitive to the distance between clusters; the small number of outlying data points may actually represent small clusters; and any practical clustering situation will necessarily involve finite data.

This paper proposes a new method, based on the idea of local fitting, that successfully solves the problem. The experimental results presented below illustrate its superiority to existing methods when small clusters are present in finite data sets. Moreover, they also suggest that it is more accurate and stable than other methods even when there are no small clusters. Existing clustering methods for semi-parametric mixture distributions are briefly reviewed in the next section. Section 3 identifies a common problem from which these current methods suffer. Then we present the new solution, and in Section 5 we describe experiments that illustrate the problem that has been identified and show how the new method overcomes it.

2 Clustering methods

The general problem of inferring mixture models is treated extensively and in considerable depth in books by Titterington et al. (1985), McLachlan and Basford (1988) and Lindsay (1995). For semi-parametric mixture distributions there are three basic approaches: *minimum distance*, *maximum likelihood*, and *Bayesian*. We briefly introduce the first approach, which is the one adopted in the paper, review the other two to show why they are not suitable for arbitrary mixtures, and then return to the chosen approach and review the minimum distance estimators for arbitrary semi-parametric mixture distributions that have been described in the literature.

The idea of the minimum distance method is to define some measure of the goodness of the clustering and optimize this by suitable choice of a mixing distribution $G_n(\theta)$ for a sample of size n. We generally want the estimator to be strongly consistent as $n \to \infty$, in the sense defined above, for arbitrary mixing distributions. We also generally want to take advantage of the special structure of semi-parametric mixtures to come up with an efficient algorithmic solution.

The maximum likelihood approach maximizes the likelihood (or equivalently the log-likelihood) of the data by suitable choice of $G_n(\theta)$. It can in fact be viewed as a minimum distance method that uses the Kullback–Leibler distance (Titterington et al., 1985). This approach has been widely used for estimating finite mixtures, particularly when the number of clusters is fairly small, and it is generally accepted that it is more accurate than other methods. However, it has not been used to estimate arbitrary semi-parametric mixtures, presumably because of its high computational cost. Its speed drops dramatically as the number of parameters that must be determined increases, which makes it computationally infeasible for arbitrary mixtures, since each data point might represent a component of the final distribution with its own parameters.

Bayesian methods assume prior knowledge, often given by some kind of heuristic, to determine a suitable *a priori* probability density function. They are often used to determine the number of components in the final distribution—particularly when outliers are present. Like the maximum likelihood approach they are computationally expensive, for they use the same computational techniques.

We now review existing minimum distance estimators for arbitrary semi-parametric mixture distributions. We begin with some notation. Let $x_1, \ldots, x_n$ be a sample chosen according to the mixture distribution, and suppose (without loss of generality) that the sequence is ordered so that $x_1 \leq x_2 \leq \ldots \leq x_n$. Let $G_n(\theta)$ be a discrete estimator of the underlying mixing distribution with a set of support points at $\{\theta_{nj}; j = 1, \ldots, k_n\}$. Each θ_{nj} provides a component of the final clustering with weight $w_{nj} \geq 0$, where $\sum_{j=1}^{k_n} w_{nj} = 1$. Given the support points, obtaining $G_n(\theta)$ is equivalent to computing the weight vector $w_n = (w_{n1}, w_{n2}, \ldots, w_{nk_n})'$. Denote by $F_{G_n}(x)$ the estimated mixture CDF with respect to $G_n(\theta)$.

Two minimum distance estimators were proposed in the late 1960s. Choi and Bulgren (1968) used

$$\frac{1}{n}\sum_{i=1}^{n}[F_{G_n}(x_i) - i/n]^2 \tag{3}$$

as the distance measure. Minimizing this quantity with respect to G_n yields a strongly consistent estimator. A slight improvement is obtained by using the Cramér-von Mises statistic

$$\frac{1}{n}\sum_{i=1}^{n}[F_{G_n}(x_i) - (i - 1/2)/n]^2 + 1/(12n^2), \tag{4}$$

which essentially replaces i/n in (3) with $(i - \frac{1}{2})/n$ without affecting the asymptotic result. As might be expected, this reduces the bias for small-sample cases, as was demonstrated empirically by Macdonald (1971) in a note on Choi and Bulgren's paper.

At about the same time, Deely and Kruse (1968) used the sup-norm associated with the Kolmogorov-Smirnov test. The minimization is over

$$\sup_{1 \leq i \leq n} \{|F_{G_n}(x_i) - (i-1)/n|, |F_{G_n}(x_i) - i/n|\}, \quad (5)$$

and this leads to a linear programming problem. Deely and Kruse also established the strong consistency of their estimator G_n. Ten years later, this approach was extended by Blum and Susarla (1977) by using any sequence $\{f_n\}$ of functions which satisfies $\sup |f_n - f_G| \to 0$ a.s. as $n \to \infty$. Each f_n can, for example, be obtained by a kernel-based density estimator. Blum and Susarla approximated the function f_n by the overall mixture pdf f_{G_n}, and established the strong consistency of the estimator G_n under weak conditions.

For reason of simplicity and generality, we will denote the approximation between two mathematical entities of the same type by $\cong$, which implies the minimization with respect to an estimator of a distance measure between the entities on either side. The types of entity involved in this paper include vector, function and measure, and we use the same symbol $\cong$ for each.

In the work reviewed above, two kinds of estimator are used: CDF-based (Choi and Bulgren, Macdonald, and Deely and Kruse) and pdf-based (Blum and Susarla). CDF-based estimators involve approximating an empirical distribution with an estimated one F_{G_n}. We write this as

$$F_{G_n} \cong F_n, \quad (6)$$

where F_n is the Kolmogorov empirical CDF—or indeed any empirical CDF that converges to it. Pdf-based estimators involve the approximation between probability density functions:

$$f_{G_n} \cong f_n, \quad (7)$$

where f_{G_n} is the estimated mixture pdf and f_n is the empirical pdf described above.

The entities involved in (6) and (7) are functions. When the approximation is computed, however, it is computed between vectors that represent the functions. These vectors contain the function values at a particular set of points, which we call "fitting points." In the work reviewed above, the fitting points are chosen to be the data points themselves.

3 The problem of minority clusters

Although they perform well asymptotically, all the minimum distance methods described above suffer from the finite-sample problem discussed earlier: they can neglect small groups of outlying data points no matter how far they lie from the dominant data points. The underlying reason is that the objective function to be minimized is defined globally rather than locally. A global approach means that the value of the estimated probability density function at a particular place will be influenced by all data points, no matter how far away they are. This can cause small groups of data points to be ignored even if they are a long way from the dominant part of the data sample. From a probabilistic point of view, however, there is no reason to subsume distant groups within the major clusters just because they are relatively small.

The ultimate effect of suppressing distant minority clusters depends on how the clustering is applied. If the application's loss function depends on the distance between clusters, the result may prove disastrous because there is no limit to how far away these outlying groups may be. One might argue that small groups of points can easily be explained away as outliers, because the effect will become less important as the number of data points increases—and it will disappear in the limit of infinite data. However, in a finite-data situation—and all practical applications necessarily involve finite data—the "outliers" may equally well represent small minority clusters. Furthermore, outlying data points are not really treated as outliers by these methods—whether or not they are discarded is merely an artifact of the global fitting calculation. When clustering, the final mixture distribution should take all data points into account—including outlying clusters if any exist. If practical applications demand that small outlying clusters are suppressed, this should be done in a separate stage.

In distance-based clustering, each data point has a far-reaching effect because of two global constraints. One is the use of the cumulative distribution function; the other is the normalization constraint $\sum_{j=1}^{k_n} w_{nj} = 1$. These constraints may sacrifice a small number of data points—at any distance—for a better overall fit to the data as a whole. Choi and Bulgren (1968), the Cramer-von Mises statistic (Macdonald, 1971), and Deely and Kruse (1968) all enforce both the CDF and the normalization constraints. Blum and Susarla (1977) drop the CDF, but still enforce the normalization constraint. The result is that these clustering methods are only appropriate for finite mixtures without small clusters, where the risk of suppressing clusters is low.

This paper addresses the general problem of arbitrary mixtures. Of course, the minority cluster problem exists for all types of mixture—including finite mixtures. Even here, the maximum likelihood and Bayesian approaches do not solve the problem, because they both introduce a global normalization constraint.

4 Solving the minority cluster problem

Now that the source of the problem has been identified, the solution is clear, at least in principle: drop both the approximation of CDFs, as Blum and Susarla (1977) do, and the normalization constraint—no matter how seductive it may seem.

Let G'_n be a discrete function with masses $\{w_{nj}\}$ at $\{\theta_{nj}\}$; note that we do not require the w_{nj} to sum to one. Since the new method operates in terms of measures rather than distribution functions, the notion of approximation is altered to use intervals rather than points. Using the formulation described in Section 2, we have

$$P_{G'_n} \cong P_n, \tag{8}$$

where $P_{G'_n}$ is the estimated measure and P_n is the empirical measure. The intervals over which the approximation takes place are called "fitting intervals." Since (8) is not subject to the normalization constraint, G'_n is not a CDF and $P_{G'_n}$ is not a probability measure. However, G'_n can be easily converted into a CDF estimator by normalizing it after equation (8) has been solved.

To define the estimation procedure fully, we need to determine (a) the set of support points, (b) the set of fitting intervals, (c) the empirical measure, and (d) the distance measure. Here we discuss these in an intuitive manner; Wang and Witten (1999) show how to determine them in a way that guarantees a strongly consistent estimator.

Support points. The support points are usually suggested by the data points in the sample. For example, if the component distribution $F(x;\theta)$ is the normal distribution with mean θ and unit variance, each data point can be taken as a support point. In fact, the support points are more accurately described as *potential* support points, because their associated weights may become zero after solving (8)—and, in practice, many often do.

Fitting intervals. The fitting intervals are also suggested by the data points. In the normal distribution example, each data point x_i can provide one interval, such as $[x_i - 3\sigma, x_i]$, or two, such as $[x_i - 3\sigma, x_i]$ and $[x_i, x_i + 3\sigma]$, or more. There is no problem if the fitting intervals overlap. Their length should not be so large that points can exert an influence on the clustering at an unduly remote place, nor so small that the empirical measure is inaccurate. The experiments reported below use intervals of a few standard deviations around each data point, and, as we will see, this works well.

Empirical measure. The empirical measure can be the probability measure determined by the Kolmogorov empirical CDF, or any measure that converges to it. The fitting intervals discussed above can be open, closed, or semi-open. This will affect the empirical measure if data points are used as interval boundaries, although it does not change the values of the estimated measure because the corresponding distribution is continuous. In small-sample situations, bias can be reduced by careful attention to this detail—as Macdonald (1971) discusses with respect to Choi and Bulgren's (1968) method.

Distance measure. The choice of distance measure determines what kind of mathematical programming problem must be solved. For example, a quadratic distance will give rise to a least squares problem under linear constraints, whereas the sup-norm gives rise to a linear programming problem that can be solved using the simplex method. These two measures have efficient solutions that are globally optimal.

It is worth pointing out that abandoning the global constraints associated with both CDFs and normalization can brings with it a computational advantage. In vector form, we write $P_{G'_n} = A_{G'_n} w_n$, where w_n is the (unnormalized) weight vector and each element of the matrix $A_{G'_n}$ is the probability value of a component distribution over an fitting interval. Then, provided the support points corresponding to w'_n and w''_n lie outside each others' sphere of influence as determined by the component distributions $F(x;\theta)$, the estimation procedure becomes

$$\begin{pmatrix} A'_{G'_n} & 0 \\ 0 & A''_{G'_n} \end{pmatrix} \begin{pmatrix} w'_n \\ w''_n \end{pmatrix} \cong \begin{pmatrix} P'_n \\ P''_n \end{pmatrix}, \tag{9}$$

subject to $w'_n \geq 0$ and $w''_n \geq 0$. This is the same as combining the solutions of two sub-equations, $A'_n w'_n \cong P'_n$ subject to $w'_n \geq 0$, and $A''_n w''_n \cong P''_n$ subject to $w''_n \geq 0$. If the relevant support points continue to lie outside each others' sphere of influence, the sub-equations can be further partitioned. This implies that when data points are sufficiently far apart, the mixing distribution G can be estimated by grouping data points in different regions. Moreover, the solution in each region can be normalized separately before they are combined, which yields a better estimation of the mixing distribution.

If the normalization constraint $\sum_{j=1}^{k_n} w_{nj} = 1$ is retained when estimating the mixing distribution, the estimation procedure becomes

$$P_{G_n} \cong P_n. \tag{10}$$

where the estimator G_n is a discrete CDF on Θ. This constraint is necessary for the left-hand side of (10) to be a probability measure. Although he did not develop an operational estimation scheme, Barbe (1998) suggested exploiting the fact that the empirical probability measure is approximated by the estimated probability measure—but he retained the normalization constraint. As noted above, relaxing the constraint has the effect of loosening the throttling effect of large clusters on small groups of outliers, and our experimental results show that the resulting estimator suffers from the drawback noted earlier.

Both estimators, G_n obtained from (10) and G'_n from (8), have been shown to be strongly consistent under weak conditions similar to those used by others (Wang & Witten, 1999). Of course, the weak convergence of G'_n is in the sense of general functions, not CDFs. The strong consistency of G'_n immediately implies the strong consistency of the CDF estimator obtained by normalizing G'_n.

5 Experimental validation

We have conducted experiments to illustrate the failure of existing methods to detect small outlying clusters, and the improvement achieved by the new scheme. The results also suggest that the new method is more accurate and stable than the others.

When comparing clustering methods, it is not always easy to evaluate the clusters obtained. To finesse this problem we consider simple artificial situations in which the proper outcome is clear. Some practical applications of clusters do provide objective evaluation functions; however, these are beyond the scope of this paper.

The methods used are Choi and Bulgren (1968) (denoted CHOI), Macdonald's application of the Cramér-von Mises statistic (CRAMÉR), the new method with the normalization constraint (TEST), and the new method without that constraint (NEW). In each case, equations involving non-negativity and/or linear equality constraints are solved as quadratic programming problems using the elegant and efficient procedures NNLS and LSEI provided by Lawson and Hanson (1974). All four methods have the same computational time complexity.

We set the sample size n to 100 throughout the experiments. The data points are artificially generated from a mixture of two clusters: n_1 points from $N(0,1)$ and n_2 points from $N(100,1)$. The values of n_1 and n_2 are in the ratios $99:1$, $97:3$, $93:7$, $80:20$ and $50:50$.

Every data point is taken as a potential support point in all four methods: thus the number of potential components in the clustering is 100. For TEST and NEW, fitting intervals need to be determined. In the experiments, each data point x_i provides the two fitting intervals $[x_i - 3, x_i]$ and $[x_i, x_i + 3]$. Any data point located on the boundary of an interval is counted as half a point when determining the empirical measure over that interval.

These choices are admittedly crude, and further improvements in the accuracy and speed of TEST and NEW are possible that take advantage of the flexibility provided by (10) and (8). For example, accuracy will likely increase with more—and more carefully chosen—support points and fitting intervals. The fact that it performs well even with crudely chosen support points and fitting intervals testifies to the robustness of the method.

Our primary interest in this experiment is the weights of the clusters that are found. To cast the results in terms of the underlying models, we use the cluster weights to estimate values for n_1 and n_2. Of course, the results often do not contain exactly two clusters—but because the underlying cluster centres, 0 and 100, are well separated compared to their standard deviation of 1, it is highly unlikely that any data points from one cluster will fall anywhere near the other. Thus we use a threshold of 50 to divide the clusters into two groups: those near 0 and those near 100. The final cluster weights are normalized, and the weights for the first group are summed to obtain an estimate $\hat{n}_1$ of n_1, while those for the second group are summed to give an estimate $\hat{n}_2$ of n_2.

Table 1 shows results for each of the four methods. Each cell represents one hundred separate experimental runs. Three figures are recorded. At the top is the number of times the method failed to detect the smaller cluster, that is, the number of times $\hat{n}_2 = 0$. In the middle are the average values for $\hat{n}_1$ and $\hat{n}_2$. At the bottom is the standard deviation of $\hat{n}_1$ and $\hat{n}_2$ (which are equal). These three figures can be thought of as measures of reliability, accuracy and stability respectively.

The top figures in Table 1 show clearly that only NEW is always reliable in the sense that it never fails to detect the smaller cluster. The other methods fail mostly when $n_2 = 1$; their failure rate gradually decreases as n_2 grows. The center figures show that, under all conditions, NEW gives a more accurate estimate of the correct values of n_1 and n_2 than the other methods. As expected, CRAMÉR shows a noticeable improvement over CHOI, but it is very minor. The TEST method has lower failure rates and produces estimates that are more accurate and far more stable (indicated by the bottom fig-

		$n_1 = 99$ $n_2 = 1$	$n_1 = 97$ $n_2 = 3$	$n_1 = 93$ $n_2 = 7$	$n_1 = 80$ $n_2 = 20$	$n_1 = 50$ $n_2 = 50$
CHOI	Failures	86	42	4	0	0
	$\hat{n}_1/\hat{n}_2$	99.9/0.1	99.2/0.8	95.8/4.2	82.0/18.0	50.6/49.4
	SD($\hat{n}_1$)	0.36	0.98	1.71	1.77	1.30
CRAMÉR	Failures	80	31	1	0	0
	$\hat{n}_1/\hat{n}_2$	99.8/0.2	98.6/1.4	95.1/4.9	81.6/18.4	49.7/50.3
	SD($\hat{n}_1$)	0.50	1.13	1.89	1.80	1.31
TEST	Failures	52	5	0	0	0
	$\hat{n}_1/\hat{n}_2$	99.8/0.2	98.2/1.8	94.1/5.9	80.8/19.2	50.1/49.9
	SD($\hat{n}_1$)	0.32	0.83	0.87	0.78	0.55
NEW	Failures	0	0	0	0	0
	$\hat{n}_1/\hat{n}_2$	99.0/1.0	96.9/3.1	92.8/7.2	79.9/20.1	50.1/49.9
	SD($\hat{n}_1$)	0.01	0.16	0.19	0.34	0.41

Table 1: Experimental results for detecting small clusters

ures) than those for CHOI and CRAMÉR—presumably because it is less constrained. Of the four methods, NEW is clearly and consistently the winner in terms of all three measures: reliability, accuracy and stability.

The results of the NEW method can be further improved. If the decomposed form (9) is used instead of (8), and the solutions of the sub-equations are normalized before combining them—which is feasible because the two underlying clusters are so distant from each other—the correct values are obtained for $\hat{n}_1$ and $\hat{n}_2$ in virtually every trial.

6 Conclusions

We have identified a shortcoming of existing clustering methods for arbitrary semi-parametric mixture distributions: they fail to detect very small clusters reliably. This is a significant weakness when the minority clusters are far from the dominant ones and the loss function takes account of the distance of misclustered points.

We have described a new clustering method for arbitrary semi-parametric mixture distributions, and shown experimentally that it overcomes the problem. Furthermore, the experiments suggest that the new estimator is more accurate and more stable than existing ones.

References

Barbe, P. (1998). Statistical analysis of mixtures and the empirical probability measure. *Acta Applicandae Mathematicae, 50(3)*, 253–340.

Blum, J. R. & Susarla, V. (1977). Estimation of a mixing distribution function. *Ann. Probab*, *5*, 200–209.

Choi, K. & Bulgren, W. B. (1968). An estimation procedure for mixtures of distributions. *J. R. Statist. Soc. B*, *30*, 444–460.

Deely, J. J. & Kruse, R. L. (1968). Construction of sequences estimating the mixing distribution. *Ann. Math. Statist.*, *39*, 286–288.

Lawson, C. L. & Hanson, R. J. (1974). *Solving Least Squares Problems.* Prentice-Hall, Inc.

Lindsay, B. G. (1995). *Mixture models: theory, geometry, and applications*, Volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics.* Institute for Mathematical Statistics: Hayward, CA.

Macdonald, P. D. M. (1971). Comment on a paper by Choi and Bulgren. *J. R. Statist. Soc. B*, *33*, 326–329.

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

Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons.

Wang, Y. & Witten, I. H. (1999). The estimation of mixing distributions by approximating empirical measures. Technical Report (in preparation), Dept. of Computer Science, University of Waikato, New Zealand.