To Appear in the AAAI Workshop on Probabilistic Search Techniques, 2002

Combining Probabilistic Search, Latent Variable Analysis and Classification Models

Ian Davidson

State University of New York, Albany, 12222
davidson@cs.albany.edu

Abstract

The EM and K-Means algorithms are two popular search techniques that converge to a local minimum of their respective loss functions. The EM algorithm uses partial assignment of instances while the K-Means algorithm uses exclusive assignment. We show that an exclusive random assignment (ERA) algorithm that performs exclusive assignment based on a random experiment can outperform both EM and K-Means for mixture modeling. We show that the ERA algorithm can obtain better maximum likelihood estimates on three real world data sets. On an artificial data set, we show that the ERA algorithm can produce parameter estimates that are more likely to be closer to the generating mechanism. To illustrate the practical benefits of the ERA algorithm we test its ability in a classification context. We propose *Latent Variable Classifier (LVC)* that combines latent variable analysis such as mixture models and classification models such as Naïve Bayes classifiers. For each mixture component (cluster) a classification model is built from those observations assigned to the component. Our experiments on three UCI data sets show LVC's obtain a greater cross-validated accuracy than building a single classifier from the entire data set and probabilistic search out-performs the EM algorithm.

Introduction and Motivation

The K-Means and EM algorithms are popular deterministic approaches used extensively in mixture modeling (clustering) and classification. Their success is in no small part due to their simplicity of implementation that involves a basic two-step process. However, both algorithms converge to local optima of their respective loss functions of distortion and likelihood that is greatly influenced by the initial starting position. This means in practice the algorithms need to be restarted many times from random initial starts in the hope that different parts of the model space will be explored.

We introduce a change to the base two-step process that makes the model space search stochastic in nature. We refer to this type of algorithm as Exclusive Random Assignment (ERA). We try the ERA algorithm for mixture modeling on three UCI data sets and an artificial data set of varying size. Our hope is that the ERA algorithm will find models with a high log likelihood and parameters close to the generating mechanism, when known. To further show the practical benefit of the ERA algorithm we test its performance at classification.

When building classification models it is the defacto standard to build a single model from the entire set of available data. However, many successful practitioners discuss dividing the observations into distinct segments and building models for each (Tuason and Parekh, 2000). This approach is prevalent in many areas of science and is known as the divide and conquer (DAC) strategy (Horowitz and Sahni, 1978). However, typically the DAC strategy requires a large amount of apriori knowledge of the domain and is often problem specific and ad-hoc in nature (Dietterich, 2000). Many problems do not conveniently breakdown according to geographic or temporal boundaries but there may be another implicit breakdown that is not known apriori but could be used in a DAC strategy. We propose *Latent Variable Classifiers (LVC)* that identify the underlying latent variables (using mixture modeling) that exist and build a classification model conditioned on the value of the latent variable. We believe this allows the application of the DAC strategy to problems where no apriori division or segmentation scheme is available and allows a principled and formal way to divide a problem and re-combine the sub-solutions for prediction. We show that for LVC the ERA algorithm outperforms the EM algorithm. The ERA algorithm does not converge to a point estimate; rather it continues to move around the model space. This enables the algorithm to escape local optima.

This paper makes two contributions. It shows that ERA the algorithm can outperform the EM and K-Means algorithm, even with multiple random restarts, in the case of finding models with a maximum likelihood estimation and the smallest Kullback Leibler distance to the generating mechanism for mixture models. It shows that this improvement carries through and has practical significance when making predictions for LVC.

We begin this paper by describing the difference between the K-Means, EM, and ERA algorithms. Next, we introduce our experimental methodology for testing the three algorithms for mixture modeling. Then we introduce Latent Variable Classifiers in graphical form and describe our experimental methodology along with results. We conclude with experimental discussion and future work.

EM, K-Means and ERA Algorithms

The Expectation Maximization (EM) algorithm (Dempster et al, 1977) and the K-Means clustering algorithm (MacQueen, 1967) are two popular search techniques. Both attempt to find the single best model within the model space though it is well known that the definition of "best" varies between the two.

The EM algorithm in the classical inference setting attempts to find the maximum likelihood estimate (MLE). The K-Means algorithm aims to find the minimum distortion within each cluster for all clusters. Both algorithms consist of two primary steps:

1) The observation assignment step: the observations are assigned to classes based on class descriptions.
2) The class re-estimation step: the class descriptions are recalculated from the observations assigned to it.

The two steps are repeated until convergence to a point estimator is achieved. The two approaches in their general form at the operational level differ only slightly. In the first step of the K-Means algorithm, the observations are assigned *exclusively* to the most probable class in a probabilistically formulated problem. In the EM algorithm an observation is assigned *partially* to each cluster, the portion of the observation assigned depending on how probable (or likely) the class generated the object.

In the second step, both algorithms use the attribute values of the observations assigned to a cluster to recalculate its class parameters. For K-Means we recompute the estimates from only those observations that are currently assigned to the class. However, in the EM algorithm if any portion of an observation is assigned to a class then its contribution to the class parameter estimates is weighted according to the size of the portion.

The K-Means algorithm aims to find the minimum distortion within each cluster for all clusters. The distortion is also known as the vector quantization error. The EM algorithm minimizes the log loss which is precisely the local maximum of the likelihood that the model (the collection of classes) produced the data.

Both algorithms converge to a local optimum of their respective loss functions.

The ERA algorithm consists of the same two-steps as the EM and K-Means algorithms. However, in the first step, we use random exclusive assignment by assigning an observation exclusively to one class by a random experiment according to the observation's normalized posterior probabilities of belonging to each cluster. In the second step, we calculate the parameter estimates based on the exclusive assignments. This process repeats as is in K-Means and the EM algorithms. However, this algorithm will not converge to a point estimate as the others, instead it will continue to explore the model space due to its stochastic nature. This allows the possibility of escaping local optima.

In earlier work (Davidson, 2000) we show that this algorithm when used in conjunction with Minimum Message Length (MML) estimators approximates a Gibbs sampler. We attempt to see if the stochastic nature of the algorithm benefits clustering/mixture modeling and classification in a non-MML setting.

Experimental Methodology for Mixture Models

The BREAST-CANCER (BC), IRIS (I), and PIMA (P) data sets available from the UCI collection (Merz et al, 1998) will be the basis of the "real world" empirical study. For each data set we compare maximum likelihood estimates found using EM, K-Means and ERA for models built for k = 1 to 8. We perform each algorithm 50 times from random restarts for one thousand iterations each.

We will then compare the algorithms on artificial data to determine their ability to resolve over-lapping classes on the univariate problem. The artificial data set consists of 1000, 500, 250 and 100 instances drawn randomly from the two component distribution, ~N(0,1) and ~N(3,1) shown in Figure 1.

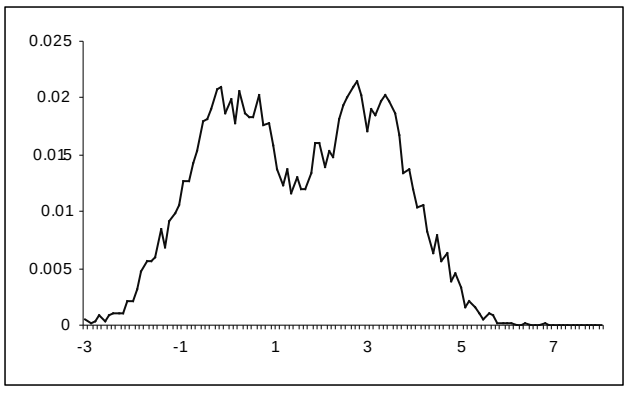

Figure 1: Sample data from a two component mixture ~N(0,1) and ~N(3,1).

Real World Data

For each data set and each technique, we report the average and best results found for the 50 experiments. For all data sets we found that the K-Means algorithm performed significantly worse that the other algorithms. For the PIMA data set (Table 1 and Figure 2) we find that the ERA algorithm consistently finds the best model. However, the EM algorithm's average results are approximately the same for the ERA algorithm except for k=5,6 when the ERA algorithm performs significantly better on average.

The ERA and EM algorithms performed almost identically on the IRIS data set (Table 2 and Figure 3). For breast-cancer data set the average results for the EM algorithm were consistently better than the ERA algorithm but the ERA algorithm consistently found the best model.

k	*EM Ave.*	*KMeans Ave.*	*ERA Ave.*	*EM Best*	*KMeans Best*	*ERA Best*
1	-23028	-23028	-23028	-23028	-23028	-23028
2	-22477	-22563	**-22474**	-22460	-22547	-22460
3	**-21645**	-21734	-21688	-21381	-21386	**-21381**
4	**-21389**	-21424	-21390	-21213	-21252	**-21192**
5	-21315	-21395	**-21216**	-20549	-20603	**-20512**
6	-21132	-21209	**-21090**	-20704	-20752	**-20487**
7	**-20896**	-20938	-20901	-19974	-20002	**-19913**
8	-20661	-20699	**-20638**	**-19889**	-19981	-19895

Table 1: Pima Data Set. Comparison of average and best Maximum Log Likelihood Estimates (MLLE) found using the EM, K-Means, and ERA algorithms. The best result for each category (average and best) is in bold.

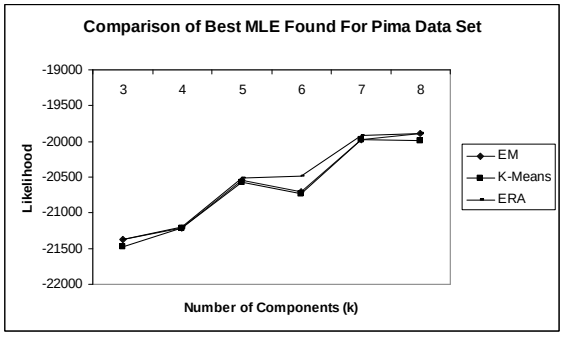

Figure 2: Best MLL estimates found for PIMA data set.

k	*EM Ave.*	*KMeans Ave.*	*ERA Ave.*	*EM Best*	*KMeans Best*	*ERA Best*
1	-741	-741	-741	-741	-741	-741
2	-472	-564	-472	-472	-481	-472
3	-419	-447	-419	-418	-508	-418
4	-402	-455	**-400**	-391	-460	-391
5	-397	-491	**-396**	-385	-482	-385
6	-395	-473	-395	-385	-418	-385
7	**-392**	-393	-393	-384	-402	-384
8	**-391**	-416	-392	-384	-484	-384

Table 2: IRIS Data Set. Comparison of average and best Maximum Log Likelihood Estimates (MLLE) found using the EM, K-Means and ERA algorithms. The best result for each category (average and best) is in bold.

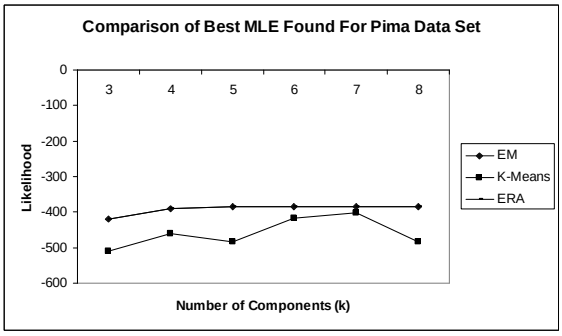

Figure 3: Best MLLE found for the IRIS data set.

k	*EM Ave.*	*KMeans Ave.*	*ERA Ave.*	*EM Best*	*KMeans Best*	*ERA Best*
1	-15156	-15156	-15156	-15156	-15156	-15156
2	**-10130**	-10160	-10143	**-10046**	-10074	-10079
3	**-9645**	-9648	-9650	-9590	-9613	-9590
4	**-9396**	-9462	-9401	**-9331**	-9400	-9341
5	**-9274**	-9281	-9280	-9187	-9266	**-9068**
6	-9138	-9228	**-9137**	-9014	-9112	**-9000**
7	**-8985**	-9022	-8996	-8869	-8947	**-8811**
8	**-8868**	-8878	-8882	-8716	-8783	**-8715**

Table 3: BREAST-CANCER Data Set. Comparison of average and best Maximum Log Likelihood Estimates (MLLE) found using the EM, K-Means, and ERA algorithms. The best result for each category (average and best) is in bold.

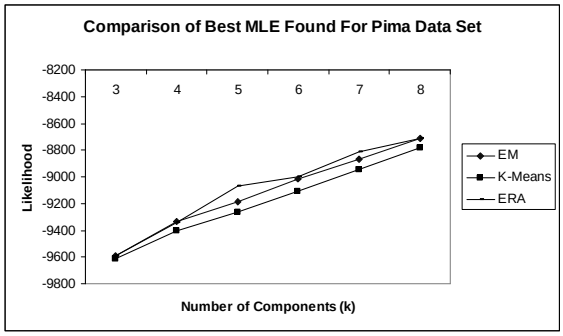

Figure 4: Best MLLE found for BREAST-CANCER data set.

Artificial Data

For the different number of observations and for each technique we report the Kullback Leibler distance of the generating mechanism (Q) to the model with the greatest likelihood (P) for 100 experiments. We place the distance for each experiment into an interval and then calculate the relative frequency of each interval. For all sizes of the data sets we found that the K-Means algorithm performed worse that the other algorithms. For all sizes of the data set except the smallest, the ERA algorithm consistently found parameter estimates that were very close to the generation mechanism as shown in Table 4.

KL Distance (Q,P)	<0.5	[0.5,2.5)	[2.5,7.5)	[7.5,10)	>10
ERA- 100	0%	66%	0%	0%	34%
ERA- 250	45%	0%	15%	8%	32%
ERA- 500	52%	0%	0%	0%	48%
ERA- 1000	54%	0%	0%	0%	46%
EM- 100	0%	0%	0%	0%	100%
EM- 250	9%	9%	3%	12%	67%
EM- 500	0%	0%	0%	0%	100%
EM- 1000	0%	0%	0%	1%	99%
KMEANS- 100	0%	0%	0%	20%	80%
KMeans- 250	0%	0%	0%	0%	100%
KMeans- 500	0%	0%	0%	0%	100%
KMeans-1000	0%	0%	0%	0%	100%

Table 4: ~N(0,1), N(3,1) Data Set. Kullback-Leibler Distances between generating mechanism and model with greatest likelihood found by interval using the EM, K-Means, and ERA algorithms. Random restarts used. The technique and the size of the data set are in the first column.

Figure 5 shows the change in log-likelihood for the ERA algorithm over the course of a single experiment. We see that the log-likelihood can monotonically increase, as is the case with the K-Means and EM algorithms but need not.

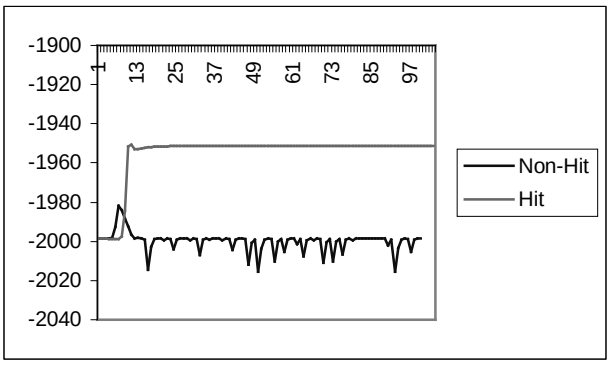

Figure 5: For the ERA algorithm the change in log likelihood as a function of iteration. The "Hit" series refers to the situation where the algorithm finds the generation mechanism as its parameter estimates.

We now examine the algorithms results when the initial parameters are the actual generating mechanism. Even in this situation, the K-Means algorithm was not able to find the generation mechanism as its best parameter estimates. Instead, it finds a solution where the overlapping "tails" of the distributions are reflected inwards.

	Log(P(D|h)	Mu_0	$Stdev_0$	Mu_1	$Stdev_1$	$Q_0\|\|P_0$	$Q_1\|\|P_1$
K-Means	-1954	-0.03	0.92	3.04	0.90	0.05	0.10
EM	-1945	-0.01	1.01	2.93	1.03	0.00	0.04
ERA	-1945	0.00	1.01	2.93	1.03	0.00	0.04

Table 5: Best model found using 1000 instances from ~N(0,1), N(3,1) data set. The generating mechanisms are provided as the algorithms initial solutions.

We have not explicitly determined the loss function of the ERA algorithm but have shown that empirically it finds good estimates for the likelihood function of the data.

Combining ERA With Simulated Annealing

At each iteration of the ERA algorithm each observation is assigned to a class j with a probability $p_j/(p_1+\ldots p_k)$. We may raise these normalized probabilities to the power of $1/c$ where c is a control parameter commonly known as the temperature in the simulated annealing community (Aarts, 1989). This allows us to perform simulated annealing, an approach that asymptotically has been shown to converge to the global optima (Aarts, 1989). However, in practice the approach is useful for obtaining good local optima as it allows the search process to more readily leave sub-optimal solutions. In our experiments we have effectively kept c a constant at 1. We have not performed any systematic experiments combing the ERA algorithm with simulated annealing.

Definition of Latent Variable Classifiers

Dividing the observations by their geographic location or time-period are examples of an obvious DAC strategy. The premise behind DAC is to divide a complex problem into a number of easier sub-problems. By solving and then combining the sub-solutions, an overall solution is obtained (Horowitz and Sahni, 1978). For example, the DAC strategy in the field of combinatorial optimization makes combinatorially large problems tractable. If we are trying to find an optimal tour in a traveling salesman problem of thirty cities in the U.S.A., the number of possible tours is 30!. If we determined the tour could be divided into an east and west coast tours of 15 cities each, with a connecting trip, then we have divided the problem into two sub-problems which together have a potential number of tours of 2.15!, a considerable saving. However, a reasonable DAC strategy for many problems is not always known.

Latent variable analysis (Everitt, 1984) attempts to find unknown classes or entities to better explain commonly occurring patterns. For example, the notions of diseases are latent classes that describe commonly occurring symptoms. The identification of latent entities can be critical in decision-making. More specific treatment regimes can be administered to a patient given they are identified as having a disease. Our aim is to capture this type of two-level reasoning for classification problems. A common form of latent variable analysis is mixture models which attempt to identify latent classes of observations.

Latent variable classifiers consist of the random variables $X = \{X_1 \ldots X_n\}$, Y and C representing the independent attributes, explicit class attribute and latent class attribute respectively. We can consider Y to represent an overt or explicit externally provided class. These random variables take on specific values for an observation written as $x_1 \ldots x_n$, y and c. Our overall aim is to accurately predict y for new unseen observations. The common naïve Bayes classification model specifies a conditional probability for y given the values of $x_{1\ldots n}$ as shown in (1). The independent variables, $X_1 \ldots X_n$, are independent of each other given knowledge of Y. A similar distribution can be specified for regression models where y is continuous. The naïve Bayes classification model is show graphically in Figure 6.

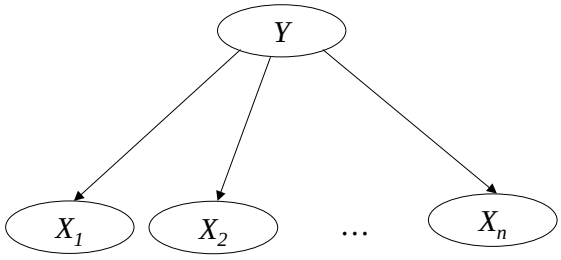

Figure 6: Graphical Model of a Naive Bayes Classifier

To make better predictions we condition predictions on both the latent variable, C, and independent variables X. The independent variables, $X_1 \ldots X_n$, are independent of each other given knowledge of C the latent attribute.

$$P(y \mid x) = P(y).\prod_{i=1}^{n} P(y \mid x_i) \qquad (1)$$

The number of latent classes, k, is typically explicitly given apriori though it may be part of one of the parameters to estimate (Davidson, 2000). The joint distribution of an observation and a latent class is given in equation (2)

$$P(c, x) = P(c)\prod_{i=1}^{n} P(x_i \mid c) \qquad (2)$$

If we assume the components of X are independent of each other given knowledge of C then the graphical model of mixture models is identical to the Naïve Bayes classifier except the top node is labeled C. Combining the two models allows us to specify the latent variable classifier in equation (3).

$$P(y \mid c, x) = \sum_{k=1}^{K} P(x \mid c_k) P(y \mid c_k, x) \qquad (3)$$

In effect, each latent class has its own classifier that makes a prediction for y. The overall prediction for y is weighted by the probability of the observation belonging to the latent class. The *LVC* is shown graphically in Figure 7.

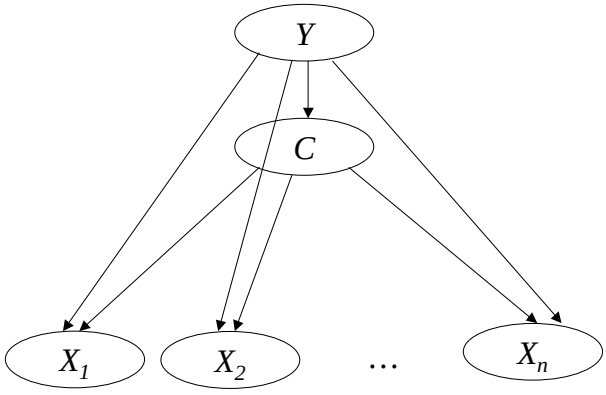

Figure 7: Graphical Model of a Latent Variable Classifier.

We estimate the parameters of the latent class and classification models simultaneously. This is extremely quick if the classification model is a simple learner such as the naive Bayes classifier but in practice any classifier can be used. The parameter estimation technique regardless of the algorithm will be identical to that of a mixture modeler except that there will an additional two terms, the probability of the extrinsic class and the probability of the extrinsic class given the observation values.

Experimental Methodology for LVC

The BREAST CANCER (BC), IRIS (I), and PIMA (P) data sets available from the UCI collection (Merz et al, 1998) will be the basis of this empirical study. For each data set we try the EM and ERA algorithms. The K-Means algorithm is not tried due to its poor prior performance.

For each data set we compare the cross-validated accuracy of models built for k = 1 to 8. Of course k=1 is just building a single model from the entire data set and is our base line for comparison. We expect that for k>2 the overall cross-validated accuracy will be an improvement over k=1. For each value of k we perform ten-fold cross validation twenty time reporting the mean of the ten-fold cross-validated accuracy and the standard deviation of the cross-validated accuracy. The mean and standard deviation of the accuracy over all folds is used to calculate the Mean Square Error (MSE).

Experimental Results

K	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Mean	66.0	75.0	72.8	68.4	68.2	68.7	68.9	69.4
MSE	11.7	6.48	7.89	10.40	10.5	10.1	10.0	9.63

Table 6: The accuracy (%) for the pima data set using the EM algorithm.

K	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Mean	66.0	76.9	74.0	69.2	69.3	72.7	71.3	71.9
MSE	11.7	4.49	5.75	8.54	8.4	8.1	7.80	8.0

Table 7: The accuracy (%) for the pima data set using the ERA algorithm.

k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Mean	53.4	72.7	94.5	95.6	95.8	96.2	96.1	96.2
MSE	23.0	9.23	0.87	0.44	0.47	0.41	0.42	0.44

Table 8: The accuracy (%) for the iris data set using the EM algorithm

k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Mean	53.4	73.7	95.5	96.6	96.6	96.8	96.7	97.6
MSE	23.0	6.31	0.67	0.21	0.24	0.33	0.32	0.32

Table 9: The accuracy (%) the iris data set using the ERA algorithm

K	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Mean	92.2	96.4	95.9	96.2	96.4	96.3	96.1	95.9
MSE	0.72	0.23	0.70	0.37	0.26	0.18	0.28	0.24

Table 10: The accuracy (%) for the breast cancer data set using the EM algorithm

K	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Mean	92.2	96.0	97.5	96.0	96.1	96.4	96.5	96.2
MSE	0.72	0.22	0.17	0.23	0.21	0.18	0.17	0.19

Table 11: The accuracy (%) for the breast cancer data set. using the ERA algorithm

Discussion

We see that the LVC obtained more accurate results than building a single model from the entire data set (k=1). Increases in accuracy as a proportion of room for improvement where 26.4%, 91.8% and 53.8% for the Pima, Iris and Breast-Cancer data sets when using the EM algorithm to find the maximum likelihood estimates. This increased to 32%, 94.5% and 67.9% respectively when using the ERA algorithm. For all data sets for nearly every value of k the MSE obtained using the ERA algorithm is less than the MSE obtained using the EM algorithm. From our experimental results for LVC we see that the ERA algorithm's improved ability at finding a better maximum likelihood estimate translate into more accurate predictions and lower mean square errors.

Conclusion and Future Work

By making a minor change to the EM and K-Means algorithms a random search algorithm that does not converge to a point estimate occurs. We illustrated how the ERA algorithm can find better maximum likelihood estimates for mixture modeling and more accurate predictive results and a lower MSE for LVCs.

As the ERA algorithm does not converge to a single point estimate but keeps on exploring the model space it is worth exploring if making predictions from a number of models found as it moves through the model space yields even better results. We also plan to determine if raising the posterior probabilities to the power of $1/c$ and slowly decreasing c yields better results. This is akin to performing simulated annealing. It is known what loss function the EM and K-Means algorithms minimize and we intend to determine the precise loss function for the ERA algorithm. Finally, we will empirically compare our approach to stochastic variations of the EM algorithm such as SEM and MCEM.

References

Aarts, E.H.L., Korst, J., *Simulated Annealing and Boltzmann machines: a Stochastic Approach to Combinatorial Optimization and Neural Computing*, Anchor Press, 1989.

Davidson, I., Minimum Message Length Clustering Using Gibbs Sampling, Uncertainty in A.I., 2000.

Dietterich T., "The Divide-and-Conquer Manifesto", *Algorithmic Learning Theory*, pp. 13-26, 2000.
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.

Everitt, B.S., *An Introduction to Latent Variable Models*, Chapman and Hall, 1984.

Horowitz E. and Sahni S., Fundamentals of Computer Algorithms, Computer Science Press, Inc, 1978.

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

Merz C, Murphy P., Machine learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html].
Irvine, CA: University of California, Department of Information and Computer Science. 1998.

Tuason, N. and Parekh, R. "Mining Operational Databases to Predict Short-Term Defection Among Insured Households." *Proceedings of the 2000 Advanced Research Techniques Forum (ART'2000)*, Monterey, CA. Jun 4-7, 2000