KDD 2000 Boston MA USA

Visualization and Interactive Feature Selection for Unsupervised Data

Jennifer G. Dy
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907
dy@ecn.purdue.edu

Carla E. Brodley
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907
brodley@ecn.purdue.edu

ABSTRACT

For many feature selection problems, a human defines the features that are potentially useful, and then a subset is chosen from the original pool of features using an automated feature selection algorithm. In contrast to supervised learning, class information is not available to guide the feature search for unsupervised learning tasks. In this paper, we introduce Visual-FSSEM (Visual Feature Subset Selection using Expectation-Maximization Clustering), which incorporates visualization techniques, clustering, and user interaction to guide the feature subset search and to enable a deeper understanding of the data. Visual-FSSEM, serves both as an exploratory and multivariate-data visualization tool. We illustrate Visual-FSSEM on a high-resolution computed tomography lung image data set.

1. INTRODUCTION

Most research in unsupervised clustering assumes that when creating the target data set, the data analyst in conjunction with the domain expert was able to identify a small relevant set of features. If this is not the case, then choosing a subset of these features will lead to better performance of the data mining algorithm when redundant and/or irrelevant features are removed. Moreover, reducing the dimensionality of the data is desirable for reducing computation time and increasing the comprehensibility of the results of the clustering algorithm.

For unsupervised learning, the goal is to find the feature subset that best discovers "natural" groupings from data (also known as clustering). To select an optimal feature subset, we need a measure to assess cluster quality. The choice of performance criterion is best made by considering the goals of the domain. In studies of performance criteria a common conclusion is: "Different classifications [clusterings] are right for different purposes, so we cannot say any one classification is best." – Hartigan, 1985 [9].

Feature subset selection coupled with unsupervised learning is a difficult task, because class labels are not available to guide the feature search. The problem is made more difficult when we do not know the number of clusters, k. Figure 1 illustrates this problem. In two dimensions (shown on the left) there are three clusters, whereas in one-dimension (shown on the right) there are only two clusters. The difficulty is in knowing which is better. Ultimately the only way to decide is to use a criterion tied to the final use of the clustering.

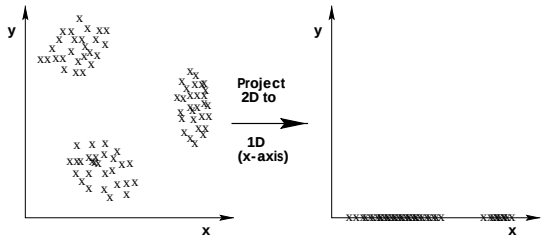


Figure 1: The number of cluster components varies with dimension.

In [5], we studied the problem of totally automated feature subset selection for unsupervised clustering. We introduced a new algorithm, FSSEM, which wraps feature subset selection around the expectation maximization [3] of a finite Gaussian mixture model (which we refer to as EM clustering). In this paper, we introduce Visual-FSSEM (Visual Feature Subset Selection using EM Clustering), which incorporates visualization techniques, clustering and user interaction to guide the feature subset search. Previous investigation on automated feature selection for unsupervised learning reveals that no single feature selection criterion is best for every application. Visual-FSSEM enables the user to evaluate clusters and features based on visual perception and evaluation measures such as scatter separability, maximum likelihood, cluster entropy and probability of error. Our implementation allows the user to perform both forward and backward searches and permits backtracking. The ability to visualize clusters in different feature subsets engenders a deeper understanding of the data. Visual-FSSEM, thus, serves both as an exploratory and multivariate-data visualization tool.

In Section 2, we introduce Visual-FSSEM, an interactive feature selection visualization environment. Specifically, we de-

scribe the chosen visualization method, we review FSSEM, and we describe three ways in which the user can guide the feature subset selection search. In Section 3, we illustrate the potential benefits of Visual-FSSEM on an HRCT-lungs (high resolution computed tomography image of the lungs) data set. Finally in Section 4, we conclude our paper and provide directions for future work.

2. VISUAL-FSSEM

Visual-FSSEM (Visual Feature Subset Selection using EM clustering) is an interactive visualization environment for feature selection for unsupervised data. As such it serves as a tool for discovering clusters on different feature subsets. Our completely automated version, FSSEM, performs a greedy sequential forward or backward search for the best feature subset as measured by the chosen performance criterion. In Visual-FSSEM, we give the user the ability to guide this process. In the remainder of this section we first describe our visualization method. We then review FSSEM and two of the four available clustering criteria. We conclude with a description of three ways in which the user can take part in the search for the best subset through interaction with Visual-FSSEM.

2.1 Visualization Method

When the data has one dimension, we display its estimated probability distribution. When the data has two dimensions, it can be displayed using scatterplots [10]. When the data has more than three dimensions, we need to apply visualization techniques. There are several multivariate-data visualization techniques. These methods can be categorized into geometric, icon-based, hierarchical, and pixel-oriented techniques. Keim and Kriegel [11] provide a short description on each of these methods.

Currently in Visual-FSSEM, we apply a geometric visualization technique. To display the data and clusters, we project the data to 2-D and display the data on a scatterplot. We choose linear discriminant analysis (LDA) [8] to project the data, because it finds the linear projection that maximizes cluster scatter. We project the original data X onto Y using a linear transformation: $Y = A^T X$, where X is a $d \times n$ matrix representing the original data with n samples and d features. Each column of X represents a single instance of dimension d. Y is an $m \times n$ matrix representing the projected data. Each column of Y is a projected instance onto m dimensions. A is a $d \times m$ matrix whose columns correspond to the largest m eigenvectors of $S_w^{-1} S_b$ (see Section 2.3). When $S_w^{-1} S_b$ turns out to be singular, we project the data using its principal components (i.e., transformation matrix A corresponds to the m eigenvectors of the data covariance) [8]. Principal components analysis (PCA) projects the data in the direction of the largest variance. We prefer LDA because it shows the projection that provides the greatest cluster separation. For example, consider the data shown in Figure 2, when projecting this two dimensional data to one dimension, LDA chooses the projection perpendicular to the largest cluster separation (the y-axis in this case). PCA, on the other hand, selects the projection representing the largest variance of the data (the x-axis in this case).

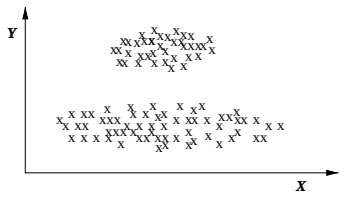


Figure 2: LDA chooses the y-axis, whereas PCA chooses the x-axis.

2.2 An Overview of FSSEM

FSSEM wraps feature subset selection around the clustering algorithm. The basic idea is to search through feature subset space, evaluating each subset, F_t, by first clustering in space F_t using EM clustering and then evaluating the resulting clusters and feature subset using the chosen feature selection criterion. An exhaustive search of the 2^d possible feature subsets (d is the number of available features) for the subset that maximizes our selection criterion is computationally intractable. Therefore, either sequential forward or backward elimination search is applied [8].

FSSEM assumes that the data comes from a finite mixture model of multivariate Gaussians. We apply the EM algorithm to estimate the maximum likelihood mixture model parameters and the cluster probabilities of each data point [3]. The EM algorithm can become trapped at a local maximum, hence the initialization values are important. We used the sub-sampling initialization algorithm proposed by Fayyad et al. [7] with 10% sub-sampling and $J = 10$ sub-sampling iterations. After initializing the parameters, EM clustering iterates until convergence (i.e., the likelihood does not change by 0.0001) or up to n (default 50) iterations whichever comes first. We limit the number of iterations because EM converges asymptotically, i.e., convergence is very slow near a maximum. EM estimation is constrained away from singular solutions in parameter space by limiting the diagonal elements of the component covariance matrices Σ_j to be greater than $\delta = 0.000001\sigma^2$, where σ^2 is the average of the variances of the unclustered data.

When we are not given the number of clusters, we apply FSSEM-k, which performs a feature subset selection search wrapped around EM-k (EM clustering with order identification). For a given feature subset, we search for k and the clusters. EM-k currently applies Bouman et al.'s method [1], which adds a minimum description length penalty term to the log-likelihood criterion. A penalty term is needed because the maximum likelihood estimate increases as more clusters are used. Without the penalty, the likelihood is at a maximum when each data point is considered as an individual cluster.

2.3 Feature Selection Criteria

In our current implementation we give the user the ability to choose from four well-known performance measures: scatter separability, maximum likelihood, cluster entropy and probability of error. Which criterion is best depends on the goals of the data mining task. Due to space limitations, we describe two criteria here and the others can be found in [4].

2.3.1 Scatter Separability Criterion

Among the many possible separability criteria, we choose the $trace(S_w^{-1}S_b)$ criterion because it is invariant under any nonsingular linear transformation [8]. S_w is the within-class scatter matrix and S_b is the between class scatter matrix, and they are defined as follows:

$$\begin{aligned} S_w &= \sum_{j=1}^{k} \pi_j E\{(X-\mu_j)(X-\mu_j)^T|\omega_j\} = \sum_{j=1}^{k} \pi_j \Sigma_j \\ S_b &= \sum_{j=1}^{k} \pi_j (\mu_j - M_o)(\mu_j - M_o)^T \\ M_o &= E\{X\} = \sum_{j=1}^{k} \pi_j \mu_j \end{aligned}$$

where π_j is the probability that an instance belongs to cluster ω_j, X is a d-dimensional random feature vector representing the data, k the number of clusters, μ_j is the sample mean vector of cluster ω_j, M_o is the total sample mean across all data points or instances in the data set, Σ_j is the sample covariance matrix of cluster ω_j, and $E\{\cdot\}$ is the expected value operator. S_w measures how scattered the samples are from their cluster means and the average covariance of each cluster. S_b measures how scattered the cluster means are from the total mean. We would like the distance between each pair of samples in a particular cluster to be as small as possible and the cluster means to be as far apart as possible with respect to the chosen similarity metric. $S_w^{-1}S_b$ is S_b normalized by the average cluster covariance. Hence, the larger the value of $trace(S_w^{-1}S_b)$ is, the larger the normalized distance between clusters is which results in better cluster discrimination. Separability is a general criterion that can be used for any clustering algorithm, but is biased towards clusters with cluster means that are far apart, and biased against clusters with equal means even though these clusters have different covariance matrices.

2.3.2 Maximum Likelihood (ML) Criterion

Maximum likelihood (ML) measures how likely our data are given the parameters and the model. Thus, it tells how well our model fits the data. Therefore, in addition to the criterion for clustering, we can employ ML to find the feature subset that models the data best using EM clustering. The maximum log-likelihood of our data, X, is $\log ML = \max_\Phi \log(f(X|\Phi)) = \max_\Phi \sum_{i=1}^{N} \log(\sum_{j=1}^{k} \pi_j f_j(X_i|\theta_j))$ where $f_j(X_i|\theta_j)$ is the probability density function for class j, π_j is the mixing proportion of class j (prior probability of class j), N is the number of data points, k is the number of clusters, X_i is a d-dimensional random data vector, θ_j is the set of parameters for class j, $\Phi = (\pi, \theta)$ is the set of all parameters and $f(X_i|\Phi)$ is the probability density function of our observed data point X_i given the parameters Φ. We choose the subset that maximizes this criterion.

2.4 Interactive Search

Currently, we have implemented three modes of interaction with FSSEM. The first method allows the user to view the clustering results of a sequential forward or backward search at each step and select the best subset. For example, given a data set described by ten features, if the user selects an SFS (sequential forward search) and the trace criterion, then Visual-FSSEM presents the user with ten (or less as specified by the user) scatterplots; one for each step in the SFS. The user can then select which of these best clusters the data. This first method is motivated by the fact that many of the criteria commonly used in clustering are biased either toward lower or higher dimensionality [5]. Therefore automated criterion-based feature selection may not lead to the best clustering. For example, the separability criterion increases as the number of features (or dimension) increases when the clustering assignments remain the same. The separability measure is biased this way because $trace(S_w^{-1}S_b)$ is basically adding d (the number of dimension) terms. Fukunaga [8] proved that a criterion of the form $X_{d\times 1}^T S_{d\times d}^{-1} X_{d\times 1}$ monotonically increases with dimension, d, assuming the same clustering assignment. [4] relates $trace(S_w^{-1}S_b)$ to this proof. Allowing the user to select the best subset ameliorates this problem. In Section 3 we show an example, in which the user selects a better cluster than FSSEM run with the $trace$ criterion.

The second method allows the user to intermingle different criteria during the search process. At each step of SFS or SBE the user is presented with the best feature to add/delete as chosen by each criterion. Our previous investigation [5] reveals that no single criterion measure works best for all applications. This option presents the user the best feature chosen by each criterion in an add/delete step, allowing the user to select the cluster one prefers to explore at each step.

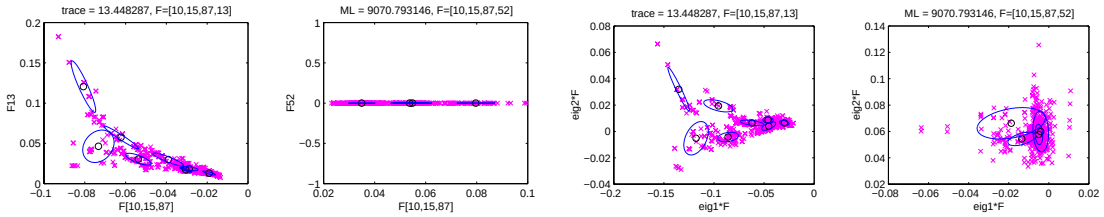
Finally, the third method allows the user to perform a beam search of the feature subset space. At each step of the SFS/SBE search, the user is presented with the b best features to add/delete as ranked by the chosen clustering criterion. The user can then select the best of these b features. FSSEM then adds/deletes this feature to/from the set of selected features. The user is given the capability to backtrack in the search space. The utility of this approach is that the user may have a preference for some features based on his/her domain knowledge and knowing how each ranks with respect to the other candidate features for addition/deletion can help focus the search. The user specifies the value for b.

In all cases, the user can select an arbitrary initial feature subset and limit the pool features to search from. The user selects the direction forward or backward in every step and the number of steps. When performing a forward sequential search we give the user an alternative visualization option. To select the nth feature for addition, we project the $n-1$ previously selected features to one dimension, and use the new feature as the other dimension. This enables the user to view the effect of the feature to be added.

3. VISUAL-FSSEM APPLIED TO THE LUNG IMAGE DATA SET

We illustrate Visual-FSSEM on the HRCT-lung data set [6]. HRCT-lung consists of 500 instances. Each of these instances are represented by 110 low-level continuous features measuring geometric, gray level and texture features. The data is classified into five disease classes (Centrilobular Emphysema, Paraseptal Emphysema, EG, IPF, and Panacinar). Feature selection is important for this data set, because EM clustering using all the features results in just one cluster. Furthermore, the data can also be grouped into different possible groupings (for example, disease class, severity or by the perceptual categories that are used to train radiologists).

The first example of Visual-FSSEM illustrates a situation in which the user has selected SFS and is examining the best feature chosen by FSSEM-k in conjunction with each of the



(a) (b)

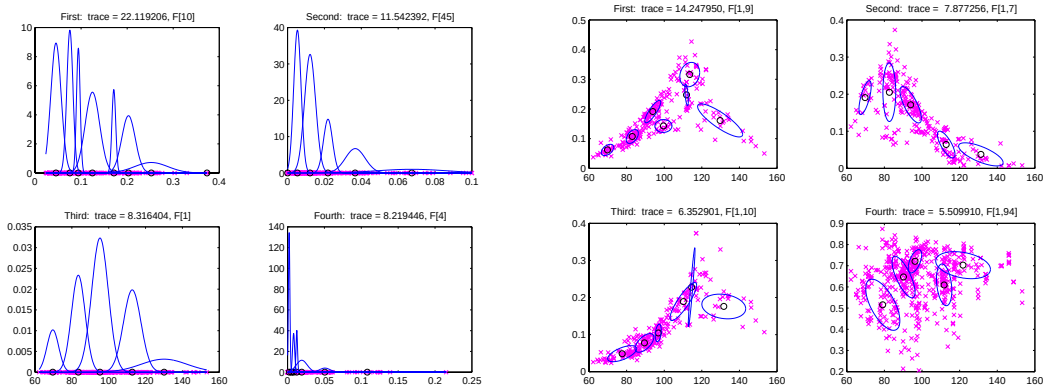
Figure 3: The two alternative displays for forward feature selection in a method 2 interaction: (a) scatterplot of the new feature versus the LDA of the previous feature set, and (b) scatterplot of the LDA on the new feature set.

Figure 4: Illustration of method 1 interaction: Scatterplot of HRCT-lung data in feature subsets F[45], F[45 ,93], F[45,93,88,102,110], F[45,93,88,102,110,84], and F=[45,93,88,102,110,84,76,29,1,68] in the sequence the features are selected.

criteria (interaction method 2). In Figure 3 we display the results for only two of the four criteria due to space limitations. At this point in the search we are determining which feature to add to the set 10, 15 and 87 (a texture homogeneity and two global gray level histogram features). This set was chosen by the three proceeding SFS steps guided by the scatter separability criterion. In the figure, the clustering results parameterized by the means and covariance matrices are displayed as $\circ$'s and ellipses respectively. FSSEM-k selects a different best feature for scatter and ML. In Figure 3a, the x-axis for both scatterplots represents a linear discriminant transformation of the previously selected three features to one-dimension. The y-axis represents the new feature being considered. The feature chosen by *trace* adds the greatest separability. Note that feature y adds separation to the current subset, if clusters can be separated in the y dimension (or clusters are spread in the y-axis). The feature chosen by the ML criterion does not aid in finding new clusters. The y-value is approximately constant for all data points. Note that the different feature subsets resulted in different numbers of clusters (eight for the scatter and four for the ML criterion). Figure 3b shows the corresponding displays using LDA projecting on to two-dimensions. Note that the plot on the ML chosen set is transformed on the first two principal components because the within cluster variance was close to singular. Figure b confirms that the *trace* plot results in a higher separability than the ML plot, since the *trace* plot has a lower cluster (ellipse) overlap than ML. The *trace* clustering results in a classification error of 24.2% (if data was hard clustered and judged based on the five disease classes). Note that class labels are not used during feature search or during clustering. We use the labels here only to aid in our evaluation of the results.

The second example of visual-FSSEM illustrates the use of domain knowledge in the interactive search. We divided the original pool of features into the following ten feature groups: gray level means and standard deviation features (1, 2, 19-21, 46-51), gray histogram features (3-18, 30-45, 52-67), area histogram (22-29), texture energy (68-72), texture entropy (73-77), texture homogeneity (78-87), texture contrast (88-92), texture correlation (92-97), texture cluster (98-102), and edginess features (103-110). We then selected the single best feature from each class of features by running FSSEM-k ten times (once on each feature group) with the *trace* criterion. Limiting our feature subset search to these ten features, we apply Visual-FSSEM (using the first method of interaction) to explore the structure of this "lump" of data. Figure 4 shows the first, second, fifth, sixth and tenth SFS steps. Note that an automated FSSEM-k using *trace* would select feature subset $F[45]$, since it has the largest *trace*. However, to a user the other clusterings might look more interesting. Even though the criterion value in feature subset $F[45]$ is the largest, it obtained the lowest class error of 37.2% which is equivalent to that obtained by a majority rule (there are 314 instances belonging to the largest class). The set with six features $F[45, 93, 88, 102, 110, 84]$ resulted in a class error of 26.8%.

Figures 5a and b illustrate an interactive beam search for which $b = 4$. We start with the empty set and apply forward selection. Figure 5a shows the four best individual features as ranked by the *trace* criterion. In this case we choose to explore starting from feature 1, because the data is nicely distributed among the clusters in feature 1 (lower left). Moreover, our domain knowledge supports this choice. Feature 1 is the mean gray level of the lung. The other three features correspond to gray level histogram bins. The mean value typically captures more information than a single histogram bin. Figure 5b shows the next step in the search when the highest ranking pairs with respect to the trace criterion are displayed. All four reveal interesting structures. The class error obtained by these four clusterings are 26.6%, 26.8%, 26.2% and 32% according to rank. This method of



(a) (b)

Figure 5: Display for the beam search method three interaction: (a) scatterplots on the best four single features, and (b) scatterplots on the best four feature pair.

interaction permits backtracking and can continue as long as the user wishes to explore other cluster structures.

4. CONCLUSION

We have illustrated that feature selection and clustering techniques can be utilized to aid in the exploration and visualization of high-dimensional unsupervised data. At the same time, visualization aids the data miner in selecting features. The ability to visualize different cluster structures on different feature subsets provides insight into understanding the data. Visual-FSSEM allows the user to select any feature subset as a starting point, select the pool of features to choose from, search forward or backward, and visualize the results of the EM clustering. Four clustering criteria are available to guide the search: separability, maximum likelihood, cluster entropy and probability of error. A more general approach would incorporate different clustering algorithms and allow the user to select the clustering method to apply. We chose to display the data and clusterings as 2-D scatterplots projected to the 2-D space using linear discriminant analysis. Other visualization techniques may be applied. An interesting technique is the manual projection method by Cook and Buja [2] which allows the user to manually control the projection of data on to different views. It would also be interesting to allow the user to cluster data within clusters (hierarchical clustering) so as to zoom in on dense groupings.

5. ACKNOWLEDGMENTS

The authors wish to thank the ML-lunch group at Purdue for helpful comments. This research is supported by NSF Grant No. IRI9711535, and NIH Grant No. 1 R01 LM06543-01A1.

6. REFERENCES

[1] C. A. Bouman, M. Shapiro, G. W. Cook, C. B. Atkins, and H. Cheng. Cluster: An unsupervised algorithm for modeling gaussian mixtures. In *http://dynamo.ecn.purdue.edu/∼bouman/software/cluster*, October 1998.

[2] D. Cook and A. Buja. Manual controls for high-dimensional data projections. *Journal of Computational and Graphical Statistics*, 6(4), 1997.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[4] J. G. Dy. In *Preliminary Report: Feature Selection for Unsupervised Learning*, Unpublished manuscript, School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, 1999.

[5] J. G. Dy and C. E. Brodley. Feature subset selection and order identification for unsupervised learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000.

[6] J. G. Dy, C. E. Brodley, A. Kak, C.-R. Shyu, and L. S. Broderick. The customized-queries approach to CBIR using EM. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 400–406, Fort Collins, CO, June 1999. IEEE Computer Society Press.

[7] U. Fayyad, C. Reina, and P. S. Bradley. Initialization of iterative refinement clustering algorithms. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 194–198, New York, August 1998. AAAI Press.

[8] K. Fukunaga. *Statistical Pattern Recognition (second edition)*. Academic Press, San Diego, CA, 1990.

[9] J. A. Hartigan. Statistical theory in clustering. *Journal of Classification*, 2:63–76, 1985.

[10] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall, 4 edition, 1998.

[11] D. A. Keim and H.-P. Kriegel. Visualization techniques for mining large databases: a comparison. *IEEE Transactions on Knowledge and Data Engineering TKDE'96, Special Issue on Data Mining*, 8(6):923–938, 1996.