Choosing Attributes for Analysis: Effective and Efficient Feature Selection for Exploring High-dimensional Data

DIANSHENG GUO

GeoVISTA Center & Department of Geography, Pennsylvania State University
302 Walker Building, University Park, PA 16802, USA
Phone: 1-814-865-3433, E-mail: dguo@psu.edu

Abstract

A large number of *attributes*[1] can cause serious problems for almost all data analysis methods, especially for exploratory clustering methods. First, irrelevant or noisy attributes often exist, especially when the data is compiled from different data sources. Using all original attributes to detect *relationships* or *patterns*[2] can be ineffective and even counter-productive. In other words, irrelevant attributes may hide rather than uncover patterns. Second, unknown relationships may exist in different *subspaces*[3], i.e., different relationships may involve different subsets of the original attributes. Therefore, it is critical for high-dimensional data analysis methods to be able to find interesting subspaces (not just a single "optimal" subspace) that contain good patterns. To effectively and efficiently identify interesting subspaces, this paper presents a novel feature selection method for exploring large and high-dimensional geographic data. A calculation of maximum conditional entropy for a 2-D data space is developed to measure the "goodness of clustering". For a data set (S) of *dimensionality*[4] d and size n, a matrix of pair-wise maximum conditional entropy values of all 2-D subspaces in S is derived. A visualization technique is also developed to visualize the entropy matrix and present a holistic picture of various relationships among dimensions. Interesting multidimensional subspaces can then be automatically extracted or interactively identified. Experiments with synthetic data and applications with real data show that the method is very useful for exploring high-dimensional data. The computational complexity of the method is $O(d^2 n \log n)$.

Keyword: feature selection, dimension reduction, multivariate clustering, high-dimensional visualization, geographic knowledge discovery, geovisualization

1. Introduction

Due to the increasing ability for data collection and the increasing complexity of problems to be addressed, datasets to be analyzed are often very large (e.g., >10,000 observations) and

[1] *Attributes*, *dimensions*, *features*, and *variables* are used interchangeably in this paper.
[2] *Patterns* and *relationships* are used interchangeably in this paper. Specifically they refer to clusters of various shapes, sizes, and densities.
[3] A *subspace* is a subset of those attributes in the original data space.
[4] The *dimensionality* of a data space is the number of attributes it contains.

have many (e.g., >50) attributes. Such data sets are commonly compiled from multiple data sources, which might be collected for different purposes. By putting them together for analysis, we are hoping to find *unknown* multivariate relationships or patterns. Nowadays, there is less and less difference between spatial datasets and non-spatial datasets because: (**1**) spatial datasets always have many non-spatial attributes, and (**2**) traditional non-spatial applications are becoming increasingly geography-aware. For example, commercial transaction data are joined with locations (addresses) to study the variation of shopping patterns across regions; clinic records are geo-coded (with confidential considerations) to study the environmental effects on the incidence of a certain disease.

Such high-dimensional data sets can cause serious problems for almost all data analysis methods, especially for clustering techniques (Hinneburg and Keim 1999). First, the quality and relevance of the original attributes can vary dramatically. Irrelevant or noisy attributes often exist. Using all original attributes in the data to derive a similarity or distance between data objects and identify clusters can be ineffective and even counter-productive. Second, for real-world applications, strong patterns that involve all original (maybe 50+) attributes are unlikely. The average density of points anywhere in the original data space is likely to be very low due to the "*curse of dimensionality*" (Hinneburg and Keim 1999; Duda, Hart et al. 2001). Third, unknown relationships or patterns may exist in different *subspaces*, i.e., different relationships may involve different subsets of the original attributes. Therefore, it is critical for high-dimensional data analysis methods to be able to find interesting subspaces (and ignore irrelevant attributes) and search relationships or patterns in each of them.

To briefly elaborate the above problem of subspace pattern, let's look at a simple 3-D data set (Figure 1). It has 3 attributes: X, Y, and Z, and contains 9000 data objects (or points). X and Y have a strong positive linear relationship (in this paper it is regarded as a special case of a cluster); Y and Z have a strong negative linear relationship; while X and Z have no significant relationship. Figure 1 shows all three 2-D scatterplots and the 3-D scatterplot. For a data analysis method, it may easily find those patterns in subspace (Abbott, Matkovsky et al. 1998) and subspace {Y, Z}, but it is difficult to find a good pattern in the original data space {X, Y, Z}. In a high-dimensional data set, subspaces that have strong patterns may either overlap with each other in terms of sharing one or more attributes, e.g. subspace {X, Y} and subspace {Y, Z}, or are totally different, sharing no attribute.

The high dimensionality problem is often bypassed by requiring the user (who should be an expert on the application problem) to specify a subspace (or several subspaces) for analysis. However, depending on the user to choose attributes for analysis makes it impossible to find *unexpected* patterns, while finding such *unexpected* patterns is one of the main purposes of data mining and knowledge discovery (Fayyad, Piatetsky-Shapiro et al. 1996). Moreover, given the high dimensional data (which may have hundreds of attributes) currently available, the above manual approach for choosing attributers is inefficient and sometime impractical for exploring various *unknown* patterns in the data.

Traditional feature selection methods have been studied in the area of supervised classification (Liu and Motoda 1998). Several unsupervised feature selection methods have recently been developed to select an "optimal" subset of attributes (Dy and Brodley 2000b; Dy and Brodley 2000a), or produce a pool of "good" subsets of attributes (Kim, Street et al. 2000), for unsupervised clustering. Since clusters can also exist in different subspaces, it can be ineffective or impossible to find a single "optimal" subset of dimensions for all clusters (Agarwal, Procopiuc et al. 1999). Several subspace clustering methods have been developed to detect clusters residing in different subspaces (Agrawal, Gehrke et al. 1998; Cheng, Fu et al. 1999a; Aggarwal and Yu 2000; Procopiuc, Jones et al. 2002).

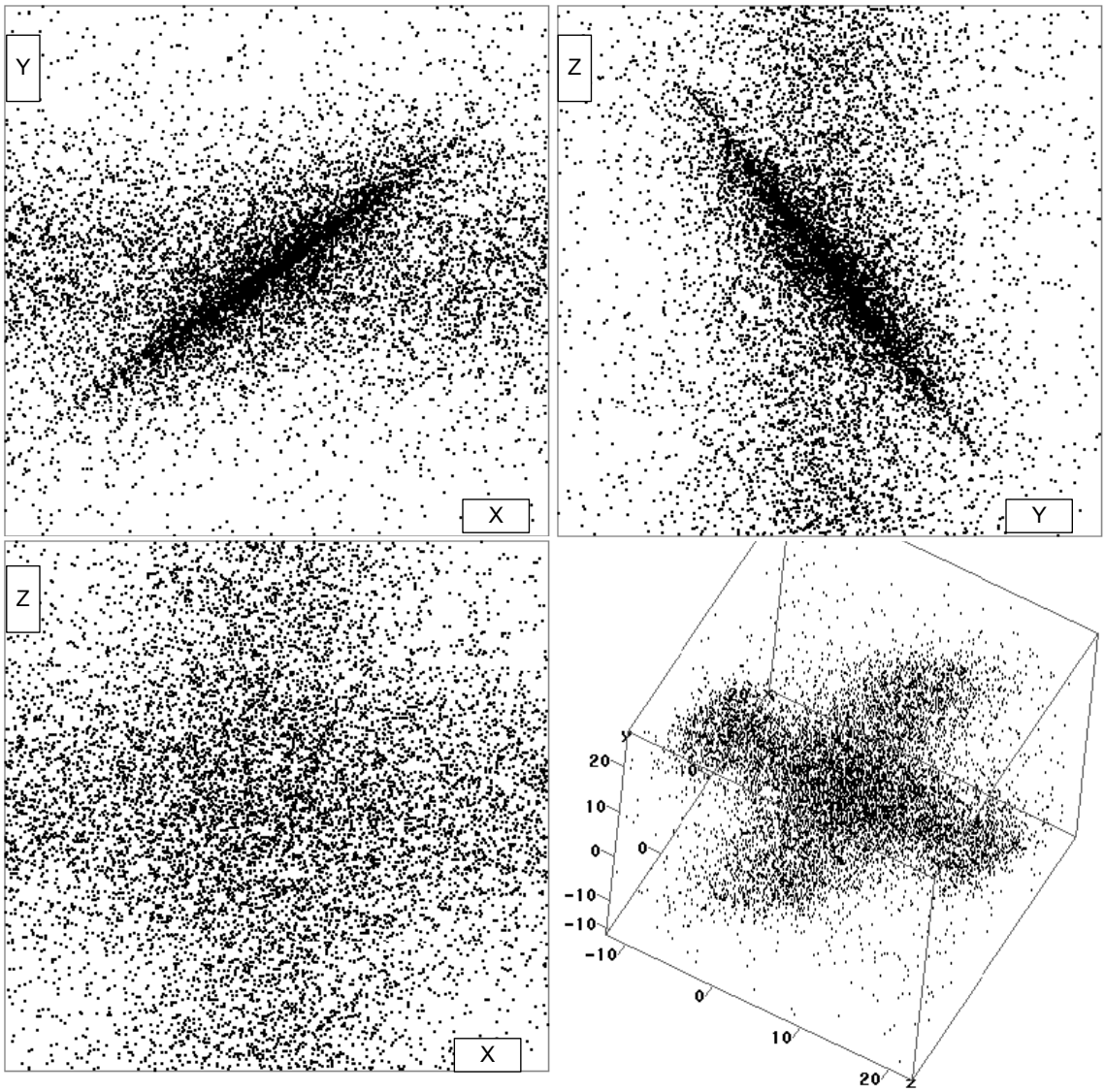


Figure 1: The problem of subspace patterns. Different patterns exist in different subspaces. Here X and Y have a strong positive relationship, Y and Z have a strong negative relationship, while X and Z have no significant relationship. In the 3-D view, those two relationships are blurred and difficult to identify. This problem gets much worse when many noisy attributes exist.

Nevertheless, choosing subsets of attributes (interesting subspaces) that contain good patterns remains a challenging research problem. First, each of existing subspace clustering methods relies on its own clustering algorithm to define and search clusters. They all try to find a cluster and the subspace contains it *simultaneously*. Due to the dramatically different definitions of clusters and searching strategies, each existing subspace clustering method is biased towards some specific types of clusters. Moreover, each of them needs several subjective input parameters, e.g., the number of clusters, the dimensionality of clusters, various thresholds, etc., which are hard to configure beforehand. The analyst needs to run the subspace clustering procedure many times with different input parameters to get a feel of which result might be more reasonable. Second, it is not easy to get an overall understanding

of various relationships among dimensions and their change from one significance level to another. It is also hard to see the impact of removing one or more dimensions on various relationships without running the procedure again.

A novel feature selection (or subspace selection) approach is developed, which is based on a new calculation of conditional entropy to measure the existence and the significance level of patterns in a 2-D space. In this paper, patterns specifically mean clusters, which are defined as contiguous, arbitrary-shaped, dense areas of data objects. A partition of the data space is not required, i.e., a data space may contain only one cluster. So, a linear relationship can be regarded as a special case of a cluster, which has an elongated shape.

For a high-dimensional data set (*S*) of *dimensionality d* and size *n*, a matrix of pair-wise conditional entropy values of all 2-D subspaces in *S* is derived. A visualization technique is also developed to visualize the entropy matrix and present a holistic picture of various relationships among dimensions. Interesting multidimensional subspaces can then be automatically extracted or interactively identified. Once a subspace is selected, any full-dimensional clustering method can be used to search for clusters in that subspace. Other than supporting effective multivariate clustering, this method can help in general explore large and high-dimensional data via: (**1**) informing various high-dimensional visualization techniques to focus on a subspace for better exploratory views, and (**2**) informing the user about the quality of each attribute and various relationships among attributes, which can help add, remove, and/or extract attributes to compile a better data set for analysis.

The rest of the paper is organized as follows. The next section will introduce related research directed to tackling the high dimensionality for clustering problems. Section 3 will present the calculation of a conditional entropy value for a 2-D data space. Then section 4 will introduce how to search interesting multidimensional subspaces based on the entropy matrix, automatic algorithms, and human interaction with the help of visualization. Section 5 includes experiments to demonstrate the ability of the subspace selection approach. The last section is conclusion and discussion.

2. Related Research

2.1. Dimension Reduction Techniques

One way to address the problem of high dimensionality is to apply a dimension reduction method to the data set (Duda, Hart et al. 2001). Methods such as principle component analysis (PCA) transform the original data space into a low dimensional space by forming dimensions that are linear combinations of given attributes. While these techniques may succeed in reducing dimensionality and are useful for information compression and classification problems, they have several shortcomings with respect to clustering. First, the new dimensions can be difficult to interpret, making it even harder to understand clusters formed by new dimensions. Secondly, these techniques are not effective in identifying clusters that reside in different subspaces of the original data set. In other words, such dimension reduction approaches can only generate *one* single optimal subspace to represent the original data space. Third, PCA methods can only work well for linear relationships. Fourth, PCA uses all of the original features in the projection to a lower dimensional space. In other words, although dimensionality is reduced, the impact of every original dimension is more or less still there.

2.2. *Feature Selection Methods*

Feature selection methods are traditionally used to select a subset of dimensions for supervised classification problems (Liu and Motoda 1998). Recently several new feature selection methods have been developed to select an "optimal" subset of features for unsupervised clustering (Dy and Brodley 2000b; Dy and Brodley 2000a), or produce a pool of "good" dimension subsets for searching clusters (Kim, Street et al. 2000). Each of these methods centers around a specific clustering method, e.g. the expectation maximization (Dy and Brodley 2000a) or the K-means (Kim, Street et al. 2000). However, it can be ineffective to rely on a specific clustering algorithm as a means to evaluate candidate subsets of dimensions. For example, K-means tends to discover equal-sized circular clusters and may fail to discover arbitrary-shaped patterns (e.g, in an extreme case, linear relationships), while an EM approach depends on a good initialization and also favors circular or ellipse clusters.

Scalability to high dimensionality (d) and very large data size (n) is another concern. Although efficient algorithms for K-means or EM-based clustering have been developed (Alsabti and Sanjay Ranka 1998; Bradley, Fayyad et al. 1998; Pelleg and Moore 1998), repeatedly using such clustering algorithms to evaluate a large number of candidates (i.e., subsets of dimensions) can still cause computational efficiency problems, especially when both d and n are large.

2.3. *Subspace Clustering Methods*

Several subspace clustering methods can detect clusters residing in different subspaces (i.e., subsets of the original dimensions). No new dimension is generated, which is important because original dimensions bear real meaning to the user. Each resultant cluster is associated with a specific subspace. CLIQUE (Agrawal, Gehrke et al. 1998), ORCLUS (Aggarwal and Yu 2000), and DOC (Procopiuc, Jones et al. 2002) are three representative methods (among many others (Cheng, Fu et al. 1999a; Goil, Nagesh et al. 1999; Aggarwal and Yu 2000; Nagesh, Goil et al. 2000)) for subspace clustering.

CLIQUE adopts a density-based approach to clustering: a cluster is a region that has higher density of points than its surrounding area. To approximate the density of data points, the data space is partitioned into a finite set of cells by partitioning each dimension into the same number of equal length intervals. A cluster is defined as a maximal set of connected dense cells in a subspace. CLIQUE uses a bottom-up searching and pruning strategy—if a k-dimensional subspace does not have any dense cell, all subspaces that contain this subspace are pruned. The result of CLIQUE critically relies on two parameters: the number of intervals (ξ) and the density threshold (τ), which are hard to configure. Equal interval discretization suffers from outliers and extreme values and cannot adapt well to various data distributions. *ORCLUS* introduces the notion of generalized projected clusters. Its performance degrades fast with increasing dimensionality. Other two drawbacks of ORCLUS are that: (1) the requirement parameters k and l, and 2) the dependency between parameters α and β. *DOC* is a Monte Carlo algorithm that computes, with high probability, a good approximation of a projective cluster. Its parameters w, α, and β, are all hard to configure but all have dramatic influences on the result. The user needs to run the algorithm many times with different settings of one or more parameters to gain an overall understanding of the data set and the overall relationships among dimensions. Moreover, the accuracy of DOC is poor for clusters of low dimensionality (e.g., <10) (Procopiuc, Jones et al. 2002).

2.4. Cluster Tendency Measures

Clustering tendency (Jain and Dubes 1988; Jain, Murty et al. 1999) refers to the problem of deciding whether the data set exhibits a tendency to cluster into natural groups without the actual identification of those clusters. In other words, clustering tendency is a quantitative measurement that indicates whether a data set has clusters and how strong those clusters are. An examination of the clustering tendency is an important part of for the overall clustering procedure because many clustering algorithms will create clusters whether or not the data are naturally clustered or just random.

In this paper, a cluster is defined as a contiguous, dense region (in multidimensional space) of arbitrary shape. A partition of the data space is not required, i.e., the data may contain only one cluster. Thus a linear relationship can be regarded as a special case of a cluster, which has an elongated shape. From this point of view, three types of cluster tendency measure exist: covariance or correlation, Quadrat analysis and chi-square test, and entropy-based measures.

The *covariance* or *correlation* measure in statistics has been long used for testing the existence of a *linear* relationship between two dimensions. However, for a data space, there may exist strong clusters while the correlation value is very low since correlation only captures linear relationships. Spatial autocorrelation can be regarded as a measure of clustering in a 2-D (usually geographic) space (Zhang and Murayama 2000).

Quadrat analysis partitions a 2-D data space into rectangles of equal size, called quadrants, and counts the number of points falling in each quadrant (Jain and Dubes 1988). If the data set contains no significant cluster, the set of counts will follow a Poisson distribution under randomness. The Chi-square test (Snedecor and Cochran 1989), which has been widely used for testing the statistical significance of bivariate tabular association, can then be used here to test a hypothesis of randomness for those quadrant counts.

Conditional entropy is another measurement for detecting the mutual interaction between two dimensions (attributes) (Pyle 1999). Cheng and others (1999) also proposed an *entropy-based approach* for evaluating and pruning subspaces (Cheng, Fu et al. 1999b).

3. Using Maximum Conditional Entropy to Measure Cluster Tendency

To efficiently tackle the problem of high dimensionality, only 2-D subspaces of a high-dimensional dataset are examined. As mentioned before, a cluster is defined as a contiguous, dense region of an arbitrary shape. To measure the cluster tendency or the "goodness of clustering" in a 2-D data space, there criteria need to be considered (Cheng, Fu et al. 1999a): high coverage (percentage of data points contained in clusters), high density (high coverage of a small region), and high dependence (the two dimensions not independent from each other). See Figure 2 for some illustrative data sets to understand these three criteria.

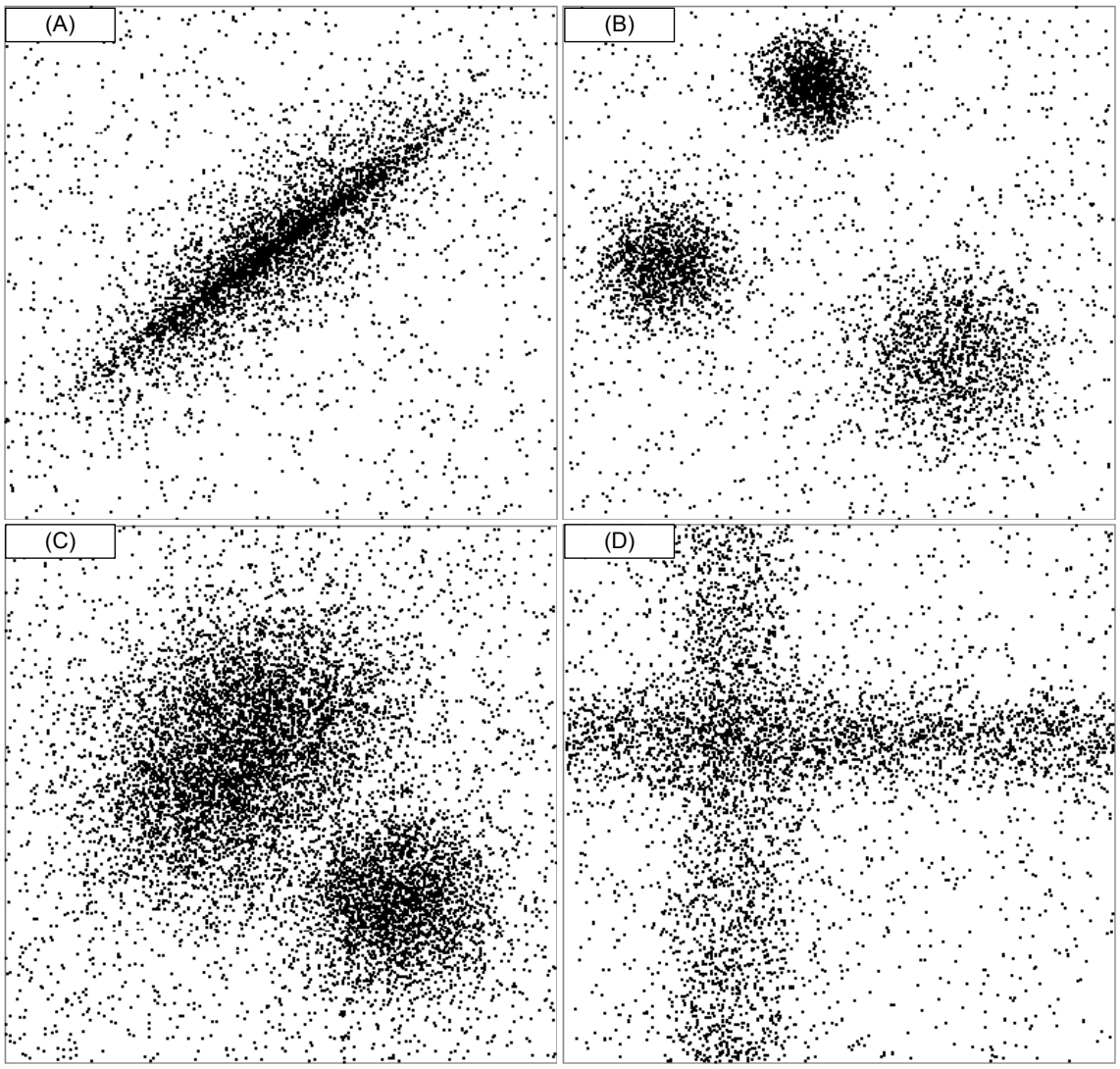


Figure 2: Various types of clusters and null cluster. Each of the four 2-D data sets has 5000 points. (A) has strong linear relationship of high coverage, very high density, and very high dependency. (B) has three circular clusters of high coverage, high densities, and high dependence. (C) has two clusters of high coverage, lower density, and lower dependence. (D) has no cluster because those two dimensions are independent from each other, i.e., for a majority of points the value of one dimension cannot be determined (to some degree) by the value of the other dimension. In all four data sets there is a certain amount of noise.

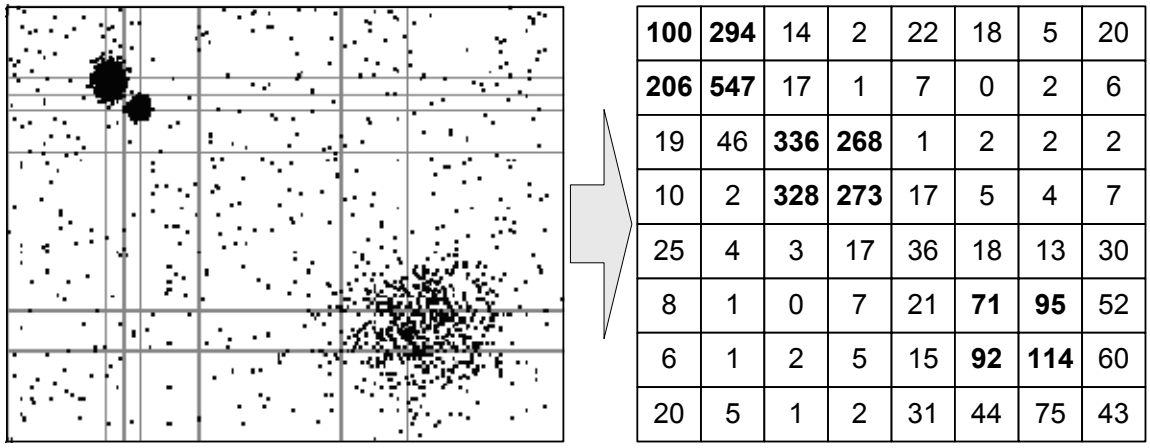
As shown in Figure 2, different subspaces can have different number of clusters of various shapes, sizes, and distributions. Noise and extreme values can also exist. A good cluster tendency measure should not be biased to any particular type of clusters and should be robust with extreme values and noise. To meet these requirements, the clustering tendency measure should be very generic.

A calculation of maximum conditional entropy for a 2-D data space is developed to measure the "goodness of clustering". To calculate a conditional entropy value, the 2-D data space is first discretized into a matrix of grid cells. Traditionally the data space is partitioned

into grids by cutting each dimension into several equal-sized intervals. However, such equal-interval discretization cannot adapt to various data distributions and is sensitive to extreme values. Here another discretization method is adopted, which is introduced below.

3.1. Dimension Discretization

There are many existing discretization (classification) methods for single-dimensional data (Slocum 1999). CLIQUE, ENCLUS and Quadrat analysis all adopt the Equal-Interval (EI) method. However, EI cannot adapt well to various data distributions and is sensitive to extreme values.



100	**294**	14	2	22	18	5	20
206	**547**	17	1	7	0	2	6
19	46	**336**	**268**	1	2	2	2
10	2	**328**	**273**	17	5	4	7
25	4	3	17	36	18	13	30
8	1	0	7	21	**71**	**95**	52
6	1	2	5	15	**92**	**114**	60
20	5	1	2	31	44	75	43

Figure 3: NM (Nested-Means) discretization. The numbers in the matrix shows the number of points that fall in each grid cell.

The Nested-Mean (NM) discretization method (Guo, Peuquet et al. 2002) is adopted here. The NM approach first calculates the mean value of a dimension and then divides the data into two halves with that mean value. Recursively, each half is divided again into halves with its own mean value (see Figure 3). The recursion stops when the required number of intervals is obtained. The NM approach is robust to outliers, extreme values, and noise.

The number of intervals (r) needed for each dimension depends on the data set size (n). A general rule adopted here is that on average each cell should contain about 35 points according to Cheng et al. (1999). Experiments showed that following this rule can reliably achieve good results. Another rule is that, for the nested-means discretization, r should equal 2^k (k is a positive integer). Since all data spaces considered here are 2-D, the $n/r^2 \approx 35$ and $r = 2^k$. For example, if $n = 10000$, then $r = 16$, because 16*16 = 256 and 256 *35 = 8960 (close to 10000). To scale well with extremely large data sets, the threshold 35 can increase by a factor of $\log_k n$, where k is a large integer (e.g., 1000). Thus the time complexity for discretizing all 2-D subspaces is $O(d^2 n \log n)$.

3.2. Calculation of Maximum Conditional Entropy

The calculation of a conditional entropy given a matrix of values can be found in (Pyle 1999). Let S be a 2-D subspace comprising of dimensions Ai and Aj. To calculate the conditional entropy of S, both Ai and Aj need to be discretized into ξ intervals. As introduce above, the NM method is used to discretize S into r intervals. Thus S is partitioned

into a matrix of grid cells. Let χ be the set of grid cells (including empty ones) for a column C in the matrix, and $d(x)$ be the density of a cell $x \in \chi$, i.e., the number of points in x divided by the total number of points in the column. Then the entropy of this column is calculated using the following equation:

$$H(C) = -\sum_{x \in \chi} [d(x) \log d(x)] / \log|\chi|.$$

Conditional entropy (Y|X) is a weighted sum of the entropy values of all columns (Figure 4). Following are three steps to calculate conditional entropy (Y|X). Similarly, conditional entropy (X|Y) can be calculated using rows instead of columns (Figure 4).

1. Calculate the *column sum* and *weight* = (*column sum*)/*n*;
2. Calculate the entropy for each column.
   $H(x2) = -[(1/36)*\log(1/36)+(9/36)*\log(9/36)+\ldots+(2/36)*\log(2/36)]/\log(6)$, where, 36 is the total number of points in *x2* and 6 is the total number of cells in *x2*.
3. Conditional entropy (Y|X) = weighted sum of all column entropy values.

	x1	x2	x3	x4	x5	x6		Sum	Wt.	CE
y1	0	1	3	0	0	0		4	.03	.314
y2	1	9	1	0	1	2		14	.09	.629
y3	7	14	3	7	6	0		37	.25	.835
y4	7	6	13	19	12	5		62	.41	.939
y5	0	4	14	5	1	1		25	.17	.668
y6	1	2	3	2	0	0		8	.05	.737
									CE(X|Y) .812	
Sum	16	36	37	33	20	8				
Wt.	.11	.24	.25	.22	.13	.05	CE(Y|X) .700		**MCE 0.812**	
CE	.597	.847	.806	.615	.540	.502				

Figure 4: The calculation of conditional entropy (Y|X) and conditional entropy (X|Y). The larger one of the two conditional entropy values is then taken as the final entropy value for the subspace.

As shown in Figure 4, a 2-D subspace has two conditional entropy values (one for columns, *CE*(Y|X), and one for rows, *CE*(X|Y)). If both values are small, that means that two dimensions are highly correlated (in terms of high dependency) and the subspace has a good clusters. If one value is small, while the other is large, that means the data is only clustered well on one dimension and thus the two dimensions is not dependent on each

other. Therefore, maximum conditional entropy, MCE = max ($CE(Y|X)$, $CE(X|Y)$) , is taken as the final cluster tendency measure of the subspace. The measure value will be only used for comparison between subspaces. The absolute value of the measure is not important although it does convey certain information. The measure is robust with noise because the noise exists in all subspaces.

4. Using Conditional Entropy Matrix to Select Subspaces

Let $A = \{A_1, A_2, \ldots, A_d\}$ be a set of dimensions and $S = A_1 \times A_2 \times \ldots \times A_d$ be a d-dimensional data space. Let $S_2 = \{A_i \times A_j \mid i = 1..d, j = 1..d, i < j\}$ be the set of all possible 2-D subspaces from S. The MCE values of these 2-D subspaces can form a matrix. This matrix can also be viewed as a complete graph with each dimension as a vertex. Each MCE value can be viewed as the distance (or dissimilarity) between two dimensions. Thus it can be imagined that there is an "edge" between any two dimensions. This "graph" will be used below for both visualizing the matrix and searching interesting multidimensional subspaces.

The rationale for selecting multidimensional subspaces based on this matrix (graph) is*: if an L-dimensional (2 < L < d) subspace S_L has good clusters, all possible 2-D subspaces of S_L should have low MCE values*. This rationale is similar to the monotonicity lemma used by CLIQUE: *if a collection of points P is a cluster in a k-dimensional space, then P is also part of a cluster in any (k-1)-dimensional projections of this space* (Agrawal, Gehrke et al. 1998). Is it possible that a subspace S_L has good clusters but some (or even all) of its projected 2-D subspaces do not have good clusters (and hence high MCE values)? It is possible, but of very low possibility. Since the method does not actually identify clusters, overlap of clusters in a projected 2-D subspace is not a problem. Overlap of clusters can actually make the conditional entropy value of the subspace even smaller (better). The only problematic case is that, all clusters are of the same density and are regularly distributed (no overlap and side by side) in the projected 2-D subspace. In such a case, the entropy value would be very high. However, the probability of such a case is very small.

4.1. Ordering Dimensions for Better Visualization

To render a better display of the entropy matrix, an optimal ordering of all dimensions is derived such that correlated dimensions (in terms of low conditional entropy) are placed as close to each other as possible in the ordering. The more correlated two dimensions are, the closer they should be in the ordering. A minimum spanning tree (MST) is constructed from the complete graph of all dimensions depicted above. From the MST, an optimal ordering of all dimensions can be derived to fully preserve the hierarchical cluster structure of all dimensions and additional proximity information between dimensions (Guo, Peuquet et al. 2002).

Figure 5 shows the matrix with ordered dimensions. Correlation values of paired variables are displayed above the diagonal and MCE values of paired variables are displayed below (in both cases, the red cells represent *good* values: low values of MCE and high values of correlation). Please note: correlation values shown here are only for reference and comparison with MCE values. The ordering of dimensions is derived using MCE values (as described above). The diagonal provides access to each variable; the user can select, add to, or subtract from a subset by simply clicking on the variable's diagonal cell. A selected subset can be broadcast to other components (sorted on the conditional entropy values for the

subspace selected) for further analysis (see section 6: application with cancer data). This visualization of matrix can accommodate a large number of attributes (e.g., >100). In such a case, each cell is a very small square with a color, but the MCE value will not be shown. With the mouse over a cell, the MCE value of that cell will pop out (see Figure 5).

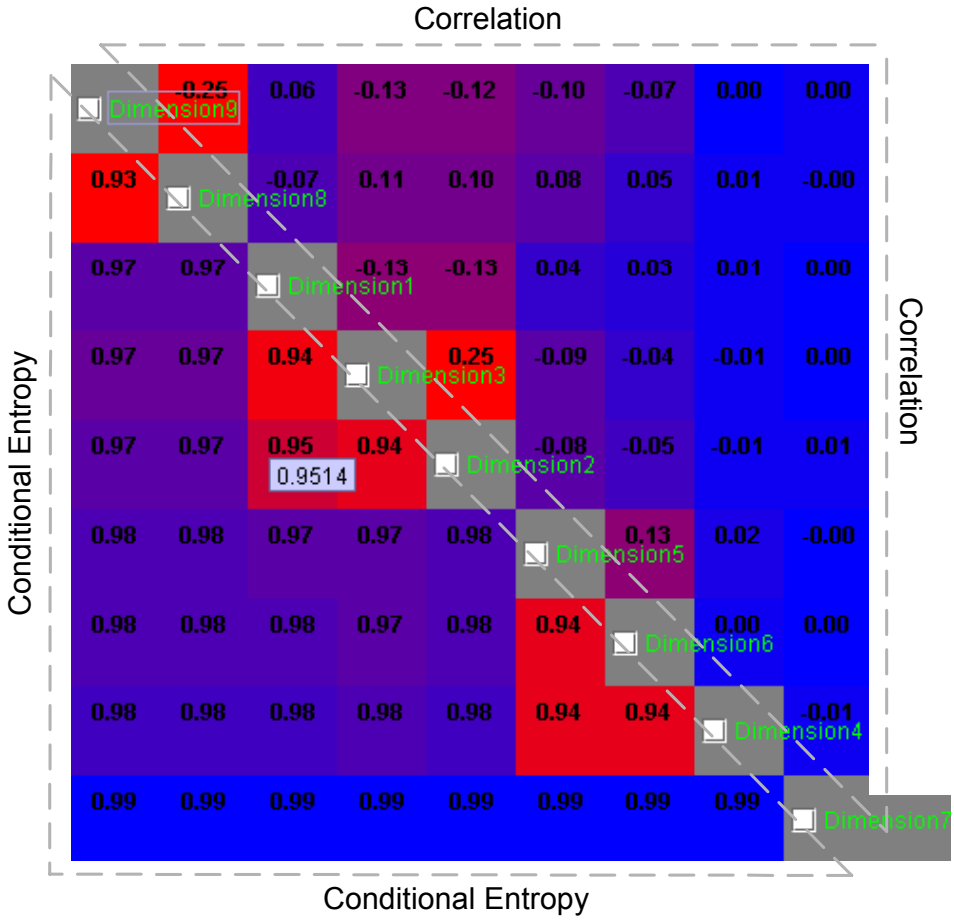


Figure 5: Visualization of the conditional entropy matrix. Dimensions are ordered according to a hierarchical clustering of those dimensions. Both entropy values (bottom-left half) and correlation values (top-right half) are shown for comparison.

From the comparison (in Figures 5, 6, 7, 9) it can observed that if a 2-D subspace has a good correlation value, it always have a good conditional entropy value as well, while those subspaces that have good MCE values do not always have good correlation values. This confirms that the MCE measure treats a linear relationship as a special case of a cluster. The time complexity for constructing the matrix is $O(d^2 n \log n)$, where d is the dimensionality and n is the size of the data set. The construction procedure of an MCE matrix is decomposable. It can adopt a distributed computing strategy to be time-efficient, or sequentially process data column by column (or row by row) to be memory-efficient. Another advantage of this approach is that, once the matrix is constructed, the user can examine various relationships among dimensions without running the procedure repeatedly.

4.2. Interactive Exploration and Interpretation of Subspaces

With the above matrix, the user can get a holistic understanding of the relationships among dimensions. The user can easily identify interesting subspaces that have good clusters based on the visual display. In Figure 5 there are three interesting subspaces: {d1, d2, d3}, {d4, d5, d6}, and {d8, d9}, which are very easy to visually identify. The data set used here only has 10 dimensions, which is designed for illustration only. The method and the matrix can effectively and efficiently handle more than 100 dimensions and clusters of various dimensionalities.

The user can interactively form a subspace according to his/her understanding, expertise, and interest. For a large number of attributes in real data set, the relationships among attributes can be very complex. Some attribute may be duplicate and identical, which definitely will show strong patterns but are not so interesting to the user. Some patterns may not be as strong as the above ones but can be of great interest to the user. With the user's interactive exploration and interpretation with the matrix, meaningful and interesting subspaces can be quickly identified (see section 6: application with cancer data).

5. Experiment with Synthetic Data

The approach is evaluated against a most recent subspace clustering method DOC (Procopiuc, Jones et al. 2002). The data generator described in DOC is used to generate high-dimensional data sets that contain U-clusters, MG*q*-clusters, or N-clusters. The developed method works well with all these types of cluster, although it works better with U-clusters and MG*q*-clusters than with N-clusters. Therefore, the results with N-clusters are presented. Points in an N-cluster conform to a normal distribution.

The dimensionality of data sets is 50. The data set size is 50,000. The *cluster dimensionality* (i.e., the dimensionality of the subspace associated with the cluster) ranges between 3 and 7 (from a Poisson distribution with a mean value of 5. The dimensions for each subspace are randomly chosen. There are always 5 clusters, which is the same as in DOC. The coverage of clusters ranges from 10% to 20%. The noise level is around 20%. All generated points have values in the range [0, 100]. The standard deviation of the clustered points for each dimension is randomly chosen from range [5, 10]. The mean value of the clustered points is also randomly chosen. Clustered points are normally distributed (with above parameters) in the associated subspace of bounded dimensions. Experiments are also conducted with cases in which clusters share a fair portion of dimensions.

Table 1: Clusters in the synthetic data.

Cluster ID	*Number of points*	*Associated subspace*
1	7443	{0, 3, 10, 12, 14, 16, 19}
2	8432	{9, 13, 19}
3	6939	{2, 3, 9, 19}
4	8403	{4, 5, 7, 13}
5	8781	{2, 4, 7, 9, 11, 14}
noise	10002	All attributes

Due to space limit, only one experiment and its synthetic data set is presented below. Table 1 shows the 5 clusters in the data. Some of those 5 subspaces heavily overlap with each other, i.e., sharing several attributes. Figure 6 shows the MCE matrix. From the matrix, one can clearly see those subspaces that contain clusters. Interestingly, one can see that those correlation values are of little use for identifying interesting subspace.

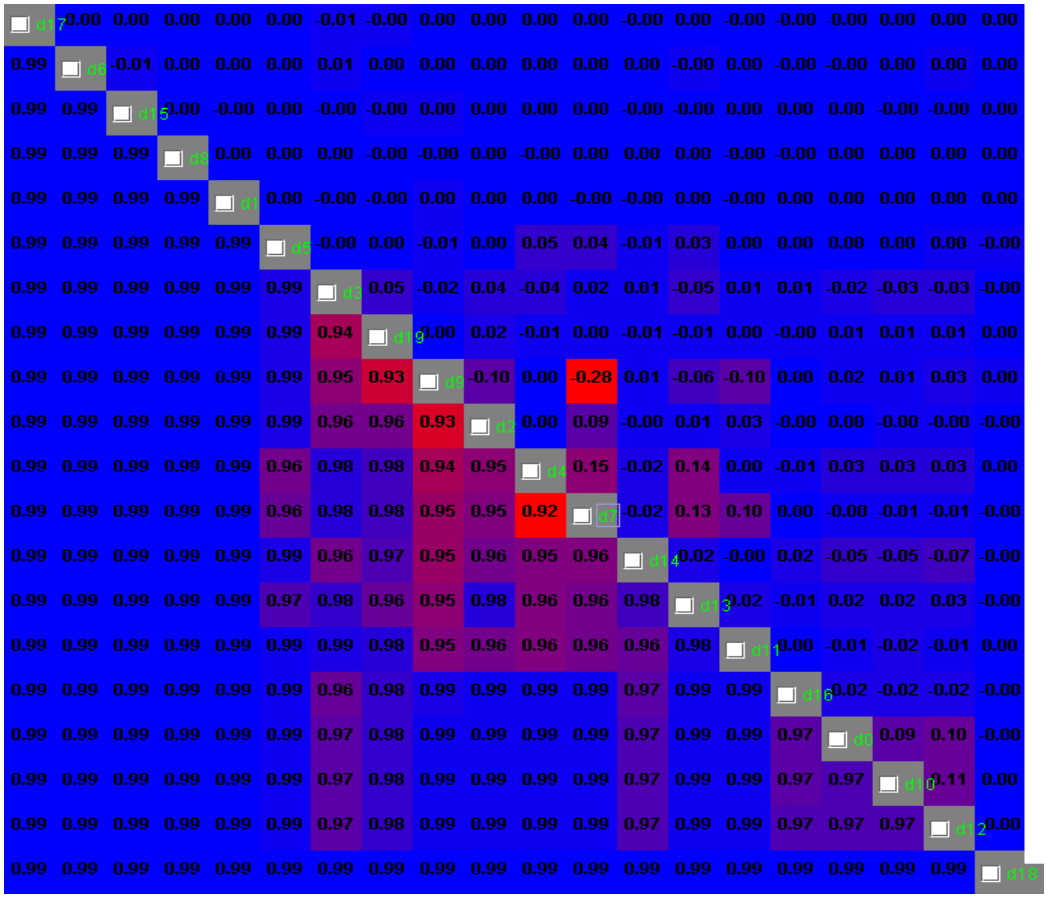


Figure 6: An experiment with synthetic data. For presentation clarity, only part of the matrix (20 attributes) is shown here. All dimensions that are involved with any cluster are covered by these 20 attributes.

6. Application with Cancer Data

This is a county-level epistemology data set. Attributes included in the data are listed and explained in Table 2. The data set contains 1156 counties in US with BRRALLZ > 45. This data set is a subset of a much larger data set of more than 100 attributes. Again, for the clarity of demonstration, only a small set of attributes included here. For this application the feature selection methods is coupled with an interactive, hierarchical, multivariate clustering method (Guo, Peuquet et al. 2002; Guo, Peuquet et al. 2003) to demonstrate how the developed method can be integrated within a comprehensive knowledge discovery environment to efficiently and effectively explore high-dimensional data and search complex patterns.

Table 2: Attributes in the breast cancer data set.

Attribute	*Explanation*
BRRALLZ	Breast cancer mortality rate per 100,000 person-years, all races, all genders, all ages, for the time period 1970 – 1994
MDRATIO	# physicians per 100,000 population
HOSP	# hospitals per 100,000 population
URBRURAL	USDA urban/rural code (0=most urban, 9=most rural)
PCINCOME	per capita income
PCTPOOR	% living below federal poverty line
PCTCOLED	% adults over 25 with 4+ years of college education
UNEMPLOY	% unemployed
POP95	1995 popopulation
MAMMOG2YSM	% women ages 50-64 who had a mammogram in past 2 years
OBESE	% of persons ages 18+ who are >120% of the median body mass index
NOINS	% of persons ages 18+ who do not have a health plan or health

The conditional entropy matrix of this breast cancer data set is shown in Figure 7. From the matrix one can see a striking red area that involves 5 attributes {PCINCOME, PCTCOLED, MDRATIO, URBRURAL, POP95}. This subspace is selected in the matrix and then passed to the clustering component to search actual clusters in it. After applying the interactive clustering method, three major clusters emerged, which are visualized and shown in Figure 8. The green cluster represents those very urban counties with high POP95, high MDRATIO, high PCTCOLED, and high PCINCOME. The red cluster represents those counties that are average URBRURAL and of high average POP95, average MDRATIO, average PCTCOLED, and average PCINCOME. The blue cluster represents those very rural counties that are of low POP95, low MDRATIO, below-average PCTCOLED, and below-average PCINCOME.

Although above clusters are valid, significant, and interesting, the analyst may have more interest in finding relationships that involve the breast cancer mortality ratio (BRRALLZ). From the matrix, one can see that BRRALLZ has a good relationship with POP95 and moderate relationships with URBRURAL, MDRATIO, PCINCOME, HOSP, and MAMMOG2YSM. Since by now the user already knows (from both the matrix and above clustering result) that URBRURAL, MDRATIO, PCINCOME, and POP95 have strong linear relationships, it is better not include them all with other attributes to avoid duplicate information. So the user forms a subspace {MDRATIO, POP95, BRRALLZ} (Figure 9). The clustering result is shown in Figure 10. The multivariate clusters also visualized on a map (Figure 10). Among those clusters found, the red cluster is particular interesting, which indicates that high breast cancer mortality is associated with low POP95 (hence rural counties) and low MDRATIO. From the map, one can also see the geographic distribution of those counties with a high breast cancer mortality ratio is also interesting—they form a spatial cluster across the north and midwest region (Figure 10).

The user can interactively and iteratively explore various patterns with this integrated discovery environment. The feature selection method makes the process more efficient, effective.

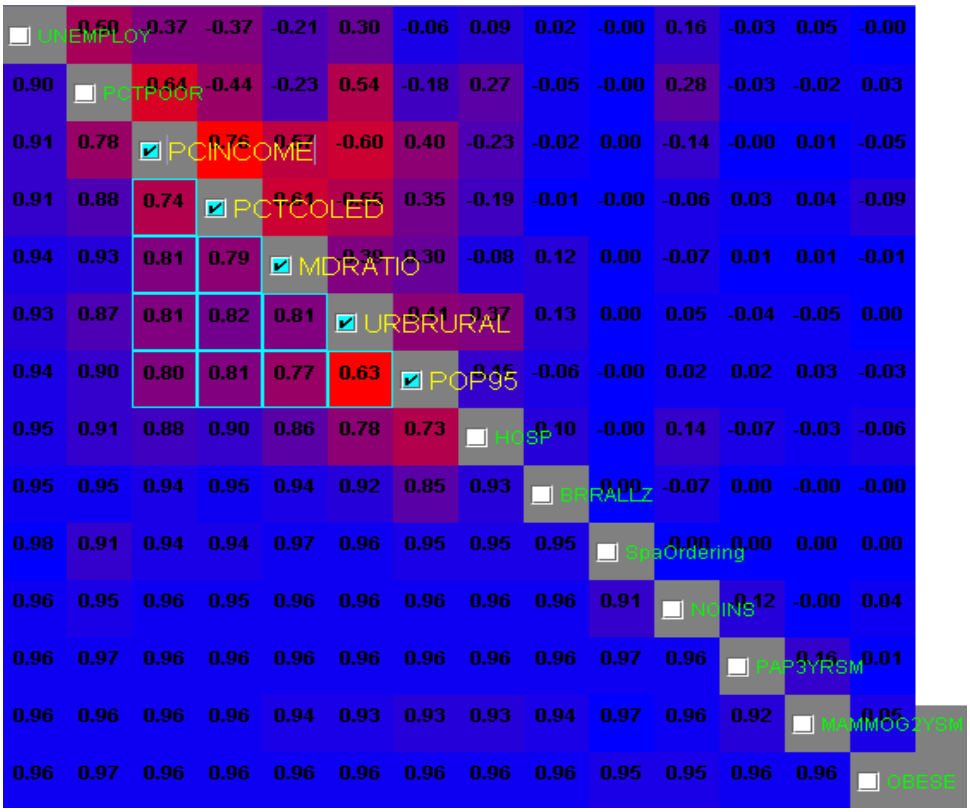


Figure 7: Conditional entropy matrix of the breast cancer data set. A subspace {PCINCOME, PCTCOLED, MDRATIO, URBRURAL, POP95} is selected.

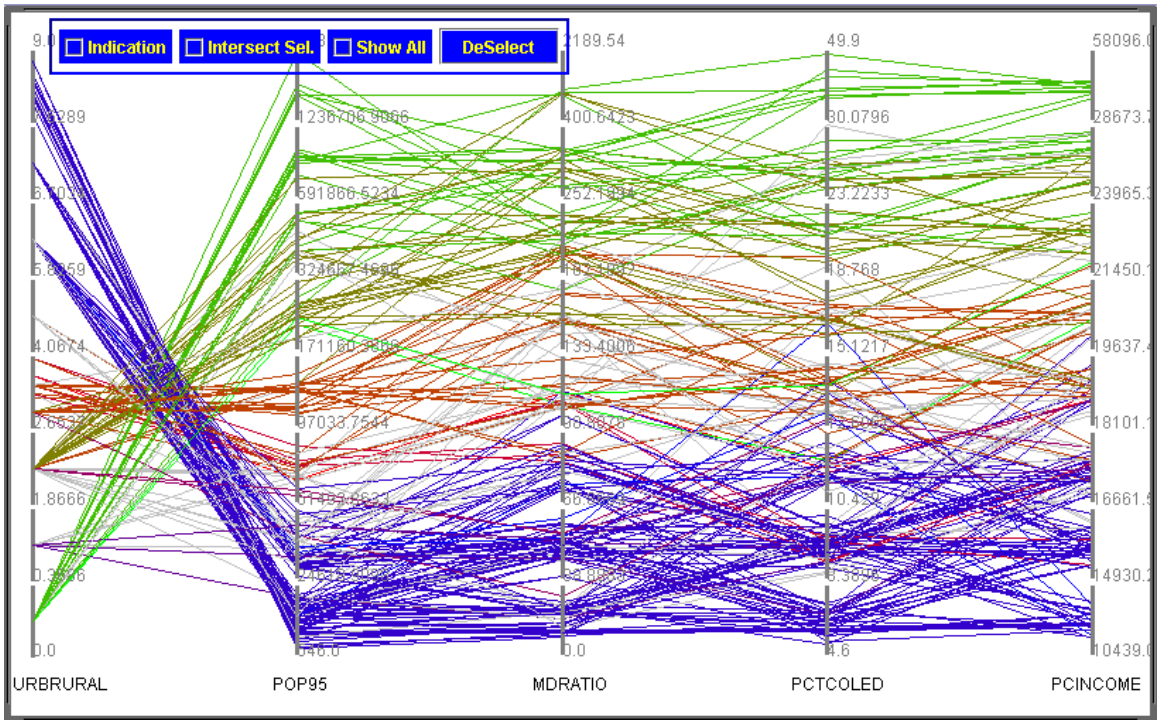


Figure 8: Visualization and clustering of the selected subspace {PCINCOME, PCTCOLED, MDRATIO, URBRURAL, POP95}.

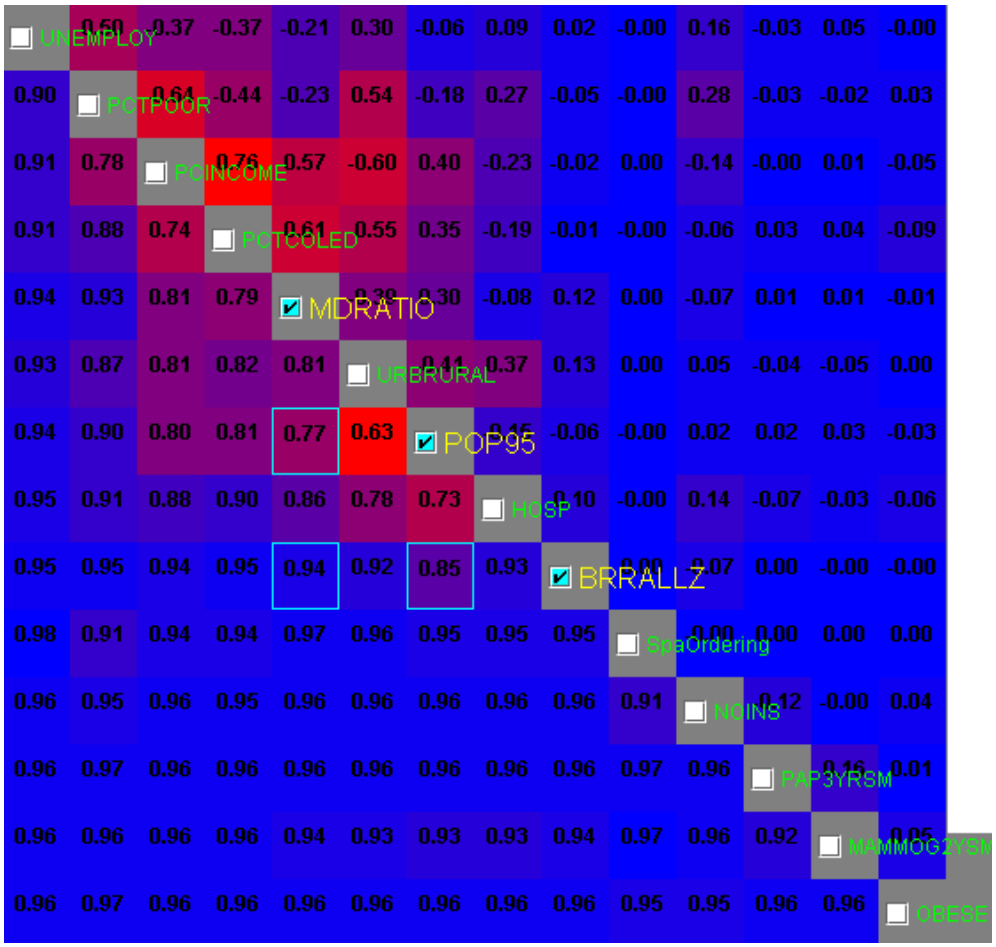


Figure 9: Another subspace. After the examination of the most significant relationship, the user then focuses on less significant but more interesting subspaces, e.g. a subspace involves the breast cancer mortality ratio. Since URBRURAL, MDRATIO, PCINCOME, and POP95 have strong linear relationships, it is better not include them all with BRRALLZ to avoid duplicate information. Here the subspace {MDRATIO, POP95, BRRALLZ} is selected for further clustering.

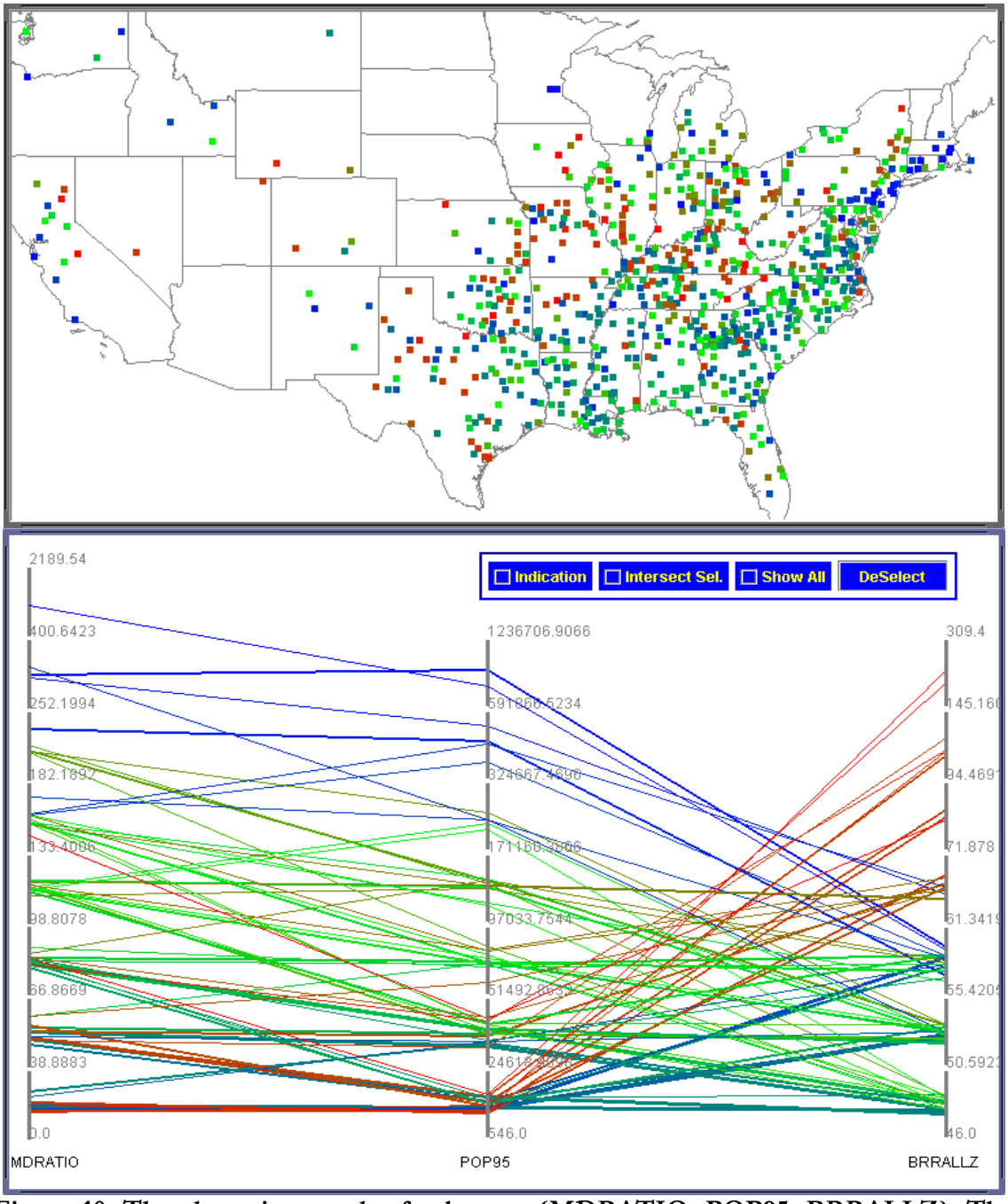


Figure 10: The clustering result of subspace {MDRATIO, POP95, BRRALLZ}. The multivariate clusters are then mapped to the geographic space. The red cluster, which is also a cluster in the geographic space, contains counties of high breast cancer morality ratio, low population, and low MDRATIO.

7. Conclusion

A feature selection method is developed to select interesting subspace for further analysis. The method is based on a matrix of pair-wise maximum conditional entropy values of 2-D

data spaces. A new calculation of the conditional entropy is proposed to reliably measure the "goodness of clustering" in a 2-D space. From the entropy matrix, the user can easily get an understanding of the overall picture of various relationships among dimensions. Multidimensional subspaces of more than two dimensions can then be interactively identified from the matrix. Other than choosing subspaces for clustering, this method can also be used to (1) inform various high-dimensional visualization techniques to focus on a subspace for better views; and (2) add, remove, and/or extract attributes to prepare a better data set for exploratory analysis (not limited to clustering).

Acknowledgement:

This paper is partly based upon work funded by NSF Digital Government grant (No. 9983445) and grant CA95949 from the National Cancer Institute.

Reference:

Abbott, D. W., I. P. Matkovsky and J. F. I. Elder (1998). An evaluation of high-end data mining tools for fraud detection. Systems, Man, and Cybernetics, IEEE International Conference on, San Diego, CA, USA**:** 2836 - 2841.

Agarwal, C., C. Procopiuc, J. Wolf, P. Yu and J. Park (1999). A Framework for Finding Projected Clusters in High Dimensional spaces. ACM SIGMOD International Conference on Management of Data.

Aggarwal, C. and P. Yu (2000). Finding generalized projected clusters in high dimensional spaces. ACM SIGMOD International Conference on Management of Data.

Agrawal, R., J. Gehrke, D. Gunopulos and P. Raghavan (1998). Automatic subspace clustering of high dimensional data for data mining applications. ACM SIGMOD International Conference on Management of Data, Seattle, WA USA**:** 94-105.

Alsabti, K. and V. S. Sanjay Ranka (1998). An Efficient K-Means Clustering Algorithm. IPPS: 11th International Parallel Processing Symposium.

Bradley, P., U. Fayyad and C. Reina (1998). Scaling clustering algorithms to large databases. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York City.

Cheng, C., A. Fu and Y. Zhang (1999a). Entropy-based subspace clustering for mining numerical data. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA.

Cheng, C., A. Fu and Y. Zhang (1999b). Entropy-based subspace clustering for mining numerical data. Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Duda, R. O., P. E. Hart and D. G. Stork. (2001). Pattern classification. New York, John Wiley & Sons.

Dy, J. G. and C. E. Brodley (2000a). Feature subset selection and order identification for unsuprervised learning. the Seventeenth International Conference on Machine Learning, Stanford University**:** 360 - 364.

Dy, J. G. and C. E. Brodley (2000b). Visualization and interactive feature selection for unsupervised data. the sixth ACM SIGKDD international conference on Knowledge

discovery and data mining, Boston, Massachusetts, United States, ACM Press New York, NY, USA**:** 360 - 364.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth (1996). From data mining to knowledge discovery-An review. Advances in knowledge discovery. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusay. Cambridge, MA, AAAI Press/The MIT Press**:** 1-33.

Goil, S., H. Nagesh and A. Choudhary (1999). "MAFIA: Efficient and Scalable Subspace Clustering for Very Large Data Sets." Technical Report CPDC-TR-9906-010 , Center for Parallel and Distributed Computing, Northwestern University, June 1999.

Guo, D., D. Peuquet and M. Gahegan (2002). Opening the Black Box: Interactive Hierarchical Clustering for Multivariate Spatial Patterns. The 10th ACM International Symposium on Advances in Geographic Information Systems, McLean, VA, USA**:** 131-136.

Guo, D., D. Peuquet and M. Gahegan (2003). "ICEAGE: Interactive Clustering and Exploration of Large and High-dimensional Geodata." GeoInformatica **7**(3): In press.

Hinneburg, A. and D. A. Keim (1999). Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering. Proceedings of the 25th VLDB Conference, Edingburgh, Scotland.

Jain, A. K. and R. C. Dubes (1988). Algorithms for clustering data. Englewood Cliffs, NJ, Prentice Hall.

Jain, A. K., M. N. Murty and P. J. Flynn (1999). "Data clustering: a review." ACM Computing Surveys (CSUR) **31**(3): 264 - 323.

Kim, Y., W. N. Street and F. Menczer (2000). Feature selection in unsupervised learning via evolutionary search. Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, Massachusetts, United States, ACM Press New York, NY, USA**:** 365-369.

Liu, H. and H. Motoda (1998). Feature selection for knowledge discovery and data mining. Boston, Kluwer Academic Publishers.

Nagesh, H. S., S. Goil and A. Choudhary (2000). A scalable parallel subspace clustering algorithm for massive data sets. Proceedings of the International Conference on Parallel Processing**:** 477 -484.

Pelleg, D. and A. Moore (1998). "Accelerating Exact k-means Algorithms with Geometric Reasoning."

Procopiuc, C. M., M. Jones, P. K. Agarwal and T. M. Murali (2002). A Monte Carlo Algorithm for Fast Projective Clustering. ACM SIGMOD, Madison, Wisconsin, USA, ACM**:** 418-427.

Pyle, D. (1999). Data preparation for data mining. San Francisco, Calif., Morgan Kaufmann Publishers.

Slocum, T. A. (1999). Thematic cartography and visualization, Upper Saddle River, N.J. : Prentice Hall.

Snedecor, G. W. and W. G. Cochran (1989). Statistical methods, Iowa State University Press.

Zhang, C. and Y. Murayama (2000). "Testing local spatial autocorrelation using k-order neighbors." International Journal of Geographical Information Science **14**(7): 681-692.