Data Mining in the Presence of Quantitatively and Qualitatively Diverse Information

NSF IDM-0415190
Anne M. Denton ¡anne.denton@ndsu.edu¿
North Dakota State University
Grant Homepage http://www.cs.ndsu.nodak.edu/∼adenton/IDM/
PI Homepage: http://www.cs.ndsu.nodak.edu/∼adenton/

The work under this grant has established several concepts for working with diverse data. As a first step, abstractions have been developed that appropriately represent the richness of data types such as sequence and graph data, and their combination with conventional data types such as Boolean (or item) data. As a second step towards integrating diverse data, techniques have been developed for establishing significance of attribute groups that correspond to a particular data source, and for finding significant relationships between attribute groups and Boolean attributes.

Extraction of patterns from graph-relational data

We have addressed the need for abstractions of graph data that go beyond characterizing the structure of the graph and rather also integrate information that is associated with graph nodes [2, 3, 1, 15]. We have developed techniques for determining which of those combined attributes are unexpected with respect to basic properties that are known from node data alone or based on simple graph correlations. Fig. 1 illustrates the concept of expected and unexpected patterns. Simple patterns are either made up of only one node, or consider only one unique set of attributes over all nodes. Complex patterns involve multiple nodes and attributes and thereby require techniques that go beyond standard single-table data mining or correlation analysis. Not all complex attributes are, however, independent of simple patterns, as can be seen in the bottom right part of the figure.

We have been able to provide a benchmark model for graph-relational data as well as developing a high-performing data mining algorithm that gives approximate answers. The benchmark model was built on the statistical theory of log-linear models [3]. We have, furthermore, adapted the concept of edge disjoint data mining techniques to graph-relational data. For

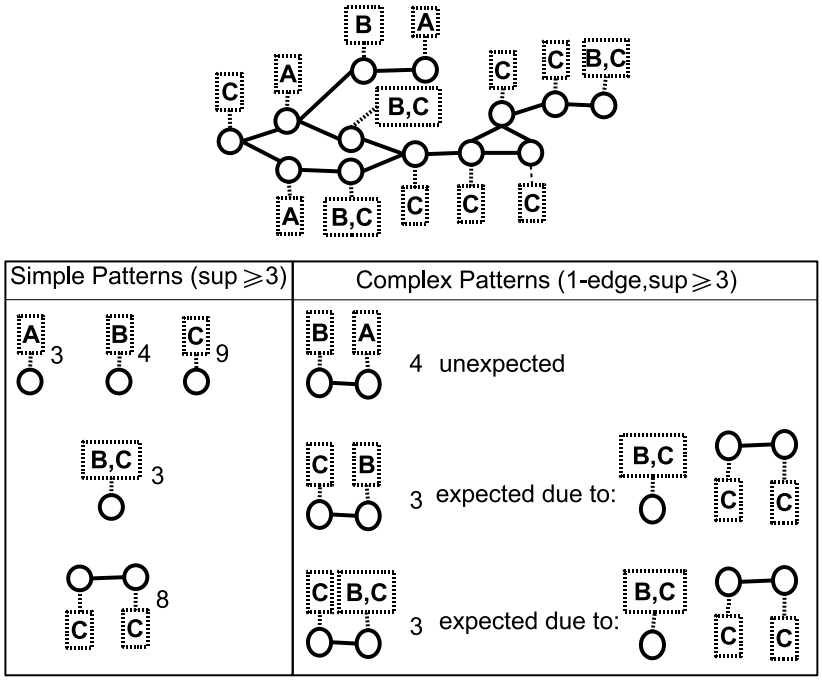

Figure 1: (Top) Example of a set-labeled graph. (Bottom-Left) Simple intra-node and correlation patterns. (Bottom-Right) Complex patterns expected or unexpected based on simple patterns.

both algorithms we have demonstrated effectiveness and efficiency on real-world applications, including protein networks with annotations and movie data. For the edge disjoint mining algorithm we gained patterns in protein regulation networks, where conventional techniques produce few or none [1].

Clustering of sequence data

We have have developed an algorithm that identifies clusters of similar sequence sections, combining benefits of two distinct ways of approaching sequence data mining. In most sequence clustering algorithms, overall pairwise similarities determine the grouping, while frequent subsequence mining and domain finding algorithms are based on identification of relevant sequence sections. In most sequence clustering algorithms, three mutually similar sequences may share no sequence section that is common to all three. Motif finding algorithms, on the other hand, depend on frequent occurrence of se-

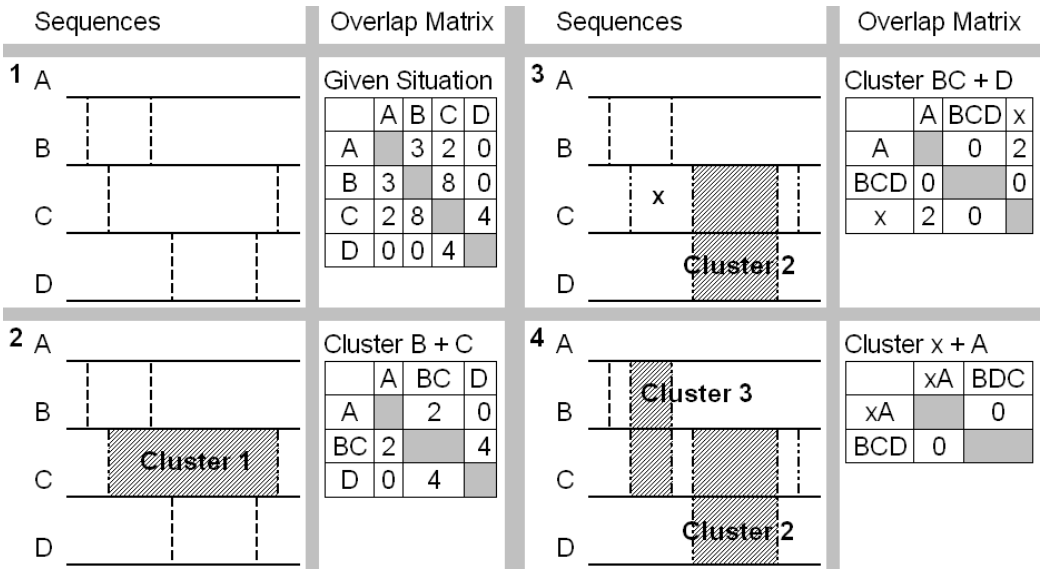

Figure 2: Our algorithm applied to multi-domain proteins. The figure shows that the algorithm correctly identifies local homology.

quence patterns or on prior knowledge. The algorithm developed under this grant uses a sliding-window-based similarity measure that clearly ties clusters to sequence sections while nevertheless capturing the alignment structure of the sequences that were clustered [12]. The usefulness of the resulting patterns is evaluated in the context of independent data.

Figure 2 shows schematically how sequences are clustered when some similarity regions are specific to only some of the sequences. For gene sequences we were able to show that our algorithm is capable of reproducing many domains already listed in the Interpro database as well as finding some new ones. As a second step we relaxed the cluster definition further and allowed clusters to be defined on the basis of any combination of windows [10, 14, 11]. Such generalized sequence signatures can be efficiently mined using clustering and pattern mining techniques. We have been able to show that annotations derived on the basis of our clusters satisfy self consistency considerations better than those based on Interpro domains.

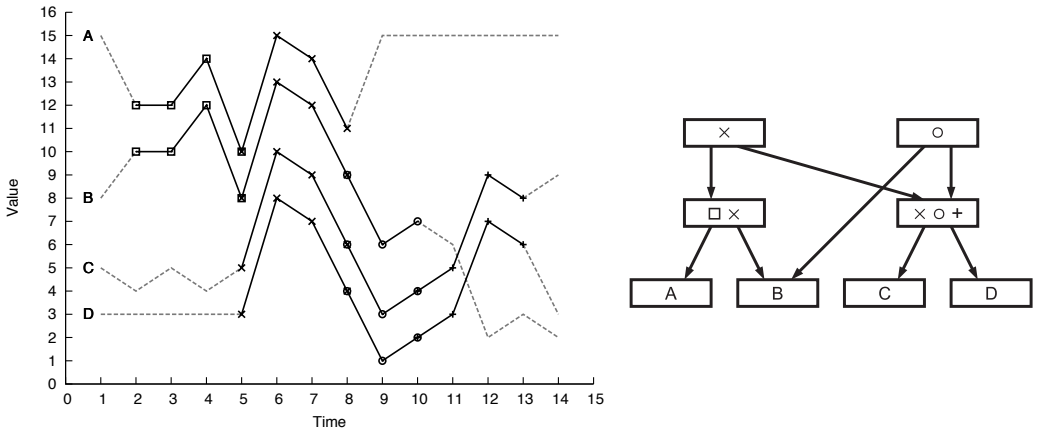

Figure 3: Example time series (a) together with the corresponding DAG (b).

Establishing relationships among patterns in stock market data

We have adapted the sequence-based clustering techniques of the previous section to stock market time series of the S&P 500 index as well as to four additional time series data sets. Sequence – subsequence relationships among patterns based on subsequence similarities are represented through a directed acyclic graph, DAG. Figure 3 shows an example of four time series that show a total of three characteristic shapes. The relationships between their regions of similarity is shown in the DAG representation in (b). Each time series is represented by a leaf node, and all three patterns are represented as internal nodes. Note that the DAG is different from similarity-based representations that are common in hierarchical clustering, where degrees of similarities are used to group sequences. In our case, length of overlap determines the position in the DAG and similarity is addressed as a single window-based threshold. As a result, the sequential nature of the underlying data is directly represented in the pattern conglomerate concept that is proposed as building block for further analysis in [13].

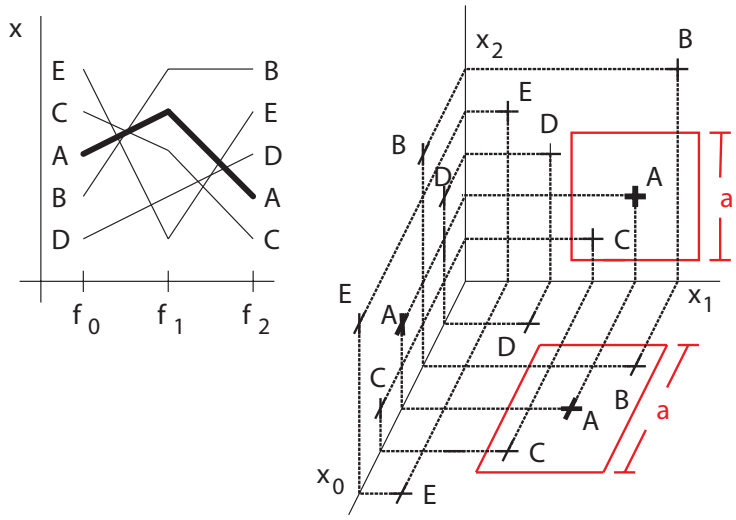

Figure 4: Example of five data points shown using parallel coordinates (left) and a vector space representation (right). The neighborhood of point A is dense for multiple subspaces. Two projections of a hypercube of side length a are shown. Note that those projections do not contain the same number of neighbors.

Detecting Non-Random Data in the Presence of Massive Noise

When combining information from many data sources, a central problem is not to include data that are randomly distributed. An algorithm for finding non-random data points in the presence of up to 95% noise has been developed [6, 5]. Rank-order-based scaling creates a flat distribution for each individual attribute, regardless of the distribution of values. Non-random data are identified through density maxima that result from the combined occurrence of similar attribute values in multiple data points. Similarity is evaluated over all axis parallel subspaces. Figure 4 illustrates how multiple subspaces can support a pattern. Its left panel shows a representation of five data points with three features (f_0,f_1, and f_2), using parallel coordinates. Alternatively one may look at the image as a set of five time series of length three that can be embedded in a three dimensional vector space. The vector space representation of the same data points can be seen in the right panel of Figure 4, where the data points are visualized through projections along

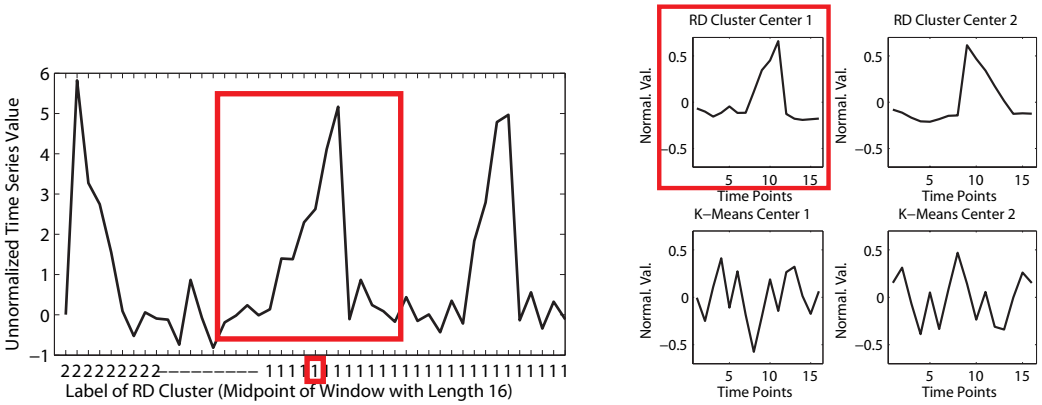

Figure 5: Left: Example of a time series together with labels corresponding to the cluster representatives from our algorithm, shown in the top right panel. Clustering is based on a section of the time series that is not shown. Bottom right: Result of k-means clustering with k=2 on same time series.

each of the three coordinate axes. Note that point A has two neighbors in the projection onto the $x_0 - x_1$ plane, and one neighbor in projections onto the $x_1 - x_2$ and $x_0 - x_2$ planes. A point is considered a neighbor if it is within a hypercube of side length a (see figure), or its projection. Note also that the neighbors differ for the two projections shown in Figure 4. In the projection along the x_0 axis only point C is a neighbor to A, while in the projection along the x_2 axis points B and C are A's neighbors. Only one point (C) is close to A in all dimensions. Subspace clustering techniques would not simultaneously consider both projections as evidence for high density. Effectiveness of the resulting algorithm is demonstrated on a real gene expression data set, for which noise has a clearly non-Gaussian distribution, as well as on time series data to which noise was added artificially. Comparison with conventional outlier detection illustrates the need for the new technique.

Data-Set-Specific Time-Series Subsequence Clustering

The problem of eliminating noisy data has also been addressed in the clustering context. We have introduced an algorithm that specifically addresses the goal of identifying source-specific clusters in time series subsequence data [7, 4]. Figure 5 illustrates the importance of disregarding noise in time series data. An artificial data set is shown that has two recognizable patterns,

one bell-shaped (pattern 1) and one funnel-shaped (pattern 2). Intermediate regions do not follow any particular pattern, and it is therefore appropriate that they should not be assigned to a cluster. Regions without obvious patterns are labeled "-", indicating that these are outliers or noise. The labeling is done based on the two cluster representatives that are shown in the top right plots. Note that the labeling in the figure is an experimental result. Conventional k-means has no mechanism for excluding noise. In fact, cluster members that are far from a cluster center have a particularly large impact on the location of k-means cluster centers, since the mean minimizes the sum square error of a set of values. Density-based clustering techniques disregard any data points that are in low-density regions of the attribute space. As a result, only those data points contribute to the cluster center definition that are in high-density regions.

We have shown that disregarding outliers is an important step in developing a robust clustering algorithm. Rather than relying on ad-hoc threshold values, our algorithm compares directly with what would be expected from random data. This comparison is done on a set of length scales to avoid limiting assumptions. In contrast to outlier detection algorithms, data points are not compared against the bulk of the data but rather against the distribution of random walk data. The rationale is that in time series data the majority of the data can be noise, and the assumption that outliers are rare may then no longer be satisfied. Our algorithm is effective over a wide range of window sizes, providing both high sensitivity and specificity in recognizing data from the same source. We have shown that kernel-density-based clustering loses its ability of identifying clusters in real-world data at window sizes of about 16-32. For our algorithm, in contrast, the best performance was seen for window sizes of w=16 to w=128. In summary, we were not only able to show that density-based techniques are able to overcome problems that had been observed in time series subsequence clustering using k-means and hierarchical clustering techniques.

Mining vector-item patterns

The previous two sections showed ways of identifying data that are significant beyond noise based on multiple continuous attributes, or vectors alone. We have taken the goal of distinguishing relevant from irrelevant information further by searching for patterns that involve vector attributes together with Boolean data [16, 8]. Subsets of the vector data, which are defined by the

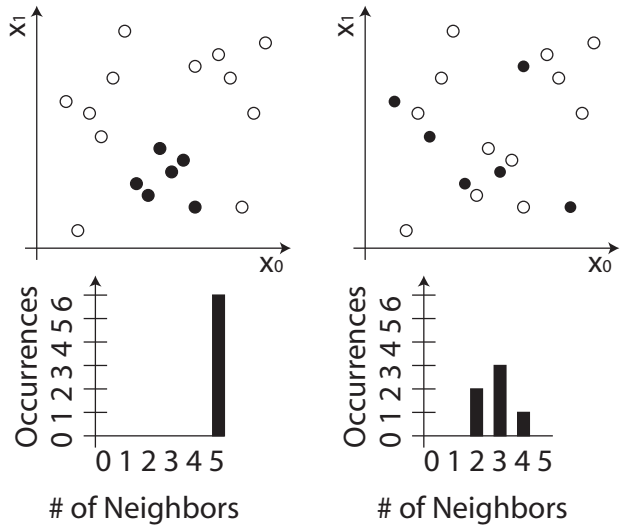

Figure 6: Schematic of a vector-item pattern between a 2-dimensional vector and one item. Records are represented by circles. The vector data determine the positions of the circles in the plane; existence of the item is represented as solid black filling. Bottom: Histograms summarize the distribution of neighbors with the item. Left: Records that have the item are close (vector-item pattern noticeable). Right: Same vector data as left but item data are distributed such that no vector-item pattern is noticeable.

presence of an item (or truth value of a Boolean attribute) define a distribution that is compared with the overall distribution of data points. Fig. 6 illustrates the problem of interest. Each of the circles represents an object or transaction. The spatial position of the object corresponds to a vector attribute, that is — in this example — two-dimensional. In general, a vector attribute is composed of D continuous attributes that are assumed to form a vector space. In Fig. 6, circles that are solid black represent objects, for which a particular item is present. Only a single item is represented in this image, but the process can be applied to many items. In the left panel the solid black circles are close together. The histogram under the left panel reflects the observation that there are six objects that have the item of interest, each of which has five neighbors that also have the item. The right panel shows objects with the same vector data as the left panel, but the item data are associated with different objects. Although the vector data are identical, the item data look far more distributed, and the histogram shows

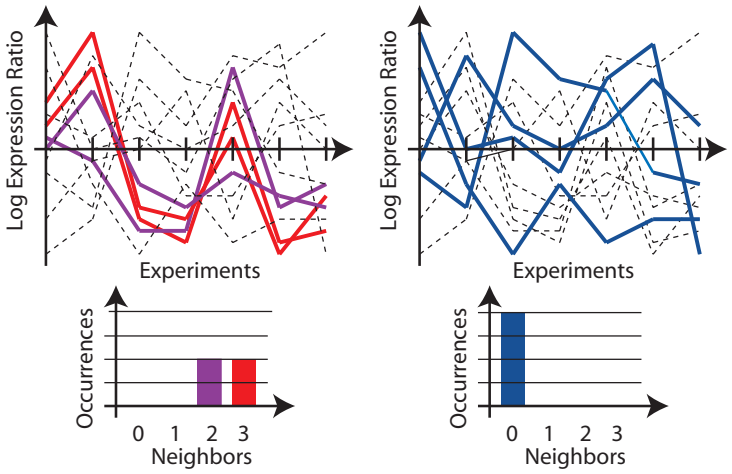

Figure 7: Sample expression profiles for two subsets of data. The top panels show gene expression profiles over multiple related experiments. The same set of curves are shown in the left and in the right panel. In each panel, a different subset of profiles is highlighted, corresponding to genes of a different functional designation. The number of neighbors, for all genes that show the function, is summarized in a histogram.

that relevant objects only have two to four neighbors. The setting on the left side illustrates what we consider a vector-item pattern. Fig. 7 shows an application of these concepts to the analysis of gene expression data together with functional information related to the proteins.

We have first introduced a histogram-based technique, and later developed a Kullback-Leibler-divergence-based approach that compares distributions directly. We have evaluated our algorithms on a variety of data sets and for a variety of problems including genomics [9], the chemistry of coatings [8], and stock market data [17]. We have shown that our algorithm typically produces more accurate results faster than comparison approaches that are based on testing whether classification leads to significant results on the data set.

Accomplishments in relationship to the proposed goals

The integration of different types of data into a coherent framework has been highly successful and dissemination has happened in many publications. The

remaining manuscripts are close to submission or under review. The research has had an unexpected but highly promising result in suggesting that the concept of aspect attributes is best generalized from just a single, traditional attribute to sets of continuous attributes, which we call a vector attribute. This outcome opens up a new perspective on the problem of data mining of diverse data, since combinations of vector attributes are not commonly discussed.

References

[1] C. Besemann and A. Denton. Mining edge-disjoint patterns in graph-relational data. In *Proc. Data Mining for Biomedical Informatics workshop in conjunction with the 7th SIAM Int'l Conf. on Data Mining*, Minneapolis, MN, April 2007.

[2] C. Besemann, A. Denton, N.J. Carr, and B.M. Prüß. BISON: bio-interface for the semi-global analysis of network patterns. *Source Code for Biology and Medicine*, 1:8, 2006.

[3] C.A. Besemann and A.M. Denton. A log-linear approach to mining significant graph-relational patterns. *Data & Knowledge Engineering (under review)*, 2009.

[4] A. Denton. Kernel-density-based clustering of time series subsequences using a continuous random-walk noise model. In *Proc. 5th IEEE Int'l Conference on Data Mining (ICDM'05)*, pages 122–129, Houston, TX, 2005.

[5] A. Denton and A. Kar. Finding differentially expressed genes through noise elimination. In *Proc. Data Mining for Biomedical Informatics workshop in conjunction with the 7th SIAM Int'l Conf. on Data Mining*, Minneapolis, MN, April 2007.

[6] A.M. Denton. Subspace sums for extracting non-random data from massive noise. *Knowledge and Information Systems*, 20:35–62, 2009.

[7] A.M. Denton, C.A. Besemann, and D.H. Dorr. Pattern-based time-series subsequence clustering using radial distribution functions. *Knowledge and Information Systems*, 18:1–27, 2009.

[8] A.M. Denton and J. Wu. Data mining of vector-item patterns using neighborhood histograms. *Knowledge and Information Systems (in press)*, 2009".

[9] A.M. Denton, J. Wu, M.K. Townsend, and B.M. Prüß. Relating gene expression data on two-component systems to functional annotations in *Escherichia coli*. *BMC Bioinformatics*, 9:294, 2008.

[10] D. Dorr and A. Denton. A pattern mining approach toward discovering generalized sequence signatures. In *Proc. of the SIAM International Conference on Data Mining (SDM08)*, Atlanta, GA, April 2008.

[11] D. Dorr and A.M. Denton. Generalized sequence signatures through symbolic clustering. In *Workshop on Machine Learning in Biomedicine and Bioinformatics of the Sixth International Conference on Machine Learning and Applications (ICMLA'07)*, Cincinnati, OH, Dec 2007.

[12] D.H. Dorr and A.M. Denton. Clustering sequences by overlap. *International Journal of Data Mining and Bioinformatics*, 3:260–279, 2009.

[13] D.H. Dorr and A.M. Denton. Establishing relationships among patterns in stock market data. *Data & Knowledge Engineering*, 68:318–337, 2009.

[14] D.H. Dorr and A.M. Denton. Generalized sequence signatures through symbolic clustering. *International Journal of Data Mining and Bioinformatics (in press)*, 2009.

[15] B.M. B.M. Prüß, C. Besemann, A. Denton, and A.J. Wolfe. A complex transcription network controls the early stages of biofilm formation. *J. Bacteriol.*, 188:3731–3739, 2006.

[16] J. Wu and A.M. Denton. Mining vector-item patterns for annotating protein domains. In *Mining Multiple Information Sources Workshop in conjunction with the ACM KDD '07 Conf. on Knowledge Discovery and Data Mining*, San Jose, CA, Aug. 2007.

[17] J. Wu, A.M. Denton, O. El-Ariss, and D. Xu. Mining core patterns in stock market data. In *Mining Multiple Information Sources Workshop in conjunction with the 2009 IEEE Int'l Conf. on Data Mining*, Miami, FL, Dec. 2009.