CLUSTERING GUIDANCE AND QUALITY EVALUATION USING RELATIONSHIP-BASED VISUALIZATION

ALEXANDER STREHL AND JOYDEEP GHOSH
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712-1084, USA
E-mail: strehl@ece.utexas.edu, ghosh@ece.utexas.edu

ABSTRACT
In this paper, we introduce CLUSION, a clustering visualization toolkit, that facilitates data exploration and validation of clustering results. CLUSION is especially suitable for very high-dimensional spaces since it is based on pair-wise relationships of the data points rather than their original vector representation. Given a similarity semantic, guidance to answer the questions 'What is the right number of clusters?', 'Which cluster should be split?', and 'Which clusters should be merged?' can be drawn from the visualization. The visualization also induces a quality metric not biased to increase with the number of clusters. We present examples from market basket data characterized by high-dimensional ($d > 10,000$), highly sparse customer-product matrices with positive ordinal attribute values and significant amount of outliers.

INTRODUCTION

Knowledge discovery in databases often requires clustering the data into a number of distinct segments or groups in an effective and efficient manner. Good clusters show high similarity within a group and low similarity between any two different groups. Automatically generated document clusters, for example, can provide a structure for organizing large bodies of text for efficient browsing and searching. Customer buying behavior groups are useful for presenting summary information to middle-level management for making marketing decisions. In recent data mining applications, a feature usually corresponds to some property (e.g., quantity, price, profit) of the objects under consideration. E.g., if we consider each customer as a data point the number of dimensions are very high and classical visualization techniques (such as 2,3–dimensional projections, parallel axis plots) become infeasable.

Let n be the number of objects (customers) in the data and d the number of features (products) for each sample $\mathbf{x}_j$ with $j \in \{1, \ldots, n\}$. The input data can be represented by a $d \times n$ product-customer matrix $\mathbf{X}$ with the j-th column representing the sample $\mathbf{x}_j$. Hard clustering[1] assigns a label λ_j to each d–dimensional sample $\mathbf{x}_j$, such that similar samples tend to

[1] In *soft* clustering, a record can belong to multiple clusters with different degrees of "association" [4].

get the same label. The number of distinct labels is k, the desired number of clusters. In general the labels are treated as nominals with no inherent order, though in some cases, such as self-organizing feature maps (SOFMs) or top-down recursive graph-bisection, the labeling may contain extra ordering information. Let $\mathcal{C}_\ell$ denote the set of all customers in the ℓ-th cluster ($\ell \in \{1, \ldots, k\}$), with $\mathbf{x}_j \in \mathcal{C}_\ell \Leftrightarrow \lambda_j = \ell$ and $n_\ell = |\mathcal{C}_\ell|$. Our batch clustering process maps from a set of raw object descriptions $\mathcal{X}$ via the vector space description $\mathbf{X}$ and similarity space description $\mathbf{S}$ to the cluster labels λ: $(\mathcal{X} \in \mathcal{I}^n) \xrightarrow{\Upsilon} (\mathbf{X} \in \mathcal{F}^n \subset \mathbb{R}^{d \times n}) \xrightarrow{\Psi} (\mathbf{S} \in \mathcal{S}^{n \times n} = [0,1]^{n \times n} \subset \mathbb{R}^{n \times n}) \xrightarrow{\Phi} (\lambda \in \mathcal{O}^n = \{1, \ldots, k\}^n)$.

FROM FEATURES TO SIMILARITY

While most of the well-known clustering techniques [3, 1] have been for numerical features, certain recent approaches assume categorical data [2]. In general, non-binary features are more informative since they capture noisy behavior better for a small number of samples. For example, in market-basket data analysis, a feature typically represents the absence (0) or presence (1) of a particular product in the current basket.

As the foundation for grouping decisions, an appropriate problem-specific similarity measure has to be defined. A similarity measures $s \in [0, 1]$ captures how related two data-points $\mathbf{x}_a$ and $\mathbf{x}_b$ are. All similarity measures considered are symmetric $s(\mathbf{x}_a, \mathbf{x}_b) = s(\mathbf{x}_b, \mathbf{x}_a)$ and define self-similarity $s(\mathbf{x}_a, \mathbf{x}_a) = 1$. Also, samples are positive $\forall i \forall j (x_{i,j} \geq 0)$ and non-zero $\forall i \exists j (x_{i,j} \neq 0)$. Clearly, if clusters are to be meaningful, the similarity measure should be invariant to transformations natural to the problem domain. Also, normalization may strongly affect clustering in a positive or negative way. The features have to be chosen carefully to be on comparable scales and similarity has to reflect the underlying semantics for the given task. A detailed discussion of the properties of the similarity measures can be found in [8] and is beyond the scope of this paper. Unlike cosine, the extended Jaccard coefficient enables us to discriminate by the total value of market-baskets *and* overcome the issues of Euclidean distances in high-dimensional sparse data. In our case study, we use the extended Jaccard similarity given by $s^{(\mathrm{J})}(\mathbf{x}_a, \mathbf{x}_b) = \frac{\mathbf{x}_a^\dagger \mathbf{x}_b}{\|\mathbf{x}_a\|_2^2 + \|\mathbf{x}_b\|_2^2 - \mathbf{x}_a^\dagger \mathbf{x}_b}$ as the relationship semantic for purchasing behavior.

Clusion: Cluster Visualization

To further support our design choices (using monetary value, extended Jaccard coefficient), as well as to visualize clustering results, we built CLUSION, which is briefly described in this section. When data is limited to 2 dimensions (or 2.5 at most [5]), the most powerful tool for clustering and for judging cluster quality is usually the human eye. CLUSION, our CLUSter visualizatION toolkit, allows us to convert high-dimensional data into a perceptually more suitable format, and employ the human vision system to explore the relationships in the data, guide the clustering process, and verify the quality of the results. In our experience with two years of DELL customer data, we found CLUSION effective for getting clusters balanced w.r.t. number of customers or net dollar amount, and even more so for conveying the results to upper management.

Given a similarity measure, the clustering problem can be moved from the original feature space ($d \times n$ data matrix $\mathbf{X}$) to similarity space ($n \times n$ similarity matrix $\mathbf{S}$). We visualize the symmetric similarity matrix $\mathbf{S}$ with entries $s_{a,b} = s(\mathbf{x}_a, \mathbf{x}_b)$ as a gray-level image where a black (white) pixel corresponds to minimal (maximal) similarity of 0 (1). The intensity (gray level value) of the pixel at row a and column b corresponds to the similarity between the samples $\mathbf{x}_a$ and $\mathbf{x}_b$. The CLUSION cluster visualization toolkit is used to verify cluster quality and fine-tune clusterings. It rearranges the columns and rows of the similarity matrix such that points with the same cluster label are contiguous: We propose CLUSION a new visualization technique for clusterings that gives a comprehensive visual impression of a high-dimensional data-set employing a *relationship-centered* approach. Common projective techniques, such as the selection of dominant features and optimal linear projections (SVD), are *object-centered*: Each object's features are transformed and visualized directly. In CLUSION, the actual features are transparent, instead, all pair-wise relationships, the relevant aspect for the purpose of clustering, are displayed. CLUSION guides the user towards appropriate design choices (such as the right number of clusters), makes splitting and merging opportunities visible, allows sanity checking the data (for outliers and bad features) as well as the imposed similarity measures.

In CLUSION, the similarity matrix $\mathbf{S}$ is permuted with a permutation matrix $\mathbf{P}$ which is defined as follows:

$$p_{i,j} = \begin{cases} 1 & \text{if } j = \sum_{a=1}^{i} l_{a,\lambda_i} + \sum_{\ell=1}^{\lambda_i - 1} n_\ell \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Please note that l are entries in the binary $n \times k$ cluster membership indicator matrix $\mathbf{L}$. Now, the permuted similarity matrix $\mathbf{S}'$ and the corresponding label vector λ' and data matrix $\mathbf{X}$' are: $\mathbf{S}' = \mathbf{PSP}^\dagger \quad , \quad \lambda' = \mathbf{P}\lambda \quad , \quad \mathbf{X}' = \mathbf{PX}$.

The similarity *within* cluster ℓ is thus represented by the average intensity within a square region with side length n_ℓ, around the main diagonal of the matrix. The off-diagonal rectangular areas visualize the relationships *between* clusters. The brightness distribution in the rectangular areas yields insight towards the quality of the clustering and possible improvements. In order to make these regions apparent, thin horizontal and vertical lines are used to show the divisions into the rectangular regions. Visualizing similarity space in this way can help to quickly get a feel for the clusters in the data. Even for a large number of points, a sense for the intrinsic number of clusters k in a data-set can be gained. Many strong relations suggest few and large natural clusters, while few relations indicate more and small clusters.

If the gray-level distribution of on-diagonal regions exposes bright as well as dark pixels the cluster should be split. On-diagonal regions which are uniformly bright are compact and should not be split further. A bright off-diagonal region may suggest that the clusters in the corresponding rows and columns should be merged. This induces an intuitive mining process that guides the user to the 'right' number of clusters k. A too small k leaves the on-diagonal regions inhomogenious. On the contrary, growing k beyond the natural number of clusters will introduce bright off-diagonal regions. OPOSSUM [7] uses the resulting insight to automatically pick the optimal number of clusters k. CLUSION has several other powerful properties. For example, it can be integrated with product hierarchies (meta-data) to provide simultaneous customer and product clustering, as well as multi-level views / summaries. It can be used with a recursive spectral decomposition method [6], so one can visualize clusters within a cluster. For details and demos, see `http://lans.ece.utexas.edu/~strehl`.

EXPERIMENTAL RESULTS

We experimented with real retail transactions of 21672 customers of a drugstore. For the illustrative purpose of this paper, we randomly selected 2500 customers. The total number of transactions (cash register scans) for these customers is 33814 over a time interval of three months. We rolled up the product hierarchy once to obtain 1236 different products purchased. 15% of the total revenue is contributed by the single item `Financial-Depts` which was removed because it was too common. 473 of these products accounted for less than \$25 each in toto and were dropped. The remaining $d = 762$ features and $n = 2466$ customers (34 customers had empty baskets after removing the irrelevant products) were clustered using OPOSSUM.

Figure 1 shows the similarity matrix (75% sparse) visualization before (a), after generalized k-means clustering using the standard Jaccard (b), after sample balanced (b), and after value balanced clustering (c). As the relationship-based CLUSION shows, OPOSSUM (c) and (d) gives more compact (better separation of on- and off-diagonal regions) and perfectly balanced clusters as compared to, for example, k-means (b). In k-means, the standard clustering algorithm (which can be generalized by using $-log(s^{(\mathrm{J})})$ as distances), the clusters contain between 48 and 597 customers contributing between \$608 and \$70443 to revenue encumbering a good overview of the customer behavior by marketing. Moreover clusters are hardly compact: Brightness is only slightly better in the on-diagonal regions in (b). All visualizations have been histogram equalized for printing purposes. In Figure 1 (c), for example, cluster 4 exposes a bimodal on-diagonal gray-level distribution indicating a splitting opportunity. On the contrary, clusters 15 and 16 are highly similar and should be merged.

CONCLUDING REMARKS

The visualization toolkit CLUSION allows managers and marketers to get an intuitive visual impression of the group relationships and customer behavior extracted from the data. This is very important for the tool to be accepted and applied by a wider community. The OPOSSUM / CLUSION combine has been successfully applied to several real-life market-baskets. We are currently working on an on-line version of OPOSSUM that incrementally updates clusters as new data points become available. Moreover, methods of scaling to very large numbers of transactions, using parallel data-flow approaches are being investigated.

Acknowledgements We want to express our gratitude to Mark Davis, Net Perceptions for providing the drugstore retail data set, and Carter Forringer of DELL, Round Rock for the DELL customer data set. This research was supported in part by the NSF under Grant ECS-9000353, a LARIAT grant from DELL, and by a gift from KD1.

References

[1] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* Wiley, New York, 1973.

[2] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Rock: a robust clustering algorithm for categorical attributes. In *Proceedings of the 15th International Conference on Data Engineering*, 1999.

[3] John A. Hartigan. *Clustering Algorithms.* Wiley, New York, 1975.

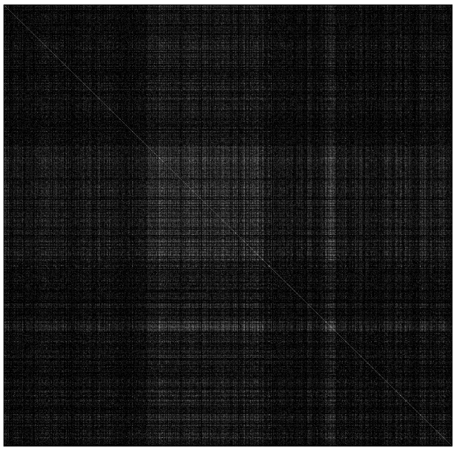
[4] S. Kumar and J. Ghosh. GAMLS: A generalized framework for associative modular learning systems. In *Proceedings of the Applications and Science of Computational Intelligence II*, pages 24–34, Orlando, Florida, 1999.

[5] D. Marr. *Vision.* W. H. Freeman Publishing Co., San Francisco, 1982.

[6] A. Pothen, H. Simon, and K. Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal of Matrix Analysis and Applications*, 11:430–452, 1990.

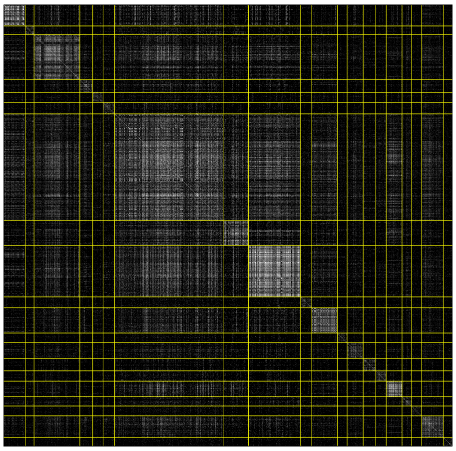
[7] Alexander Strehl and Joydeep Ghosh. Value-based customer grouping from large retail data-sets. In *Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery, 24-25 April 2000, Orlando, Florida, USA*, volume 4057, pages 33–42. SPIE, April 2000.

[8] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney. Impact of similarity measures on web-page clustering. In *Proc. Workshop of Artificial Intelligence for Web Search (AAAI 2000)*, 30-31 July 2000, Austin, Texas, USA. AAAI, July 2000, pp. 58-64.



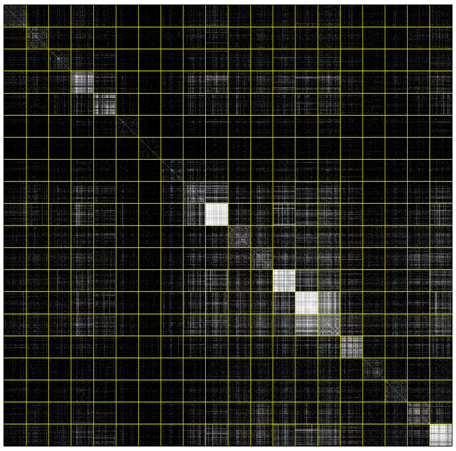
2466 customers, \$126899 revenue

(a)



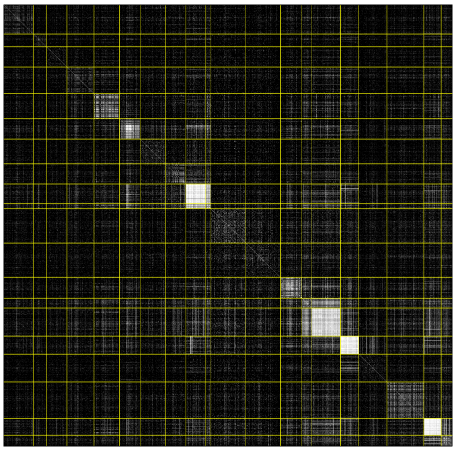
48 - 597, \$608 - \$70443

(b)



122 - 125, \$1624 - \$14361

(c)



28 - 203, \$6187 - \$6609

(d)

Figure 1: Results of clustering drugstore data. Relationship visualizations using CLUSION (a) before, (b) after k-means binary Jaccard, (c) after sample balanced OPOSSUM (d) value balanced OPOSSUM clustering with $k = 20$. In (b) clusters are neither compact nor balanced. In (c) and (d) clusters are very compact as well as balanced.