HHS Public Access

Author manuscript

Int Conf Comput Netw Commun. Author manuscript; available in PMC 2016 September 22.

Published in final edited form as:

Int Conf Comput Netw Commun. 2016 February ; 2016: .

ESammon: A Computationaly Enhanced Sammon Mapping based on Data Density

Chanpaul Jin Wang[1,2], **Hua Fang**[1,*], and **Honggang Wang**[2]

[1]Department of Quantitative Health Science, University of Massachusetts Medical School, Worcester, USA

[2]Department of Electrical and Computer Engineering, University of Massachusetts Dartmouth, North Dartmouth, MA, USA

Abstract

Sammon mapping is a widely used visualization technique to display complex data from high- to low-dimensional space. However, its extensive computational cost may pose potential computational challenges to big data visualization. This paper proposes a computationally-enhanced Sammon mapping (ESammon) by leveraging the characteristics of spatial data density. Unlike the conventional Sammon, ESammon preserves critical pairwise distances between data points in the process of projection, instead of all distances. Specifically, we integrated the Directed-Acyclic-Graph (DAG) based data density characterization method to select the critical distances. The numerical results demonstrated that our ESammon can achieve comparable projection results as the conventional Sammon mapping while reducing the computational cost from $O(N^2)$ **to** $O(N)$.

Index Terms

Sammon mapping; data density; Multidimensional scaling (MDS)

I. Introduction

Dimension reduction plays an integral role in high-dimensional data mining, especially for big data [1]. The cognitive capabilities of humans enable us to rapidly identify data structures such as clusters, homogeneous regions, or outliers. Multi-dimensional scaling (MDS) is one of the most widely used tools for dimension reduction [2], [3]. It builds upon the pairwise relations between individual data points, and can help characterize the structure of high-dimensional data. Sammon mapping is a typical MDS technique to map complex data from high- to low-dimensional space and has been widely applied in health and medical data analyses [4], [5]. However, preserving all pair-wise data distances in dimension reduction can cause significant computational complexity especially for big data. Besides, optimization based on all pairwise distance could be redundant. As shown in Fig. 1, $x_j \in R^d$ denotes a d-dimension data point, $C_j \in R^d$ as a Cluster centroid S_j. The solid dots A and B

* huajulia.fang@umassmed.edu.
chanpaul.wang@umassmed.edu, hwang1@umassd.edu

are two local cluster centroids, and the hollow dot C is near B. $d(\cdot)$ denotes the Euclidean distance. The distance $d(A, B)$ between two local Cluster Centroids A and B represents a critical part of the global cluster structure. Additionally, the hollow dot C is close to B, and the distance $d(B, C)$ represents the local cluster structure. However, compared with $d(A, B)$ and $d(B, C)$, $d(A, C)$ is less important in characterizing the cluster structure. In fact, the structural information described by $d(A, C)$ can be approximated by $d(A, B)$ and $d(B, C)$, and therefore redundant. Since dimension reduction aims at describing the cluster structure, it would be sufficient to preserve critical distances, e.g., $d(A, B)$ and $d(B, C)$ instead of all distances.

Besides, the typical gradient decent optimization algorithm used by Sammon mapping may lead to a local optimum, thus affecting the entire data structure. For example, as shown in Fig. 2, in the original high-dimensional space, point A is located near data points B, C, and D; on the low-dimensional plane, A',B', C', and D' are corresponding to A, B, C, and D, respectively. A' is initialized outside the corresponding center of the three other points, and the distance between A' and D' is much greater than its original distance $d(A, D)$. Then, when A' moves close to D' during the gradient decent optimization, $d(A', B')$ and $d(A', C')$ will also become shorter than the corresponding original distances, and easily cause the overall Sammon stress converging to a local optimum. If we ignore some pair-wise distances, we may reduce the possibility of trapping into local optimum.

To solve these problems, this paper revises the Sammon mapping while achieving similar projection performance. We argue that it is sufficient to keep the spatial data structure by preserving some critical distances during Sammon projection, instead of all pairwise distances. Specifically, we propose a new method for critical distance selection based on data density, and then launch the Sammon mapping based on these selected critical distances. Overall, the main contribution of this paper is to reduce the computational complexity of Sammon mapping from $O(N^2)$ to $O(N)$ while achieving a comparable projection quality.

The remainder of this paper is organized as follows: Section II describes the background of conventional Sammon mapping, Section III demonstrates the design of our ESammon mapping, Section IV evaluates the performance of this new algorithm, and Section V concludes our work.

II. Background of Sammon Mapping

Sammon mapping aims at mapping complex data from high-to low-dimensional space (e.g., 2 dimensions). Specifically, for a $N \times n$ dataset $X \subseteq R^n$, where N is the number of points, ans n is the number of attributes of each data, Sammon mapping implements the corresponding low-dimensional projection $Y \subseteq R^q$ of X by finding N data points in the q-dimension data space ($q << n$). The pairwise distances $d^*(y_i, y_j)$, $y_i, y_j \in Y$ in the q-dimension space approximate its corresponding high-dimensional inter-data point distances $d(x_i, x_j)$, $x_i, x_j \in X$. To achieve this data projection process, Sammon mapping defines the error criterion (a.k.a Sammon stress) E as Eq. 1, and solves the minimization problem $Y_{optimal} = argmin_Y E$.

$$E=\frac{1}{\lambda}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{\left(d_{ij}-d_{ij}^{*}\right)^{2}}{d_{ij}} \quad (1)$$

where $\lambda=\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}d_{ij}$, and $d(\cdot)$ and $d^{*}(\cdot)$ denotes the Euclidean distance in high and low-dimensional space, respectively.

Sammon mapping typically applies the steepest decent method to minimize the stress E, and obtain the optimal solutions consisting of $N \times q$ variables y_{il}, $i = 1, 2, \cdots, N$, $l = 1, 2, \ldots, q$ on the projected space.

Assuming the *t-th* iteration of y_{il} as $y_{il}(t)$, the update of y_{il} can be achieved by Eq. 2-4.

$$y_{il}(t+1)=y_{il}(t)-\alpha\left[\frac{\frac{\partial E(t)}{\partial y_{il}(t)}}{\frac{\partial^{2}E(t)}{\partial^{2}y_{il}(t)}}\right] \quad (2)$$

where α is a nonnegative scalar constant, representing the step size for the gradient search. Some other searching related parameters can be computed as follows:

$$\frac{\partial E(t)}{\partial y_{il}(t)}=-\frac{2}{\lambda}\sum_{k=1,k\neq i}^{N}\left[\frac{d_{ki}-d_{ki}^{*}}{d_{ki}d_{ki}^{*}}\right](y_{il}-y_{kl}) \quad (3)$$

$$\frac{\partial^{2}E(t)}{\partial^{2}y_{il}(t)}=-\frac{2}{\lambda}\sum_{k=1,k\neq i}^{N}\frac{1}{d_{ki}d_{ki}^{*}}\left[(d_{ki}-d_{ki}^{*})-\left(\frac{(y_{il}-y_{kl})^{2}}{d_{ki}^{*}}\right)\left(1+\frac{d_{ki}-d_{ki}^{*}}{d_{ki}}\right)\right] \quad (4)$$

III. Algorithm Design of ESammon

The core idea of ESammon is to preserve critical pairwise distances, instead of all distances. First, we design a data density-based critical distance selection method, and then provide an optimal projection algorithm based on these critical distances.

A. How to select critical distances?

We define critical distances as the one that can characterize the profile of spatial data structures. For example, as demonstrated in Fig. 3, the solid dots represent the cluster centroids of corresponding clusters represented by the blue shadowed circles, and the hollow dots represent the data points in the clusters. The critical distances includes global and local critical distances. As shown in the graph. The pairwise distances between the solid dots are defined as the global critical distances, which characterize the global data structure. The

distances between the hollow dots to its corresponding centroids in the same shadowed circle is the local critical distance, which characterizes the local data structure.

In order to select the critical distances, we adapted data density definition defined by Rodriguez et al. [6]. They [6] characterized the data distribution with local density and the minimum distance between the data point and any other data points with higher local density as Eq. 5 and 6.

Assuming the dataset $X = \{x_1, x_2, \cdots, x_N\}$, for any data point $x_i = \left\{x_i^1, x_i^2, \ldots, x_i^n\right\}$ where n denotes the number of the attributes. Rodriguez et al. [6] computes the local data density ρ_i with Gaussian kernel function as Eq. 5.

$$\rho_i = \sum_{j=1}^{N} e^{-\left(\frac{\|x_i - x_j\|}{d_c}\right)^2} \quad (5)$$

Where $\|x_i - x_j\|$ denotes the Euclidean distance, and d_c as the cut-off distance.

The minimum distance between the point i and any other point with higher local density δ_i is defined as Eq. 6. In this paper, we call δ_i as the critical distance of data point i.

$$\delta(x_i) = min_{x_j:\rho(x_j)>\rho(x_i)}(\|x_i - x_j\|) \quad (6)$$

In particular, local density represents the compactness of spatial data structure. Large local density means more data points are distributed nearby, and may form a cluster. The critical distances describe the relationship between low-density data points and the nearest high-density data point. More generally, the local density and critical distance can describe the spatial data structure.

Based on the above data density characteristics, we further integrated our newly developed directed neighbor and Directed-Acyclic-Graph (DAG) method [7] to maintain the spatial data structure. Below we briefly show the definition of directed neighbor.

Definition 1—The directed neighbor of point x_I is point $x_j = argmin_{j:\rho(x_j)>\rho(x_i)} d(x_i, x_j)$, where $d(x_i, x_j)$ denotes the Euclidean distance of data x_i and x_j .

In particular, the data point with the highest density of the dataset has no directed neighbors. Hence, we set its critical distances $\delta = 0$. The other three properties of directed neighbors are as follows: (1) assuming the directed neighbor of data point p as q, then $\rho_p < \rho_q$; (2) each data point has only one directed neighbor except the data point having the largest local density; (3) the directed neighbor relation is not reversible. If the directed neighbor of Point p is point q, then point p is not the directed neighbor of point q. Fig. 4 shows the construction of our DAG method, where the vertices denote the data points, and the directed edge denotes the directed neighbor relationship. The critical distance is used to define the weight of directed edge. For example, the data point with the largest local density is treated

as the global anchor of other data points. Once it is initialized in the low-dimensional space, we can find the low-dimensional coordinates for the other data points based on their critical distances.

B. Design of ESammon based on data density graph

Assuming the original dataset $X = \{x_1, x_2, \ldots, x_N\}$, for any data point $x_i = \left\{x_i^1, x_i^2, \ldots, x_i^n\right\}$ where n denotes the number of the attributes, and the corresponding low-dimensional as $Y = \{y_1, y_2, \ldots, y_N\}$, $y_i = \left\{y_i^1, y_i^2, \ldots, y_i^q\right\}$, $q << n$, $q = 2$. We denote the directed neighbor of x_i as $neig(x_i)$, and the original critical distance of x_i as $\delta_i = min_{x_j:\rho_j>\rho_i}(\|x_i - x_j\|)$, and $\delta_i^* = min_{y_j:\rho_j>\rho_i}(\|y_i - y_j\|)$. Thus, our ESammon stress is expressed as follows:

$$E = \frac{1}{\lambda}\sum_{i=1}^{N}\frac{(\delta_i - \delta_i^*)^2}{\delta_i} \quad (7)$$

where $\lambda = \sum_{i=1}^{N}\delta_i$. Since the data point with the highest density of the dataset has no directed neighbors, and its critical distance $\delta = 0$, its low-dimensional position does not distort the structural information of the final projection. Our ESammon does not update it in the low-dimensional space. Thus, the minimization of E is an optimization problem with $2*(N-1)$ variables y_{il}, $i = 1, 2, \ldots, N$, $l = 1, 2$. We also adopt the steepest decent method to minimize E, and the update of y_{il} at the t-th iteration is expressed as Eq. 2. Particularly, we solve the parameters in the steepest decent method based on our ESammon stress as shown below.

$$\frac{\partial E(t)}{\partial y_{il}(t)} = -\frac{2}{\lambda}\left[\frac{\delta_i - \delta_i^*}{\delta_i \delta_i^*}\right](y_{il} - y_{kl}) \quad (8)$$

$$\frac{\partial^2 E(t)}{\partial^2 y_{il}(t)} = -\frac{2}{\lambda}\frac{\delta_i - \delta_i^*}{\delta_i \delta_i^*} + \frac{2}{\lambda}\frac{1}{(\delta_i^*)^2}\frac{(y_{il} - y_{kl})^2}{\delta_i^*} \quad (9)$$

where $k = neig(x_i)$.

Moreover, our ESammon preserves the main structural information of special data density. As demonstrated in Fig. 4, the data points located near the center of high-density areas usually have much larger critical distances than other data points, and thus influence the ESammon stress E more than other data points. In other words, these larger critical distances would be preserved more than any other data points in one cluster during our ESammon projection.

IV. ESammon Performance Evaluation

We evaluated ESammon with some real datasets as shown in Tab. I. Particularly, TDTA [8] is a longitudinal dataset with missing values less than 25%. We used multiple imputation-based method to deal with this incomplete data [9], [10]. The other datasets are from UCI [11]. IRIS includes three classes (three IRIS species: Setosa, Versicolor, and Virginica) with 50 samples each and four features (the length and the width of sepal and petal). Wine consists of 3 classes (3 types of wines grown in the same region in Italy but derived from three different cultivars); each class is characterized by 12 continuous-valued features, such as Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline. The evaluation was performed on Windows 7 platform with *Intel(R) Core(TM) i7-3520M 2.9 GHz* CPU and 8 *G* memory. Fig. 5 demonstrates the numerical results. To visualize the comparison between conventional Sammon and ESammon, we denote different clusters with numbers *1*, *2* and *3*. For TDTA, it consists of 3 clusters. Cluster *1* is perfectly separated, yet the other 2 clusters are overlapped. Compared with the conventional Sammon mapping, although our ESammon has several data points overlapped between Cluster *1* and *2*, we can still clearly detect the difference between Cluster *1* and *2*. For IRIS, its spatial data structure is similar to TDTA, where one cluster is perfectly separated and the other two are overlapped. As demonstrated by the experimental results, our ESammon achieved almost the same projection as the conventional Sammon for IRIS. The clusters of the WINE dataset are more overlapped, where both Sammon mappings seem to project the three clusters close to each other.

V. Conclusions and Future Work

Aiming at reducing the computational cost of the conventional Sammon Mapping, this paper proposed a newly designed Sammon mapping algorithm (ESammon). ESammon integrates the Directed-Acyclic-Graph (DAG) and Rodriguez's data density characterization methods to select the critical distances in order to reduce the computational complexity from $O(N^2)$ to $O(N)$. As the numerical results demonstrated, our ESammon achieved comparable projection performance as the conventional Sammon mapping.

However, for highly overlapped datasets, both Sammon and ESammon might not project data well. In the future, we could consider more sophisticated dimension-reduction techniques such as subspace clustering [12].

Acknowledgment

This research was supported by NIH grant R01 DA033323-01A1, 1UL1RR031982-01 Pilot Project to Dr. Fang.

REFERENCES

1. Fang H, Zhang Z, et al. A survey on big data research. IEEE Network Magazine. 2015; 29(5):6–9. [PubMed: 26504265]
2. Borg, L.; Groenen, P. Modern Multidimensional Scaling. Springer; New York: 1997.
3. Cox, T.; Cox, M. Multidimensional Scaling. Chapman & Hall; 1994.

4. Fang, H.; Zhang, Z., et al. Using visualization-aided trajectory pattern validation in a longitudinal dietary data. 36th Annual Meeting & Scientific Sessions; San Antonio, TX, USA. 2015; p. 22-25.
5. Fang, H.; Kim, S., et al. Visualization-aided trajectory pattern recognition approach to smoking and depression among asian americans. Society for Research on Nicotine and Tobacco 2015; Philadelphia, PA. 2015; p. 25-28.
6. Rodriguez A, Laio A. Clustering by fast search and find of density peaks. Science. Jun; 2014 344(6191):226–231.
7. Wang, CJ.; Fang, JH., et al. Dag-searched and density-based initial centroid location method for fuzzy clustering of big biomedical data. Proc. International Conference on Bio-inspired Information and Communications Technologies '14; Boston, USA. Dec. 1–3, 2014; p. 290-293.
8. Kim SS, Kim SH, et al. A culturally adapted smoking cessation intervention for korean americans: A mediating effect of perceived family norm toward quitting. J. Immigr. Minor Health. 2015; 17(4): 1120–11. 129. [PubMed: 24878686]
9. Fang H, Johnson C, et al. A new look at quantifying tobacco exposure during pregnancy using fuzzy clustering. Neurotoxicol Teratol. Jan-Feb;2011 33(1):155–165. [PubMed: 21256430]
10. Fang H, Dukic V, et al. Detecting graded exposure effects: a report on an east boston pregnancy cohort. Nicotine Tob. Apr; 2012 14(9):1115–1120.
11. [Online]. Available: http://archive.ics.uci.edu/ml/
12. Muller, E.; Assent, I., et al. Relevant subspace clustering: Mining the most interesting non-redundant concepts in high dimensional data. Proc. International Conference on Data Mining (ICDM) '14; Shenzhen, China. Dec. 14–17, 2014;

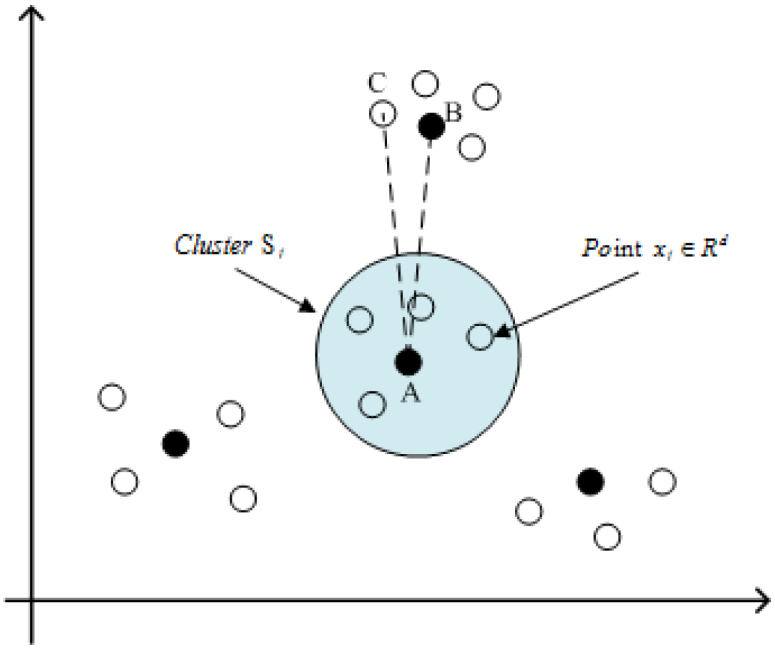


Fig. 1.
Demonstration of redundant pairwise distance in Sammon mapping

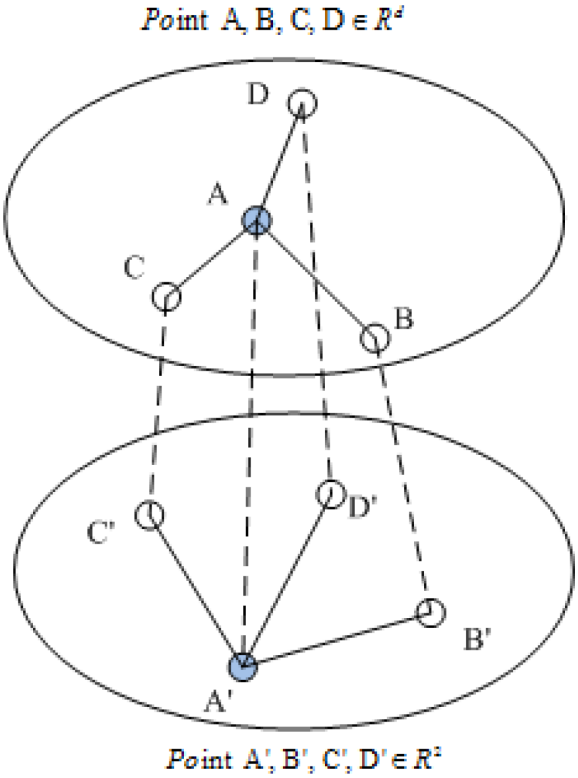


Fig. 2.
Demonstration of local convergence of Sammon mapping

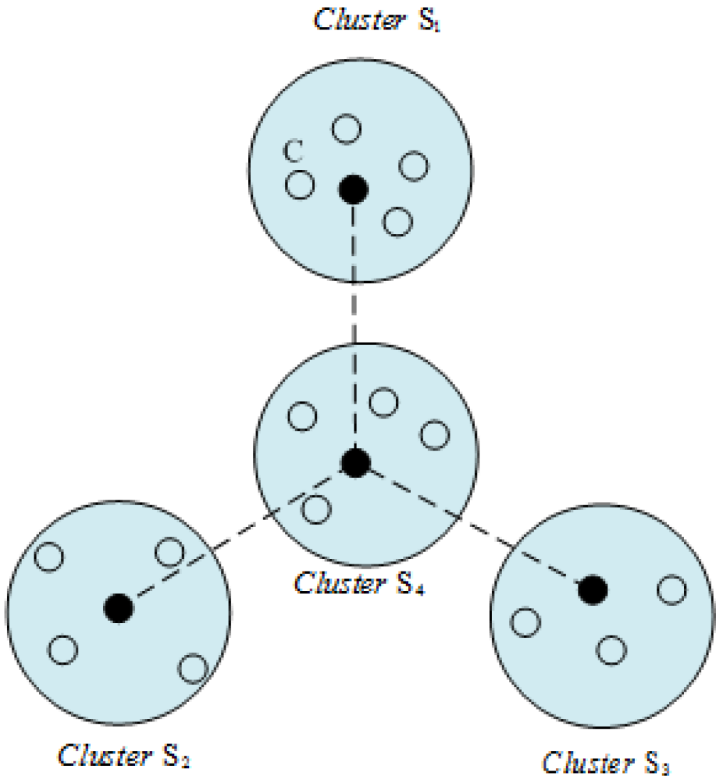


Fig. 3.
Illustration of critical distances

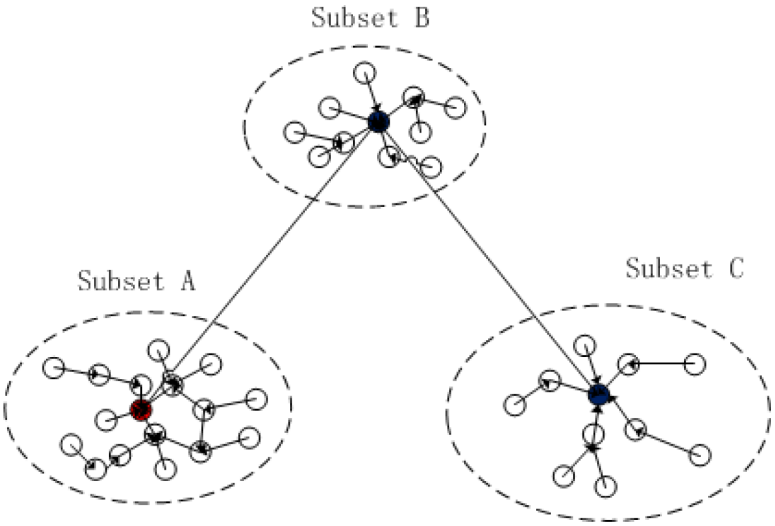
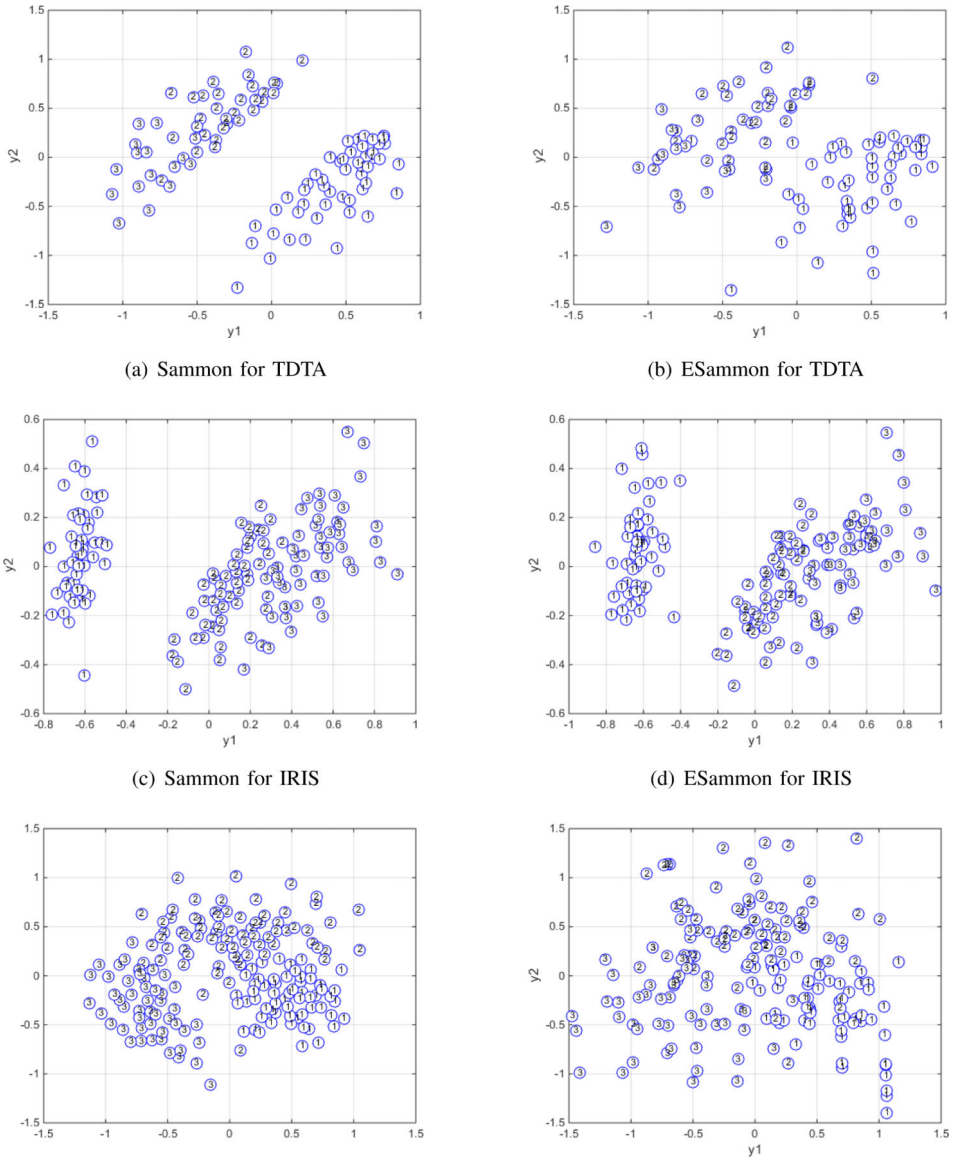


Fig. 4.
Demonstration of the DAG-based data density graph



(a) Sammon for TDTA (b) ESammon for TDTA

(c) Sammon for IRIS (d) ESammon for IRIS

(e) Sammon for Wine (f) ESammon for Wine

Fig. 5.

Sammon mapping vs. ESammon mapping for a) TDTA, b) IRIS, c) Wine

TABLE I

DESCRIPTION OF REAL DATASET

	Num. clusters	Size	Num. of Attributes
TDTA	3	97	9
IRIS	3	150	4
Wine	3	178	12