International Journal of Applied Mathematical Research, 4 (2) (2015) 336-344
www.sciencepubco.com/index.php/IJAMR
©*Science Publishing Corporation*
doi: 10.14419/ijamr.v4i2.4419
Research Paper

More faithfulness graph embedding

Alaa A. Najim

Mathematics Department, Science College, Basrah University, Basrah, Iraq
Email: alaanajim68@yahoo.co.uk

Copyright ©2015 Alaa A. Najim. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Using dimensionality reduction idea to visualize graph data sets can preserve the properties of the original space and reveal the underlying information shared among data points. Continuity Trustworthy Graph Embedding (CTGE) is new method we have introduced in this paper to improve the faithfulness of the graph visualization. We will use CTGE in graph field to find new understandable representation to be more easy to analyze and study. Several experiments on real graph data sets are applied to test the effectiveness and efficiency of the proposed method, which showed CTGE generates highly faithfulness graph representation when compared its representation with other methods.

Keywords: *Graph drawing, Information visualization and dimensionality reduction method.*

1. Introduction

Graph visualization is a sub-field of information visualization uses to visualize the set of data and the related information in a graph. Graph is simple model to directly explain the things without more interpretation. It has been used in various fields, as mathematics, computer science, biology, art, and psychology, because its ability to represent a system in a simple structure with a set of nodes and edges. Recently, the size of data is exponential growth, and representing this large data in a graph is essentially to reach the idea to reader in an easy way. However, large graph is incomprehensible and analysis it might be infeasible, where its time complexity might be greater than $O(n^3)$. Understanding large graph is essential, therefore, embedding it into simple representation is necessary. The main challenge encounters graph embedding is: what is an optimum low dimension space to represent graph in different tasks in order to effectively deliver information to readers?. Many methods have been introduced for graph visualization to discover its low dimension space in order to aid the reader to see and easy explore the relations among items [1], [2], [3].

In this paper, we concentrate on the straight line edges way to draw undirected connected graph, which represent the nodes coordinates of graph in two dimension space. A force directed model is a popular method defines a stress (energy) function among nodes to find their good positions [4]. This method starts with large energy and then iteratively minimized by updating the node positions according to the force direction. There are different versions of force directed methods, where the differences among them in selecting and minimizing the cost function [5], [6], [7]. In general, force-directed method works efficiently with small graph, but it is impossible to find good space of large graph because it is difficult to minimize its energy. In addition, the complexity of the force-directed method is very high, which probably require $O(n^2)$ where n is the number of nodes, because it updates all node positions in each iteration. Spectral method [8], [9] and multidimensional scaling [10] are also well-known linear methods used

to reduce the dimensionality of graph. These methods fail to find acceptable graph visualization when graph has nonlinear structure because their linearity [11].

In general, the dimensionality reduction of graph without lost some of information is impossible. In addition, representing relations among data in large graph is still incomplete solvable by graph drawing methods because the complexity of these relations [12]. Dimensionality reduction methods are classified into two types: continuity and trustworthy [13]. Continuity methods generate continuity projection where closed neighbor data points in the original space are also closed neighbors in the projected space. This type of those method concentrate on increasing the continuity of the projected space, as in MDS [10], PCA [9], SPE [14] and Isomap [15]. On the other hand, the trustworthy methods, as CCA and CDA, generate trustworthy projected space where nearest neighbor data points in the projected space are also nearby in the original space. Although the continuity error causes the projection be tearing, and trustworthy error causes the projections be fattening, the trustworthiness is more important than continuity in projected space [16]. The main drawbacks of CCA and its improved version CDA are might generate tearing projection because the several local optima might be found in their cost function [17], which causes the projection graph be usefulness.

2. Related works

3. Quality of the embedded space

Computing and comparing the quality of a given embedded space with the original data sets by visual inspection are difficult due to its high dimensionality. Thus, the best formal measurements should evaluate the amount of preserving neighborhood distances in the embedded space with their corresponding in the original data. Correlation (γ) [18], stress [15] and local continuity (LC) [19] are the well-known methods used in this matter. If we suppose X is a vector of all pairwise distance of the data points in the original space and Y is a vector of their corresponding pairwise point distances in the embedded space, then their definitions are:

Correlation function (γ): this function compute the linear correlation between original input distances and point distances in embedded space. The value of correlation is equal to 1 when all distances are perfect preserved, where positive slope between two vectors with perfect linear. In the other hand, the value equal to -1 if the two vectors have prefect linear relationship with negative slope. The correlation is defined by:

$$\gamma = \frac{X^T\ Y/|X| - \overline{X}\ \overline{Y}}{\sigma_X\ \sigma_Y} \tag{1}$$

where $|X|$ is the length of X, and $\overline{X}$ and σ_X are the mean and standard deviation of X, respectively.

Local Continuity (LC) computes the degree of similarity between two corresponding nearest neighbors sets in projected and original spaces. The average of all cases represents the efficiency measurement of a projected space. Formally, let suppose the k nearest neighbors set in the original space to data point i is $\mathcal{N}_k^X(i) = \{\, j_1, j_2, ..., j_k \,\}$, and the k nearest neighbors set to i in the projected space is $\mathcal{N}_k^Y(i) = \{\, l_1, l_2, ..., l_k \,\}$. The measurement of overlapping between two sets is evaluated by

$$N_k(i) = |\mathcal{N}_k^X(i) \cap \mathcal{N}_k^Y(i)| \tag{2}$$

$N_k(i)$ is normalized to the [0,1] interval in order to compute the faithfulness measure of i:

$$Faithfulness_k(i) = \frac{1}{k} N_k(i) \tag{3}$$

In this case the faithfulness value of projected space will be :

$$Faithfulness_k = \frac{1}{N} \sum_{i=1}^{N} Faithfulness_k(i) \tag{4}$$

where N is total number of data points in dataset.

3.1. Continuity and Trustworthiness

Dimensionality reduction methods are used to reduce the high-dimensionality space into the low-dimensionality space which preserves the point relations of the original data. The goal is to explore the useful representation aids the reader to analysis and process a low space directly. In general, there are two types dimension reduction methods: continuity and trustworthy. Continuity methods try to find low dimension space hold the global structure of original space as much as possible. Thus, these methods focus on original point relations in original space to update their corresponding relations in low space. For example, SPE generates low-dimensionality space preserves the original space structure. However, the low-dimensionality space might encounter tearing and/or fattening problems which change the low space figure and make it be distorted. These problems are happened when its difficult to unfold the sharing relations of points in the original space.

On the other hand, trustworthiness idea depends on the point relations in the low-dimensionality space rather than those relations in the original space to find perfect low space. The points, in the trustworthy idea, can be updated their coordinates in low space in a flexible way without constrains to preserve the structure of the original space. CDA method has succeeded to prove this idea, where it explores efficient low-dimensionality space by depending on the point relations in low-dimensionality space. However, the computing geodesic distances leads CDA is not suitable to deal with very high-dimensionality data sets. Furthermore, the pure working under low-dimensionality space might cause low-dimensionality space be tearing.

In this paper, the ideas of continuity and trustworthy will be used together to give more robustly for a low-dimensionality space, where trustworthy takes a master role and assigning secondary role to continuity.

4. Notation and definition

Let us suppose there is a indirect connected graph $G = (X, V)$, where $X = \{x_1, x_2, ..., x_n\}$; and each node at maximum has n indirect edges. The reduction method attempts to find a low-dimensional graph space by placing x_j onto the projected space in such a way that their euclidean distance $d_{ij} = ||x_i - x_j||$ is closed to the corresponding distance (r_{ij}) in the original high-dimensional space. Therefore, if the points in a local region are preserved there distances, the final projected space preserves all the local distances.

5. Methodology: Continuity Trustworthy Graph Embedding (CTGE)

The aim of dimensionality reduction method is to preserve the original graph structure as much as possible by minimize the following stress function:

$$Stress = \sum_{i \neq j} (r_{ij} - d_{ij})^2 \tag{5}$$

Although this thing is very important to show more details, some times, we want to show the main important relations among points without going to show everything. To do that, we require to:

1. Classify graph points into several groups depending on their connections similarity.
2. Separate these groups as much as possible in order to focus on the general relations among groups.

In order to satisfy the above two goals, we require to add continuity and trustworthy modes to the dimensionality reduction.

The trustworthiness part can be defined: the trustworthy stress function of the dimensionality reduction is divided on the distance d_{ij} and multiplied it with trustworthiness weight function to decrease the trustworthy error as in Equation 6. This devision gives points, which are preserving their distances in original and projected spaces, high weight:

$$f_{trustworthy} = \sum_{i \neq j} \frac{(r_{ij} - d_{ij})^2}{d_{ij}} T_f(r_{ij}, d_{ij}, d_c(t_k)) \tag{6}$$

where the trustworthy local function T_f is defined as:

$$T_f(r_{ij}, d_{ij}, d_c(t_k)) = \begin{cases} 1 & if(d_{ij} \leq d_c(t_k)) or((d_{ij} > d_c(t_k)) \\ & \quad and(d_{ij} < r_{ij})), \\ 0 & Otherwise \end{cases} \tag{7}$$

The role of $f_{trustworthy}$ is to classify the points into groups according to their connections similarities, where the points are updated to be as groups when they satisfy the following constrain:

$$if(d_{ij} \leq d_c(t_k))or((d_{ij} > d_c(t_k))and(d_{ij} < r_{ij})) \tag{8}$$

$$x_j \leftarrow x_j + \lambda(t_k)\frac{r_{ij}^2 - d_{ij}^2}{d_{ij}^2 + \epsilon} \tag{9}$$

To preserve the distance among points in each group, we should define continuity part. In Equation 10, the stress function $\sum_{i \neq j}(r_{ij} - d_{ij})^2$ is divided on the distance r_{ij} and multiplying it with continuity weight function:

$$f_{continuity} = \sum_{i \neq j} \frac{(r_{ij} - d_{ij})^2}{r_{ij}} T_c(r_{ij}, d_{ij}, r_c) \tag{10}$$

where the continuity weight function T_c is defined as:

$$T_c(r_{ij}, d_{ij}, r_c) = \begin{cases} 1 & if(r_{ij} \leq r_c)or((r_{ij} > r_c) \\ & \quad and(d_{ij} < r_{ij})), \\ 0 & Otherwise \end{cases} \tag{11}$$

Thus, the points can be updated to preserve their distances among groups as much as possible:

$$if(r_{ij} \leq r_c)or((r_{ij} > r_c)and(d_{ij} < r_{ij})) \tag{12}$$

$$x_j \leftarrow x_j + \lambda(t_k)\frac{r_{ij} - d_{ij}}{r_{ij} + \epsilon} \tag{13}$$

The trustworthiness definition in Equation 6 can be merged with the continuity definition in Equation 10 to generate continuity and trustworthy embedding method. This method will be applied on the graph data sets to reduce their dimensionality, thus, the proposed method will be called continuity trustworthy graph embedding (CTGE):

$$f_{CTGE} = \sum_{i \neq j}(\alpha \frac{(r_{ij} - d_{ij})^2}{d_{ij}} T_f(r_{ij}, d_{ij}, d_c(t_k)) \ + (1 - \alpha)\frac{(r_{ij} - d_{ij})^2}{r_{ij}} T_c(r_{ij}, d_{ij}, r_c)) \tag{14}$$

where α is important to get an optimum two dimensionality graph representation, where $(0 < \alpha < 1)$. According to α value, the distance among groups are specify, so the points relations within a group are determine. When increasing α value the method be near to trustworthy mode, otherwise it goes to continuity mode.

By derivation Equation 14, we get the updating function which put original points in a proper locations in two dimension representation of graph:

$$\frac{\partial f_{CTGE}}{\partial d_{ij}} = (-1)\sum_{i \neq j}(\alpha \frac{r_{ij}^2 - d_{ij}^2}{d_{ij}^2} \ T_f(r_{ij}, d_{ij}, d_c(t_k)) \ + (1 - \alpha)\frac{2(r_{ij} - d_{ij})}{r_{ij}} \ T_c(r_{ij}, d_{ij}, r_c))) \tag{15}$$

6. Data sets

Two graph data sets will be embedded by dimensionality reduction. The first data sets, American College Football, is formed by teams in nodes, and the two teams are connected by edge if they have played each other that season. This graph has 115 nodes (teams) and 613 edges (games). The goal is to find a two-dimensional representation to classify these data sets. It is available at https://Networkdata.ics.uci.edu/data.php?id=5. The Primary School graph is the second data sets and represents the relationship between students in different classes. This graph has 10 classes, with nine of them being student classes and one being a teacher class. This data sets is available at http://www.sociopatterns.org/datasets/primary-school-cumulative-Networks/.

7. Experimental results

We compared the proposed method (CTGE) with 19 dimensionality reduction methods, where this comparisons are evaluated in quantitative and visual manners.

First Data Sets

The American College Football Graph data sets considers the games played between 115 teams in the year 2000. There is a priori knowledge about gathering these teams into groups. Dimensionality reduction methods can classify this data sets by reducing its dimensionality into a small number of dimensions where most of the information is concentrated. In this section, we will evaluate the performance of CTGE in classifying this graph, and compare them with 19 dimensionality reduction methods.

Table 1: Measuring the low-dimensional representation of the American College Football graph, which was carried out by 20 methods, by using correlation, LC and stress metrics. (Correlation and LC: the highest is better, Stress: the lowest is better)

Method	Correlation	LC	Stress
PCA	0.673	0.579	0.429
CCA	0.640	0.529	0.543
CDA	0.591	0.594	0.176
Factor_analysis	0.462	0.515	0.616
FastMVE	0.312	0.362	9.363
Hessian_LLE	0.294	0.417	0.579
Isomap	0.582	0.512	0.836
Kernel_PCA	0.156	0.356	0.948
Laplacian	0.613	0.598	0.976
LLC	0.189	0.246	0.880
LLE	0.568	0.545	0.926
LLTSA	0.306	0.575	0.941
LPP	0.449	0.499	0.983
LTSA	0.306	0.575	0.946
NPE	0.463	0.493	0.952
Prob_PCA	0.670	0.577	0.327
SPE	0.638	0.567	0.290
SNE	0.705	**0.638**	0.285
tSNE	0.667	0.571	1.693
CTGE	**0.766**	0.585	**0.083**

Figure 1 shows five of the low-dimensional representations, which are carried by PCA, Prob_PCA, SNE, tSNE and CTGE, of the American College Football graph data sets. The classification of this data sets is clear in this figure. We can see the classification by SNE, tSNE and CTGE is better than by PCA and Prob_PCA. We measured the efficiency of 20 results by using correlation, LC and stress metrics, as shown in Table 1. Although CTGE has got higher correction and lower stress values, SNE is better when using the LC metric. Thus, in classification, we cannot say that our methods are the best, because some other methods are better. However, our methods obtained an acceptable classification for this data sets, while some methods failed in this task. The relationship between correlation and stress metrics is shown in Figure 2, which shows that the efficiency of some visualizations are not good, in that these visualizations generate a large amount of error, as in the case of Kernel_PCA, LLC, FastMVU, LTSA, LLTSA and Hess_LLE. In addition, we can show that some methods have a high degree of efficiency using the correlation metric and, at the same time, a very large error, as in the case of the tSNE and the Laplacian method. Other methods have remained in the middle in that they are not as efficient as expected for their results to be relied on, but their results are not so bad that they cannot be rejected at all.

Second Data Sets

The dimensionality of the Primary School graph data sets is 236. The dimensionality reduction methods will be used to reduce this high degree of dimensionality into a 2-dimensional space. This graph consist of 10 classes, and

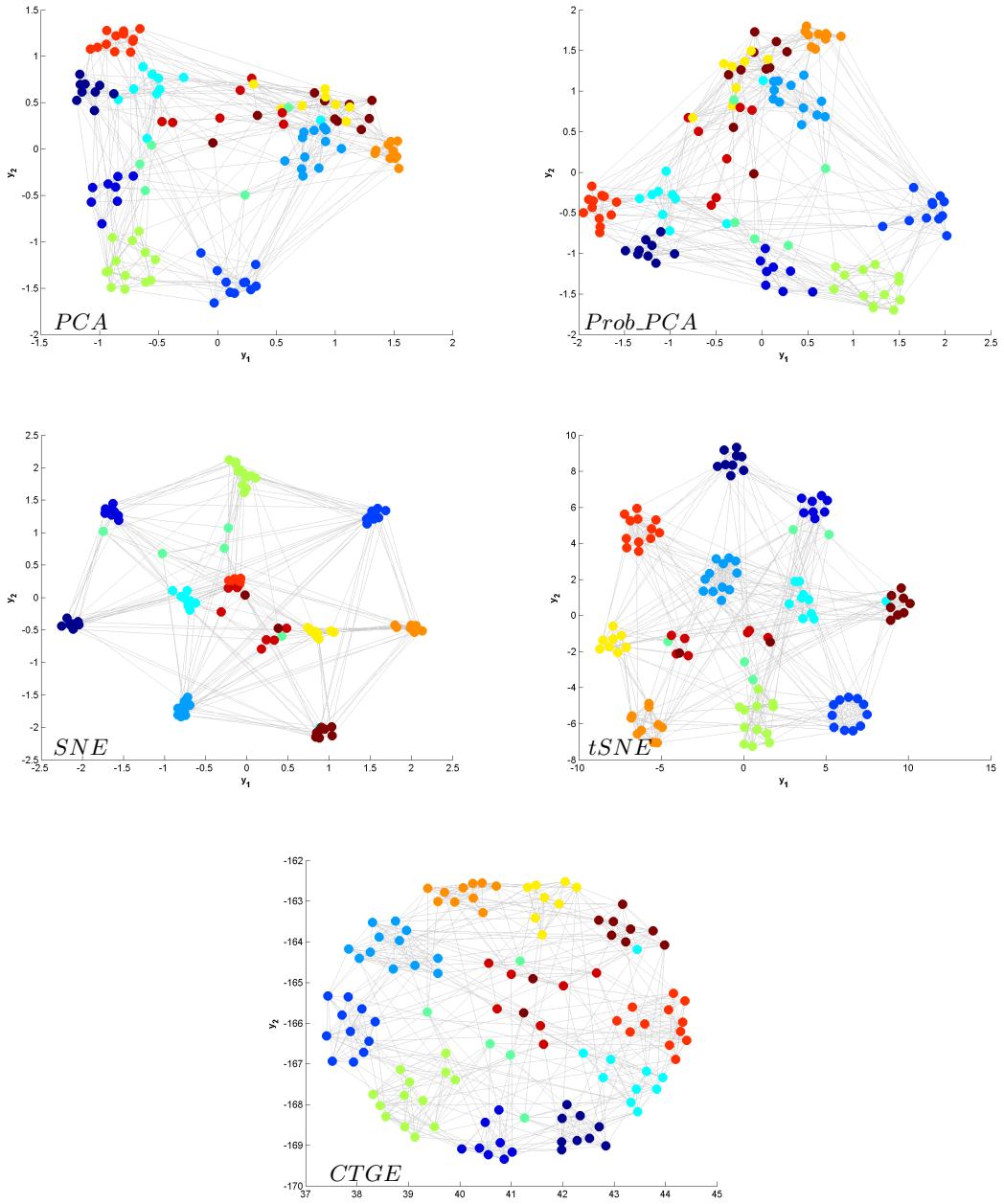

Figure 1: Low-dimensional representation of the American College Football graph data sets by PCA, Prob_PCA, SNE, tSNE and CTGE. The best representation is the one that preserves neighbourhood relations between points. The points which have the same colours are teams belonging to the same group.

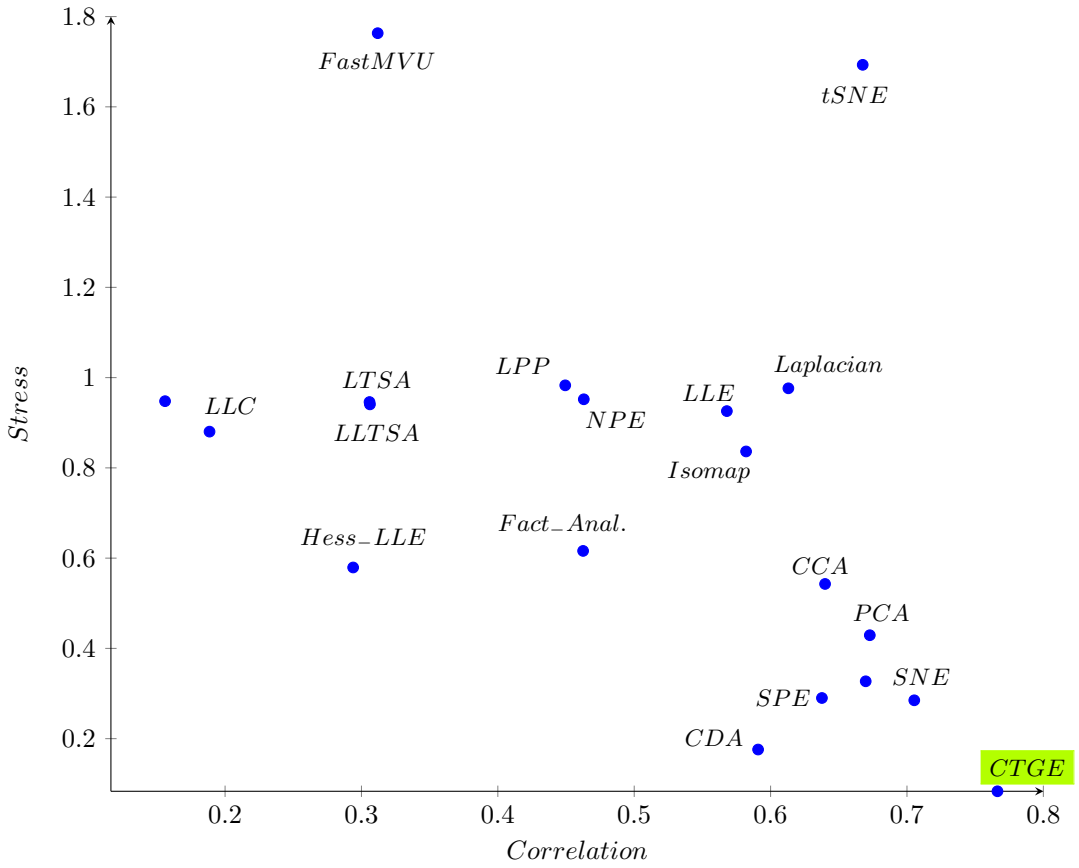

Figure 2: The results of the comparisons, in Table 1, of our method (CTGE) gets the highest correlation and the lowest stress values.

Table 2: Measuring the low-dimensional representation of the second graph data sets, which was carried out by 20 methods, by using correlation, LC and stress metrics. (Correlation and LC: the highest is better, Stress: the lowest is better)

Method	Correlation	LC	Stress
PCA	0.745	0.596	0.046
CCA	0.821	0.538	0.368
CDA	0.796	0.612	0.487
Factor_analysis	0.712	0.566	1.035
FastMVE	0.386	0.240	0.115
Hessian_LLE	0.107	0.103	0.117
Isomap	0.540	0.702	0.914
Kernel_PCA	0.048	0.108	0.117
Laplacian	0.503	0.716	0.117
LLC	0.268	0.178	0.100
LLE	0.347	0.521	0.118
LLTSA	0.590	0.393	0.118
LPP	0.414	0.708	0.117
LTSA	0.132	0.104	0.117
NPE	0.380	0.433	0.118
Prob_PCA	0.743	0.545	0.117
SPE	0.376	0.200	0.676
SNE	0.103	0.106	0.053
tSNE	0.635	**0.760**	0.112
CTGE	**0.841**	0.662	**0.040**

the low-dimensional space should allow the user to show these classes. Our method ($CTGE$) will be compared with 19 DR methods in the classification of this network.

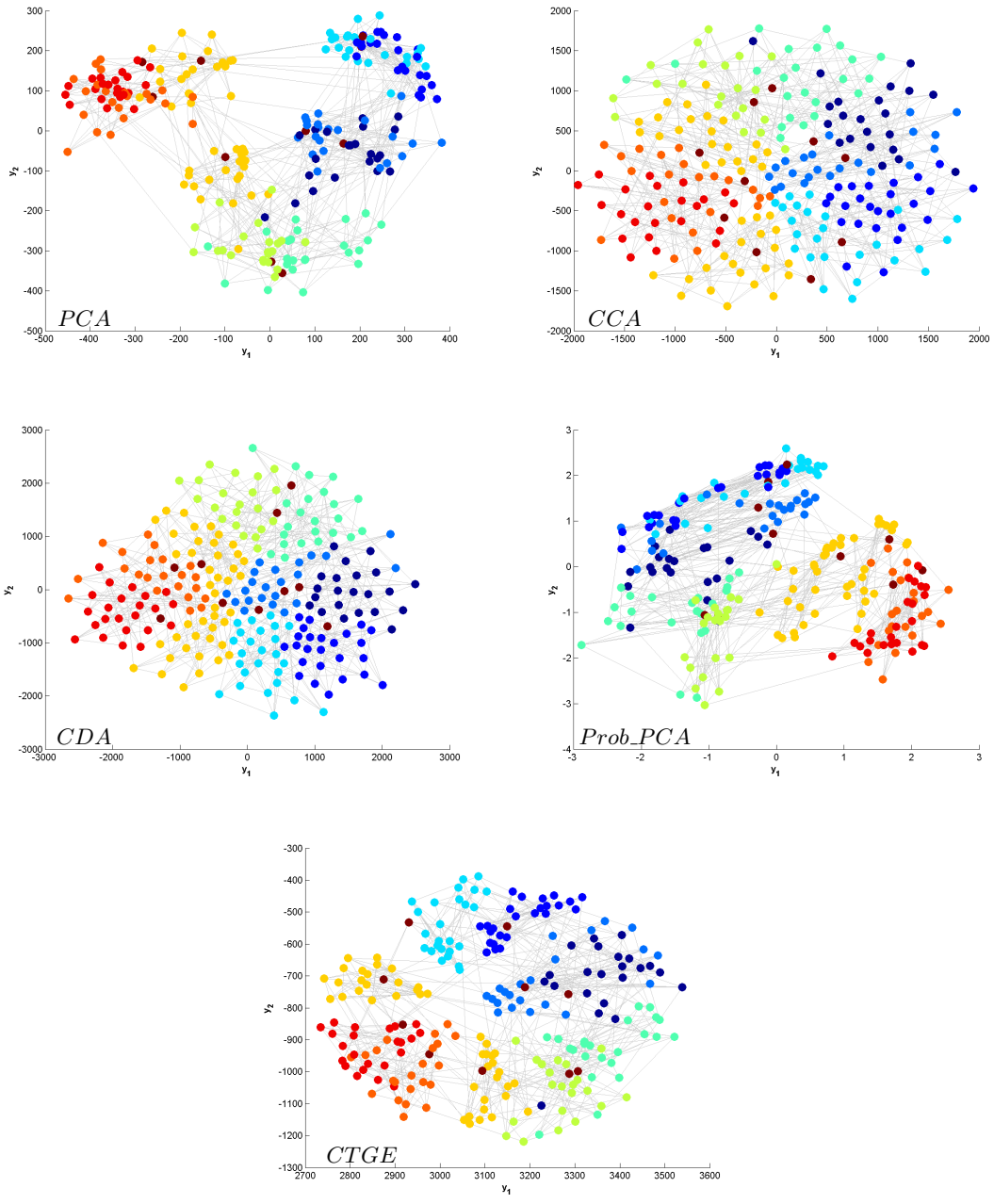

Figure 3: Low-dimensional representation of the second graph data sets by PCA, CCA, CDA, Prob_PCA and CTGE. The best representation is the one that preserves neighborhood relations between points. The points which have the same colors belong to the same class.

Figure 3 shows the five representations of the Primary-School graph data sets by using PCA, CCA, CDA, Prob_PCA and CTGE. We used three metrics to measure the efficiency of the results. These are correlation, LC and stress, and Table 2 shows the details. The results show that our method (CTGE), CCA and CDA, have higher correction values, with our methods the highest. By using LC metric, we find tSNE, Laplacian, LPP and Isomap have good accuracy, and the efficiency of our methods are inferior to them. Using the stress metric, the amount of error resulting from CFGE, PCA and SNE, are the less than the 20 DR methods.

8. Conclusion

In this paper, we applied our proposed methods, CTGE, to reduce the dimensionality of the graph data sets. The comparison of 20 dimensionality reduction methods would be based on the three measurement metrics: correlation, LC and stress.

The results showed the ability of our method to preserve neighborhood relationships, and to reveal much interesting information. The large dimensionality of the graph data sets were reduced into 2-dimensionality by our method, and they are better when compared with the results of other methods. The comparisons show the ability of the CTGE to unfold these complex data sets and, at the same time, to preserve most of the information of the original data sets.

References

[1] D. Harel, Y. Koren, Graph drawing by high-dimensional embedding, Graph Algorithms and Applications 8 (2004) 195–214.

[2] D. A. R. Bourqui, P. Mary, How to draw clustered weighted graphs using a multilevel force-directed graph drawing algorithm, in: In Proc. of the 11 Int. Conf. on Information Visualisation (IV'07), pages 757-764, Washington, USA., 2007.

[3] J. P. JOSEP D'IAZ, M. SERNA, A survey of graph layout problems, ACM Computing Surveys 34 (2002) 313–356.

[4] Y. Hu, Efficient, high-quality force-directed graph drawing, The Mathematica Journal 10 (2006) 37–71.

[5] P. Eades, M. Huang, Navigating clustered graphs using force-directed methods, Graph Algorithms and Applications 4 (2000) 183–210.

[6] M. G. P. Gajer, S. Kobourov, A multi-dimensional approach to forcedirected layouts of large graphs, Computational Geometry: Theory and Applications 29 (2004) 3–18.

[7] K. M. Tim Dwyer, M. Wybrow, Integrating edge routing into force-directed layout, Lecture Notes in Computer Science 4372 (2007) 8–19.

[8] Y. Koren, On spectral graph drawing, COCOON (2003) 496508.

[9] I. T. Jolliffe, Principal Component Analysis, New York: Springer-Verlag, 2002.

[10] D. Harel, Y. Koren, Graph drawing by high-dimensional embedding, GraphDrawing (2002) 207219.

[11] T. B. H. Albert D. Shieh, E. M. Airoldi, Tree preserving embedding, in: Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 2011.

[12] B. H. Junker, F. Schreiber (Eds.), Analysis of Biological Networks, wiley & Sons Inc., 2008.

[13] V. V. Onclinx, M.Verleysen, Nonlinear data projection on non-euclidean manifolds with controlled trade-off between trustworthiness and continuity, Neurocomputing 72 (2009) 14441454.

[14] D. K. Agrafiotis, Stochastic proximity embedding, Computational Chemistry 24.

[15] J. T. V. De Silva, Global versus local methods in nonlinear dimensionality reduction, Advances in Neural Information Processing Systems 15 (2003) 705–712.

[16] M. O. J. V. P. T. Samuel Kaski, Janne Nikkil, E. Castrn, Trustworthiness and metrics in visualizing similarity of gene expression, BMC Bioinformatics 4: 48 (2003) (2003) 4:48.

[17] J. Venna, S. Kaski, Comparison of visualization methods for an atlas of gene expression data sets, Information Visualization 6 (2007) 139–154.

[18] M. Mignotte, A bicriteria optimization approach based dimensionality reduction model for the color display of hyperspectral images, IEEE Transactions on Geoscience and Remote Sensing 50 (2012) 501–513.

[19] L. Chen, A. Buja, Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis, Journal of the American Statistical Association 104 (2009) 209–219.