2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

K-Associated Optimal Network for Graph Embedding Dimensionality Reduction

Murillo G. Carneiro, Thiago H. Cupertino and Liang Zhao

***Abstract*— In machine learning, dimensionality reduction aims at reducing the dimension of the input data in order to achieve a small set of features that keeps the most important original relationships among data samples. In this paper, we investigate the usage of a non-parametric network formation algorithm into a graph embedding framework to perform supervised dimensionality reduction. Specifically, our technique maps data into networks and constructs two network adjacency matrices which convey information about intra-class components and inter-class penalty connections. Both matrices are inserted into an optimization framework in order to achieve a projection vector that is used to project high-dimension data samples into a low-dimensional space. One advantage of the technique is that no parameter is required, that is, there is no need to select a model for the input data. Computer simulations on real-world data sets have been performed to compare the proposed technique to some classical network formation methods such as k-NN and ϵ-radius, and to well-known dimensionality reduction algorithms such as PCA and LDA. Statistical tests have shown that our approach outperforms those algorithms.**

I. Introduction

TRADITIONAL algorithms used in machine learning and pattern recognition applications are often susceptible to the well-known problem of the "curse of dimensionality" [1]. In this situation, similarity measures among data suffers from distortions, that is, when the dimensionality increases, the volume of the space increases so fast that the available data samples becomes sparse. As way to alleviate this problem, dimensionality reduction techniques [2], [3], [4], [5], which aims at reducing the dimension of the input data in order to achieve a small set of features that keeps the most important original relationships among data samples, are often applied as a data pre-processing step or as part of the data analysis to simplify the data model. The great advantages is that by working with a reduced representation, tasks such as classification or clustering can often yield more accurate and readily interpretable results, while computational costs may also be significantly reduced [6].

Murillo G. Carneiro is with the Institute of Mathematics and Computer Science, The University of São Paulo, São Carlos, Brazil. He is also with the Faculty of Computing, The Federal University of Uberlândia, Uberlândia, Brazil (email: carneiro@icmc.usp.br).

Thiago H. Cupertino is with the Institute of Mathematics and Computer Science, The University of São Paulo, São Carlos, Brazil (email: thiagohc@icmc.usp.br).

Liang Zhao is with the Department of Computing and Mathematics in the School of Philosophy, Sciences and Languages, The University of São Paulo, Ribeirão Preto, Brazil (email: zhao@usp.br).

The authors would like to acknowledge the Minas Gerais State Research Foundation (FAPEMIG), the São Paulo State Research Foundation (FAPESP) and the Brazilian National Council for Scientific and Technological Development (CNPq) for the financial support given to this research.

Techniques for dimensionality reduction often lie in the unsupervised or in the supervised learning. A classical example of unsupervised technique is the Principal Component Analysis (PCA) [7]. PCA is an orthogonal transformation that represent data by using the so called principal components. Usually, a small number of principal components is sufficient to account for most of the structure in the data. It maximizes the mutual information between the original high-dimensional Gaussian distributed measurements and the projected low-dimensional measurements. As an unsupervised technique, PCA does not use the class label information of the input data. In the supervised setting, data instances are marked with label information that guides the formation of the low-dimensional space. The labels often take discrete class values, indicating which data points have to be grouped together (same class) or set far apart from the other (different classes) in the embedded space. In the group of supervised techniques, Linear Discriminant Analysis (LDA) [8] plays an important role. As a supervised technique, it uses the class label information of the input data samples. LDA finds a projection matrix that maximizes the trace of the between-class scatter matrix and minimizes the trace of the within-class scatter matrix in the projected subspace simultaneously.

Supervised dimensionality reduction can also be performed by using a graph embedding framework [9]. Graphs are powerful tools to represent data relationships and have been applied to a variety of learning tasks [10], [11], [12], [13], [14], [15], [16]. The purpose of graph embedding is to represent each vertex (data sample) of a network as a low-dimensional vector that preserves similarities between the vertex pairs, where similarity is measured by a graph similarity matrix that characterizes certain statistical or geometric properties of the data set. The usage of graph embedding for dimensionality reduction can overcome some limitations of the LDA technique such as the number of available projection directions lower than the number of classes, and the assumption that data is approximately Gaussian distributed [9].

In this paper, we investigate the usage of the recently proposed K-Associated Optimal Graph (KAOG) [17] into the graph embedding framework for dimensionality reduction. A preliminary study essentially applied to image classification was conducted in [15] and the results were considered promising. The KAOG is a network construction technique which relies on two concepts: a purity measure, which uses the graph representation to measure mixing levels of the original data samples regarding their classes given a k-neighborhood; and the k-associated graph, which can be

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

considered as an improved adaptive k-Nearest Neighbor (k-NN) graph. The network construction process consists of building the k-associated optimal graph, that represents the data set as a sparse network in which components carry local information about the underlying data distribution. Furthermore, we propose a modification of the KAOG network formation to construct a penalty graph, which is required for the graph embedding framework. The penalty graph conveys information about which data samples (class components) should not be close together (different classes) in the reduced feature space.

Computer simulations were performed to evaluate the proposed technique. In addition, it was compared, in terms of predictive performance, to traditional dimensionality reduction techniques: PCA and LDA; and classical network construction methods: k-NN and ϵ-radius. Statistical tests have shown that the new approach outperforms these algorithms.

This paper is organized as follows. Section II introduces the problem setting of dimensionality reduction. Section III describes the network formation methods to construct the scatter-matrices to be used into the graph embedding framework. Section IV shows the experimental results and section V concludes the paper.

II. Dimensionality reduction problem setting

Given a training data set $\mathcal{X}^{(l)} = \{\mathbf{x}_i^{(l)},\ i = 1, \ldots, n\}$, containing labeled instances, and a test data set $\mathcal{X}^{(u)} = \{\mathbf{x}_i^{(u)},\ i = 1, \ldots, m\}$, containing unlabeled instances, each instance is described by q attributes, that is, a vector $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iq}]^T$, and belongs to a single class $c \in \{1, \ldots, C\}$, where C is the number of classes. The goal of the proposed technique is to perform dimensionality reduction by using the information provided by the labeled data set $\mathcal{X}^{(l)}$ in order to improve classification accuracy or, at least, to speed up the classification process of the unlabeled data set $\mathcal{X}^{(u)}$ without decreasing the accuracy, given that a small number q' of projected attributes is used ($q' < q$).

Usually, the feature dimension q can be very high, and transforming the data from the original high-dimensional space to a low-dimensional space can alleviate the curse of dimensionality [1]. To accomplish that, it is needed to find a mapping function F that transforms $\mathbf{x}$ into the desired low-dimensional representation $\mathbf{y}$, so that $\mathbf{y} = F(\mathbf{x})$ ($\mathbf{y} \in \mathbb{R}^{q'}$). By using an underlying network to find such function F, the dimensionality reduction process can be viewed as a graph-preserving criterion of the following form [9]:

$$Y^* = arg\,min \sum_{i \neq j} ||\mathbf{y}_i - \mathbf{y}_j||^2 W_{ij} = arg\,min\, Y^T L Y, \quad (1)$$

constrained to $Y^T B Y = \mathbf{d}$. In this formulation, $\mathbf{d}$ is a constant vector, W_{ij} is adjacency matrix of the network, B is the constraint matrix and L is the Laplacian matrix. The Laplacian matrix can be found via the following operation:

$$L = D - W,\ Dii = \sum_{i \neq j} W_{ij}, \forall i.$$

The constraint matrix B can be viewed as the adjacency matrix of a penalty network W^P, so that $B = L^P = D^P - W^P$. The penalty network conveys information about which vertices should not be linked together, that is, which instances should be far apart after the dimensionality reduction process. The similarity preservation property from the graph-preserving criterion has a two-fold explanation. For larger similarity between samples $\mathbf{x}_i$ and $\mathbf{x}_j$, the distance between $\mathbf{y}_i$ and $\mathbf{y}_j$ should be smaller to minimize the objective function. Likewise, smaller similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ should lead to larger distances between $\mathbf{y}_i$ and $\mathbf{y}_j$ for minimization [9].

In this paper, we assume that the low-dimensional attribute space can be found by using a linear projection such as $Y = X^T \mathbf{w}$, in which $\mathbf{w}$ is the projection vector. The objective function in Eq. 1 becomes:

$$\begin{aligned} \mathbf{w}^* = & \quad arg\,min \sum_{i \neq j} ||\mathbf{w}^T x_i - \mathbf{w}^T x_j||^2 W_{ij} \\ = & \quad arg\,min\, \mathbf{w}^T X L X^T \mathbf{w}, \end{aligned} \quad (2)$$

constrained to $\mathbf{w}^T X L X^T \mathbf{w} = d$. By using the Marginal Fisher Criterion [9] and the penalty network constraint, Eq. 2 becomes:

$$\mathbf{w}^* = arg\,min_{\mathbf{w}} \frac{\mathbf{w}^T X L X^T \mathbf{w}}{\mathbf{w}^T X L^P X^T \mathbf{w}}, \quad (3)$$

which can be solved by the generalized eigenvalue problem by using the equation $XLX^T\mathbf{w} = \lambda XL^PX^T\mathbf{w}$.

III. Network formation techniques

The construction of the underlying networks is an elementary step of the proposed dimensionality reduction technique. In this section, we provide the concepts related to the most relevant network formation techniques and propose the adaptations that we develop to employ them in a new dimensionality reduction technique.

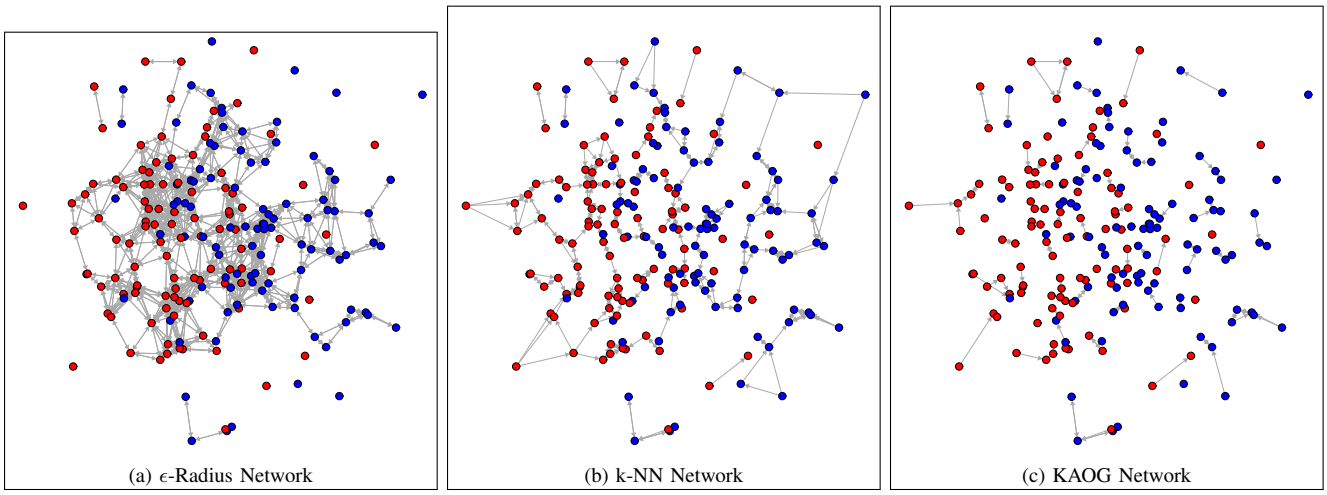
In a brief overview, there are few techniques related to network construction in the literature. The most used techniques are ϵ-radius and k-Nearest Neighbors (k-NN) [18]. However, both require parameter selection. On the other hand, a recently proposed technique, named k-Associated Optimal Graph (KAOG), provides a network that is constructed from a purity measure, without requiring the usage of parameter selection [17]. Fig. 1 shows a visual comparation among the networks obtained by, respectively, ϵ-radius, k-NN and KAOG from a data set composed of two mixed Gaussians. Next subsections introduces these network formation algorithms.

A. ϵ-Radius Network

In data classification, the ϵ-radius technique creates a link between two vertices i and j if two conditions are satisfied: i and j are within a distance ϵ and they belong to the same class:

$$E = E \cup \{e_{i,j} \mid d_{i,j} \leq \epsilon\ \&\ c_i = c_j\} \quad (4)$$

The ϵ-radius technique provides a network with higher density when compared to other graph formation techniques.



(a) ϵ-Radius Network (b) k-NN Network (c) KAOG Network

Fig. 1. Network formation algorithms applied to a data set composed of two mixed Gaussians.

An example of the ϵ-radius network is illustrated in the Fig. 1a. Note that there are a large number of links among the vertices.

As explained in section II, our technique requires the construction of two matrices: the adjacency matrix and the penalty matrix. The adjacency matrix (E) is obtained directly from (4). Alg. 1 presents a simple way to obtain the penalty matrix B. The algorithm creates a link between i and j in B if the vertices are within a distance ϵ and belong two different classes. In this case, a link means that these vertices should be far apart after the dimensionality reduction process.

Algorithm 1 ϵ-radius algorithm

Require: ϵ and a data set X
1: $E, B \Leftarrow \emptyset$
2: **for all** $i, j \in X$ **do**
3: **if** $d_{i,j} \leq \epsilon$ & $c_i = c_j$ **then**
4: $E \Leftarrow E \cup e_{i,j}$
5: **else if** $c_i \neq c_j$ **then**
6: $B \Leftarrow B \cup e_{i,j}$
7: **end if**
8: **end for**
9: **return** E and B

B. k-NN Network

The k-NN network construction creates a link between vertices i and j if two conditions are satisfied: j is one of the k-nearest neighbors of i and the classes of i and j are the same, as showed by 5:

$$E = E \cup \{e_{i,j} \mid j \in K\text{-NN}(i) \ \& \ c_i = c_j\}. \tag{5}$$

Unlike ϵ-radius network formation, k-NN is able to represent sparse regions of the network. Fig. 1b illustrates the application of the k-NN technique on a data set composed of two mixed Gaussians. Note that vertices in sparse regions, which do not link using ϵ-radius technique (Fig. 1a), are able to make connections using k-NN formation graph.

We propose a simple way to obtain the penalty matrix B for the k-NN network as follows. There is a link from i to j in B only if the j is one of the k-nearest neighbors of i and their classes are distinct. In consequence, the adjacency matrix E is obtained from (5). Alg. 2 presents the steps to obtain E and B.

Algorithm 2 k-NN algorithm

Require: K and a data set X
1: $E, B \Leftarrow \emptyset$
2: **for all** $i, j \in X$ **do**
3: **if** $j \in K$-NN(i) & $c_i = c_j$ **then**
4: $E \Leftarrow E \cup e_{i,j}$
5: **else if** $c_i \neq c_j$ **then**
6: $B \Leftarrow B \cup e_{i,j}$
7: **end if**
8: **end for**
9: **return** E and B

C. KAOG Network

Differently from the previous graph formation techniques, the KAOG technique constructs a network guided by a measure named purity. This measure expresses the level of mixture of a component in relation to other components of distinct classes and it is given by:

$$\Phi_\alpha = \frac{D_\alpha}{2K_\alpha}, \tag{6}$$

where D_α and K_α denote, respectively, the average degree and the K value associated to the component α. In this way, KAOG uses the purity measure to construct and optimize each component of the network.

Alg. 3 shows step by step the construction of KAOG networks. Note that no parameter is needed by the algorithm. After the initial setting, a loop starts to merge the subsequent k-associated graphs by increasing k, while improving the purity of the network encountered so far, until the optimal network measured by the purity degree is reached. Basically, the k-associated graph (KAG) algorithm links a vertex i to all its k nearest neighbors that belong to the same class of

i (a set denoted by $\Lambda_{i,K}$). More details about the algorithm are presented in [17].

Algorithm 3 K-associated optimal graph algorithm

Require: data set X
$K \Leftarrow 1$
$G^{(op)} \Leftarrow K\text{-}associated\ graph(K, X)$
repeat
  $lastAvgDegree \Leftarrow D^{(K)}$
  $K \Leftarrow K + 1$
  $G^{(K)} \Leftarrow K\text{-}associated\ graph(K, X)$
  for all $C_\beta^{(K)} \subset G^{(K)}$ **do**
    if $\Phi_\beta^{(K)} \geq \Phi_\alpha^{(op)}$ for all $C_\alpha^{(op)} \subseteq C_\beta^{(K)}$ **then**
      $G^{(op)} \Leftarrow G^{(op)} - \cup_{C_\alpha^{(op)} \subseteq C_\beta^{(K)}} C_\alpha^{(op)}$
      $G^{(op)} \Leftarrow G^{(op)} \cup \{C_\beta^{(K)}\}$
    end if
  end for
until $D^{(K)} - lastAvgDegree < D^{(K)}/K$
return $G^{(op)}$

Furthermore, we develop a fast way to obtain the penalty matrix B for the KAOG network as follows. There is a link between i and j in B if j is one of the k nearest neighbors of i, and j belong to a different class of i. Alg. 4 shows step by step how the links of the adjacency matrix E and the constraint matrix are done in the k-associated graph. It is worth noting that the constraint matrix is optimized by the purity measure too.

Algorithm 4 K-associated graph algorithm

Require: K and a data set X
$E, B \Leftarrow \emptyset$
for all $i \in V$ **do**
  if $j \in \Lambda_{i,K}$ & $c_i = c_j$ **then**
    $E \Leftarrow E \cup e_{i,j}$
  else if $c_i \neq c_j$ **then**
    $B \Leftarrow B \cup e_{i,j}$
  end if
end for
$C \Leftarrow findComponents(E)$
for all $\alpha \in C$ **do**
  $\Phi_\alpha \Leftarrow$ Eq. (6)
  $G^{(K)} \Leftarrow G^{(K)} \cup \{(\alpha(V', E', B'); \Phi_\alpha)\}$
end for
return K-associated graph $G^{(K)}$

Fig. 1c illustrates the construction of the KAOG network. The resulted network is distinct from the ϵ-radius and k-NN techniques. The main advantages on these algorithms is that KAOG network is obtained without any parameter. In addition, its vertices are linked according to the maximization of the purity measure. This provides an optimized network and a robust mechanism to avoid noisy and outliers [17].

IV. Experimental results

The proposed dimensionality reduction technique was evaluated by using the KAOG network formation method as described in Sec. III. Firstly, the proposed technique was compared to two well-known dimensionality reduction techniques, PCA[1] and LDA. Secondly, it was compared to other two classical network formation algorithms, k-NN and ϵ-radius. After the dimensionality reduction step, the projected data set was classified by using the nearest-neighbor classification rule. In the experiments, we used 5 high-dimensional data sets comprising data from diverse and different nature. These data sets can be found in the UCI machine learning repository [20], and Table I summarizes their corresponding meta-data. Following, we present a brief overview about each data set:

- **Sonar** The Sonar data set contains patterns obtained by bouncing sonar signals off a metal cylinder at various angles and under various conditions. The transmitted sonar signal is a frequency-modulated chirp, rising in frequency. Each pattern is a set of 60 numbers in the range 0.0 to 1.0. Each number represents the energy within a particular frequency band, integrated over a certain period of time;
- **Libras** The Libras data set contains movements from the visual language of hear-impaired people. From recorded videos, the movements were mapped in a representation with 90 features, with representing the coordinates of the movements;
- **Hill** In the Hill data set, each record represents 100 points on a two-dimensional graph. When plotted in order (from 1 through 100) as the y co-ordinate, the points create either a Hill or a Valley;
- **Musk1** The Musk1 data set describes a set of 92 molecules of which 47 are judged by human experts to be musks and the remaining 45 molecules are judged to be non-musks. To generate this data set, the low-energy conformations of the molecules were generated and then filtered to remove highly similar conformations, resulting in 476 conformations;
- **CNAE** The CNAE data set contains 1080 documents of free text business descriptions of Brazilian companies categorized into a subset of 9 categories cataloged in a table called National Classification of Economic Activities. Each document was represented as a vector, where the weight of each word is its frequency in the document.

TABLE I
META-DATA OF THE SIMULATED DATA SETS.

Data set	# Instances	Dimension
Sonar	208	60
Libras	360	90
Hill	606	100
Musk1	476	166
CNAE	1080	856

[1]According to the study presented in [19], PCA presented the best results on real-world data sets when compared with traditional nonlinear dimensionality reduction techniques, such as Isomap, Locally Linear Embedding (LLE) and others.

Each experiment was performed by using a 10-fold stratified cross-validation process [21]. In this process, the data set is split in 10 disjoint sets and, in each run, 9 sets are used as training data and 1 set is used as the test data, resulting in a total of 10 runs. The results are averaged over 30 runs, totaling $10 \times 30 = 300$ runs. Following, we show how the parameters were adjusted for each technique:

- For the principal component analysis algorithm (PCA), parameter $number\ of\ components$ was optimized according to the heuristic proposed in [22]. Note that PCA does not consider the class label information, but it is usually applied in supervised learning as part of the data analysis to simplify the data model.
- For the linear discriminant analysis algorithm (LDA), parameter $number\ of\ components$ was optimized in the interval $\{1, 2, \ldots, C-1\}$, where C is the number of classes in the data set;
- For the k-NN network formation technique, parameter k was optimized in the interval from 1 to the number of instances of the largest class in the training data set;
- For the ϵ-radius network formation method, parameter ϵ was optimized in the interval $\{5\%, 10\%, \ldots, 100\%\}$, concerning the average distance among instances in the training data set;
- For the KAOG method, remember that selection of parameters is unnecessary.

Table II shows the results of classification accuracy after dimensionality reduction by using the three different network formation methods and two traditional dimensionality reduction techniques. In the last column of this table it can be seen the classification accuracy without any dimensionality reduction process for comparison purposes. In order to analyze statistically the results, we adopted a statistical test that compares multiple classifier over multiple data sets [23]. Firstly, Friedman test is calculated to check whether the performance of the classifiers are significantly different. Using a significance level of 5%, the null hypothesis is rejected. This means that the algorithms under study are not equivalent. Following a post-hoc test, Nemenyi test is employed (also considering a significance level of 5%). The results of this test is showed in Table III and it indicates that KAOG outperforms all other techniques, including the traditional PCA and LDA algorithms. In addition to the good predictive results obtained by our approach, the use of the dimensionality reduction method present other advantages when compared to the original dimension of large data set, such as: (i) the reduced number of features which implies in decreasing the training and/or classification time; and (ii) the robustness to noisy and irrelevant features that can impact negatively on accuracy. A visual comparison of the proposed technique results to the original number of attributes can be seen in Fig. 2.

Now, we move on to an interesting analysis about k-NN, ϵ-radius and KAOG network formation techniques which were adapted into a graph embedding framework to perform supervised dimensionality reduction in this paper. Table IV

TABLE III

RESULTS OF THE FRIEDMAN/NEMENYI STATISTICAL TEST. "BETTER/WORSE" INDICATES THAT THE ALGORITHM CORRESPONDING TO ITS COLUMN IS BETTER/WORSE THAN THE ALGORITHM CORRESPONDING TO ITS ROW (A REJECTION OF THE NULL HYPOTHESIS). "NO" SUGGESTS THAT BOTH THE COLUMN ALGORITHM AND THE ROW ALGORITHM PERFORM EQUALLY WELL (A FAILURE TO REJECT THE NULL HYPOTHESIS AT THE 5% SIGNIFICANCE LEVEL).

	PCA	**LDA**	**KAOG**	**k-NN**	**ϵ-radius**
LDA	No	-	-	-	-
KAOG	Worse	Worse	-	-	-
k-NN	Better	Better	Better	-	-
ϵ-radius	Worse	Worse	Better	Worse	-
Orig. dimen.	Worse	Worse	Better	Worse	No

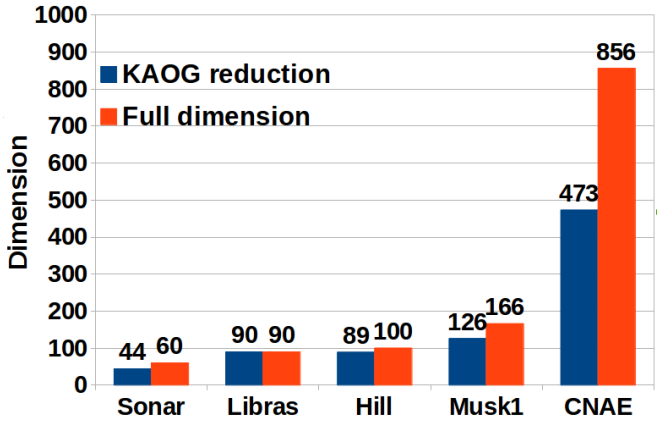


Fig. 2. Results of the proposed dimensionality reduction technique for the 5 data sets for comparison to the original attribute space size. It can be seen that for the largest data set (CNAE) the reduction is relevant, almost half the original space.

shows the number of attributes after dimensionality reduction for the results in Table II. The proposed technique achieved dimensionality reductions up to 55.26% (CNAE data set) of the number of the original feature space. The other network formation methods also achieved good dimensionality reduction rates, but with a smaller classification accuracy (see Table II).

For the sake of completeness, Fig. 3 shows the classification accuracy after dimensionality reduction performed by the proposed technique by using from 1 to the number of attributes of the original feature space. It can be seen that the highest accuracy is achieved by using a small number of projected attributes, specially for data sets Sonar (44 attributes) and CNAE (473 attributes).

V. CONCLUSION

We have studied the usage of a modified version of the recently proposed K-Associated Optimal Graph (KAOG) to perform supervised dimensionality reduction. The proposed technique derives two adjacency matrices which represent the intra-class and the inter-class information of the input data. Both matrices are used into a graph embedding framework which is optimized in terms of a projection vector.

TABLE II

CLASSIFICATION ACCURACY (%) BY USING THE REDUCED PROJECTED ATTRIBUTE SPACE AFTER DIMENSIONALITY REDUCTION PERFORMED BY 3 UNDERLYING NETWORKS: KAOG, K-NN AND ϵ-RADIUS. THE ACCURACY BY USING THE ORIGINAL DIMENSION IS SHOWED FOR COMPARISON PURPOSES. THE BEST RESULTS ARE IN BOLDFACE.

Data set	PCA	LDA	KAOG	k-NN	ϵ-radius	Orig. dimen.
Sonar	83.16 ± 7.92	71.39 ± 9.07	**84.90 ± 20.90**	79.01 ± 24.44	84.65 ± 9.57	83.30 ± 10.60
Libras	84.86 ± 5.27	67.61 ± 7.31	**85.06 ± 9.26**	73.93 ± 17.94	84.89 ± 7.24	84.89 ± 5.76
Hill	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.15**	**100.00 ± 0.00**	**100.00 ± 0.15**	**100.00 ± 0.00**
Musk1	85.92 ± 4.00	80.32 ± 5.05	**86.23 ± 11.25**	80.21 ± 13.57	85.93 ± 8.66	85.93 ± 8.17
CNAE	84.75 ± 3.40	**87.32 ± 10.59**	87.30 ± 9.08	83.63 ± 10.02	86.19 ± 5.34	86.24 ± 5.61

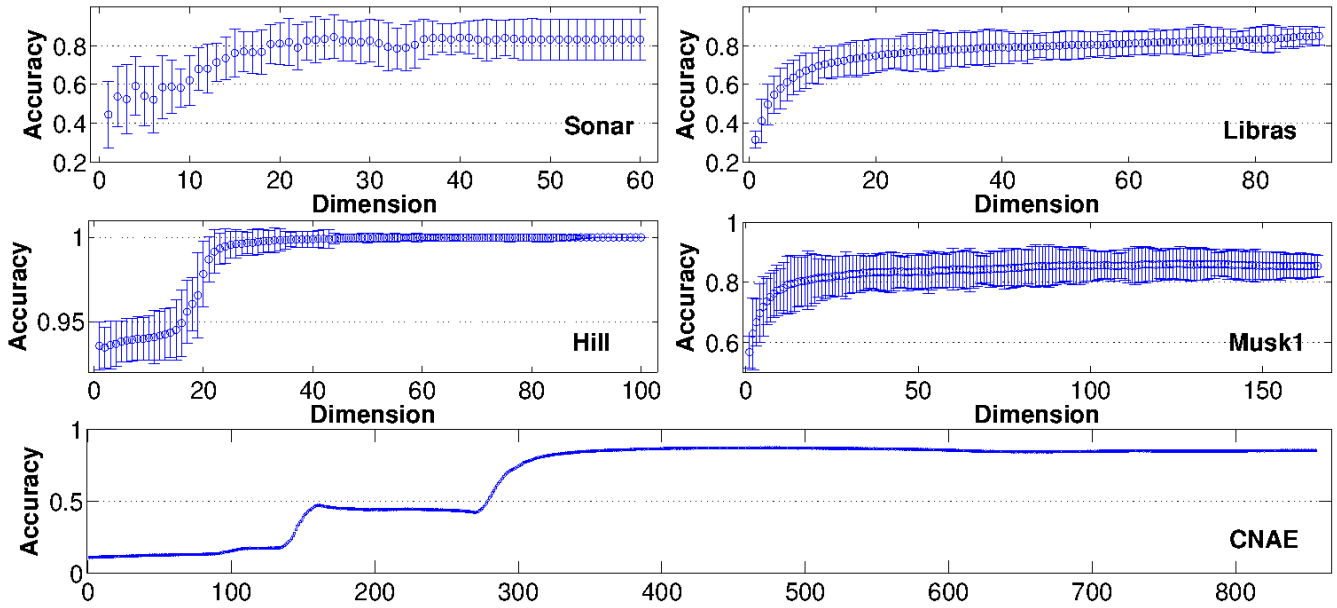


Fig. 3. Classification accuracy by using different numbers of projected dimensions. It can be seen that the highest accuracy is achieved by using a small number of projected attributes, specially for data sets Sonar (44 attributes) and CNAE (473 attributes).

TABLE IV

PROJECTED LOW-DIMENSION USED FOR CLASSIFICATION AFTER DIMENSIONALITY REDUCTION FOR THE RESULTS IN TABLE II. THE PERCENTAGES OF THE NUMBER OF PROJECTED ATTRIBUTES COMPARED TO THE ORIGINAL FEATURE SPACE ARE IN PARENTHESIS.

Data set	KAOG (%)	k-NN (%)	ϵ-radius (%)
Sonar	44 (73.33)	50 (83.33)	38 (63.33)
Libras	90 (100.00)	86 (95.56)	90 (100.00)
Hill	89 (89.00)	24 (24.00)	89 (89.00)
Musk1	126 (75.90)	165 (99.40)	126 (75.90)
CNAE	473 (55.26)	729 (85.16)	473 (55.26)

Experimental studies and statistical tests have showed that the proposed technique achieves competitive dimensionality reduction and better predictive results compared to some other traditional techniques and network formation methods. As future studies we suggest computer simulations on more data sets, comparative analysis with other algorithms for dimension reduction, and new experimental studies including other classification rules.

REFERENCES

[1] R. O. Duda, D. G. Stork, and P. E. Hart, *Pattern classification*, 2nd ed. Wiley-Interscience, 2000.

[2] H. Cai, K. Mikolajczyk, and J. Matas, "Learning linear discriminant projections for dimensionality reduction of image descriptors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 2, pp. 338–352, 2011.

[3] S. Chen and D. Zhang, "Semisupervised dimensionality reduction with pairwise constraints for hyperspectral image classification," *Geoscience and Remote Sensing Letters, IEEE*, vol. 8, no. 2, pp. 369–373, 2011.

[4] B. Raducanu and F. Dornaika, "A supervised non-linear dimensionality reduction approach for manifold learning," *Pattern Recognition*, vol. 45, no. 6, pp. 2432 – 2444, 2012.

[5] B.-D. Liu, Y.-X. Wang, Y.-J. Zhang, and B. Shen, "Learning dictionary on manifolds for image classification," *Pattern Recognition*, vol. 46, no. 7, pp. 1879 – 1890, 2013.

[6] P. Cunningham, "Dimension reduction," in *Machine Learning Techniques for Multimedia*, ser. Cognitive Technologies, M. Cord and P. Cunningham, Eds. Springer Berlin Heidelberg, 2008, pp. 91–112.

[7] I. T. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.

[8] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Academic Press, 1991.
[9] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 1, pp. 40–51, 2007.
[10] L. Zhao, T. H. Cupertino, and J. R. B. Jr., "Chaotic synchronization in general network topology for scene segmentation," *Neurocomputing*, vol. 71, no. 16-18, pp. 3360–3366, 2008.
[11] T. H. Cupertino, T. C. Silva, and L. Zhao, "Classification of multiple observation sets via network modularity," *Neural Computing and Applications*, pp. 1–7, 2012.
[12] T. Silva, L. Zhao, and T. H. Cupertino, "Handwritten data clustering using agents competition in networks," *Journal of Mathematical Imaging and Vision*, pp. 1–13, 2012.
[13] T. H. Cupertino, J. Huertas, and L. Zhao, "Data clustering using controlled consensus in complex networks," *Neurocomputing*, vol. 118, pp. 132–140, 2013.
[14] M. G. Carneiro and L. Zhao, "High level classification totally based on complex networks," in *Proceedings of the 1st BRICS Countries Congress*, 2013, pp. 1–8.
[15] T. H. Cupertino, M. G. Carneiro, and L. Zhao, "Dimensionality reduction with the k-associated optimal graph applied to image classification," in *Proceedings of the IEEE International Conference on Imaging Systems and Techniques*, 2013, pp. 1–6.
[16] M. G. Carneiro, J. L. Rosa, A. A. Lopes, and L. Zhao, "Network-based data classification: Combining k-associated optimal graphs and high level prediction," *Journal of the Brazilian Computer Society*, pp. 1–12, 2014.
[17] J. R. Bertini, L. Zhao, R. Motta, and A. de Andrade Lopes, "A nonparametric classification method based on k-associated graphs," *Information Sciences*, vol. 181, pp. 5435–5456, 2011.
[18] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.
[19] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality reduction: A comparative review," *Technical Report TiCC TR 2009-005*, 2009.
[20] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml
[21] J.-H. Kim, "Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap," *Computational Statistics and Data Analysis*, vol. 53, pp. 3735–3745, 2009.
[22] T. P. Minka, "Automatic choice of dimensionality for PCA," in *Advances in Neural Information Processing Systems*, 2000, pp. 598–604.
[23] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.