International Journal of Machine Learning and Cybernetics (2019) 10:397–411
https://doi.org/10.1007/s13042-018-0784-y

ORIGINAL ARTICLE

Semi-supervised rough fuzzy Laplacian Eigenmaps for dimensionality reduction

Minghua Ma[1] · Tingquan Deng[1] · Ning Wang[1,3] · Yanmei Chen[2]

Received: 28 May 2017 / Accepted: 9 January 2018 / Published online: 22 January 2018

Abstract
Laplacian Eigenmaps is a popular nonlinear dimensionality reduction technique and there exist various scenarios of its extensions. In this paper, a semi-supervised rough fuzzy Laplacian Eigenmaps (SSRFLE) approach is developed for dimensionality reduction of high dimensional hybrid data. In the proposed method, a set of semi-supervised fuzzy similarity granules are constructed to characterize the similarity between samples according to the principle that homogeneous samples have higher similarity degrees than heterogeneous samples. A neighborhood rough fuzzy set model of such fuzzy similarity granules is built to assess the degrees two samples belong to the same class. A Laplacian nearest neighborhood graph and a class-related neighborhood graph are constructed to characterize the topological structure between samples and between each sample and its prototype to ensure homogeneous samples being mapped closer to and more compact around the prototypes in a lower dimensional space. In view of the fact that different features bring out distinct impacts on performances of feature extraction and clustering, the significance of each feature is assessed by designing an information entropy measure and the weighted distance between samples is incorporated into the proposed technique. A series of simulation experiments on real world hybrid datasets are carried out. Experimental results show superior performance of the proposed method in classification accuracy and data visualization compared with other state of the art semi-supervised methods.

Keywords Laplacian Eigenmaps · Dimensionality reduction · Information entropy · Significance of feature · Neighborhood rough fuzzy sets

1 Introduction

Due to rapid emergence of high dimensional data in recent years, more and more scholars pay close attention to dimensionality reduction techniques, an important research issue in machine learning community. Dimensionality reduction aims to reduce redundant or irrelevant features, and extract salient characteristics in order to compress data, decrease computing complexity, and improve efficiency and accuracy of data classification and recognition [14, 37, 38, 43].

Feature extraction is a distinguished way to dimensionality reduction that maps a high dimensional dataset into a lower dimensional space where the essential characterization of original data is preserved as much as possible. It can be divided into two categories, the linear and nonlinear. Principal component analysis (PCA) [3], independent component analysis (ICA) [32], linear discriminant analysis (LDA) [24, 46], and local preserving projection (LPP) [13] are typical linear dimensionality reduction techniques, developed under different optimization criteria. In practice, almost all datasets do not have linear structures and linear techniques cannot handle nonlinear data. Various nonlinear dimensionality reduction techniques have been developed and falls into two types, based on kernel function methods [35] and based on manifold learning methods [21]. Nonlinear dimensionality reduction techniques based on kernel function methods have a difficulty of choosing suitable kernel functions. An appropriate kernel function can make data be linearly separable or approximately linearly separable in a lower dimensional space, but it is not applicable to every dataset. Nonlinear

✉ Tingquan Deng
Deng.tq@hrbeu.edu.cn

[1] College of Science, Harbin Engineering University, Harbin 150001, People's Republic of China

[2] Department of Mathematics, Harbin Institute of Technology, Harbin 150001, People's Republic of China

[3] Chengdu Aircraft Industrial (Group) Co., Ltd, Chengdu 610073, People's Republic of China

dimensionality reduction techniques based on manifold learning have been extensively investigated in recent years. ISOMAP [37], local tangent space analysis (LTSA) [52], local linear embedding (LLE) [31], Laplacian Eigenmaps (LE) [4], and their generalizations [2, 6, 19] have been successfully achieved.

Among the manifold learning based techniques, Laplacian Eigenmaps is a frequently used nonlinear dimensionality reduction method due to its superior property of preserving local neighborhood structure of data. However, it has lots of deficiencies, including sensitivity to noise, difficulty in choosing size of neighborhood, and no capability of preserving class structures of data. Many scholars focus on developing its improved scenarios. Raducanu [29] and Keyhanian et al. [17] extended, separatively, LE by constructing adaptive neighborhood graphs to avoid the puzzle of choosing the parameter of size of neighborhood. Wang et al. [40] proposed a distinguishing variance embedding (DVE) method by introducing the idea of minimum variance unbiased (MVU) to the classical LE. Liu et al. [25] proposed a local linear Laplacian Eigenmaps by combining LE with LLE. Malik et al. [26] explored a generalized incremental LE that can be applied to dynamic data.

Meanwhile, several researchers introduced supervised information of datasets into Laplacian Eigenmaps to improve the performances of dimensionality reduction [5]. Jiang et al. [16] and Li [20] introduced class information of datasets into LE and developed a supervised LE algorithm applied to fault diagnosis and face recognition. Xu et al. [48] combined the idea of LDA with LE in the framework of marginal patch alignment. Such supervised LE techniques can not only preserve local neighborhood structures of samples, also strengthen class structures of datasets. In the mean while, Costa et al. [8] proposed a classification constrained dimensionality reduction (CCDR) that ensures samples tending to collapse into the prototypes. In that method, two neighborhood graphs, a k-nearest neighborhood graph and a sample-class neighborhood graph, were constructed. The weights between vertices (samples) in the first graph were assessed according to distances between samples, while the weights in the second were set to be 1 or 0, depending on whether a sample had a class label or not. In virtue of whether samples are labelled or not, Kim et al. [18] proposed a semi-supervised Laplacian Eigenmaps (SSLE) by constructing two neighborhood graphs. The weights were assigned to 1, 0.5, or 0, depending on whether the corresponding vertices are labelled and whether one sample is in the neighborhood of the other. SSLE can make homogeneous samples pull each other and heterogeneous samples push each other. The performance of dimensionality reduction can be strengthened.

In both CCDR and SSLE algorithms, fixed weights in the sample-class neighborhood graph are assigned to unlabelled samples, ignoring the membership degrees of them belonging to each class and cannot exactly express the topological structures of data. Wang et al. [41] developed a T-S norm neural network to train weights for fuzzy if-then rules, where the T-S norms are fundamental ingredients in the theory of fuzzy sets. Fuzzy set is a generalization of classical set for modelling imprecise and vague information [50]. A fuzzy similarity relation is a typical notion describing association of objects. The derived fuzzy similarity classes are fuzzy information granules characterizing topological structures of data [47]. Another notion, rough sets, was initiated by Pawlak [27] for modelling and processing incomplete information. It has been found extensive and successful applications in the field of artificial intelligence. Various extensions of the Pawlak rough set model have been exposed, such as a general relation based rough set [36], a dominance relation based rough set [34], a similarity or tolerance relation based rough set [15, 33], covering rough set [53], a neighborhood rough set [44, 49], and a decision-theoretic rough set [22, 45]. The integration of both granular computing frameworks brings out the models of fuzzy rough sets, rough fuzzy sets, and various variations [1, 9, 10, 30, 39, 51] to solve the problems with imprecise and incomplete information.

In this paper, the notion of granular computing is introduced into LE and a semi-supervised LE for dimensionality reduction is developed. In this method, a set of semi-supervised fuzzy similarity granules are constructed to characterize the similarity between samples according to the principle that homogeneous samples have higher similarity degrees than heterogeneous samples. A neighborhood rough fuzzy set model of such fuzzy similarity granules is built to assess the degrees two samples belong to the same class. A class-related neighborhood graph of dataset is created based on the semi-supervised fuzzy similarity granules for classes to describe the relationship between samples and their prototypes, whereas a Laplacian k-nearest neighborhood graph is established according to both the semi-supervised information for classes and the association degrees of samples derived from the neighborhood rough fuzzy lower approximations. Simultaneously, the feature significance is assessed by building an information entropy measure and the weighted distance is incorporated into the establishment of Laplacian neighborhood graph. The proposed semi-supervised rough fuzzy based Laplacian Eignmaps (SSRFLE) model for dimensionality reduction of hybrid data not only inherits the advantages of classical LE, also preserves class characterization of the original dataset.

The rest of this paper is organized as follows. In Sect. 2, the classical Laplacian Eigenmaps model and its semi-supervised extensions for dimensionality reduction are recalled. Some basic notions related to fuzzy sets, rough sets and their integrations are briefly reviewed. Section 3 presents an approach to determining the weights of features of a dataset.

A semi-supervised Laplacian Eigenmaps model based on a neighborhood rough fuzzy sets is exposed. In Sect. 4, various comparative experiments on real world datasets are implemented. Parameters and performance analysis on the proposed method are presented sequentially. Conclusions and further work follow in Sect. 5.

2 Related work

In this section, we first recall the classical Laplacian Eigenmaps method and its two typical semi-supervised versions for dimensionality reduction. And then, some basic notions related to fuzzy sets, rough sets and their integrations are briefly reviewed. These notions will be used in the sequent sections of this work.

2.1 Laplacian Eigenmaps (LE)

Laplacian Eigenmaps [4] is a typical nonlinear dimensionality reduction technique based on spectral graph theory. It has remarkable properties of preserving local neighborhood structure of data. A k-nearest neighborhood graph or an ε-ball neighborhood graph is constructed and weights of edges (between vertices) are assigned using the Gaussian kernel function or 0-1 weighting method.

Given a dataset $X = \{x_1, x_2, \dots, x_n\}$ with n samples. Each sample $x_i \in X$ has m features, namely, $A = \{a_1, a_2, \dots, a_m\}$. Let $\{y_1, y_2, \dots, y_n\}$ be the $d\,(d \ll m)$ dimensional representations of X. That is, each y_i is a d dimensional row vector. With LE, the lower dimensional representation of X can be achieved by solving the following optimization problem

$$\min \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| y_i - y_j \right\|^2 W_{ij} = 2tr(Y^T L Y) \tag{2.1}$$

where $Y = (y_1^T\, y_2^T\, \dots\, y_n^T)^T$ is the $n \times d$ embedded matrix of X, $W = (W_{ij})_{n \times n}$ is the weight matrix of the k-nearest neighborhood graph, $D = (D_{ij})_{n \times n}$ is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n} W_{ij}$, and $L = D - W$ is the Laplacian matrix, a symmetric and positive semi-definite matrix. In order to ensure the problem (2.1) having a unique solution, the constraints $Y^T D Y = I$ and $Y^T D \mathbf{1} = 0$ are imposed to remove arbitrary scaling factor and translational degree of freedom in the lower dimensional embedding, where I is the $n \times n$ identity matrix and $\mathbf{1}$ is a column vector with all components being 1. By the Lagrange multiplier method, the optimization problem (2.1) can be translated to solve the following generalized eigenvalue problem

$$LY = \lambda DY \tag{2.2}$$

If $0 \neq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$ are the d smallest positive eigenvalues of Eq. (2.2) and the column vectors $y^1, y^2, \dots, y^d$ are the corresponding d eigenvectors, then the lower dimensional embedding of x_i is as follows

$$y_i = (y^1(i), y^2(i), \dots, y^d(i)), i = 1, 2, \dots, n$$

meaning that the ith row vector of Y is right the d dimensional embedding of x_i.

2.2 Classification constrained dimensionality reduction (CCDR)

CCDR [8], proposed by Costra et al. in 2005, is an extension of classical LE by fusing class information. It can make samples with the same class label collapse into corresponding prototypes.

Suppose each sample of a dataset X (or a subset of X) is labelled, namely x_i has a class label $l_i \in \{1, 2, \dots, f\}$, where f is the number of classes of X. Two neighborhood graphs G^N and G^C were constructed, where G^N is a k-nearest neighborhood graph and the weights of edges are computed by using the Gaussian kernel function. G^C is a graph regarding the class information of X, called the class-related neighborhood graph. Inserting edges between prototypes and samples with the same class label and the weights of such edges were set to be 1. Herein, the prototypes of X were computed by the way of maximum alignment between samples and classes.

In order to achieve the goal that samples with the same label were clustered together around the prototypes and simultaneously the neighborhood structures of X were preserved, a cost function was constructed as follows.

$$E = \beta \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} \left\| y_i - y_j \right\|^2 + \sum_{i=1}^{n} \sum_{k=1}^{f} C_{ki} \left\| y_i - z_k \right\|^2 \tag{2.3}$$

where $W = (W_{ij})_{n \times n}$ is the weight matrix of G^N, $C = (C_{ki})_{f \times n}$ is the weight matrix of G^C, $\{y_1, y_2, \dots, y_n\}$ is the lower dimensional representation of X, $z_1, z_2, \dots, z_f$ are the prototypes of X in the lower dimensional space, and β is a parameter trading off the influences between preserving local neighborhood structures and keeping class structures.

Let $Z = (z_1^T\, \dots\, z_f^T\, y_1^T\, \dots\, y_n^T)^T$, then minimizing Eq. (2.3) is equivalent to solving the following optimization problem

$$\min_{Z^T \mathbb{D} Z = I, Z^T \mathbb{D} 1 = 0} tr(Z^T \mathbb{L} Z) \tag{2.4}$$

where $\mathbb{L} = \mathbb{D} - \mathbb{W}$ is a $(f + n) \times (f + n)$ Laplacian matrix associated with the weight matrix $\mathbb{W} = (\mathbb{W}_{ij})_{(f+n)\times(f+n)} = \begin{pmatrix} I & C \\ C^T & 2\beta W \end{pmatrix}$ and $\mathbb{D} = (\mathbb{D}_{ij})_{(f+n)\times(f+n)}$ with $\mathbb{D}_{ii} = \sum_{j=1}^{f+n} \mathbb{W}_{ij}$, $\mathbb{D}_{ij} = 0$ when $i \neq j$. By the Lagrange multiplier method, the problem (2.4) can be transformed to solve the following generalized eigenvalue problem

$$\mathbb{L}Z = \lambda \mathbb{D}Z \tag{2.5}$$

If $z^1, z^2, \dots, z^d$ are the eigenvectors corresponding to the d smallest positive eigenvalues of Eq. (2.5), then the first f rows of $Z = [z^1\ z^2\ \dots\ z^d]$ correspond to the coordinates of prototypes and the following n rows determine the embedding of the original samples.

2.3 Semi-supervised Laplacian Eigenmaps (SSLE)

Kim et al. proposed a semi-supervised Laplacian Eigenmaps algorithm, called SSLE, which is suitable for sentiments analysis [18]. Execution of SSLE algorithm needs two premises. One is that the labelled information about similarity among samples and the other is that the similarity between homogeneous samples should be larger than that between heterogenous samples.

For a dataset X, let $X_c = \{x_1, x_2, \dots, x_s\}$ be a subset of X, in which each sample is labelled as a cluster among f clusters, namely $x_i \in X_c$ is assigned a label $l_i \in \{1, 2, \dots, f\}$. Two k-nearest neighborhood graphs, G_u and G_l, were built, where G_u was a k-nearest neighborhood graph without label information and the weights W^u_{ij} of edges were computed by using the Gaussian kernel function, while G_l was a k-nearest neighborhood graph with label information and the weights of edges were assessed as

$$W^l_{ij} = \begin{cases} 1, & \text{if } x_i \in N_{l+}(x_j) \text{ or } x_j \in N_{l+}(x_i) \\ 0, & \text{if } x_i \in N_{l-}(x_j) \text{ or } x_j \in N_{l-}(x_i) \\ 0.5, & \text{if } x_i \in N_{l0}(x_j) \text{ or } x_j \in N_{l0}(x_i) \\ 0, & \text{otherwise} \end{cases}$$

where $N_{l+}(x_j)$ and $N_{l-}(x_j)$ were the homogeneous neighborhood set and heterogeneous neighborhood set of a labelled sample x_j, respectively. $N_{l0}(x_j)$ was the neighborhood set of an unlabelled sample x_j. In SSLE, the objective function was defined as follows

$$\Phi(Y) = (1-\mu)tr(Y^T L^u Y) + \mu tr(Y^T L^l Y) = tr(Y^T((1-\mu)L^u + \mu L^l)Y) \tag{2.6}$$

where $\{y_1, y_2, \dots, y_n\}$ is the lower dimensional embedding of X and $Y = (y_1^T\ y_2^T\ \dots\ y_n^T)^T$, $L^u = D^u - W^u$ and $L^l = D^l - W^l$ are the Laplacian matrices of G_u and G_l, respectively, and D^u and D^l are two diagonal matrices with $D^u_{ii} = \sum_{j=1}^n W^u_{ij}$ and $D^l_{ii} = \sum_{j=1}^n W^l_{ij}$.

Let $W = (1-\mu)W^u + \mu W^l$, $D = (1-\mu)D^u + \mu D^l$, and $L = (1-\mu)L^u + \mu L^l$, then $L = D - W$ and $D_{ii} = \sum_{j=1}^n W_{ij}$. Meanwhile, the same constraint conditions $Y^T DY = I$ and $Y^T D\mathbf{1} = 0$ were imposed. The lower dimensional embedding of X can be obtained by solving the generalized eigenvalue problem $LY = \lambda DY$.

If $0 \neq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_d$ are the d smallest eigenvalues of Equation $LY = \lambda DY$ and $y^1, y^2, \dots, y^d$ are corresponding column eigenvectors, then the lower dimensional embedding of sample x_i is $y_i = (y^1(i), y^2(i), \dots, y^d(i)), i = 1, 2, \dots, n$.

2.4 Fundamentals of fuzzy sets and rough sets

Both fuzzy sets and rough sets are fundamental components of granular computing theory for uncertain information analysis and processing in the fields of decision analysis and artificial intelligence. In this subsection, we review some basic notions related to the granular computing framework that are indispensable in our proposal.

Let X be a universe of discourse (the dataset aforementioned), a map μ_F from X to [0, 1] models a fuzzy concept F on X [50]. For any $x \in X$, $\mu_F(x)$, or $F(x)$ briefly, denotes the membership degree of x belonging to the fuzzy concept F.

A fuzzy relation R on X is a fuzzy set on $X \times X$. It is referred to be reflexive if $\mu_R(x,x) = 1$ for all $x \in X$ and symmetric when $\mu_R(x,y) = \mu_R(y,x)$ for any $x, y \in X$. A fuzzy relation is called a fuzzy similarity relation if it is reflexive and symmetric. A fuzzy similarity relation characterizes the similarity degrees between objects. A fuzzy similarity relation R on X is associated with a group of fuzzy sets (fuzzy similarity classes) $\{[x]_R \mid x \in U\}$, reflecting the topological and granular structures of X, where $\mu_{[x]_R}(y) = \mu_R(x,y)$ for any $y \in X$.

There exist lots of ways to determine a fuzzy similarity relation describing the associations between objects in a dataset, such as, the correlation coefficient method, the similarity coefficient method, and a kind of distance-based method [50].

The concept of rough sets, introduced by Pawlak [27], is different from fuzzy sets to interpret and handle objects by using an indiscernibility relation.

For a dataset X with its attribute (feature) set A, the pair (X, A) is called an information system. For any $B \subseteq A$, $R_B = \{(x,y) \in X^2 \mid \forall b \in B(x(b) = y(b))\}$ is an indiscernibility relation on X, which is a crisp equivalence relation partitioning X into a family of disjoint subsets X/R_B, called the quotient set of X with respect to R_B or B, where $x(b)$ denotes the value of x on b. Let $Y \subseteq X$, the two sets

$$\underline{R_B}(Y) = \{Z \in X/R_B \mid Z \subseteq Y\}, \quad \overline{R_B}(Y) = \{Z \in X/R_B \mid Z \cap Y \neq \emptyset\}$$

are referred to as the lower approximation and upper approximation of Y with respect to R_B or B, respectively.

The lower approximation of a set consists of granules of indiscernible objects totally contained in the set, whereas its upper approximation is composed of granules of indiscernible objects partly contained in the set. The pair of the approximation sets can be used to discern and analyze such a set. It is the cores of rough sets theory for knowledge representation and discovery.

When the class information of X is known and D is the class label feature, called the decision feature of X, the information system $(X, A \cup D)$ is referred to be a decision information system. For any $B \subseteq A$,

$$Pos_B(D) = \cup_{Y \in X/_{R_D}} \underline{R_B}(Y)$$

is said to be the positive domain of D with respect to B, where R_B and R_D are indiscernibility relations derived from B and D, respectively. The dependency of D to B can be described by

$$\gamma_B(D) = |Pos_B(D)|/|X|$$

where $|Y|$ denotes the cardinality of Y. For any $a \in B$, the measure

$$sig(a) = \gamma_B(D) - \gamma_{B\setminus\{a\}}(D))$$

can be used to describe the significance of feature a in X with respect to B [11].

In the case of an information system having continuous-values attributes, on one hand, some discretization methods [42] have been developed to process such information systems. On the other hand, the classical indiscernibility relation has been extended to general mathematical notions and various generalized rough set models are constituted, such as a general relation based rough set [36], a similarity or tolerance relation based rough set [15, 33], a dominance relation based rough set [12, 34], a covering rough set [53], and a neighborhood rough set [44, 49]. Herein the neighborhood rough set model is outlined and others can be referred to the literature.

Let X be a finite universe, for each $x \in X$, we associate it with a subset $n(x) \subseteq X$, called the neighborhood of x. A neighborhood system $N(X) = \{n(x) \mid x \in X\}$ of X is a family of neighborhoods associated with all $x \in X$. The pair $(X, N(X))$ is referred to as a neighborhood approximation space. In the neighborhood approximation space $(X, N(X))$, several models of neighborhood rough sets have been exposed [44, 49], where the following two sets of formulae

$$\underline{N_1}(Y) = \{x \in X \mid n(x) \subseteq Y\}, \quad \overline{N_1}(Y) = \{x \in X \mid n(x) \cap Y \neq \emptyset\}$$

and

$$\underline{N_2}(Y) = \cup\{n(x) \in N(X) \mid n(x) \subseteq Y, x \in X\}, \quad \overline{N_2}(Y) = \cup\{n(x) \in N(X) \mid n(x) \cap Y \neq \emptyset, x \in X\}$$

are two kinds of typical depictions of neighborhood rough lower approximations and neighborhood rough upper approximations. Each pair of them brings ones different characterizations about objects and has specific properties.

In the mean time, the connection between fuzzy sets and rough sets has been extensively investigated and several versions of fuzzy rough set models [9, 10, 30] have been worked out, in which

$$\mu_{\underline{R}(F)}(x) = \inf_{y \in X} I(\mu_R(x, y), \mu_F(y)), \mu_{\overline{R}(F)}(x) = \sup_{y \in X} T(\mu_R(x, y), \mu_F(y))$$

are the typical definitions of generalized fuzzy rough lower approximation and upper approximation of a fuzzy set F with respect to a fuzzy relation R on X, $x \in X$, where I is a kind of fuzzy implication and T is a t-norm.

When the fuzzy relation R reduces to a crisp relation on X, or even a neighborhood system $N(X) = \{n(x) \mid x \in X\}$, the model of rough fuzzy set

$$\mu_{\underline{N}(F)}(x) = \inf_{y \in n(x)} \mu_F(y), \mu_{\overline{N}(F)}(x) = \sup_{y \in n(x)} \mu_F(y)$$

is achieved.

In the following section, the neighborhood rough fuzzy set model is introduced to assess the membership degree of an object belonging to a fuzzy association class in a neighborhood system. Such membership measures are incorporated with the similarity measures between objects into the renovation of weight matrix of Laplacian neighborhood graph.

3 A rough fuzzy sets based semi-supervised Laplacian Eigenmaps

Both CCDR and SSLE are semi-supervised nonlinear dimensionality reduction methods. In CCDR, the weights of labelled samples belonging to their classes were all 1 and those of unlabelled samples were all 0, regardless of the membership degrees of samples belonging to their classes. In SSLE, two neighborhood graphs are complementary to each other. The weights were directly assigned to 0.5 between unlabelled neighbors, which cannot exactly express the relationship between labelled and unlabelled samples. The fixed weights disregard the influences of distances and similarity between samples and their prototypes. Thereafter, both methods cannot preserve the class structures of datasets well. In this section we introduce the ideas of fuzzy sets and rough fuzzy sets into the design of weight matrices of the Laplacian neighborhood graph and class-related neighborhood graph, and propose a novel semi-supervised LE method for dimensionality reduction.

3.1 Assessment of significance of features

In a high dimensional data space, data distributions are sparse. Each feature contributes to different ingredients for data clustering and parts of features usually cause serious impacts on clustering effect. In this subsection, motivated by the rough set theory [11], the significance of each feature is adaptively assessed by introducing an information entropy measure.

Given a continuous, discrete, or hybrid dataset $X = \{x_1, x_2, \dots, x_n\}$ with the feature (attribute) set $A = \{a_1, a_2, \dots, a_m\}$, we assume that there are s labelled samples in X, and $l_1, l_2, \dots, l_s \in \{1, 2, \dots, f\}$ are their class labels. Due to the fact that the ranges of different features vary in magnitude, the dataset needs to be normalized for all continuous values features before carrying out lower dimensional embedding. The commonly used approaches to data normalization are as follows [28].

(1) The range normalization (min–max method)

$$x^*_{ij} = \frac{x_{ij} - \min_k x_{kj}}{\max_k x_{kj} - \min_k x_{kj}}, \quad i = 1, 2, \dots, n, j = 1, 2, \dots, m$$

(2) The Z-score normalization (mean-deviation method)

$$x^*_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

where

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \mu_j)}$$

After data normalization, we establish a semi-supervised fuzzy similarity matrix regarding the given dataset. For convenience, the normalized dataset is also denoted by X. The similarity degrees between samples with the same class labels are set as 1, while the similarity between samples with different class labels is set as 0, and the similarity degrees between the rests are computed by a certain kind of distance method. That is, for two samples, if one is labelled and the other is not, or neither of them is labelled, then the similarity degree between them is evaluated by their distance. Mathematically, the semi-supervised fuzzy similarity matrix R_A of X is obtained by

$$\mu_{R_A}(x_i, x_j) = \begin{cases} 1, & \text{if } x_i \text{ and } x_j \text{ have the same label} \\ 1 - \alpha\sqrt{\sum_{k=1}^{m} d_k^2(x_{ik}, x_{jk})}, & \text{if at least one of } x_i \text{ and } x_j \text{ is not labelled} \\ 0, & \text{if } x_i \text{ and } x_j \text{ have different labels} \end{cases} \tag{3.1}$$

where α is an appropriate positive number such that $\mu_{R_A}(x_i, x_j) \in [0, 1]$ for all $x_i, x_j \in X$, and $d_k(x_{ik}, x_{jk})$ is the distance between the values of kth feature of samples x_i and x_j. If a_k is a continuous attribute, we define $d_k(x_{ik}, x_{jk}) = |x_{ik} - x_{jk}|$, and when a_k is a discrete or categorial attribute, we set

$$d_k(x_{ik}, x_{jk}) = \begin{cases} 0, & \text{if } x_{ik} = x_{jk} \\ 1, & \text{if } x_{ik} \neq x_{jk} \end{cases}$$

It is clear that R_A is a reflexive and symmetrical relation, or a fuzzy similarity relation. The family of fuzzy similarity classes $\{[x]_{R_A}\}$ characterize the granular structures of X, where $[x]_{R_A}$ is a fuzzy set on X and its membership degree at $y \in X$ is $\mu_{R_A}(y, x)$, thus, $\mu_{[x]_{R_A}}(y)$ or $\mu_{R_A}(y, x)$ can be interpreted as the membership degree of sample y belonging to the fuzzy similarity class of sample x. Let

$$H(A) = -\sum_{j=1}^{n} \frac{\sum_{i=1}^{n} \mu_{R_A}(x_i, x_j)}{|X|^2} \log \frac{\sum_{i=1}^{n} \mu_{R_A}(x_i, x_j)}{|X|^2}$$

then $H(A)$ is referred to as the information entropy of the family of fuzzy information granules (similarity classes) $\{[x]_{R_A} | x \in X\}$, reflexing the distributions or fluctuation of membership degrees of the fuzzy information granules. Obviously, $H(A) \in [0, \log n)$.

According to (3.1), it is evident that $\sum_{i=1}^{n} \mu_{R_A}(x_i, x_j) \leq \sum_{i=1}^{n} \mu_{R_{A\backslash\{a\}}}(x_i, x_j) \leq |X|$ for any $a \in A$ and $x_j \in X$. Thus,

$$\frac{\sum_{i=1}^{n} \mu_{R_A}(x_i, x_j)}{|X|^2} \leq \frac{\sum_{i=1}^{n} \mu_{R_{A\backslash\{a\}}}(x_i, x_j)}{|X|^2} \leq \frac{1}{n}$$

Therefore, when $n > 2$, $H(A) \leq H(A\backslash\{a\})$. If $n \leq 2$, this result is trivial. Hence, for any $a \in A$, let

$$sig(a) = \frac{H(A\backslash\{a\}) - H(A)}{\max_{a_i \in A}(H(A\backslash\{a_i\}) - H(A))} \tag{3.2}$$

then $sig(a) \in [0, 1]$. $sig(a)$ can characterize the significance of a in A while considering the problem of preserving the same indiscernibility. According to (3.2), if $sig(a)$ is smaller, then the fluctuation of all of the kth attribute values is smaller, and for any x_i and x_j, the value of $d_k(x_i, x_j)$ is relatively smaller. Thus it has a higher possibility that both x_i and x_j are in the same class. On the other hand, if $sig(a)$

is bigger, the fluctuation of all of the kth attribute values is larger and the value of $d_k(x_i, x_j)$ is relatively bigger, thus the possibility of samples x_i and x_j belonging to different classes is higher and therefore the attribute a has a stronger discernibility ability. Based on such a fact, we modify the fuzzy similarity matrix (3.1) by introducing the significance of features into the computation of distances between samples and obtain the following weighted fuzzy similarity relation

$$\mu_{R_W}(x_i, x_j) = \begin{cases} 1, & \text{if } x_i \text{ and } x_j \text{ have the same label} \\ 1 - \alpha\sqrt{\sum_{k=1}^{m} S_k d_k^2(x_{ik}, x_{jk})}, & \text{if at least one of } x_i \text{ and } x_j \text{ is unlabelled} \\ 0, & \text{if } x_i \text{ and } x_j \text{ have different labels} \end{cases} \tag{3.3}$$

where $S_k = e^{sig(a_k)}$.

From (3.3) we know that if $sig(a_k)$ is smaller, then $\mu_{R_W}(x_i, x_j)$ is almost equal to $\mu_{R_A}(x_i, x_j)$. If $sig(a_k)$ is bigger, then $\mu_{R_W}(x_i, x_j)$ will be smaller than $\mu_{R_A}(x_i, x_j)$, assuming that there are no other variable factors. That is, if two samples belong to different classes with a high possibility, then we impose a bigger weight to the computation of distance and derive a smaller similarity degree between both samples. Therefore, if the similarity degree of two samples is small, then R_W can push the two samples farther when samples are embedded in a lower dimensional space.

If two samples x_i and x_j have the same label, we assume that the membership degree of any sample $x \in X$ belonging to the fuzzy similarity class of x_i is equal to that of x belonging to the fuzzy similarity class of x_j. Under such an assumption, we let

$$\mu_{R_S}(x_i, x_j) = \begin{cases} \mu_{R_W}(x_i, x_j), & \text{if } x_j \text{ is not labelled} \\ \max_{y \in l(t)} \mu_{R_W}(x_i, y), & \text{if } x_j \text{ is labelled} \end{cases} \tag{3.4}$$

where t is the class label of sample x_j and $l(t)$ is the subset of X, whose elements have the class label t.

From (3.4) we know that the similarity between homogeneous samples is larger than that between heterogeneous samples. We call R_S a fuzzy association relation. It is obvious that R_S is no longer a fuzzy similarity relation. We refer the fuzzy set X_j, defined by $\mu_{X_j}(x) = \mu_{R_S}(x, x_j)$, $x \in X$, to as the fuzzy association class of x_j. Thus, $\mu_{R_S}(x_i, x_j)$, denoting the membership degree of x_i belonging to X_j, is not necessarily equal to the membership degree of x_j belonging to X_i. That is, a sample has a possibility of belonging to some class, but there may exist samples in this cluster which have different degrees of possibility belonging to the class that the given sample is in. These interpretations are rational in clustering analysis.

3.2 A semi-supervised rough fuzzy Laplacian Eigenmaps (SSRFLE)

For a hybrid dataset $X = \{x_1, x_2, \dots, x_n\}$ with its feature set $A = \{a_1, a_2, \dots, a_m\}$, we assume that there are s samples being labelled as $l_1, l_2, \dots, l_s$, where $l_i \in \{1, 2, \dots, f\}$ and f is the number of classes, $f \leq s$. For a given positive integer k, we construct a k-nearest neighborhood graph G^N of X, where the distance between samples x_i and x_j is computed by using the following weighted distance

$$d_W(x_i, x_j) = \sqrt{\sum_{k=1}^{m} S_k d_k^2\left(x_{ik}, x_{jk}\right)} \tag{3.5}$$

In order to determine the membership degree of a sample belonging to a certain class, we construct a neighborhood rough fuzzy set model for the fuzzy association classes $\{[x]_{R_S} \mid x \in X\}$.

Let $N_k(x)$ be the k-nearest neighborhood set of $x \in X$ and $N_k = \{N_k(x) | x \in X\}$ be the k-nearest neighborhood system on X, the pair $K = (X, N_k)$ is referred to as a neighborhood approximation space. Let X_j denote the fuzzy association class of x_j and is indeed the jth column of R_S. The rough fuzzy lower and upper approximation sets of X_j with respect to the neighborhood system N_k are two fuzzy sets on X, and their membership functions are defined by, $x_i \in X$,

$$\begin{aligned} \mu_{\underline{N}(X_j)}(x_i) &= \bigwedge_{y \in N_k(x_i)} \mu_{X_j}(y) \\ \mu_{\overline{N}(X_j)}(x_i) &= \bigvee_{y \in N_k(x_i)} \mu_{X_j}(y) \end{aligned} \tag{3.6}$$

Due to the fact that the similarity degrees between homogeneous samples are larger than those between heterogeneous samples, according to (3.6), we know that $\mu_{\underline{N}(X_j)}(x_i)$ is bigger if the samples in the k-nearest neighborhood of x_i belongs to the class that x_j belongs to, whereas $\mu_{\underline{N}(X_j)}(x_i)$ is smaller if the neighbors of x_i and x_j belong to heterogeneous classes. So $\mu_{\underline{N}(X_j)}(x_i)$ can be used to assess the degree what two samples x_i and x_j belonging to the same class.

In order to ensure homogeneous samples being mapped closer in the lower dimensional space, in the k-nearest neighborhood graph G^N of X, the weight between two vertices should be large when both vertices are in the homogeneous neighborhoods, while it is small if they are in heterogeneous neighborhoods. Motivated by this idea, we designate the weights (similarity degrees) between samples with the same class label in a neighborhood as the largest value 1, while the similarity between samples with different labels as the smallest value 0. The similarity between samples that are not included in each other neighborhood is also set as the smallest value 0, and the similarity degree between samples that only one is labelled or neither is labelled is evaluated by combining the weighted Gaussian kernel distance between samples and the membership degrees of both samples belonging to the same class. As a result, we establish the weight measure between vertices x_i and x_j as

$$W_{ij}^N = \begin{cases} 1, & \text{if } x_i \text{ and } x_j \text{ have the same label} \\ & \text{and } x_i \in N_k(x_j) \text{ or } x_j \in N_k(x_i) \\ \mu_{\underline{N}(X_i)}(x_j)e^{-d_W(x_i,x_j)^2/\sigma}, & \text{if at least one of } x_i \text{ and } x_j \text{ is not labelled} \\ & \text{and } x_i \in N_k(x_j) \text{ or } x_j \in N_k(x_i) \\ 0, & \text{otherwise} \end{cases} \tag{3.7}$$

where σ is a scale factor of adjusting the Gaussian kernel function, usually determined by the average similarity degrees between samples. The weight matrix of G^N is denoted by $W^N = (W_{ij}^N)_{n\times n}$.

Since the similarity degrees between each sample and samples with the same label are uniformed in Eq. (3.4), the samples with the same label can be regarded as one sample, namely the prototype. If a sample x_j has a class label $t \in \{1, 2, \dots, f\}$, then $\mu_{R_S}(x_i, x_j)$ can be interpreted as the membership degree of x_i belonging to the tth class. For that, we build a weighted class-related neighborhood graph G^C to depict the relationship between samples and their prototypes and the corresponding weight matrix is denoted by $W^C = (W_{it}^C)_{n\times f}$, where the weight W_{it}^C, defined by

$$W_{it}^C = \mu_{R_s}(x_i, y) \tag{3.8}$$

represents the membership degree of x_i belonging to the tth class, here y is an arbitrary element in $l(t)$, whose labels are all t.

Let the row vectors $y_1, y_2, \dots, y_n \in \mathbb{R}^d$ be the $d\,(d < n)$ dimensionality representations of dataset $X = \{x_1, x_2, \dots, x_n\}$. In order to achieve the goal that homogeneous samples are mapped closer and more compact around the prototypes in the lower dimensional space, the following optimization problem

$$\min(1-\gamma)\sum_{i,j}^{n} W_{ij}^N \left\|y_i - y_j\right\|^2 + \gamma \sum_{i=1}^{n}\sum_{t=1}^{f} W_{it}^C \left\|y_i - c_t\right\|^2 \tag{3.9}$$

is approached, where $c_1, c_2, \dots, c_f \in \mathbb{R}^d$ are the prototypes of the dataset X in the lower dimensional space and γ is a parameter trading off the performance of preserving local neighborhood structure and maintaining clustering structure.

Let

$$Z = \begin{pmatrix} c_1 \\ \vdots \\ c_f \\ y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1d} \\ \vdots & \vdots & & \vdots \\ c_{f1} & c_{f2} & \cdots & c_{fd} \\ y_{11} & y_{12} & \cdots & y_{1d} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nd} \end{pmatrix} = [z^1\, z^2\, \dots\, z^d]$$

then the optimization problem (3.9) can be translated to

$$\min tr(Z^T \mathbb{L} Z) \tag{3.10}$$

where $\mathbb{L} = \mathbb{D} - \mathbb{W}$ is the $(f+n)\times(f+n)$ Laplacian matrix, which is a symmetric and positive semidefinite matrix, $\mathbb{D}$ is a $(f+n)\times(f+n)$ diagonal matrix with $\mathbb{D}_{ii} = \sum_{j=1}^{f+n} \mathbb{W}_{ij}$, and $\mathbb{W} = \begin{pmatrix} I & \gamma(W^C)^T \\ \gamma W^C & 2(1-\gamma)W^N \end{pmatrix}$.

In order to ensure that the optimization problem (3.10) has a unique solution, two restricted conditions $Z^T\mathbb{L}Z = I$ and $Z^T\mathbb{D}\mathbf{1} = 0$ are imposed to remove scaling and translation factors in the lower dimensional embedding. By the Lagrange multiplier method, the optimization problem (3.10) can be transformed to solve the following generalized eigenvalue problem

$$\mathbb{L}Z = \lambda \mathbb{D} Z \tag{3.11}$$

The column vectors $z^1, z^2, \dots, z^d$, corresponding to the d smallest positive eigenvalues $0 \neq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$, are the solutions to Eqs. (3.11) or (3.9). The first f rows of the matrix $[z^1\, z^2\, \dots\, z^d]$ correspond to the coordinates of prototypes and the following n rows determine the d dimensional embedding of the original samples.

4 Comparative experiments and analysis

Due to diversities and complex natures of captured datasets, there is no single method being able to deal successfully with all situations. In this section, we compare the proposed SSRFLE with the classical (non-supervised) LE, DVE, and the state of the art semi-supervised CCDR and SSLE, for dimensionality reduction in the aspects of classification performance and data visualization through several experiments on benchmark datasets. All of the experiments are implemented on the platform of Matlab 7.0.

Table 1 The details of five datasets used in the experiments

Datasets	Size	#Condition attributes	#Decision attribute	#Classes
Wine	178	13	1	3
Seeds	210	7	1	3
WDBC	569	31	1	2
Heart	270	13	1	2
Mushroom	1650	22	1	2

We take five benchmark datasets from UCI Machine Learning Repository [23] for our experiments. Three of which, Wine, Seeds, and Wisconsin Diagnostic Breast Cancer (WDBC), are continuous datasets. Statlog Heart (Heart) is a hybrid dataset, while Mushroom is a categorical dataset. Due to the fact that the Mushroom dataset has missing values, actually 5643 complete samples, and is classified into two classes, we randomly select 30% complete samples in each class as the test samples in the follow-up experiments. The details of the five datasets are illustrated in Table 1.

4.1 Performance analysis

Since the values of continuous features in the first four datasets have different scales, all of these datasets are preprocessed by the normalized methods listed in Sect. 3.1. We compare the proposed SSRFLE with the classical LE and DVE, and the state of arts CCDR and SSLE. For consistence, we set the parameters $\beta = 1$ for CCDR, $\mu = 0.5$ for SSLE, and $\gamma = 0.5$ for SSRFLE.

The first two methods (LE and DVE) are non-supervised and the last two and the proposal are semi-supervised. Thus we randomly label a percentage of samples in each class of every dataset as the semi-supervised information in the simulation experiments. Based on the analysis of establishing the proposed method, one of the largest advantage of the proposed SSRFLE is to preserve class characterization of original data, the classification (clustering) accuracy is used as one of the criteria to test the performance of these techniques. With the proposed SSRFLE, the lower dimensional representations of original data as well as the prototypes of the lower dimensional embedding can be derived. Therefore, the commonly used fuzzy C-means (FCM) clustering method is introduced to cluster the dataset in the lower dimensional embedding. The fuzzification factor m in FCM is set to be 2 in all experiments.

Although there were several approaches to dimensionality estimation of a data manifold [7], the real intrinsic dimensionality of the dataset may not be correctly determined. In the following experiments, we assume the embedded dimensionality d varying from 2 to 4, instead of estimating its dimensionality. For every fixed d, the size of neighborhood of each sample varies from 4 to 14. In each semi-supervised method, 5% samples in each class of every dataset are randomly labelled. For every set of such parameters, the tenfold cross-validation is carried out and the average classification accuracy is used to evaluate the listed methods for dimensionality reduction. Tables 2, 3 and 4 show the average classification accuracies when d is set as 2, 3 and 4, respectively. The bold number in each column (each dataset) of every table shows the highest classification accuracy among the tested methods.

Table 2 The average classification accuracies when $d = 2$

	Wine	Seeds	WDBC	Heart	Mushroom
LE	69.05	84.98	85.91	62.95	65.07
DVE	70.64	85.67	87.38	65.84	72.59
CCDR	71.66	85.98	87.56	62.31	83.26
SSLE	69.25	85.11	85.70	66.40	65.60
SSRFLE	**95.96**	**89.05**	**91.25**	**75.40**	**84.28**

Table 3 The average classification accuracies when $d = 3$

	Wine	Seeds	WDBC	Heart	Mushroom
LE	68.44	82.86	67.50	60.69	67.04
DVE	70.24	84.63	83.00	62.71	73.94
CCDR	70.13	83.96	74.26	64.74	84.55
SSLE	68.23	83.59	70.01	66.08	68.83
SSRFLE	**90.57**	**86.53**	**91.20**	**75.68**	**90.36**

Table 4 The average classification accuracies when $d = 4$

	Wine	Seeds	WDBC	Heart	Mushroom
LE	64.66	84.03	52.04	58.23	69.62
DVE	64.23	**84.69**	82.47	63.75	74.28
CCDR	67.39	83.32	70.84	60.70	81.25
SSLE	65.18	83.37	55.73	58.05	74.78
SSRFLE	**89.26**	84.11	**86.61**	**78.68**	**85.14**

From these Tables we know that the semi-supervised dimensionality reduction methods, CCDR, SSLE, and the proposed SSRFLE, are all superior to the classical LE in classification performance for the implemented five datasets and for the chosen three dimensionalities. These results are consistent with our intuition that semi-supervised methods outperform non-supervised ones. The non-supervised DVE can bring out better classification accuracies than LE and SSLE in most cases. However, its time cost is rather larger since it is a global method of unfolding the data manifold by maximizing the global variance. Nevertheless, the experimental results show that the proposed SSRFLE brings out the highest average classification accuracies among the tested methods for each of the five datasets when the embedded dimensionality is taken 2 or 3. When $d = 4$, although SSRFLE does not produce the best classification accuracies for all of the five datasets, the best classification accuracies can be achieved by SSRFLE for four out of five datasets and the rest one (for the dataset Seeds) is close to the best one obtained by DVE. These facts indicate that different embedded dimensionalities taken produce slightly distinct impacts on the classification accuracy for a given dataset. One may choose empirically a suitable embedded dimensionality by experiments in practical applications.

As for the facts that each method aforementioned produces distinct classification accuracies for different datasets, the main reasons arise from the diversities of capture ways, mechanism and topological structures of the datasets. It is the difference between datasets that can be used to verify the performance of an algorithm.

To further test the semi-supervised performance of the proposed SSRFLE, we compare the proposed SSRFLE with the semi-supervised CCDR and SSLE through the five datasets aforementioned with the same parameter settings, except that the rates of labelled samples are set to be 5, 15 and 30%, respectively, in each class of every dataset. The number of embedded dimensionality is set as 2. Figure 1 shows the experimental results.

In each subfigure of Fig. 1, the bars with blue, green and brown in each group represent the classification accuracies in the cases of 5, 15 and 30% labelled samples, respectively. The results show that for each of the five datasets, the classification accuracy of each method increases when the rate of labelled samples increases. These facts are consistent with our intuition that more supervised information brings out higher classification accuracy. Furthermore, the proposed method achieves the highest classification accuracy among the three implemented methods for each dataset and for each rate of labelled samples in the same experimental setting.

4.2 Data visualization

Data visualization is an important research issue in the field of machine learning. Especially, it makes ones perceive and inspect high dimensional data or complicated phenomena in an intuitionistic and visible way. An effective dimensionality reduction and visualization method brings ones to a vive and convictive demonstration on high dimensional data. In this subsection, we test the visualization of the Wine dataset by using LE, DVE, SSLE, CCDR, and the proposed SSRFLE method. This dataset has 178 samples that fall into 3 classes. We randomly label 5% samples in each class. Figure 2 shows the visualized results by using these methods.

In Fig. 2, three classes of samples are depicted by different colors. From Fig. 2 the first two algorithms lead to that the classes have not been completely separated. Although

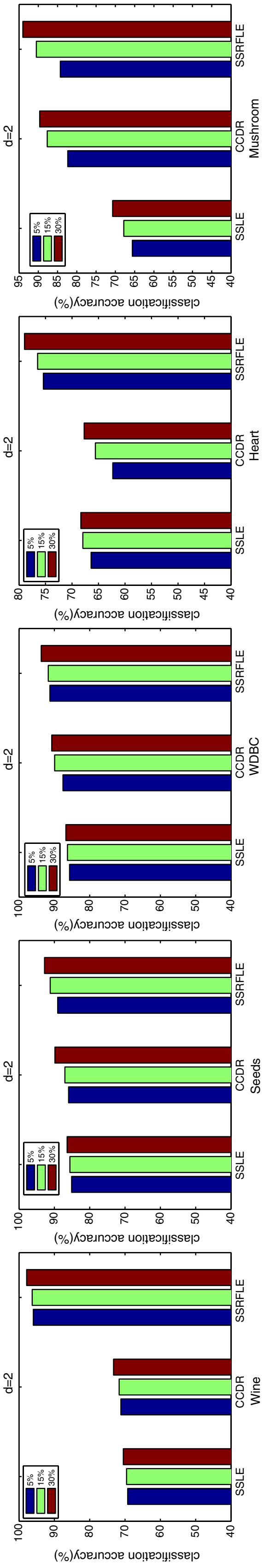

Fig. 1 The influence of rate of labelled samples on the classification accuracy (for interpretation of the references to color in the text, the reader is referred to the web version of this article)

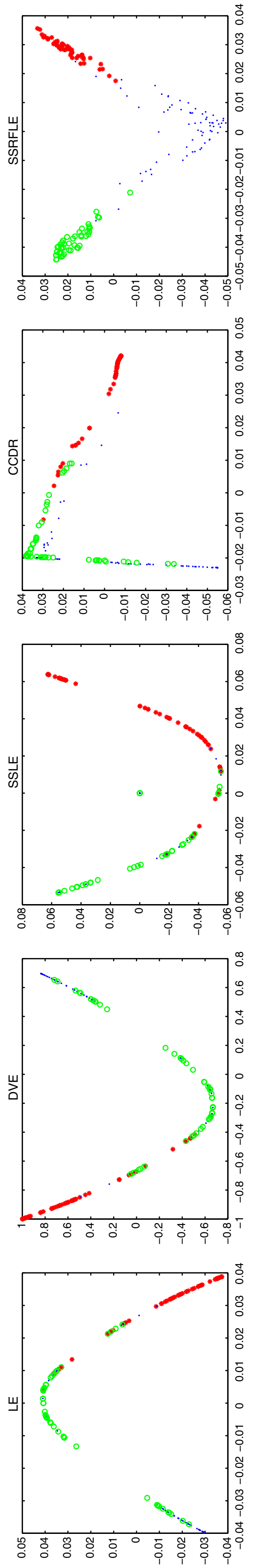

Fig. 2 The results of visualization obtained by LE, DVE, SSLE, CCDR, and SSRFLE (for interpretation of the references to color in the text, the reader is referred to the web version of this article)

SSLE and CCDR are semi-supervised and the label information of dataset has been incorporated into, the similarity between homogeneous samples and dissimilarity between heterogeneous samples have not been involved. Homogeneous samples scatter, but have no compact embedding. The last figure in Fig. 2 plots the visualization of this dataset by using the proposed method. It is evident that homogeneous samples have compact distributions and heterogeneous samples can be better distinguished.

4.3 Impact of feature weights on classification accuracy

It is natural that different features have distinct impacts on clustering structures of a dataset. In the proposed SSRFLE, an adaptive weight has been designated to each feature based on the information entropy theory in the computation of similarity, or alternatively, distance, between two samples. In order to verify the rationality of using the weighted distance in SSRFLE, we repeat the comparative experiments above in the environments of with and without weights of features (in the case of without weighting, we set $S_k = 1$ or $sig(a_k) = 0$ for all $k = 1, 2, \ldots, m$ in (3.1), and sign this method as 'SSRFLE without weighting'). In these experiments, we set $d = 2$ and let the size of neighborhood vary from 4 to 14. We also randomly label 5% samples of each class in each dataset. Figure 3 shows the experimental results of the five datasets.

From Fig. 3 one sees that, on the whole, the classification accuracies of SSRFLE are higher than those of SSRFLE without weighting for the first four datasets. They are fluctuant for the Mushroom dataset. The reason may arise from that Mushroom is a discrete dataset and the weights of features are computed according to the frequencies of feature values appeared. Features with different significance may have similar frequent distributions of feature values. On the whole, it is convinced that the weighted distance can improve the effectiveness of dimensionality reduction and greatly increase the classification accuracy, especially for the continuous and hybrid datasets.

4.4 Parameters selection

In the proposed method, there are two parameters to be assigned before experiments, the size of neighborhood and the number of embedded dimensionality. The choice strategy of the latter has been accounted for in Sect. 4.1. In this subsection, we discuss the influence of the size of neighborhood on classification accuracy by comparing the five methods on the five datasets for dimensionality reduction and clustering performance.

We label 5% samples in each class of every dataset. The number of embedded dimensionality d is taken as 2, 3, and 4, and the size of neighborhood k varies from 4 to 14. The

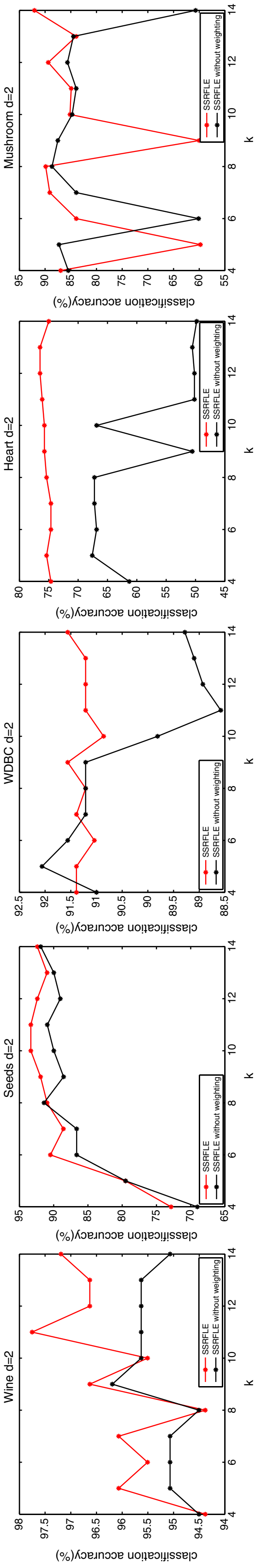

Fig. 3 The effects of with and without weighting in SSRFLE for five datasets (for interpretation of the references to color in the text, the reader is referred to the web version of this article)

tenfold cross-validation is executed for each set of parameters and the average classification accuracy is recorded. Figure 4 shows the experimental results.

In Fig. 4, each of the three rows displays the experimental results for different embedded dimensionality, $d = 2$, 3, and 4, respectively. The five subfigures in each row show the relationship between classification accuracy and size of neighborhood of five datasets, namely, Wine, Seeds, WDBC, Heart, and Mushroom, separatively. Every subfigure plots five groups of classification accuracies derived from the aforementioned five methods with the variation of size of neighborhood from 4 to 14. It is shown that the proposed method produces the highest classification accuracies among all the five methods for all datasets, for all sizes of neighborhood sets, and for most of given numbers of embedded dimensionality. It also illustrates that the classification accuracy by using the proposed SSRFLE is not very sensitive to the choice of size of neighborhood. The essential may be the intervention of adaptive weighted distance and the rough fuzzy approximation characterization on neighborhood of samples.

5 Conclusions and future work

In this work, a semi-supervised rough fuzzy Laplacian Eigenmaps (SSRFLE) is proposed for nonlinear dimensionality reduction of hybrid data. In this method, a semi-supervised fuzzy similarity relation is introduced and the weights of features of a dataset are adaptively assessed by designing an information entropy measure based on this fuzzy similarity relation. A new fuzzy association relation is derived from the weighted distances between samples together with the labelled information of the dataset. The rough fuzzy lower approximations of the fuzzy association classes related to the fuzzy association relation together with the gaussian Kernel weighted distances between samples are used to characterize the similarities between vertices of the Laplacian neighborhood graph. At the mean time, the fuzzy association classes are considered as the descriptions of similarities between samples and their prototypes, which produce the weighted class-related neighborhood graph. The combination of both neighborhood graphs ensures homogeneous samples being embedded closer and more compact around the prototypes in the lower dimensional space of a dataset.

A series of comparative experiments on real world hybrid datasets are implemented. Experimental results show that the proposed method outperforms the state of arts methods in the aspect of classification accuracy and data visualization due to the superior performance of preserving class structures when a dataset is embedded into a lower dimensional space.

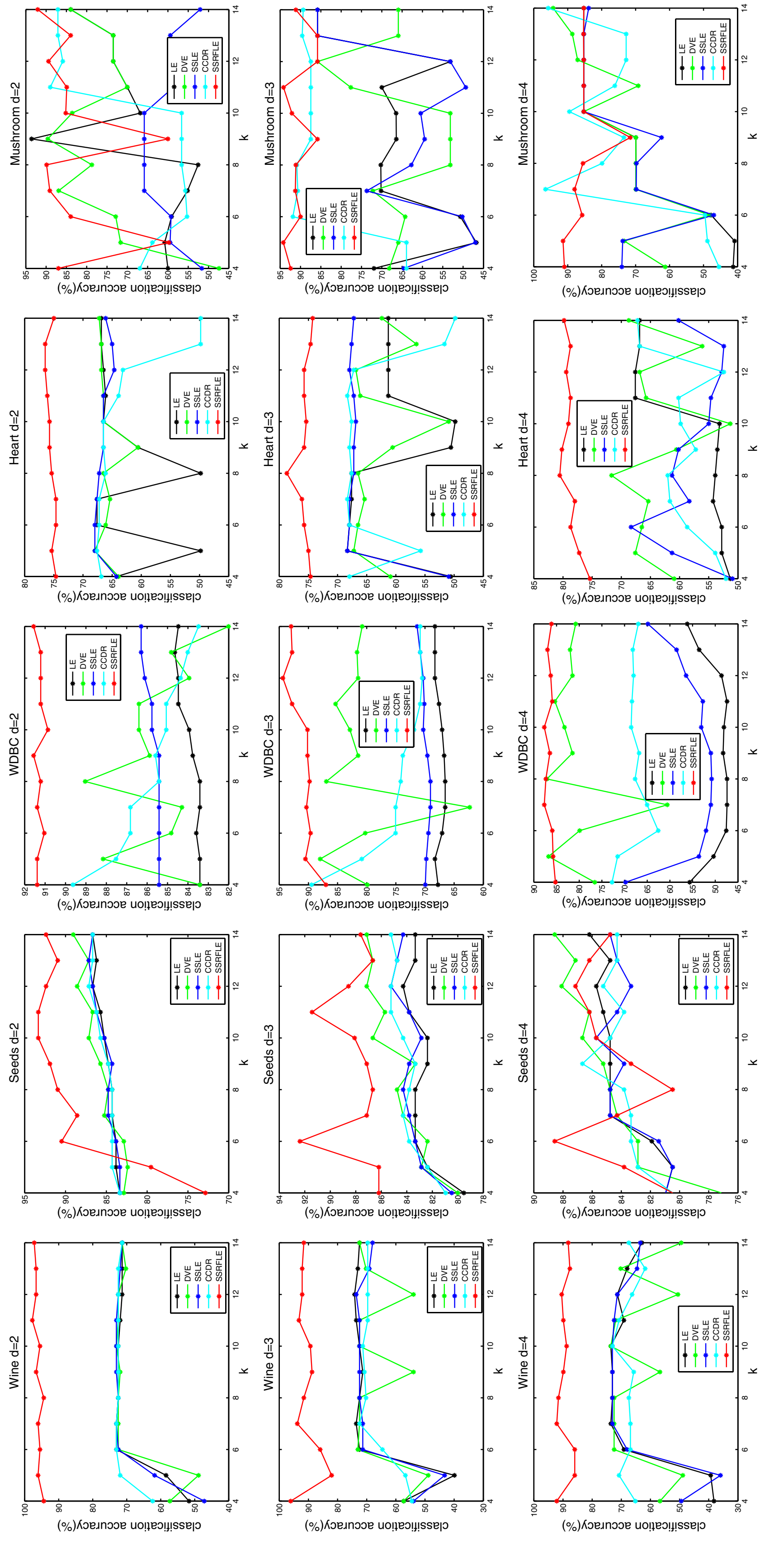

Fig. 4 The influences of parameters k and d on classification accuracies for five datasets (for interpretation of the references to color in the text, the reader is referred to the web version of this article)

Although the proposed method is not very sensitive to the choice of size of neighborhood, we will lucubrate on appropriate approaches in theory to the choice of such a parameter in the future. Theoretical analysis and applications of the related approaches to incremental and dynamic data are under consideration.

Acknowledgements This work was supported by the National Natural Science Foundation of China (11471001).

References

1. Abd El-Monsef ME, El-Gayar MA, Aqeel RM (2017) A comparison of three types of rough fuzzy sets based on two universal sets. Int J Mach Learn Cybern 8:343–353
2. Abdel-Mannan O, Ben Hamza A, Youssef A (2007) Incremental hessian locally linear embedding algorithm. IEEE Int Sympo Signal Process Appl 1–4
3. Bartholomew DJ (1983) Principal components analysis probability, statistical optics, and data testing. Springer, Berlin, Heidelberg
4. Belkin M, Niyogi P (2003) Laplacian Eigenmaps for dimensionality reduction and data representation. Neural Comput 15:1373–1396
5. Cai X, Wen G, Wei J, Li J, Yu Z (2014) Perceptual relativity-based semi-supervised dimensionality reduction algorithm. Appl Soft Comput 16:112–123
6. Chen C, Zhang L, Bu J, Wang C, Chen W (2010) Constrained Laplacian Eigenmap for dimensionality reduction. Neurocomputing 73:951–958
7. Costa JA, Hero AO (2004) Geodesic entropic graphs for dimension and entropy estimation in manifold learning. IEEE T Signal Process 52:2210–2221
8. Costa JA, Hero AO (2005) Classification constrained dimensionality reduction. Proceedings of (ICASSP '05). IEEE Int Conf Acoust Speech Signal Process 5:1077–1080
9. Deng TQ, Chen YM, Xu WL, Dai QH (2007) A novel approach to fuzzy rough sets based on a fuzzy covering. Inf Sci 177:2308–2326
10. Dubois D, Prade H (1990) Rough fuzzy sets and fuzzy rough sets. Int J Gen Syst 17:191–209
11. Estévez PA, Tesmer M, Perez CA, Zurada J (2009) Normalized mutual information feature selection. IEEE T Neural Network 20:189–201
12. Greco S, Matarazzo B, Slowinski R (2002) Rough approximation by dominance relations. Int J Intell Syst 17:153–171
13. Huang S, Zhuang L (2016) Exponential discriminant locality preserving projection for face recognition. Neurocomputing 208:373–377
14. Hsu CC, Huang WH (2016) Integrated dimensionality reduction technique for mixed-type data involving categorical values. Appl Soft Comput 43:199–209
15. Järvinen J, Radeleczki S (2014) Rough sets determined by tolerances. Int J Approx Reason 55:1419–1438
16. Jiang Q, Jia M, Hu J (2009) Machinery fault diagnosis using supervised manifold learning. Mech Syst Signal Process 23:2301–2311
17. Keyhanian S, Nasersharif B (2014) Laplacian Eigenmaps modification using adaptive graph for pattern recognition. Int Sympo Telecommun 25–29
18. Kim K, Lee J (2014) Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction. Pattern Recogn 47:758–768
19. Lai ZH, Wong WK, Xu Y, Yang J, Zhang D (2016) Approximate orthogonal sparse embedding for dimensionality reduction. IEEE T Neural Net Learn 27:723–735
20. Li R (2013) A new supervised Laplacian Eigenmap for expression recognition. J Inf Comput Sci 10:4445–4451
21. Li K, Kwong S (2014) A general framework for evolutionary multiobjective optimization via manifold learning. Neurocomputing 146:65–74
22. Li WT, Xu WH (2015) Double-quantitative decision-theoretic rough set. Inf Sci 316:54–67
23. Lichman M (2013) UCI machine learning repository. University of California, School of Information and Computer Science, Irvine, CA. http://archive.ics.uci.edu/ml
24. Lin RS, Yang MH, Levinson SE (2004) Object tracking using incremental Fisher discriminant analysis. Int Conf Pattern Recogn 2:757–760
25. Liu F, Zhang W, Gu S (2016) Local linear Laplacian Eigenmaps: a direct extension of LLE. Pattern Recogn Lett 75:30–35
26. Malik ZK, Hussain A, Wu J (2016) An online generalized eigenvalue version of Laplacian Eigenmaps for visual big data. Neurocomputing 173:127–136
27. Pawlak Z (1982) Rough sets. Int J Comput Inf Sci 11:341–356
28. Pollesch NL, Dale VH (2016) Normalization in sustainability assessment: methods and implications. Ecol Econ 130:195–208
29. Raducanu B, Dornaika F (2012) A supervised non-linear dimensionality reduction approach for manifold learning. Pattern Recogn 45:2432–2444
30. Radzikowskaa AM, Kerre EE (2002) A comparative study of fuzzy rough sets. Fuzzy Sets Syst 126:137–155
31. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326
32. Singer A (2006) Spectral independent component analysis. Appl Comput Harmon Anal 21:135–144
33. Slowinski R, Vanderpooten D (2000) A generalized definition of rough approximations based on similarity. IEEE T Knowl Data En 12:331–336
34. Susmaga R (2014) Reducts and constructs in classic and dominance-based rough sets approach. Inf Sci 277:45–54
35. Suykens JA (2008) Data visualization and dimensionality reduction using kernel maps with a reference point. IEEE T Neural Network 19:1501–1517
36. Tan A, Li J (2015) A kind of approximations of generalized rough set model. Int J Mach Learn Cybern 6:455–463
37. Tenenbaum JB, De Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323
38. Thangavel K, Pethalakshmi A (2009) Dimensionality reduction based on rough set theory: a review. Appl Soft Comput 9:1–12
39. Tsang ECC, Sun B, Ma W (2017) General relation-based variable precision rough fuzzy set. Int J Mach Learn Cybern 8:891–901
40. Wang Q, Li J (2009) Combining local and global information for nonlinear dimensionality reduction. Neurocomputing 72:2235–2241
41. Wang XZ, Dong CR, Fan TG (2007) Training T-S norm neural networks to refine weights for fuzzy if-then rules. Neurocomputing 70:2581–2587
42. Wang XZ, Hong JR (1998) On the handling of fuzziness for continuous-valued attributes in decision tree generation. Fuzzy Sets Syst 99:283–290
43. Wang XZ, Li CG (2005) A new definition of sensitivity for RBFNN and its applications to feature reduction. Lect Notes Comput Sci 3496:81–86
44. Wu WZ, Zhang WX (2002) Neighborhood operator systems and approximations. Inf Sci 144:201–217

45. Xu WH, Guo YT (2016) Generalized multigranulation double-quantitative decision-theoretic rough set. Knowl Based Syst 105:190–205
46. Xu J, Gu ZH, Xie K (2016) Fuzzy local mean discriminant analysis for dimensionality reduction. Neural Process Lett 44:701–718
47. Xu W, Li WT (2016) Granular computing approach to two-way learning based on formal concept analysis in fuzzy datasets. IEEE Trans Cybern 46:366–379
48. Xu J, Xie SL, Zhu WK (2017) Marginal patch alignment for dimensionality reduction. Soft Comput 21:2347–2356
49. Yao YY (1998) Relational interpretations of neighborhood operators and rough set approximation. Inf Sci 111:239–259
50. Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353
51. Zhai J, Zhang Y, Zhu H (2017) Three-way decisions model based on tolerance rough fuzzy set. Int J Mach Learn Cybern 8:35–43
52. Zhang Y, Li B, Wang W, Sun T, Yang X (2014) Supervised locally tangent space alignment for machine fault diagnosis. J Mech Sci Tech 28:2971–2977
53. Zhu W (2007) Topological approaches to covering rough sets. Inf Sci 177:1499–1508