Pattern Analysis and Applications (2019) 22:999–1013
https://doi.org/10.1007/s10044-018-0707-2

THEORETICAL ADVANCES

Unsupervised feature selection in linked biological data

Elham Hoseini[1] · **Eghbal G. Mansoori**[1]

Received: 21 April 2017 / Accepted: 8 April 2018 / Published online: 27 April 2018

Abstract
Feature selection techniques have become an apparent need in many bioinformatics applications, especially when there exist a huge number of features. For instance, classification of hereditary disease genes/proteins plays a significant role in prediction and diagnosis of diseases. In this regard, some knowledge of features' goodness in making predictions is needed. Apparently, distinctive features and their relevancy to class labels are determinant in designing efficient classifiers. Indeed, excluding redundant and/or irrelevant features, without incurring much loss of information, can reduce the processing cost while improving the predictor's performance. Consequently, feature selection is a preliminary task in most biological studies. Traditionally, biological data analysis methods also use the common feature selection techniques which imagine the data instances as independent objects and so not consider their possibly inter-relations. For instance, protein–protein interactions (PPIs) handle a wide range of biological processes including cell-to-cell interactions and metabolic and developmental control. Apparently, linked data have more similar characteristics than uncorrelated ones and so accounting these inter-relations beside to data content will be beneficial in feature selection. To incorporate the data inter-relations (e.g., PPIs in biological data) along with the data content in selecting more effective features, a novel feature selection algorithm is proposed. This method works in unsupervised manner to handle the unlabeled biological data since most of the real-world genes/proteins have no label. For this purpose, we try to optimize a novel objective function which incorporates both the inter-relations of data instances and their content. The proposed method tries to identify the most relevant and non-redundant features and extract the top-ranked ones. For this purpose, an efficient iterative algorithm is developed to optimize the objective function. To assess our methods, three well-known evaluation criteria are examined on some real-world biological datasets and the results are compared against some of the state-of-the-art feature selection methods. The experiments demonstrate the effectiveness of our proposed approach.

Keywords Feature selection · Unsupervised feature selection · Biological linked data · Protein–protein interaction

1 Introduction

A genetic disease is any disease that is caused by an abnormality in an individual's genome. The abnormality can range from minuscule to major; from a discrete mutation in a single base in the DNA of a single gene to a gross chromosome abnormality involving the addition or subtraction of an entire chromosome or set of chromosomes. Some genetic disorders are inherited from the parents, while other genetic diseases are caused by acquired changes or mutations in a preexisting gene or group of genes. Mutations can occur either randomly or due to some environmental exposure.

Contemporary classification of human disease dates to the late 19th century and derives from observational correlation between pathological analysis and clinical syndromes. Characterizing disease in this way established a classification schema that has served clinicians well to the current time, relying on observational skills to define the syndrome phenotype. Throughout the last century, this approach became more objective, as the molecular underpinnings of many disorders were identified and definitive laboratory tests became an essential part of the overall diagnostic paradigm [1].

In bioinformatics, various large projects, such as the human genome project, together with new techniques, such as the microarray, have created enormous amount of data. These data often come with high dimensionality so that they

✉ Elham Hoseini
hoseini-e@shirazu.ac.ir

Eghbal G. Mansoori
mansoori@shirazu.ac.ir

[1] Shiraz University, Shiraz, Iran

can involve a huge number of genes with many dimensions. This condition can significantly increase the computational burden, even to the extent that it renders some data mining approaches impossible. For example, it would be very difficult to train a neural network or support vector machine with tens of thousand input nodes. Furthermore, many of these tremendous input features are redundant and/or irrelevant to a given task and can act like noise to decrease performance. Feature selection [3, 14] is a useful technique since it can help alleviate the curse of dimensionality, speed up the learning process, and provide better interpretability.

Network data have become increasingly popular in the past decades, because of the proliferation of various social and information networks. Social networks such as Facebook and Twitter have millions of users all across the world. Different forms of information networks such as co-author networks, citation networks, and protein interaction networks also attract considerable research attention [4, 5]. In addition to the link structure, these network data are usually accompanied with content information on the nodes. For example, one can extract thousands of profiling features for users in social networks or ontology features for genes in protein interaction networks.

Proteins rarely act alone as their functions tend to be regulated. Many molecular processes within a cell are carried out by molecular machines that are built from a large number of protein components organized by their protein–protein interactions (PPIs). Direct PPIs are one of the strongest manifestations of a functional relation between genes/proteins, so interacting proteins may lead to the same disease phenotype when mutated. Protein–protein interactions refer to lasting or ephemeral physical contacts of high specificity established between two or more protein molecules as a result of biochemical events steered by electrostatic forces including the hydrophobic effect. These interactions make up the so-called interactomes of the organism, while aberrant PPIs are the basis of multiple aggregation-related diseases. A recent study showed that interacting proteins tend to lead to similar disease phenotypes when mutated. Therefore, protein–protein interactions might in principle be used to identify potentially interesting disease gene candidates.

Accordingly, we have incorporated PPI networks in feature selection process. This is because two linked proteins are more likely to have similar properties than two randomly picked ones. Using this network information along with the features themselves, we have tried to select more discriminative features of proteins. However, in most of the existing feature selection methods on gene/protein data, they seldom consider their inter-relations. This is due to lack of relationship information among instances in most biological datasets.

On the other hand, in the genomics setting, an increasingly common data configuration consists of a small set of sequences possessing a targeted property (positive instances) among a large set of sequences for which class label is unknown. Therefore, our proposed feature selection method tries to work in unsupervised manner.

By coinciding the link information of unlabeled proteins besides their abundant features, we have proposed an Unsupervised Feature Selection Framework for Linked Biological data (UFLB), in order to facilitate hereditary disease genes classification. For this purpose, we try to optimize a novel objective function via an efficient iterative algorithm in order to identify the most relevant and non-redundant features.

The rest of this paper is arranged as follows. The related work is presented in Sect. 2. Our new framework for unsupervised feature selection in biological data, UFLB, is introduced in Sect. 3, including approaches to capture protein–protein interactions, clustering the proteins iteratively, and optimization analysis. The experimental results with discussion are presented in Sect. 4. Finally, we conclude this work in Sect. 5.

2 Related work

Feature selection is an important operation in processing the data stored in gene microarrays. The most relevant features increase our understanding of the mechanism of disease formation and allow to predict the potential danger of being affected by such disease. The application of feature selection methods allows to identify a subset of important features that can be used as biomarkers of the appropriate disease. In the following, we introduce some related work on feature selection for both non-linked and linked data.

2.1 Feature selection for non-linked data

Recently, many learning techniques have been proposed to solve the problem of feature selection. It is certainly worth mentioning a number of methods that have emerged empirically for their effectiveness. One of the differences among various feature selection procedures is the way they perform the search in the feature space. Three categories of feature selection methods can be distinguished as follows: filter [7, 13], wrapper [8], and embedded methods [9, 10].

Filter methods assess features by calculating a relevant score for each one of them. The low-relevant features are then removed, and the selected features may then be used to serve classification via many types of classifiers. Feature selection filter-based methods can scale easily to high-dimensional datasets since they are computationally simple and fast compared with the other approaches. Various examples for filter-based approaches are ReliefF [11], mRMR [12], SPEC [13], Laplacian score [14], and its extensions [13].

Wrapper methods evaluate feature subsets using a predictive model which is run on the dataset partitioned into training and testing sets. Each subset is used with training dataset to train the model, which is then tested on the test set. Calculating a model prediction error from the test set gives a score for that feature subset. The subset with the highest evaluation is selected as the final set on which to run this particular model. The wrapper methods are computationally expensive since they need a new model to be fitted for each subset [15, 16]. In the embedded models, however, feature selector is a combination of both filter and wrapper. They are less computationally expensive than the wrapper methods.

Depending on the availability of class labels, feature selection algorithms can be categorized into supervised methods and unsupervised methods [8]. In the supervised methods such as Fisher score [17] and ReliefF [11], class labels provide a clear guidance to the feature selection process. This is because supervised methods usually are more reliable than unsupervised ones. However, these methods suffer from two main restrictions. First, since they evaluate each feature independently, they ignore the correlation between features. Second, access to labeled training data in real world is too expensive. Nevertheless, much attention has been paid to unsupervised feature selection in recent years.

Unsupervised feature selection becomes more challenging problem due to the absence of class labels. Unsupervised filter methods usually assign each feature a score which can indicate the feature's capacity to preserve the structure of data. Top-ranked features are selected since they can best preserve the structure of data. The typical methods include maximum variance [18], Laplacian score [14], and SPEC [13]. Unsupervised wrapper methods [19] require a learning algorithm to evaluate the candidate feature subsets. Unsupervised embedded methods perform feature selection as a part of model training process, e.g., UDFS [20] and NDFS [21].

State-of-the-art approaches introduce the notion of pseudo-labels [20–22] to guide the feature selection process. Unsupervised Discriminative Feature Selection (UDFS) [20] introduces pseudo-labels to better capture the discriminative information, and the sparsity-inducing $l_{2,1}$ norm is used to select features in an iterative manner. NDFS [21] performs non-negative spectral analysis and feature selection simultaneously.

The basic idea is to imitate supervised methods by generating pseudo-labels via certain clustering methods (e.g., spectral clustering and non-negative matrix factorization) and performing sparse regression toward these cluster labels. However, the generated pseudo-labels are usually inaccurate and could further mislead the feature selection process.

2.2 Feature selection for linked data

Traditional feature selection approaches assume that data instances are independent and identically distributed (i.i.d). Several methods have been proposed in recent years in which the relationships among data are also considered. In the network data, however, instances are implicitly or explicitly related to certain correlations and dependencies. For example, in research collaboration networks, researchers who collaborate with each other (i.e., connections in the network) tend to share more similar research topics (i.e., close distances in the feature space) than researchers without such collaboration. Most existing feature selection approaches fail to exploit the rich information contained in the links.

In [23], a supervised feature selection algorithm (called FSNet) is proposed. It adopts linear regression to fit the content information. Moreover, it uses graph regularization to capture the link information. On the other hand, LinkedFS [24] selects features in social media data in a semi-supervised manner. A supervised feature selection framework, CoSelect, for social media data is proposed in [25]. Instance selection is incorporated into feature selection in CoSelect in order to select relevant instances while selecting features simultaneously.

Linked unsupervised feature selection (LUFS) [26] is an unsupervised feature selection method that utilizes both content and link information. LUFS exploits network information through incorporating social dimension-based regularization [27] into the UDFS framework [20]. It enforces the nodes within the same social dimension to have similar pseudo-labels. But the social dimensions generated from links (e.g., by modularity [28] or spectral clustering [29]) and pseudo-labels generated from attributes are usually far from accurate, which could mislead the feature selection process.

In our previous work, an unsupervised feature selection method in social media data (called UFSS) is presented [43]. UFSS incorporates the inter-relationship of objects in addition to their feature values. By using graph partitioning, the objects are labeled and then are applied in the objective function. An iterative algorithm is designed to optimize the proposed objective function. However, in UFSS, the labeling of objects is a pre-processing step and these labels do not change during the algorithm's run; which is a constraint.

In this paper, however, the labels of objects are assigned in a dynamic manner. Unlike UFSS and LUFS which use graph partitioning and social dimension incorporation for static labeling, the proteins are labeled dynamically in the consecutive iterations of UFLB so that, after its convergence, an appropriate clustering of proteins is achieved. Our unsupervised feature selection method for linked biological data takes into account both inter-protein relationship information and feature content

of proteins. It tries to select some features which effectively discriminate proteins in the reduced space by using PPIs.

3 The proposed method

Our proposed approach is categorized in hybrid methods since combines both filter and wrapper methods. In this section, we present several concepts as preliminaries of our unsupervised feature selection method. We aim to select a set of effective features which can highly discriminate the protein classes.

3.1 Notations

In this work, we use $P = \{p_1, p_2, \ldots, p_n\}$ to denote the set of n proteins and $F = \{f_1, f_2, \ldots, f_m\}$ the set of m features. Also, let $A \in \mathcal{R}^{m\times n}$ holds the feature values of these proteins. That is, the vector $A(:,j)$ represents the features of protein and $A(i,:)$ is the values of feature f_i in all proteins. Additionally, $R \in \mathcal{R}^{n\times n}$ denotes the link information for protein–protein network where $R(i,j)$ is set to 1 if protein p_i and p_j are linked and 0 otherwise. We imagine there are undirected connections between proteins, that is, $R = R^T$. By applying the centering matrix $H = I_n - \frac{1}{n}1_n 1_n^T$ on A via $X = AH$, we obtain the data matrix $X \in \mathcal{R}^{m\times n}$. This matrix is centered, that is, $\sum_{j=1}^{n} X(:,j) = 0$. In H, I_n is the identity matrix and 1_n is a column vector of n ones.

3.2 Unsupervised feature selection for linked biological data

Supposing the n proteins are sampled from c classes/clusters, we assume that there is a mapping matrix $M \in \mathcal{R}^{m\times c}$ which assigns the proteins with a cluster label indicator matrix $C \in \mathcal{R}^{c\times n}$. In this matrix, $C(:,i) \in \{0,1\}^{c\times 1}$ represents the cluster indicator vector for protein p_i. To use its scaled version $G(:,i)$, we define the scaled cluster indicator matrix $G \in \mathcal{R}^{c\times n}$ where $G = \left(CC^T\right)^{-\frac{1}{2}} C$ [30] and $GG^T = \left(CC^T\right)^{-\frac{1}{2}} CC^T \left(CC^T\right)^{-\frac{1}{2}} = I_c$.

Accordingly, our aim was to learn the scaled cluster indicator matrix G and the feature selection matrix M simultaneously. In this regard, we propose to optimize the following objective function:

$$\min_{M} \left\| M^T X - G \right\|_F^2 + \lambda \|M\|_{2,1} \tag{1}$$

where $\|.\|_F^2$ is the Frobenius norm [32] and $\|M\|_{2,1}$ is the $l_{2,1}$ -norm of M [31] which controls the capacity of this matrix. The parameter λ is used to control the sparsity of M. Due to the nature of the $l_{2,1}$-norm penalty, some coefficients will be shrunk to exact 0 if λ is large enough.

In (1), M essentially contains the combination coefficients for different features in approximating G. The joint minimization of the regression model and $l_{2,1}$-norm regularization term enables M to evaluate the correlation between cluster indicator and features. Also, minimizing $\|M\|_{2,1}$ ensures that M is sparse in rows. These reasons, altogether, make M particularly suitable for feature selection.

By considering matrix R and the fact that linked proteins are likely to have similar cluster label indicator, we are going to minimize the following term:

$$\min_{G} \frac{1}{2} \sum_{i,j=1}^{n} R_{ij} G\|(:,i) - G(:,j)\|_2^2 = Tr\left[GLG^T\right] \tag{2}$$

where $L = D - R$ is a Laplacian matrix and D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n} R_{ij}$ on diagonal elements. Including (2) in (1), we obtain the new version of objective function:

$$\min_{M,G} \left\| M^T X - G \right\|_F^2 + \lambda \|M\|_{2,1} + Tr\left[GLG^T\right] \tag{3}$$

According to the definition of G, its elements are constrained to be discrete values, making the optimization of (3) an NP-hard problem [33]. A well-known solution is to relax it from discrete values to continuous ones [33, 34], so the objective function in (3) is relaxed to:

$$\min_{M,G} \left\| M^T X - G \right\|_F^2 + \lambda \|M\|_{2,1} + Tr\left[GLG^T\right]. \tag{4}$$

$$s.t. \quad GG^T = I_c$$

The last part of our objective function is formed by taking into account the (centered) protein information matrix X for discrimination. A well-known method to utilize discriminative information is to find a low-dimensional subspace in which the between-class scatter matrix Q_b is maximized while minimizing the total scatter matrix Q_t [35].

As in [43], the maximum of $Tr\left(\frac{Q_b}{Q_t}\right)$ (minimum of its negative) is included in (4) and the new objective function is given by:

$$\min_{M,G} \left\| M^T X - G \right\|_F^2 + \lambda \|M\|_{2,1} + Tr\left[GLG^T\right] - \gamma Tr\left(\frac{Q_b}{Q_t}\right) \tag{5}$$

$$s.t. \quad GG^T = I_c$$

where parameter γ controls the discrimination value. In order to use the definitions of Q_b and Q_t in [43], we define $Y = M^T X$ and so:

$$\min_{M,G} \left\|M^T X - G\right\|_F^2 + \lambda\|M\|_{2,1} + Tr\left[GLG^T\right] + \gamma Tr\left(YY^T - YG^T GY^T\right) \tag{6}$$

$$s.t. \quad GG^T = I_c$$

Note that all the elements of G are non-negative by definition. However, the optimal G of (6) has mixed signs which violates its definition and makes G severely deviate from the ideal cluster indicators. As a result, we cannot directly assign labels to data using the cluster indicator matrix G. To address this problem, it is reasonable to impose a non-negative constraint into the objective function. When both non-negative and orthogonal constraints are satisfied, there is only one positive element in each row of G, while all others are zero. In that way, the learned G is more accurate and more capable to provide discriminative information. Therefore, by rewriting (6), the new objective function is obtained as follows:

$$\min_{M,G} \left\|M^T X - G\right\|_F^2 + \lambda\|M\|_{2,1} + Tr\left[GLG^T\right] + \gamma Tr(M^T X\left(I_n - G^T G\right)X^T M \tag{7}$$

$$s.t. \quad GG^T = I_c \text{ and } G \geq 0$$

To optimize this function, we propose an iterative optimization algorithm. In this regard, we rewrite the objective function in (7) as follows:

$$\min_{M,G} \left\|M^T X - G\right\|_F^2 + \lambda\|M\|_{2,1} + Tr\left[GLG^T\right] + \gamma Tr(M^T X\left(I_n - G^T G\right)X^T M + \alpha\left\|GG^T - I\right\|_{cF}^2 \tag{8}$$

$$s.t. \quad G \geq 0$$

where $\alpha > 0$ is a parameter to control the orthogonality condition. In practice, α should be large enough to insure the orthogonality satisfied. For the ease of representation, the last objective function $J(M, G)$ is defined as follows:

$$J(M,G) = \left\|M^T X - G\right\|_F^2 + \lambda\|M\|_{2,1} + Tr\left[GLG^T\right] + \gamma Tr(M^T X\left(I_n - G^T G\right)X^T M + \alpha\left\|GG^T - I\right\|_{cF}^2 \tag{9}$$

Theorem 1 The mapping matrix M in $J(M, G)$ can be updated as follows:

$$M^{(\text{new})} = \left(XX^T + \gamma X\left(I_n - G^T G\right)X^T + \lambda D_M^{(\text{old})}\right)^{-1} XG^T \tag{10}$$

where D_M is an $m \times m$ diagonal matrix with $\frac{1}{\|2M(i,:)\|_2}$ on its i th row.

Proof In order to minimize $J(M, G)$ in (9), its derivative is taken as follows:

$$\frac{\partial J(M,G)}{\partial M} = 2XX^T M - 2XG^T + 2\lambda D_M M + 2\gamma X\left(I_n - G^T G\right)X^T M$$

By setting this derivative to zero, the update rule in (10) is obtained. □

Theorem 2 The scaled cluster indicator matrix, G, is updated by this rule:

$$G_{ij}^{(\text{new})} = G_{ij}^{(\text{old})} . \frac{U_{ij}^{(\text{old})}}{V_{ij}^{(\text{old})}} \tag{11}$$

where $U = M^T X + M^T XG^T M^T X + 2\alpha G$ and $V = G + GL + 2\alpha GG^T G$.

Proof Following [36–38], we introduce multiplicative updating rules. Setting derivative of $J(M, G)$ with respect to G_{ij} to 0 and using the Karush–Kuhn–Tucker condition [39], we obtain the updating rule in (11). □

Using the updating rule of M in (10) and of G. in (11), we have developed the iterative algorithm of UFLB:

Algorithm: UFLB

Inputs:

X: feature-protein matrix, $X \in \mathcal{R}^{m \times n}$
R: protein-protein relationship matrix, $R \in \mathcal{R}^{n \times n}$
λ: parameter to control sparsity
γ: parameter to control discrimination
α: parameter to control orthogonality condition
ε: parameter of stopping condition
c: number of clusters
q: number of required features

Output: list of q best features

1. Initialize G randomly or use a prior knowledge (e.g. use graph partitioning)
2. Initialize the diagonal matrix $D_M = I_m$
3. Initialize the mapping matrix $M = (XX^T + \gamma X(I_n - G^T G)X^T + \lambda D_M)^{-1} XG^T$
4. Compute the objective function $J(M, G)$ using (9)
5. Repeat
6. Update D_M via computing its i^{th} diagonal element using $\frac{1}{2\|M(i,:)\|_2}$
7. Update M using (10)
8. Update G using (11)
9. Compute $J(M, G)$ using (9)
10. Until $J^{(\text{old})} - J^{(\text{new})} < \varepsilon$
11. Rank, in descending order of $\|M(i,:)\|_2$, the features f_i $(i = 1, \ldots, m)$
12. Return q top-ranked features

The larger the norm of $\|M(i,:)\|_2$, the more informative the feature f_i is.

4 Experiments and discussion

In this section, we present experiment details to verify the effectiveness of the proposed framework, UFLB. In this regard, it is compared against the state-of-the-art unsupervised feature selection with/without link information. We evaluate the effectiveness of the selected features using both accuracy measure and clustering quality. Then, the effects of parameters on performance of UFLB are discussed. At last, its convergence analysis is conducted via experiments.

4.1 Datasets

In this work, some labeled genes from Online Mendelian Inheritance in Man (OMIM) with six groups of confirmed diseases are selected. The labels are cardiovascular disease, endocrine disease, cancer disease, metabolic disease, neurological disease, and ophthalmological disease [40, 45]. With respect to the quality and the performance of disease gene classification methods, the data are derived from multiple biological sources [41]. This dataset consists of 949 genes with 4004 features and 956 links. The features are extracted from gene ontology (3000 features), protein domain (1000 features), and protein–protein interactions (4 features) to construct the feature vector of each gene [45]. We have reduced the dataset to uncover the features which none of the genes contain them. So, the dataset is reduced to 3522 features.

The second dataset consists of a subset of IntAct[1] with three groups of diseases. IntAct provides an open source database and toolkit for the storage, presentation, and analysis of protein interactions. We extract 846 genes/proteins from cancer, Alzheimer, and Parkinson databases with 1980 links. The sequence of each gene/protein is obtained from UniProt[2] database. Then, using the distribution of 1 gram, 2 grams and 3 grams in each protein sequence, 8420 features are extracted from the combinations of amino acids [44]. By reducing the dataset to uncover the features which none of the genes/proteins contain them, the dataset is reduced to 8404 features.

A subset of HPRD[3] database is selected as third dataset. This dataset contains 234 genes from four disease classes: diabetes, myopathy, syndrome, and cancer. Each gene has 8420 features which are extracted from the combinations of amino acids. In our experiments, a small number of samples

[1] http://www.ebi.ac.uk/intact/.
[2] http://www.uniprot.org/uniprot/.
[3] http://www.hprd.org/.

Table 1 Statistics of four linked biological datasets

Dataset	No. of genes/proteins	No. of features	No. of links	No. of classes
OMIM	949	3522	956	6
IntAct	846	8404	1980	3
HPRD	234	8420	716	4
Hetio	966	127	2124	6

(genes) versus too many features are selected to evaluate UFLB.

The fourth dataset contains fewer features than the three datasets. In this dataset, 966 genes with 127 features are used. Its six classes, Parkinson, Alzheimer, vitiligo, chronic lymphocytic leukemia, schizophrenia, and type I diabetes mellitus, are extracted from Hetio[4] database.

The statistics of datasets is shown in Table 1.

4.2 Evaluation measures

Following the convention of clustering study, we have evaluated the clustering quality of UFLB by two commonly used metrics, Unsupervised Accuracy Measure (UAM) and Normalized Mutual Information (NMI).

Denoting q_i as the clustering result and g_i as the ground-truth label of protein p_i, UAM is computed as [20]:

$$\mathrm{UAM} = \frac{1}{n}\sum_{i=1}^{n}\delta\big(g_i, map(q_i)\big) \tag{12}$$

where

$$\delta(g,q) = \begin{cases} 1, & if\ g = q \\ 0, & if\ g \neq q \end{cases} \tag{13}$$

and $map(q)$ is the best mapping function that permutes clustering labels to match the ground-truth labels using the Kuhn–Munkres algorithm. The larger UAM indicates the better performance.

Also, NMI is computed as [42]:

$$\mathrm{NMI} = \sum_{l}\sum_{h} t_{l,h}\log\left(\frac{n\times t_{l,h}}{t_l \times t_h}\right) \Big/ \sqrt{\left(\sum_{l} t_l \log\left(\frac{t_l}{n}\right)\right)\left(\sum_{h} t_h \log\left(\frac{t_h}{n}\right)\right)} \tag{14}$$

where t_l is the number of proteins in l th cluster ($l = 1, \ldots, c$)., and t_h is the number of proteins in h th ground-truth class ($h = 1, \ldots, c$). Also, $t_{l,h}$ is the number of proteins in both l th cluster and h th ground-truth class. Agn, higher values of NMI, report the better clustering results.

4.3 Experimental results

In this subsection, we compare the quality of features, selected by different algorithms, using NMI and accuracy metrics. For baseline methods with some parameters, we have tried different parameter values and reported the best performance. UFLB has three important parameters: γ, λ, and α which control the discriminative information, sparsity, and orthogonality, respectively. To select the best parameters, each one is set while holding the others fixed to see how accuracy of UFLB varies when different number of features is selected. Figure 1 depicts the NMI measure when three parameters of UFLB are examined for OMIM dataset.

As shown in Fig. 1, it is clear that for $\alpha > 0.5$, NMI values degrade. In the case of γ, the values are closer to each other though for $\gamma = 0.9$, they are least. Also for λ, the values of NMI decrease when λ decreases. However, for $\lambda = 0.9$, NMI reduces drastically. Consequently, setting the parameters of orthogonality, sparsity, and discrimination to high values cannot generate the desired results. At last, these parameters are set as $\alpha = 0.5$, $\gamma = 0.7$, and $\lambda = 0.3$ in the experiments on OMIM dataset. Similarly, for three other datasets, these settings are achieved: $\alpha = 0.3$, $\gamma = 0.5$, $\lambda = 0.3$ for IntAct; $\alpha = 0.1$, $\gamma = 0.2$, $\lambda = 0.3$ for HPRD, and $\alpha = 0.3$, $\gamma = 0.1$, $\lambda = 0.1$ for Hetio.

After setting appropriately the parameters of UFLB, its performance is compared against seven unsupervised feature selection algorithms in terms of UAM and NMI metrics. These methods are described briefly here.

(1) UDFS is proposed in [20] to optimize the $l_{2,1}$-norm regularized minimization problem with orthogonal constraint. It selects the most discriminative feature subset from the whole feature set in batch mode.
(2) UFSS [43] utilizes both the relationship between instances and information of features to propose an objective function. This function seeks for a mapping matrix in which the discriminative information of each feature exists. Finally, the ranked features are obtained by utilizing this mapping matrix.
(3) NMF [37] is a matrix factorization method which approximately decomposes a known matrix into two unknown matrices with much lower dimensions.

[4] http://het.io/disease-genes/.

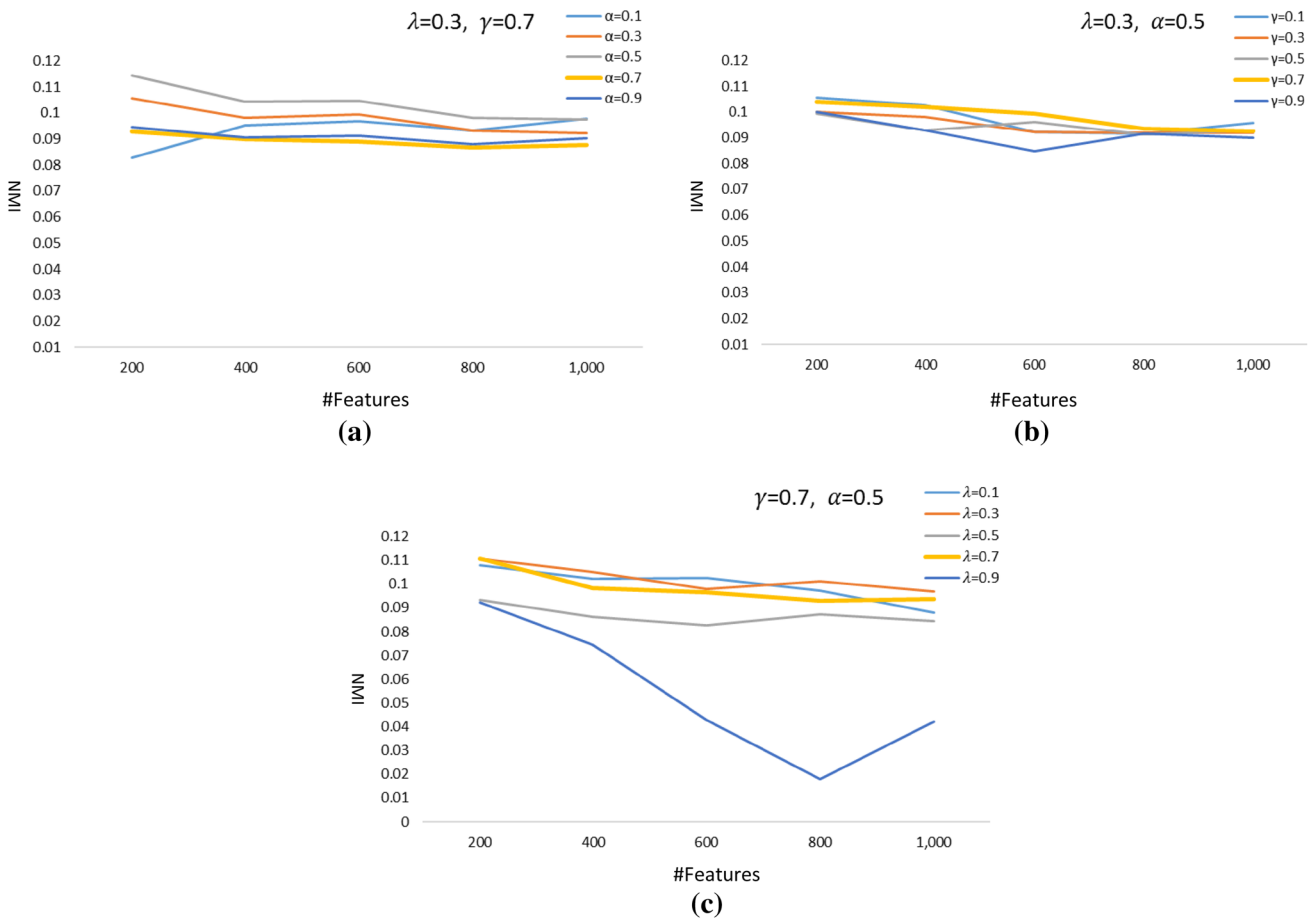

Fig. 1 NMI measure with different values of **a** α, **b** γ, and **c** λ

Table 2 NMI measure of different feature selection methods for OMIM dataset

Method	Number of selected best features						3522
	200	400	600	800	1000	2000	
UFLB	**0.1186**	**0.1169**	**0.1181**	**0.1061**	0.0965	0.0848	0.0795
UFSS	0.0446	0.0944	0.0859	0.0854	0.0870	0.0750	
UDFS	0.0372	0.0406	0.0398	0.0428	0.0459	0.0547	
SPEC	0.0421	0.0454	0.0403	0.0403	0.0441	0.0765	
Laplacian	0.0184	0.0267	0.0621	0.0759	0.0667	0.0694	
CGSSL	0.1075	0.0889	0.0882	0.0938	**0.0976**	0.0818	
NMF	0.0652	0.0817	0.0866	0.0845	0.0831	0.0880	
LUFS	0.0796	0.0881	0.0872	0.0828	0.0884	**0.0955**	

Bold values represent the best result compare to other results

(4) Laplacian score [14] is based on the observation that, in many real-world classification problems, data from the same class are often close to each other. The importance of a feature is evaluated by its power of locality preserving via Laplacian score.

(5) LUFS [2] is an unsupervised feature selection framework for linked data in social media. This framework utilizes a concept of social dimensions from social network analysis to extract relations among linked data as groups. Then, it defines a social dimension regularization inspired by linear discriminant analysis to mathematically model these relations.

(6) SPEC [13] is a general framework of spectral feature selection for both supervised and unsupervised learning. It is based on sparse multi-output regression by considering $l_{2,1}$-norm. This algorithm performs well in both redundant features removing and relevant features preserving.

Table 3 UAM of different feature selection methods for OMIM dataset

Method	Number of selected best features						3522
	200	400	600	800	1000	2000	
UFLB	**0.3404**	0.3267	**0.3298**	**0.3171**	**0.3120**	**0.3446**	0.3135
UFSS	0.3219	**0.3305**	0.3259	0.3151	0.3118	0.3213	
UDFS	0.2544	0.2551	0.2430	0.2512	0.2745	0.2822	
SPEC	0.2450	0.2473	0.2408	0.2507	0.2701	0.2880	
Laplacian	0.2415	0.2786	0.2976	0.2932	0.2900	0.3020	
CGSSL	0.3066	0.2982	0.2972	0.2961	0.3066	0.2856	
NMF	0.2898	0.3024	0.2845	0.2761	0.2782	0.2929	
LUFS	0.2613	0.2782	0.2929	0.3035	0.2972	0.2887	

Bold values represent the best result compare to other results

Table 4 Classification accuracy of different feature selection methods for OMIM dataset

Method	Number of selected best features						3522
	200	400	600	800	1000	2000	
UFLB	**0.6196**	**0.6812**	0.7160	0.7544	**0.7981**	**0.8661**	0.9020
UFSS	0.5653	0.6668	0.7260	0.7412	0.7653	0.8165	
UDFS	0.3898	0.4278	0.4373	0.4942	0.5057	0.7323	
SPEC	0.4836	0.5110	0.5468	0.5690	0.6122	0.8061	
Laplacian	0.3466	0.4351	0.4762	0.5395	0.5911	0.7249	
CGSSL	0.5932	0.6585	0.7070	0.7523	0.7839	0.8514	
NMF	0.5047	0.5268	0.6111	0.6406	0.6701	0.8155	
LUFS	0.5911	0.6680	**0.7260**	**0.7618**	0.7829	0.8351	

Bold values represent the best result compare to other results

Table 5 NMI measure of different feature selection methods for IntAct dataset

Method	Number of selected best features						8404
	500	1000	1500	2000	4000	6000	
UFLB	**0.0239**	0.0095	**0.0121**	**0.0181**	**0.0183**	0.0107	0.0094
UFSS	0.0084	**0.0124**	0.0094	0.0087	0.0050	0.0150	
UDFS	0.0073	0.0084	0.0071	0.0071	0.0071	0.0085	
SPEC	0.0078	0.0094	0.0079	0.0094	0.0082	0.0107	
Laplacian	0.0097	0.0106	0.0078	0.0087	0.0101	0.0123	
CGSSL	0.0102	0.0122	0.0066	0.0086	0.0147	0.0083	
NMF	0.0071	0.0073	0.0072	0.0072	0.0073	0.0071	
LUFS	0.0205	0.0105	0.0096	0.0046	0.0052	**0.0197**	

Bold values represent the best result compare to other results

(7) CGSSL [6] jointly exploits non-negative spectral analysis and structural learning with sparsity. In this unsupervised feature selection approach, the cluster indicators, learned by non-negative spectral clustering, are used to provide label information for the structural learning.

The comparison results *w.r.t* both UAM and NMI are demonstrated in Table 2 and Table 3 for OMIM dataset. In these tables, q best features of each method are examined. Moreover, the clustering performance with all features (i.e., without feature selection) is also reported. Note that the results of each evaluation criterion are reported in two forms: in tabular (left) and in plot (right).

According to results in Table 2 and 3, UFLB mostly outperforms seven compared methods. Although NMI's results for UFLB are much better than other methods, its values are generally small. This is because those proteins, which are placed in a ground-truth class, are not necessarily grouped in the same cluster. This in turn leads to a noticeable decrease in NMI. This also occurs for UAM.

In order to evaluate the accuracy measure more reliable, ground-truth labels are considered. By utilizing these labels, Multi-class Support Vector Machine (SVM) classifier [46] is trained and the classification accuracy is calculated according to the obtained results in Table 4. As depicted here, the accuracy of UFLB is near to LUFS and

Table 6 UAM of different feature selection methods for IntAct dataset

Method	Number of selected best features						8404
	500	1000	1500	2000	4000	6000	
UFLB	**0.3853**	**0.3830**	0.3690	**0.3917**	0.3700	**0.3747**	0.3700
UFSS	0.3775	0.3570	0.3629	0.3593	0.3700	0.3725	
UDFS	0.3570	0.3700	0.3569	0.3700	0.3688	0.3700	
SPEC	0.3688	0.3725	0.3582	0.3570	0.3672	0.3700	
Laplacian	0.3383	0.3582	0.3441	0.3476	0.3547	0.3606	
CGSSL	0.3593	0.3593	0.3511	0.3546	**0.3759**	0.3558	
NMF	0.3582	0.3569	0.3570	0.3570	0.3570	0.3570	
LUFS	0.3558	0.3712	**0.3696**	0.3697	0.3712	0.3695	

Bold values represent the best result compare to other results

Table 7 Classification accuracy of different feature selection methods for IntAct dataset

Method	Number of selected best features						8404
	500	1000	1500	2000	4000	6000	
UFLB	**0.7133**	0.6733	**0.6866**	**0.6800**	**0.6766**	0.6466	0.6333
UFSS	0.6766	**0.6866**	0.6666	0.6533	0.6533	0.6533	
UDFS	0.6266	0.6400	0.6600	0.6533	0.6466	0.6400	
SPEC	0.5133	0.6266	0.5266	0.5866	0.5733	0.5333	
Laplacian	0.6066	0.6400	0.5733	0.6066	0.5666	0.5933	
CGSSL	0.6866	0.6600	0.6600	0.6733	0.6600	0.6533	
NMF	0.6733	0.5933	0.6600	0.6533	0.6466	0.6400	
LUFS	0.6800	0.6800	0.6733	0.6600	0.6600	**0.6600**	

Bold values represent the best result compare to other results

Table 8 NMI measure of different feature selection methods for HPRD dataset

Method	Number of selected best features						8420
	500	1000	1500	2000	4000	6000	
UFLB	0.1332	**0.1323**	**0.1324**	**0.1346**	0.1295	0.1295	0.1303
UFSS	0.1317	0.1275	0.1295	0.1323	**0.1323**	**0.1323**	
UDFS	0.1204	0.1295	0.1227	0.0989	0.1060	0.0989	
SPEC	0.1022	0.1216	0.1241	0.1317	0.1295	0.0989	
Laplacian	**0.1397**	0.1235	0.1205	0.1205	0.1096	0.1096	
CGSSL	0.0968	0.0989	0.1060	0.0837	0.0837	0.0847	
NMF	0.1204	0.1204	0.1204	0.1204	0.0964	0.1204	
LUFS	0.1264	0.1258	0.1295	0.1302	0.1204	0.1295	

Bold values represent the best result compare to other results

Table 9 UAM of different feature selection methods for HPRD dataset

Method	Number of selected best features						8420
	500	1000	1500	2000	4000	6000	
UFLB	**0.4487**	**0.4444**	**0.4444**	**0.4487**	**0.4444**	**0.4444**	0.4444
UFSS	0.4389	0.4256	0.4265	0.4308	0.4269	0.4239	
UDFS	0.4226	0.4252	0.4145	0.3974	0.3718	0.3932	
SPEC	0.3761	0.3932	0.3632	0.3803	0.3974	0.3932	
Laplacian	0.3846	0.3675	0.3889	0.4359	0.3761	0.3803	
CGSSL	0.3932	0.3675	0.3718	0.3932	0.4188	0.4017	
NMF	0.4145	0.4017	0.3932	0.4060	0.4060	0.3718	
LUFS	0.4246	0.4275	0.3932	0.4017	0.3975	0.3918	

Bold values represent the best result compare to other results

Table 10 Classification accuracy of different feature selection methods for HPRD dataset

Method	Number of selected best features						8420
	500	1000	1500	2000	4000	6000	
UFLB	**0.3437**	**0.3437**	0.3750	**0.3750**	**0.4062**	**0.3958**	0.3437
UFSS	0.3125	0.3333	0.3437	0.3541	0.3541	0.3541	
UDFS	0.3125	0.3125	0.3125	0.3437	0.3541	0.3541	
SPEC	0.2708	0.3333	0.3541	0.3541	0.3333	0.3125	
Laplacian	0.3125	0.2916	0.2916	0.3125	0.3125	0.3125	
CGSSL	0.3333	0.3125	0.3125	0.3125	0.3125	0.3541	
NMF	0.3333	0.3333	**0.3958**	0.3541	0.3125	0.3125	
LUFS	0.3225	0.3325	0.3425	0.3525	0.3333	0.3541	

Bold values represent the best result compare to other results

Table 11 NMI measure of different feature selection methods for Hetio dataset

Method	Number of selected best features						127
	20	40	60	80	100	120	
UFLB	0.0696	**0.0761**	**0.0778**	**0.0756**	**0.0773**	0.0723	0.0768
UFSS	0.0619	0.0562	0.076	0.0734	0.0735	**0.0745**	
UDFS	0.0719	0.0608	0.0753	0.0723	0.0717	0.0743	
SPEC	0.0722	0.073	0.0738	0.073	0.0738	0.0727	
Laplacian	0.0609	0.0685	0.0676	0.0714	0.0723	0.0743	
CGSSL	0.0577	0.0744	0.0672	0.0736	0.0735	0.0739	
NMF	0.0741	0.071	0.07	0.0732	0.0735	0.0731	
LUFS	**0.0751**	0.0727	0.0754	0.0739	0.0741	0.0736	

Bold values represent the best result compare to other results

Table 12 UAM of different feature selection methods for Hetio dataset

Method	Number of selected best features						127
	20	40	60	80	100	120	
UFLB	0.2295	**0.2276**	**0.2307**	**0.2286**	**0.2254**	**0.2236**	0.2226
UFSS	**0.2429**	0.2205	0.2214	0.2249	0.2225	0.221	
UDFS	0.218	0.2215	0.2195	0.2217	0.2197	0.2217	
SPEC	0.2164	0.2197	0.2189	0.2209	0.2216	0.2203	
Laplacian	0.2236	0.2201	0.2236	0.2236	0.22	0.2206	
CGSSL	0.2153	0.2195	0.2215	0.2205	0.2184	0.2195	
NMF	0.2233	0.2201	0.2264	0.2209	0.2206	0.2205	
LUFS	0.2274	0.2195	0.2274	0.2254	0.2212	0.2219	

Bold values represent the best result compare to other results

UFSS in most cases. This is because these algorithms take into account the link information. However, the execution time of UFLB is less than both LUFS and UFSS.

In the same way, the NMI, UAM, and classification accuracy of eight methods for IntAct dataset are shown in Tables 5, 6, and 7, respectively.

It is clear that in most cases for IntAct dataset, UFLB is better than the other methods. It outperforms almost all traditional feature selection methods which do not take into account link information. Also, in comparison with UFSS and LUFS, our proposed method obtains the better results in most cases.

In Tables 8, 9, and 10, the NMI, UAM, and classification accuracy of eight feature selection methods are shown for HPRD dataset. Clearly, UFLB works the best in most cases, though the sample size is not large. This depicts that the performance of UFLB is acceptable in medium-sized datasets with huge dimensions.

The results of NMI, UAM, and accuracy on Hetio dataset are reported in Tables 11, 12, and 13 where UFLB is compared against seven unsupervised feature selection methods. According to these results, UFLB outperforms the other methods, in most cases, for this dataset with many genes and a few features, because of taking into account the link information.

Table 13 Classification accuracy of different feature selection methods for Hetio dataset

Method	Number of selected best features						127
	20	40	60	80	100	120	
UFLB	**0.6555**	**0.5400**	0.3888	0.2866	**0.2861**	**0.2555**	0.2500
UFSS	0.4611	0.3761	**0.3944**	**0.3000**	0.2722	0.2502	
UDFS	0.4166	0.3111	0.2777	0.2555	0.2444	0.2500	
SPEC	0.1444	0.1777	0.1722	0.1722	0.2277	0.2500	
Laplacian	0.4833	0.4333	0.3277	0.2888	0.2533	0.2500	
CGSSL	0.4166	0.4166	0.3222	0.2944	0.2722	0.2505	
NMF	0.1888	0.1833	0.1833	0.2611	0.2444	0.2516	
LUFS	0.5111	0.4888	0.3611	0.2866	0.2633	0.25	

Bold values represent the best result compare to other results

Table 14 CPU time (in seconds) of UFLB against seven methods, run on four datasets

Method	Dataset			
	OMIM	IntAct	HPRD	Hetio
UFLB	139	1160	704	2.2
UFSS	153	1590	1480	3.9
UDFS	700	4970	7010	3.2
SPEC	63	107	5	6.1
Laplacian	2	3	1	0.4
CGSSL	468	4223	4410	3.3
NMF	4	10	2	0.6
LUFS	762	5716	4520	2.6

Table 15 Classification accuracy of different feature selection methods for the smallest subset of selected features on four datasets

Method	Dataset			
	OMIM	IntAct	HPRD	Hetio
UFLB	0.6196	0.7133	0.3437	0.6555
UFSS	0.5653	0.6766	0.3125	0.4611
UDFS	0.3898	0.6266	0.3125	0.4166
SPEC	0.4836	0.5133	0.2708	0.1444
Laplacian	0.3466	0.6066	0.3125	0.4833
CGSSL	0.5932	0.6866	0.3333	0.4166
NMF	0.5047	0.6733	0.3333	0.1888
LUFS	0.5911	0.6800	0.3225	0.5111

4.4 Time complexity

To evaluate the time complexity of UFLB, all unsupervised feature selection methods are implemented in MATLAB R2017, and then, their required CPU time is compared in the experiments. The codes are run on a Core i7, 1.8 GHz CPU with 8 GB of memory in 64-bit Windows 10. Table 14 shows the CPU time of UFLB against seven compared methods when run on four datasets.

From the results, it is clear that UFLB is considerably faster than LUFS and UFSS which consider the link information. Also, it is faster than CGSSL and UDFS. However, UFLB is not faster than SPEC, Laplacian, and NMF since they do not use the link information in feature selection. Clearly, those methods which select the features by considering the link information between samples require more CPU time to process these communications to improve their performance.

4.5 Non-parametric test

In order to assess statistically the compared methods, we have used the non-parametric statistical test to justify the significant differences among them. Friedman's test [47] with confidence level of 0.05 is used in the experiment.

Table 16 Average rankings of the compared methods

Method	Ranking
UFLB	8.0
LUFS	6.0
CGSSL	6.0
UFSS	4.5
NMF	4.1
Laplacian	3.0
UDFS	2.8
SPEC	1.5

This test is usually applied to show any significant difference among more than two results.

In Table 15, the classification accuracy of all algorithms for the smallest subset of selected features (as reported in Tables 4, 7, 10, and 13) on four datasets is displayed. We use the Friedman's test on these accuracies to examine the rejection of the hypothesis that all the feature selection methods perform equally well for a given level. It ranks the methods for each dataset separately, and the best performing method gets the highest rank.

By applying the Friedman's non-parametric test, we get the p value < 0.01. It can be concluded that at least two of the algorithms are significantly different from each other.

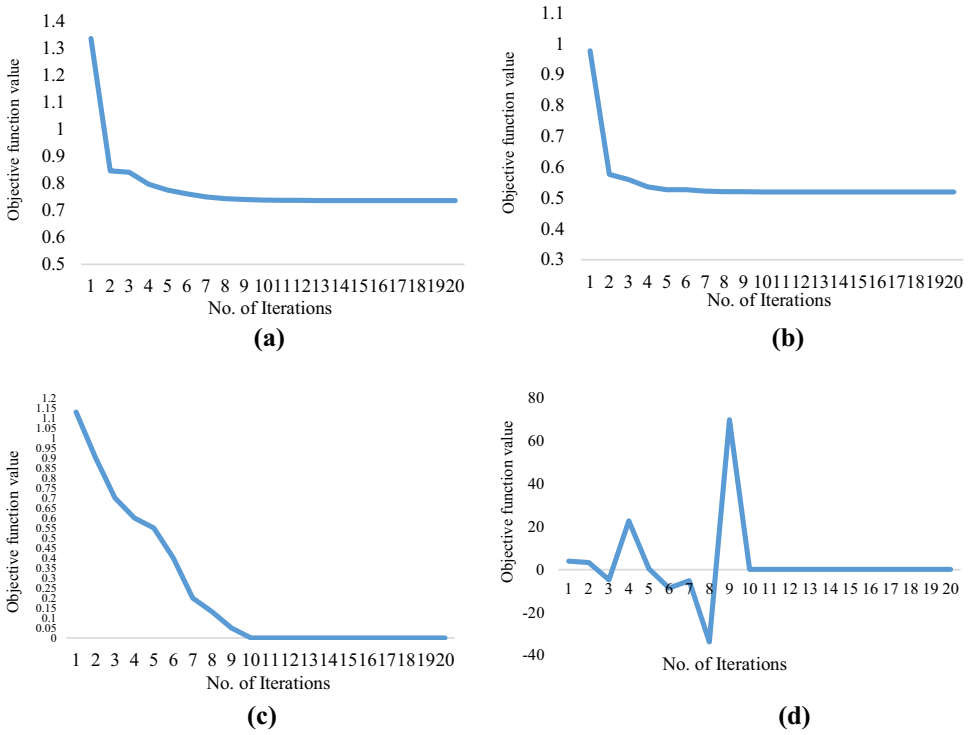

Fig. 2 Convergence curves of UFLB for **a** OMIM, **b** IntAct, **c** HPRD, and **d** Hetio datasets

Average rankings of the eight methods on four biological datasets, examined by Friedman's test, are shown in Table 16. These rankings reveal that UFLB is the most influential for classification tasks.

5 Discussion

In biological studies, it is important to know that which features can better discriminate the groups of hereditary diseases. In this subsection, we explain the nature of selected features in four datasets. First, by examining the top-ranked features for OMIM dataset, we found that 1000 top features are related to gene ontology (specifically biological process subontology). According to [46], gene ontology includes three subontologies: molecular function as the elemental activities of a gene product at the molecular level, biological process as a set of molecular functions, and cellular components which represent some parts of a cell or its extracellular environment. So, the features which are derived from biological process can do better discrimination.

Similarly, for IntAct and HPRD datasets, the top-ranked features are investigated. Since the large number of features is related to 3-gram distribution, compared to 1 gram and 2 grams, it is rational that a considerable portion of top-ranked features belong to this category. In IntAct dataset, among 1000 top features, 202 features are from 2-gram and only 3 features belong to 1-gram distribution. This means that the 1-gram features are not discriminative in disease genes.

Furthermore for Hetio dataset, two groups of features are extracted: (1) the features computed for each gene–disease pair and (2) the processed version of the GNF BodyMap [47] providing a gene's expression value for 77 specific tissues which can do discrimination more precise.

As stated before, UFLB mostly outperforms the other methods. Usually, for small number of best features in four datasets, UFLB is the best method. This means that UFLB is able to recognize the top discriminative features in comparison with the other methods. It is worth mentioning that the methods, which consider link information, usually outperform the other methods. This is because the interacted genes/proteins have more similar characteristics than uncorrelated ones. Thus, it is beneficial to incorporate PPIs beside to features in the feature selection process.

5.1 Convergence study

In this part, we justify it experimentally via plotting the convergence speed. Figure 2 shows the value of objective function, in (9), in consecutive iterations of algorithm. From these plots, it is clear that UFLB converges only after few iterations in four datasets. This justifies that the algorithm of UFLB converges certainly and quickly.

6 Conclusion

Classification of hereditary disease genes/proteins plays a significant role in prediction and diagnosis of diseases. Diseases with the same or similar phenotype have the same biological features which describe them. Since there are often a large number of features related to biological data (genes/proteins), it is important to find out which input features are useful in diagnosis of a given disease. This is because feature selection is an important tool in biological researches.

On the other hand, almost all methods presented so far for feature selection in biological data have not considered the inter-relationship between data. However, interacted proteins have more similar characteristics than uncorrelated ones; it is beneficial to incorporate PPI in addition to features in feature selection.

Therefore, an unsupervised method for feature selection is proposed here because of the existing a huge subset of unlabeled data in biological studies. For this purpose, by optimizing a novel objective function, which incorporates both the inter-relationship of genes/proteins in addition to their features, the top-ranked features are extracted. Also, unlike other methods, in this paper, the data are labeled dynamically in the consecutive iterations of proposed algorithm so that, after its convergence, an appropriate clustering of proteins is achieved.

We compare our proposed method with some well-known evaluation criteria on two real-world datasets. The experimental results demonstrated the effectiveness of our proposed method in exploiting link information for selecting informative features in comparison with the state-of-the-art methods.

References

1. Loscalzo J, Kohane I, Barabási A-L (2007) Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Mol Syst Biol 3(124):1–11
2. Tang J, Liu H (2012) Unsupervised feature selection for linked social media data. In: KDD
3. Nie F, Huang H, Cai X, Ding CHQ (2010) Efficient and robust feature selection via joint l2, 1-norms minimization. In: NIPS, pp 1813–182
4. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113
5. Backstrom L, Leskovec J (2011) Supervised random walks: predicting and recommending links in social networks. In: WSDM, pp 635–644
6. Li Z, Liu J, Yang Y, Zhou X, Lu H (2014) Clustering-guided sparse structural learning for unsupervised feature selection. IEEE Trans Knowl Data Eng 26(9):2138–2150
7. Peng H, Long F, Ding CHQ (2005) Feature selection based on mutual information: criteria of maxdependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238
8. Dy JG, Brodley CE (2004) Feature selection for unsupervised learning. J Mach Learn Res 5:845–889
9. Cawley GC, Talbot NLC, Girolami M (2006) Sparse multinomial logistic regression via bayesian l1 regularisation. In: NIPS, pp 209–216
10. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc (Ser B) 58:267–288
11. Robnik-Sikonja M, Kononenko I (2003) Theoretical and empirical analysis of ReliefF and RReliefF. Mach Learn 53(1):23–69
12. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238
13. Zhao Z, Liu H (2007) Spectral feature selection for supervised and unsupervised learning. In: Proceedings of the 24th international conference on machine learning, ACM, pp. 1151–1157
14. He X, Cai D, Niyogi P (2006) Laplacian score for feature selection. In: NIPS, vol. 18, no. 507
15. Constantinopoulos C, Titsias M, Likas A (2006) Bayesian feature and model selection for gaussian mixture models. In: TPAMI, pp. 1013–1018
16. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46(1):389–422
17. Pehro D, Stork DG (2001) Pattern Classification. Wiley, London
18. Krzanowski W (1987) Selection of variables to preserve multivariate data structure, using principal components. Appl Stat 26:22–33
19. Cai D, Zhang C, He X (2010) Unsupervised feature selection for multi-cluster data. In: KDD. ACM pp. 333–342
20. Yang Y, Shen H, Ma Z, Huang Z, Zhou X (2011) L21-norm regularized discriminative feature selection for unsupervised learning. In: Proceedings of the twenty-second international joint conference on artificial intelligence
21. Li Z, Yang Y, Liu J, Zhou X, Lu H (2012) Unsupervised feature selection using nonnegative spectral analysis. In: AAAI
22. Qian M, Zhai C (2013) Robust unsupervised feature selection. In: IJCAI
23. Gu Q, Han J (2011) Towards feature selection in network. In: CIKM
24. Tang J, Liu H (2012) Feature selection with linked data in social media. In: SIAM international conference on data mining
25. Tang J, Liu H (2013) CoSelect: feature selection with instance selection for social media data. In: SIAM international conference on data mining
26. Tang J, Liu H (2014) An unsupervised feature selection framework for social media data. IEEE Trans Knowl Data Eng 26(12):2914–2927
27. Tang L, Liu H (2009) Relational learning via latent social dimensions. In: KDD
28. Newman ME (2006) Modularity and community structure in networks. Proc Natl Acad Sci USA 103(23):8577–8582
29. Ng AY, Jordan MI, Weiss Y (2002) On spectral clustering: analysis and an algorithm. In: NIPS. MIT Press, pp 849–856
30. Yang Y, Shen HT, Nie F, Ji R, Zhou X (2011) Nonnegative spectral clustering with discriminative regularization. In: AAAI
31. Ding C, Zhou D, He X, Zha H (2006) R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization. In: ICML
32. Golub GH, Van Loan CF (1996) Matrix computations, 3rd edn. Johns Hopkins, Baltimore
33. Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Trans PAMI 22:888–905
34. Yu SX, Shi J (2003) Multiclass spectral clustering. In: ICCV

35. Fukunaga K (1990) Introduction to statistical pattern recognition, 2nd edn. Academic Press Professional Inc, San Diego
36. Lee D, Seung H (1999) Learning the parts of objects by nonnegative matrix factorization. Nature 401:788–791
37. Lee D, Seung H (2001) Algorithms for nonnegative matrix factorization. In: NIPS
38. Liu Y, Jin R, Yang L (2006) Semi-supervised multilabel learning by constrained non-negative matrix factorization. In: AAAI
39. Kuhn H, Tucker A (1951) Nonlinear programming. In: Berkeley symposium on mathematical statistics and probabilistics
40. Goh K et al (2007) The human disease network. PNAS 104(21):8685–8690
41. Gill N, Singh S, Aseri TC (2014) Computational disease gene prioritization: an appraisal. J Comput Biol 21(6):456–465
42. Strehl A, Ghosh J (2002) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 3:583–617
43. Hoseini E, Mansoori EG (2016) Selecting discriminative features in social media data: an unsupervised approach. Neurocomputing 205(C):463–471
44. Mansoori EG, Zolghadri MJ, Katebi SD (2009) Protein superfamily classification using fuzzy rule-based classifier. IEEE Trans Nanobiosci 8(1):92–99
45. Jowkar GH, Mansoori EG (2016) Perceptron ensemble of graph-based positive unlabeled learning for disease gene identification. Comput Biol Chem 64:263–270
46. Hsu CW, Lin CJ (2002) A comparison of methods for multiclass support vector machines. IEEE Trans Neural Netw 13(2):415–425
47. Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Am Stat Assoc 32(200):675–701