*An unsupervised approach for feature selection in linked data*

Elham Hoseini

Department of Computer Science and Engineering

Shiraz University

Shiraz, Iran

hoseini-e@shirazu.ac.ir

Eghbal G. Mansoori

Department of Computer Science and Engineering

Shiraz University

Shiraz, Iran

mansoori@shirazu.ac.ir

Abstract—**Most of the data in the field of social media has many features. Accordingly, one of the main challenges in this field is processing such high-dimensional data. Researchers are motivated to propose novel approaches in order to overcome this problem. One of the best solutions is extracting the effective information from data pool and discard unnecessary one. Feature selection is a known technique which aims to distinguish discriminative features. Because of the unlabeled nature of datasets in social network, an Unsupervised Feature Selection algorithm might be a good scenario. In addition to features information, we confront inherently linked users in social network datasets. This is because a stronger unsupervised feature selection is needed which ignores the independent and identically distributed assumption. Hence, by optimizing a novel objective function in this paper, feature ranking is done and top features are extracted for further processing. This objective function incorporates both the relationship between users and information of users' features. The experimental results on real-world social network datasets demonstrate the effectiveness of our proposed approach.**

Keywords-Unsupervised feature selection; social media; link information; graph partitioning.

# Introduction

The rapid growing of social network services such as Facebook and Twitter in recent years, allows millions of users to participate in online social activities. Social network services have to be able to handle high-dimensional and massive data. However, due to curse of dimensionality and lack of scalability, processing data via classification or clustering would be an undeniably challenge. One effective approach to handle high-dimensional data is feature selection[1], [2], which tries to select a subset of features from pool of features that minimize redundancy and maximize relevance to the target (e.g., class label). Feature selection aims to improve the performance of learning models by alleviating the curse of dimensionality, speeding up the learning process, and improving the generalization capability of a learning model [3], [4].

Feature selection algorithms are categorized into supervised and unsupervisedmodels. In supervised feature selection, the training data is labeled and the correlated features are accessed based on distinguishing different classes. Supervised feature selection algorithms usually suffer from high complexity. On the other hand, unsupervised feature selection algorithms use unlabeled data and they are particularly difficult due to the definition of relevancy of features becomes unclear [5], [6], [7]. Furthermore, with high-dimensional data, it is likely to find many sets of features that seem equally good without considering additional constraints [5], [6]. Most existing feature selection algorithms work with independent and identically distributed data, while social media data is inherently linked; adding further challenges to feature selection.

As stated above, social network datasets consist of some users which are probably connected to some like-minded participates. Each user has its own features which might be composed of its provided tags, personal information or etc. Figure 1 shows a small example of linked data in social network consists of ten distinct users with twelve features.

Clearly, in most social networks, only some features are available for a single user. This is because the users usually fill in their basic fields such as name and gender, but seldom introduce their interests and other detailed information. Additionally, due to the privacy issues, most social network sites limit the access to some personal information. Accordingly, in the example shown in Fig.1, user has features, and but not features, and .

1. A sample of linked social media data with six users and eight features.

*u*1

*u*2

*u*3

*u*4

*u*5

*f*4, *f*8

*f*1, *f*6, *f*7

*u*6

*f*4

*f*1, *f*3, *f*6

*f*8

*f*2, *f*5

In this paper, we would like to develop a novel feature selection algorithm by considering linked property of social network data. In particular, we aim to exploit and model the relations among data instances, and to take advantage of these relations for feature selection in an unsupervised scenario. In supervised learning, label information plays the role of constraint. Without labels, alternative constraints can be used, such as data variance and separability [8], [1], [9]. Because of the importance of discriminative information in data analysis, it is beneficial to exploit discriminative information for feature selection. However, how to select discriminative features in unsupervised scenarios is a significant but hard task due to the lack of labels.

In order to consider both relationship between users and discriminative information, we propose an Unsupervised Feature Selection framework for Social network datasets (UFSS) which takes into account the followings:

First, users are clustered by applying relationship between them and we search for a subset of features by which the clustering results become reachable.

Second, we seek for some features by which separability between users with respect to their features becomes as much as possible.

The rest of this paper is organized as follows. The related work is presented in next section. Then, we formally define the problem of unsupervised feature selection for linked data in social network in Section 3. Our new framework of unsupervised feature selection for social network, UFSS, is introduced in Section 4, including approaches to capture link information, optimization, and convergence analysis. At the end, experimental result with discussion is presented in Section 5 and concludes this work in Section 6.

# Related Work

Traditionally, feature selection algorithms can be either supervised or unsupervised [1, 4] based on the training data being labeled or unlabeled.

Supervised feature selection methods [4] can be broadly categorized into the *wrapper* models [10, 3] and the *filter* models [11, 12]. The wrapper model uses the predictive accuracy of a predetermined learning algorithm to determine the quality of selected features. These methods can be egregiously expensive to run for data with a large number of features [13]. The filter model separates feature selection from classifier learning so that the bias of a learning algorithm does not interact with the bias of a feature selection algorithm. It relies on measures of the general characteristics of the training data such as distance, consistency, dependency, information, and correlation [11]. Many researchers paid great attention to developing unsupervised feature selection [14]. Unsupervised feature selection [15, 1] is a less constrained search problem without class labels, depending on clustering quality measures, and can eventuate many equally valid feature subsets. With high dimensional data, it is likely to find many sets of features that seem equally good without considering additional constraints. Another key difficulty is how to objectively measure the results of feature selection. A wrapper model is proposed in [1] to use a clustering algorithm in evaluating the quality of feature selection.

Recently, sparsity regularization, such as the 2,1-norm of a matrix [16], in dimensionality reduction has been widely investigated and applied to feature selection including multi-task feature selection [17, 4], robust joint 2,1-norms [18], spectral feature selection [19], discriminative unsupervised feature selection [20]. Through sparsity regularization, feature selection can be embedded in the learning process.

LinkedFS [21] is a supervised algorithm which attempt to select features on social media data. various relations (coPost, coFollowing, coFollowed and Following) are extracted following social correlation theories [22]. LinkedFS significantly improves the performance of feature selection by incorporating these relations into feature selection. The first attempt to select features in an unsupervised manner in social media is LUFS [25, 29]. LUFS utilizes graph regularization and social dimension regularization to capture the individual and group behaviors of linked instances, separately. Also, it introduces the concept of pseudo-class labels to guide extracting constraints from link information and feature-value information.

# Proposed method

We first introduce the notations and definitions in our paper. Let U = {u1, u2… un} be the set of users where n is the number of users and f = {f1, f2 . . . fm} be the set of features where m is the number of features.

We have X = [x1, x2 … xn] ϵ where irepresents ui and xij is the frequency of j used by iand R ϵ denotes the link information for social network data. The value of Rij would be one if ui and uj are linked, zero otherwise. Here, there are undirected connections among users, i.e., R = RT. We assume that the data, X, is centered, where , which can be realized as where H = In −.

Given and , we aim to find a mapping matrix W Є which provides Y=WTX where Y ϵ .

Since X is centered, it is easy to verify that Y is also centered,

 (1)

In order to find matrix W, a proposed objective function in this paper should be minimized. This objective function captures both information of X and R matrices.

By capturing users’ relationship matrix , we apply a known graph partitioning based on spectral factorization [ref] and make matrix ϵ . It is clear that, we aim to find W such that the difference between and is minimized. So, we form the first part of our objective function as follows:

 (2) (2)

where is the -norm of and controls the capacity of this matrix. Also, minimizing ensures that is sparse in rows, making it particularly suitable for feature selection and is defined as follows,

(3)

We use Ferobenius norm to calculate the difference between and , which is defined as the [square root](http://mathworld.wolfram.com/SquareRoot.html) of the sum of the absolute squares of its elements. The Parameter is used to control the sparsity of .

The second part of our objective function is formed by taking into account the users information matrix and apply discriminative information. A well-known method to utilize discriminative information is to find a low dimensional subspace in which is maximized while is minimized [27]. The between class scatter matrix and total scatter matrix are defined as follows:

 (4)

where is the mean of all samples, is the mean of samples in the ith class, ni is the number of samples in the ith class, is the data matrix after being centered, and TT-1/2 is the scaled label matrix.

On the other hand, the between class scatter matrix and total scatter matrix in the reduced space are defined as follows:

By maximizing the following term (minimizing the negative) and merge it by the previous part, objective function is completed.

So,

 (6)

The parameter controls the discriminative information value.

Also we have,

W can be updated as:

W 🡸 (7)

where is a diagonal matrix with the ith diagonal element as .

*Proof:* The Lagrangian function of the objective function is:

To minimize , we take the derivative:

Setting the derivative to zero, we can obtain the update rule of equation 7, which completes the proof.

Based on calculated update rule for W, we develop algorithm 1.

**Algorithm 1: UFSS**

Input: ()

Output: features

1. Do graph partitioning on R and make P
2. Calculate F = PT (PPT)-1/2
3. Initialize Dw as identity matrix
4. Initialize
5. While not convergent do
6. Update diagonal matrix Dw, where the ith diagonal element is
7. Update W as
8. End while
9. Sort each feature according to in descending order and select top\_q ranked ones

The larger the norm of , the more relevant the ith feature is.

# Experiments

In this section, we present experiment details to verify the effectiveness of the proposed framework, UFSS. After introducing real-world social media datasets, we first evaluate the quality of selected features in terms of clustering performance, and then we study the effects of parameters on performance.

We collect two datasets from real-world social media web- sites, i.e., BlogCatalog and Flickr, which are the subsets of two public available datasets used in [28] to uncover overlapping groups in social media. Some statistics of the datasets are shown in Table 1.

1. Statistics of two real world datasets.

|  |  |  |
| --- | --- | --- |
|  | BlogCatalog | Flickr |
| #Users | 88,784 | 35,313 |
| #Features | 5,413 | 77,263 |
| #Links | 2,818,224 | 3,017,530 |
| #Classes | 60 | 200 |

We reduce BlogCatalog dataset to uncover the users which are being in more than one category. Also, to have balanced distribution of users, 14 categories are selected in which between 200 and 500 users exist. Indeed, the number of users is reduced to 4923 with 14998 links. In a similar manner, Flickr is also reduced to cover about 7,000 users with 9 classes. Since the original Flickr is a multi-class dataset, those users which are in the first nine classes are only selected.

LUFS is compared with the following three unsupervised feature selection algorithms,

* UDFS [20] selects features in batch mode by simultaneously exploiting discriminative information and feature correlation.
* Laplacian Score [24] evaluates the importance of a feature through its power of locality preservation.
* LUFS [25, 29] which is an unsupervised feature selection and employs relations as groups via social dimensions.

The clustering quality is evaluated by two commonly used metrics, Mean Silhouette [23] and Entropy.

Silhouette refers to a method of interpretation and validation of [clusters of data](http://en.wikipedia.org/wiki/Cluster_analysis). *a(i)* be the average dissimilarity of *i* with all other data within the same cluster. Any measure of dissimilarity can be used but [distance measures](http://en.wikipedia.org/wiki/Distance) are the most common. We can interpret *a(i)* as how well *i* is assigned to its cluster (the smaller the value, the better the assignment). We then define the average dissimilarity of point *i* to a cluster *c* as the average of the distance from to points in c.

Let *b(i)* be the lowest average dissimilarity of i to any other cluster which *i* is not a member. The cluster with this lowest average dissimilarity is said to be the "neighboring cluster" of *i* because it is the next best fit cluster for point *i*. We now define:

 (7)

Mean silhouette is calculated over all users.

Furthermore, given two variables P and Q, NMI is defined as:

 (9)

Where is the mutual information between P and Q, and H(P) and H(Q) are the entropies of P and Q [26]. Denote as the number of data in the cluster according to clustering results and be the number of data in the *h*-th ground-truth class . NMI is defined as follows [26].

 (8)

where is the number of samples that are in the intersection between the cluster and the *h*-th ground truth label. Again, a larger NMI indicates a better clustering result.

UFSS has three important parameters: the number of spectral clusters (), controlling discriminative information (), and controlling the sparsity of W (). Each parameter is set by fixing the others to see how the performance of UFSS varies with the number of selected features. The best value of are estimated 8, 0.9, 0.3 for BlogCatalog and 9, 0.5, 0.1 for Flickr respectively.

The comparison results w.r.t both Mean silhouette and NMI are demonstrated in Table 2 and 3 for Flickr dataset and in Table 4 and 5 for BlogCatalog.

1. Mean Silhouette versus different feature selection algorithms in BlogCatalog

|  |  |
| --- | --- |
|  | **No. of features** |
| Methods | 200 | 400 | 600 | 800 | 1000 | 5413 (All features) |
| LUFSLaplacianUDFSUFSS | 0.52290.50420.87600.9355 | 0.49520.52980.84440.8928 | 0.47510.47470.79920.8420 | 0.42780.50560.75890.7665 | 0.42750.47200.73870.7498 | 0.3806 |

1. NMI versus different feature selection algorithms in BlogCatalog

|  |  |
| --- | --- |
|  | **No. of features** |
| Methods | 200 | 400 | 600 | 800 | 1000 | 5413 (All features) |
| LUFSLaplacianUDFSUFSS | 0.01020.01100.01250.0298 | 0.00780.01120.01480.0292 | 0.00910.01000.01350.0270 | 0.01000.00960.01380.0242 | 0.00900.01050.01410.0206 | 0.0081 |

1. Mean Silhouette versus different feature selection algorithms in Flickr

|  |  |
| --- | --- |
|  | **No. of features** |
| Methods | 200 | 400 | 600 | 800 | 1000 | 9045 (All features) |
| **LUFS****Laplacian****UDFS****UFSS** | 0.73690.89930.72090.9893 | 0.68740.82000.76320.9805 | 0.64780.73940.68310.9782 | 0.64600.68590.65520.9628 | 0.66470.66100.67630.9538 | 0.6265 |

1. NMI versus different feature selection algorithms in Flickr

|  |  |
| --- | --- |
|  | **No. of features** |
| Methods | 200 | 400 | 600 | 800 | 1000 | 9045 (All features) |
| **LUFS****Laplacian****UDFS****UFSS** | 0.00880.00350.00340.0138 | 0.01480.00380.00420.0164 | 0.00660.00320.00390.0154 | 0.01330.00400.00360.0147 | 0.01430.00320.00370.0158 | 0.0031 |

The results demonstrate that UFSS outperforms three other algorithms. With the increasing number of selected features, the value of mean silhouette always degrades. Although NMI for UFSS compared to other algorithms is more, generally its value is small. This is because the samples are placed in a ground-truth class, at the end are not necessarily clustered in a same cluster and this leads to a decrease in NMI.

Like the previous, the value of mean silhouette is always keeping increasing with the rise of features. Notice that the difference between mean silhouette of our approach and three other algorithms is relatively high.

# Conclusion

Because of the independent and identically distributed nature of social network datasets, it presents new challenges to traditional feature selection algorithms. In this paper, we propose a novel unsupervised feature selection framework, for linked data.

We utilize both the relationship between users and information of users' features and propose an objective function. In this function, we seek for a mapping matrix in which the discriminative information of each features exists. Finally, the ranked features are obtained utilizing this mapping matrix.

Experimental results on two datasets from real-world social media websites show that the proposed method can effectively exploit link information in comparison with the state-of-the-art unsupervised feature selection methods.

##### References

1. J. Dy and C. Brodley. Feature selection for unsupervised learning. Journal of Machine Learning Research, 5:845–889, 2004.
2. H. Liu and H. Motoda. Computational methods of feature selection. Chapman & Hall, 2008.
3. G. John, R. Kohavi, and K. Pfleger. Irrelevant feature and the subset selection problem. ICML, 1994.
4. H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. TKDE, 2005.
5. J. G. Dy, C. E. Brodley, A. C. Kak, L. S. Broderick, and A. M. Aisen. Unsupervised feature selection applied to contentbased retrieval of lung images. TPAMI, 2003.
6. J. G. Dy and C. E. Brodley. Visualization and interactive feature selection for unsupervised data. In KDD, pages 360–364, 2000.
7. D. Cai, C. Zhang, and X. He. Unsupervised feature selection for multi-cluster data. In KDD, pages 333–342. ACM, 2010.
8. V. Roth and T. Lange. Feature selection in clustering problems. NIPS, 16:473–480, 2004.
9. C. Constantinopoulos, M. Titsias, and A. Likas. Bayesian feature and model selection for gaussian mixture models. TPAMI, pages 1013–1018, 2006.
10. J. G. Dy and C. E. Brodley. Feature subset selection and order identification for unsupervised learning. In ICML, 2000.
11. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. Machine learning, 46(1):389–422, 2002.
12. M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. Physical review E, 69(2):26113, 2004.
13. Q. Gu and J. Han. Towards feature selection in network. In CIKM, 2011.
14. L. Tang and H. Liu. Relational learning via latent social dimensions. In KDD, pages 817–826. ACM, 2009.
15. M. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In ICML, 2000.
16. C. Ding, D. Zhou, X. He, and H. Zha. R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization. In ICML, 2006.
17. A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. NIPS, 19:41, 2007.
18. F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint l21-norms minimization. NIPS, 2010.
19. Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In Proceedings of the 24th international conference on Machine learning, pages 1151–1157. ACM, 2007.
20. Y. Yang, H. Shen, Z. Ma, Z. Huang, and X. Zhou. L21-norm regularized discriminative feature selection for unsupervised learning. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, 2011.
21. J. Tang and H. Liu. Feature selection with linked data in social media. In SIAM International Conference on Data Mining, 2012.
22. P. Marsden and N. Friedkin. Network studies of social influence. Sociological Methods and Research, 22(1):127–151, 1993.
23. P. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” J. Comput. Appl. Math., vol. 20, no. 1, pp. 53–65, 1987.
24. X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. NIPS, 18:507, 2006.
25. J. Tang, H. Liu. Unsupervised Feature Selection for Linked Social Media Data. KDD 2012.
26. A. Strehl and J. Ghosh. Cluster ensembles–a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 3:583–617, 2002.
27. K. Fukunaga. Introduction to statistical pattern recognition (2nd Edition). Academic Press Professional, Inc, San Diego, USA, 1990.
28. X. Wang, L. Tang, H. Gao and H. Liu. Discovering Overlapping Groups in Social Media. ICDM 2010.
29. J. Tang, H. Liu, An Unsupervised Feature Selection Framework for Social Media Data, Knowledge and Data Engineering, IEEE Transactions on 26(12), 2014.