Unsupervised Feature Selection with Heterogeneous Side Information

Xiaokai Wei, Bokai Cao and Philip S. Yu
Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA
weixiaokai@gmail.com, {caobokai, psyu}@uic.edu

ABSTRACT

Compared to supervised feature selection, unsupervised feature selection tends to be more challenging due to the lack of guidance from class labels. Along with the increasing variety of data sources, many datasets are also equipped with certain side information of heterogeneous structure. Such side information can be critical for feature selection when class labels are unavailable. In this paper, we propose a new feature selection method, SideFS, to exploit such rich side information. We model the complex side information as a heterogeneous network and derive instance correlations to guide subsequent feature selection. Representations are learned from the side information network and the feature selection is performed in a unified framework. Experimental results show that the proposed method can effectively enhance the quality of selected features by incorporating heterogeneous side information.

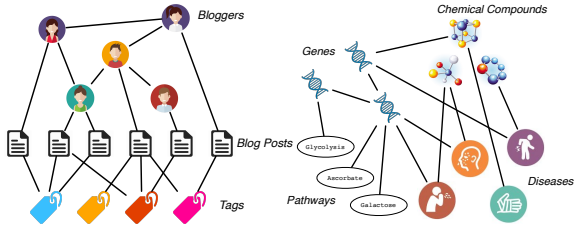
1 INTRODUCTION

High dimensionality of data poses challenges to many machine learning tasks and feature selection [2] [13], by retaining a set of high-quality ones, can help alleviate the curse of dimensionality and make machine learning models more interpretable.

Supervised feature selection [6] selects features by measuring their correlations with class labels which are usually expensive to obtain. Therefore, we mainly focus on unsupervised feature selection in this paper. Unsupervised feature selection methods [2] [16] [5] [10] [4] [13] [12] [11] usually aim to exploit the information embedded in the unlabeled data. However, using the potentially noisy features alone as guidance might be insufficient for selecting high-quality features.

In the era of big data, one can often collect various forms of side information associated with the entities of interest. Such side information can usually provide abundant information about the data instances. Hence, to select high-quality features, it is highly desirable to incorporate the side information. However, side information usually comes in a complex form, where different objects are interrelated. Such inter-connected complex side information poses additional challenges on how to use it effectively (Figure 1).

- In blog websites (e.g., BlogCatalog), each blog can be represented as a high dimensional feature vector from the text data. Besides such text features, blog posts are also equipped with complex side information as shown in Figure 1a where each blog post is associated with a user who writes the blog and a set of tags describing the post. In addition, there are social relationships between users. Considering the cost of supervised feature selection with labels from human experts, it would be a worthwhile effort to utilize the side information to guide feature selection.
- In bioinformatics, each chemical compound can be represented by its substructures in the vector space. Consider the task of predicting the side effect of drugs (i.e., chemical compounds) based on these substructure features. The supervision information (i.e., side effect) can be very expensive to obtain through clinical trials (e.g., sometimes even at the cost of human lives), and thus supervised feature selection is less desirable. Fortunately, one can have a set of side information (e.g., gene-chemical compound interaction, gene-pathway interaction and gene-disease interaction, as in Figure 1b) which provides rich information for selecting informative substructure features.
- In news articles, there are different concepts or entities, such as people, places and organizations. The hetereogeneous relationships (e.g., extracted from Freebase [1]) can also be useful for guiding the selection of text features.



(a) Blogs (b) Chemical compounds

Figure 1: Examples of data with complex side information

Such heterogeneous side information can provide valuable information for feature selection, especially when class labels are unavailable. To handle the increasingly complicated form of side information, we propose a new method, SideFS (Complex Side Information-guided Feature Selection), in this paper. Since different types of relationship may exist among the objects in the side information, we model them as a heterogeneous information network [7] [9] [14]. We then derive similarity measures between instances based on the concept of meta-path. Information is derived from the meta-paths by learning network based representations and such representations are used to guide feature selection. The contributions of this paper can be summarized as follows.

CIKM'17 , November 6–10, 2017, Singapore, Singapore

ACM ISBN 978-1-4503-4918-5/17/11…$15.00
https://doi.org/10.1145/3132847.3133055

- To our best knowledge, we are the first to formulate the problem of unsupervised feature selection with heterogeneous form of side information.
- We propose a novel method, SideFS, which performs joint feature selection and representations learning from the complex side information by modeling it as a heterogeneous information network.

2 PROPOSED METHOD

We denote n data samples as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ and the dimensionality of original feature space is D. So $\mathbf{x}_i \in \mathbb{R}^D$ and x_{ip} denotes the value of p-th ($p = 1, \ldots, D$) feature of $\mathbf{x}_i$. Our goal is to select d ($d \ll D$) high-quality features.

2.1 Knowledge Extraction from Complex Side Information

We model the complex relationship of entities in the side information as a heterogeneous *side information network*. The key idea of this knowledge extraction step is to first derive meta-paths from the side information network and encode the side information via embedding learning.

Definition 1. **Side Information Network** *The heterogeneous side information of data instances can be represented as a Side Information Network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ denotes the set of nodes, which includes t types of entities, $\mathcal{V}_1 = \{v_{11}, v_{12}, \ldots, v_{1n_1}\}, \ldots, \mathcal{V}_t = \{v_{t1}, v_{t2}, \ldots, v_{tn_t}\}$. $\mathcal{E}$ denotes the set of (multiple types of) links $\mathcal{E} \subset V \times V$.*

The target data instances are also one type of nodes in the side information network and we refer to them as **instance nodes**.

Definition 2. **Meta-path** *A meta-path $\mathcal{P}$ of length l represents a sequence of relations $\mathcal{R}_i$ ($i = 1, \ldots, l$), i.e., $\mathcal{T}_1 \xrightarrow{\mathcal{R}_1} \mathcal{T}_2 \xrightarrow{\mathcal{R}_2} \cdots \xrightarrow{\mathcal{R}_l} \mathcal{T}_{l+1}$, where $\mathcal{T}_i$ ($i = 1, \ldots, l+1$) are the types of nodes. A unique sequence of nodes is referred to as a* **path instance** *of $\mathcal{P}$.*

For each pair of instances, various meta-paths can be extracted to provide information about their correlations. Different types of meta-paths usually have different semantic meanings. For example, meta-path *Compound-Disease-Compound* means chemical compounds that can cure the same disease, while meta-path *Compound-Gene-Pathway-Gene-Compound* indicates chemical compounds binding with the genes that are involved in the same pathway.

Inspired by the path-counting measure in [7], we define the following side information-based similarity measure by counting the meta-path instances between the target data points.

Definition 3. **SideSim** *Given a side information network, we define the following similarity measure from the side information w.r.t meta-path $m \in M$ as follows:*

$$s_{ij}^{(m)} = \frac{2 \cdot |\mathcal{P}^{(m)}(i \rightsquigarrow j)|}{|\mathcal{P}^{(m)}(i \rightsquigarrow \cdot)| + |\mathcal{P}^{(m)}(j \rightsquigarrow \cdot)|} \tag{1}$$

where $|\mathcal{P}^{(m)}(i \rightsquigarrow j)|$ denotes the number of path instances with type m between data instances i and j, and $|\mathcal{P}^{(m)}(i \rightsquigarrow \cdot)|$ denotes the number of out-going path instances of type m from instance i.

These multiple types of meta-paths depict the correlations among target data instances from complementary perspectives, and it is desirable to ensemble them to obtain a more comprehensive view of correlations. We consider the following two ways of aggregation, which we refer to as **Micro Aggregation** and **Macro Aggregation**. We will compare the performance of these two aggregation methods in experiments.

Definition 4. **Micro SideSim Aggregation** *We define the micro-aggregation of SideSim as follows:*

$$s_{ij} = \frac{\sum_{m \in M} \cdot 2w^{(m)} |\mathcal{P}^{(m)}(i \rightsquigarrow j)|}{\sum_{m \in M} w^{(m)} |\mathcal{P}^{(m)}(i \rightsquigarrow \cdot)| + \sum_{m \in M} w^{(m)} |\mathcal{P}^{(m)}(j \rightsquigarrow \cdot)|} \tag{2}$$

Definition 5. **Macro SideSim Aggregation** *We define the macro-aggregation of SideSim as follows:*

$$s_{ij} = \sum_{m \in M} w^{(m)} s_{ij}^{(m)} \tag{3}$$

where $w^{(m)}$ is the weight assigned to meta-path with type m. In the unsupervised scenario, one could just use equal weights for all types of meta-paths, as the simplest form of ensemble. Alternatively, one could rely on domain experts to provide prior knowledge to determine the importance of different meta-paths. We adopt the former approach in our experiments.

We further define the transition probability $\mathbf{P}$ based on the aggregated SideSim

$$P_{ij} = \frac{s_{ij}}{\sum_{j=1}^{n} s_{ij}} \tag{4}$$

Considering that the SideSim between instances are not quite reliable for non-nearest pairs, we truncate the fused full similarity graph to a kNN graph G^f based on $\hat{P} = (P_{ij} + P_{ij}^T)/2$ which tends to have better performance in our preliminary experiments. We use $k = 10$ in this paper.

To further extract information from the fused graph G^f, we learn embeddings from this graph structure. Since the connected instances tend to have larger correlations, we learn the embeddings $\mathbf{u}_i \in \mathbb{R}^c$ ($i = 1, 2, \ldots, n$) for each instance to make the embeddings of neighbors in G^f close and embeddings of non-neighbors far apart. Hence, we minimize the negative log-likelihood as follows:

$$\min_{\mathbf{U}} L^g = -\sum_{(i,j) \in E} \log(f_{ij}) - \sum_{(i,j) \in NE} \log(1 - f_{ij}) + \gamma ||\mathbf{U}||_F^2 \tag{5}$$

where $\mathbf{U} = [\mathbf{u}_1^T, \ldots, \mathbf{u}_n^T]^T$ and γ controls the complexity of $\mathbf{U}$ ($||\cdot||_F$ denotes Frobenius norm). f_{ij} should be a monotonic function that transforms the similarity or distance between $\mathbf{u}_i$ and $\mathbf{u}_j$ into the range of $(0, 1)$. For example, f_{ij} could be $\frac{1}{1+\exp(-\mathbf{U}_i \mathbf{U}_j^T)}$ or $\frac{1}{1+||\mathbf{U}_i - \mathbf{U}_j||_F^2}$. We found these two functions have similar performance in our preliminary experiments and we use the former one in the rest of paper. For the set of negative edges NE in Eq (5), we perform negative sampling as in [8] and retain $|E|$ number of negative edges.

2.2 Joint Representation Learning and Feature Selection

Side information could also be noisy, so the representations derived from the side information network might not be high-quality for every data instance. Under such scenario, it is desirable to also

incorporate information from the instance features for the representation learning. Meanwhile, we perform feature selection jointly in this process.

To utilize these representations for feature selection, we learn a linear projection of $\mathbf{U}$.

$$\min_{\mathbf{V}} ||\mathbf{U}\mathbf{V}^T - \mathbf{X}||_F^2 \tag{6}$$

A projection matrix $\mathbf{V} \in \mathbb{R}^{D\times c}$ is introduced to establish the connection between the representations $\mathbf{U}$ and the feature matrix $\mathbf{X}$ in Eq (6).

To perform joint feature selection when learning the representation, we employ a feature selection indicator vector $\mathbf{s} \in \{0,1\}^D$, where $s_p = 1$ indicates the p-th feature is selected and $s_p = 0$ otherwise.

$$\begin{aligned} \min_{\mathbf{V},\mathbf{s}} \ & ||\mathbf{U}\mathbf{V}^T \operatorname{diag}(\mathbf{s}) - \mathbf{X}||_F^2 \\ \text{s.t. } & s_p \in \{0,1\}, \forall p = 1,\dots,D \\ & \sum_{p=1}^{D} s_p = d \end{aligned} \tag{7}$$

where $\operatorname{diag}(\mathbf{s})$ is the diagonal matrix with $\mathbf{s}$ as the diagonal elements. The constraint $\sum_{p=1}^{D} s_p = d$ enforces that only d ($d < D$) features are retained. Intuitively, the representations leverage information from both the original features and the rich information from the side information. The features that cannot be well represented by the latent representations through linear projection tend to be noisy features and will be removed. s_p of such features tend to be 0 under the constraint of $\sum_{p=1}^{D} s_p = d$. $\operatorname{diag}(\mathbf{s})\mathbf{V}$ is a matrix with d non-zero rows and hence it achieves the effect of feature selection.

Since the optimization problem in Eq (7) is difficult to solve, we employ $L_{2,1}$ norm to achieve the similar effect of feature selection. We further write the constraint in the form of Lagrangian as follows:

$$\min_{\mathbf{V}} L^a = ||\mathbf{U}\mathbf{V}^T - \mathbf{X}||_F^2 + \lambda||\mathbf{V}||_{2,1} \tag{8}$$

where λ is the regularization parameter on $L_{2,1}$ norm.

We combine the side information-based loss and feature-based loss together, and the final objective function becomes the following:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} L &= L^g + L^a \\ &= -\sum_{(i,j)\in E} \log(f_{ij}) - \sum_{(i,j)\in NE} \log(1 - f_{ij}) + \\ &\quad \gamma||\mathbf{U}^g||_F^2 + \alpha||\mathbf{U}\mathbf{V}^T - \mathbf{X}||_F^2 + \lambda||\mathbf{V}||_{2,1} \end{aligned} \tag{9}$$

where α is the parameter that controls the relative importance of consensus learning.

2.3 Optimization

In this section, we discuss how to solve the optimization problem for SideFS. We decompose the objective function into two subproblems and develop an alternating optimization approach (Algoritm 1) to solve the problem in Eq (9). One can use gradient-based method (e.g., steepest descent or L-BFGS) to solve each subproblem. To make the regularization term $||V||_{2,1}$ differentiable at 0, we add a very small positive number ϵ in the denominator as in [15].

Algorithm 1 Alternating Optimization for SideFS

Initialize: $\mathbf{U} = rand(0,1)$, $\mathbf{V} = \mathbf{0}$, $t = 1$.
while not converged **do**
  Fixing $\mathbf{V}$, find the optimal $\mathbf{U}$ by L-BFGS
  Fixing $\mathbf{U}$, find the optimal $\mathbf{V}$ by L-BFGS
  $t = t + 1$
end while
Output: Rank all the features ($i = 1,\dots,D$) by $||\mathbf{V}_{i\cdot}||$ and return the top d features.

Table 1: Statistics of two datasets

Statistics	BlogCatalog	Chemical Compound
# of instances	3083	105
# of features	3170	290
# of labels	5	550

The objective function in Eq (9) monotonically decreases in each iteration and it is lowered bounded. So the alternating framework in Algorithm 1 would converge.

3 EXPERIMENTS

In this section, we compare the proposed method with several baselines with applications on clustering and multi-label prediction.

3.1 Datasets

- BlogCatalog Dataset[1]: A subset of blog post dataset from six categories. A blog post can have side information such as users, tags and relationships between users.
- Chemical Compound Dataset [3]: A bioinformatics network in which each chemical compound has subgraph features mined from the compound structure (statistics shown in 1). Besides, we also have heterogeneous side information such as genes, diseases, pathways and PPIs (protein-protein interactions).

3.2 Baselines

We compare our method to the following unsupervised feature selection methods: Laplacian Score [2], UDFS [16], RSFS and [5] and SNFS [13].

For all methods (except SNFS), we do grid search for the regularization parameter in the range of $\{0.1, 1, 10\}$ and report the best performance. For SNFS, we follow the author's suggestion and choose the λ that makes $N_{0.9}$ close to the desired number of features. We use $c = 5$ as the latent dimension size in our method and the baselines. For the proposed SideFS, we use all the non-redundant meta-paths with length less than 5, since previous work [7] suggests meta-paths with large length tend to be not as useful.

3.3 Clustering Blog Posts

For the BlogCatalog dataset, we evaluate the feature quality by the clustering performance. We use Accuracy and Normalized Mutual Information (NMI) as evaluation metrics following the convention

[1] http://dmml.asu.edu/users/xufei/datasets.html

Table 2: Clustering performance on BlogCatalog

Accuracy (All Features: 0.6224)						
# of Features	LS	UDFS	RSFS	SNFS	Micro-SideFS	Macro-SideFS
100	0.2809	0.4490	0.3973	0.6103	**0.6796**	0.6740
200	0.3730	0.5489	0.5292	0.6670	0.7284	**0.7333**
300	0.3958	0.6147	0.5752	0.6840	**0.7247**	0.7157
400	0.4311	0.6380	0.6059	0.6020	**0.7430**	0.7351
NMI (All Features: 0.4667)						
# of Features	LS	UDFS	RSFS	SNFS	Micro-SideFS	Macro-SideFS
100	0.0193	0.2113	0.1453	0.3870	0.4595	**0.4671**
200	0.1195	0.3391	0.3268	0.4673	**0.5348**	0.5252
300	0.1503	0.4346	0.3837	0.5000	**0.5387**	0.5377
400	0.2109	0.4451	0.4298	0.4357	**0.5628**	0.5570

Table 3: 1-NN performance on side effect prediction. ↑ indicates that larger value is better while ↓ indicates that smaller value is better. The best result on each metric is in bold font.

Micro-F1 ↑ (All Features: 0.0913)					
LS	UDFS	RSFS	SNFS	Micro-SideFS	Macro-SideFS
0.0799	0.0866	0.0936	0.0825	**0.1041**	0.0978
Macro-F1 ↑ (All Features: 0.1061)					
LS	UDFS	RSFS	SNFS	Micro-SideFS	Macro-SideFS
0.0946	0.1023	0.1094	0.1029	**0.1177**	0.1127
Hamming Loss ↓ (All Features: 0.0456)					
LS	UDFS	RSFS	SNFS	Micro-SideFS	Macro-SideFS
0.0477	0.0459	0.0456	0.0477	0.0453	**0.0431**

in existing work [16] [13] [5]. For all methods, we report the average performance of 20 K-means runs[2] on the selected features.

Results The clustering performance is shown in Table 2. When comparing the two variants of SideFS, Macro-SideFS and Micro-SideFS performs similarly with different numbers of features. Compared with other feature selection methods, both Micro-SideFS and Macro-SideFS outperform the baseline methods significantly with different feature sizes. This suggests both micro and macro aggregation methods can be effective to incorporate side information.

3.4 Predicting Side Effect of Chemical Compounds

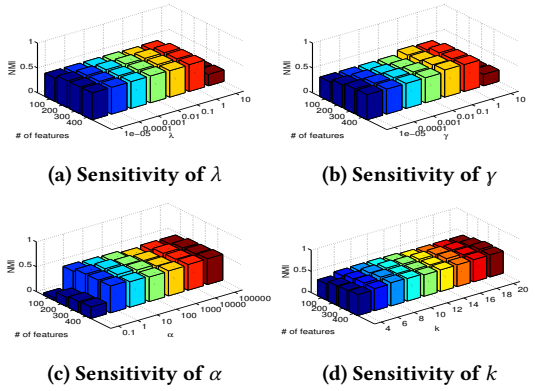
In this subsection, we evaluate the feature quality by their performance in predicting side effects for chemical compounds. Selecting informative substructures can help human experts develop better insights on the mechanisms of compound structures and their potential risks on incurring side effects.

We use 1-NN as the classifier for the prediction task. Since a chemical compound might cause more than one side effects, we use the micro-F1, macro-F1 and Hamming Loss as the performance measures (Table 3). The features selected by SideFS usually outperform baseline methods by 5% ∼ 10%.

3.5 Sensitivity Analysis

We investigate how the proposed method performs under different values of parameters (vary one parameter when fixing $c = 5$ and others equal to 1) with feature sizes $\{100, 200, 300, 400\}$. The

[2]We use the code at http://www.cad.zju.edu.cn/home/dengcai/Data/Clustering.html



(a) Sensitivity of λ **(b) Sensitivity of γ**

(c) Sensitivity of α **(d) Sensitivity of k**

Figure 2: Parameter sensitivity w.r.t. different parameters

NMI results on the BlogCatalog dataset with micro-aggregation are shown in Figure 2. We can observe that SideFS is not very sensitive to these parameter values and performs consistently well for a wide range of parameter values.

4 CONCLUSION

In this paper, we propose a novel method, SideFS, for unsupervised feature selection with heterogeneous side information by learning representations from the meta-paths. Experimental results show that incorporating side information can effectively enhance the quality of selected features in real-world applications.

REFERENCES

[1] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, pages 1247–1250, 2008.
[2] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In *NIPS*, 2005.
[3] X. Kong, B. Cao, and P. S. Yu. Multi-label classification by mining label and instance correlations from heterogeneous information networks. In *KDD*, pages 614–622, 2013.
[4] J. Li, J. Tang, and H. Liu. Reconstruction-based unsupervised feature selection: An embedded approach. In *IJCAI*, 2017.
[5] L. Shi, L. Du, and Y.-D. Shen. Robust spectral learning for unsupervised feature selection. In *ICDM*, 2014.
[6] L. Song, A. J. Smola, A. Gretton, K. M. Borgwardt, and J. Bedo. Supervised feature selection via dependence estimation. In *ICML*, volume 227, pages 823–830, 2007.
[7] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In *VLDB*, 2011.
[8] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. Line: Large-scale information network embedding. In *WWW*, pages 1067–1077, 2015.
[9] C. Wang, Y. Song, H. Li, M. Zhang, and J. Han. Knowsim: A document similarity measure on structured heterogeneous information networks. In *ICDM*, pages 1015–1020, 2015.
[10] S. Wang, J. Tang, and H. Liu. Embedded unsupervised feature selection. In *AAAI*, pages 470–476, 2015.
[11] X. Wei, B. Cao, and P. S. Yu. Multi-view unsupervised feature selection by cross-diffused matrix alignment. In *IJCNN*, pages 494–501, 2017.
[12] X. Wei, S. Xie, and P. S. Yu. Efficient partial order preserving unsupervised feature selection on networks. In *SDM*, pages 82–90, 2015.
[13] X. Wei and P. S. Yu. Unsupervised feature selection by preserving stochastic neighbors. In *AISTATS*, 2016.
[14] L. Xu, X. Wei, J. Cao, and P. S. Yu. Embedding of embedding (eoe) : Embedding for coupled heterogeneous networks. In *WSDM*, 2017.
[15] L. Yan, W.-J. Li, G.-R. Xue, and D. Han. Coupled group lasso for web-scale ctr prediction in display advertising. In *ICML*, volume 32, pages 802–810, 2014.
[16] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou. l2, 1-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*, pages 1589–1594, 2011.