Semi-Supervised Feature Selection with Universum Based on Linked Social Media Data

Junyang QIU\textsuperscript{a)}, Yibing WANG\textsuperscript{b)}, Nonmembers, Zhisong PAN\textsuperscript{c)}, Member, and Bo JIA\textsuperscript{d)}, Nonmember

SUMMARY Independent and identically distributed (i.i.d) assumptions are commonly used in the machine learning community. However, social media data violate this assumption due to the linkages. Meanwhile, with the variety of data, there exist many samples, i.e., Universum, that do not belong to either class of interest. These characteristics pose great challenges to dealing with social media data. In this letter, we fully take advantage of Universum samples to enable the model to be more discriminative. In addition, the linkages are also taken into consideration in the means of social dimensions. To this end, we propose the algorithm Semi-Supervised Linked samples Feature Selection with Universum (U-SSLFS) to integrate the linking information and Universum simultaneously to select robust features. The empirical study shows that U-SSLFS outperforms state-of-the-art algorithms on the Flickr and BlogCatalog.

key words: universum, feature selection, social media, semi-supervised learning

1. Introduction

With the rapid development of the Internet, communication ways have been greatly changed. For many people, traditional network communication tools, such as e-mail and forums have been replaced by social media, e.g. Facebook, Twitter, etc, which are more convenient and effortless for people to express themselves. However, the pervasive use of social media produces massive, high-dimensional and unlabeled data, which presents a new challenge to feature selection. Feature selection is used to reduce the feature space dimension and improve the learning performance.

With the advent of the era of Big Data, it is easy to collect large amounts of unlabeled data. But it is time-consuming to complete the labeling task. How to use large amounts of unlabeled samples to improve the generalization capability of the learned model is an important issue in the machine learning field. For example, Zheng Zhao, et al. \cite{1} proposed a semi-supervised feature selection method based on spectral analysis. Zenglin Xu, et al. \cite{2} discussed a novel discriminative semi-supervised feature selection method based on the idea of manifold regularization. The above feature selection methods all work with data which is typically assumed to be independent and identically distributed (i.i.d.) \cite{3}. However, the social media data does not meet the i.i.d. assumption which presents a new challenge to feature selection. The Universum learning concept was first proposed to improve the binary classification accuracy with the help of Universum—the samples that do not belong to either target classes but belong to the same domain as the classification problem at hand \cite{4, 5}. Universum learning can achieve better performance than other learning methods in classification and clustering scenarios. For example, Weston, et al. \cite{4} proposed Universum support vector machine which aims to leverage the Universum by maximizing the number of observed contradictions \cite{6}. Zhang, et al. \cite{7, 8} introduced the Universum samples to the graph method, in which the Universum samples help to depict the prior knowledge for the construction of classifiers.

However, although Universum learning outperforms other traditional classification and clustering methods, to the best of our knowledge, there is no similar work applying it to the feature selection on the linked social media data. Owing to its favorable performance, we propose an algorithm called U-SSLFS, which integrates Universum regularization with the optimization object to select features on the linked data in a semi-supervised scenario. U-SSLFS learns from the labeled, unlabeled and the Universum samples simultaneously so as to select the most relevant features from the high-dimensional data.

2. Semi-Supervised Feature Selection with Universum

The motivation of semi-supervised methods is to utilize the unlabeled samples to improve the learning performance. Now we obtain extra samples that do not belong to any classes but in the same domain with the problem at hand, called Universum samples. These Universum samples can be used to improve the learning accuracy. For example, in the binary classification, the Universum samples can be used to improve the classification accuracy by maximize the number of contradictions (on the Universum). The Universum samples can also be used in feature selection and extended to multi-class classification. In this letter, we introduce the Universum samples to feature selection based on linked social media data. The Universum samples are integrated to the semi-supervised feature-selection framework as a regularization term.

3. Problem Statement and Notations

Given a graph $G$ consists of nodes and edges, which can be
formulated as $G = (E, V)$, $E$ is the set of edges connecting certain pairs of nodes, and $V = \{u_1, u_2, \ldots, u_n\}$ is the set of $n$ nodes. The $n$ nodes belong to $c$ classes and the corresponding attribute-value matrix is $X = [x_1, x_2, \ldots, x_n] \in R^{m \times n}$, where $x_i$ represents the sample vector of the $i$th node, $m$ is the sample dimension. We assume that the first $l$ samples in $n$ is labeled and can be donated as $L = [x_1, x_2, \ldots, x_l] \in R^{m \times l}$. Let $Y = [y_1, y_2, \ldots, y_l]$ be the label vector of the labeled samples, where $l < n$. The size of unlabeled samples is $n - l$. Let $U = [z_1, z_2, \ldots, z_q] \in R^{m \times q}$ denote the Universum samples and $q$ is the number of the Universum samples. The Universum samples are considered as the $(c + 1)$th class.

4. Semi-Supervised Feature Selection with Universum

4.1 The Social Dimension Regularization

Tang, et al. [3] proposed the concept of pseudo-class labels to guide the learning process. Suppose there is a mapping matrix $W \in R^{m \times (c + 1)}$, the matrix $X$ can be projected into $Y \in R^{(c + 1) \times n}$ where $(c + 1)$ is the number of classes, and $Y = W^T X$. If and only if each column of $Y$ has one non-zero element, i.e. $\|Y(:, i)\|_0 = 1$, we call $Y$ the pseudo-class label matrix. The modularity maximization algorithm which is formulated as Eq. (1) can obtain the sample’s social dimensions.

$$\max_{H} \ \text{tr}(H^T MH), \quad \text{s.t.} \quad H^T H = I.$$  \hspace{1cm} (1)

Where $H \in R^{K \times m}$ is the social dimension indicator matrix which is constituted of the eigenvectors corresponding to the top $K$ ranked eigenvalues of the modularity matrix $M$ [9]. The K-means algorithm can be used to acquire the discrete-valued social dimension denotation, $H(i, j) = 1$ if node belong to the dimension, otherwise, $H(i, j) = 0$. The formation of the weighted social dimension indicator matrix is $F = H(H^T H)^{-1/2}$. The within and between social dimension scatter matrix are $S_w = YY^T - YFF^T Y^T$, $S_b = YFF^T Y^T$, and the total scatter matrix is $S = S_w + S_b$. In [3], the authors utilize the samples’ link information to restrict the object function with the social dimension regularization term: $\max_w \text{tr}(S_b/S_i)$.

4.2 Constraint from the Labeled and Universum Samples Regularization

We can construct the label indicator matrix [10] of the Universum and labeled samples. The label indicator matrix for labeled samples is $T^l \in R^{c \times (c + 1)}$. $T^l(i, j) = 1$, if $y_i = c_j$, and $T^l(i, j) = 0$, otherwise. The label indicator matrix for Universum samples is $T^u \in R^{q \times (c + 1)}$. As Universum samples do not belong to any classes, $T^u(i, c + 1) = 1$, and $T^u(i, j) = 0$, where $j \neq c + 1$.

By introducing the $l_{2,1}$ norm of $\overline{W}$, that is $\|\overline{W}\|_{2,1}$. We can obtain the sparse row vector of $\overline{W}$. The Universum regularization term is $\|U^T \overline{W} - T^u\|_F^2$ and the regularization term of the labeled samples is $\|L^T \overline{W} - T^l\|_F^2$. Here $\| \cdot \|_F$ is the Frobenius norm.

4.3 Manifold Learning Model

In graph spectrum analysis, we can construct weighted matrix $G \in R^{m \times m}$ according to the attribute-value matrix. The weight between sample $u_i$ and $u_j$ can be defined as $G(i, j) = \exp(-||x_i - x_j||^2/\sigma^2)$. The object function can be formulated as $\sum_{i, j} (y_i - y_j)^2 G_{ij} = \text{tr}(Y^T L E Y)$. Here $Y_i \in R^{(c + 1)}$ is the embedding coordinate of the $i$th sample, and $L_p = D - G$ is the Laplacian Matrix [11]. The matrix $D$ is a diagonal matrix with its elements defined as $D(i, i) = \sum_j G(i, j)$. The object function can be concluded as minimizing the following problem:

$$\min_w \text{tr}(Y^T L_p Y), \quad \text{s.t.} \quad \|Y(:, i)\|_0 = 1, 1 \leq i \leq n.$$  \hspace{1cm} (2)

5. Semi-Supervised Feature Selection with Universum Based on Linked Social Media Data

5.1 The Framework of the Model

Now we can integrate the manifold model, the social dimension regularization term and the supervised learning regularization term simultaneously to obtain the feature selection framework based on linked data:

$$\min_{W, w} f(\overline{W}, W) = \text{tr}(YL_p Y^T) + \alpha \|\overline{W}\|_{2,1}$$
$$- \beta \text{tr}(S_l^{-1} S_b) + \gamma \|U^T \overline{W} - T^u\|_F^2$$
$$+ \varepsilon \|L^T \overline{W} - T^l\|_F^2$$
$$\text{s.t.} \|Y(:, i)\|_0 = 1, 1 \leq i \leq n.$$  \hspace{1cm} (3)

We can simplify the problem by making $\overline{W} = W$ and $YY^T = I$. In addition, $T^l$ and $T^u$ are combined into $T$. $U$ and $L$ are combined into $E$. After the transformation, the optimization problem can be converted to:

$$\min_w f(W) = \text{tr}(YL_p Y^T) + \alpha \|W\|_{2,1}$$
$$- \beta \text{tr}(S_l^{-1} S_b) + \theta \|E^T W - T\|_F^2$$
$$\text{s.t.} \|Y^T\|_F = 1.$$  \hspace{1cm} (4)

Since $\beta \text{tr}(YY^T) + \beta \text{tr}(T^T T)$ is a constant, so the above problem can be formulated as follows:

$$\min_w f(W) = \text{tr}(W^T MW - 2NW) + \alpha \|W\|_{2,1},$$  \hspace{1cm} (5)

$$M = XL_p X^T - \beta XFX^T + \theta EE^T,$$
$$N = \theta E^T E^T.$$  \hspace{1cm} (6)

5.2 Optimization Algorithm for U-SSLFS

The derivation of the object function $f(W)$ is:
\[ \frac{\partial f(W)}{\partial W} = 2MW - 2N^T + 2\alpha DW. \] (7)

Where \( D \) is diagonal matrix, the \( i^{th} \) diagonal element is \( 1/(2||W(i,:)||_2^2) \). Set the derivation to 0, then \( W = (M + \alpha D)^{-1}N^T \). Because each element of \( ||W(i,:)||_2 \) corresponds to one feature, the relevance of features is determined according to the value of \( ||W(i,:)||_2 \). In the final step of the algorithm, we sort each feature according to \( ||W(i,:)||_2 \) in descending order and select the top-\( p \) ranked ones. The details of the optimization algorithm are given below.

Next we give the brief convergence analysis of the algorithm. It is easy to verify that \( W_{r+1} \) in line 9 of Algorithm 1 is the solution to the following problem,

\[ W_{r+1} = \arg \min_w tr(W^T(M + \alpha D)W - 2NW) \] (8)

which means that,

\[ tr(W_{r+1}^T(M + \alpha D)W_{r+1} - 2NW_{r+1}) \leq tr(W_{r}^T(M + \alpha D)W_{r} - 2NW_{r}) \] (9)

According to Lemma 3.2 in [12], we have

\[ tr(W_{r+1}^TW_{r+1} - 2NW_{r+1}) + \alpha \sum ||W_{r+1}(i,:)||^2 \leq tr(W_{r}^TW_{r} - 2NW_{r}) + \alpha \sum ||W_{r}(i,:)||^2 \] (10)

that is \( f(W_{r+1}) \leq f(W_{r}) \), which indicates that Algorithm 1 converges to the optimal \( W \).

### 6. Experiments and Results Analysis

In this section, we evaluate the effectiveness of our proposed framework on Flickr and BlogCatalog datasets used in [3]. We compare U-SSLFS with SSLFS (Semi-Supervised Linked samples Feature Selection) [10], Laplacian Score and Fisher Score. The classification experiments are done using the k-NN (k-Nearest Neighbors) algorithm.

In the setup of the experiments, we choose the 9th and the 6th class samples as Universum for the Flickr and BlogCatalog respectively. For the remaining samples of each dataset, 50% serves as the classification samples, other 50% and Universum samples is used to select features. The U-SSLFS algorithm has four parameters: \( \{K, \alpha, \beta, \theta\} \). In the process of selecting parameters, we adjust one parameter while fix the others. The ranges of parameters are: \( K \in \{10, 15, 20, \ldots, 70\} \), \( \alpha \in \{0.01, 0.1, 0.5\} \), \( \beta \in \{0.01, 0.1, 0.5\} \), \( \theta \in \{0.0005, 0.005, 0.05, 0.1, 0.5\} \). The resulting parameter values are: \( \alpha = 0.1, \beta = 0.5, \theta = 5, K = 65 \) for Flickr while \( \alpha = 0.1, \beta = 0.1, \theta = 0.0005, K = 10 \) for BlogCatalog. The SSLFS parameters are set according to [10]. In feature selection process, 30% of the samples are labeled. The ratio of training and testing samples in classification process is 3:1. The classification accuracies along with changed feature numbers are shown in Table 1 and 2.

From the above experiment results, we can obtain the following conclusion: On the Flickr dataset, the classification accuracy of the unsupervised Laplacian Score is obvious lower than other methods. This demonstrates that the label information is very important to feature selection. When the number of selected features is less than 900, SSLFS achieves better performance than Fisher Score, because of

<table>
<thead>
<tr>
<th>Features Algorithm</th>
<th>500</th>
<th>600</th>
<th>700</th>
<th>800</th>
<th>900</th>
<th>1000</th>
<th>1500</th>
<th>2000</th>
<th>2500</th>
<th>5000</th>
<th>8000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher</td>
<td>0.2274</td>
<td>0.0358</td>
<td>0.1444</td>
<td>0.284</td>
<td>0.3389</td>
<td>0.3699</td>
<td>0.5465</td>
<td>0.4308</td>
<td>0.4463</td>
<td>0.3807</td>
<td>0.2768</td>
</tr>
<tr>
<td>Laplacian</td>
<td>0.0274</td>
<td>0.0358</td>
<td>0.0465</td>
<td>0.1098</td>
<td>0.142</td>
<td>0.1826</td>
<td>0.3294</td>
<td>0.3699</td>
<td>0.3055</td>
<td>0.2947</td>
<td>0.2458</td>
</tr>
<tr>
<td>SSLFS</td>
<td>0.3222</td>
<td>0.327</td>
<td>0.3401</td>
<td>0.3317</td>
<td>0.3305</td>
<td>0.3425</td>
<td>0.3532</td>
<td>0.352</td>
<td>0.3353</td>
<td>0.2816</td>
<td>0.2709</td>
</tr>
<tr>
<td>U-SSLFS</td>
<td>0.3341</td>
<td>0.3473</td>
<td>0.3586</td>
<td>0.3636</td>
<td>0.3699</td>
<td>0.3914</td>
<td>0.4224</td>
<td>0.3807</td>
<td>0.3914</td>
<td>0.3902</td>
<td>0.2434</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Features Algorithm</th>
<th>500</th>
<th>600</th>
<th>700</th>
<th>800</th>
<th>900</th>
<th>1000</th>
<th>1500</th>
<th>2000</th>
<th>2500</th>
<th>5000</th>
<th>8000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher</td>
<td>0.0155</td>
<td>0.0191</td>
<td>0.1169</td>
<td>0.2387</td>
<td>0.2852</td>
<td>0.327</td>
<td>0.4737</td>
<td>0.5263</td>
<td>0.5274</td>
<td>0.4356</td>
<td>0.2673</td>
</tr>
<tr>
<td>Laplacian</td>
<td>0.015</td>
<td>0.0191</td>
<td>0.0227</td>
<td>0.0716</td>
<td>0.1098</td>
<td>0.1348</td>
<td>0.321</td>
<td>0.3246</td>
<td>0.364</td>
<td>0.284</td>
<td>0.1241</td>
</tr>
<tr>
<td>SSLFS</td>
<td>0.3151</td>
<td>0.3198</td>
<td>0.3198</td>
<td>0.3186</td>
<td>0.3103</td>
<td>0.3186</td>
<td>0.3258</td>
<td>0.3186</td>
<td>0.3103</td>
<td>0.2518</td>
<td>0.1408</td>
</tr>
<tr>
<td>U-SSLFS</td>
<td>0.3246</td>
<td>0.3532</td>
<td>0.3532</td>
<td>0.3273</td>
<td>0.3675</td>
<td>0.395</td>
<td>0.3711</td>
<td>0.3236</td>
<td>0.3425</td>
<td>0.2706</td>
<td>0.1277</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Features Algorithm</th>
<th>500</th>
<th>600</th>
<th>700</th>
<th>800</th>
<th>900</th>
<th>1000</th>
<th>1500</th>
<th>2000</th>
<th>2500</th>
<th>5000</th>
<th>8000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher</td>
<td>0.015</td>
<td>0.022</td>
<td>0.1229</td>
<td>0.2697</td>
<td>0.3174</td>
<td>0.3437</td>
<td>0.4964</td>
<td>0.5072</td>
<td>0.5191</td>
<td>0.4296</td>
<td>0.1444</td>
</tr>
<tr>
<td>Laplacian</td>
<td>0.0143</td>
<td>0.0215</td>
<td>0.0263</td>
<td>0.0611</td>
<td>0.1205</td>
<td>0.1492</td>
<td>0.216</td>
<td>0.3005</td>
<td>0.3028</td>
<td>0.279</td>
<td>0.2804</td>
</tr>
<tr>
<td>SSLFS</td>
<td>0.2924</td>
<td>0.29</td>
<td>0.2721</td>
<td>0.2733</td>
<td>0.2721</td>
<td>0.2735</td>
<td>0.2749</td>
<td>0.255</td>
<td>0.2351</td>
<td>0.1786</td>
<td>0.1671</td>
</tr>
<tr>
<td>U-SSLFS</td>
<td>0.2601</td>
<td>0.2792</td>
<td>0.2864</td>
<td>0.4105</td>
<td>0.4045</td>
<td>0.3986</td>
<td>0.3807</td>
<td>0.2721</td>
<td>0.2888</td>
<td>0.1659</td>
<td>0.5103</td>
</tr>
</tbody>
</table>
the utilizing of linking information. However, when the number of features ranges from 1500 to 2500, the supervised information of Fisher Score can greatly improve the classification accuracy. From the results of U-SSLFS, it can be seen that the classification accuracy of U-SSLFS outperforms the other methods when the number of features ranges from 700 to 1000. Besides, the U-SSLFS method performs better than SSLFS, due to the Universum samples integrated. However, when the number of selected features is more than 5000, all the methods get lower accuracies. This phenomenon reveals that more selected features than necessary may not improve the performance, but rather do harm to it. Experiments on the BlogCatalog also demonstrate the effectiveness of the proposed algorithm.

7. Conclusion

In this letter, we propose a Universum-based semi-supervised method, named U-SSLFS, which integrates Universum samples into a model. U-SSLFS can utilize the information not only from Universum samples, but also from the linked properties of social media data. Furthermore, U-SSLFS can harness the advantages of a portion of labeled samples and manifold structures of all these samples. Through the learning process, U-SSLFS can finally select the most robust features among the feature space. The experiments on the Flickr and BlogCatalog prove the effectiveness of our proposed method.

Acknowledgments

The authors would want to express the sincere appreciation to Dr. Jiliang Tang from Arizona State University, USA, for his kindly offering of the dataset used in this letter. This work is supported by the National Technology Research and Development Program of China (863 Program) with grant number 2012AA01A510, and the Natural Science Foundation of Jiangsu Province, China (Grant NO.BK20140073).

References


<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Fisher</th>
<th>Laplacian</th>
<th>SSLFS</th>
<th>U-SSLFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
<td>50</td>
<td>100</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>SSLFS</td>
<td>0.372</td>
<td>0.3856</td>
<td>0.321</td>
<td>0.4871</td>
</tr>
<tr>
<td>U-SSLFS</td>
<td>0.2851</td>
<td>0.2982</td>
<td>0.4566</td>
<td>0.4668</td>
</tr>
<tr>
<td>Fisher</td>
<td>0.237</td>
<td>0.2463</td>
<td>0.267</td>
<td>0.274</td>
</tr>
<tr>
<td>Laplacian</td>
<td>0.233</td>
<td>0.2435</td>
<td>0.288</td>
<td>0.2739</td>
</tr>
<tr>
<td>SSLFS</td>
<td>0.2518</td>
<td>0.2869</td>
<td>0.3376</td>
<td>0.3976</td>
</tr>
<tr>
<td>U-SSLFS</td>
<td>0.3856</td>
<td>0.4548</td>
<td>0.4686</td>
<td>0.4751</td>
</tr>
</tbody>
</table>