Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 5874161, 7 pages
http://dx.doi.org/10.1155/2016/5874161

Research Article

Semisupervised Feature Selection with Universum

Junyang Qiu and Zhisong Pan

College of Command Information System, PLA University of Science and Technology, Nanjing 210007, China

Correspondence should be addressed to Zhisong Pan; hotpzs@hotmail.com

Received 6 April 2016; Revised 13 July 2016; Accepted 21 July 2016

Academic Editor: Yaguo Lei

The Universum data, defined as a set of unlabeled examples that do not belong to any class of interest, have been shown to encode some prior knowledge by representing meaningful information in the same domain as the problem at hand. Universum data have been proved effective in improving learning performance in many tasks, such as classification and clustering. Inspired by its favorable performance, we address a novel semisupervised feature selection problem in this paper, called semisupervised feature selection with Universum, that can simultaneously exploit the unlabeled data and the Universum data. The experiments on several UCI data sets are presented to show that the proposed algorithms can achieve superior performances over conventional unsupervised and supervised methods.

1. Introduction

With the rapid accumulation of high-dimensional data such as financial time series, gene expression microarrays, and digital images, feature selection has been a significant preprocessing step to data mining and machine learning. In many real-word applications, feature selection has been very effective in reducing feature space dimension, removing irrelevant and redundant features, and improving learning performance [1, 2].

1.1. Background. Typically, according to whether supervised information is used or not, feature selection methods can be divided into two categories: unsupervised ones and supervised ones. The supervised feature selection methods evaluate feature relevance by the correlation between features and class labels, while unsupervised methods measure feature relevance by the capability of keeping certain properties of the data, for example, the locality or sparsity preserving ability [3, 4]. In general, supervised feature selection methods need the label information of the training sets. However, in some cases it is time consuming and expensive to complete the labeling task and the amount of labeled data is often quite limited. Most conventional supervised feature selection methods cannot work on such situation. To deal with such problem, semisupervised feature selection is a common option when unlabeled data is available since unlabeled data helps model data distribution of the whole data. The popular supervised information used in semisupervised feature selection is class labels.

In fact, besides class labels, there exist other forms of supervised information, for example, the pairwise constraints or the Universum data. The Universum learning concept was first proposed to increase the binary classification with the help of Universum data, the data that do not belong to either target classes but belong to the same domain as the classification problem at hand [5, 6]. Universum data can carry additional valuable prior knowledge from the domain of the problem into the training process. Moreover, Universum based learning can better model the whole data set since Universum data stays in the same domain of learning problem with which we are concerned, while the unlabeled data may be too general and stay outside of the domain. The purpose of Universum provided a new way to alleviate the problem of insufficient labeled samples.

Universum learning has attracted many scholars since it was proposed and it has been used for classification, clustering, and other machine learning scenarios and obtained favorable improvements with the help of Universum data [5–12]. In [5], Universum was firstly proposed to enhance the performance of support vector machine; in the paper USVM (Universum Support Vector) was introduced to leverage the

Universum by maximizing the number of observed contractions. Experiments showed that the performance of USVM outperformed SVM. Besides, the results also confirmed that the Universum can be an important instrument for boosting performance, especially in the small sample size regime. In 2012, Universum was introduced to improve the classification performance of TSVM (Twin Support Vector Machine); the results demonstrate that Universum samples are helpful to improve the generalization ability of the model and the training time of Universum based TSVM is faster than USVM [7]. Universum learning was extended to dimensionality reduction by incorporating it with linear discriminant analysis; the proposed method was termed as Universum linear discriminant analysis (ULDA) which aimed to find discriminant directions by maximizing the distance between two target classes and simultaneously minimizing the distance between the Universum and the mean of the target classes [8]. A novel semisupervised classification problem, called semisupervised Universum, that can simultaneously utilize the labeled data, unlabeled data, and the Universum data to improve the classification performance was addressed in [9]. In 2012, Weston's principle of maximal contradiction on Universum data was extended to boosting learning; however, as pointed in the paper, in some scenarios poorly generated Universum data may not help [12]. A maximum margin clustering method was proposed to model both target samples and Universum samples for document clustering and the method performed substantially better than state-of-the-art methods in most cases [10]. Universum was also introduced to multiview learning and obtained satisfactory performance.

In terms of data acquisition, Universum data can be obtained more easily with different methods [5–7, 9]. The generation methods can be divided into three categories, including $\mathcal{U}_{\text{rest}}$ (other samples that are not included in the learning tasks serve as Universum data), $\mathcal{U}_{\text{mean}}$ (each Universum is generated by first randomly selecting two samples from two different classes and then combined with a specific combination coefficient), and $\mathcal{U}_{\text{gene}}$ (generate Universum according to the statistic of the labeled and unlabeled samples). What is more, an algorithm was proposed about how to evaluate and select the informative and useful Universum data in [13].

1.2. Contributions and Novelty. However, to the best of our knowledge, there is few work about introducing Universum learning to feature selection. Inspired by the favorable performance of Universum data in guiding learning tasks, we address a novel semisupervised feature selection problem in this paper; the main contributions can be listed as follows:

(i) A semisupervised feature selection technique with Universum is proposed which can simultaneously exploit the unlabeled data and the Universum data.

(ii) Based on Variance Score, we integrate Universum into it and introduce a Universum based Variance Score algorithm so as to select features with larger variances as well as maximizing the margins between Universum and target samples while minimizing the margins on the Universum.

(iii) An improved Laplacian Score named ULS is proposed to select features with stronger locality preserving ability as well as exploiting the supervised information provided by the Universum data.

(iv) What is more, we add the Universum data into the Sparsity Score in order to combine the sparse structure of data and the prior knowledge encoded by the Universum data.

The three improved semisupervised methods can inherit the merits of traditional unsupervised methods as well as the valuable prior information carried by the Universum data so as to select more discriminative features. Experiments carried out on several UCI data sets validate the effectiveness of Universum in enhancing the performance of feature selection.

2. Traditional Feature Selection Algorithms

In this selection, we briefly present several algorithms popularly used in feature selection, including Variance Score [14], Laplacian Score [15], Sparsity Score [16], and Fisher Score [14]. Among them, the former three are unsupervised, while the last one is supervised.

Variance Score uses the variance along a feature dimension to evaluate its representative power and those features with the maximum variance are selected. The rth feature's score can be computed as follows [14]:

$$\text{VS}_r = \frac{1}{m}\sum_{i=1}^{m}\left(f_{ri} - \mu_r\right)^2, \tag{1}$$

where f_{ri} is the rth feature of the ith sample $\mathbf{x}_i$ and μ_r is the mean of the rth feature, $i = 1, \ldots, m$, $r = 1, \ldots, n$.

Laplacian Score aims to select features not only with larger variances but also with stronger locality preserving ability. Laplacian Score is under the assumption that data with the same label are close to each other. The rth feature's score can be computed by minimizing the following formula [15]:

$$\text{LS}_r = \frac{\sum_{i,j}\left(f_{ri} - f_{rj}\right)^2 \mathbf{Q}_{i,j}}{\sum_i \left(f_{ri} - \mu_r\right)^2 \mathbf{D}_{i,i}}. \tag{2}$$

Here $\mathbf{D}$ is a diagonal matrix and $\mathbf{D}_{i,i} = \sum_j \mathbf{Q}_{i,j}$; the definition of $\mathbf{Q}_{i,j}$ can be defined as follows:

$$\mathbf{Q}_{i,j} = \begin{cases} \exp\left(-\dfrac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{\sigma^2}\right), & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are neighbors} \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where σ is a constant to be set. In this paper, the parameter σ is set to default value 1 according to [15].

Another unsupervised feature selection algorithm Sparsity Score is based on sparse representation; it aims to identify

an optimal feature subset that is most useful in capturing the intrinsic sparse structure of data. The objective function of Sparsity Score can be defined as the following formulation [16]:

$$\mathrm{SS}_r = \sum_{i=1}^{m} \left(f_{ri} - \sum_{i=1}^{m} \widehat{s}_{ij} f_{rj} \right)^2, \tag{4}$$

where $\widehat{s}_{ij}$ is the estimated value of s_{ij} and s_{ij} indicates the jth sample's contributions to the reconstruction of $\mathbf{x}_i$; thus $\mathbf{x}_i$ can be reconstructed by other samples in the training data as $\mathbf{x}_i = s_{i1}\mathbf{x}_1 + \cdots + s_{i,i-1}\mathbf{x}_{i-1} + s_{i,i+1}\mathbf{x}_{i+1} + \cdots + s_{im}\mathbf{x}_m$. An optimal feature for $\mathbf{x}_i$ implies the following equality: $f_{ri} = s_{i1}f_{r1} + \cdots + s_{i,i-1}f_{r,i-1} + s_{i,i+1}f_{r,i+1} + \cdots + s_{im}f_{rm}$.

With full class labels, the supervised Fisher Score prefers features with best discriminant ability. The Fisher Score of the rth feature FS_r, which should be maximized, is computed as follows [14]:

$$\mathrm{FS}_r = \frac{\sum_{i=1}^{c} n_i \left(\mu_r^i - \mu_r \right)^2}{\sum_{i=1}^{c} n_i \left(\sigma_r^i \right)^2}. \tag{5}$$

Here c is the number of classes and n_i is number of samples in class i; μ_r^i and $(\sigma_r^i)^2$ denote the mean and variance of class i corresponding to the rth feature.

3. Semisupervised Feature Selection with Universum

Here we formulate the Universum data guided feature selection as follows: given a set of data samples $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]$, m is the number of samples. The Universum data can be donated as $\mathbf{U} = [\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_u]$ while u is the number of Universum data samples.

Based on Variance Score, we are now in the position to derive a semisupervised Variance Score feature selection algorithm, called Variance Score with Universum (UVS). The intuition is that the Universum data can serve as supervised information since they are known not belonging to any target classes. UVS prefers those features with larger variances but also prefers to selecting features which can maximize the margins between Universum and target samples while minimize the margins on the Universum simultaneously. The rth feature score of UVS, which should be maximized, is computed as follows:

$$\begin{aligned} \mathrm{UVS}_r = {} & \frac{\alpha}{m * u} \sum_{\mathbf{x}'_i \in \mathbf{U}} \sum_{\mathbf{x}_j \in \mathbf{X}} \left(f_{ri} - f_{rj} \right)^2 \\ & - \frac{\beta}{u} \sum_{\mathbf{x}'_i, \mathbf{x}'_j \in \mathbf{U}} \left(f_{ri} - f_{rj} \right)^2 + \frac{1}{m} \sum_{i=1}^{m} \left(f_{ri} - \mu_r \right)^2, \end{aligned} \tag{6}$$

where α and β are two scaling parameters, whose functions are to balance the contributions of the three terms in (6). The first term of (6) constrains the selected features to maximize the margins between Universum and target samples. On the contrary, the second term aims to minimize the margins among Universum. The last term expresses the variance between the selected features, which is equivalent to the Variance Score criterion.

Now we give the formal representation of Laplacian Score with Universum (ULS). ULS can select features with stronger locality preserving ability as well as exploiting the supervised information provided by the Universum data. The objective function of ULS can be maximized as follows:

$$\begin{aligned} \mathrm{ULS}_r = {} & \frac{\alpha}{m * u} \sum_{\mathbf{x}'_i \in \mathbf{U}} \sum_{\mathbf{x}_j \in \mathbf{X}} \left(f_{ri} - f_{rj} \right)^2 \\ & - \frac{\beta}{u} \sum_{\mathbf{x}'_i, \mathbf{x}'_j \in \mathbf{U}} \left(f_{ri} - f_{rj} \right)^2 \\ & - \frac{\sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}} \left(f_{ri} - f_{rj} \right)^2 \mathbf{Q}_{i,j}}{\sum_{\mathbf{x}_i \in \mathbf{X}} \left(f_{ri} - \mu_r \right)^2 \mathbf{D}_{i,i}}. \end{aligned} \tag{7}$$

The motivation of the former two terms of (7) is to use Universum to enhance performance of feature selection. The last term aims to improve the locality preserving ability of selected features.

Similarly, we add the Universum data into the Sparsity Score and propose a semisupervised Sparsity Score with Universum, called USS. USS combines the sparse structure of data and the prior knowledge encoded by the Universum data. The rth feature score of USS can be obtained as follows:

$$\begin{aligned} \mathrm{USS}_r = {} & \frac{\alpha}{m * u} \sum_{\mathbf{x}'_i \in \mathbf{U}} \sum_{\mathbf{x}_j \in \mathbf{X}} \left(f_{ri} - f_{rj} \right)^2 \\ & - \frac{\beta}{u} \sum_{\mathbf{x}'_i, \mathbf{x}'_j \in \mathbf{U}} \left(f_{ri} - f_{rj} \right)^2 \\ & - \sum_{i=1}^{m} \left(f_{ri} - \sum_{i=1}^{m} \widehat{s}_{ij} f_{rj} \right)^2. \end{aligned} \tag{8}$$

The motivation of the former two terms of (8) is similar to (6) and (7). The third term is used to preserve the sparse structure of the data.

To sum up, the advantages of the proposed three methods are very clear. Firstly, they inherit the merits of traditional unsupervised methods as well as the valuable prior information carried by the Universum data so as to select more discriminative features. Secondly, no label information is needed, which saves the great cost of labeling task. Besides, the Universum samples are easy to obtain, which makes the proposed methods have great potentials in those applications where labeled data is rare. The disadvantages are that the proposed methods cost more running time than the corresponding unsupervised methods since we add the Universum constraint terms in the objective functions.

4. Experiments and Results Analysis

To evaluate the performance of our proposed algorithms, we apply them on five UCI data sets, that is, ionosphere, sonar,

Table 1: Statistics information of the data sets.

Data sets	Size	Dimension	Class
Ionosphere	351	34	2
Sonar	208	60	2
Wine	178	13	3
Zoo	101	16	7
Vehicle	848	18	4

wine, zoo, and vehicle. Some statistics information of the data sets can be found in Table 1.

4.1. Experimental Setting. We compare our proposed algorithms (UVS, ULS, and USS) with existing unsupervised algorithms including Variance Score (VS), Laplacian Score (LS), Sparsity Score (SS), and supervised Fisher Score (FS). The LibSVM package and a 10-fold cross validation strategy are adopted to perform classification and compute the average classification accuracy, respectively. Each classification experiment is repeated 10 times and the average accuracy is the final result. In this paper, we consider two ways to generate the Universum data samples. For the two-class data sets, each Universum is generated by first randomly selecting two samples from two different categories and then combined with a specific combination coefficient (here the combination coefficient is set to 0.5). The generated Universum plus 50% of the data sets is used in features selection process while the other 50% serves as the classification samples. For the multiclass data sets, each time one class of samples is treated as Universum; for the remaining samples of each data set, 50% is employed in the classification task; the other 50% and the selected Universum are used to select features.

The proposed algorithms have two parameters: α and β. We search one parameter while fixing the other one. The ranges of parameters are set as follows: $\alpha = [0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100]$ and $\beta = [0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100]$.

4.2. Experimental Results. In order to show the changing trend of the accuracy along with the parameters, in Figure 1 we plot the accuracy versus different values of parameters of UVS on ionosphere data set. It is easy to see that the accuracy first rises and then declines with the increase of α and β. The accuracy reaches its peak when α is around 1 while β is around 10.

Tables 2 and 3 give the best accuracy of different algorithms on ionosphere, sonar, and wine. For ionosphere and sonar, we add 150 and 200 constructed Universum samples, respectively. As to wine, we select each class sample as Universum. Here the numbers in parentheses after the accuracy represent the optimal features. The averaged accuracy versus different numbers of selected features and different class of samples used as Universum on vehicle is summarized in Table 4. The three tables indicate that the performances of the improved algorithms are significantly better than that of unsupervised algorithms, even outperforming the supervised Fisher Score. This verifies that Universum is very useful in learning feature scores. It is noted that sometimes the proposed semisupervised methods get the better results. We try to give the following reasons. Firstly, the proposed methods not only exploit the prior information carried by

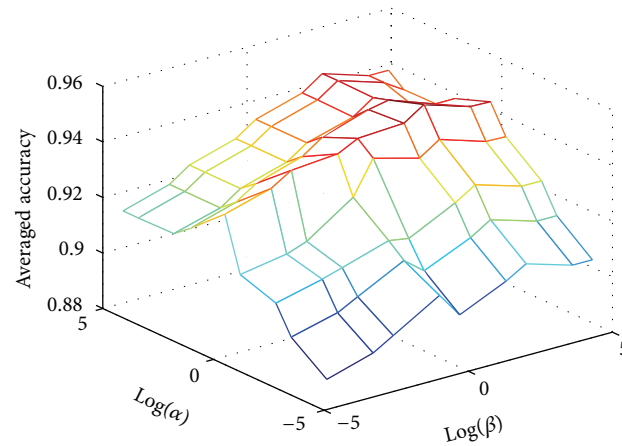

Figure 1: Accuracy versus different values of parameters.

Table 2: The best accuracy of different algorithms.

Algorithm	Data sets	
	Ionosphere (150)	Sonar (200)
VS	0.9489 (23)	0.8269 (39)
LS	0.9659 (27)	0.8269 (51)
SS	0.9432 (17)	0.8269 (39)
FS	0.9659 (13)	0.8365 (25)
UVS	0.9659 (23)	0.8269 (39)
ULS	0.9659 (23)	**0.8558 (49)**
USS	**0.9716 (25)**	0.8269 (39)

Table 3: The best accuracy on wine data set.

Algorithm	Wine (1)	Wine (2)	Wine (3)
VS	0.8500 (9)	0.8704 (1)	0.8769 (3)
LS	0.9667 (6)	0.9630 (7)	0.9077 (6)
SS	0.8833 (8)	0.9259 (11)	0.8923 (11)
FS	0.9833 (8)	0.9444 (1)	**0.9385 (1)**
UVS	0.9000 (12)	0.8889 (2)	0.9077 (6)
ULS	**0.9833 (7)**	0.9815 (9)	0.9077 (6)
USS	0.9000 (12)	**0.9815 (8)**	**0.9385 (1)**

Table 4: Averaged accuracy on different numbers of selected features.

Algorithm	Vehicle (1)	Vehicle (2)	Vehicle (3)	Vehicle (4)
VS	0.8081	0.7787	0.6001	0.5792
LS	0.8566	0.7698	0.5582	0.5582
SS	0.7979	0.7771	0.6109	0.5475
FS	0.7886	0.8072	0.6014	**0.6163**
UVS	0.8302	0.7802	0.6079	0.5943
ULS	**0.8587**	**0.8323**	0.5913	0.5972
USS	0.8356	0.8083	**0.6139**	0.5753

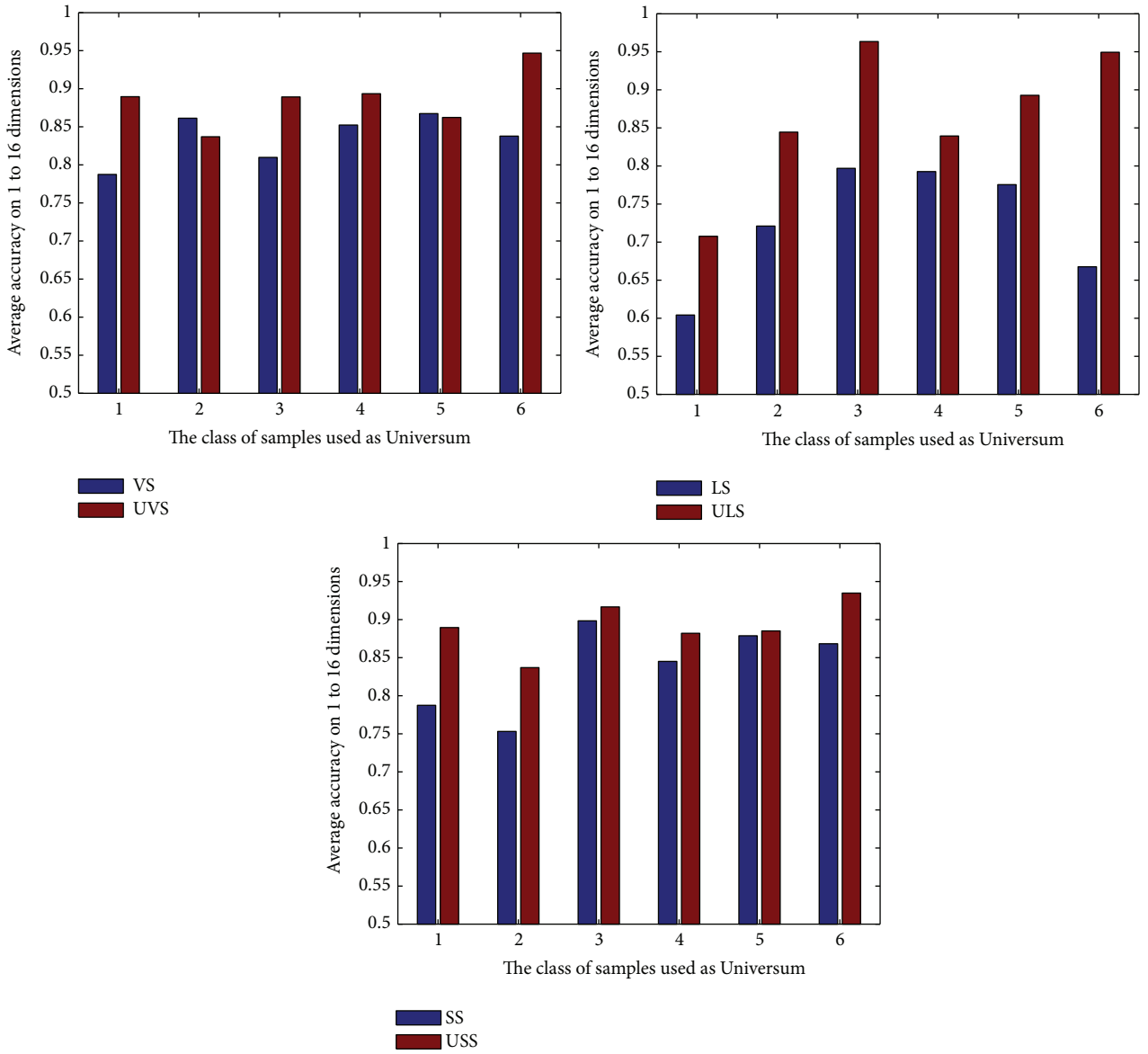

Figure 2: Accuracy versus different class of samples used as Universum.

Universum, but also preserve the structure (such as the local or sparse structure) of the data. However, the supervised Fisher Score only utilizes the label information of the data. Secondly, there does not exist the feature selection algorithm which is best, but the most appropriate for specific data set; we should select the most suitable method according to the property of the data.

To investigate the effectiveness of Universum in guiding feature selection, we compare the original algorithm with its corresponding improved algorithm on zoo in Figure 2. As shown in the figure, the performances of the improved algorithms are significantly superior to the original algorithms in most cases. This verifies again the usefulness of Universum in feature selection.

Figure 3 shows the plots for accuracy versus different numbers of selected features on zoo. Here we randomly select the 1st- and 6th-class samples as Universum. Obviously, the proposed methods obtain a satisfying accuracy and reach their peaks with only a few features being selected.

To uncover the influence of the Universum data on feature selection, we plot the distribution of the former six selected features and the feature score curves of different algorithms in Figures 4 and 5. It is surprising to see that the selected feature distribution of the improved algorithms (UVS, ULS, and USS) is similar to each other and dissimilar to the original algorithms (VS, LS, SS, and FS). The underlying reason is that the prior knowledge encoded by Universum reflects certain kind of structure information of the data and dominates the feature selection process. The feature score curves of the improved algorithms are also similar to each other. This phenomenon reveals that the constraint information of Universum plays a major role in feature selection while

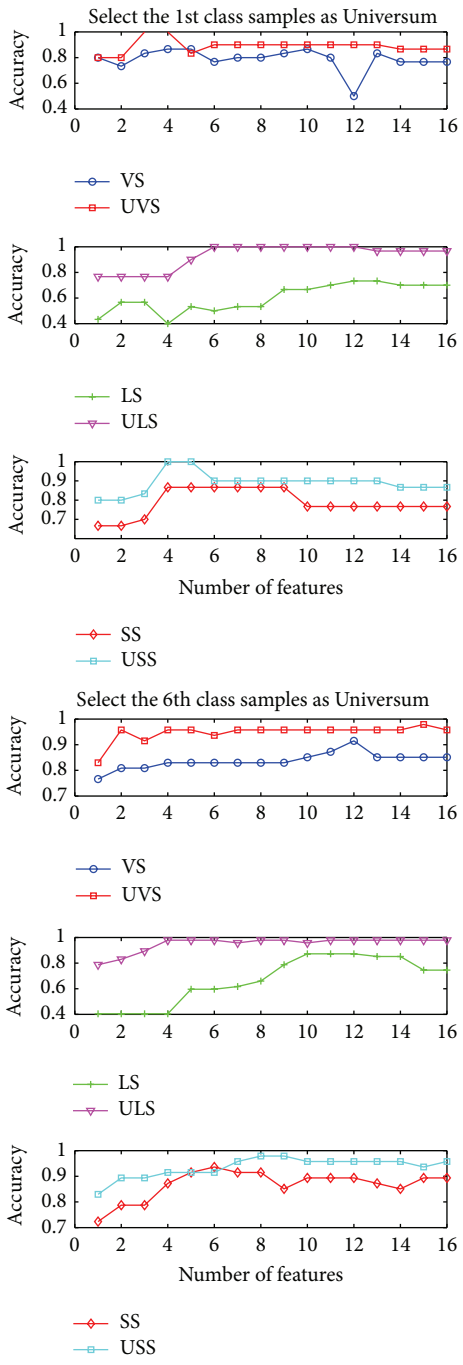

Figure 3: Accuracy versus different numbers of selected features on zoo.

other constraint information (the samples variance, locality structure, or sparse structure information) plays a secondary role. Meanwhile, we find that the feature score curves of the original algorithms (VS, LS, SS, and FS) are different to each other, because they use different criterion function to compute feature score.

5. Conclusions

In this paper, we address a novel semisupervised feature selection problem, called semisupervised feature selection

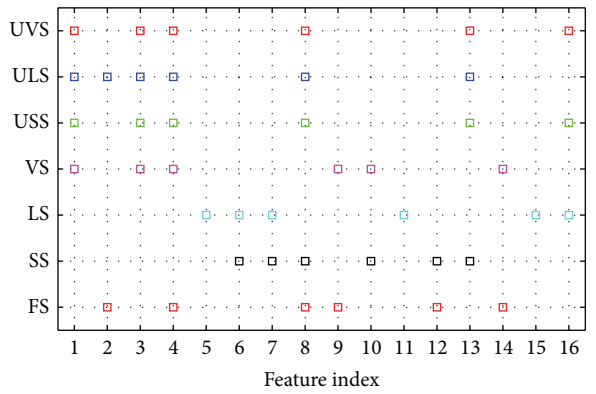

Figure 4: The distribution of the former six selected features.

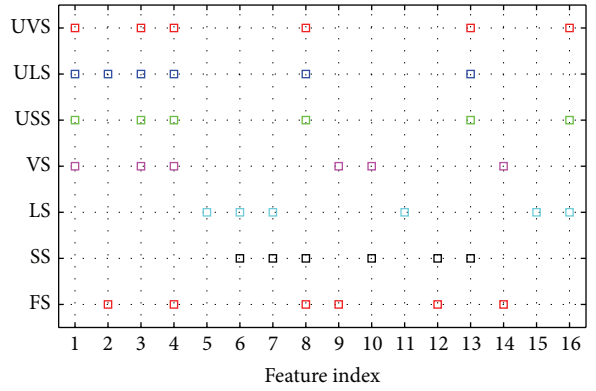

Figure 5: The feature score curve of different algorithms.

with Universum. Three new score functions are presented to evaluate features based on Universum. The proposed algorithms can inherit the merits of traditional unsupervised methods as well as exploit the valuable prior information carried by the Universum data so as to select more discriminative features from the high-dimensional data. The experiments on five UCI data sets demonstrate that the improved algorithms achieve similar or higher accuracy to supervised Fisher Score and significantly outperform unsupervised methods. Finally, because in many real applications generating Universum data is much easier than obtaining labeled or unlabeled data, our improved algorithms have great potentials in those applications.

In the future, how to generate more discriminative Universum will be an interesting research idea. Besides, our work will also include Universum motivated active learning and metric learning. What is more, how to quantitative evaluate the effectiveness remains to be solved.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The authors' work is supported by the National Natural Science Fund with Grant no. 61473149 and China Postdoctoral Science Foundation Project with Grant no. 2015M572731.

References

[1] D. Zhang, S. Chen, and Z.-H. Zhou, "Constraint Score: a new filter method for feature selection with pairwise constraints," *Pattern Recognition*, vol. 41, no. 5, pp. 1440–1451, 2008.

[2] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[3] X. Niyogi, "Locality preserving projections," in *Neural Information Processing Systems*, vol. 16, p. 153, MIT, 2004.

[4] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognition*, vol. 43, no. 1, pp. 331–341, 2010.

[5] J. Weston, R. Collobert, F. Sinz, L. Bottou, and V. Vapnik, "Inference with the universum," in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pp. 1009–1016, ACM, Pittsburgh, Pa, USA, June 2006.

[6] O. Chapelle, A. Agarwal, F. H. Sinz, and B. Schölkopf, "An analysis of inference with the universum," in *Advances in Neural Information Processing Systems*, pp. 1369–1376, 2007.

[7] Z. Qi, Y. Tian, and Y. Shi, "Twin support vector machine with universum data," *Neural Networks*, vol. 36, pp. 112–119, 2012.

[8] X. H. Chen, S. C. Chen, and H. Xue, "Universum linear discriminant analysis," *Electronics Letters*, vol. 48, no. 22, pp. 1407–1409, 2012.

[9] D. Zhang, J. Wang, F. Wang, and C. Zhang, "Semi-supervised classification with universum," in *Proceedings of the 8th SIAM International Conference on Data Mining (SDM '08)*, pp. 323–333, Atlanta, Ga, USA, April 2008.

[10] D. Zhang, J. Wang, and L. Si, "Document clustering with universum," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*, pp. 873–882, ACM, Beijing, China, July 2011.

[11] Z. Wang, Y. Zhu, W. Liu, Z. Chen, and D. Gao, "Multi-view learning with Universum," *Knowledge-Based Systems*, vol. 70, pp. 376–391, 2014.

[12] C. Shen, P. Wang, F. Shen, and H. Wang, "UBoost: boosting with the universum," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 825–832, 2012.

[13] S. Chen and C. Zhang, "Selecting informative universum sample for semi-supervised learning," in *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI '09)*, vol. 6, pp. 1016–1021, Pasadena, Calif, USA, July 2009.

[14] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, New York, NY, USA, 1995.

[15] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS '05)*, pp. 507–514, December 2005.

[16] M. Liu, D. Sun, and D. Zhang, "Sparsity Score: a new filter feature selection method based on graph," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR '12)*, pp. 959–962, IEEE, Tsukuba, Japan, November 2012.