A Dictionary-Based Method for Classification with Universum Data

Zhiyong Che
Department of Automation,
Guangdong University of Technology,
Guangzhou, China
952860969@qq.com

Bo Liu*
Department of Automation,
Guangdong University of Technology,
Guangzhou, China
csbliu@gmail.com

Yanshan Xiao
Department of Computer Science,
Guangdong University of Technology,
Guangzhou, China
xiaoyanshan@gmail.com

ABSTRACT

In fact, the collected examples included the third-class examples, they do not belong to positive samples or negative samples, which are referred as the Universum data. And Universum data can make better performance for the classifier. In this paper, a dictionary-based method for classification with Universum data is proposed to construct a unified model. In the proposed method, we embed the dictionary and Universum data to construct a unified framework, and the Universum data is introduced into the framework by the ϵ-insensitive loss. For the optimization, the SVD algorithm and gradient-based optimization methods are utilized to alternately optimize and update the dictionary, and the Lagrangian function is used to iteratively optimize the unified framework to obtain the classifier. Finally, extensive experiments are conducted on the benchmark datasets to evaluate the performance of the proposed U-DL method and baselines. The results have shown that the proposed U-DL method makes better performance than previous methods.

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**;

KEYWORDS

Dictionary learning, Universum data

ACM Reference Format:
Zhiyong Che, Bo Liu*, and Yanshan Xiao. 2021. A Dictionary-Based Method for Classification with Universum Data. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3487075.3487115

1 INTRODUCTION

In machine learning, the researchers always pay attention to the samples with label or samples without labels for supervised learning and semi-supervised learning. However, the existing additional examples that do not belong to positive samples or negative samples are ignored, which can provide the help to the classification problem. The additional examples included in collected data are termed as third-class examples, we refer them as Universum data [1, 2], in which they are neither positive samples nor negative samples. The prior knowledge corresponding to the classification problem, which is obtained from the Universum data, is utilized to assist the classification model. To date, Universum data has been widely studied and has made great achievement in Universum data for supervised learning [3–5] and semi-supervised learning [6–8].

For supervised learning, Vapnik et al. [3] first incorporate the standard SVM and Universum data to construct U-SVM model, and the model obtains the maximum number of Universum data closed to the classification hyperplane by adjusting the classification hyperplane. After the U-SVM model is proposed, researchers have pay attention to study the work. The authors in [5] utilize least squares SVM to replace SVM to construct a new U-LSSVM model, in which the Fisher discri-minant analysis and PCA methods are introduced to solve the classification problem of the Universum data in the model. The above-mentioned methods construct the unified frame-work by incorporating the SVM and Universum data. However, the solutions of optimization method are complex which increase the computation time cost and weaken the performance. Thus, a new variant of U-SVM is proposed in [1], it embeds the Universum data into twin SVMs to build the unified U-TSVM model, in which two Hinge Loss functions are utilized to deal with the Universum data and assist the classification of the model. Further, in order to improve the generalization performance, Xu et al. [2] propose a new U-LSTSVM model and minimize the structural risk by applying a regularization term into the unified model. And two small sized linear equations are utilized to replace a larger sized quadratic programming problem to optimize the classification model.

For the semi-supervised learning, the work in [6] builds a new Universum-data-based semi-supervised model, and utilizes the graph-based method to obtain the prior knowledge embedded in Universum data, in which prior knowledge is related to the classification problem. Based on the extended U-SVM algorithm, the authors in [8] propose a new semi-supervised algorithm to handle the classification problem with Universum data. In the paper, the Universum data without labels are generated from the trained process, and classification error of samples with labels and contradiction of Universum data are adopted to train the classifier. Tian et al. [7] embed the self-constructed Universum data into SVM to design a semi-supervised learning model, in which Universum data constructed by the classifier are utilized to assist the classification model to distinguish positive and negative class from unlabeled data.

Existing methods introduce the Universum data into the model to construct a unified framework by the ϵ-insensitive loss. With

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CSAE 2021, October 19–21, 2021, Sanya, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487115

this strategy, these prior knowledge related to the classification problem that embedded in Universum data are more flexible for us to utilize. In our proposed method, the strategy is also adopted to introduce the Universum data into SVM to design a unified model. In addition, collected data always have the problems of data noise, data redundancy and uncertainty. However, the dictionary method can be utilized to solve these problems. The authors in [11] propose a projective and discriminative dictionary learning method for high-dimensional process monitoring. Huang et al. [12] propose a method based on distributed dictionary learning for high-dimensional process monitoring. Therefore, it is necessary for us to study that embedding the dictionary into the unified model.

In this paper, we focus on the problem that dictionary learning and Universum data are considered to construct a unified model. A dictionary-based method for classification with Universum data is proposed, which is termed as U-DL. In the proposed U-DL method, we first introduce the Universum data into the SVM model by the ϵ-insensitive loss function, which can provide the prior knowledge related to the classification problem to improve the performance. We then embed the dictionary into the model to construct a unified framework, in which the dictionary can solve the problem of the data noise. In all, the main contributions of our model can be listed as follows:

- We propose a dictionary-based method for classification with Universum data, in which we first introduce Universum data into the model by ϵ-insensitive loss, which can improve the performance. We then embed the dictionary into the U-SVM model to design a unified framework.
- For the optimization, considering the difficulty of optimizing dictionary, the gradient-based optimization and SVD algorithm are adopted to alternately optimize and update the dictionary. Besides, the Lagrangian function is utilized to alternately optimize the model to obtain the classifier.
- Extensive experiments are conducted to on the benchmark datasets, and the results have shown that the proposed U-DL method obtains better performance than baselines.

The rest of the paper is organized as follows. The related work of dictionary learning is presented in Section 2, the objective function of the proposed U-DL method is given in Section 3, and we conduct the experiments in Section 4. Finally, we present the conclusion in Section 5.

2 RELATIVE WORK

Dictionary learning (DL), it is a representation method, in which the given sample X is utilized the sparse matrix learned by the dictionary to approximate represent. To date, DL method has been widely studied by considerable researchers, and has made great success in person re-identification [13], face recognition [14] and image classification [15]. For example, Hu et al. [15] incorporate the DL and nonlinear structure of samples to build a nonlinear learning model for image classification, and utilize the labels of training samples to learn a class-specific dictionary to extend as supervised NDL method.

Recently, extensive researches have been done on dictionary learning and has made great success in many fields. The existing DL methods are grouped into two categories from the study on the dictionary learning: unsupervised dictionary learning [16, 17] and discriminant dictionary learning [18–20].

For the unsupervised dictionary learning method [16, 17], it cannot utilize the prior knowledge of the given data, i.e. the label information of the training samples. In unsupervised learning method, it obtains an approximate dictionary by minimizing the residual error of the original data. For example, Mai et al. [16] embed the reconstructive and discriminative capabilities into the learned dictionary to build an unsupervised dictionary learning model, which can improve the discriminative abilities, and the different sub-dictionaries are learned by the training patches. Yang et al. [17] incorporate the analysis dictionary and synthesis dictionary to construct a novel unsupervised dictionary learning model, in which the discrimination representation of universality and particularity are represented by the learned dictionaries. Besides, representation of universality is the label preserving term.

For the discriminant dictionary learning method [18–20], it is referred as the supervised method because it can utilize the prior information embedded in the training samples to assist the classification problem. The work in [18] introduces the classification error and label consistency constraint to construct a discriminant dictionary model, and the class labels corresponding to the dictionary are utilized to obtain a learned discriminative dictionary. This method just constructs a small size dictionary which cannot ensure the discriminative capability and lead to computationally expensive. In order to improve the discriminative ability of the dictionary, the discriminative terms are introduced to the dictionary learning model. Guo et al. [19] embed a regularization term into the analysis dictionary learning model, in which it is utilized to ensure consistent between the code and dictionary. And a triplet-constraint-based local topology is adopted to embed the discriminative information into the training samples, which can improve the discriminative ability of the learned dictionary. Further, in order to improve the performance, an incoherence promoting term is introduced into the discriminative dictionary learning model, in which it can ensure that the learned class-specific dictionaries are independent. Based on the existing dictionary methods, Gu et al. [20] utilize a pair of synthesis and analysis dictionaries to design a discriminative DL framework, in which the l_0 or l_1 -norm sparsity constraint is utilized to ensure the sparse ability of training samples and discriminative ability. To date, discriminative dictionary learning methods have been proposed to promote the discriminative power of the learned dictionary.

3 THE OBJECTIVE FUNCTION

Suppose we have a set of data set$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \cup \{x^*_{n+1}, x^*_{n+2}, \ldots, x^*_m\}$, in which $u_s = (x^*_{n+1}, x^*_{n+2}, \ldots, x^*_m)$ is the Universum data, where $x_u \in \mathbb{R}^N$, $u = n+1, n+2, \ldots, m$. y_i is utilized to represent as the label of samples, where $y_i \in \{1, -1\}$. If x_i belongs to plus class, then y_i equals to 1; otherwise y_i equals to -1. Given arbitrary i-th sample, and we use the dictionary to make the feature learning of the given samples. $\phi(x_i) = S_i = [s_1, s_2, \ldots, s_m] \in \mathbb{R}^n$ denotes the operation of mapping. For example, $\phi : x \rightarrow S$. During the process of test or verification, we have the optimal sparse code S_i of the

given sample x_i, which is expressed as follows:

$$S_i = arg\min_{S_i} X - DS_i{}_F^2 + \tau S_{i2,1} \quad s.t.\ d_i{}_2^2 \le 1 \tag{1}$$

Suppose we have arbitrary i-th sample x_i and its sparse code representationS_i, we then construct a SVM-based binary classifier, and the formulation expresses as follows:

$$y_i = sgn\left(g\left(S_i\right)\right) = \omega S_i + b \tag{2}$$

We present the detail formulation of the proposed U-DL method in this section. Firstly, the Universum data is introduced into the learning model by ε-insensitive loss to improve the performance. Second, we embed the dictionary to construct a unified framework to reduce the effect of the noise and enhance the sparsity of the samples. The unified learning model is proposed as follows:

$$\begin{aligned}\min_{\omega,D,S,b,\xi} \frac{1}{2}\omega^2 + C_1\sum_i \xi_i + C_2\sum_{n+1}^{m}\left(\varphi_m + \varphi_m^*\right) + \frac{\theta}{2}X - DS_F^2 + \tau S_{2,1}\\ s.t. d_i{}_2^2 \le 1, y_i\left(\omega^T S_i + b\right) \ge 1 - \xi_i, \xi_i \ge 0, \forall i,\\ -\epsilon - \varphi_m^* \le \omega^T S_m + b\_m \le \epsilon + \varphi_m\ ,\ \ u = n+1, \cdot s, m.\end{aligned} \tag{3}$$

C_1 and C_2 are the penalty parameters. τ is a scalar constant and θ is the control parameter. The first constraint is the basic constraints for the stand SVM model. And the second constraint denotes ε-insensitive loss for Universum data, which is used to maximize the margin between the separating hyperplanes, and make Universum data distribute near the hyperplane. Next, we have a detail optimization of the objective function.

4 OPTIMIZATION ALGORITHM

1). Fixed D and S , optimize ω , b and ξ

Through the operation, the optimization formulation can be expressed as follows:

$$\begin{aligned}\min_{\omega,D,S,b,\xi} \frac{1}{2}\omega^2 + C_1\sum_i \xi_i + C_2\sum_{n+1}^{m}\left(\varphi_m + \varphi_m^*\right)\\ s.t. \quad y_i\left(\omega^T S_i + b\right) \ge 1 - \xi_i, \xi_i \ge 0, \forall i,\\ -\varepsilon - \varphi_m^* \le \omega^T S_m + b_m \le \varepsilon + \varphi_m, u = n+1, \cdot s, m.\end{aligned} \tag{4}$$

The Lagrangian multipliers are introduced into Eq. (5), in which α, β, and λ are the Lagrangian multipliers for the optimization formulation. We obtain the Lagrangian function as follows:

$$\begin{aligned}L = \frac{1}{2}\omega^2 + C_1\sum_i \xi_i + C_2\sum_{n+1}^{m}\left(\varphi_m + \varphi_m^*\right)\\ + \sum_i \alpha^T 1 - \xi_i - y_i\left(\omega^T S_i + b\right)\\ - \sum_u \beta^T\left\{\omega^T S_m + b_m + \varepsilon + \varphi_m^*\right\}\\ + \sum_u \lambda^T\{\omega^T S_m + b_m - \varepsilon - \varphi_m\}\\ + \gamma^T\xi_i + \mu^T\varphi_m + \rho^T\varphi_m^*\end{aligned} \tag{5}$$

Then considering the Karush-Kuhn-Tucker (K.K.T) conditions, we obtain the formulas as follows:

$$\begin{aligned}\frac{\partial L}{\partial \omega} = \omega - \alpha y_i^T S_i^T - \beta S_m^T + \lambda S_m^T = 0\\ \frac{\partial L}{\partial b} = \alpha y_i^T - \beta + \lambda = 0\\ \frac{\partial L}{\partial \xi_i} = C_1 - \alpha + \gamma = 0\\ \frac{\partial L}{\partial \varphi_m} = C_2 - \lambda + \mu = 0\\ \frac{\partial L}{\partial \varphi_m^*} = C_2 - \beta + \rho = 0\end{aligned} \tag{6}$$

By introducing the Lagrangian function and taking the Karush-Kuhn-Tucker (K.K.T) conditions into account, we get the partial derivative of $\omega, b, \xi_i, \varphi_m$ and φ_m^*, and set the partial derivatives to zero. We obtain ω and b as follows:

$$\omega = \alpha y_i^T S_i^T + \beta S_m^T - \lambda S_m^T \tag{7}$$

$$b = y_i - \left(\alpha y_i^T S_i^T + \beta S_m^T - \lambda S_m^T\right) S \tag{8}$$

Considering the K.K.T conditions, the dual problem of is written as Eq. (9):

$$\begin{aligned}\max_{\alpha,\beta,\lambda} \sum_i\sum_j \alpha_i\alpha_j y_i y_j S_i S_j - \sum_i\sum_j (\beta-\lambda)_i(\beta-\lambda)_j S_i S_j\\ - \sum_i (\beta-\lambda)\alpha_i y_i S_i + \sum_i \alpha_i\\ s.t. \alpha y_i^T - \beta + \lambda = 0,\\ \alpha \ge 0,\ \beta \ge 0,\ \lambda \ge 0.\end{aligned} \tag{9}$$

2). Fixed ω , b and ξ , optimize S

For example, a sample $x_i \in X$, we can use the sparse representation S_i to represent the sample x_i. In addition, we set $S_{2,1} = 2tr(S^T\Lambda S)$ from the definition of $l_{2,1}$-norm, where the diagonal matrix Λ meets the condition $\Lambda_{ii} = 1/2S^i{}_2$. The optimization problem of S can be expressed as follows:

$$\begin{aligned}\min_S \frac{1}{2}X - DS_F^2 + 2\tau tr\left(S^T\Lambda S\right)\\ s.t. \quad y_i\left(\omega^T S_i + b\right) \ge 1 - \xi_i, \xi_i \ge 0, \forall i,\\ -\varepsilon - \varphi_m^* \le \omega^T S_m + b_m \le \varepsilon + \varphi_m, u = n+1, \cdot s, m.\end{aligned} \tag{10}$$

Considering the hinge losses, we rewrite the Eq. (10) into the following formulation:

$$\begin{aligned}\min_S \frac{1}{2}X - DS_F^2 + 2\tau tr\left(S^T\Lambda S\right)\\ + \frac{1}{|S|}\sum_i max\left(0, 1 - y_i\left(\omega^T S_i + b\right) - \xi_i\right)\\ \frac{1}{|S|}\sum_u max\left(0, \left|\omega^T S_m + b_m\right| - (\varepsilon + \varphi_m)\right)\end{aligned} \tag{11}$$

The stochastic gradient descent (SGD) algorithm is utilized to optimize S_i, the gradient of S_i in has the expression as follows ∇:

$$\mathrm{S} = \mathrm{D}^{\mathrm{T}}(\mathrm{X} - \mathrm{DS}) + 2\tau\Lambda\mathrm{S} - \frac{1}{|S|}(\Gamma_i + \Gamma_u) \tag{12}$$

Where we let

$$\begin{aligned}\Gamma_i = \Gamma\left(y_i\omega^T S_i + b < 1\right) y_i\omega\\ \Gamma_u = \Gamma\left(y_i\omega^T S_m + b < 1\right) y_i\omega\end{aligned} \tag{13}$$

Γ is an indication function, which is utilized to represent Γ_i and Γ_u. Let ∇_s be 0, S is expressed as follows:

$$S = \left(D^T D + 2\tau\Lambda S + \sigma I\right)^{-1} \left(D^T X - \frac{1}{|S|}\left(\Gamma_i + \Gamma_u\right)\right) \tag{14}$$

where σ is small constants, and $\sigma = 1e^{-4}$. In the feature space, the dimension is larger than the number of samples, in order to avoid the problem of inverse singular, we add σI to avoid this singularity issue.

3). Fixed , b and ξ, optimize D

Notable, θ is the trade-off parameters which are used to denote the weights of reconstruction loss, and we set $\theta = 1$. For the optimization problem of dictionary, we adopt the singular value decomposition (SVD) method to optimize the dictionary D. Next, the optimization of the dictionary D can be expressed as follows:

$$\min_D \frac{1}{2} X - DS_F^2 \qquad s.t.\ \ s.t.\ d_{i2}^2 \le 1,\ \forall i. \tag{15}$$

Next, we use the SVD with $S = UVW^T$ where $U \in R^{K_1 \times K_1}$ is an orthogonal matrix and $W \in R^{K_1 \times K_2}$ is a thin (since $K1 \gg K2$) orthonormal matrix. Let $U^T U = UU^T = I$ and $W^T W = I$ derive the following updated dictionary.

Using SVD method, a new error cost function can be expressed as follows:

$$X - DUVW_F^2 = XW - DUV_F^2 \tag{16}$$

Let$X^{new} = XW$and$B = DU$and then obtain the optimization formula which is equivalent to Eq. (16) as Eq. (17).

$$\min_B X^{new} - BV_F^2 \tag{17}$$

where we can know that x_i^{new} is the ith column of X_{new}, and b_i is the i-th column of B.

According to the SVD, V is a diagonal matrix. For each b_i, we have the expression as follows:

$$\min_{b_i} x_i^{new} - b_i v_i{}_F^2 \tag{18}$$

where v_i is the i-th diagonal element in V . A unique solution is considered to solve this minimization problem, which shows as $b_i = x_i^{new}/v_i$. From $b_i = x_i^{new}/v_i$, we can obtain the estimation B and the i-th column in b_i. Thus, the estimation of D is obtained by

$$D = BU^T \tag{19}$$

Once ω is obtained from Eq. (7), we can obtain the separating hyperplane as follows:

$$\omega S + b = 0 \tag{20}$$

Therefore, we can classify the new data. A new data sample is assigned to class "+" or "-", depending on the Eq. (21).

$$\omega S + b = arg \min_{i=1,2} |\omega S_i + b| \tag{21}$$

The proposed algorithm is summarized in Algorithm 1.

Algorithm 1 Optimization of the proposed method

Input: Training dataset, parameters α, β, λ , γ, ρ, μ and τ.
Output: D, S, ω , b
1: **Initialize** D, S , ω and b, and initialize parameters α, β, λ , γ, ρ , μ and τ.
2: **While** not converge **do**
3: Update ω and b by Eq. (9).
4: Update S by Eq. (14).
5: Update D by Eq. (19).
6: **End while**
7: Given a new test data x.
8: **If** Eq. (21) ≥0,
9: Assign the new sample as positive class.
10: **Else**
11: Assign it as negative class.
12: **End If**
13: **End**

5 EXPERIMENTAL RESULTS

Since the proposed U-DL method is a kind of dictionary-based with Universum data learning model, in order to verify the effectiveness of the proposed method, we have conducted the experiments on the real-world datasets to compare the performance of the proposed method with other baselines (i.e. U-SVM [21], U-AdaBoost[4], RUTSVM-CIL[9], SDTSL[10]).

1) U-SVM [21]: It utilizes a framework which embedded the Universum data into the SVM classification.
2) U-AdaBoost [4]: An Universum data method based on AdaBoost which assigns different weight to the samples, in which the Universum data is utilized to improve the performance.
3) RUTSVM-CIL [9]: It is an Universum data method based on twin SVM, and use a small sized rectangular kernel matrix to reduce the computation time.
4) SDTSL [10]: It is a dictionary learning method, in which the samples are represented by a shared dictionary matrix in dimensional subspace.

Datasets. We select four datasets (i.e. 20 Newsgroup[1], Reuters-21578[2], MNIST[3] and USPS[4]) to conduct the experiments. 20 Newsgroup dataset is widely used for text classification, text mining and information retrieval research. According to different topics, this dataset can be grouped into seven categories, which include more than 20,000 documents; Reuters-21578 dataset has 135 large categories according to their respective themes or contents, which has a total of 21,578 documents; MNIST dataset has more than 60,000 examples and over 10,000 examples in training set and test set, respectively; USPS dataset has more than 7200 images and about 2000 images in training set and test set, respectively. MNIST dataset and USPS dataset have ten classes' digits from "0" to "9". We use the four datasets to construct 10 groups of datasets and implement the experiments, and the basic information of datasets is listed as Table 1.

Evaluation Metrics. We select two widely used evaluation metrics (i.e. classification accuracy (AC) and F_1-measure (F_1)) to evaluate the classification performance of the proposed method and baselines. For all evaluation metrics, the larger value means the better performance.

[1] http://qwone.com/~jason/20Newsgroups/
[2] http://www.kdd.ics.uci.edu/databases/reuters21578/reuters21578.html
[3] http://yann.lecun.com/exdb/mnist/
[4] http://www.kaggle.com/bistaumanga/usps-dataset

Table 1: Description of Dataset

Dataset	Positive class	Negative class	Universum data
1	comp	sci	talk,rec
2	rec	comp	sci,talk
3	sci	talk	Rec,comp
4	Orgs	People	Places
5	Places	People	Orgs
6	Places	Orgs	People
7	"2" in MNIST	"3" in MNIST	"8" in MNIST
8	"5" in MNIST	"8" in MNIST	"6" in MNIST
9	"2" in USPS	"5" in USPS	"5" in USPS
10	"3" in USPS	"5" in USPS	"8" in USPS

Experiment Setup. In order to make a fair comparison of the proposed method and baselines, we introduce the conFigureurations of the proposed U-DL method and baselines in this part, and then we conduct five-folder cross validation to obtain the performance. The detailed conFigureurations of the proposed method and baselines as follows.

(1). For U-SVM method, ε-insensitive loss is chosen from the set $\{0, 0.1, \ldots, 1\}$, C_t and C_u are the penalty parameters, which are chosen from the set $\{10^{-5,-4,\ldots,5}\}$. (2). For U-AdaBoost method, the number of weak classifiers is selected in set $\{1, 2, \ldots, 1000\}$, while the maximum number of iterations T_{max} limited to 1000. Besides, c is the regularization parameter, which is selected in set $\{2^{-15,-13,\ldots,-5}\}$. (3). For RUTSVM-CIL method, the penalty parameters $C = C_1 = C_2 = C_u$ are searched in the set $\{10^{-5,-4,\ldots,5}\}$, and the value μ is selected in the set $\{2, \ldots, 10\}$. (4). For SDTSL method, λ is used to control the sparsity, $\lambda = 10^{-3}$. μ is the penalty parameter, in which $\mu = 0.1$ and $\mu_{max} = 10^6$. Besides, the error tolerance ε= 10^{-4}, and $\rho = 3.6$. In addition, the maximum number of iterations $T = 10^3$. (5). For the proposed U-DL method, C_1 and C_2 are the penalty parameters C_1 is selected in set $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, while and C_2 is chosen from the set $\{10^{-5,-4,\ldots,4}\}$. In addition, and ε-insensitive loss is selected in set $\{1/10, 1/9, \ldots, 1/2, 3/5, 3/4\}$. Besides, the scalar constants τ is selected in set $\{1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-1}, 5 \times 10^{-1}, 1, 5\}$.

Results. We can also obtain the optimal performance (i.e. AC and F_1 measure), and the corresponding results have shown in Table 2 and Table 3. For example, for the dataset 2, the performance of U-SVM, U-Adaboost, RUTSVM-CIL and SDTSL methods is 72.84, 73.41, 68.91 and 62.67, respectively, and the proposed method U-DL method is 78.21, which is higher than compared methods. This result occurs because the proposed U-DL method introduces the Universum data into the classification model by ε-insensitive loss function, in which the prior knowledge related to the classification problem are embedded in the Universum data, which can improve the performance of the proposed U-DL method. Besides, the dictionary is embedded into the learning model to construct a unified framework to improve the classification performance, in which dictionary learning can solve the problem of noise in collected data and enhance the sparse ability of original data. In addition, we adopt the F_1-measure evaluation metric to evaluate the performance of the proposed U-DL method and baselines. We present the

Table 2: AC Comparison of Different Methods on Four Datasets (The Bold Means the Best Results)

Dataset	U-SVM	U AdaBoost	RUTSVM-CIL	SDTSL	Ours
1	71.77	78.27	73.52	65.35	**80.87**
2	72.84	73.41	68.91	62.67	**78.21**
3	70.21	71.92	68.29	55.67	**75.26**
4	72.57	76.66	73.93	69.91	**79.28**
5	70.15	73.87	68.87	66.37	**78.37**
6	71.74	74.25	69.29	65.57	**79.63**
7	71.11	77.93	75.12	68.83	**80.85**
8	70.21	78.43	74.59	68.43	**80.21**
9	64.83	74.28	70.16	62.22	**77.58**
10	62.57	69.32	65.17	58.79	**72.69**

Table 3: F_1 Comparison of Different Methods on Four Datasets (The Bold Means the Best Results)

	U-SVM	U-AdaBoost	RUTSVM-CIL	SDTSL	Ours
K-means	64.89	65.63	63.11	52.89	**70.25**
EM Clustering	61.20	60.55	60.06	51.63	**65.66**
DBSCAN	60.45	62.01	58.09	52.26	**67.63**
GrabCut	59.08	60.50	61.81	55.86	**65.27**
MILCut	60.78	62.34	59.32	53.21	**63.29**

results for the algorithms with different feature extraction methods in Tabel 3. From the Table 3, it is observed that the proposed U-DL method obtains better performance than baselines. We select five features extraction methods to conduct the above experiments on four datasets, which make a performance comparison of the proposed U-DL method with U-SVM, U-AdaBoost, RUTSVM-CIL and SDTSL methods. Finally, the results have shown that the proposed U-DL method performs better than baselines. In addition, the results indicate that the proposed U-DL method has a better classification performance.

Finally, we calculate the average computational time of the proposed U-DL method and baselines on the ten datasets and present them on the Figure 1. From the Figure 1, it is observed that the proposed U-DL method spends more time than U-SVM, U-AdaBoost and RUTSVM-CIL methods, but less than SDTSL method. This occurs because that the proposed U-DL incorporates the dictionary and Universum data to construct a unified model, in which DL method can be utilized to enhance the sparsity of original data and improve the discriminative ability. Therefore, it leads to complex calculations and computation time cost.

6 CONCLUSION

In this paper, a dictionary-based method for classification with Universum data is proposed firstly. In the proposed method, we first introduce the Universum data into the classification model by the hinge loss function, in which the prior knowledge provided

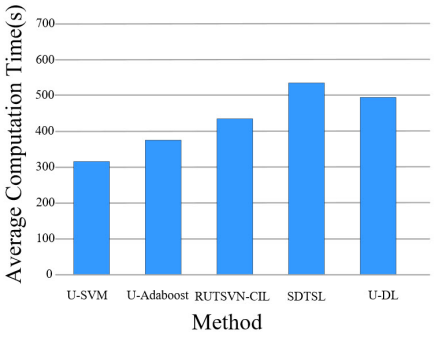

Figure 1: Average Computation Time.

by the Universum data can improve the performance. Second, the dictionary is embedded into the learning model to construct a unified framework. The dictionary is used to solve the problem of data noise, data redundancy and uncertainty. For the optimization, the sparse representations of samples are obtained by the learned dictionary, then are fed into SVM classifier. We introduce the Lagrangian function to optimize the SVM classifiers iteratively. Besides, the gradient-based optimization method and SVD algorithm are used to alternately optimize and update the dictionary. Finally, we conduct the experiments to evaluate the proposed U-DL method and baselines, and the results have shown that the proposed U-DL method performs better than the baselines.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their very useful comments and suggestions. This work was supported in part by the Natural Science Foundation of China under Grant 62076074, 61876044 and Grant 61672169, in part by Guangdong Basic and Applied Basic Research Foundation Grant 2020A1515010670 and 2020A1515011501, in part by the Science and Technology Planning Project of Guangzhou under Grant 202002030141.

REFERENCES

[1] Z. Qi, Y. Tian and Y. Shi (2012). Twin support vector machine with universum data, Neural Networks, 112–119.
[2] Y. Xu, M. Chen and G. Li (2016). Least squares twin support vector machine with universum data for classification, International Journal of Systems Science 47, 3637–3645.
[3] V. Vapnik (2006). Estimation of dependence based on empirical data, in: Springer, Berlin Heidelberg New York pp. 1–479.
[4] J. Xu, Q. Wu, J. Zhang and Z. Tang (2014). Exploiting universum data in adaboost using gradient descent, Image & Vision Computing, 32 550–557.
[5] F. H. Sinz, O. Chapelle, A. Agarwal and B. Schölkopf (2007). An analysis of inference with the universum, Neural Information Processing Systems 1 1369–1376.
[6] D. Zhang, J. Wang, F. Wang and C. Zhang (2008). Semi-supervised classification with universum, in: Proceedings of the 2008 SIAM International Conference on Data Mining pp. 323–333.
[7] C. Shen, P. Wang, F. Shen and H. Wang (2012). Uboost: Boosting with the universum, International Journal of Systems Science 34 825–832.
[8] Y. Tian, Y. Zhang and D. Liu (2016). Semi-supervised support vector classification with self-constructed universum, Neurocomputing 189 33–42.
[9] K. Huang, Y. Wu, C. Wang, *et al.* (2020). A Projective and Discriminative Dictionary Learning for High-Dimensional Process Monitoring With Industrial Applications[J]. IEEE Transactions on Industrial Informatics, PP(99),1-1.
[10] K. Huang, Y. Wu, H. Wen, *et al.* (2020). Distributed dictionary learning for high-dimensional process monitoring[J]. Control Engineering Practice, 98, 104386.
[11] H. Li, J. Xu, J. Zhu, D. Tao and Z. Yu (2019). Top distance regularized projection and dictionary learning for person re-identification, Information Sciences 502 472–491.
[12] Z. Chen, X. Wu, H. Yin and J. Kittler (2020). Noise-robust dictionary learning with slack block-diagonal structure for face recognition, Pattern Recognition 100 107118.
[13] J. Hu and Y. Tan (2018). Nonlinear dictionary learning with application to image classification, Pattern Recognition 75, 282–291.
[14] X. Mai and Z. Wang (2015). A novel double-layer sparse representation approach for unsupervised dictionary learning, Computer Vision & Image Understanding 143 (C) (2015) 1-10.
[15] M. Yang, W. Liu, W. Luo and L. Shen (2016). Analysis-synthesis dictionary learning for universality-particularity representation-based classification, AAAI Conference on Artificial Intelligence 2251-2257.
[16] Z. Jiang, Z. Lin and L. S. Davis (2013). Label consistent k-svd: Learning a discriminative dictionary for recognition, IEEE Transactions on Pattern Analysis & Machine Intelligence 35 (11) (2013) 2651-2664.
[17] J. Guo, Y. Guo, X. Kong, M. Zhang and R. He (2016). Discriminative analysis dictionary learning, in: Association for the Advancement of Artificial Intelligence (AAAI) pp. 1617-1623.
[18] S. Gu, L. Zhang, W. Zuo and X. Feng (2014). Projective dictionary pair learning for pattern classification, Neural Information Processing Systems 793-801.
[19] W. Long, Y. Tang and Y. Tian (2016). Investor sentiment identification based on the universum svm, Neural Computing & Applications 30 661–670.
[20] B. Richhariya and M. Tanveer (2020). A reduced universum twin support vector machine for class imbalance learning, Pattern Recognition 102 107150.
[21] A. Zhang and X. Gao (2019). Supervised dictionary-based transfer subspace learning and applications for fault diagnosis of sucker rod pumping systems, Neurocomputing 338 293–306.