Computer Science and Application 计算机科学与应用, 2021, 11(3), 672-681
Published Online March 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.113069

Hans汉斯

基于Universum数据的多视角学习算法

曾　博[1]，肖燕珊[1]，刘　波[2]

[1]广东工业大学计算机学院，广东　广州
[2]广东工业大学自动化学院，广东　广州
Email: 2756115625@qq.com

收稿日期：2021年2月25日；录用日期：2021年3月19日；发布日期：2021年3月29日

摘　要

多视角学习是以不同方法获得的特征集表示的数据中学习的问题，其中双视角学习是一种仅由双视角数据组成的多视角学习。由于多视角学习可能会忽略一些多视角数据的原始信息，这些数据之间存在着内在的联系和不同视角之间的差异。因此，为了解决多视角数据之间存在的问题，我们引入了既不属于正类又不属于负类的无标签数据Universum数据。本文提出了一种基于Universum数据的多视角学习算法，将Universum数据和多视角学习结合到一个目标模型中，其中Universum数据被认为是该模型的先验知识。为了解决提出的算法模型，我们推导了该算法模型的对偶问题并得到了预测分类器。最后，通过大量的实验对该方法的性能进行了研究，结果表明，所提算法的性能优于传统的方法。

关键词

多视角学习，Universum数据，支持向量机

Multi-View Learning Algorithm Based on Universum Data

Bo Zeng[1], Yanshan Xiao[1], Bo Liu[2]

[1]Department of Computer, Guangdong University of Technology, Guangzhou Guangdong
[2]Department of Automation, Guangdong University of Technology, Guangzhou Guangdong
Email: 2756115625@qq.com

Received: Feb. 25th, 2021; accepted: Mar. 19th, 2021; published: Mar. 29th, 2021

Abstract

Multi-View Learning (MVL) focuses on the problem of learning from the data represented by fea-

文章引用: 曾博, 肖燕珊, 刘波. 基于 Universum 数据的多视角学习算法[J]. 计算机科学与应用, 2021, 11(3): 672-681.
DOI: 10.12677/csa.2021.113069

ture sets obtained from different approaches, in which two-view learning is a kind of MVL which only consists of two-view data. Since multi-view learning may ignore the original information of some multi-view data, there are inherent connections between these data and the differences between different perspectives. Therefore, in order to solve the problems between multi-view data, we introduce the unlabeled data Universum data that neither belongs to the positive category nor the negative category. This paper proposes a multi-perspective learning algorithm based on Universum data, which combines Universum data and multi-perspective learning into a target model, in which Universum data is considered as the prior knowledge of the model. In order to solve the proposed algorithm model, we derive the dual problem of the algorithm model and get the predictive classifier. Finally, the performance of the method is studied through a large number of experiments, and the results show that the performance of the proposed algorithm is better than that of the traditional method.

Keywords

Multi-View Learning, Universum Data, Support Vector Machine

1. 引言

在传统机器学习的分类问题中，总是利用样本的单一视角特征来建立机器学习模型，进而对未知的样本进行预测，该模型也可以称为单视角学习[1] (Single-View Learning, SVL)。但在现实中和大数据环境下，数据的收集和存储能力得到了极大的提高，在多样化的信息获取技术的推动下，我们总是可以更容易和更便捷的从不同的途径或者不同的方法获得样本，其被称为多视角数据，呈现出多态性和多描述性等特点[2]。如图 1 所示，左图含有图像特征和文本描述，文本特征是对图像的一种描述和解释，右图表示不同语种表示你好的多种表现形式。然而，我们可以在多视角数据上构建机器学习模型，这被称为多视角学习[3] (Multi-View Learning, MVL)。

Figure 1. Schematic diagram of multi-view data
图 1. 多视角数据示意图

多视角学习最初由 De sa [4]等人提出，表明模式之间的不同感觉方式可以被看作是多种视角的。这里的“视角”表示来自多个源或不同特征子集的数据。多视角学习的应用被扩展到半监督学习[5]、监督学习[6]、降维[7]和聚类[8]。Farquhar 等人[6]提出将内核典型相关分析和支持向量机链接到一个名为

SVM-2K (KCCA followed by Support Vector Machine)的单一优化。Chen 等人[7]提出了一种新的降维方法，称为半配对半监督广义相关分析，称为 semi-paired and semi-supervised generalized correlation analysis (S2GCA)，既能保持未标记数据的全局结构，又能找出标记数据的最大可分性。Bickel 和 Scheffer [8]发现多视角聚类算法可以改进单视角聚类。尽管在多视角学习中取得了巨大的进步，但多视角学习可能会忽略一些具有内部连接以及不同视图之间存在差异的多视角数据的原始信息。在实践中，我们可能会遇到在训练阶段没有标签的数据，并且它们不同时属于正负类，这被称为 Universum 数据。Universum 数据可以视为其他数据，以提高学习工具的泛化性能。

近年来，国内外研究者对 Universum 学习也逐渐广泛关注。Vapnik 等人[9]首先提出了一个新的研究框架，这是一个替代大范围方法的能力概念，以便在学习过程中整合领域信息。Weston 等人[10]在 Vapnik 研究的基础上，首先提出了 Universum 的概念，特别是在二元分类问题中，将一组有标记的样本和一组不属于任何兴趣类别的无标记样本作为输入信息，未标记的样本称为 Universum 样本。Liu 等人[11]提出了一种半监督学习算法，利用增强型技术和 Universum 实例来提高文本分类性能，该算法获取分类函数的先验信息。Richhariya 和 Gupta [12]提出了一种基于牛顿法的迭代 Universum twin support vector machine (IUTWSVM)，该方法利用 Universum 数据从人脸图像中对人脸情绪进行不同层次的分类。Wang 等人[13]开发了半监督学习方法，将 Universum 数据与社交媒体信息相结合称为 Semi-Supervised Feature Selection with Universum (U-SSLFS)，将 Universum 样本相结合，使模型更具辨识性。

综上所述，为了解决多视角学习可能会忽略一些不同视角之间存在差异的多视角数据的原始信息，我们引入了 Universum 数据，提出了一种基于 Universum 数据的多视角学习方法。使得算法更加正则化，提高了多视角分类器的性能，此外，为每个视角引入正则化项意味着每个视角中都有先验知识。我们工作的主要贡献总结如下：

1) 我们首次提出了一种新颖的算法模型，基于 Universum 数据的多视角学习算法，以提高多视角学习的分类性能。对于 Universum 数据，我们首先在每个视角上有相应的特征，考虑原始数据和 Universum 数据的特征构建超平面，并让 Universum 数据分布在正类和负类之间，从而得到一个准确的预测分类器。

2) 构造拉格朗日函数讲算法模型转化为对偶问题，求解相应的优化问题。

3) 我们进行了大量的实验来评估我们提出的算法的性能。统计结果表明，提出的算法比现有的方法更能提高分类精度。

2. 基于 Universum 数据的多视角学习算法

2.1. 目标函数

假设给定相同数据的两个视角，一个视角可以通过相应的内核函数 k_A 的特征投影 ϕ_A 表示，另一个通过具有相应的内核函数 k_B 的特征投影 ϕ_B 表示。对于分类任务，每个数据项还应该包含标签[6]。然后，通过一组给定的配对数据集：

$$S=\left\{\left(\phi_A\left(x_1\right),\phi_B\left(x_1\right)\right),\cdots,\left(\phi_A\left(x_l\right),\phi_B\left(x_l\right)\right)\right\} \tag{1}$$

通过引入 Universum 数据，我们将多视角学习和 Universum 数据相结合。假设除了多视角训练数据外，我们还获得了另一组数据，称为 Universum 数据，该数据是一组不属于任何感兴趣类别的未标记样本，被用作输入信息，一个 Universum 数据可以表示为 $\left\{\phi\left(x_\mu^1\right),\phi\left(x_\mu^2\right),\cdots,\phi\left(x_\mu^\mu\right),\phi\left(x_\mu^{\mu+1}\right),\cdots,\phi\left(x_\mu^{2\mu}\right)\right\}$，$x_\mu^m\left(m=1,2,\cdots,2\mu\right)$，每一个多视角数据都包含一个 Universum 数据。所以训练数据可以被表示为：

$$T=\left\{\left(\phi_A\left(x_1\right),\phi_B\left(x_1\right)\right),\cdots,\left(\phi_A\left(x_l\right),\phi_B\left(x_l\right)\right)\right\}\cup\left\{\phi\left(x_\mu^1\right),\phi\left(x_\mu^2\right),\cdots,\phi\left(x_\mu^\mu\right),\phi\left(x_\mu^{\mu+1}\right),\cdots,\phi\left(x_\mu^{2\mu}\right)\right\} \tag{2}$$

对于使用 Universum 数据进行多视角学习的问题，我们首次提出了以下学习模型。我们将有标签的多视角数据与另一组没有标签的 Universum 样本结合起来，该算法模型可以表示为：

$$\min_{w,b,\xi_i,\psi_m} \quad \frac{1}{2}\left(\|w_A\|^2+\|w_B\|^2\right)+C^A\sum_{i=1}^{l}\xi_i^A+C^B\sum_{i=1}^{l}\xi_i^B+C\sum_{i=1}^{l}\eta_i+D\sum_m\left(\psi_m+\psi_m^*\right) \tag{3}$$

约束条件：

$$\begin{aligned}
&\left|\left(w_A\cdot\phi_A\left(x_i\right)+b_A\right)-\left(w_B\cdot\phi_B\left(x_i\right)+b_B\right)\right|\le\varepsilon+\eta_i\\
&y_i\left(w_A\cdot\phi_A\left(x_i\right)+b_A\right)\ge1-\xi_i^A\\
&y_i\left(w_B\cdot\phi_B\left(x_i\right)+b_B\right)\ge1-\xi_i^B\\
&-\delta-\psi_m^*\le\left(w_Ax_\mu^m+b_A\right)\le\delta+\psi_m\\
&-\delta-\psi_m^*\le\left(w_Bx_\mu^m+b_B\right)\le\delta+\psi_m\\
&\xi_i^A\ge0,\xi_i^B\ge0,\eta_i\ge0,\quad i=1,\cdots,l\\
&\psi_m,\psi_m^*\ge0,\quad m=1,2,\cdots,\mu,\mu+1,\cdots,2\mu
\end{aligned}$$

对于上述提出的算法，我们给出如下详细解释。$\|w_A\|$ 和 $\|w_B\|$ 分别是视角 A 和视角 B 的正则化项，用于防治过拟合。参数 C^A,C^B,C,D 是惩罚参数。参数 η_i 是非负松弛变量，用于控制两个分类器之间的间隙，希望两个视角的预测相似。ξ_i^A 和 ξ_i^B 是视角 A 和视角 B 的非负松弛变量。ψ_m 和 ψ_m^* 是 Universum 样本的非负松弛变量。参数 δ 是用户定义的参数，代表 Universum 样本的不敏感损失。约束 $-\delta-\psi_m^*\le\left(w_Ax_\mu^m+b_A\right)\le\delta+\psi_m$ 和 $-\delta-\psi_m^*\le\left(w_Bx_\mu^m+b_B\right)\le\delta+\psi_m$ 表示的是 Universum 数据定义了不敏感损耗区域，Universum 数据位于支持超平面之间的不敏感区域。

2.2. 对偶问题

为了解决公式(3)的优化问题，首先构造拉格朗日函数，对于公式(3)中的每个不等式约束，通过引入拉格朗日乘子 $\alpha_i^A,\alpha_i^B,\mu_i^A,\mu_i^B,\beta_i^A,\beta_i^B,\beta_m^{(*)},\gamma_m^{(*)},k_m^{(*)},\lambda$。拉格朗日函数被定义为：

$$\begin{aligned}
L(\Theta)=&\frac{1}{2}\left(\|w_A\|^2+\|w_B\|^2\right)+C^A\sum_{i=1}^{l}\xi_i^A+C^B\sum_{i=1}^{l}\xi_i^B+C\sum_{i=1}^{l}\eta_i+D\sum_m\left(\psi_m+\psi_m^*\right)\\
&-\sum_i\alpha_i^A\left(y_i\left(w_A\cdot\phi_A\left(x_i\right)+b_A\right)-1+\xi_i^A\right)\\
&-\sum_i\alpha_i^B\left(y_i\left(w_B\cdot\phi_B\left(x_i\right)+b_B\right)-1+\xi_i^B\right)\\
&+\sum_i\beta_i^A\left(\left(w_A\cdot\phi_A\left(x_i\right)+b_A\right)-\left(w_B\cdot\phi_B\left(x_i\right)+b_B\right)-\varepsilon-\eta_i\right)\\
&+\sum_i\beta_i^B\left(\left(w_A\cdot\phi_A\left(x_i\right)+b_A\right)-\left(w_B\cdot\phi_B\left(x_i\right)+b_B\right)+\varepsilon+\eta_i\right)\\
&+\sum_m\beta_m\left(w_Ax_\mu^m+b_A-\delta-\psi_m\right)-\sum_m\beta_m^*\left(w_Ax_\mu^m+b_A+\delta+\psi_m^*\right)\\
&+\sum_m\gamma_m\left(w_Bx_\mu^m+b_B-\delta-\psi_m\right)-\sum_m\gamma_m^*\left(w_Bx_\mu^m+b_B+\delta+\psi_m^*\right)\\
&-\sum_i\mu_i^A\xi_i^A-\sum_i\mu_i^B\xi_i^B-\sum_mk_m\psi_m-\sum_mk_m^*\psi_m^*-\sum_i\lambda\eta_i
\end{aligned} \tag{4}$$

根据朗格朗日的对偶性，原始问题的对偶性是极大极小问题，因此，为了解决对偶问题，我们首先对拉格朗日函数 $L(\Theta)$ 对 w,b,ξ_i,ψ_m,η_i 进行求偏导并设置等式为 0，拉格朗日函数的微分如下：

$$
\begin{aligned}
&w_A = \sum_i \alpha_i^A y_i \phi_A(x_i) - \sum_i \beta_i^A \phi_A(x_i) + \sum_i \beta_i^B \phi_A(x_i) - \sum_m \beta_m x_\mu^m + \sum_m \beta_m^* x_\mu^m,\\
&w_B = \sum_i \alpha_i^B y_i \phi_B(x_i) - \sum_i \beta_i^A \phi_B(x_i) + \sum_i \beta_i^B \phi_B(x_i) - \sum_m \gamma_m x_\mu^m + \sum_m \gamma_m^* x_\mu^m,\\
&C^A = \alpha_i^A + \mu_i^A, \quad C^B = \alpha_i^B + \mu_i^B,\\
&D = \beta_m + \gamma_m + k_m, \quad D = \beta_m^* + \gamma_m^* + k_m^*, \quad C = \beta_i^A + \beta_i^B + \lambda
\end{aligned} \tag{5}
$$

将公式(5)代入公式(4)中，得到对偶问题：

$$
\begin{aligned}
\max W = &-\frac{1}{2}\sum_i\sum_j\left(g_i^A g_j^A\left(\phi_A(x_i)\phi_A(x_j)\right)\right) - \frac{1}{2}\sum_i\sum_j\left(g_i^B g_j^B\left(\phi_B(x_i)\phi_B(x_j)\right)\right)\\
&+\frac{1}{2}\sum_m\sum_n\left(\beta_m\beta_n + \gamma_m\gamma_n\right)x_\mu^m x_\mu^n + \frac{3}{2}\sum_m\sum_n\left(\beta_m^*\beta_n^* + \gamma_m^*\gamma_n^*\right)x_\mu^m x_\mu^n\\
&+\sum_i\sum_j\left(\alpha_i^A\beta_j^A - \alpha_i^A\beta_j^B\right)y_i y_j\phi_A(x_i)\phi_A(x_j) - \sum_i\sum_j\left(\alpha_i^B\beta_j^A - \alpha_i^B\beta_j^B\right)y_i y_j\phi_B(x_i)\phi_B(x_j)\\
&-2\sum_m\sum_n\left(\beta_m\beta_n^* + \gamma_m\gamma_n^*\right)x_\mu^m x_\mu^n + \sum_i\left(\alpha_i^A + \alpha_i^B\right) - \delta\sum_m\left(\beta_m + \gamma_m\right) - \delta\sum_m\left(\beta_m^* + \gamma_m^*\right)
\end{aligned} \tag{6}
$$

约束条件：

$$
\begin{aligned}
&g_i^A = \alpha_i^A y_i - \beta_i^A + \beta_i^B, \quad g_i^B = \alpha_i^B y_i + \beta_i^A - \beta_i^B, \quad \sum_i g_i^A = \sum_i g_i^B = 0,\\
&0 \le \alpha_i^{A/B} \le C^{A/B}, \quad 0 \le \beta_m + \gamma_m \le D, \quad 0 \le \beta_m^* + \gamma_m^* \le D\\
&i = 1,\cdots,l, \quad m = 1,2,\cdots,\mu,\mu+1,\cdots,2\mu
\end{aligned}
$$

基于上述具有 Universum 数据的多视角学习模型，我们提出了该算法的完整过程，具体的算法实现步骤如表 1 所示。

Table 1. Algorithm implementation steps
表 1. 算法实现步骤

输入：带有标签的训练集和 Universum 样本
输出：w_A, w_B, b_A, b_B

1：通过相应的核函数得到训练集的特征投影；
2：融合多视角数据与 Universum 数据；
3：初始化 $\xi_i^A, \xi_i^B, C, C^A, C^B, D$；
4：构造并求解凸二次规划问题(6)；
5：通过求解 Quadratic programming 问题得到解 $\alpha^A, \beta^B, \beta_m, \beta_m^*, \gamma_m, \gamma_m^*$；
6：通过公式(5)计算 w_A, w_B, b_A, b_B；
7：返回结果 w_A, w_B, b_A, b_B；

2.3. 时间复杂度分析

我们将讨论该算法的时间复杂度并给出一个估计。该算法可以归结为凸二次规划问题(Convex quadratic programming problem)，所以该算法的时间复杂度为 $\mathcal{O}\left((l+m)^3\right)$ (l 为训练样本个数，m 为 Universum 样本个数)。

3. 实验与分析

3.1. 实验数据

我们已经试验了多个数据集，这些数据集广泛应用于多视角学习中。数据集包括 Pascal Visual Object

Classes、NUS-WIDE-OBJECT、Handwritten Digit 和 Image Segmentation，其详细描述如下：

- Pascal Visual Object Classes (VOC2007)：该数据集是图像数据集，其中包含 9963 个真实世界的图像，这些图像分为 20 类，例如人，鸟，自行车，椅子等。在本实验中，该数据集被划分为 5011 张训练图像和 4952 张测试图像。
- NUS-WIDE-OBJECT：该数据集包含 30,000 个对象图像，并被分为 30 类，例如玩具，花朵，山脉，旗帜等。在本实验中，该数据集被随机分为 17,927 个训练图像和 12,073 个测试图像。
- Handwritten Digit：该数据集由手写数字(‘0’-‘9’)组成，手写数字有 2000 个图像，共 10 个类别，每个类别有 200 个图像。每张图片均已用二进制图像进行数字表示。在此实验中，我们从每个数字中随机选择 50%的图像进行训练。其余图像是测试图像。
- Image Segmentation：是从 7 个户外图像的数据库中随机抽取的图像数据集，该数据库由 2310 个随机选择的对象组成，这些对象分为 7 个类，即砖墙，天空，树叶，水泥，窗户，路径和草。数据集包含 19 个连续属性，可以自然分为多个视角数据。

对于 Universum 数据而言，有几种方法可以收集 Universum 样本[10] [14]。在本文中，我们使用的是 U_{Rest} 方法来生成 Universum 样本，并通过 Universum 样本的先验知识来提高该算法的性能。U_{Rest} 方法表示分类任务中不包括的其他数据。例如，如果分类任务是对数字 0 和 1 进行分类，并且有其他数字(从 2 到 9)的图像，这些图像可以用作 Universum 示例。总体而言，实验中使用的实验组和如表 2 所示。

Table 2. Experimental data combination
表 2. 实验数据组合

Subdataset	Data Set	Positive	Negative	Universum
dataset 1	Pascal VOC2007	bird	cat, cow, dog	horse, sheep
dataset 2	Pascal VOC2007	boat	bus, car, train, aeroplane	motorbile, bicycle
dataset 3	NUS-WIDE-OBJECT	bear	birds, cat, tiger	cow, sun
dataset 4	NUS-WIDE-OBJECT	boats	cars, plane, train	vehicle, sand
dataset 5	NUS-WIDE-OBJECT	flags, sign	whales, cars, plane, train	tower, toy
dataset 6	Handwritten Digit	Number 3	Number 5, Number 0	Number 9
dataset 7	Handwritten Digit	Number 2	Number 6	Number 7
dataset 8	Image Segmentation	brickface	cement, path	window
dataset 9	Image Segmentation	sky	grass	foliage

3.2. 实验设置

为了验证所提算法的有效性，在实验设置阶段我们采用与其他四种多视角学习算法进行对比，对比算法如下：

- SVM-2K [6]：该方法结合了标准 SVM 和 KCCA 算法，通过利用两视图数据之间的关系来提高分类器性能。
- USVM [10]：它将具有先验知识的 Universum 数据与标准 SVM 算法结合使用，以在模式识别问题上获得更好的性能。
- MvTSVMs [15]：它引入了两个一维投影之间的相似性约束，并结合了多视角学习和双支持向量机方法。

- MvNPSVM [16]：该方法结合了非平行支持向量机(Nonparallel support vector machine, NPSVM)算法和多视角学习的优点，并将 NPSVM 扩展到多视角学习领域，带来了新的见解。

在实验中，我们对所有实验在两个视角上使用高斯 RBF 核，并且将 RBF 核参数 σ 设置为 $\{0.25,0.5,0.75,1\}$。在提出的方法中，我们设置惩罚参数 $C^A=C^B=C$ 和 D 为 $\{10^{-3},10^{-2},10^{-1},1,10,10^2\}$ 和 $\{0.05,0.5,1,3,5\}$，参数 ε 在 $\{0.01,0.1\}$ 集合中调整。

对于四个对比算法，我们设置与他们的研究相似的参数，并且实验中算法的配置如下。在 USVM 算法中，惩罚参数 C 和 D 分别设置为 $\{10^{-2},10^{-1},1,10,10^2\}$ 和 $\{0.05,0.5,1,3,5\}$。在 SVM-2K 算法中，在 $\{10^{-2},10^{-1},1,10,10^2\}$ 集合中均等设置惩罚参数 C^A 和 C^B，参数 D 和 ε 分别在 $\{0.05,0.5,1,3,5\}$ 和 $\{0.01,0.1\}$ 上调整。对于 MvTSVM 和 MvNPSVM 算法，我们将惩罚参数 $C_1=C_2=C_3=C_4$ 和 $D=H$ 分别从 $\{10^{-3},10^{-2},10^{-1},1,10,10^2\}$ 和 $\{0.05,0.5,1,3,5\}$ 集合中选择。参数 ϵ 和 ε 的取值范围为 $\{0.01,0.1\}$。对于所有方法，为了避免实验中的采样偏差，我们使用五次交叉验证，选择四层作为训练集，另一层被视为每一轮的测试集。另外，五次交叉验证用于确定实验中的适当参数。例如，对于提出的方法，我们在表 2 中填写适当的参数，该参数在五次训练和测试集中处于最高性能。

3.3. 实验结果分析

3.3.1. 性能比较

在本节中，我们将比较提出的方法和四个对比方法的性能。表 3 汇总了四个数据集中不同方法的分类准确性以及标准差。

Table 3. Classification accuracy and standard deviation result statistics
表 3. 分类准确度和标准差结果统计

Subdataset	SVM-2K	USVM	MvTSVMs	MvNPAVM	所提方法
dataset 1	82.60 ± 1.35	83.50 ± 1.26	81.65 ± 1.95	86.50 ± 1.62	**90.45 ± 1.01**
dataset 2	86.00 ± 1.86	82.60 ± 1.06	80.90 ± 1.76	85.80 ± 1.75	**88.50 ± 1.25**
dataset 3	74.60 ± 1.78	77.65 ± 1.68	72.50 ± 2.29	**84.45 ± 1.59**	83.02 ± 1.84
dataset 4	72.50 ± 1.94	78.60 ± 1.64	74.90 ± 1.95	80.60 ± 1.78	**84.50 ± 1.28**
dataset 5	73.60 ± 1.36	76.40 ± 1.28	74.90 ± 1.46	79.90 ± 1.89	**83.90 ± 1.34**
dataset 6	76.65 ± 1.76	80.60 ± 1.95	75.50 ± 2.05	79.45 ± 1.80	**81.00 ± 1.54**
dataset 7	77.50 ± 1.89	78.80 ± 1.26	74.00 ± 1.95	78.50 ± 1.57	**80.65 ± 1.05**
dataset 8	82.50 ± 1.20	83.00 ± 1.34	85.30 ± 1.91	89.02 ± 1.25	**90.50 ± 0.81**
dataset 9	84.52 ± 1.54	83.60 ± 1.25	82.50 ± 2.08	**91.50 ± 1.24**	89.80 ± 1.02

我们可以观察到，提出的方法始终可以比其他方法表现更好。例如，对于数据集 1，SVM-2K，USVM，MvTSVMs 和 MVNPSVM 方法分别获得“82.60”，“83.50”，“81.65”，“86.50”的精度；但是，提出的方法可以达到“90.45”的精度，优于其他方法。发生这种情况的原因是，这是因为提出的方法将 Universum 数据考虑到多视角学习中，从而可以修改多视角学习的决策边界。但是，在构造分类器时，SVM-2K，MvTSVMs 和 MVNPSVM 方法不会考虑 Universum 数据。因此，提出的方法可以比其他方法执行得更好。对于 USVM 方法和提出的方法，它们都考虑了 Universum 数据。但是，提出的方法仍然比 USVM 具有更好的性能。发生这种情况是因为所提出的方法将多视角数据合并到学习中，可以提供更好的特征表示，因此，所提出的方法比 USVM 方法更好。对于标准偏差比较，我们可以进一步观察到，对

于大多数数据集，所提出的方法可以提供比其他方法更少的标准偏差。例如，对于数据集 1，所提出的方法的标准偏差为 1.01，而其他方法则大于 1.01。这表明，提出的方法可以提供相对稳定的性能。

3.3.2. 训练样本数量分析

本文在 Pascal VOC2007，NUS-WIDE-OBJECT，Handwritten Digit 和 Image Segmentation 的不同大小的训练集，实现了 SVM-2K，USVM，MvTSVMs，MVNPSVM 和所提出的方法。我们以上述四个数据集为例。对于 NUS-WIDE-OBJECT 数据集，训练大小从{6000, 8000, 10,000, 12,000, 14,000}集合变化。同样，我们改变 Pascal VOC2007，Handwritten Digit 和 Image Segmentation 训练样本大小，如图 2 的 x 轴所示，Universum 样本的数量是恒定的。另外，图 2 显示了根据上述变化的训练量数据集的 SVM-2K，USVM，MvTSVMs，MVNPSVM 和所提出的方法的分类精度。我们可以发现，在几乎所有情况下，提出的方法显然都优于其他比较算法，并且随着训练样本数量的增加，所有已实现算法的分类精度都会提高。

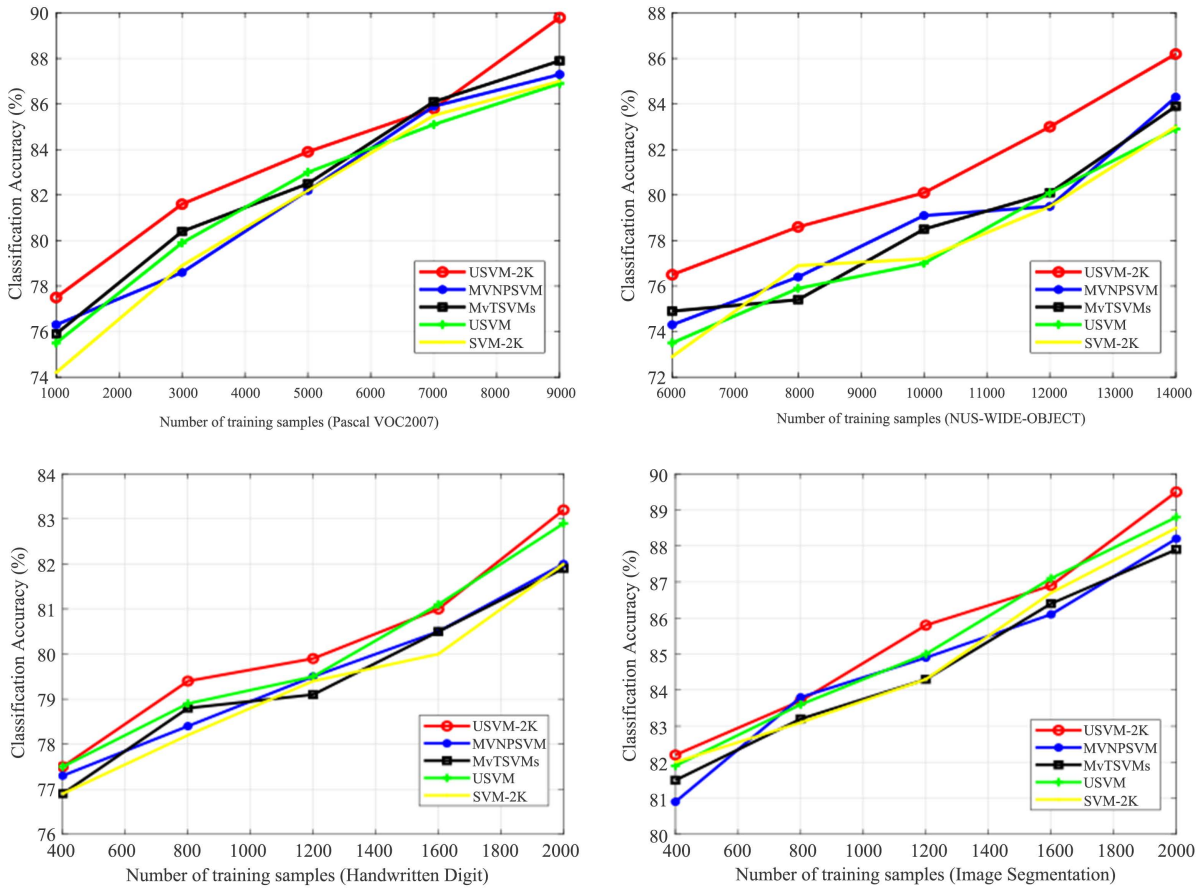

Figure 2. Classification accuracy under different number of training samples
图 2. 不同训练样本数量下的分类准确率

3.3.3. Universum 样本数量分析

研究结果表明，Universum 样本的大小不同也会影响算法的性能[14]。因此，我们在此讨论所提出的算法在 Universum 样本数量变化的情况。实验中，所有参数保持不变，即 $C^A=1, C^B=1, C=1, D=0.5, \varepsilon=0.1$。图 3 显示了所提出的方法和 USVM 方法在不同数量的 Universum 样本下的分类准确率。在图 3 中，水平轴表示 Universum 实例的大小，即 NUS-WIDE-OBJECT 的 Universum 大小在{500, 1000, 1500, 2000, 2500}

的集合中变化。同样，我们改变 Pascal VOC2007、Handwritten Digit 和 Image Segmentation 数据集的 Universum 样本大小，如图 3 的 x 轴所示。纵轴表示对应的分类精度。从图中可以看出，不同数量的 Universum 数据对分类有一定的影响。例如，可以发现在 NUS-WIDE-OBJECT 的数据集中，当 Universum 的数据量较小时，分类准确率较低，然后性能曲线随着样本数量的增加而增加。而当 Universum 数据量达到 1000 时，分类准确率达到最高。当 Universum 数据量不断增加时，分类精度在总体上仍能保持相对稳定。类似地，其他数据集也显示出类似的效果。

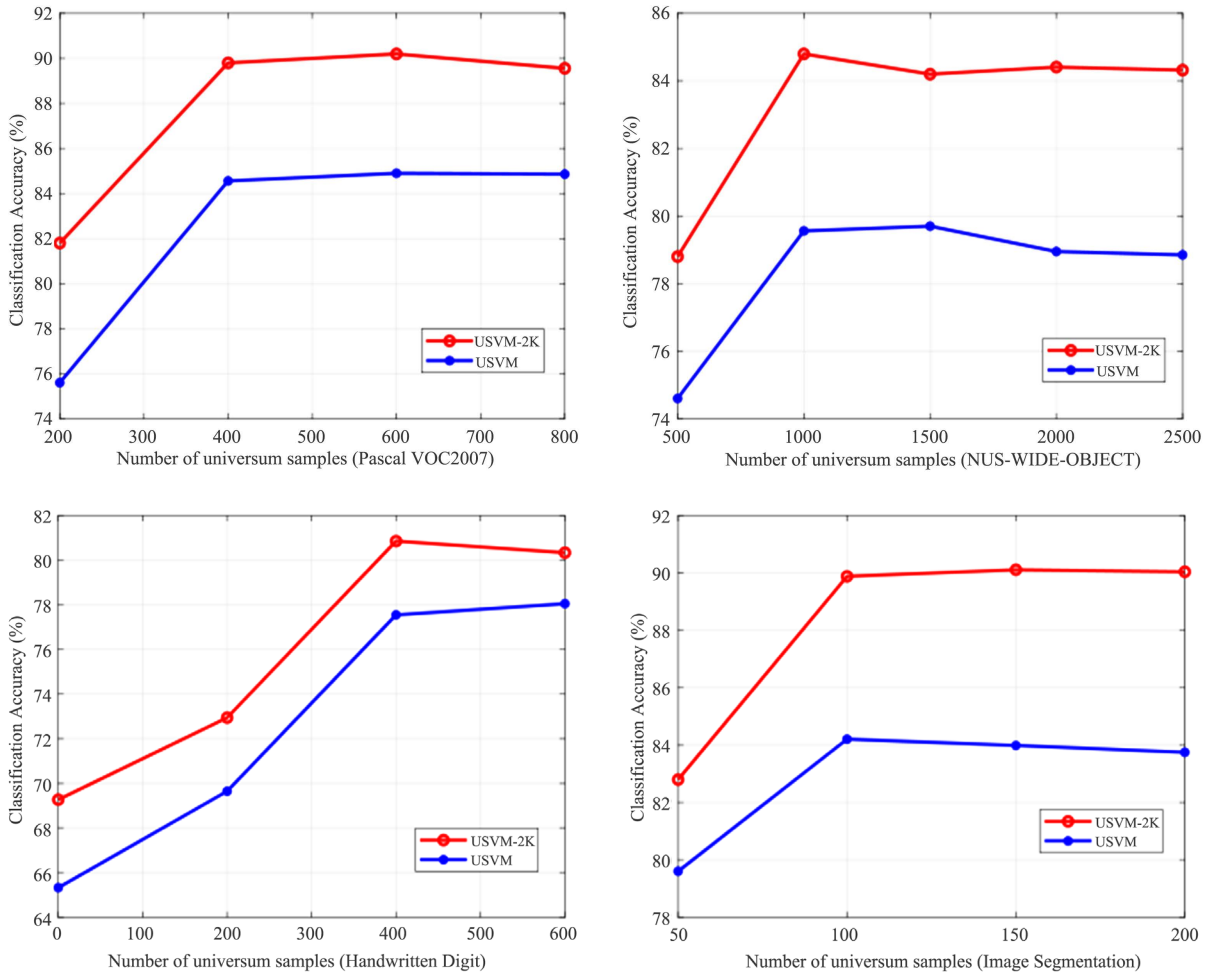

Figure 3. Classification accuracy under different number of universum samples
图 3. 不同 Universum 样本数量下的分类准确率

4. 总结与展望

在本文中，我们提出了一种基于 Universum 数据的多视角学习算法。新提出的方法借助于不属于任何一类分类问题的 Universum 示例，既继承了先前的多视角学习的优势，而且可以获取更多的关于整个数据分布的先验知识。为了有效的求解该算法，我们推导了该算法的对偶形式，为了得到更有效的预测模型。为了验证所提出方法的有效性，我们在真实的数据集上进行了实验。在图像数据集的情况下，我们讨论了所有方法的分类准确率，并分析了不同训练规模下的分类性能。在未来，我们希望研究在数据流环境中的多视角学习和 Universum 数据的结合。

基金项目

本文得到国家自然科学基金资助项目(No.62076074)的资助。

参考文献

[1] Qin, B., Xia, Y., Wang, S. and Du, X.Y. (2011) A Novel Bayesian Classification for Uncertain Data. *Knowledge-Based Systems*, **24**, 1151-1158. https://doi.org/10.1016/j.knosys.2011.04.011

[2] 唐静静, 田英杰. 多视角学习综述[J]. 数学建模及其应用, 2017, 6(3): 1-15, 25.

[3] Sun, S. (2013) A Survey of Multi-View Machine Learning. *Neural Computing and Applications*, **23**, 2031-2038. https://doi.org/10.1007/s00521-013-1362-6

[4] DeSa, V.R. (1993) Learning Classification with Unlabeled Data. 6*th International Conference on Neural Information Processing Systems*, Denver, November 1993, 112-119.

[5] Jiang, Y., Liu, J., Li, Z. and Lu, H. (2014) Semi-Supervised Unified Latent Factor Learning with Multi-View Data. *Machine Vision & Applications*, **25**, 1635-1645. https://doi.org/10.1007/s00138-013-0556-3

[6] Farquhar, J.D.R., Hardoon, D.R., Meng, H., Shawe-Taylor, J. and Szedmák, S. (2006) Two View Learning: SVM-2K, Theory and Practice. In: Weiss, Y., Schölkopf, B. and Platt, J., Eds., *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, 355-362.

[7] Chen, X., Chen, S., Xue, H. and Zhou, X. (2012) A Unified Dimensionality Reduction Framework for Semi-Paired and Semi-Supervised Multi-View Data. *Pattern Recognition*, **45**, 2005-2018. https://doi.org/10.1016/j.patcog.2011.11.008

[8] Bickel, S. and Scheffer, T. (2004) Multi-View Clustering. 2004 4*th IEEE International Conference on Data Mining*, Brighton, 1-4 November 2004, 19-26. https://doi.org/10.1109/ICDM.2004.10095

[9] Vapnik, V. (2006) Estimation of Dependences Based on Empirical Data. 2nd Edition, Springer, Berlin.

[10] Weston, J., Collobert, R., Sinz, F.H., Bottou, L. and Vapnik, V. (2006) Inference with the Universum. Machine Learning, *Proceedings of the* 23*rd International Conference* (*ICML* 2006), Pittsburgh, June 25-29 2006, 1009-1016. https://doi.org/10.1145/1143844.1143971

[11] Liu, C.L., Hsaio, W.H., Lee, C.H., Chang, T.-H. and Kuo, T.-H. (2017) Semi-Supervised Text Classification with Universum Learning. *IEEE Transactions on Cybernetics*, **46**, 462-473. https://doi.org/10.1109/TCYB.2015.2403573

[12] Richhariya, B. and Gupta, D. (2018) Facial Expression Recognition Using Iterative Universum Twin Support Vector Machine. *Applied Soft Computing*, **76**, 53-67. https://doi.org/10.1016/j.asoc.2018.11.046

[13] Qiu, J., Wang, Y., Pan, Z. and Jia, B. (2014) Semi-Supervised Feature Selection with Universum Based on Linked Social Media Data. *IEICE Transactions on Information and Systems*, E97, 2522-2525. https://doi.org/10.1587/transinf.2014EDL8033

[14] Zhang, D., Wang, J., Wang, F. and Zhang, C. (2008) Semi-Supervised Classification with Universum. *Proceedings of the* 2008 *SIAM International Conference on Data Mining*, Atlanta, 24-26 April 2008, 323-333. https://doi.org/10.1137/1.9781611972788.29

[15] Xie, X. and Sun, S. (2015) Multi-View Twin Support Vector Machines. *Intelligent Data Analysis*, **19**, 701-712. https://doi.org/10.3233/IDA-150740

[16] Tang, J., Li, D., Tian, Y. and Liu, D. (2018) Multi-View Learning Based on Nonparallel Support Vector Machine. *Knowledge-Based Systems*, **158**, 94-108. https://doi.org/10.1016/j.knosys.2018.05.036