ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 1,Jan. 2015,pp. 106－116
DOI:10.16337/j.1004-9037.2015.01.010

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

一种基于多视图数据的半监督特征选择和聚类算法

汪荆琪　徐林莉

(中国科学技术大学计算机科学与技术学院,合肥,230027)

摘　要:高维数据中许多特征之间互不相关或冗余,这给传统的学习算法带来了巨大的挑战。为了解决该问题,特征选择应运而生。与此同时,许多实际问题中数据存在多个视图而且数据的标签难以获取,多视图学习和半监督学习成为机器学习中的热点问题。本文研究怎样从"部分标签"的多视图数据中选择最大相关最小冗余的特征子集,提出一种基于多视图的半监督特征选择方法。为了剔除冗余和无关的特征,探索蕴含于多视图数据中的互补信息以及每个视图中不同特征之间的冗余关系,并利用少量标签数据蕴含的信息协同未标签数据同时进行特征选择。实验结果验证了本算法能够获得很好的特征选择效果及聚类效果。

关键词:聚类;半监督;特征选择;多视图

中图分类号:TP181　　**文献标志码**:A

Semi-supervised Feature Selection and Clustering for Multi-view Data

Wang Jingqi, Xu Linli

(School of Computer Science and Technology, University of Science and Technology of China, Hefei, 230027, China)

Abstract: Lots of features in high-dimensional data are redundant or irrelevant. To tackle this problem, the concept of feature selection is introduced. In the meantime, many problems in machine learning involve examples that are naturally comprised of multiple views and with a limited number of labels. Multi-view learning and semi-supervised learning become the hotspots in machine learning. Hence authors investigate how to select relevant features with minimum redundancy from multi-view data with a limited number of labels, and propose a semi-supervised feature selection and clustering framework. To remove redundant and irrelevant features, authors exploit relations among views and relations among features in each view, and use a limited number of labeled data to help feature selection. The proposed framework in multi-view datasets is systematically evalated, and the results demonstrate the effectiveness and potential of the proposed method.

Key words: clustering; semi-supervised; feature selection; multi-view

引　　言

在很多实际的应用领域,经常会遇到许多高维数据,如图像视频、Web 文本和基因序列等。高维数据

基金项目:国家自然科学基金(61375060)资助项目;中央高校基本科研业务费专项资金(WK0110000036)资助项目。
收稿日期:2014-03-25;**修订日期**:2014-05-12

中许多特征之间是互不相关或冗余的,它们的存在给传统的学习算法带来了巨大的挑战。特征选择作为一种数据预处理技术,能够有效处理大规模高维数据。特征选择通常是指根据某种评估标准,从原始特征空间中选择一个最优或最有效的特征子集代替原始特征空间的过程。其目的就是尽可能剔除数据中的不相关或冗余的特征,降低计算复杂度,提高学习算法的性能。特征选择是数据挖掘、机器学习中一个活跃的研究领域,而且有着非常广泛的应用,比如基因分析、文本挖掘、图像处理和目标识别[1-3]等。

在一些实际问题中,同一事物可以从多种不同的途径或不同的角度进行描述,即所获取的数据常常可以由多个特征集合表示,这些不同的特征空间从不同视角揭示了事物的不同属性,这类数据通常被称为多视图数据。对此类多视图数据的研究被称为多视图学习。随着多视图数据的增加,研究如何找出多视图数据所包含的重要信息已成为当前机器学习领域研究的热点,研究者已经在多视图学习方面取得了一定的成果,并运用到实际的应用中[4-7]。研究结果表明,合理地探索运用蕴含于多视图数据中的互补信息和关系可以大大提升学习效果。

按照所使用数据集数目的不同,特征选择可分为单视图与多视图两种。现有的特征选择方法[8-15]大多用于处理传统的单视图数据,如 mRMR[8],Laplacian Score[9],SPEC[10],MCFS[11]等。2013 年,Tang 等人提出了处理多视图数据的特征选择算法 MVFS[16],该方法对每个视图独立地执行特征选择,并通过谱分析约束不同的视图,使其满足多视图学习的一致性原则。MVFS 方法虽然考虑了不同视图之间的相互联系,却忽略了每个视图中不同特征之间的关系。同年,Wang 等人提出多视图聚类和特征学习方法 wang_MVFS[17],该方法将多视图数据组合成单视图数据,并通过 Group Lasso[18]正则化项约束每个视图的重要性,该方法既适用于监督学习也适用于无监督学习。

另一方面,随着采集数据越来越容易,采集一个样本远远要比标签一个样本来得容易,所以现实情况的训练数据集往往是由少数的有标签样本加上大量的无标签样本构成,半监督特征选择就是处理这种数据集的特征选择方法,它利用样本先验信息来改善无监督特征选择算法的性能。

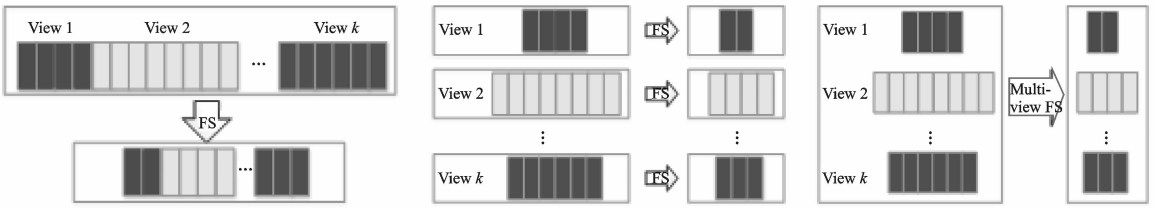
本文提出了一种新的基于多视图学习的半监督特征选择方法(Semi-supervised feature selection for multi-view data,SSMVFS),该方法不仅探索了蕴含于多视图数据之间的关系和互补信息,还考虑了每个视角中不同特征之间的冗余关系,并利用少量标签信息协同大量未标签数据一起学习,用于处理"部分标签"的多视图数据。SSMVFS 改进了 MVFS 中忽略每个视图中不同特征间冗余关系的问题,并且基于半监督学习,利用样本的先验信息来改善无监督特征选择算法的性能。SSMVFS 算法在选择最大相关最小冗余特征子集的同时,能够学习到数据的类别分配信息,即可同时进行特征选择和聚类学习。在 5 个多视图数据集上的实验表明本算法能够获得较好的特征选择效果及聚类效果。

1 基于多视图的半监督特征选择和聚类基本原理

1.1 特征选择算法

现有的特征选择方法大多用于处理传统的单视图数据。多视图数据给传统的特征选择算法带来了挑战:如何表示不同视图之间的关系;如何利用这些关系来提高特征选择的性能。应用传统的特征选择方法处理多视图数据有两个简单的策略是:(1)将多视图数据组合成单视图数据,再利用单视图特征选择方法处理,如图 1(a)所示;(2)对多视图数据的每个视图独立地执行传统的单视图特征选择,如图 1(b)所示。组合策略明显忽略了不同特征空间之间的差异,而分离策略却认为各视图之间相互独立。然而,不同视图之间是有内在关联的,因为它们描述相同的一组对象。一般情况下,多个视图可以相互补充信息,如标签和文字描述为图片提供语义信息,可以帮助获取更好的学习性能。本文研究的多视图特征选择方法通过探索不同视图之间的关系的同时对所有视图进行特征选择,如图 1(c)所示。

在处理多视图数据时,多视图与单视图特征选择问题具有明显的区别:(1)多视图特征选择通过探索不同视图之间的关系同时在多个视图数据上学习,而单视图特征选择则在每一个视图上独立学习;



(a) 组合策略 (a) Concatenating strategy (b) 分离策略 (b) Separation strategy (c) 多视图特征选择 (c) Multi-view feature selection

图 1 处理多视图数据的策略

Fig. 1 Different strategies for multi-view data

(2)多视图特征选择可以在异构的特征空间中选择特征,而单视图特征选择只能从同构的特征空间中选取特征。

文献[16]提出了一种多视图特征选择方法 MVFS。给定数据集$(\boldsymbol{X}^{(1)},\cdots,\boldsymbol{X}^{(k)})$(有 k 个视图),该方法将多视图数据的特征选择问题转化为求解如下问题

$$\min_{\boldsymbol{W},\boldsymbol{Z}} J(\boldsymbol{W},\boldsymbol{Z})=\sum_{v=1}^{k}\lambda_v(\mathrm{tr}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{L}^{(v)}\boldsymbol{Z})+\alpha(\|\boldsymbol{X}^{(v)\mathrm{T}}\boldsymbol{W}^{(v)}-\boldsymbol{Z}\|_F^2+\beta\|\boldsymbol{W}^{(v)}\|_{2,1})) \tag{1}$$

$$\text{s.t.}\quad \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}=\boldsymbol{I},\boldsymbol{Z}\geqslant 0$$

式中:$\boldsymbol{W}=(\boldsymbol{W}^{(1)},\cdots,\boldsymbol{W}^{(k)})$表示特征对样本点类别确定的贡献度,$\boldsymbol{L}^{(v)}$为 $\boldsymbol{X}^{(v)}$ 的拉普拉斯矩阵,$\boldsymbol{Z}$ 为指示矩阵,$\{\lambda_v\}_{v=1}^{k}$,α 和 β 为调节参数。

文献[17]提出的视图聚类和特征学习方法 wang_MVFS 可描述为解如下优化问题

$$\min_{\boldsymbol{W},\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}=\boldsymbol{I}}\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}+\boldsymbol{1}_n\boldsymbol{b}^{\mathrm{T}}-\boldsymbol{Z}\|_F^2+\alpha\|\boldsymbol{W}\|_{G1}+\beta\|\boldsymbol{W}\|_{2,1} \tag{2}$$

式中:$\boldsymbol{W}$ 表示特征对样本点类别确定的贡献度,$\boldsymbol{Z}$ 为指示矩阵,α 和 β 为调节参数。

1.2 基本原理

在多视图的半监督学习中,一方面数据有多个视图,即 $\boldsymbol{x}=\{\boldsymbol{x}^{(1)},\cdots,\boldsymbol{x}^{(k)}\}$,另一方面数据集由标签数据集$(\boldsymbol{X}_L,\boldsymbol{Y}_L)=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{x}_l,\boldsymbol{y}_l)\}$和未标签数据集 $\boldsymbol{X}_U=\{\boldsymbol{x}_{l+1},\cdots,\boldsymbol{x}_{l+u}\}$组成。假设数据集为$(\boldsymbol{X},\boldsymbol{Y}_L)$,其中 $\boldsymbol{X}=\{\boldsymbol{X}^{(1)},\cdots,\boldsymbol{X}^{(k)}\}=(\boldsymbol{X}_L;\boldsymbol{X}_U)=\{\boldsymbol{x}_i\}_{i=1}^{n}\in\mathbf{R}^{d\times n}$,$n$ 为样本数,$\boldsymbol{x}_i\in\mathbf{R}^d$ 是由 k 个视图组成的 d 维特征空间,其中第 v 视图 $\boldsymbol{X}^{(v)}$ 有 d_v 个特征$(d=\sum_{v=1}^{k}d_v)$,$\boldsymbol{Y}_L\in\mathbf{R}^{l\times c}$是标签数据集 $\boldsymbol{X}_L$ 的类别信息(一般地,$l<<n-l$),c 为类别数。多视图半监督学习应考虑如何利用多个视图蕴含的信息和标签数据及未标签数据蕴含的信息。

现有的多视图学习大多假设所有的视图共享标签信息,通过共享的标签信息探索不同的视图数据之间的关系。然而,数据集 $\boldsymbol{X}$ 中大部分数据的标签都难以获取,所以本文利用伪标签来探索不同视图之间的关系。这里假设伪标签为 $\boldsymbol{Z}=[\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_n]^{\mathrm{T}}\in\mathbf{R}^{n\times c}$,若 $\boldsymbol{x}_i$ 属于第 j 类,$\boldsymbol{Z}(i,j)=1$;否则,$\boldsymbol{Z}(i,j)=0$。因此,$\boldsymbol{Z}$ 满足条件:$\boldsymbol{Z}(i,:)\in\{0,1\}^n$,$\|\boldsymbol{Z}(i,:)\|_0=1$,$\forall i,1\leqslant i\leqslant n$。另外,已标签数据 $\boldsymbol{X}_L$ 对应的伪标签 $\boldsymbol{Z}(1:l,:)$应尽量与已知的标签信息 $\boldsymbol{Y}_L$ 一致;而且同一类别中的样本点应相似,则对每一个视图数据的限制条件可被描述为如下优化问题

$$\min_{\boldsymbol{Z}}\mathrm{tr}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{L}^{(v)}\boldsymbol{Z})+\delta\|\boldsymbol{Z}-\boldsymbol{Y}\|_F^2 \tag{3}$$

式中:$\boldsymbol{Y}=[\boldsymbol{Y}_L;\boldsymbol{Z}(l+1:n,:)]$为扩充的标签,$\boldsymbol{L}^{(v)}=\boldsymbol{D}^{(v)}-\boldsymbol{S}^{(v)}$为第 v 视图数据 $\boldsymbol{X}^{(v)}$ 的拉普拉斯矩阵,对角阵 $\boldsymbol{D}^{(v)}=\sum_{i=1}^{n}\boldsymbol{S}^{(v)}(i,j)$,$\boldsymbol{S}^{(v)}$ 为 $\boldsymbol{X}^{(v)}$ 的相似度矩阵,实验中采用 RBF 核计算相似度矩阵,δ 为调节参数。

令 $\boldsymbol{s}^{(v)}=\pi\left(\overbrace{0,\cdots,0}^{d_v-k_v},\overbrace{1,\cdots,1}^{k_v}\right)$,$\pi$ 是置换函数,k_v 表示在视图 v 中选择的特征数目,$\boldsymbol{s}^{(v)}(j)=1$ 表示视图 v 特征空间中的第 j 特征被选中。经过特征选择后的视图 v 数据可从原始数据 $\boldsymbol{X}^{(v)}$ 变换为 $\hat{\boldsymbol{A}}=\mathrm{diag}(\boldsymbol{s}^{(v)})\boldsymbol{X}^{(v)}$。则视图 v 的特征选择过程可表示成如下优化目标

$$\begin{aligned}&\min_{\boldsymbol{W},\boldsymbol{Z},\boldsymbol{s}^{(v)}}\left\|\hat{\boldsymbol{X}}^{(v)\mathrm{T}}\boldsymbol{W}^{(v)}-\boldsymbol{Z}\right\|_F^2+\gamma\mathrm{tr}(\boldsymbol{W}^{(v)\mathrm{T}}\boldsymbol{R}^{(v)}\boldsymbol{W}^{(v)})\\&\text{s.t.}\quad \boldsymbol{s}^{(v)}\in\{0,1\}^n,\boldsymbol{s}^{(v)\mathrm{T}}\boldsymbol{1}_n=k_v\\&\boldsymbol{Z}(i,:)\in\{0,1\}^n,\left\|\boldsymbol{Z}(i,:)\right\|_0=1,1\leqslant i\leqslant n\end{aligned}\tag{4}$$

式中:$\boldsymbol{W}=(\boldsymbol{W}^{(1)},\cdots,\boldsymbol{W}^{(k)})\in\mathbf{R}^{d\times c}$,$\boldsymbol{W}^{(v)}\in\mathbf{R}^{d_v}$ 表示视图 v 特征空间中的特征对样本点类别确定的贡献度,值越大,对应的特征越相关;$\boldsymbol{C}^{(v)}=(c_{kl})\in[-1,1]^{d_v\times d_v}$ 表示视图 v 特征空间中不同特征之间的关系,体现特征之间的冗余性,令

$$\begin{aligned}&\boldsymbol{R}^{(v)}=\boldsymbol{C}^{(v)}\circ\boldsymbol{C}^{(v)}(r_{kl}=c_{kl}^2)\\&c_{kl}=\frac{\sum_{j=1}^{n}x_{kj}x_{lj}}{\sqrt{\sum_{j=1}^{n}x_{kj}^2}\sqrt{\sum_{j=1}^{n}x_{lj}^2}}\end{aligned}\tag{5}$$

γ 控制特征子集的冗余性。零范数与整数的限制条件使得式(2)中的优化问题难以求解,所以放松 $\boldsymbol{Z}$ 的限制条件为

$$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}=\boldsymbol{I},\boldsymbol{Z}\geqslant 0\tag{6}$$

另外,式(4)中 $\mathrm{diag}(\boldsymbol{s}^{(v)})$ 均与 $\boldsymbol{W}^{(v)}$ 一起以 $\mathrm{diag}(\boldsymbol{s}^{(v)})\boldsymbol{W}^{(v)}$ 的形式出现,其中 $\mathrm{diag}(\boldsymbol{s}^{(v)})$ 有 d_v-k_v 行为全零值,则 $\mathrm{diag}(\boldsymbol{s}^{(v)})\boldsymbol{W}^{(v)}$ 是行稀疏的。记 $\boldsymbol{W}^{(v)}=\mathrm{diag}(\boldsymbol{s}^{(v)})\boldsymbol{W}^{(v)}$,并采用 $l_{2,1}$ 范数保证 $\boldsymbol{W}^{(v)}$ 的行稀疏性质($\|\boldsymbol{W}\|_{2,1}=\sum_{i=1}^{d}\|\boldsymbol{W}_i\|_2$),从而达到特征选择的目的。

通过以上分析,多视图数据的半监督特征选择问题的实现可以转化为求解如下优化问题

$$\begin{aligned}&\min_{\boldsymbol{W},\boldsymbol{Z}}J(\boldsymbol{W},\boldsymbol{Z})=\sum_{v=1}^{k}(\mathrm{tr}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{L}^{(v)}\boldsymbol{Z})+\alpha(\left\|\boldsymbol{X}^{(v)\mathrm{T}}\boldsymbol{W}^{(v)}-\boldsymbol{Z}\right\|_F^2+\beta\left\|\boldsymbol{W}^{(v)}\right\|_{2,1}+\gamma\mathrm{tr}(\boldsymbol{W}^{(v)\mathrm{T}}\boldsymbol{R}^{(v)}\boldsymbol{W}^{(v)}))+\\&\qquad\delta\left\|\boldsymbol{Z}-\boldsymbol{Y}\right\|_F^2\\&\text{s.t.}\quad \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}=\boldsymbol{I},\boldsymbol{Z}\geqslant 0\end{aligned}\tag{7}$$

式中:参数 λ_v 用于控制不同视图数据的重要性,且 $\sum_{v=1}^{k}\lambda_v=1$;$\alpha$ 为控制不同视角的一致性原则和每个视角下的特征学习的可调节参数;β 为控制 $l_{2,1}$ 范数正则化项的调节参数,值越大,求得的权值矩阵 $\boldsymbol{W}$ 越稀疏,从而达到控制所选特征数据的效果;δ 为半监督项的系数,控制已标签样本指导特征学习的程度。

为了解得 $\boldsymbol{W}$ 和 $\boldsymbol{Z}$,这里采用迭代交替优化方法解决上述优化问题,即交替更新 $\boldsymbol{W}$ 和 $\boldsymbol{Z}$。

(1)固定 $\boldsymbol{Z}$,求 $\{\boldsymbol{W}^{(v)}\}_{v=1}^{k}$。此时,可对每一个 $\boldsymbol{W}^{(v)}$ 独立求解,即求解如下优化问题

$$\begin{aligned}&\min_{\boldsymbol{W}^{(v)}}J(\boldsymbol{W}^{(v)})=\left\|\boldsymbol{X}^{(v)\mathrm{T}}\boldsymbol{W}^{(v)}-\boldsymbol{Z}\right\|_F^2+\beta\left\|\boldsymbol{W}^{(v)}\right\|_{2,1}+\\&\qquad\gamma\mathrm{tr}(\boldsymbol{W}^{(v)\mathrm{T}}\boldsymbol{R}^{(v)}\boldsymbol{W}^{(v)})\end{aligned}\tag{8}$$

目标函数 $J(\boldsymbol{W}^{(v)})$ 对 $\boldsymbol{W}^{(v)}$ 求导,并令导数等于 0,得到 $\boldsymbol{W}^{(v)}$ 的更新规则

$$\boldsymbol{W}^{(v)}=(\boldsymbol{X}^{(v)\mathrm{T}}\boldsymbol{X}^{(v)}+\beta\boldsymbol{D}^{(v)}+\gamma\boldsymbol{R}^{(v)})^{-1}\boldsymbol{X}^{(v)}\boldsymbol{Z}\tag{9}$$

式中,$\boldsymbol{D}^{(v)}$ 为角矩阵,其第 i 个对角元素为 $\boldsymbol{D}^{(v)}(i,i)=1/2(\left\|\boldsymbol{W}^{(v)}(i,:)\right\|_2)$。

定理 1 根据式(9)中规则更新 $\boldsymbol{W}^{(v)}$,可以保证目标函数 $J(\boldsymbol{W}^{(v)})$ 的值单调递减。

证明:见文献[14,15]。

(2)固定 $\{\boldsymbol{W}^{(v)}\}_{v=1}^{k}$,求 $\boldsymbol{Z}$。

当 $\boldsymbol{W}$ 固定时,$\boldsymbol{Z}$ 通过求解如下优化问题求得

$$\min_{\boldsymbol{Z}} \mathrm{J}(\boldsymbol{Z}) = \sum_{v=1}^{k} \lambda_v (\mathrm{tr}(\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{L}^{(v)} \boldsymbol{Z}) + \alpha \left\| \boldsymbol{X}^{(v)\mathrm{T}} \boldsymbol{W}^{(v)} - \boldsymbol{Z} \right\|_F^2) + \delta \left\| \boldsymbol{Z} - \boldsymbol{Y} \right\|_F^2$$
$$\text{s. t.} \quad \boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z} = \boldsymbol{I}, \boldsymbol{Z} \geqslant 0 \tag{10}$$

记 $\boldsymbol{M} = \sum_{v=1}^{k} \lambda_v \boldsymbol{L}^{(v)}$, $\boldsymbol{Q} = \frac{1}{\tau} (\sum_{v=1}^{k} \lambda_v \alpha \boldsymbol{X}^{(v)\mathrm{T}} \boldsymbol{W}^{(v)} + \delta \boldsymbol{Y})$, $\tau = \alpha + \delta$, 则式(10)中的优化目标变换为

$$\min_{\boldsymbol{Z}} J(\boldsymbol{Z}) = \mathrm{tr}(\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{M} \boldsymbol{Z}) + \tau \left\| \boldsymbol{Z} - \boldsymbol{Q} \right\|_F^2$$
$$\text{s. t.} \quad \boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z} = \boldsymbol{I}, \boldsymbol{Z} \geqslant 0 \tag{11}$$

引入拉格朗日算子 $\boldsymbol{\Gamma}$ 和 $\boldsymbol{\Lambda}$, 得到拉格朗日公式为

$$L(\boldsymbol{Z}) = \mathrm{tr}(\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{M} \boldsymbol{Z}) + \tau \mathrm{tr}(\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z} - 2\boldsymbol{Q}^{\mathrm{T}} \boldsymbol{Z}) + \mathrm{tr}(\boldsymbol{\Gamma}(\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z} - \boldsymbol{I})) - \mathrm{tr}(\boldsymbol{\Lambda} \boldsymbol{Z}) \tag{12}$$

然后对 $\boldsymbol{Z}$ 求导, 并令导数等于 0

$$\frac{\partial L(\boldsymbol{Z})}{\partial \boldsymbol{Z}} = 2\boldsymbol{M}\boldsymbol{Z} + 2\tau(\boldsymbol{Z} - \boldsymbol{Q}) + 2\boldsymbol{\Gamma}\boldsymbol{Z} - \boldsymbol{\Lambda} = 0。$$

利用 KKT 条件, $\boldsymbol{\Lambda}(i,j)\boldsymbol{Z}(i,j) = 0$, 可以得到

$$(\boldsymbol{M}\boldsymbol{Z} + \tau(\boldsymbol{Z} - \boldsymbol{Q}) + \boldsymbol{\Gamma}\boldsymbol{Z})(i,j)\boldsymbol{Z}(i,j) = 0 \tag{13}$$

则 $\boldsymbol{Z}$ 的更新规则为

$$\boldsymbol{Z}(i,j) \leftarrow \boldsymbol{Z}(i,j) \sqrt{\frac{(\boldsymbol{M}^{-}\boldsymbol{Z} + \tau\boldsymbol{Q}^{+} + \boldsymbol{Z}\boldsymbol{\Gamma}^{-})(i,j)}{(\boldsymbol{M}^{+}\boldsymbol{Z} + \tau(\boldsymbol{Q}^{-} + \boldsymbol{Z}) + \boldsymbol{Z}\boldsymbol{\Gamma}^{+})(i,j)}} \tag{14}$$

式中: $\boldsymbol{X}^{+}(i,j) = (|\boldsymbol{X}(i,j)| + \boldsymbol{X}(i,j))/2$; $\boldsymbol{X}^{-}(i,j) = (|\boldsymbol{X}(i,j)| - \boldsymbol{X}(i,j))/2$。

定理 2 根据式(14)中规则更新 $\boldsymbol{Z}$, 可以保证目标函数 $J(\boldsymbol{Z})$ 的值单调递减。

证明: 见文献[16]。

2 基于多视图的半监督特征选择和聚类算法

给定 n 个包含 k 个视图的 d 维多视图样本, 其中第 v 视图包含 d_v 个特征($1 \leqslant v \leqslant k$), 记样本集为 $\boldsymbol{X} = (\boldsymbol{X}^{\langle 1 \rangle}, \cdots, \boldsymbol{X}^{\langle k \rangle}) = (\boldsymbol{X}_L; \boldsymbol{X}_U) = \{\boldsymbol{x}_i\}_{i=1}^{n} \in \mathbf{R}^{d \times n}$, $\boldsymbol{Y}_L \in \mathbf{R}^{l \times c}$ 为 $\boldsymbol{X}_L$ 的标签信息, c 为类别数。特征选择算法从视图 v 特征空间中选择包含 k_v 个特征的最大相关最小冗余的特征子集。本文提出的基于多视图的半监督特征选择算法 SSMVFS 表述如下。

算法 1 SSMVFS

输入: 数据集 $(\boldsymbol{X}, \boldsymbol{Y}_L)$, 参数 $\{k_v, \lambda_v\}_{v=1}^{k}$, $\alpha, \beta, \gamma, \delta$

输出: $\sum_{v=1}^{k} k_v$ 个特征, $\boldsymbol{Z}$

第 1 步: 初始化

for $v = 1$ to k

构建 $\boldsymbol{X}^{(v)}$ 的拉普拉斯矩阵 $\boldsymbol{L}^{(v)}$;

根据式(5)构建 $\boldsymbol{X}^{(v)}$ 中特征间的关系矩阵 $\boldsymbol{R}^{(v)}$;

初始化 $\boldsymbol{D}^{(v)} = \boldsymbol{I}_{d_v}$;

end for

构建 $\boldsymbol{M} = \sum_{v=1}^{k} \lambda_v \boldsymbol{L}^{(v)}$, $\boldsymbol{M}^{+}$ 和 $\boldsymbol{M}^{-}$;

初始化 $\boldsymbol{Z}$ 为 $\boldsymbol{X}$ 的 k-means 聚类结果;

构建扩充的标签矩阵 $\boldsymbol{Y} = [\boldsymbol{Y}_L; \boldsymbol{Z}(l+1:n, :)]$;

第 2 步: 交替迭代更新 $\boldsymbol{W}$ 和 $\boldsymbol{Z}$。

while 不收敛 do

① 更新 $\boldsymbol{W}$(独立更新每个 $\boldsymbol{W}^{(v)}$)

for $v=1$ to k

根据式(9)更新 $\boldsymbol{W}^{(v)}$;

更新 $\boldsymbol{D}^{(v)}$,$\boldsymbol{D}^{(v)}(i,i)=1/(2\|\boldsymbol{W}^{(v)}(i,:)\|_2)$;

end for

② 更新 $\boldsymbol{Z}$。

构建 $\boldsymbol{Q}=\frac{1}{\alpha+\delta}(\sum_{v=1}^{k}\lambda_v\alpha\boldsymbol{X}^{(v)\mathrm{T}}\boldsymbol{W}^{(v)}+\delta\boldsymbol{Y})$,$\boldsymbol{Q}^{+}$ 和 $\boldsymbol{Q}^{-}$;

计算 $\boldsymbol{\Gamma}=\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Q}-\boldsymbol{I}-\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{Z}$,$\boldsymbol{\Gamma}^{+}$ 和 $\boldsymbol{\Gamma}^{-}$;

根据式(14)更新 $\boldsymbol{Z}$;

更新 $\boldsymbol{Y}=[\boldsymbol{Y}_L;\boldsymbol{Z}(l+1:n,:)]$;

endwhile

第 3 步:依据 $\boldsymbol{W}$ 的值选择特征。

for $v=1$ to k

计算 $\|\boldsymbol{W}^{(v)}(i,:)\|_2$,按降序排列,选择前 k_v 个对应的特征组成特征子集;

end for

聚类规则:算法 1 中的输出 $\boldsymbol{Z}$ 可视作聚类结果,$\boldsymbol{Z}$ 的每行只有一个正值,选择该值对应的列号作为类别号。

3 实验结果与分析

3.1 数据集

实验采用 5 个多视图的数据集,表 1 列举了这些数据集的概要信息,如每个视图的特征数、样本数、和类别数。

表 1 实验数据的简要描述

Table 1 Summary of data sets

数据集	Digit	3-Source	Movie	ML Text	Cora
视图 1	Fourier(76)	B(3 315)	Key(1 878)	EN(2 000)	CI(2 708)
视图 2	Pixel(240)	G(3 358)	Act(1 398)	FR(2 000)	CO(1 433)
视图 3	—	R(2 813)	—	GR(2 000)	—
样本数	2 000	169	617	1 200	2 708
类别数	10	6	17	6	7

Digit 是来自公用的 UCI 机器学习存储库[19]的手写数字(0～9)数据集,包含 2 000 个样本点,视图 1 数据 Fourier 有 76 个特征,视图 2 数据 Pixel 有 240 个特征。3-Source 是一个多视图的文本数据集,共有 948 篇新闻报道,其中 169 个样本同时出现在 3 个新闻网站——BBC,Reuters 和 The Guardian。Movie 数据集包含 17 类的 617 个电影样本,每个样本由 1 878 个关键词描述,且有 1398 个相关演员。ML Text 数据集取自 Reuters RCV1/RCV2 数据集,Reuters 数据集包括英语、法语、德语、意大利语和西班牙语 5 种语言的文档和它们对应的翻译为其他语言的平行文档,所有语言的文档都有相同的类标结构,每种语言有 6 种类别(CCAT, C15, ECAT, E21,GCAT 和 M11);本文使用英语描述作为第一视图,英译法和英译德分别作为第二和三视图,且随机选择 1 200 个样本点,每类 200 个。Cora 数据集由

3 0714篇学术论文组成，其中关于机器学习的论文分为 7 个类别，每篇论文由标题、作者、摘要、参考文献等内容构成，实验采用其中 2 708 个样本点，并将论文内容和引文作为两个视图。

3.2 实验设置

为了验证 SSMVFS 特征选择算法的性能，SSMVFS 方法将与 5 种典型的单视角特征选择算法(Laplacian score，SPEC 和 MCFS，sSelect[20] 和 mRMR)和两种多视图特征选择算法(MVFS 和 wang_MVFS)在上述 5 个数据集上进行实验比较。其中，mRMR 是经典的基于最大相关最小冗余的特征选择算法；Laplacian Score，SPEC，MCFS 和 sSelect 都是基于谱分析的特征选择方法，既可用于监督学习又可用于非监督学习；sSelect 是基于半监督的谱特征选择方法。

实验中，首先执行特征选择算法选择特征，令 $\boldsymbol{s}^{(v)}=\pi\left(\overbrace{0,\cdots,0}^{d_v-k_v},\overbrace{1,\cdots,1}^{k_v}\right)$，$k_v$ 表示在视图 v 中选择的特征数目，$\boldsymbol{s}^{(v)}(j)=1$ 表示视图 v 中的第 j 特征被选中。经过特征选择后的视图 v 可从原始数据 $\boldsymbol{X}^{(v)}$ 变换为 $\hat{\boldsymbol{X}}^{(v)}=\mathrm{diag}(\boldsymbol{s}^{(v)})\boldsymbol{X}^{(v)}$，并记 $\hat{\boldsymbol{X}}=(\hat{\boldsymbol{X}}^{(1)},\cdots,\hat{\boldsymbol{X}}^{(k)})$ 为特征选择之后的多视图数据。然后，对 $\hat{\boldsymbol{X}}$ 进行聚类，采用两种常用的指标覆盖率(Accuracy，ACC)和归一化互信息(Normalized mutual information，NMI)[9] 评价聚类性能，并用聚类结果来验证特征选择算法的性能。

实验中 SSMVFS 方法采用的参数如下：在 Digit，Movie 和 Cora 数据集上，$\lambda_1=0.5$，$\lambda_2=0.5$，在 ML Text 和 3-Source 数据集上，$\lambda_1=0.4$，$\lambda_2=0.3$，$\lambda_3=0.3$；在所有 5 个数据集上，$\beta=0.1$，并使用交叉验证挑选适当的参数 α，γ 和 δ。sSelect 算法有一个参数 λ，根据 Zhao 等人的实验结果，实验中采用 $\lambda=0.1$。MVFS 算法有 $k+2$ 个参数 $\{\lambda_v\}_{v=1}^{k}$，α 和 β，其取值方法同 SSMVFS。wang_MVFS 算法中有两个参数 α 和 β，实验通过交叉验证选择参数。

另外，SSMVFS 和 sSelect 半监督特征选择算法中在每个数据集中随机选择 10%的样本作为标签数据，剩下的作为无标签数据。

3.3 特征选择的多视图聚类性能

令 $k_v=0.1*d_v$，执行特征选择算法，即在每个视图中选择 10%最相关的特征，再对经过特征选择后保留的多视图数据集 $\hat{\boldsymbol{X}}$ 执行多视图聚类算法 CMSC[21]。CMSC 的聚类覆盖率和归一化互信息分别如表 2，3 所示。表中“All Features”表示原始特征空间的聚类结果，“—”表示 MCFS 无法得到特征选择结果，黑体字表示每列的最大值。

表 2 特征选择算法的多视图聚类 ACC 比较

Table 2 Multi-view clustering performance comparison measured by ACC with different feature selection algorithms

%

数据集	Digit	3-Source	Movie	ML Text	Cora
LaplacianScore	45.95	18.93	16.05	18.00	14.70
SPEC	34.40	13.61	16.21	20.42	19.61
MCFS	70.05	—	15.07	20.17	19.35
sSelect	69.15	**30.18**	15.24	21.42	21.60
mRMR	49.15	27.81	15.40	23.67	22.23
MVFS	38.10	26.04	16.69	21.67	15.51
wang_MVFS	68.85	27.22	15.56	19.92	18.69
SSMVFS	**73.85**	29.59	**17.99**	**23.92**	**26.26**
All Features	71.05	25.44	14.10	21.33	25.26

表 3 特征选择算法的多视图聚类 NMI 比较

Table 3 Multi-view clustering performance comparison measured by NMI with different feature selection algorithms

%

数据集	Digit	3-Source	Movie	ML Text	Cora
LaplacianScore	47.68	6.71	12.55	3.12	2.00
SPEC	31.08	4.36	13.16	3.77	5.49
MCFS	62.51	—	12.23	4.48	4.41
sSelect	60.74	4.79	12.46	6.07	12.21
mRMR	46.11	7.11	**18.60**	**7.95**	12.92
MVFS	31.77	6.57	12.20	6.41	3.06
wang_MVFS	60.55	4.36	11.54	1.42	2.37
SSMVFS	**65.86**	**7.41**	13.92	5.85	**13.01**
All Features	64.26	4.49	11.10	2.39	0.98

表 2 的结果显示,除在 3-Source 上的聚类覆盖率略低于 sSelect 外,SSMVFS 算法在其他多视图数据集上的 ACC 是 8 种算法中最高的,这验证了本算法能够获得更好的特征选择效果。其中 SSMVFS 的结果要优于另外两个无监督的多视图特征选择算法 MVFS 和 wang_MVFS,这说明少量的标签信息有助于学习特征选择模型。表 3 中结果显示,SSMVFS 在 Digit,3-Source 和 Cora 数据集上的 NMI 是最大的,在 Movie 和 ML Text 数据集上的 NMI 值小于 mRMR。

另外,对比表 2,3 中 SSMVFS 与 All Features 的聚类结果,前者的聚类覆盖率和归一化互信息在 5 个数据集上均大于后者,这说明了采用 SSMVFS 特征选择算法能有效地剔除数据中不相关和冗余的特征,有助于提高聚类学习的性能。

3.4 特征选择的单视图聚类性能

以 Cora 数据集为例,Cora 有两个视图,第一个视图是引文信息(记为 CI),第二个视图是文本内容(记为 CO),设所选特征数目为 $k=k_1+k_2$,k_1 为视图 CI 中所选特征数,k_2 为视图 CO 中所选特征数。在 CI 和 CO 上分别执行 Laplacian Score,SPEC,MCFS,sSelect 和 mRMR 特征选择算法,在 Cora 上执行 SSMVFS,MVFS 和 wang_MVFS 特征选择算法。然后在特征选择之后的两个视图 $\hat{\boldsymbol{X}}^{(1)}$ 和 $\hat{\boldsymbol{X}}^{(2)}$ 上分别执行 k-means 聚类算法,且实验重复 20 次,取平均值作为最终结果。两个视图特征选择后的 ACC 和 NMI 随着所选特征数($k_1\in\{20,40,60,\cdots,1000\}$,$k_2\in\{10,30,50,\cdots,990\}$)的变化曲线平滑后如图 2 所示。

图 2 中结果显示,SSMVFS 算法在 CI 视角下的 k-means 聚类准确率始终优于其他 7 种算法,在 CO 视角下的 ACC 在所选特征数较少时要低于 sSelect 算法,但当所选特征数超过 270 时,SSMVFS 的聚类准确率要远远优于其它特征选择算法。SSMVFS 算法在 CI 和 CO 视角下的 NMI 在所选特征数较少时结果较差,只有在所选特征数较大时 SSMVFS 的 NMI 值要大于其他特征选择算法。

3.5 SSMVFS 算法的聚类性能

SSMVFS 算法也可被视为一个多视图的聚类算法。多视图特征选择算法 MVFS,wang_MVFS 和 SSMVFS 求解过程中都能得到指示矩阵 $\boldsymbol{Z}$。MVFS 和 SSMVFS 中 $\boldsymbol{Z}$ 可直接作为聚类结果,$\boldsymbol{Z}$ 每行有且仅有一个正值,选择该正值对应的列号作为类别号即可;wang_MVFS 算法获得 $\boldsymbol{Z}$ 后,对 $\boldsymbol{Z}$ 做 k-means 得到聚类结果。另外,这里也对比了多视图聚类算法 CMSC 的性能。4 种算法的聚类性能如表 4,5 所示。

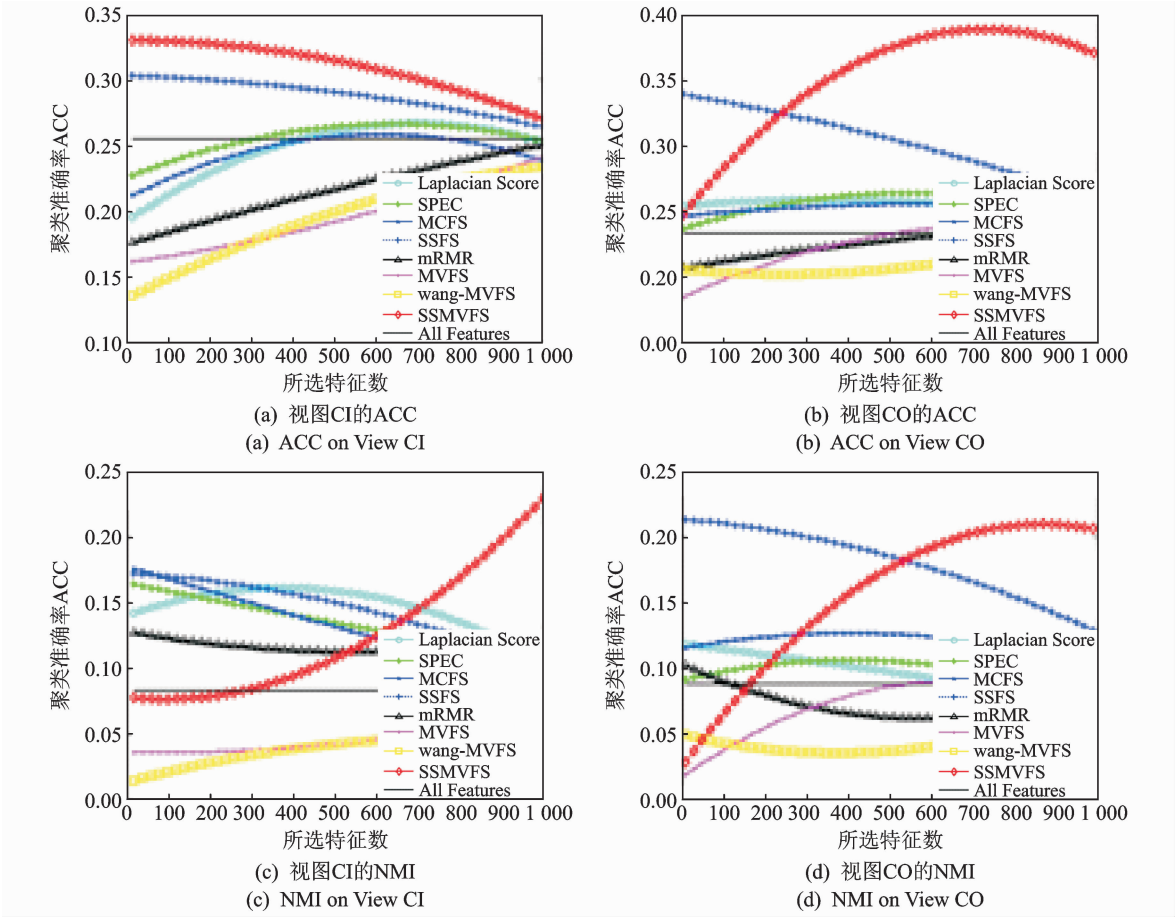


图 2 Cora 数据集上单视图聚类性能比较

Fig. 2 Single-view clustering performance comparison on Cora

表 4 多视图聚类算法的 ACC 比较

Table 4 Multi-view clustering performance comparison measured by ACC %

数据集	Digit	3-Source	Movie	ML Text	Cora
MVFS	34.30	24.26	12.28	**32.50**	19.83
wang_MVFS	18.72	21.27	14.50	28.83	28.21
SSMVFS	**68.05**	**29.59**	**23.50**	28.92	**32.16**
CMSC	71.05	25.44	14.10	21.33	25.26

表 5 多视图聚类算法的 NMI 比较

Table 5 Multi-view clustering performance comparison measured by NMI %

数据集	Digit	3-Source	Movie	ML Text	Cora
MVFS	40.68	6.32	0.34	20.30	12.63
wang_MVFS	18.46	4.45	0.84	16.20	22.05
SSMVFS	**73.82**	**6.38**	**25.33**	**21.93**	**22.66**
CMSC	64.26	4.49	11.10	2.39	0.98

表 4,5 的结果显示,SSMVFS 聚类覆盖率在除 ML Text 外的 4 个数据集上是最优的,SSMVFS 聚类的归一化互信息在所有数据集上都是最大的,这可能是因为 SSMVFS 利用了少量的标签信息。表中的实验结果表明,SSMVFS 作为多视图聚类算法时,其聚类准确率和归一化互信息均优于多视图聚类算法 wang_MVFS 和 CMSC,这验证了 SSMVFS 算法作为聚类算法的有效性。

3.6 SSMVFS 与 MVFS 所选特征子集的冗余性

图 3 给出了 5 个数据集上两种特征选择算法所选的特征子集的冗余度随着所选特征数目的变化曲线。其中横坐标表示所选特征数占总特征数的比例。图 3 中对应 SSMVFS 的虚线在 Digit,ML Text 和 Cora 数据集上大多在对应 MVFS 的实线之下,这说明在这三个数据集上 SSMVFS 选择的特征子集的冗余度小于 MVFS 所选特征子集的冗余度。在 3-Source 数据集上,选择 43%以上的特征时,SSMVFS 的冗余性较低。SSMVFS 和 MVFS 在 Movie 数据集上的冗余度相差不大。图中的结果显示,SSMVFS 选择的特征子集的冗余度基本比 MVFS 的低,这说明本文提出的 SSMVFS 算法对 MVFS 中不足的改进有效。

从上述实验结果可知,SSMVFS 算法能够获得较好的特征选择效果及聚类效果。比较 3 种多视图特征选择算法,可以发现 SSMVFS 与 MVFS 在实验中的运行速度要快于 wang_MVFS。

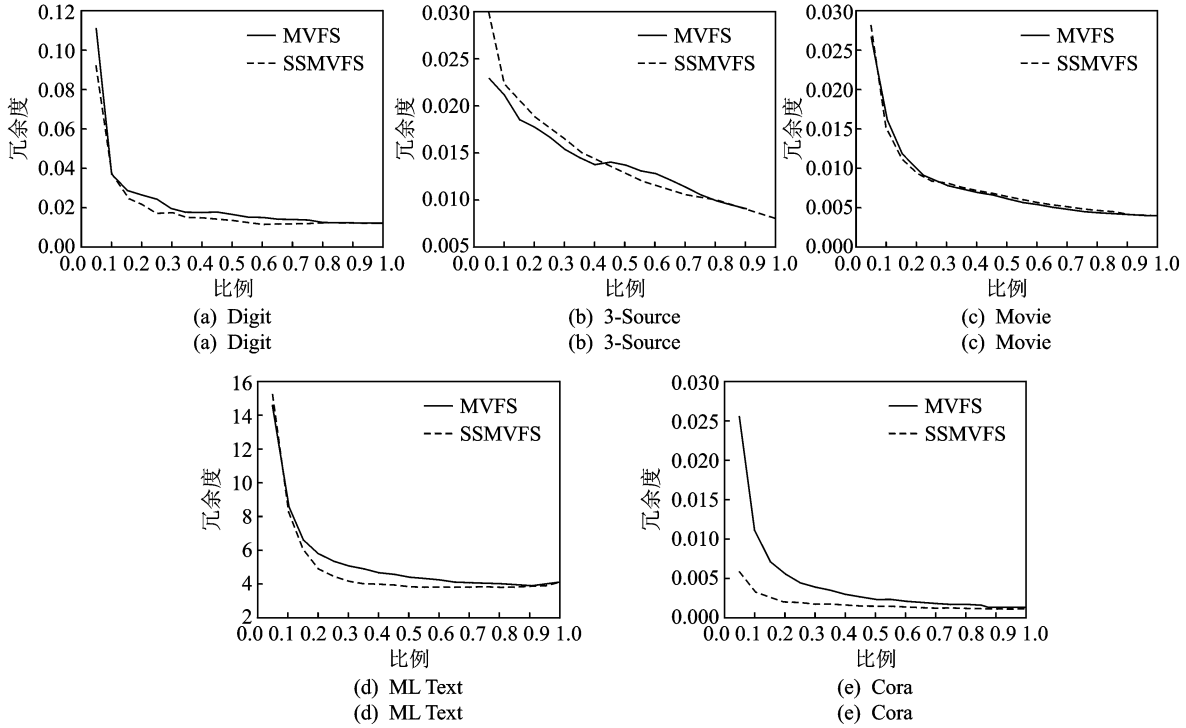


图 3 SSMVFS 与 MVFS 方法所选特征子集冗余性的比较

Fig. 3 Comparison of selected features vs. redundancy rate with SSMVFS/MVFS

4 结束语

本文结合多视图学习和半监督学习提出一种特征选择算法,利用 $l_{2,1}$ 范数实现每个视图的特征选择,基于谱分析衡量多个视图之间的关系,并比较伪类标与扩展的标签信息来实现半监督学习。与现有的多视图特征选择方法相比,本文提出的方法不仅考虑了蕴含于多视图数据中的互补信息和关系以及每个视图中不同特征之间的冗余关系,而且利用少量标签信息协同大量未标签数据一起学习,提高了特

征选择算法的性能。实验表明,SSMVFS算法能够获得很好的特征选择效果及聚类效果。

参考文献:

[1] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data[J]. Journal of Bioinformatics and Computational Biology, 2005, 3(02): 185-205.

[2] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C]// ICML. [S. l.]: Morgan Kaufmann Publishers, 1997: 412-420.

[3] 李士进,仇建斌,於慧. 基于视觉单词选择的高分辨率遥感图像飞机目标检测[J]. 数据采集与处理,2014,29(1):60-65.
Li Shijin, QiuJianbin, Yu Hui. Aircraft detection in high-resolution remote sensing imagery based on visual words selection [J]. Journal of Data Acquisition and Processing, 2014,29(1):60-65.

[4] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]// Proceedings of the Eleventh Annual Conference on Computational Learning Theory. [S. l.]: ACM, 1998: 92-100.

[5] Heckmann M, Berthommier F, Kroschel K. Noise adaptive stream weighting in audio-visual speech recognition[J]. EURASIP Journal on Applied Signal Processing, 2002, 2002(1): 1260-1273.

[6] La Cascia M, Sethi S, Sclaroff S. Combining textual and visual cues for content-based image retrieval on the world wide web [C]// Content-Based Access of Image and Video Libraries. [S. l.]: IEEE, 1998: 24-28.

[7] Wu Y, Chang E Y, Chang K C C, et al. Optimal multimodal fusion for multimedia data analysis[C]// Proceedings of the 12th Annual ACM International Conference on Multimedia. [S. l]: ACM, 2004: 572-579.

[8] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2005, 27(8): 1226-1238.

[9] He X, Cai D, Niyogi P. Laplacian score for feature selection[C]// Advances in Neural Information Processing Systems 18. [S. l]: MIT Press, 2005: 507-514.

[10] Zhao Z, Liu H. Spectral feature selection for supervised and unsupervised learning[C]// Proceedings of the 24th international conference on Machine learning. [S. l.]: ACM, 2007: 1151-1157.

[11] Cai D, Zhang C, He X. Unsupervised feature selection for multi-cluster data[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S. l.]: ACM, 2010: 333-342.

[12] Zhao Z, Wang L, Liu H. Efficient spectral feature selection with minimum redundancy[C]// Proceedings of the Twenty-Fourth Conference on Artificial Intelligence. [S. l.]: American Association for Artificial Intelligence, 2010.

[13] Jiang Y, Ren J. Eigenvalue sensitive feature selection[C]// Proceedings of the 28th International Conference on Machine Learning. [S. l.]: Association for Computing Machinery, 2011: 89-96.

[14] Nie F, Huang H, Cai X, et al. Efficient and robust feature selection via joint l2, 1-norms minimization[J]. Advances in Neural Information Processing Systems, 2010, 23: 1813-1821.

[15] Tang J, Liu H. Feature selection with linked data in social media[C]// Proceedings of the Twelfth International Conference on Data Mining. [S. l.]: Society for Industrial and Applied Mathematics, 2012: 118-128.

[16] Tang J, Hu X, Gao H, et al. Unsupervised feature selection for multi-view data in socail media[C]// SDM, SIAM International Conference on Data Mining. [S. l.]: SIAM, 2013.

[17] Wang H, Nie F, Huang H. Multi-view clustering and feature learning via structured sparsity[C]// Proceedings of the 30th International Conference on Machine Learning. [S. l.]: JMLR Org, 2013: 352-360.

[18] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006, 68(1): 49-67.

[19] Uci machine learning repository[M]. Irvine, CA: University of California, 2010.

[20] Zhao Z, Liu H. Semi-supervised feature selection via spectral analysis[C]// Proceedings of the Seventh International Conference on Data Mining. Minneapolis, Minnesota, USA: Society for Industrial and Applied Mathematics Publications, 2007: 641-646.

[21] Kumar A, Rai P, Daumé III H. Co-regularized multi-view spectral clustering[C]// Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2011: 1413-1421.

作者简介:**汪荆琪**(1990-),女,硕士研究生,研究方向:机器学习,数据挖掘,E-mail: wjingqi@mail.ustc.edu.cn;**徐林莉**(1980-),女,副教授,研究方向:机器学习与数据挖掘。