计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20190150
56(8):1677-1685,2019

样本加权的多视图聚类算法

洪 敏[1,2] 贾彩燕[1,2] 李亚芳[3] 于 剑[1,2]
[1](交通数据分析与挖掘北京市重点实验室(北京交通大学) 北京 100044)
[2](北京交通大学计算机与信息技术学院 北京 100044)
[3](北京工业大学信息学部 北京 100124)
(16120372@bjtu.edu.cn)

Sample-Weighted Multi-View Clustering

Hong Min[1,2], Jia Caiyan[1,2], Li Yafang[3], and Yu Jian[1,2]
[1](*Beijing Key Laboratory of Traffic Data Analysis and Mining* (*Beijing Jiaotong University*), *Beijing* 100044)
[2](*School of Computer and Information Technology*, *Beijing Jiaotong University*, *Beijing* 100044)
[3](*Faculty of Information Technology*, *Beijing University of Technology*, *Beijing* 100124)

Abstract In the era of big data, the ability of humans to collect, store, transmit and manage data has been increasingly improved. Various industries have accumulated a large amount of data resources, which are often multi-source and heterogeneous. How to effectively cluster these multi-source data (also known as multi-view clustering) has become one of the focuses of today's machine learning research. The existing multi-view clustering algorithms mainly pay attention to the contribution of different views and features to the cluster structure from the "global" perspective, without considering the "local" information complementary differences between different samples. Therefore, this paper proposes a new sample-weighted multi-view clustering (SWMVC). The method weights each sample with different views and adopts alternating direction method of multipliers (ADMM) to learn sample weight, which can not only learn the "local" difference of weights among multiple views in different sample points, but also reflect the "global" difference of the contribution of different views to the cluster structure, and has better flexibility. Experiments on multiple datasets show that the SWMVC method has a better clustering effect on heterogeneous view data.

Key words data mining; multi-view; cluster; K-means; sample weights

摘 要 大数据时代,人类收集、存储、传输、管理数据的能力日益提高,各行各业已经积累了大量的数据资源,这些数据常呈现出多源性和异构性.如何对这些多源数据进行有效的聚类(也称为多视图聚类)已成为当今机器学习研究关注的焦点之一.现有的多视图聚类算法主要从"全局"角度关注不同视图和特征对簇结构的贡献,没有考虑不同样本间存在的"局部"信息间的差异.因此,提出一种新的多视图样本加权聚类算法(sample-weighted multi-view clustering, SWMVC),该算法对每个样本的不同视图进行加权,采用交替方向乘子法自适应学习样本权值,不仅可以学习不同样本点间不同视图权重的"局部"差异,还可以从学习到的"局部"差异反映出不同视图对簇结构贡献的"全局"差异,具有较好的灵活性.

收稿日期:2019-03-11;**修回日期**:2019-05-16
基金项目:国家自然科学基金项目(61876016,61632004);中央高校基本科研业务费专项资金项目(2018JBZ006)
This work was supported by the National Natural Science Foundation of China (61876016, 61632004) and the Fundamental Research Funds for the Central Universities (2018JBZ006).
通信作者:贾彩燕(cyjia@bjtu.edu.cn)

多个数据集上的实验表明:SWMVC 方法在异质视图数据上具有较好的聚类效果.

关键词 数据挖掘;多视图;聚类;K-means;样本权重

中图法分类号 TP301

飞速发展的信息技术使现实世界中的数据不仅呈现出规模庞大的特性,而且多源信息采集技术和多样化的特征表示使得多视图数据在众多实际应用中越来越普遍.例如 Web 网页可以从 3 个不同的角度描述:词向量视图直观刻画了网页文本中单词的出现情况;网页中的图像提供了丰富的视觉特征视图;网页与网页之间的链接关系展示了网页内容彼此之间的相关性.多视图数据的互补性和一致性有利于从不同方位协同完成特定的机器学习任务.因此,多视图学习逐渐受到人们的关注.

多视图研究最早起源于 Yarowsky[1] 和 Blum 等人[2] 提出的消除单词歧义算法和协同训练算法. Yarowsky[1] 将文档中单词信息和单词所在文档信息定义为 2 个视图,并将不同视图加入各自的分类器进行交互学习.Blum 等人[2] 在对多视图定义分类器的同时,基于视图独立性的原则引入协同训练. Bickel 等人[3] 以协同 EM(expectation maximization)算法为基础,提出了多视图 EM 和两视图球状 K-means 算法. Cleuziou 等人[4] 提出了 CoFKM (centralized method for multiple-view clustering) 算法,通过集中式策略使多个视图的信息尽可能的一致.Sa[5] 构造了一个二视图谱聚类算法,创建一个描述 2 个视图共现信息的二部图,以最小化视图之间的不一致性.Kumar 等人[6-7] 相继提出了在多视图谱聚类加入协同训练和协同正则的方法,保持了不同视图上聚类结果的一致性.

但是,上述方法平等看待各个视图,难以体现不同视图信息对簇结构重要性的差异,因此出现了一些视图权重学习的经典算法.较早的工作是 Tzortzis 等人[8] 给出的多视图带权凸混合模型(convex mixture models, CMM).该模型对不同视图训练相应的 CMM 模型,同时为各个视图分配自适应权重,并将自学习视图权重思想推广至核 K-means,设计了多视图核函数 K-means 算法(multi-view kernel K-means, MVKKM)[9].该算法通过预先定义的核函数将原始数据映射到新的特征空间,并根据视图对簇划分结果的贡献度为每一个视图分配权重. 为了适用大规模多视图数据,Cai 等人[10] 给出了 RMKMC 模型,不仅在目标函数中新增视图权值参数,同时 $\ell_{2,1}$ 范式还增强了算法的鲁棒性.最近, Nie 等人[11-12] 从图的角度提出了 2 种适用聚类和半监督分类的多视图学习模型:AMGL 和 MLAN. AMGL[11] 凭借视图近邻图学习,构建出了多视图谱聚类权重学习框架;MLAN[12] 通过学习多图局部流形结构,不断修正簇结构矩阵以优化聚类性能.

多视图学习不仅是不同视图之间的有效性融合,而且视图特征的选择也是多视图学习的重要组成部分.Chen 等人[13] 提出了自适应双权多视图算法 (TW-K-means).该算法能同时学习视图权重和特征权重对簇结构的贡献.Xu 等人[14] 提出了具有特征选择的加权多视图聚类算法(weighted multi-view clustering with feature selection, WMCFS).Zhang 等人[15] 在多视图权重学习模型中增加了视图间的一致性约束,设计了性能良好的 TW-Co-K-means 模型.

除了以上的学习不同视图和特征对簇结构贡献差异的"全局"模型外,Li 等人以节点属性和节点的链接关系为 2 种数据源,学习了不同样本点上不同类型信息对社区结构(即簇结构)贡献的"局部"差异[16],给出了 Adapt-SA 模型,取得了不错的效果. 但该模型需要对 2 种信息下的数据进行同维变换,不但缺乏灵活性且同维变换易造成原有数据上信息的损失.因此,我们基于两视图同维变换样本局部权值学习模型 Adapt-SA,提出了一种新的多视图样本加权聚类算法(sample-weighted multi-view clustering, SWMVC).该算法不仅可以学习不同样本点对多个视图间权重的差异,具有"局部"学习能力,而且还可以从"全局"角度反映不同视图对簇结构的贡献差异.另外,该方法不仅克服了 Adapt-SA 方法同维变换的约束,而且将 Adapt-SA 的思想从 2 个视图扩展到多个视图上.

1 相关算法介绍

本节,我们介绍 4 个经典的加权多视图聚类算法: MVKKM[9], WMCFS[14], TW-Co-K-means[15] 和 Adapt-SA[16].给定多视图数据集 $X=\{\boldsymbol{X}^1,\boldsymbol{X}^2,\cdots,\boldsymbol{X}^V\}$,其中,$\boldsymbol{X}^v\in\mathbb{R}^{N\times D_v}$,即每个数据视图包括 N 个数据.$\boldsymbol{x}_i^v$ 是 $\boldsymbol{X}^v$ 的第 i 个样本.

1.1　MVKKM

MVKKM(multi-view clustering kernel K-means)[9]是一种基于核函数的多视图权值模型.相比传统的视图权值方法,该算法使用事先定义的核函数对数据进行映射,然后在多视图 K-means 中学习各个视图的重要性.其目标函数:

$$\widetilde{K}=\sum_{v=1}^{V}(\omega_v^p)\sum_{i=1}^{N}\sum_{k=1}^{K}\delta_{ik}\|\phi^{(v)}(x_i^{(v)})-m_k^{(v)}\|^2,\quad(1)$$

其中,ω_v,$\phi^{(v)}$ 和 $m_k^{(v)}$ 分别表示各个视图的权重、核函数和中心,p 是控制视图权重分布的超参数,δ_{ik} 是各个视图的共有划分结果.对于核函数,可以选用线性核、高斯径向基核等.

1.2　WMCFS

基于 MVKKM 思想,WMCFS(weighted multi-view clustering with feature selection)[14]模型在多视图 K-means 聚类过程中同时考虑不同视图和各个特征的差异,设计了目标函数:

$$\varepsilon_H=\sum_{v=1}^{V}(\omega_v)^p\sum_{i=1}^{N}\sum_{k=1}^{K}\delta_{ik}\|diag(\tau^v)(x_i^v-m_k^v)\|^2+\beta\sum_{v=1}^{V}\|\tau^v\|^2,\quad(2)$$

其中,ω_v 和 τ^v 分别表示视图权重和视图中各个特征的权重,p 和 β 是分别控制视图和特征稀疏性的参数.

1.3　TW-Co-K-means

TW-Co-K-means(two-level weighted collaborative K-means)[15]采用协同策略约束不同视图之间的一致性,进而学习视图和特征对簇结构的重要性.其目标函数:

$$J=\sum_{t=1}^{T}v^{(t)}\Big(\sum_{j\in G^{(t)}}w_j^{(t)}\sum_{k=1}^{K}\sum_{i=1}^{N}u_{ik}^{(t)}\|x_{ij}^{(t)}-m_{kj}^{(t)}\|^2\Big)+\frac{\eta}{(T-1)}\Delta+\alpha\sum_{t=1}^{T}\sum_{j\in G^{(t)}}w_j^{(t)}\log w_j^{(t)}+\beta\sum_{t=1}^{T}v^{(t)}\log v^{(t)},\quad(3)$$

$$\Delta=\sum_{t=1}^{T}\sum_{t'\neq t}\Big(\sum_{i=1}^{N}\sum_{k=1}^{K}|u_{ik}^{(t')}-u_{ik}^{(t)}|(v^{(t)}\xi^{(t)}-v^{(t')}\xi^{(t')})\Big),\quad(4)$$

$$\xi^{(t)}=\sum_{j\in G^{(t)}}w_j^{(t)}\|x_{ij}^{(t)}-m_{kj}^{(t)}\|,\quad(5)$$

其中,$v^{(t)}$,$w^{(t)}$,$u_{ik}^{(t)}$ 和 $m^{(t)}$ 分别表示各个视图权重、特征权重、各个视图的指派和中心,Δ 衡量不同视图之间的不一致性.α,β 和 η 分别控制 $v^{(t)}$,$w^{(t)}$ 的分布,以及 Δ 的影响.根据各个视图的指派,融合视图权值,便可获得最终的共有指派.

1.4　Adapt-SA

Adapt-SA(adaptive fusion of structural and attribute information)[16]是面向图聚类问题提出的一种学习样本上不同种类信息(链接信息和节点属性)对节点簇结构重要性差异的 K-means 型算法.该算法首先将 2 类信息映射到同维空间上,再进行加权融合.因此,融合后的表示具有统一的簇中心.其目标函数:

$$F=\sum_{i=1}^{n}\sum_{j=1}^{K}\boldsymbol{U}_{ij}\|\alpha_i\bar{\boldsymbol{A}}_i+(1-\alpha_i)\bar{\boldsymbol{C}}_i-z_j\|_2^2,\quad(6)$$

其中,$\boldsymbol{U}$ 和 z 分别表示划分矩阵和融合中心.$\bar{\boldsymbol{A}}_i$ 是经过 Signal 相似度[17]变换的链接向量,$\bar{\boldsymbol{C}}_i$ 是使用余弦相似度变换的属性向量.

2　基于样本加权的多视图聚类算法

Adapt-SA 虽然可以学习不同样本间 2 类信息重要性的差异,但该方法在进行信息融合时,需要对原空间数据进行同维变换,可能会造成一定的信息损失.并且,同维变化也会带来算法复杂性的增加.在实际应用中,可能某个视图的维度适当,最宜在原空间进行簇结构学习,使得 Adapt-SA 缺乏灵活性.因此,本部分我们给出了一种普适性更强且能够学习不同样本点权值的多视图 K-means 算法,以弥补 Adapt-SA 算法的不足.

2.1　基本定义

为方便介绍所提出的算法模型,本节所用符号的具体含义如表 1 所示:

Table 1　The Symbol Meaning

表 1　符号含义

Symbol	Meaning
V	Number of views
N	Number of samples in each view
K	Number of clusters
δ_{ik}	Cluster assignment showing whether the i-th sample belongs to the k-th cluster
x_i^v	The i-th sample of the v-th view
m_k^v	The cluster center of the k-th cluster in the v-th view
α_i^v	Weight for the i-th sample of the v-th view
λ	Parameter controlling the sparisity of sample weight

2.2　SWMVC 算法

受 Adapt-SA 思想的启发,并结合多视图 K-means 算法,为了描述多视图聚类过程中每个样本

点的不同视图对簇结构的贡献,SWMVC 算法优化模型:

$$\varepsilon=\sum_{v=1}^{V}\sum_{i=1}^{N}\alpha_i^v\sum_{k=1}^{K}\delta_{ik}\|x_i^v-m_k^v\|^2+\lambda\sum_{v=1}^{V}\sum_{i=1}^{N}\alpha_i^v\log\alpha_i^v$$

$$\text{s.t.}\quad \sum_{k=1}^{K}\delta_{ik}=1,\delta_{ik}\in\{0,1\},1\leqslant i\leqslant N;\quad \sum_{v=1,i\in N}^{V}\alpha_i^v=1,\ 0\leqslant\alpha_i^v\leqslant 1,\quad 1\leqslant i\leqslant N,v=1,2,\cdots,V. \tag{7}$$

SWMVC 的目标函数由 2 部分组成:首先是在多视图 K-means 的基础上增加了样本权参数 α_i^v,用于分析样本 x_i 对第 v 个视图的重要性.同时,为了使模型能够在所有可能解的分布中得到 α_i^v 的最优分布,增加了基于信息熵的加权项,并通过 λ 控制每个数据点权值的大小,有利于调节样本权 α_i^v 的稀疏性.

因此,SWMVC 模型具备 3 方面的特点:1)能使每一个视图拥有独立的簇中心;2)每个样本点对多个视图拥有自己对簇结构贡献的权重影响,具有局部学习能力,同时通过样本点在各视图上的不同重要程度,易于统计各视图对簇结构的全局重要性;3)不但可以在多个视图的原空间进行操作,也可以在变换后的视图空间(如使用 Adapt-SA 模型中的同维变换策略)进行学习,灵活性高.

2.3 SWMVC 模型求解

求解 SWMVC 目标函数式(7)时使用交替方向乘子法.对于 3 个未知项 δ_{ik},α_i^v 和 m_k^v,在固定 2 个未知项的同时使用梯度下降法进行求解.具体求解流程为:

1) 固定所有视图节点权重 α_i^v 和视图类中心 m_k^v,更新视图的公共隶属度元素 δ_{ik}.

在求解时,针对每一个类寻求所有视图到每个类中心的带权距离最小值.即按照式(8)进行计算,之后根据得到的最小值结果对每个数据点进行指派.

$$\arg\min_{\delta_{ik}}\sum_{v=1}^{V}\sum_{k=1}^{K}\alpha_i^v\|x_i^v-m_k^v\|^2. \tag{8}$$

2) 固定视图共有节点隶属度 δ_{ik} 和视图类中心 m_k^v,更新每个视图的节点权重 α_i^v.

对每个视图而言,可看成是求解 K 个彼此独立的子问题,视图 v 的节点 x_i^v 隶属于第 k 个类时,隶属度 $\delta_{ik}=1$.为了最小化目标函数,利用对视图节点权重在多个视图上施加的约束,采用拉格朗日乘子法求解,加入拉格朗日乘子 β 后,目标函数:

$$L(\alpha_i^v,\beta)=\sum_{v=1}^{V}\sum_{i=1}^{N}\alpha_i^v\sum_{k=1}^{K}\delta_{ik}\|x_i^v-m_k^v\|^2+\lambda\sum_{v=1}^{V}\sum_{i=1}^{N}\alpha_i^v\log\alpha_i^v+\beta\Big(\sum_{v=1,i\in N}^{V}\alpha_i^v-1\Big). \tag{9}$$

通过对式(9)中 α_i^v 求导,得到:

$$\alpha_i^v=\exp\left(-\frac{\|x_i^v-m_k^v\|^2-\lambda-\beta}{\lambda}\right), \tag{10}$$

利用对 α_i^v 的现有约束条件,可以推出最终的更新公式

$$\alpha_i^v=\frac{\exp\left(-\dfrac{\|x_i^v-m_k^v\|^2}{\lambda}\right)}{\sum\limits_{v=1}^{V}\exp\left(-\dfrac{\|x_i^v-m_k^v\|^2}{\lambda}\right)}, \tag{11}$$

3) 固定视图节点权重 α_i^v 和视图共有节点隶属度 δ_{ik},更新每个视图的类中心 m_k^v:

$$m_k^v=\frac{\sum\limits_{i=1}^{N}\delta_{ik}\alpha_i^v x_i^v}{\sum\limits_{i=1}^{N}\alpha_i^v\delta_{ik}}. \tag{12}$$

根据上述模型求解过程知,本节构建的基于样本加权的多视图聚类算法(SWMVC)的详细步骤如算法 1 所示:

算法 1. SWMVC.

输入:多视图数据集 $X=\{\boldsymbol{X}^1,\boldsymbol{X}^2,\cdots,\boldsymbol{X}^V\}$、类个数 K、参数 λ、最大迭代次数 $t_{\max}$;

输出:X 的共有隶属度 δ、样本权集 α^v.

① 初始化:每个视图随机选取 K 个中心,$\alpha_i^v=1/V$,$t=0$;

② 根据式(8)更新 δ;

③ 根据式(11)更新 α_i^v;

④ 根据式(12)更新 m_k^v;

⑤ $t=t+1$;

⑥ 如果 $t>t_{\max}$,停止算法,返回 δ 和 α^v.

3 实验设计与结果分析

3.1 数据集描述

为了评估本文提出的基于样本加权的多视图聚类模型(SWMVC)算法的效果,对不同的多视图权值聚类模型进行实验和对比分析.我们选取了视图之间是异质特性的 6 个数据集,分别是 WebKB 中的 Cornell,Texas,Washington,Wisconsin 网络/文本数据集和图像/文本数据集 Wiki 和 VOC 以及

视图具有同质特性的 2 个数据集:Handwritten numerals 和 Caltech101-7.

1) WebKB 数据集

Cornell,Texas,Washington 和 Wisconsin 是 WebKB 网页数据集的子集,分别包含 195,187,230,265 个样本.每个样本通过引用关系视图和文本内容属性视图进行描述,对应的成对视图维数分别是 195 维和 1 703 维、187 维和 1 703 维、230 维和 1 703 维、265 维和 1 626 维.该数据集涉及到 5 个类别课程、学院、学生、工程和员工.

2) Wiki 数据集

Wiki 是从维基百科专题文章中提取的数据集,常用于跨模态检索中.该数据集由 2 173 个训练样本和 693 个测试样本组成的图片-文本对数据,共有 10 个类.其中,每一张图片使用 128 维 SIFT 特征向量视图表示,并含有相关图片的 10 维文本 LDA 主题描述向量视图.

3) Pascal Visual Object Classes 2007 数据集

Pascal Visual Object Classes 2007(VOC)是一个自然图像数据集.选用其中的 2 808 张图片,每一张图像有 512 维的 GIST 特征和 399 维的 TF-IDF (term frequency-inverse document frequency)文本特征.涉及到 20 个类 aeroplane,bicycle,bird,boat,bottle,bus,car,cat,chair,cow,dining table,dog,horse,motorbike,person,potted plant,sheep,sofa,train,tv/monitor.

4) Handwritten numerals 数据集

Handwritten numerals(HW)是一个含有 10 个类的 2 000 个手写体数字数据集.本实验选取其中的 5 个视图数据,分别是维数 76 维 FOU 特征、216 维 FAC 特征、64 维 KAR 特征、240 维 PIX 特征和 47 维 ZER 特征.

5) Caltech101-7 数据集

Caltech101-7 是一个对象识别数据集.包括 1 474 张图片,被划分为 7 个类,分别是 Face,Motorbike,DollaBill,Garfield,Snoopy,Stop-Sign,Windsor-Chair,含有图像的 Gabor,Wavelet,CENTRIST,HOG,GIST 和 LBP 6 个特征,并将其作为 Caltech101-7 的 6 个视图,对应视图的特征维数分别为 48 维、40 维、254 维、1984 维、512 维和 928 维.

3.2　实验方案

为了全面评估本文提出的基于样本加权的多视图聚类模型的性能,在实验部分中对比研究了多视图聚类领域中的多个代表性算法:

1) WMCFS[14] 是一种具有特征选择功能的多视图聚类算法.该方法的 2 个参数 p 和 β 利用网格贪心搜索方法在 3.1 节介绍的 8 个数据集上依次设置为{5,0.5},{30,0.0005},{20,0.5},{30,0.0002},{10,0.5},{20,0.0005},{10,0.1}和{5,0.0005}.

2) MLAN[12] 是一种面向图聚类的多视图聚类算法.与传统图聚类不同的是,它考虑局部流形结构在每一轮迭代不断修改相似性矩阵,直至获得最优矩阵.该方法可以自动学习各个视图的权重系数.其隐含参数近邻数 $k=9$.

3) AMGL[11] 该方法假设所有图共享潜在簇结构,是一种无参数的多图学习框架,即在学习时能够为各个图分配合适的权值,可以应用于多视图聚类和半监督分类任务.其隐含参数近邻数 $k=5$.

4) TW-Co-K-means[15] 是一种融合视图和特征的多视图算法.该方法全面考虑不同视图的互补性和一致性,以协同方式挖掘不同视图间的共享信息,同时考量了每个视图的多样性特点.参数 η,α 和 β 在各个数据集分别采用网格贪心搜索方法设置为{0.8,10,60},{0.8,30,60},{0.1,20,80},{0.9,50,80},{0.3,30,8},{0.01,80,50},{0.6,10,50}和{0.2,30,8}.

5) SWMVC 是本文提出的基于样本加权的多视图聚类算法.模型中含有样本权稀疏系数,参数 λ 经过网格贪心搜索方法在各数据集的值分别设为 100,75,35,60,0.5,1,10,45.

3.3　实验结果分析

本节首先展示不同方法在各种多视图数据集上的聚类效果,然后展示 WebKB,Wiki 和 VOC 所有样本在给定视图的权重分布,最后研究了参数 λ 的敏感性.

1) 多视图聚类效果对比

表 2～4 分别展示了各多视图聚类方法在 8 个数据集上的聚类评价指标 *ACC*,*NMI* 和 *F*-measure.其中,各个算法精度选用 20 次运行结束的最大值,各多视图聚类算法在同一数据集上的最好结果用黑体表示.

由表 2～4 可以看出:

① 相比传统多视图聚类方法,本文提出的基于样本加权的多视图聚类算法(SWMVC)在 Cornell,Texas,Washington,Wisconsin,Wiki 和 VOC 数据集上有较好的性能提升.具体来说,SWMVC 的 *ACC* 和 *F*-meansure 指标值比传统的多视图聚类算法 TW-Co-K-means 在前 5 个数据上分别提高了约

0.03 和 0.02,在 VOC 上取得了次优结果.在 *NMI* 指标上,SWMVC 在异质数据 Texas,Washington,Wisconsin 和 VOC 获得了最高结果,在 Cornell 和 Wiki 以 0.2082 和 0.5388 取得了第 2 高结果.究其原因,我们发现:WebKB 的 4 个数据集由链接结构视图和文本内容视图 2 种异构视图描述,能够形成较强的互补关系;Wiki 和 VOC 由图像特征视图和文本特征视图 2 种异构视图组成,同样具有较好的互补性,且文本特征视图解释力强于图像特征描述子.

② SWMVC 在手写数据集 HW 和图像数据集 Caltech101-7 表现欠佳,原因在于尽管这 2 个数据集的多个视图从各个方面进行了描述,但是视图之间

Table 2 *ACC* of Different Clustering Algorithms on Multi-View Datasets

表 2 不同聚类算法在多视图数据集上的 *ACC* 值

Dataset	WMCFS	MLAN	AMGL	TW-Co-*K*-means	SWMVC
Cornell	0.4872	0.4308	0.4667	0.5333	**0.5744**
Texas	0.6150	0.6043	0.5346	0.6364	**0.6738**
Washington	0.6435	0.6565	0.5783	0.6696	**0.7130**
Wisconsin	0.5396	0.4981	0.5245	0.6792	**0.7094**
Wiki	0.5806	0.1800	0.5464	0.5429	**0.5998**
VOC	0.6742	0.3712	0.1524	**0.7130**	0.6802
HW	0.8025	**0.9730**	0.8530	0.8960	0.8135
Caltech101-7	0.4640	**0.7802**	0.6771	0.5360	0.5556

Note: The best results are highlighted in bold.

Table 3 *NMI* of Different Clustering Algorithms on Multi-View Datasets

表 3 不同聚类算法在多视图数据集上的 *NMI* 值

Dataset	WMCFS	MLAN	AMGL	TW-Co-*K*-means	SWMVC
Cornell	0.1446	**0.2204**	0.0621	0.1613	0.2082
Texas	0.2320	0.2520	0.0449	0.2432	**0.2716**
Washington	0.3454	0.3571	0.1537	0.2310	**0.3585**
Wisconsin	0.2718	0.2773	0.0935	0.3351	**0.3876**
Wiki	**0.5456**	0.0393	0.5193	0.5370	0.5388
VOC	0.6412	0.2571	0.0172	0.6558	**0.6618**
HW	0.7953	**0.9390**	0.8843	0.8217	0.7956
Caltech101-7	0.4539	**0.6304**	0.6075	0.4659	0.4427

Note: The best results are highlighted in bold.

Table 4 *F*-measure of Different Clustering Algorithms on Multi-View Datasets

表 4 不同聚类算法在多视图数据集上的 *F* 值

Dataset	WMCFS	MLAN	AMGL	TW-Co-*K*-means	SWMVC
Cornell	0.3789	0.3916	0.4403	0.4761	**0.5147**
Texas	0.5353	0.5540	0.5042	0.5956	**0.6146**
Washington	0.5677	0.5947	0.5110	0.5890	**0.6372**
Wisconsin	0.4818	0.4709	0.4797	0.6068	**0.6120**
Wiki	0.5056	0.1800	0.4208	0.4604	**0.5108**
VOC	0.6005	0.2571	0.1520	**0.6171**	0.6060
HW	0.8370	**0.9466**	0.8448	0.8229	0.7680
Caltech101-7	0.5043	**0.7365**	0.7102	0.5518	0.5452

Note: The best results are highlighted in bold.

表现出同质特性,视图间的信息互补性相对较弱,且对簇结构的刻画能力在局部样本上的差异性不明显.

③ 根据实验结果可以发现:MLAN 模型在 HW 和 Caltech101-7 效果显著.究其原因,MLAN 学习了数据的局部流形结构,多个同质视图上局部流形结构的整合较好地反映出了全局的簇结构信息,是目前较为突出的模型.

④ 综合 WMCFS 和 TW-Co-K-means 多视图特征算法在上述 8 个数据集的实验结果可见:良好的特征选择机制可以提高多视图聚类算法的性能.

2）节点权重分析

为了进一步验证本文提出的 SWMVC 算法,我们在图 1 中展示了该方法在 6 个异构视图数据集 Cornell,Texas,Washington,Wisconsin,Wiki 和 VOC

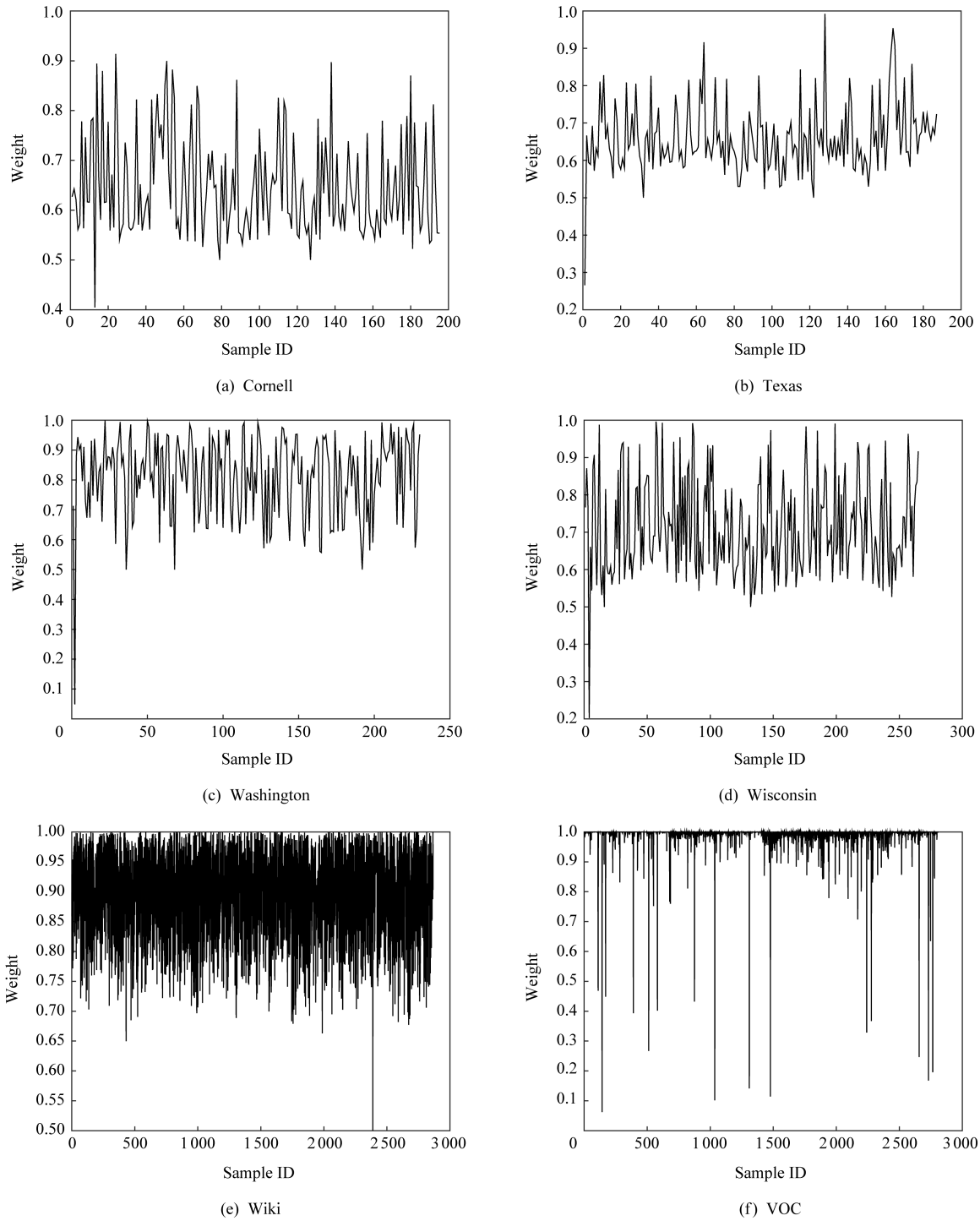

Fig. 1　Weights of samples in a given view

图 1　各个样本在给定视图上的权重

上学习到的样本权重.其中,横坐标表示样本编号,纵坐标表示各样本在 WebKB 四个数据链接视图上的权重及 Wiki,VOC 数据集的各样本点在文本视图上的权重.

由图 1 可以看出:整体上 WebKB 四个数据集的样本在链接视图所占权重较大,即拓扑结构更能刻画数据特性;Wiki 和 VOC 各样本点在文本视图权重大于图像特征视图.同时视图包含重要度样本越多,说明该视图该视图越重要,实现了对数据"全局"信息重要性的学习.

对于同质的多视图数据集 HW 和 Caltech101-7,SWMVC 算法学习到的各样本点的权重差异性小.由于这 2 个数据集视图多,考虑到空间限制,只简单描述如下:HW 数据集所有样本在 FAC,FOU,KAR 和 ZER 视图权重基本分布在 0.21 左右,在 PIX 特征视图权值约为 0.16;Caltech101-7 所有样本点在前 3 个视图权重约为 0.20,第 4 个视图权值约是 0.06,最后 2 个权值为 0.17 左右.从而进一步可见,我们的方法更适于差异性较大的异构互补视图的簇结构学习问题.

3.3 参数敏感性分析

SWMVC 样本重要度正则项参数 λ 对数据集有不同程度的影响.具体而言:参数 λ 对异质数据集的样本在视图权重分布上有明显影响.以 Wiki 数据集为例,图 2 显示了 λ 为 0.5 和 50 时所有样本在文本视图上的权重分布情况.可以发现:λ 取值较小时,文本特征权重相对较大,聚类结果也较好,印证了 λ 可以调节样本权 α_i^v 的稀疏性.

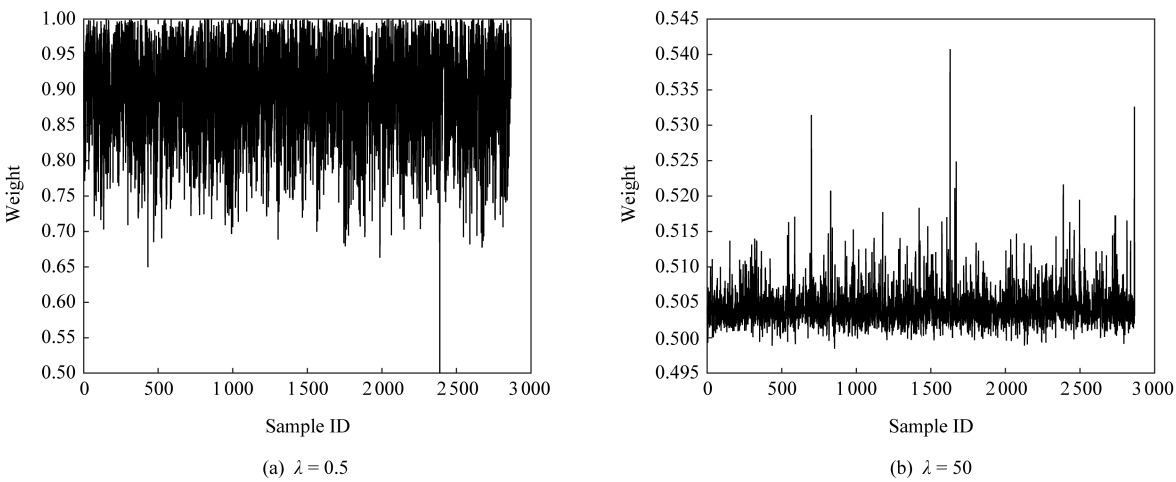

Fig. 2 The influence of λ on sample weights

图 2 参数 λ 对样本权重分布的影响

4 总 结

本文主要提出了一种基于样本加权的多视图聚类算法(SWMVC).为了更好地学习多视图蕴含的信息,本文受 Adapt-SA 算法启发,引进多视图学习思想,从不同视图角度学习样本对簇结构的贡献度.这种多视图样本加权学习机制不仅体现了样本差异性,而且还刻画了视图的重要度,实现了对数据"局部"和"全局"性质的学习.与此同时,SWMVC 能够保留数据原始信息,有助于降低模型复杂度.在 6 个异质视图数据集和 2 个同质视图数据集上的大量实验对比表明:本文提出的模型在异质视图数据上具有不错的聚类效果.

参 考 文 献

[1] Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods [C] //Proc of the 33rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics. Stroudsburg: ACL, 1995: 189-196

[2] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training [C] //Proc of the 11th Annual Conf on Computational Learning Theory. New York: ACM, 1998: 92-100

[3] Bickel S, Scheffer T. Multi-view clustering [C] //Proc of the 4th IEEE Int Conf on Data Mining. Los Alamitos: IEEE Communicatoins Society, 2004: 19-26

[4] Cleuziou G, Exbrayat M, Martin L, et al. CoFKM: A centralized method for multiple-view clustering [C] //Proc of

the 9th IEEE Int Conf on Data Mining. Los Alamitos: IEEE Communicatoins Society, 2009: 752-757

[5] De Sa V R. Spectral clustering with two views [C] //Proc of the Workshop on Learning with Multiple Views. New York: ACM, 2005: 20-27

[6] Kumar A, Daumé H. A co-training approach for multi-view spectral clustering [C] //Proc of the 28th Int Conf on Machine Learning. New York: ACM, 2011: 393-400

[7] Kumar A, Rai P, Daume H. Co-regularized multi-view spectral clustering [C] //Proc of the 24th Int Conf on Neural Information Processing System. Vancouver: Curran Associates, 2011: 1413-1421

[8] Tzortzis G F, Likas A C. Multiple view clustering using a weighted combination of exemplar-based mixture models [J]. IEEE Transactions on Neural Networks, 2010, 21(12): 1925-1938

[9] Tzortzis G, Likas A. Kernel-Based Weighted Multi-view Clustering [C] //Proc of the 12th IEEE Int Conf on Data Mining. Los Alamitos: IEEE Communicatoins Society, 2012: 675-684

[10] Cai Xiao, Nie Feiping, Huang Heng. Multi-view *K*-means clustering on big data [C] //Proc of the 23rd Int Joint Conf on Artificial Intelligence. San Francisco: Morgan Kaufmann, 2013: 2598-2604

[11] Nie Feiping, Cai Guohao, Li Jing, et al. Auto-weighted multi-view learning for image clustering and semi-supervised classification [J]. IEEE Transactions on Image Processing, 2018, 27(3): 1501-1511

[12] Nie Feiping, Cai Guohao, Li Xuelong, et al. Multi-View Clustering and Semi-Supervised Classification with Adaptive Neighbours [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2017: 2408-2414

[13] Chen Xiaojun, Xu Xiaofei, Huang Joshuazhexue, et al. TW-*K*-means: Automated two-level variable weighting clustering algorithm for multiview data [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(4): 932-944

[14] Xu Yumeng, Wang Changdong, Lai Jianhuang. Weighted multi-view clustering with feature selection [J]. Pattern Recognition, 2016, 53: 25-35

[15] Zhang Guangyu, Wang Changdong, Huang Dong, et al. TW-Co-*K*-means: two-level weighted collaborative *K*-means for multi-view clustering [J]. Knowledge-Based Systems, 2018, 150: 127-138

[16] Li Yafang, Jia Caiyan, Kong Xiangnan, et al. Locally weighted Fusion of structural and attribute information in graph clustering [J]. IEEE Transactions on Cybernetics, 2017, 49(1): 247-260

[17] Hu Yanqing, Li Menghui, Zhang Peng, et al. Community detection by signaling on complex networks [J]. Physical Review E, 2008, 78(1): 139-143

Hong Min, born in 1993. Master in Beijing Jiaotong University. Her main research interest is multi-view clustering.

Jia Caiyan, born in 1976. Professor and PhD supervisor in Beijing Jiaotong University. Her main research interests include data mining, complex network analysis, text mining, bioinformatics, text mining and etc.

Li Yafang, born in 1988. Postdoctoral fellow in the Faculty of Information Technology, Beijing University of Technology. Her main research interests include data mining and complex network analysis, etc.

Yu Jian, born in 1969. Professor and PhD supervisor in Beijing Jiaotong University. Senior member of CCF. His main research interests include machine learning, data mining and image segmentation and computational intelligence.