

优化加权多视角 K-means 聚类算法

贺艳芳¹, 梁书田²

(1. 广东理工学院 信息工程学院, 广东 肇庆 526100;

2. 河南理工大学 电气工程与自动化学院, 河南 焦作 454000)

摘 要: 现存的多视角聚类算法能够充分利用多个视角的信息进行聚类, 因而其聚类效果较单视角聚类算法更优, 但是绝大多数多视角聚类算法在聚类过程中为各个视角赋予了同等的权重值, 这对于划分不明确的视角, 会严重影响聚类的最终结果。目前的加权 K-means 聚类算法在面对多视角聚类任务时, 能解决上述权重的取值分配问题, 但其权重在迭代过程中会出现除以零错误, 造成相关视角的丢失。针对这个问题, 提出了一种优化加权多视角 K-means 聚类算法 (MKSC)。该算法给每个视角分配权重, 利用加权策略有效地控制各个视角的重要程度, 通过引入常数对每个视角的权重进行优化, 使用 K-means 进行聚类。通过基于人工数据集和真实数据集的实验对该算法进行验证, 实验结果表明该算法较已有的多视角聚类技术具有更好的聚类性能。

关键词: 加权; 优化; 多视角; 聚类; K-means

中图分类号: TP181

文献标识码: A

文章编号: 1673-629X(2019)03-0081-04

doi: 10.3969/j.issn.1673-629X.2019.03.017

Optimizing Weighted Multi-view K-means Clustering Algorithm

HE Yan-fang¹, LIANG Shu-tian²

(1. School of Information Engineering, Guangdong Polytechnic College, Zhaoqing 526100, China;

2. School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo 454000, China)

Abstract: The existing multi-view clustering algorithm can make full use of multi-view information to cluster, so its effect is better than that of the single view clustering algorithm. However, in the clustering process, most of the multi-view clustering algorithms assign the same weight values for each view, which will seriously affect the final result of clustering. The current weighted K-means clustering algorithm can solve the above problem of weight assigning for the multi-view clustering tasks, but its weight will be divided by zero in the iteration process, which leads to the loss of related perspectives. For this, we propose an optimizing weighted multi-view K-means clustering algorithm (MKSC) which assigns weight for each view and uses the weighted strategy to effectively determine the importance of the various perspectives, optimizing the weight of each view by introducing a constant and with K-means to cluster. The algorithm is verified by experiments based on artificial data set and real dataset, results of which have shown that it has better clustering performance than the existing multi view clustering technology.

Key words: weighted; optimization; multi-view; clustering; K-means

1 概述

传统的聚类算法根据数据集中存在的特征将未知的数据样本进行划分, 它根据这些特征按照某种相似性度量, 使同一类的数据集具有相似性, 而不同类的数据集尽可能不相似。目前较为传统的聚类算法包括基于划分的方法^[1]、基于层次的方法^[2]、基于网格的方法^[3]和基于密度的方法^[4]等。这些常见的聚类算法均是围绕单一视角的聚类分析。然而当前信息技术的发展越来越快, 人们在现实世界会遇到越来越多具有重要意义的多特征数据, 即又称为多视角数据。多视角数据存在于社会、经济和科学等方面。例如在医学上, 通过不同的视角来描述红核细胞的密度、颜色、纹理、几何特征等不同的特征, 其中每个视角表示数据集的一种不同的度量值。多视角聚类通过分析同一数据簇的不同特征, 利用特征之间的相似性成分, 协调处理这些关系, 让多视角中的多特征形成互补, 得到尽可能一

展越来越快, 人们在现实世界会遇到越来越多具有重要意义的多特征数据, 即又称为多视角数据。多视角数据存在于社会、经济和科学等方面。例如在医学上, 通过不同的视角来描述红核细胞的密度、颜色、纹理、几何特征等不同的特征, 其中每个视角表示数据集的一种不同的度量值。多视角聚类通过分析同一数据簇的不同特征, 利用特征之间的相似性成分, 协调处理这些关系, 让多视角中的多特征形成互补, 得到尽可能一

收稿日期: 2018-03-07

修回日期: 2018-07-17

网络出版时间: 2018-12-19

基金项目: 河南省重点科技公关项目(142102210231); 广东理工学院校级项目(GKJ2017016)

作者简介: 贺艳芳(1988-), 女, 硕士, 研究方向为数据挖掘、人工智能和机器学习等。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20181219.1510.006.html>

致的聚类结果^[5]。

经过对传统聚类分析方法的研究,挖掘出更有效的多视角聚类技术,该技术在聚类过程中使得具有多特征的多视角数据在聚类过程中协同学习,解决了复杂数据多特征问题。传统聚类算法可能仅仅能处理复杂数据的一个特征,而早期的多视角聚类技术是考虑数据的每个视角,将每个视角作为独立的聚类任务进行处理,在得到每个视角对应的聚类结果后,再利用集成学习机制选择一个合适的集成学习策略将多个视角结果进行集成,进而得到最终的聚类结果^[6-7]。

当前,多视角聚类技术得到了迅速发展,越来越多的人对该技术感兴趣,并建立了各种数学模型解决当前面临的问题。例如,Wang等^[8]利用数据间的联系进行聚类,增加了各个数据属性的关联性;Bickel等^[9]提出的多视角聚类算法,将每个视角作为独立的数据集进行K-means聚类,再将每个视角的聚类结果提供给其他视角使用,完成多视角聚类;Bickel等^[10]提出的基于EM算法适用于多视角应用场景的协同聚类算法Co-EM,文献[11]将高维数据映射到不同低维空间,在低维空间对这些数据进行多视角聚类;Cleuziou等^[12]等提出了协同多视角模糊聚类算法Co-FKM。

通过研究多视角聚类算法,学习多视角算法的各种模型,发现多视角学习是利用数据集中的各种特征之间的联系,充分利用各构成视角的差异性和关联性,使最终的学习结果趋于一致。多视角学习模型如图1所示。

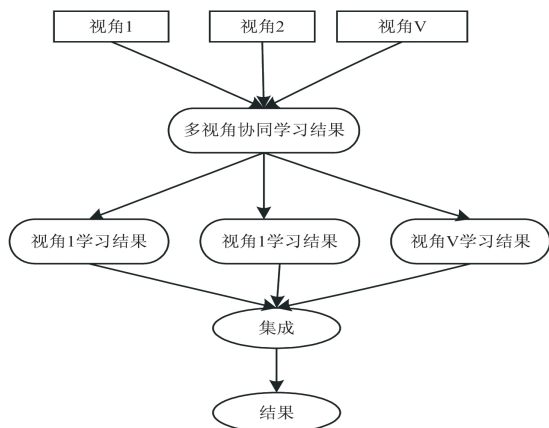


图1 多视角学习模型

该模型的多视角聚类算法同时把每个视角的聚类结果融合在一起,在互补性和一致性原则下让它们相互作用,实现各个视角间的协同学习,提高聚类的精确度,使得多视角聚类算法优于传统的单视角聚类,具有更有效的聚类性能。

在上述研究基础上,文中提出了一种多视角K-means聚类算法,利用权重值分配各个视角的重要程度,引入常数确保权重值不为零,从而确保受噪声干扰

视角或异常视角不被忽视,所有视角协同学习,从而得到好的聚类结果。

2 单视角聚类算法

给定数据样本(N 为样本总量, D 为样本维数), z 为样本的类中心,目标函数为:

$$J_M = \sum_{i=1}^C \sum_{j=1}^N \delta_{ij} \|x_j - z_i\|^2 \quad (1)$$

其中, z_i 为第 i 类的中心点; ε_0 为一个标量,当 ε_0 属于第 j 个样本点, ε_0 等于1,否则等于0。根据拉格朗日求解法得到的聚类中心迭代表达式为:

$$z_i = \frac{\sum_{j=1}^N \delta_{ij} x_j}{\sum_{j=1}^N \delta_{ij}} \quad (2)$$

K-means算法的思想是首先从一簇数据集中随机选取若干个合适的对象作为初始簇类中心,其次让其他对象根据它们与簇类中心的距离加入到该类,让一个类中的所有对象具有相似性,最后再根据距离重新计算每个簇的聚类中心,不断重复这一过程直到收敛为止。

3 多视角K-means聚类算法

为了把具有多特征的数据用K-means表示,给每个视角的聚类分配一个权重向量 W ,满足以下条件:

$$\sum_{v=1}^V w_v = 1, 0 \leq w_v \leq 1, 1 \leq v \leq V \quad (3)$$

给出的多特征数据集 X 有 N 个样本和 V 个视角, $X = \{x_i\}_{i=1}^N$,其中 $x_i = \{x_i^{(v)}\}_{v=1}^V$, $x_i \in R^{d^{(v)}}$ 是样本 x_i 的视角向量,给每个视角加权。多视角K-means的目标函数定义为:

$$\sum_{v=1}^V \sum_{i=1}^C \sum_{j=1}^N w_v^p \delta_{ij} (x_j^{(v)} - z_i^{(v)})^2 \quad (4)$$

其中, $z_i^{(v)} = (z_{i1}^{(v)}, z_{i2}^{(v)}, \dots, z_{iN}^{(v)})$ 为第 i 个聚类中心; C 为聚类的分类数; N 为数据总数; ε_0 为一个标量,当 ε_0 属于第 k 个聚类时, ε_0 等于1,否则等于0; w_v^p 为第 v 视角的权重向量, p 为权重指数。

给每个视角分配一个权重值,用来衡量视角的重要程度。当某个视角的数据较分散时,某个视角受噪声干扰较大时,该视角容易被忽略,造成该视角的权重为0。文中通过引入了常数 ε_0 ,构造新的目标函数,如下:

$$J_H = \sum_{v=1}^V \sum_{i=1}^C \sum_{j=1}^N w_v^p \delta_{ij} (x_j^{(v)} - z_i^{(v)})^2 + \varepsilon_0 \sum_{v=1}^V w_v^p \quad (5)$$

利用拉格朗日乘子最优化方法最小化式5,得到了MKSC算法聚类中心 v_k ,权重向量 w_v 的迭代表达式为:

$$z_i^{(v)} = \frac{\sum_{j=1}^N \delta_{ij} x_j^{(v)}}{\sum_{j=1}^N \delta_{ij}} \quad (6)$$

$$w_v = \frac{1}{\sum_{v=1}^V \left[\frac{\sum_{j=1}^N \delta_{ij} (x_{jv} - z_{iv})^2 + \varepsilon_0}{\sum_{j=1}^N \delta_{ij} (x_{jv} - z_{iv})^2 + \varepsilon_0} \right]^{1/(p-1)}} \quad (7)$$

MKSC 的算法流程如下:

输入: 多视角样本集 $\text{view} = \{\text{view}_1, \text{view}_2, \dots, \text{view}_v\}$ (共 v 个视角), 而任意视角对应的样本集为 $\text{view}_k = \{x_1, x_2, \dots, x_N\}$, 聚类数 $C (2 \leq C \leq N)$, 权重指数 p , 协同学习参数 ε_0 , 迭代阈值 τ 。

输出: 划分聚类 δ , 聚类中心矩阵 Z , 权重矩阵 W

Step1: 在数据集中随机选择 C 个样本作为初始聚类中心, 权重值 w_v 初始值为 $w_v^p = 1/V$;

Step2: 通过式 6 更新计算聚类中心矩阵 Z ;

Step3: 通过式 7 更新计算权重值矩阵 W ;

Step4: 通过式 5 更新计算目标函数 J_H ;

Step5: 如果 $\|J^{k+1} - J^k\| < \tau$, 则算法结束, 跳出循环, 返回 Step2。

4 实验结果与分析

4.1 实验结果的评价标准

为了评价 MKSC 算法的性能, 应用 normalized mutual information (NMI)^[12] 和 茆氏指标 rand index^[13] 作为评价聚类质量的测量标准。它们的取值范围在 0 到 1 之间。值越高, 算法中得到的聚类结果越接近真实的聚类结果。

$$\text{NMI} = \frac{2 \sum_{i=1}^C \sum_{h=1}^M \frac{n_{ih}}{N} \log \frac{n_{ih} N}{\sum_{i=1}^C n_{ih} \sum_{h=1}^M n_{ih}}}{H(\pi) + H(\zeta)} \quad (8)$$

其中, N 为聚类的样本数量; C 为聚类的类别数; M 为真实的聚类数目, n_{ih} 为聚类 i 来自真实类 h 的数据点数目; $H(\pi) = -\sum_{i=1}^C (n_i/N) \log(n_i/N)$ 是聚类的

熵; $H(\zeta) = -\sum_{i=1}^M (n^i/N) \log(n^i/N)$ 是真实类的熵。

(2) 茆氏指标 (rand index, RI)。

$$\text{RI} = 2(d_{00} + d_{11}) / (N(N-1)) \quad (9)$$

其中, d_{00} 为数据点具有不同的类标签并且属于不同类的配对数目; d_{11} 为数据点具有相同的类标签并且属于同一类的配对点数目; N 为数据集所有的对象数目。

4.2 实验结果对比

为了验证算法的有效性, 将 MKSC 算法和多视角

模糊聚类算法 Co-FKM^[14]、基于多任务的组合 K-means 算法 CombKM^[15] 和在特征空间进行协同聚类算法 Co-clustering^[16] 在性能上进行比较。

使用人工数据集和真实数据集 (MF) 验证 MKSC 算法的有效性。其中人工数据集中含有两个视角, 每个视角中包含 700 个数据样本, 每个数据样本维度为 2, 每个视角中含有三个数据类簇, 其中第二个视角数据存在噪声点的干扰, 这两个视角如图 2 所示。

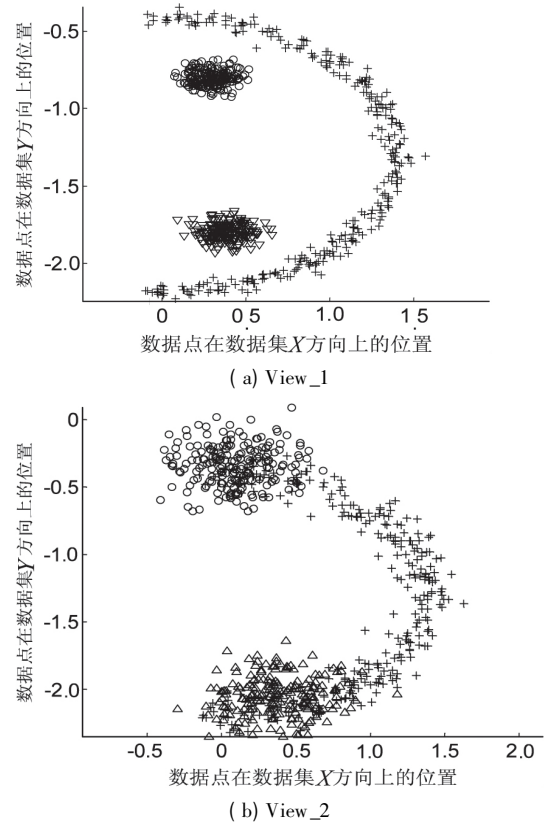


图2 人工数据集

表 1 为人工数据集在 MKSC 算法中的应用。实验选取 $\varepsilon_0 = 0.0004$, 通过 p 值的大小来对比聚类结果和权重值。其中 View_1 和 View_2 表示两个视角的权重值。

表 1 MKSC 算法取不同 p 值的 NMI 值和权重

p	NMI	RI	View_1	View_2
1.2	1	1	0.812	0.188
1.3	1	1	0.80	0.2
1.5	1	1	0.752	0.248
2	0.712	0.675	0.58	0.42
2.5	0.705	0.653	0.55	0.45
3	0.705	0.621	0.5	0.5

从实验结果可以看出, MKSC 算法得到的聚类结果受 p 值的影响, p 值越大, 权重值分配越均匀。 p 值在 $[1.2, 1.5]$ 之间时, NMI 值为最大。从图 2 可以看出, 视角 2 受噪声干扰较严重, 该视角的权重较小。视角

1 权重值大,对聚类结果的划分较明显。

为了进一步分析 MKSC 算法在处理真实数据时的有效性,在 UCI 数据集集中的 multiple features dataset (MF),即数字手写体多特征数据集,进行实验验证。该数据集包含 6 个视角,分别为 Mfeat-fou (Mfo)、Mfeat-fac (Mfa)、Mfeat-kar (Mfk)、Mfeat-pix (Mfp)、Mfeat-zer (Mfz)、Mfeat-mor (Mfm)。结果如表 2 和图 3 所示。

表 2 各种算法在手写体多特征数据集上的聚类结果

指标	CombKM	Co-clustering	Co-FKM	MKSC
RI-mean	0.885	0.915	0.932	0.942
RI-std	0.021	0.013	0.023	0.032
NMI-mean	0.655	0.705	0.745	0.856
NMI-std	0.033	0.072	0.075	0.022

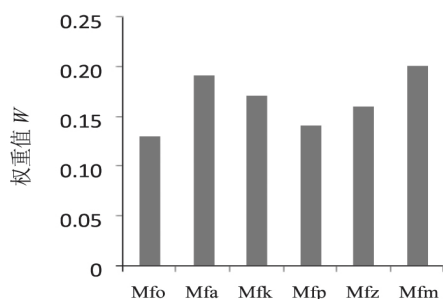


图 3 算法 MKSC 在数据集 MF 的权重值

从实验结果可以看出,与算法 CombKM、Co-clustering 和 Co-FKM 相比, MKSC 算法聚类效果更好,更接近于真实数据的分布,说明该算法的各视角协作能力更强。从权重值上看各视角的重要程度较均衡,上述原因使得 Co-FKM 算法的结果更接近于文中算法,但是对比下 Mfa 和 Mfm 视角权重值较大,说明该视角具有更好的划分。

5 结束语

通过研究多视角聚类算法的各种模型,发现多视角聚类算法聚类结果更优,精度更精确,在解决多视角任务时提出了一种优化加权多视角 K-means 聚类算法。该算法利用权重的策略,为每个视角分配权重,通过引入常数优化权重来解决受噪声干扰视角和较分散的视角容易丢失的问题。实验结果表明,该算法具有较好的聚类效果。

参考文献:

[1] 熊子源,徐振海,张亮,等.基于聚类算法的最优子阵划分方法研究[J].电子学报,2011,39(11):2615-2621.

- [2] 罗恩韬,王国军.大数据中一种基于语义特征阈值的层次聚类方法[J].电子与信息学报,2015,37(12):2795-2801.
- [3] 刘卓,杨悦,张健沛,等.不确定度模型下数据流自适应网格密度聚类算法[J].计算机研究与发展,2014,51(11):2518-2527.
- [4] MIYAHARA S, MIYAMOTO S. A family of algorithms using spectral clustering and DBSCAN [C]//IEEE international conference on granular computing. Noboribetsu, Japan: IEEE, 2014: 196-200.
- [5] 邱保志,贺艳芳.多视角核 K-means 聚类算法的收敛性证明[J].郑州大学学报:理学版,2017,49(3):32-38.
- [6] 蒋亦樟,邓赵红,王骏,等.熵加权多视角协同划分模糊聚类算法[J].软件学报,2014,25(10):2293-2311.
- [7] 刘正,张国印,陈志远.基于特征加权和非负矩阵分解的多视角聚类算法[J].电子学报,2016,44(3):535-540.
- [8] WANG Jidong, ZENG Huajun, CHEN Zheng, et al. ReCoM: reinforcement clustering of multi-type interrelated data objects [C]//Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval. Toronto, Canada: ACM, 2003: 274-281.
- [9] BICKEL S, SCHEFFER T. Multi-view clustering [C]//IEEE international conference on data mining. Brighton, UK: IEEE, 2004: 19-26.
- [10] BICKEL S, SCHEFFER T. Estimation of mixture models using Co-EM [C]//Proceedings of the 16th European conference on machine learning. Porto, Portugal: Springer-Verlag, 2005: 35-46.
- [11] CHAUDHURI K, KAKADE S M, LIVESCU K, et al. Multi-view clustering via canonical correlation analysis [C]//Proceedings of international conference on machine learning. New York, NY, USA: ACM, 2009: 129-136.
- [12] TZORTZIS G F, LIKAS C L. The global kernel k-means algorithm for clustering in feature space [J]. IEEE Transactions on Neural Networks, 2009, 20(7): 1181-1194.
- [13] DENG Zhaohong, CHOI K S, CHUNG F L, et al. Enhanced soft subspace clustering integrating within-cluster and between-cluster information [J]. Pattern Recognition, 2010, 43(3): 767-781.
- [14] CLEUZIOU G, EXBRAYAT M, MARTIN L, et al. CoFKM: a centralized method for multiple-view clustering [C]//Ninth IEEE international conference on data mining. Miami, FL, USA: IEEE, 2009: 752-757.
- [15] GU Quanquan, ZHOU Jie. Learning the shared subspace for multi-task clustering and transductive transfer classification [C]//IEEE international conference on data mining. Miami, FL, USA: IEEE, 2009: 159-168.
- [16] GU Quanquan, ZHOU Jie. Co-clustering on manifolds [C]//ACM SIGKDD international conference on knowledge discovery and data mining. Paris, France: ACM, 2009: 359-368.