文章编号: 1001-0920 (2014) 02-0236-05 DOI: 10.13195/j.kzyjc.2012.1646

基于最大内聚度基准的加权投票聚类集成

陈晓云, 陈 刚

(福州大学 数学与计算机科学学院，福州 350116)

摘 要: 提出一种基于投票的聚类集成方法. 通过分析聚类结构与聚类准确率的关系, 将内聚度最高的聚类成员作为重新标记的基准以实现簇标记的统一; 同时, 根据数据点在不同聚类成员中与所划分簇中心的距离确定权值, 最终实现加权投票. 实验结果表明, 该算法在准确率和稳定性上均有较大提高.

关键词: 聚类集成；加权投票；内聚度

中图分类号: TP391 **文献标志码:** A

Weighted voting clustering ensemble based on maximum cohesion

CHEN Xiao-yun, *CHEN Gang*

(School of Mathematics and Computer Science，Fuzhou University，Fuzhou 350116，China. Correspondent：CHEN Xiao-yun，E-mail：c_xiaoyun@21cn.com)

Abstract: A voting-based clustering ensemble method is presented. By analyzing the relationship between the clustering structure and accuracy, the highest cohesive cluster member is considered as the benchmark of relabel algorithm to unify cluster labels. Then the voting weights are determined by the distance from cluster centers which data points in different cluster members are divide into. The experimental results show that, the proposed algorithm is greatly improved in accuracy and stability.

Key words: clustering ensemble；weighted voting；cohesion

0 引 言

聚类分析是数据分析中的基础问题, 通过对无标签数据间的相似度进行计算将数据集划分为若干个自然簇, 达到簇内对象尽可能相似、簇间差异尽可能大的目的. 单一的聚类算法易受参数影响, 且适用的数据集类型有限. 聚类集成利用共识函数对多个不同的聚类结果进行综合得到新结果, 比单个聚类算法更具有鲁棒性、新颖性、稳定性和可扩展性[1]. 目前, 许多学者都结合实际问题提出了多种聚类集成算法对共识函数进行拓展, 这主要分为两类: 一类规避了簇标记对应的问题, 但着眼于数据点间的关系, 普遍具有较高的复杂度, 如共协矩阵[2]、超图法[3]、分类特征空间[4]等; 另一类对不同聚类成员的簇标记进行重新对应, 简单易行, 时间复杂度低, 如投票法[5-10].

已有的基于投票的聚类集成方法受聚类成员质量的影响较大, 且研究重点集中在如何提高聚类成员簇标记的匹配质量, 忽略了作为基准的聚类成员对整体集成效果可能带来的影响; 而且加权方法或是赋予固定权值, 或是复杂度较高, 未能在时效上取得平衡.

本文提出一种基于投票法的聚类集成方法, 将内聚度最高的聚类成员作为重新标记的基准, 并用互信息值确定不同聚类成员的最佳匹配簇, 实现聚类成员间簇标签的统一. 最后计算数据点在不同聚类成员中与所划分簇中心的距离, 将该距离的倒数作为权值加权投票.

1 聚类集成中的投票问题

投票法在分类集成中有着广泛的应用, 但在聚类集成中却存在簇标记对应的问题. 例如聚类成员的标记向量 $[1,1,2,2,1,1]^{\mathrm{T}}$ 与 $[2,2,1,1,2,2]^{\mathrm{T}}$ 指代的是同样的聚类结果, 但由于簇标记的不同, 无法识别为相同的划分. 文献 [8] 提出一种不同聚类成员簇标签的重新标记法, 该方法基于一种启发式思想, 即不同聚类成员中标记对应的簇所共有的数据点个数应该是最多的. 假设有两个聚类成员 U_1 和 U_2, 分别将原始数据集划分为 k 个簇, 得到两个不同的聚类标记 $S_a =$

收稿日期: 2012-11-10；**修回日期:** 2012-12-30.

基金项目: 国家自然科学基金项目(71273053).

作者简介: 陈晓云(1970−), 女, 教授, 博士, 从事数据挖掘、机器学习、模式识别等研究；陈刚(1986−), 男, 硕士生, 从事数据挖掘、模式识别的研究.

$[S_a(1),\ S_a(2),\ \cdots,\ S_a(k)]^{\mathrm{T}}$ 和 $S_b=[S_b(1),\ S_b(2),\ \cdots,\ S_b(k)]^{\mathrm{T}}$. 重新标记法步骤如下:

1) 将 S_a 与 S_b 中每一对 $S_a(i)$ 和 $S_b(j)$ 所包含的相同数据点个数输出到 $k\times k$ 的矩阵 M.

2) 选择该矩阵中最大的值, 将该对应关系输出并删除该值所处的行和列的元素.

3) 重复上一步直到矩阵没有元素.

当聚类成员个数为 N 时, 任意选择其中一个成员作为基准, 其他成员按上述方法与基准成员的簇标记进行统一, 时间复杂度仅为 $O(k^2N)$.

文献 [11] 认为, 仅凭两个簇共有的数据点个数来判断对应关系容易导致信息丢失, 因此引入互信息值来代替原本的共有数据点个数, 即

$$\mathrm{MI}(S_a(i),S_b(j))=\log\frac{S_{i,j}}{S_i+S_j-S_{i,j}}. \tag{1}$$

其中: $S_{i,j}$ 表示 $S_a(i)$ 与 $S_b(j)$ 共有的数据点个数, S_i 和 S_i 分别表示 $S_a(i)$ 和 $S_b(j)$ 各自的数据点个数. 按照此匹配方法可以使使用的数据点总个数最大.

2　基于最大内聚度基准的加权投票聚类集成

2.1　最大内聚度基准

传统的投票法采用随机选基准的方式进行重新标记, 虽然提高了簇标签问题的解决效率, 但由于基准的质量参差不齐, 很大程度上影响了聚类集成的总体效果. 本文以 k 均值算法[12]作为研究对象, 研究基准的确定方法.

k 均值算法是一种经典的聚类算法, 通过对聚类中心进行反复迭代直到满足收敛条件来得到最终聚类结果. k 均值聚类结果可能受以下 3 个指标的影响:

1) 样本的总类内离散度

$$S_w=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(m_j^{(i)}-c_i)^2; \tag{2}$$

2) 样本的类间离散度

$$S_b=\sum_{i=1}^{k}\sum_{j=1\ j\neq i}^{k}(c_i-c_j)^2; \tag{3}$$

3) Fisher 准则函数

$$J_F=\frac{S_b}{S_w}. \tag{4}$$

其中: k 为聚类结果的簇个数, $n_i(i=1,2,\cdots,k)$ 为每个簇的样本数, $m_j^{(i)}$ 为该簇的样本向量, c_i 为该簇的中心向量, S_w 为聚类结果中样本到其所属簇中心的距离之和, S_b 为簇中心两两间的距离之和.

因为类间离散度仅仅反映聚类结果的簇中心之间的关系, 而簇内样本的紧密程度则决定了样本被误分的概率, 所以聚类结果的类内离散度 S_w 与类间离散度 S_b 共同发生作用. 因为 k 均值算法本身是以最小化类内离散度为目的, 所以聚类结果的质量与类内离散度关系更为密切. 本文实验正说明了这点. 实验在 UCI 机器学习数据库的 wine 数据集上进行, 共运行 20 组实验, 以欧氏距离作为距离度量, 每组实验由 Matlab R2010b 自带的 k-means 程序采用随机种子的方式生成 15 个聚类结果, 分别求出每组成员中 S_w 最小、S_b 最大、J_F 最大的 3 个成员, 最后统计这 3 个成员平均聚类准确率及其分别在每组实验中准确率最高的概率, 实验结果如表 1 所示.

表 1　3 种成员的聚类效果比较

比较对象	平均准确率/%	准确率最高的概率/%
$\mathrm{Member}_{\min(S_w)}$	**70.22**	**100**
$\mathrm{Member}_{\max(S_b)}$	56.74	10
$\mathrm{Member}_{\min(J_F)}$	**70.22**	**100**

在实验中, 所有聚类结果的总体平均准确率为 67.35%. 从表 1 中可以看出, S_b 最大的成员平均准确率仅为 56.74%, 低于总体平均准确率, 可以认为 S_b 与 k 均值聚类结果不存在密切关系. 为考察 S_w 与 J_F 的关系, 可近似地将 J_F 的分子 S_b 设为 1, 则有

$$\mathrm{Member}_{\max(J_F(S_b=1))}=\mathrm{Member}_{\max(\frac{1}{S_w})}=\mathrm{Member}_{\min(S_w)}. \tag{5}$$

从式 (5) 可以看出, 排除 S_b 的影响后 S_w 与 J_F 取极值得到的成员是相同的, 这点反映在实验中二者的成员结果具有相同的平均聚类准确率, 且准确率最高的概率也相同. 因此, 本文把具有最小的总类内离散度 (即内聚度最大的聚类结果) 作为重新标记法的基准.

2.2　加权投票方法

文献 [8] 提到采用平均互信息作为投票的权值, 虽然能在一定程度上刻画聚类个体间的紧密程度, 但需要计算不同成员间的不同簇所共有的数据点个数来获得权值, 时间复杂度高. 考虑到距离簇中心越远的数据点被误划分的概率越大, 反映在投票上其所得到的票数越小, 因而本文从单个聚类成员的距离信息出发, 计算不同成员中数据点到簇中心的距离对投票进行加权.

假设有 N 个聚类成员 $[S_1,S_2,\cdots,S_N]$, 每个成员被划分为 k 个簇 $S_i=[S_i(1),S_i(2),\cdots,S_i(k)]^{\mathrm{T}}$, 数据点 p 在 N 个聚类成员中分别划分到簇 $S_1(1),S_2(2),\cdots,S_N(N)$ 上, 这 N 个簇的中心分别为 $c_1(1),c_2(2),\cdots,c_N(N)$. 数据点 p 的加权投票方法如下:

1) 计算出数据点 p 在不同聚类成员中距离所在簇中心的距离, 即

$$r_i=\|p-c_i(i)\|, \tag{6}$$

其中 $i=1,2,\cdots,N$.

2) 对式 (6) 得到的距离进行变换得出权值

$$v_i = \frac{Z}{r_i}, \tag{7}$$

其中: Z 用于将权值规范化, 使 $\sum_{i=1}^{N} \mathrm{vote}_i = N$.

3) 生成标签与权值的对应关系

$$T(S_i(i)) = \mathrm{vote}_i, \tag{8}$$

其中 T 为 2.1 节中将不同聚类成员的簇标签进行重新标记的转换函数. 将权值作为票数, 将转换后标记相同的票数相加, 票数最高的标记即为该数据点所属的簇.

该方法避免了簇间两两对应所需要的高昂的计算代价, 仅考虑簇内部的距离信息进行加权投票, 有效降低了算法的时间复杂度.

2.3 算法描述

利用上述重新标记法和加权方法进行聚类集成, 具体流程如下.

算法 1 基于最大内聚度基准的加权投票法 (WVMC).

输入: n 个数据点构成的集合 Data, 聚类成员个数 t;

输出: 聚类结果 Result.

Step 1: 通过随机参数的方式对 Data 执行 t 次 k 均值聚类, 得到 t 个聚类成员.

Step 2: 重新标记.

Step 2.1: 计算 t 个成员的总类内离散度 $s_1, s_2, \cdots, s_t$;

Step 2.2: 选取总类内离散度最小的第 i 个成员作为基准进行重新标记, 得到统一的簇标记向量 $C = [C_1, C_2, \cdots, C_k]^{\mathrm{T}}$ 以及各个聚类成员的标记转换函数 T.

Step 3: for $j = 1$ to n, //加权投票

Step 3.1: 计算数据点 p_j 在不同聚类成员中距离所在簇中心的距离, 得到重新标记后的标签与规范化票数的对应关系向量 M;

Step 3.2: 在 M 中找到票数最高的簇标记, 将其与 p_j 关联并存储到结果矩阵 Result.

Step 4: 输出结果矩阵.

3 实验与分析

3.1 实验环境

实验使用 Matlab R2010b 自带的 kmeans 程序进行测试, 共分为 3 个部分: 第 1 部分针对 WVMC 与主流的共识函数方法进行比较, 以测试算法的性能; 第 2 部分针对 TVA (文献 [11] 提出的投票法) 和本文的 WVMC 在多个数据集上进行比较, 以测试算法的改进程度; 第 3 部分采取不同的距离度量方式进行实验, 以测试 WVMC 算法的准确率与聚类成员数量间的关系.

已知数据集的原始类标签向量 $C = [C(1), C(2), \cdots, C(k)]^{\mathrm{T}}$, 实验所得的簇标记向量 $L = [L(1), L(2), \cdots, L(k)]^{\mathrm{T}}$, 使用重新标记法进行匹配, 假设标记间的对应关系为 $L(i) \rightarrow C(i)$. 评价标准采用 Micro-precision 进行衡量, 定义为

$$\text{Micro-p} = \frac{1}{n}\sum_{i=1}^{k} N_i. \tag{9}$$

其中: n 为数据集中对象的个数, N_i 为 $L(i)$ 与 $C(i)$ 共有的数据点个数.

3.2 5 种共识函数比较

实验选择了 CSPA[3]、HGPA[3]、MCLA[3]、TVA[11] 四种主流的共识函数算法与本文算法进行比较, 对象为 UCI 机器学习数据库的 iris 数据集, 通过 20 次聚类集成得出平均准确率, 距离为欧氏距离. CSPA、HGPA、MCLA 为文献 [11] 提供的实验数据. 实验结果如表 2 所示.

表 2 5 种共识函数的比较

N	k	CSPA/%	HGPA/%	MCLA/%	TVA/%	**WVMC**/%
5	3	87.73	62.00	89.33	85	**89.4**
10	3	86.93	53.33	**89.33**	88.7	89.3
15	3	87.60	61.80	89.33	88	**89.4**
20	3	87.13	59.73	89.33	85.5	**89.7**
30	3	86.87	59.13	89.33	87	**89.5**
平均准确率/%	87.25%	59.20	89.33	87.01	**89.53**	
时间复杂度	kn^2N	knN	k^2nN^2	$\boldsymbol{k^2N}$	$\boldsymbol{k^2N}$	

N 表示聚类成员的个数, k 为聚类算法生成簇的个数, n 为数据点个数. 由表 2 可以看出, 本文算法的平均准确率最高, MCLA 算法略低, 最差的是 HGPA 算法. 从时间复杂度上考虑, CSPA、HGPA、MCLA 这 3 个算法由于涉及到数据点操作, 普遍具有较高的时间复杂度, 尤其是 CSPA 达到 $O(kn^2N)$, 当数据集规模较大时, 运算时间也相应增加, 严重影响聚类的效率; TVA 虽然与 WVMC 算法同样保持较低的时间复杂度, 但易受 N 值变化影响. 本文提出的 WVMC 算法聚类效果最好, 且时间复杂度仅为 $O(k^2N)$, 不会随数

据集规模变化而变化.

3.3　WVMC 与 TVA 的比较

由于算法 TVA 和 WVMC 同样具有较低的时间复杂度, 为进一步地深入比较, 构造了 8 个数据集进行测试, 分别是 UCI 机器学习数据库中的 Segment-Challenge、Landsat Satellite、Image Segmentation、Iris、Class、Vehicle Silhouettes、Wine Quality Red 数据集, 以及一个由纹理图像 (图 1) 提取 CS-LBP[13] 纹理谱生成的数据集 Outex_TC, 表 3 列出了各个数据集的具体信息. 实验的距离度量为欧氏距离, 其中单一聚类算法为 500 次聚类的平均准确率, 而 TVA 和 WVMC 算法则是对 20 次聚类集成求平均准确率, 每次集成包含 25 个聚类成员.

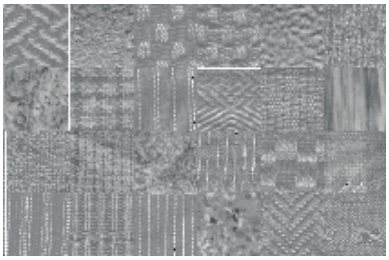

图 1　Outex_TC_00003 纹理图像

表 3　数据集的信息

数据集	数据点个数	簇个数	数据分布类型	特征维数
Iris	150	3	均匀分布	4
Outex_Tc	480	24	均匀分布	16
Glass	214	6	非均匀分布	9
Segment-Challenge	1 500	7	非均匀分布	19
Landsat Satellite	4 435	6	非均匀分布	36
Image Segmentation	2 310	7	非均匀分布	19
Vehicle Silhouettes	846	4	非均匀分布	18
Wine Quality Red	1 599	4	非均匀分布	11

表 4 给出了数据集上的测试结果, WVMC 和 TVA 的平均准确率均高于单一的聚类算法, 而 WVMC 的表现更为优秀, 在 8 个数据集上比 TVA 分别领先 0.98%～6.17%, 尤其是在 iris 数据集上差距最大, 平均准确率比 TVA 高了 2.44%. 实验数据集在簇数量和数据集规模上均不相同, 再次体现了 WVMC 算法的优势.

表 4　WVMC 与 TVA 在不同数据集上的比较

数据集	单一聚类/%	TVA/%	**WVMC/%**
Iris	82.14	83.20	**89.37**
Outex_Tc	66.77	68.19	**70.73**
Glass	51.70	53.06	**54.04**
Segment-Challenge	52.05	54.43	**56.90**
Landsat Satellite	52.56	55.40	**56.44**
Image Segmentation	52.46	54.23	**57.32**
Vehicle Silhouettes	44.72	44.53	**44.96**
Wine Quality Red	20.76	20.39	**23.20**
平均值	52.90	54.18	**56.62**

3.4　稳定性比较

为了测试 WVMC 算法的准确率与聚类成员数量的关系, 以及研究不同距离度量对加权投票的影响, 选择 4 种距离度量方式对 WVMC 和 TVA 进行比较, 分别是欧氏距离、城区距离、余弦距离和皮尔森相关系数; 针对不同聚类成员数量在 iris 数据集上分别进行 20 次集成并比较二者的平均准确率, 实验结果见图 2.

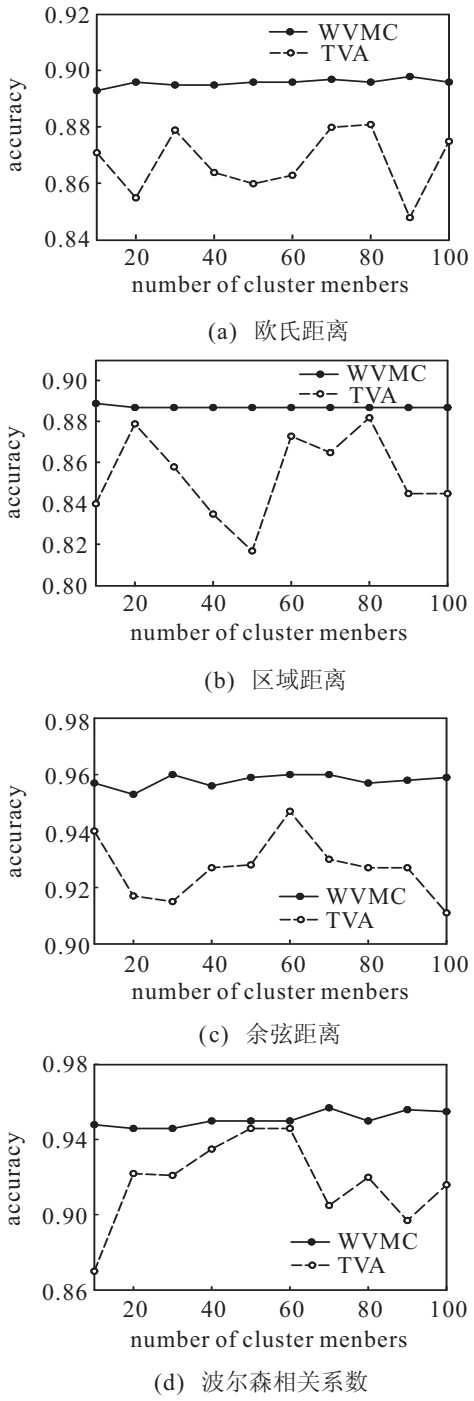

图 2　不同距离度量下 WVMC 与 TVA 的比较

实验中一方面对比了 WVMC 和 TVA 的性能, 可以看出, 不论在何种距离度量下 WVMC 都具有明显的优势, 且当距离度量为皮尔森相关系数情况下聚类

成员数量为10时, 二者差距可以达到7.8%. 另一方面, 由于基准的质量较优, WVMC重新标记过程的效率得到显著提高, 与TVA的准确率随着聚类成员数量变化出现较大幅度波动不同, WVMC算法始终保持了相对稳定的聚类准确率, 这点在以城区距离作为度量时表现得尤为明显.

4 结 论

本文将内聚度最高的聚类成员作为重新标记的基准, 以互信息值来确定不同聚类成员的最佳匹配簇进行类标签的转换, 并根据数据点在不同聚类成员中与所划分簇中心的距离得出权值, 得到了一种新的基于投票的聚类集成算法WVMC. 本文算法充分利用了投票法时间复杂度低、无需计算数据点间关系的特点, 使其能够应用在较大规模的数据集; 同时具有较高的稳定性和聚类准确率, 因此仅需较少的聚类成员即可实现聚类集成. 实验结果表明, WVMC与TVA、CSPA、HGPA、MCLA相比, 在聚类准确率和时间复杂度及稳定性方面都具有很大的优势, 因此WVMC算法是有效可行的. 如何更好地提高基准成员的质量, 设计计算量更小的权值生成途径是今后进一步研究的方向.

参考文献(References)

[1] Topchy A, Jain A K, Punch W. A mixture model for clustering ensembles[C]. Proc of the 4th SIAM Int Conf on Data Mining. Lake Buena Vista, 2004: 379-390.

[2] Fred A, Jain A K. Combining multiple clusterings using evidence accumulation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(6): 835-850.

[3] Strehl A, Ghos J. Cluster ensembles-a knowledge reuse framework for combining multiple partitions[J]. J of Machine Learning Research, 2002(3): 583-617.

[4] Topchy A, Jain A K, Punch W. Clustering ensembles: Models of consensus and weak partitions[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(12): 1866-1881.

[5] Dudoit S, Fridlyand J. Bagging to improve the accuracy of a clustering procedure[J]. Bioinformatics, 2003, 19(9): 1090-1099.

[6] Fischer B, Buhmann J M. Bagging for path-based clustering[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2003, 25(11): 1411-1415.

[7] Topchy A, Law M, Jain A K, et al. Analysis of consensus partition in clustering ensemble[C]. Proc of the IEEE Int Conf on Data Mining 2004. Brighton, 2004: 225-232.

[8] 唐伟, 周志华. 基于Bagging的选择性聚类集成[J]. 软件学报, 2005, 16(4): 496-502.
(Tang W, Zhou Z H. Bagging-based selective clusterer ensemble[J]. J of Software, 2005, 16(4): 496-502.)

[9] Ayad H G, Kamel M S. Cumulative voting consensus method for partitions with a variable number of clusters[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2008, 30(1): 160-173.

[10] 李春生, 王耀南. 基于模糊简单多数票法则的模糊聚类组合模型[J]. 控制与决策, 2010, 25(3): 394-398.
(Li C S, Wang Y N. Ensemble of fuzzy clusterings based on fuzzy simple majority vote[J]. Control and Decision, 2010, 25(3): 394-398.)

[11] Meng F R, Tong X J, Wang Z X. A clustering-ensemble approach based on voting[C]. Proc of the IEEE Int Conf on Artificial Intelligence and Computational Intelligence 2011. Taiyuan, 2011: 421-427.

[12] Mac Q J. Some methods for classification and analysis of multivariate observations[C]. Proc of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, 1967: 281-297.

[13] Heikkilä M, Pietikäinen M, Schmid C. Description of interest regions with local binary patterns[J]. Pattern Recognition, 2009, 42(3): 425-436.

(责任编辑: 孙艺红)