计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190700044

基于 Xie-Beni 指数的选择性聚类集成

邵 超 马进家
河南财经政法大学计算机与信息工程学院 郑州 450046

摘 要 选择性聚类集成是选择一部分精度高、差异性大的基聚类结果进行集成,从而得到更为有效的聚类集成结果。然而,聚类结果的准确性难以客观度量。为此,文中提出了一种基于 Xie-Beni 指数的选择性聚类集成算法,该算法采用 Xie-Beni 指数来度量基聚类结果的有效性,利用并结合 NMI(互信息)选择出精度较高的基聚类结果,从而提升聚类结果的准确性。实验结果证实了该算法的有效性。

关键词:选择性聚类集成;聚类有效性指数;Xie-beni;NMI
中图法分类号 TP181

Selective Clustering Ensemble Based on Xie-Beni Index

SHAO Chao and MA Jin-jia
School of Computer & Information Engineering, Henan University of Economics and Law, Zhengzhou 450046, China

Abstract Selective clustering ensemble is to select some of the basic clustering results with high accuracy and large diversity for integration, so as to obtain more effective clustering ensemble results. In the cluster analysis application, the cluster validity index is used to measure the goodness of the clustering results. In this paper, a selective clustering ensemble algorithm based on Xie-Beni index is proposed. The algorithm uses Xie-Beni index to measure the validity of the basic clustering results, and uses NMI(normalized mutual information) to select the better basic clustering results to enhance the aggregation, thereby improving the accuracy of the clustering results. Experimental results confirm the effectiveness of the algorithm.

Keywords Selective clustering ensemble, Clustering validity index, Xie-beni, NMI

1 引言

聚类分析(Cluster Analysis)按照最大化类内相似性、最小化类间相似性的原则把数据划分成多个簇,可以揭示数据的隐含特征[1],聚类算法旨在通过寻找有限的聚类集合根据其对象之间的相似性来划分数据集[2]。聚类分析在生物信息、医学诊断、计算机视觉等领域得到了广泛的应用。在过去的十年中,聚类集合已成为一种流行的技术来处理数据聚类问题。由于良好的性能和灵活的流程,集群整合已经应用于机器学习的许多领域,如文档数据集学习[3]、高维数据聚类[4]、流数据聚类[5]、噪声数据分析[6],以及不平衡的数据分析[7]。与分类不同,在聚类分析中,数据对象没有相应的标签信息,因此,聚类是一种无监督学习方法。由于数据的实际分布情况未知,实验前一般需要先对数据分布做出假设,再用某种规则对数据进行划分。但这种假设并不都符合数据集自身特征,也不存在某一种聚类算法能满足任意形状、任意分布的数据[8]。因此,面对这种不确定性,如何选择恰当的聚类算法对具体的数据集进行划分是聚类分析中一个不可避免的挑战[9]。

聚类集成[10]最早由 Strehl 等[11]于 2002 年提出,对于有 n 个对象的数据集 $X=\{x_1,x_2,x_3,x_4,\cdots,x_n\}$,通过不同的生成方法产生 m 个基聚类结果 $\Pi=\{\pi_1,\pi_2,\cdots,\pi_m\}$,其中 π_j 是第 j 次划分结果的类标签集合,$\pi_j(x_i)$ 是第 i 个对象在 π_j 中类标签,然后通过一个共识函数(即一致性函数)对 m 个基聚类结果进行集成,得到最终的聚类集成结果。人们普遍认为聚类集成的结果比只使用单一聚类算法的结果要更好,聚类集成的优点主要包括以下几个方面[12-13]:1)鲁棒性,聚类集成比单一聚类方法对噪声、孤立点和其他影响的抗干扰能力强。2)适用性,聚类集成与单一聚类算法相比能够适应更多类型的数据集。3)可扩展性和并行性,聚类集成能对数据子集进行并行聚类且能够进行合并;能对分布式数据源或数据属性的聚类结果进行合并。与此同时,聚类集成也存在着一定的缺点:与使用单个基聚类结果相比[14],聚类集成虽然可以提高学习精度,但对于如何选择个体精度高、差异性大的基聚类结果仍存在许多困难。

由于在无监督学习中,传统的聚类算法缺乏客观的精度度量,所以针对如何选择个体精度高、差异性大的基聚类结果,本文提出一种基于 Xie-Beni 指数的选择性聚类集成算法(Selective Clustering Ensemble Based On Xie-Beni Index, SCEX)。因为 K-means 算法在解决聚类问题时具有简单、快速的特点,在处理大规模数据集时,可保持伸缩性和高效性,所以本文运用此算法产生基聚类结果,然后利用 Xie-Beni[15]指数计算基聚类结果的有效性,利用 NMI(normalized mutual

基金项目:国家自然科学基金(61806073,61907011)
This work was supported by the National Natural Science Foundation of China (61806073,61907011).
通信作者:邵超(sc_flying@163.com)

information,互信息)计算相应基聚类结果的相似性,从中选择出差异性大、个体精度高的基聚类结果,最后通过 CSPA[9](Cluster-based Similarity Partitioning Algorithm)算法得到最终聚类结果。

2 相关工作

2.1 聚类有效性指数

Xie 等[15]通过引用数据集合,并且考虑数据元素的几何结构和内在的联系方式,给出了类内聚和类间距的定义,提出了有效性指数 V_{xb},如式(1)所示:

$$V_{xb}=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}|x_j-v_i|^2}{n*\min_{i,j}|v_i-v_j|^2} \tag{1}$$

其中,x_j 表示第 j 个数据点,v_i 表示第 i 个簇中心,u_{ij} 表示 x_j 隶属于 v_i 的隶属度,在本文中 m 为权重指数。该有效性指标中分子表示模糊划分的紧凑程度,分母表示类间距的分离程度。

好的聚类结果应具有较小的类内距和相对较大的类间距,因此 V_{xb} 的最小值对应的 c 为最佳聚类个数 c^*,如式(2)所示:

$$c^*=\arg\min V_{xb} \tag{2}$$

V_{xb} 有效性指标得到了广泛的应用,成为以后众多改进算法的基础。然而由于指标的分子随着 c 的增大而逐渐减小,当 c 逐渐逼近 n 时,分子趋向于 0,分母则逐渐增大,指标同样会因为过分单调而失去判断能力。

2.2 NMI 相似性

NMI[4]是用来度量两个事件关联性的一种方法。在聚类集成中,NMI 用来度量两个聚类结果间的差异性,对于两个聚类对象 π_a 和 π_b,其互信息如式(3)所示[8]:

$$NMI(\pi_a,\pi_b)=\frac{\sum_{h=1}^{k_a}\sum_{l=1}^{k_b}n_{h,l}/N\log_2\left(\frac{n_{h,l}/N}{n_h*n_l/N*N}\right)}{\sqrt{\left(\sum_{h=1}^{k_a}n_h/N\log_2 n_h/N\right)\left(\sum_{l=1}^{k_b}n_l/N\log_2 n_l/N\right)}} \tag{3}$$

其中,k_a 和 k_b 分别是 π_a 和 π_b 中类的个数,n_h 是 π_a 的 h 类中对象的个数,n_l 是 π_b 的 l 类中的对象个数,$n_{h,l}$ 是同时在 π_a 的 h 类和 π_b 的 l 类中对象个数。

2.3 CSPA 集成算法

CSPA 需要把每次 K-means 聚类产生的初始基聚类结果构建成一个 $n*n$ 大小的相似性矩阵 $\boldsymbol{S}$。具体构建方法如下[5]:在同一个初始基聚类结果中,如果两个对象在相同的簇中,则相似性为 1,否则相似性为 0,则可以容易地为每个聚类构建一个 $n*n$ 大小的相似性矩阵。在得到矩阵 $\boldsymbol{S}$ 后可以使用任何基于相似性的聚类算法来完成集成。

3 基于 Xie-Beni 指数的选择性聚类集成

针对如何选择具有差异性大、个体精度高的基聚类结果的问题,本文提出了基于 Xie-Beni 指数的选择性聚类集成的 SCEX 算法来解决此问题。

SCEX 算法的具体描述如算法 1 所示。

算法 1 SCEX 算法

输入:包含有 n 个对象的数据集 X

输出:包含有 n 个对象类标签的集合

1. 使用 K-means 算法得到 m 个基聚类结果,组成基聚类成员集 k_0;
2. 根据式(1)计算 k_0 内基聚类成员的 V_{xb} 有效性,并筛选出若干个 V_{xb} 值最大的基聚类成员组成基聚类成员核 k_1;
3. 利用式(3)计算 k_0 与 k_1 内各成员之间的 NMI 相似性,并筛选出 n_2 个 NMI 值最大的基聚类成员组成个体精度高的基聚类成员集 k_2;
4. 根据式(3)计算 k_2 内各成员之间的 NMI 相似性,并筛选出 n_3 个相似性小的基聚类成员组成候选集成成员集 k_3;
5. 通过 CSPA 算法得到候选集成成员集 k_3 对应的相似性矩阵,然后通过凝聚的层次聚类算法进行集成,得到最终聚类结果。

3.1 利用 K-means 算法产生基聚类成员集 k_0

利用 K-means 算法对初始簇中心敏感这一特点产生 m(本文实验中 $m=20$)个有差异的基聚类结果:原始数据集的 K 值等于原始数据集标签个数,然后进行 K-means 聚类,组成基聚类成员集 k_0。

3.2 计算 k_0 中每个基聚类成员的有效性,并结合 NMI 相似性筛选出个体精度高的基聚类成员集

利用式(1)来计算 k_0 中每个聚类成员的 V_{xb}[13]。然后选出 V_{xb} 最大的 n_1 个基聚类成员,组成基聚类成员核 k_1。这里选出多个基聚类成员组成基聚类成员核是为了克服 Xie-Beni 指数的局限性。

然后利用式(2)分别计算 k_0 与 k_1 内各成员之间的 NMI 相似性,并选择出 n_2 个相似性值最大的基聚类成员,组成个体精度高的基聚类成员集 k_2。

3.3 计算 k_2 内各成员之间的相似性,筛选出差异性大的候选集成成员集 k_3

利用式(2)计算 k_2 内各成员之间的相似性,挑选出相似性最小的 n_3 个基聚类成员组成成员集 k_3。k_3 中的各基聚类成员具有个体精度高、差异性大的特点,作为候选集成成员集。

3.4 CSPA 集成

计算候选集成成员集对应的相似性矩阵,然后再通过凝聚的层次聚类算法进行集成,得到最终聚类结果。

4 实验仿真

为了验证 SCEX 算法的有效性,本文在 5 个数据集上分别运行谱聚类算法(NJW)、CSPA 聚类集成算法、文献[11]中提出的 CAS 算法和本文的 SCEX 算法。SCEX 算法中 $m=20$,$n_1=3$,$n_2=10$,$n_3=5$。在 CSPA 聚类算法中 $m=20$。

4.1 实验数据

测试本文算法聚类质量的实验数集采用人工数据集,分别为 Aggregation,Two_Cluster,Three_Cluster,Five_Cluster 和 Flame 5 个人工数据集。各个数据集的属性如表 1 所列。

表 1 5 个数据集的基本属性

Table 1 Basic attributes of five data sets

实验数据集	数据集属性		
	维数	类数	数据量
Aggregation	2	7	788
Two_Cluster	2	2	400
Three_Cluster	2	3	600
Five_Cluster	2	5	2000
Flame	2	2	240

4.2　实验评价方法及对比试验

由于本文使用的是人工数据集，其每个数据集的原始类标签是已知的，因此采用外部评价法中的 F-measure[8] 作为聚类集成结果的评价指标。F-measure 的值越大，说明聚类集成的效果越好。

表 2　实验结果 F-measure 值对比

Table 2　Comparison of experimental results F-measure value

数据集	F_measure 值			
	NJW	CSPA	CAS	SCEX
Aggregation	0.2088	0.2125	0.2376	**0.2388**
Two_Cluster	0.4239	0.6660	0.6748	**0.6824**
Three_Cluster	0.3683	0.5000	**0.5598**	0.5000
Five_Cluster	0.2701	0.3216	0.3381	**0.3386**
Flame	0.3160	0.6915	**0.8711**	0.7155

从表 2 中可以看出：相较于传统的基聚类算法，使用集成算法后的 F-measure 要好于传统的基聚类算法。除了 Three_Cluster 这个数据集以外，其余的数据集使用 CAS 和 SCEX 算法以后的 F-measure 值都要好于使用 CSPA 集成后的 F-measure 值。而 CAS 算法和 SCEX 算法在不同的数据集上表现出不同的优越性。对于 Aggregation，Two_Cluster，Five_Cluster 数据集，SCEX 算法的 F-measure 值最大；对于 Three_Cluster，Flame 数据集，CAS 算法的 F-measure 值最大。

实验结果直方图如图 1 所示（其中横坐标是 5 个数据集，纵坐标是不同算法得到的结果与每个数据标签之间的 F-measure 值）。

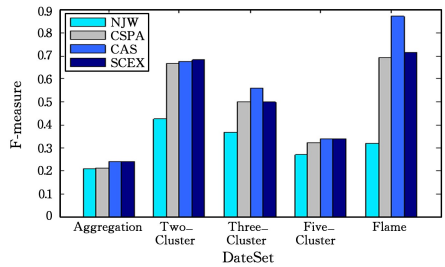
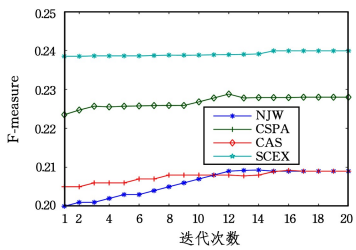


图 1　实验结果直方图

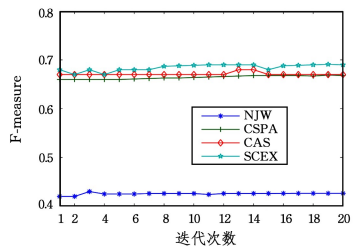
Fig. 1　Histogram of experimental results

从图 1 可以看出，目前并没有哪种选择性聚类集成算法能在所有数据集上展现出比其他算法的优越性，造成这种现象的原因是多样的，比如自身的特性、在选取基聚类结果进行集成时选择的个数不同、使用的集成算法不同等。

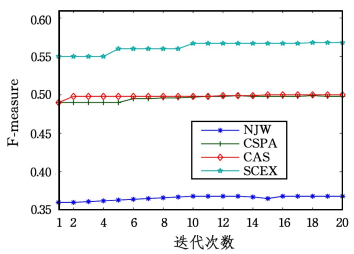
图 2 给出在 Aggregation，Two_Cluster，Three_Cluster，Five_Cluster 和 Flame 5 数据集上每次迭代，即每次选择后的 F-measure 值。其中横坐标是迭代次数，纵坐标是每次的集成结果的 F-measure 值。



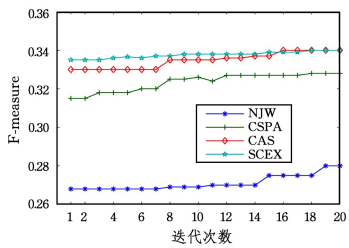
(a) Aggregation



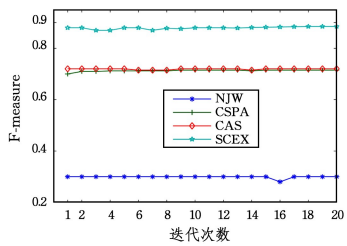
(b) Two_Cluster



(c) Three_Cluster



(d) Five_Cluster



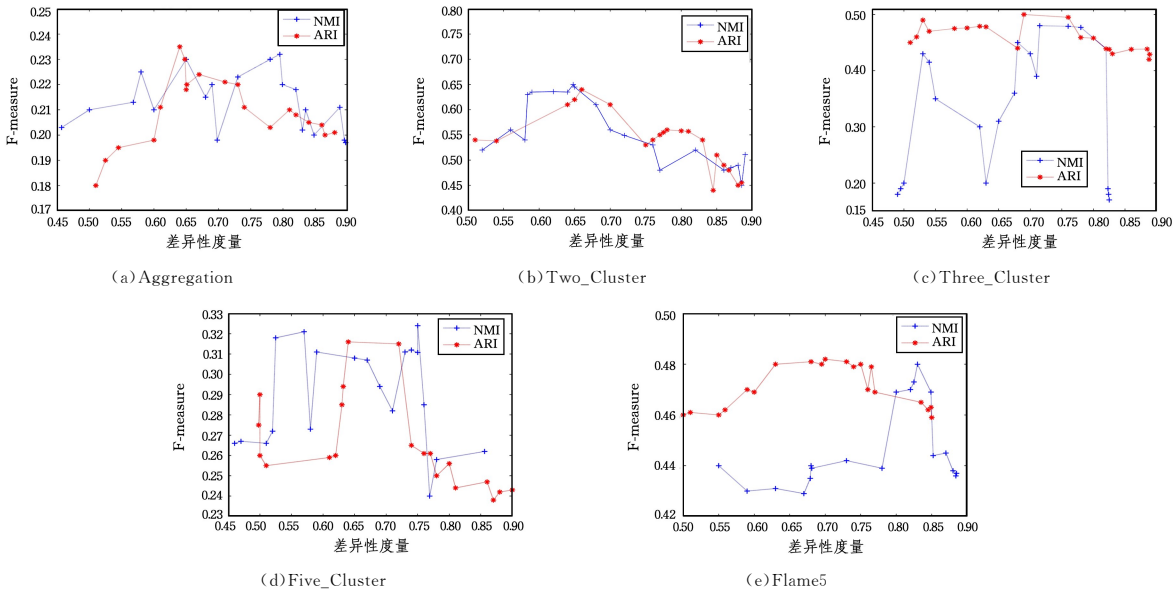
(e) Flame5

图 2　各类算法的迭代次数对比

Fig. 2　Comparison of iteration times of various algorithms

4.3　差异性大小与聚类集成性能分析

首先使用 K-means 算法采用不同初始化的方法产生 20 个基聚类结果之间的差异性，再使用 CSPA 对不同差异性的基聚类结果进行集成。这里我们使用 F-measure 作为聚类集成评价指标，对于上述过程，我们重复进行 20 次，再求其平均值作为最后结果，实验结果如图 3 所示，其中横坐标是差异性度量方法的值，蓝色“+”是 NMI（基于 NMI 的差异性度量方法），红色“*”的是 ARI（基于 Adjusted Rand Index 的差异性度量方法），纵坐标是聚类集成结果的 F-measure 值。



(a) Aggregation (b) Two_Cluster (c) Three_Cluster

(d) Five_Cluster (e) Flame5

图 3 不同数据集在各种差异性度量方法的聚类集成性能比较图(电子版为彩色)

Fig. 3 Comparison of clustering integration performance of different data sets in various difference measurement methods

结束语 本文提出了基于 Xie-Beni 指数的选择性聚类集成的 SCEX 算法,在给出的 5 个人工数据集上做了相应的对比实验,最后通过 F-measure 值进行了实验结果有效性的对比,证实了该算法的有效性。虽然本文在如何挑选个体精度高、差异性大的基聚类成员的选择策略的研究上取得了一些进展,但是单一的聚类有效性指数在应对不同的数据时存在着一定的缺点,例如 Xie-Beni 指数中指标的分子部分随着 c 的增大而逐渐减小,指标同样会因为过分单调而失去判断能力。因此在今后的研究学习中,可以设计一个新的有效性指标,使其借鉴现有经典的有效性指标的优点,同时改进其不足,形成一个高效的有效性指标。利用此新有效性指标进行选择性聚类集成,将会选择出个体精度更高的基聚类结果。

参 考 文 献

[1] HAN J W, KAMBER M, PEI J. Data Ming and Technology (Third Edition)[M]. Beijing: Mechanical Industry Press, 2012.

[2] NALDI M, ANDRE C P L, CARVALHO R. Campello Cluster ensemble selection based on relative validity indexes[J]. Data Min Konwl Disc, 2013, 27: 259-289.

[3] XU S, CHAN K S, GAO J, et al. An integrated K-means-Laplacian cluster ensemble approach for document datasets[J]. Neurocomputing, 2016, 214: 495-507.

[4] FERN B, ZHANG X L, BRODLEY C E. Random projection for high dimensional data clustering: A cluster ensembleapproach [C]// Proceedings of the International Conference on Machine Learning (ICML). 2003: 186-193.

[5] KHAN Y, CHEN Y Y, KE C. Temporal data clustering via weighted clustering ensemble with different representations[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(2): 307-320.

[6] CHEN S, GUO G D, CHEN L F. A new over-sampling method based on cluster ensembles[C]// Proceedings ofthe 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (WAINA). IEEE, 2010: 599-604.

[7] JAIN A K, FLYNN P J. Data Clustering, A Review [J]. ACM Computing Surveys, 1999, 31(3): 264-323.

[8] YANG L Z, ZHOU H J, ZHUO Q, et al. Weighted Clustering Fusion Based on Attribute Importance[J]. Computer Science, 2009, 36(4): 243-245.

[9] LU X Y. Research on Selective Clustering Integration Based on Covariance[D]. Chengdu: Southwest Jiaotong University, 2013.

[10] STREHL A, GHOSH J, CARDIE C. Cluster ensembles: A knowledge reuse framework for combining multiple partitions [J]. Journal of Machine Learning Research, 2002(3): 583-617.

[11] TOPCHY A, JAIN A K, PUNCH W. A Mixture Model for Clustering Ensembles[C]// Proceedings of the 4th SIAM International Conference on Data Mining. 2004: 379-390.

[12] YANG L Z, WANG W Y. Overview of clustering fusion methods[J]. Application Research of Computers, 2005, 22(12): 8-10.

[13] LI S. Selective Clustering Integration Research[D]. Jinan: Shandong Normal University, 2010.

[14] HOU S S. Research and Analysis of Clustering Effectiveness Index[D]. Qingdao: China University of Petroleum, Master of Engineering, 2016.

[15] XIE X L, BENI G. A validity measure for fuzzy clustering[J]. IEEE Trans. Pattern Anal. Mach. Intell, 1991, 13: 841-847.

SHAO Chao, born in 1977, professor, is a member of China Computer Federation. His main research interests include machine learning and so on.