

基于新的成员选择方法的聚类融合算法

冯中慧¹, 何 亮¹, 王 栋^{1,2}

(1 西安交通大学 电子与信息工程学院, 陕西 西安 710049;

2 洋浦经济开发区经济发展局, 海南 海口 578001)

摘 要: 聚类融合是聚类分析领域的一个研究热点, 它将针对同一问题产生的多个聚类成员(即聚类结果)融合成一个结果, 以提高聚类算法的鲁棒性和准确性。在聚类融合过程中, 聚类成员的“质量”非常重要, 一些“质量”比较差的聚类成员会直接影响聚类融合的结果。在深入研究聚类融合算法基础上, 提出新的加权 JP(Jaccard index-Precision) 聚类成员选择方法和基于该方法的聚类融合算法。在多个不同数据集上的实验结果表明, 利用这种新的聚类成员选择方法可以有效地改善聚类融合算法结果的准确性和鲁棒性, 性能显著提高。

关键词: 聚类融合; 聚类成员选择方法; 聚类精度; 差异度

中图分类号: TP274

文献标识码: A

文章编号: 1000—7180(2016)11—0025—05

Clustering Ensemble Algorithm Based on New Members Selection Method

FENG Zhong-hui¹, HE Liang¹, WANG Dong^{1,2}

(1 School of Electronic and Information Engineering, Xi'an Jiaotong University,
Xi'an 710049, China; 2 Hainan Province Yangpu Economic Development Zone Economic
and Development Department, Haikou 578001, China)

Abstract: Clustering ensemble is one attractive field of cluster analysis. Focus on improving the robustness and precision of clustering algorithms, clustering ensemble methods combine multiple clustering members (clustering results) of the same issue into one result. During the process of clustering ensemble, the quality of clustering members is very important, some poor clustering members will directly affect the results of clustering ensemble. In this paper, clustering ensemble algorithms are studied, clustering ensemble algorithms based on new clustering members selection method using weighted JP(Jaccard index-Precision) are proposed. Experimental results show that precision and robustness of clustering are both improved on different datasets. The clustering performance is improved significantly.

Key words: clustering ensemble; clustering members selection method; clustering precision; diversity

1 引言

聚类分析是一种非常有效的数据分析工具, 已广泛应用于众多领域, 并取得了丰硕的研究成果。然而, 在这一领域还存在以下困难: 现有的聚类算法^[1]对被处理的数据都有一个特定的假设, 即每一种聚类分析方法适用的情形是特定的, 但是现实生活中

数据的形式是千变万化的, 如何增强聚类算法的鲁棒性是聚类分析领域必须要解决的一个问题。聚类融合技术是集成学习在无监督学习领域的应用, 其基本思想是将同一问题的多个聚类结果进行融合进而形成一个统一的聚类结果, 这些同一问题的不同聚类结果称为聚类成员。

有一些学者提出在聚类融合的过程中选择一部

分聚类成员进行聚类融合,这种聚类融合的模式称为选择性聚类融合.选择性聚类融合算法的核心是聚类成员的选择方法,现有的选择方法基本都是利用某项评价指标对聚类成员进行排序,然后根据排序结果对聚类成员进行选择.周志华等^[2]提出了一种基于 Bagging 的选择性聚类融合算法,这种算法在进行聚类成员选择的过程中利用平均标准互信息来表示聚类成员之间的差异度,剔除掉那些与其它聚类成员差异度比较小的聚类成员.Kuncheva 等^[3]研究了聚类成员间差异度对聚类融合结果的影响并提出了一种根据聚类成员之间的差异度对聚类成员进行选择的方法.传统选择性聚类融合算法中难以消除劣质聚类成员的干扰,聚类融合结果存在准确性不高的问题^[4].究其原因,聚类成员的选择标准普遍存在的一个问题是考虑的因素往往比较单一,在聚类成员选择的过程中只注重考察聚类成员的某一方面的指标.

本文提出了一种基于加权 JP 聚类成员选择方法的聚类融合算法.在多个不同数据集上的实验结果表明,同时考虑聚类成员的精度和聚类成员之间差异度的效果要好于只考虑聚类成员的某一项指标,利用这种聚类成员选择方法有效地改善聚类融合结果的准确性和鲁棒性,使得基于加权 JP 聚类成员选择方法的聚类融合算法性能要优于传统的聚类融合算法以及传统的聚类算法.

2 基于加权 JP 聚类成员选择方法的聚类融合算法

2.1 加权 JP 聚类成员选择方法

新的基于加权 JP 聚类成员选择方法综合考虑聚类成员的精度和差异度对聚类融合结果的影响,其中“J”表示 Jaccard index,“P”表示 Precision.该方法充分利用加权的精度和差异度评价指标对聚类成员进行排序,然后根据排序结果对聚类成员进行选择.相关定义如下:

定义 1 聚类成员平均差异度定义为:

$$D_i = \frac{1}{r-1} \sum_{m=1, m \neq i}^r [1 - J(\lambda_i, \lambda_m)] \quad (1)$$

式中:设有 r 个聚类成员 $\lambda_1, \lambda_2, \dots, \lambda_r$, D_i 表示聚类成员 λ_i 和其它聚类成员之间的平均差异度. $J(\lambda_i, \lambda_m)$ 表示聚类成员 λ_i 和 λ_m 的 Jaccard index 值; $1 - J(\lambda_i, \lambda_m)$ 表示聚类成员 λ_i 和 λ_m 的差异度, Jaccard index 的值越小,两个聚类成员间的差异度越大.选择 Jaccard index 作为聚类成员差异度的度

量方法的原因是:Jaccard index 对差异度的敏感度比较高,用 Jaccard index 得到的结果区分度大, Jaccard index 的这些特性有利于聚类成员的选择.

在聚类成员的选择过程中计算精度时需要实际的类标号,但是在聚类问题中往往不知道数据的类标号,因此必须通过别的方法来估计聚类成员的精度值. Levine^[5]提出了一种通过可重复抽样技术来估计聚类成员精度的方法,通过可重复抽样技术得到关于数据集 X 的 r 个数据子集 $T_1, T_2, \dots, T_r$, 在每个数据子集上运行聚类算法,得到 r 个聚类结果 $\lambda_1, \lambda_2, \dots, \lambda_r$, 第 i 个聚类结果的精度可以通过式 (2) 来进行估计.

$$P_i = \frac{\sum_{k=1}^r \text{sim}(\lambda_i, \lambda_k)_{T_k}}{r}, \quad i = 1, 2, \dots, r \quad (2)$$

式中: $\text{sim}(\lambda_i, \lambda_k)_{T_k}$ 表示在数据子集 T_k 上的聚类结果 λ_k 和 λ_i 的相似度.

在聚类融合算法中,可以将聚类成员当做在数据子集上产生的聚类结果 $\lambda_1, \lambda_2, \dots, \lambda_r$, 然后用同样的方法估计每个聚类成员的精度.

定义 2 聚类成员精度定义为:

$$P_i = \frac{\sum_{k=1}^r \text{sim}(\lambda_i, \lambda_k)}{r}, \quad i = 1, 2, \dots, r \quad (3)$$

式中: P_i 表示聚类成员 λ_i 的精度.

定义 3 加权 JP 聚类成员选择评价指标定义为:

$$\mathbf{W} = (1 - \theta) * \mathbf{D} + \theta * \mathbf{P} \quad (4)$$

式中: $\mathbf{W}$ 是一个长度为 r 的向量, $\mathbf{W}$ 的第 i 个元素表示聚类成员 λ_i 的 P_i 与 D_i 的加权和; θ 表示在聚类成员的选择方法中精度占的权值; $\mathbf{D}$ 是一个长度为 r 的向量, $\mathbf{D}$ 的第 i 个元素表示聚类成员 λ_i 与其它聚类成员之间的平均差异度; $\mathbf{P}$ 是一个长度为 r 的向量, $\mathbf{P}$ 的第 i 个元素表示聚类成员 λ_i 的估计精度.

2.2 基于加权 JP 聚类成员选择方法的聚类融合算法

本文对传统的聚类融合算法进行改进^[6], 提出了基于 JP 聚类成员选择方法的聚类融合算法,具体工作基于以下两方面进行:

(1) 修改聚类成员的产生方式

聚类成员的产生阶段,本文采用了一种混合式的聚类成员产生方法,具体方法如下:

首先利用特征抽取方法对原有数据集进行抽取得到若干个特征子集,然后交替使用 k-means 算法、Single-link 算法、Complete-link 算法在这些特征子集上产生聚类成员.

这种聚类成员的产生方法综合考虑了若干种影响聚类成员之间的差异度的因素,可以使得聚类成员之间保持一定的差异度。

(2)添加新的聚类成员选择方法

得到聚类成员之后,利用基于加权 JP 聚类成员选择方法对聚类成员进行选择,即综合考虑聚类成员的精度和差异度对聚类融合结果的影响。

该方法充分利用加权的精度和差异度评价指标对聚类成员进行排序,然后根据排序结果对聚类成员进行选择,保证只让一部分“质量”好的聚类成员参与到聚类融合的过程中。

基于 JP 聚类成员选择方法的聚类融合算法描述如下:

输入: X, θ, σ

输出: λ

//对数据集 X 进行特征抽取得到若干个特征子集 X_i , 交替使用不同聚类算法在这些特征子集 X_i 上产生聚类成员 λ_i

For $i=1:r$ // r 为聚类成员数

$X_i = \text{FE}(X)$

$\lambda_i = \text{clusterdata}(X_i)$

End

//聚类结果 $\lambda_1, \lambda_2, \dots, \lambda_r$ 随机排列

$\lambda_{\text{base}} = \text{random}(\lambda_1, \lambda_2, \dots, \lambda_r)$

//使用解决聚类成员一致性问题的 max-agreement 算法^[7]

For each λ_i in $\{\lambda_1, \lambda_2, \dots, \lambda_r\}$

$\text{max-agreement}(\lambda_{\text{base}}, \lambda_i)$

End

//按照给定加权参数 θ 值计算聚类成员 λ_i 的精度 P_i 和平均差异度 D_i 的加权和 W_i

For each λ_i in $\{\lambda_1, \lambda_2, \dots, \lambda_r\}$

$D_i = \frac{1}{r-1} \sum_{m=1, m \neq i}^r [1 - J(\lambda_i, \lambda_m)]$

$P_i = \frac{\sum_{k=1}^r \text{sim}(\lambda_i, \lambda_k)}{r}$

$W_i = (1-\theta) * D_i + \theta * P_i$

End

//按照给定比例参数 σ 值选择聚类成员

$\{\lambda_1, \lambda_2, \dots, \lambda_m\} = \text{selectclustering}(W, \sigma)$

//聚类融合产生最终聚类结果

$\lambda = \text{Ensemble}(\lambda_1, \lambda_2, \dots, \lambda_m)$

基于 JP 聚类成员选择方法的聚类融合算法采用新的聚类成员选择方法,在聚类成员的选择过程中充分考虑了聚类成员的精度以及聚类成员之间的差异度,可以有效地改善聚类融合的结果。

3 实验结果与分析

3.1 聚类结果评价标准

为了衡量算法正确性,实验中选用的聚类评价标准是经典的 micro-p^[4] 指标. 计算如下:

$$\text{micro-p} = \frac{1}{N} \sum_{i=1}^k n_i \tag{5}$$

式中: k 表示簇的个数, n_i 表示第 i 个簇中划分正确的样本的数目, N 表示数据集中的样本总数。

micro-p 是一种有监督的评价指标, micro-p 值表示了聚类结果中被划分到正确簇的样本数目占整个数据集样本总数的比例. micro-p 的值越高,说明聚类结果越准确。

3.2 实验数据集

为了对基于加权 JP 聚类成员选择方法的聚类融合算法的鲁棒性进行评价,实验选取了 UCI 公共数据集中不同类型的 5 个数据集:iris、wine、SAT、ionosphere、waveform, 5 个数据集簇的规模、形状、个数以及簇之间的边界的清晰度都不尽相同. 通过利用 PCA 技术将数据从高维空间映射到低维空间然后运用可视化技术描绘出各个数据集的散点图,如图 1 所示. 实验选用的数据集的簇的规模各异,如数据集 waveform 和数据集 iris;簇的形状各异,如数据集 SAT 和数据集 ionosphere;大部分数据集簇间的边界都很模糊. 表 1 描述了这些数据集的具体信息。

表 1 实验所用数据集

数据集名称	样本数目	属性数目	类的个数
iris	150	5	3
wine	178	14	3
SAT(Satellite)	2 000	37	7
ionosphere	351	35	2
waveform	5 000	41	2

3.3 实验结果分析

本实验选取 5 个传统的聚类融合算法、k-means 算法、2 个层次聚类算法 Single-link 和 Complete-link 作为实验参照组,与本文提出的基于 JP 聚类成员选择方法的聚类融合算法进行比较. 为叙述方便,将加入基于 JP 聚类成员选择方法的修改后聚类融合算法命名为: JPSel-w-vote、JPSel-CSPA、JPSel-HGPA、JPSel-MCLA、JPSel-Voting-k-means. 为了实验结果的可靠性,各算法重复十次,基于 JP 聚类成员选择方法聚类融合算法的参数 σ 和 θ 随机取值. 表 2 描述了相关聚类结果的平均 micro-p 值。

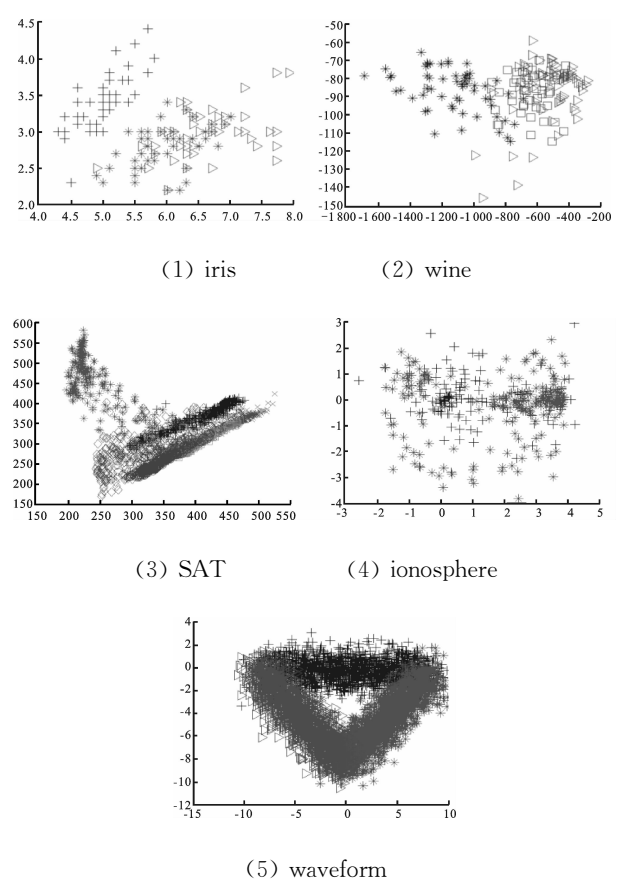


图 1 各个数据集的散点图

表 2 不同聚类算法的准确性比较

	iris	wine	SAT	ionosphere	waveform
JPSel-w-vote	0.90	0.72	0.72	0.78	0.59
w-vote	0.86	0.70	0.61	0.72	0.52
JPSel-CSPA	0.83	0.68	0.79	0.73	0.54
CSPA	0.81	0.65	0.74	0.70	0.50
JPSel-HGPA	0.65	0.60	0.59	0.56	0.53
HGPA	0.62	0.58	0.55	0.50	0.48
JPSel-MCLA	0.90	0.71	0.75	0.75	0.57
MCLA	0.86	0.70	0.70	0.72	0.52
JPSel-Voting-k-means	0.85	0.69	0.59	0.68	0.56
Voting-k-means	0.76	0.62	0.49	0.62	0.47
k-means	0.78	0.59	0.52	0.70	0.49
Single-link	0.68	0.42	0.33	0.64	0.40
Complete-link	0.80	0.60	0.48	0.68	0.53

从表 2 可以看出,选择性聚类融合算法 JPSel-w-vote、JPSel-CSPA、JPSel-MCLA、JPSel-Voting-

k-means 在 5 个不同类型的数据集上平均 micro-p 值要优于传统的聚类融合算法以及传统的聚类算法 k-means、Single-link、Complete-link, 算法 JPSel-HGPA 的平均 micro-p 值要优于聚类融合算法 HGPA. 这样的实验结果也充分说明基于 JP 聚类成员选择方法的聚类融合算法综合考虑聚类成员的精度和差异度对聚类融合结果的影响,充分利用加权的精度和差异度评价指标对聚类成员进行排序,然后根据排序结果对聚类成员进行选择,能够选择出“质量”比较好的聚类成员,从而可以有效地改善聚类融合算法结果的准确性和鲁棒性,性能显著提高.

加权参数 θ 表示在聚类成员的选择方法中精度占的权值,在验证参数 θ 对实验结果的影响时,设定聚类成员的选择比例 $\sigma=0.5$, θ 的取值为 0, 0.1, 0.2, ..., 1. 图 2 描述了随着 θ 的变化,不同的选择性聚类融合算法在不同的数据集上的 micro-p 值的变化情况. 可以发现,当 θ 值比较适中时,选择性聚类融合算法的 micro-p 值比较高. 当 θ 值比较大或者比较小时,选择性聚类融合算法的 micro-p 值比较低.

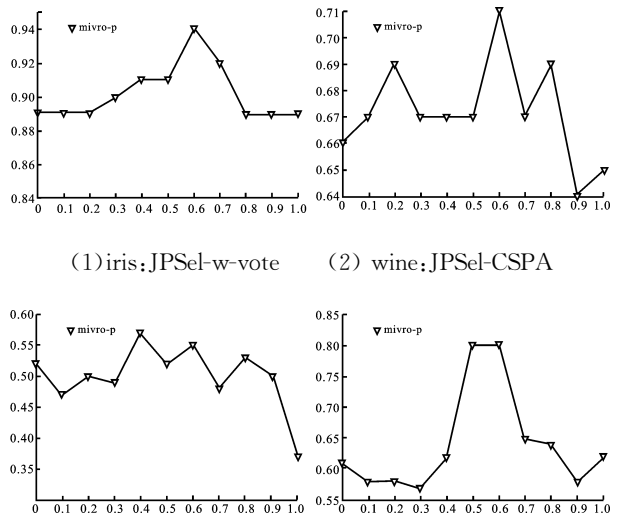
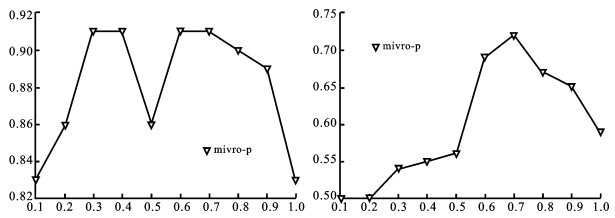


图 2 聚类成员选择方法中加权参数 θ 对实验结果的影响

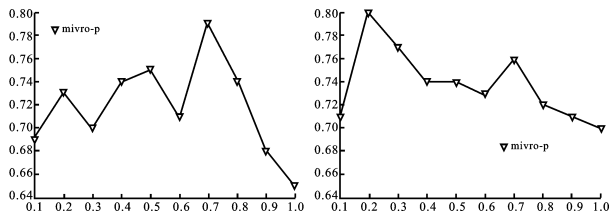
实验结果表明,在聚类成员的选择过程中,选择方法中参数 θ 设定一个比较适中的值有利于提高选择性聚类融合算法结果的准确度. 在聚类成员的选择过程中,如果只关注聚类成员的差异度,即 θ 值比较小,那么精度比较差的聚类成员就会对聚类融合的结果产生影响. 如果只关注聚类成员的精度,即 θ 值比较大,可能由于聚类成员之间的差异度不够,聚类融合的效果也不会太好. 在聚类成员选择的过程

中从聚类成员的精度和聚类成员之间的差异度两方面对聚类成员进行衡量,要比单纯地考虑聚类成员的某一个指标的结果要好。

参数 σ 表示被选择的聚类成员占有所有聚类成员的比例,在聚类成员选择的过程中, σ 会对选择性聚类融合算法的结果产生影响. 在实验过程中, 设定 $\theta = 0.5$, σ 的取值为 $0.1, 0.2, \dots, 1$. 当 $\sigma = 1$ 时, 此时没有对聚类成员进行选择, 即是传统的聚类融合算法. 图 3 描述了聚类成员的选择比例参数 σ 对实验结果的影响. 当 σ 的取值适中时, 各种选择性聚类融合算法的结果比较好.



(1) iris; JPSel-CSPA (2) wine; JPSel-Voting-k-means



(3) SAT; JPSel-MCLA (4) ionosphere; JPSel-w-vote

图 3 聚类成员选择比例参数 σ 对实验结果的影响

实验结果说明,在聚类成员选择的过程中,设定一个比较适中的 σ 值有利于提高选择性聚类融合算法的精度. 当 σ 的值偏小时,选择的聚类成员数目比较少,不能体现出聚类融合算法的优势. 当 σ 的值偏大时,选择的聚类成员数目又过多,一些质量不好的聚类成员就会参与到聚类融合过程中,影响聚类融合结果.

4 结束语

聚类分析是一种常用的数据分析方法,它可以根据数据中点与点之间的关系自动地对数据进行组织,然而并没有一种通用的聚类算法可以处理所有类型的数据. 聚类融合算法是聚类分析领域的一个研究热点,它可以显著地提高聚类算法的正确性和鲁棒性. 传统的聚类融合算法中对聚类成员进行选择的方法存在不能消除劣质聚类成员干扰的问题,一些“质量”比较差的聚类成员直接影响聚类融合结

果的准确性. 本文根据影响聚类融合结果的两个因素,提出了一种新的聚类成员选择方法——基于加权 JP 的聚类成员选择方法,以及基于该方法的聚类融合算法. 这种选择方法在对聚类成员进行选择时充分考虑了聚类成员两方面的因素:聚类成员的精度和聚类成员之间的差异度. 实验结果表明,基于新的聚类成员选择方法的聚类融合算法性能要优于传统的聚类融合算法以及传统的聚类算法;在聚类成员选择过程中,同时考虑聚类成员的精度和聚类成员之间的差异度的效果要好于只考虑聚类成员的其中一项指标. 对聚类成员的选择方法可以进一步的优化,本文提出的基于 JP 的聚类成员选择方法只在聚类成员的精度和聚类成员之间的差异度上进行了考察,是否存在影响聚类融合结果的其它因素值得进一步的研究.

参考文献:

[1] Jiawei Han, Micheline Kamber, Jian Pei. Data mining: concept and techniques[M]. 3rd ed. Beijing: China Machine Press, 2012.

[2] 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.

[3] Hadjitodorov S T, Kuncheva L I, Todorova L P. Moderate diversity for better cluster ensembles [J]. Information Fusion, 2006, 7(3): 264-275.

[4] 刘丽敏, 樊晓平, 廖志芳. 选择性聚类融合新方法研究[J]. 计算机应用研究, 2012, 29(11): 1-5.

[5] Levine E, Domany E. Resampling method for unsupervised estimation of cluster validity[J]. Neural Comput, 2008, 13(11): 2573-2593.

[6] Fischer B, Buhmann J M. Path-based clustering for grouping of smooth curves and texture segmentation [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2003, 25(4): 513-518.

[7] Constantinos Boulis, Mari Ostendorf. Combining multiple clustering systems [M]. Berlin Heidelberg: Springer, 2004: 63-74.

作者简介:

冯中慧 女, (1977-), 博士, 讲师. 研究方向为数据挖掘、自然语言处理.

何 亮(通讯作者) 男, (1975-), 博士, 讲师. 研究方向为数据挖掘、机器学习. E-mail: lhe@mail. xut. edu. cn.

王 栋 男, (1988-), 硕士. 研究方向为数据挖掘.