DOI:10.7544/issn1000-1239.2018.20180575
55(12):2611-2619,2018

基于群体智慧的簇连接聚类集成算法

张恒山[1,2] 高宇坤[1] 陈彦萍[1,2] 王忠民[1,2]
[1](西安邮电大学计算机学院 西安 710121)
[2](陕西省网络数据分析与智能处理重点实验室(西安邮电大学) 西安 710121)
(hengshzhang@foxmail.com)

Clustering Ensemble Algorithm with Cluster Connection Based on Wisdom of Crowds

Zhang Hengshan[1,2], Gao Yukun[1], Chen Yanping[1,2], and Wang Zhongmin[1,2]
[1](*School of Computer Science & Technology, Xi'an University of Posts and Telecommunications, Xi'an* 710121)
[2](*Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing (Xi'an University of Posts and Telecommunications), Xi'an* 710121)

Abstract The accuracy and stability of clustering will be obviously improved when a lot of independent clustering results for the same data set are aggregated by utilizing the principle of wisdom of crowds. In this paper, clustering ensemble algorithm with cluster connection based on wisdom of crowds (CECWOC) is proposed. Firstly, the independent clustering results are produced by the different clustering algorithms, which is guided by utilizing the independency, decentralization, diversity of wisdom of crowds. Secondly, the clustering ensemble algorithm based on connecting triple is developed to grouping aggregate the produced independent clusters, and the obtained results are aggregated again and the final cluster set is produced. The advantages of proposed algorithm are that: 1) The produced clusters by base clustering is grouping aggregated and weights of clusters are adjusted so that the selection of clusters is avoided, as a result, information on the produced clusters are not ignored; 2) Similarities of data are computed by using connected triple algorithm, the relations of data that their similarities are zero can be used. The experimental results at the different data sets show that the proposed algorithm can obtain the more accurate and stable results than other clustering ensemble algorithms, including the ones based on framework of wisdom of crowds.

Key words wisdom of crowds (WOC); clustering ensemble; connecting triple; clustering ensemble select (CES); data mining

摘 要 利用群体智慧原理,将多个相互独立的聚类算法的结果进行聚合,将显著提高聚类结果的准确性.基于群体智慧的簇连接聚类集成算法,首先使用群体智慧理论的独立性、分散性、多样性原则引导个体聚类结果的生成,然后提出基于连接三元组的聚类集成算法对个体聚类结果进行分组聚合,将分组聚合的结果再次进行聚合得到最终的聚类结果.该算法的优点包括:1)通过簇的分组和权重调整,避免了

收稿日期:2018-08-17;**修回日期**:2018-10-23
基金项目:国家自然科学基金项目(61373116);陕西省科技统筹创新工程基金项目(2016KTZDGY04-01)
This work was supported by the National Natural Science Foundation of China (61373116) and the Shaanxi Science and Technology Coordination and Innovation Project (2016KTZDGY04-01).
通信作者:高宇坤(821566504@qq.com)

对基聚类生成的簇进行选择,有利于充分利用已生成簇的信息;2)采用连接三元组算法计算数据之间的相似性,可以充分挖掘数据点之间的关系.对不同数据集的实验研究表明:该算法相对传统的集成聚类算法以及群体智慧与机器学习相结合的集成聚类算法,可以进一步提高集成聚类结果的准确性.

关键词 群体智慧;聚类集成;连接三元组; 聚类集成选择;数据挖掘

中图法分类号 TP399

聚类问题是机器学习领域中一个极具挑战性的研究问题.聚类分析通过计算数据对象间的相似度把数据集划分成若干个簇,使在相同簇的对象具有较高的相似度,不同簇的对象则差异较大[1].在聚类的过程中存在着许多问题:1)同种聚类算法的不同参数和初始化会影响聚类的结果;2)大部分聚类算法都很难得出数据集中真实类的数目;3)不同的聚类算法会产生不同的聚类结果.为了解决这些问题,研究者提出了聚类集成算法,通过合并不同的聚类结果得到一个鲁棒性好、稳定性高的最终结果[2].聚类集成选择(clustering ensemble select, CES)使用一致性度量来评估和选择个体聚类结果[3],通过对选择的个体聚类结果进行集成,可以提高最终结果的准确性、稳定性.一般来说,聚类集成选择包含4个组成部分:生成、评价、选择和组合.首先,通过使用不同的聚类算法或重复一种算法生成多个聚类结果,这些结果可以在每次运行时随机产生;其次,一个共识度量(如归一化互信息)来评估产生的结果;再次,通过阈值选择评估结果;最后,通过聚集机制得到最终的聚类结果[3-6].

在聚类集成选择算法中有3大问题:1)生成策略;2)度量评价;3)阈值生成.为了解决聚类集成选择算法中存在的3个问题,研究者利用了群体智慧理论来引导个体聚类的生成以及最终结果的集成,并基于此提出了一种框架——群体智慧聚类集成(wisdom of crowds ensemble, WOCE)框架[7].在该框架中,通过实现群体智慧的4个必备条件:独立性、分散性、多样性、聚集性,可以有效解决聚类集成选择算法中存在的3个问题.然而该框架存在的问题是没有充分利用数据点之间的关系,部分具有相似性的数据点之间的相似度计算为零,导致最终得到的聚类结果的准确性不足.

本文提出了一种基于群体智慧框架的簇链接聚类集成算法(clustering ensemble algorithm with cluster connection based on wisdom of crowds, CECWOC),该算法先将原始数据集进行转化,得到一个满足群体智慧标准的可使用数据集,然后利用不同聚类算法构造多样性的聚类成员,并采用连接三元组算法(connected triple algorithm, CTA)得到数据点之间的相似度矩阵.最后,对相似度矩阵使用层次聚类算法得到聚类结果.该算法充分挖掘了尽管计算得到的相似度为0但实际存在某些相似性的数据点之间的关系,提高了最终聚类结果的准确性.

1 相关算法与概念

1.1 群体智慧

群体智慧是由Russell(1983), Atlee(1993), Mayer-Kress (2003)等人与其他理论家共同描述.群体智慧对多个参与者独立给出的评价意见进行聚合,可以得到准确性好于其中任何一个参与者的评价意见.Surowiecki[8]于2006年提出将群体智慧作为做出优化决策的一种框架,他对智慧群体提出了4个标准:1)独立性.人们的观点不会被周围人的观点所影响.2)分散性.人们能够专注和利用自己的知识.3)多样性.每个人都有各自的私人信息,即使对已知事实的认知偏差较大.4)聚集性.将私人的判断转化为集体决定的机制.

1.2 聚类集成

2003年Strehl等人提出“聚类集成”(cluster ensembles, CE)的概念,并给出了定义.聚类集成是指将一个对象集合的多个划分组合成为一个统一聚类结果的算法[2].在文献[9]中作者也给出该问题的一种描述:给定一个聚类结果的集合,聚类集成CE的目标就是要寻找一个聚类,相对于所有的输入聚类结果来说,尽可能多地符合(或一致)[9].由此可见,聚类集成CE是利用多个聚类结果找到一个新的数据划分,这个划分在最大程度上共享了所有输入的聚类结果对数据集的聚类信息.聚类集成过程为:假设数据集 $\boldsymbol{X}$ 有 n 个数据对象 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,首先对数据集 $\boldsymbol{X}$ 使用 N 次聚类算法,得到 N 个聚类结果 $\boldsymbol{P}=\{\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_N\}$,其中 $\boldsymbol{p}_i(i=1,2,\cdots,N)$ 为第 i 个聚类算法得到的聚类结果.然后一

致性函数 T 对 $\boldsymbol{P}$ 中的聚类结果进行集成得到一个新的数据划分 $\boldsymbol{P}'$,将此作为最终的聚类结果.

1.3 基于连接的相似度矩阵

相似度矩阵简单易生成,但是它存在一个缺点,那就是只得到部分数据点之间的关系,对于具有弱相似度的数据点只能显示为 0.为了解决这一问题,得到更多数据点间的关系,文献[10]提出了一种基于连接的相似度矩阵构造算法:连接三元组 $\Delta=(\boldsymbol{V}_\Delta,\boldsymbol{W}_\Delta)$ 是 G 的一个子图,包含 3 个顶点 $\boldsymbol{V}_\Delta=(A_1,A_2,A_3)\subset\boldsymbol{V}$ 和 2 条边 $\boldsymbol{W}_\Delta=(e_{A_1,A_2},e_{A_1,A_3})\subset\boldsymbol{W}$,连接其他 2 个顶点的顶点称为这个三元组的中心.其基本思想是:如果 2 个节点都与第 3 个节点有连接,则认为这 2 个节点之间存在相似性.

2 基于群体智慧框架的簇连接聚类集成算法

本文提出的基于群体智慧框架的簇连接聚类集成算法 CECWOC,有效地扩充了相似度矩阵中数据点间的潜在信息,对其中相似度为 0 但实际还存在某些关联关系的数据点进行了处理,使相似度矩阵包含更多的有效信息,提高了最终结果的准确率.

2.1 数据预处理

文献[7,11]中分别使用群体智慧理论中的人群、信息、观点来代替聚类问题中的算法、数据和结果.基于群体智慧框架的定义,人们必须采用独立的信息来做决定.本质上就是通过删除原始数据特征之间的相关性来生成新的数据特征,这些数据特征之间是相互独立的.在使用聚类算法之前,有各种算法去除数据的相关性,如主成分分析或线性判别分析法等,删除特征间相关性能够显著提高聚类结果的性能[12].在群体智慧框架中,分散性标准可以增加群体的智慧,减少最终结果的误差,提高结果的准确率.本文在数据预处理阶段通过整合多种算法,可以实现数据之间的独立性和分散性.

首先,本文通过基于主成分分析的算法将数据映射到不同的维度,以使其特征之间的相关性较小.

给定一个数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,对数据集 $\boldsymbol{X}$ 求平均值:

$$\bar{\boldsymbol{X}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i, \tag{1}$$

其中,n 表示数据集 $\boldsymbol{X}$ 中数据的个数,$\boldsymbol{x}_i$ 表示数据集 $\boldsymbol{X}$ 中第 i 个数据.此时,就可以求出

$$\boldsymbol{X}'=\boldsymbol{X}-\bar{\boldsymbol{X}}=\{(\boldsymbol{x}_1-\bar{\boldsymbol{x}}_1),(\boldsymbol{x}_2-\bar{\boldsymbol{x}}_2),\cdots,(\boldsymbol{x}_n-\bar{\boldsymbol{x}}_n)\}. \tag{2}$$

定义 1. $\boldsymbol{Q},\boldsymbol{X}'\in\boldsymbol{R}^{m\times n},\boldsymbol{Y}\in\boldsymbol{R}^{m\times n}$,其中 m,n 分别表示特征的个数和数据点的数目.这个映射的目标是极小化特征间的关联性,根据主成分分析法,这个问题可以转换为

$$\boldsymbol{Y}=\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{X}'. \tag{3}$$

对于 $\boldsymbol{R}$,计算为

$$\boldsymbol{R}=\boldsymbol{E}\{\boldsymbol{X}'\boldsymbol{X}'^{\mathrm{T}}\}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i'\boldsymbol{x}_i'^{\mathrm{T}}. \tag{4}$$

其次,本文利用本地知识构造散布矩阵,通过数据转换实现数据的分散性,从而增加最终聚类结果的准确性、稳定性.在构造散布矩阵时用到的基本概念描述如下:

成对约束. Must-link 规定 2 个点必须属于同一类,即 $\boldsymbol{M}=\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\}$;Cannot-link 规定 2 个点不能在同一类中,即 $\boldsymbol{C}=\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\}$.以成对约束为基础,本文运用约束投影,即一组约束向量 $\boldsymbol{W}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_d\}$,在数据转换时,把 $\boldsymbol{M}$ 和 $\boldsymbol{C}$ 保存到被转化的低维度表示 $\boldsymbol{z}_i=\boldsymbol{W}^{\mathrm{T}}y_i$ 中.定义函数 $\boldsymbol{J}(\boldsymbol{W})$ 为

$$\boldsymbol{J}(\boldsymbol{W})=\mathrm{trace}(\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{S}_C-\gamma\boldsymbol{S}_M)\boldsymbol{W}), \tag{5}$$

其中,$\boldsymbol{S}_C$ 和 $\boldsymbol{S}_M$ 分别定义为

$$\boldsymbol{S}_C=\frac{1}{2n_{C(\boldsymbol{y}_i,\boldsymbol{y}_j)\in\boldsymbol{C}}}\sum(\boldsymbol{y}_i-\boldsymbol{y}_j)(\boldsymbol{y}_i-\boldsymbol{y}_j)^{\mathrm{T}}, \tag{6}$$

$$\boldsymbol{S}_M=\frac{1}{2n_{M(\boldsymbol{y}_i,\boldsymbol{y}_j)\in\boldsymbol{M}}}\sum(\boldsymbol{y}_i-\boldsymbol{y}_j)(\boldsymbol{y}_i-\boldsymbol{y}_j)^{\mathrm{T}}, \tag{7}$$

其中,n_C 和 n_M 分别代表 $\boldsymbol{C}$ 和 $\boldsymbol{M}$ 的基数,γ 是比例系数.γ 计算为

$$\gamma=\frac{\dfrac{1}{n_{C(\boldsymbol{y}_i,\boldsymbol{y}_j)\in\boldsymbol{C}}}\sum\|\boldsymbol{y}_i-\boldsymbol{y}_j\|^2}{\dfrac{1}{n_{M(\boldsymbol{y}_i,\boldsymbol{y}_j)\in\boldsymbol{M}}}\sum\|\boldsymbol{y}_i-\boldsymbol{y}_j\|^2}. \tag{8}$$

将 $\boldsymbol{S}_C$ 和 $\boldsymbol{S}_M$ 称为不连散布矩阵和必连散布矩阵,与线性判别分析法[12]中簇间散布矩阵和簇内散布矩阵的概念相似.区别在于后者使用簇标签生成散布矩阵,而前者使用成对约束生成散布矩阵.式(5)表达的问题可以通过计算 $\boldsymbol{S}_C-\gamma\boldsymbol{S}_M$ 对应特征值的特征向量来有效地解决.分别记 $\boldsymbol{Z}=\{\zeta_1,\zeta_2,\cdots,\zeta_p,\cdots,\zeta_d\}$ 和 $\overline{\boldsymbol{W}}$ 为 $\boldsymbol{S}_C-\gamma\boldsymbol{S}_M$ 的特征值和特征向量.对 ζ 做降序排序($\zeta_1\geqslant\zeta_2\geqslant\cdots\geqslant\zeta_p\geqslant0\geqslant\cdots\geqslant\zeta_d$),选取大于 0 的特征值所对应的特征向量,记为 $\boldsymbol{W}$.至此,转换的待使用数据集可计算为

$$\boldsymbol{Z}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{Y}. \tag{9}$$

综上所述,本文中采用的数据预处理算法流程为:

算法 1. 数据预处理.

输入:数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$、特征数 d、Must-link 集 $\boldsymbol{M}$、Cannot-link 集 $\boldsymbol{C}$;

输出:待使用数据集 $\boldsymbol{Z}$.

① 使用式(1)计算平均值 $\bar{\boldsymbol{X}}$;

② 使用式(2)计算 $\boldsymbol{X}'$;

③ 生成 $\boldsymbol{R}=\boldsymbol{E}\{\boldsymbol{X}'\boldsymbol{X}'^{\mathrm{T}}\}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i'\boldsymbol{x}_i'^{\mathrm{T}}$;

④ 计算 $\boldsymbol{R}$ 的特征值 Λ 和特征向量 $\boldsymbol{Q}$,并基于特征值对特征向量做降序排序;

⑤ $\boldsymbol{Y}=\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{X}'$;

⑥ 若 $\boldsymbol{M}$ 和 $\boldsymbol{C}$ 为空集时,$\boldsymbol{Z}=\boldsymbol{Y}$;否则使用式(6)~(8)计算 $\boldsymbol{S}_C,\boldsymbol{S}_M,\gamma$;

⑦ 计算 $\boldsymbol{S}_C-\gamma\boldsymbol{S}_M$ 的特征值 ζ 和特征向量 $\bar{\boldsymbol{W}}$;

⑧ 通过 $\zeta\geqslant 0$,得到 $\boldsymbol{W}$;

⑨ $\boldsymbol{Z}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{Y}$.

2.2 基聚类结果的生成

事实上,多样性是群体智慧理论和聚类集成选择算法中共有的概念.多样性增加了最终结果的准确性和稳定性.生成多样的基聚类结果的方法有很多,如给同一种聚类算法取不同的参数;使用多种不同的聚类算法;选取数据集的不同子集进行聚类等.本文利用多种聚类算法生成基聚类结果,将生成的结果表示为一个参考集:$\boldsymbol{E}=\{\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_{i-1},\boldsymbol{P}_i,\boldsymbol{P}_{i+1},\cdots,\boldsymbol{P}_T\}$,其中 T 代表个体聚类结果的数目,$\boldsymbol{P}_i$ 表示生成结果中的第 i 个分区(第 i 个基聚类结果).

本文运用均匀性[13]来表示分区 $\boldsymbol{P}$ 与参考集中所有分区的多样性:

$$U(\boldsymbol{P},\boldsymbol{E})=1-\frac{-2\eta(\boldsymbol{P})}{\xi(\boldsymbol{P})+\Theta(\boldsymbol{P},\boldsymbol{E})},\tag{10}$$

其中,$\boldsymbol{E}$ 是参考集,$\boldsymbol{P}$ 是参考集中的分区,

$$\eta(\boldsymbol{P})=\max_{C_i\in P}\left(n_i\ln\left(\frac{n}{n_i}\right)\right),$$

$$\xi(\boldsymbol{P})=\max_{C_i\in P}\left(n_i\ln\left(\frac{n_i}{n}\right)\right),$$

$$\Theta(\boldsymbol{P},\boldsymbol{E})=\max_{P_i\in E}\left(\max_{C_j\in P_i}n_i^j\ln\left(\frac{n_i^j}{n}\right)\right),$$

$\boldsymbol{C}_i$ 表示分区 $\boldsymbol{P}$ 中第 i 个簇,n 表示分区 $\boldsymbol{P}$ 中样本的基数,n_i 表示簇 $\boldsymbol{C}_i$ 中样本的基数,$\boldsymbol{C}_j$ 表示分区 $\boldsymbol{P}_i$ 中第 j 个簇,n_i^j 表示簇 $\boldsymbol{C}_j$ 中样本的基数.

对于生成的基聚类结果,本文对其进行了分组处理,分组的原理是分别计算每个聚类结果与其他聚类结果的相似度 C_{ij}:

$$C_{ij}=\frac{|\boldsymbol{P}_i\cap\boldsymbol{P}_j|}{|\boldsymbol{P}_i\cup\boldsymbol{P}_j|},$$

其中,$\boldsymbol{P}_i,\boldsymbol{P}_j\in\boldsymbol{E}$.

根据相似度生成相似度矩阵 $\boldsymbol{C}$:

$$\boldsymbol{C}=\begin{pmatrix}C_{1,1}&\cdots&C_{1,20}\\ \vdots&\ddots&\vdots\\ 0&\cdots&C_{20,20}\end{pmatrix}.$$

然后使用 K-means 算法把基聚类结果分为若干组:

$$\boldsymbol{G}=\{\boldsymbol{G}_1,\boldsymbol{G}_2,\cdots,\boldsymbol{G}_i\},\tag{11}$$

其中,i 代表分组的数目.根据文献[14]关于群体智慧中群体分组的基本结论,一般将群体随机分为4~5组,分组的差异对最终结果影响甚微.因而在实际应用中,建议利用 K-means 算法随机将基聚类结果分为5组进行分组聚合.

2.3 基于连接三元组的聚类集成算法

聚类集成的核心问题之一是如何根据这些由聚类成员得到的聚类结果构造数据点之间的相似度矩阵.本文选用能得到数据点之间更多相似性信息的连接三元组算法[10]和群体智慧框架[13]来构造数据点之间的相似度矩阵.

在聚类问题上基于连接三元组的相似度矩阵的应用如图1所示,黑心圆代表数据点,虚线圆代表不同的分区,空的正方形代表分区中的簇,带影印的正方形表示通过连接三元组具有相似性的簇.

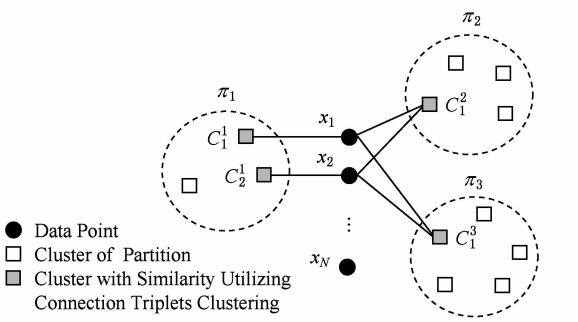

Fig. 1 Connection triplet clustering diagram

图1 连接三元组示意图

如果数据点 $\boldsymbol{x}_i\in\boldsymbol{C}_j^k$,则点与簇之间会有一条边.对于分区 π_2 和 π_3 来说,簇 $\boldsymbol{C}_1^2$ 和簇 $\boldsymbol{C}_1^3$ 都包含点 $\boldsymbol{x}_1$ 和点 $\boldsymbol{x}_2$,所以可以认为这2个点具有相似性.而对于分区 π_1 来说,点 $\boldsymbol{x}_1$ 和点 $\boldsymbol{x}_2$ 分别属于簇 $\boldsymbol{C}_1^1$ 和簇 $\boldsymbol{C}_2^1$,理论上它们是不相似的,但是这2个簇如果有相似性,则这两个点之间也应该存在相似性.根据连接三元组的算法,由于簇 $\boldsymbol{C}_1^1$ 和簇 $\boldsymbol{C}_2^1$ 具有2个连接三元组,并且簇 $\boldsymbol{C}_1^2$ 和簇 $\boldsymbol{C}_1^3$ 分别是这2个连接三元组的中心,因此,簇 $\boldsymbol{C}_1^1$ 和簇 $\boldsymbol{C}_2^1$ 是相似的,从而对于分区 π_1 来说,$\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 也是相似的,只是相似度较低.可见该算法扩充了数据点之间的相似性信息,有利于具有复杂结构的数据聚类.

连接簇 $\boldsymbol{C}_i$ 和簇 $\boldsymbol{C}_j$ 的边的权重 W_{ij} 由这 2 个簇共同包含的数据点个数得到:

$$W_{ij}=\frac{|\boldsymbol{x}_i\cap\boldsymbol{x}_j|}{|\boldsymbol{x}_i\cup\boldsymbol{x}_j|}, \tag{12}$$

其中,$\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 分别为属于簇 $\boldsymbol{C}_i$ 和簇 $\boldsymbol{C}_j$ 的数据点的集合.邻接点为簇 $\boldsymbol{C}_k$ 的 2 个簇 $\boldsymbol{C}_i$,$\boldsymbol{C}_j$ 之间连接三元组的值为

$$WCT_{ij}^{k}=\min(w_{ik},w_{jk}). \tag{13}$$

簇 $\boldsymbol{C}_i$,$\boldsymbol{C}_j$ 之间所有的三元组$(1,2,\cdots,q)$可以计算为

$$WCT_{ij}=\sum_{k=1}^{q}WCT_{ij}^{k}. \tag{14}$$

簇 $\boldsymbol{C}_i$,$\boldsymbol{C}_j$ 之间的相似度可以计算为

$$Sim(i,j)=\frac{WCT_{ij}}{WCT_{\max}}\times DC, \tag{15}$$

其中,$WCT_{\max}$是任何 2 个簇 WCT 中最大的值,而 DC 是一个衰减因子.此时数据点 $\boldsymbol{x}_i$,$\boldsymbol{x}_j$ 之间的相似度为

$$S_m(\boldsymbol{x}_i,\boldsymbol{x}_j)=\begin{cases}1, & C(\boldsymbol{x}_i)=C(\boldsymbol{x}_j),\\ Sim(i,j), & C(\boldsymbol{x}_i)\neq C(\boldsymbol{x}_j).\end{cases} \tag{16}$$

WCTE 矩阵的计算为

$$S(\boldsymbol{x}_i,\boldsymbol{x}_j)=\begin{cases}\dfrac{1}{M}\sum N_{ij}*\rho_{ij}, & C(\boldsymbol{x}_i)=C(\boldsymbol{x}_j),\\ \dfrac{1}{M}\sum_{m=1}^{M}S_m(\boldsymbol{x}_i,\boldsymbol{x}_j), & C(\boldsymbol{x}_i)\neq C(\boldsymbol{x}_j),\end{cases} \tag{17}$$

其中,M 为个体聚类结果的个数,N_{ij} 表示样本 i 和样本 j 在 M 种划分中属于同一个簇时值为 1,ρ_{ij} 为权重,本文将 2 种聚类算法的平均均匀性[13]当作权重系数:

$$\rho_{ij}=\frac{1}{2}(U(\boldsymbol{P}_i,\boldsymbol{E})+U(\boldsymbol{P}_j,\boldsymbol{E})). \tag{18}$$

当 2 种聚类算法均具有较高的均匀性时就生成了有效的结果,同时当 2 种聚类算法在均匀性度量中值较小时,对生成结果的影响接近于 0.因此,本文使用这种忽略低质量个体结果影响的机制代替通过生成阈值进行选择.相对于传统的互关联矩阵来说,本文中相似度矩阵的计算考虑了当 2 个数据点不属于同一个簇时,它们之间的相似度,扩充了数据点之间的潜在信息,有利于结构复杂的数据聚类.

综上所述,CECWOC 算法流程如下:

算法 2. CECWOC 算法.

输入:数据集 $\boldsymbol{Z}$;

输出:最终聚类结果 T.

① 用不同的聚类算法对数据集 $\boldsymbol{Z}$ 进行聚类,聚类的结果放入一个参考集 $\boldsymbol{E}$ 中;

② 对参考集 $\boldsymbol{E}$ 进行分组,得到 $\boldsymbol{G}$;

③ 对$\boldsymbol{G}$ 中的每组成员通过使用式(12)~(17)得到各自的 WCTE 矩阵;

④ 对 $\boldsymbol{G}$ 中分组成员的 $\boldsymbol{S}$ 矩阵使用 Average-Linkage 算法进行聚类,得到结果 $\boldsymbol{P}$;

⑤ 使用协相关矩阵对 $\boldsymbol{P}$ 进行整合,得到矩阵 $\boldsymbol{C}$;

⑥ 对矩阵 $\boldsymbol{C}$ 使用 Average-Linkage 算法进行最终聚类,得到结果 T.

算法的流程如图 2 所示.在本算法中,层次聚类的收敛时机一般设为达到某一聚类簇数量.在具体应用中,应根据应用场景尽可能客观地确定聚类簇的数量.在本文的实验中,因为试验使用的数据集为标准数据集,已经包含对簇数量的说明.

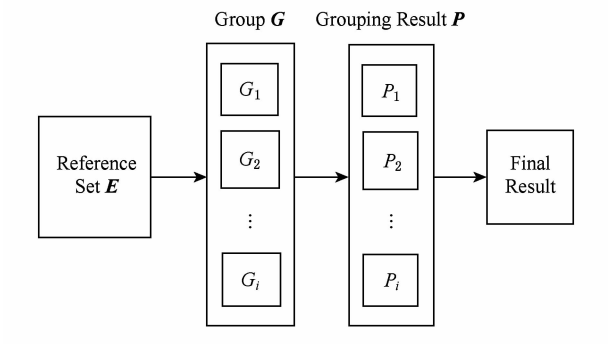

Fig. 2 CECWOC algorithm flow chart

图 2 CECWOC 算法流程图

3 实 验

在本文中,我们利用标准数据集和它们的真实类比较了该算法与其他个体聚类算法和聚类集成选择算法的性能.当然,监督信息将根据真实的类标签生成.所有的算法将在 MATLAB R2016a 中实现.首先对算法进行 10 次独立运算,然后将得到的聚类结果取平均值,作为该算法的最终结果.表 1 是实验中用来产生基聚类的个体聚类算法,实验中选用的数据集为 UCI(university of california-irvine)数据库中的真实数据集.

3.1 UCI 数据集上无监督信息的聚类实验

表 2 给出了实验中选用的 UCI 数据库中的真实数据集.在这 8 个数据集上,分别采用无监督信息的聚类集成算法:EAC[15],WPCK[16],GKPC[17],HCSS[18],GP-MGLA[19],WOCE[13]和本文提出的

CECWOC 算法进行聚类集成. 其中,EAC 算法作为对比的基础算法,WPCK,GKPC,HCSS 和 GP-MGLA 算法为 4 种效果良好的加权聚类集成算法,WOCE 算法为一种使用群体智慧框架的聚类集成算法. 通过计算准确率[12](最终聚类结果的标签与数据集真实标签的正确率)来评价聚类集成算法的性能.

Table 1 Individual Clustering Algorithm Used in the Experiments

表 1 实验所用的个体聚类算法

No.	Algorithm
1	*K*-means
2	Fuzzy *C*-means
3	Median *K*-flats
4	Gaussian mixture
5	Subtract Clustering
6	Single-linkage Euclidean
7	Single-linkage cosine
8	Single-linkage hamming
9	Complete-linkage Euclidean
10	Complete-linkage cosine
11	Complete -linkage hamming
12	Ward-linkage Euclidean
13	Ward-linkage cosine
14	Ward-linkage hamming
15	Average-linkage Euclidean
16	Average-linkage cosine
17	Average-linkage hamming
18	Spectral using a sparse similarity matrix
19	Spectral using Nystrom method with orthogonalization
20	Spectral using Nystrom method without orthogonalization

实验得到的结果如图 3 所示. 由于在 EAC 算法中,并没有使用群体智慧理论的 4 个标准,所以这是一个未利用群体智慧框架的例子,WPCK,GKPC,HCSS 和 GP-MGLA 为 4 种加权聚类集成算法,WoCE 算法使用了群体智慧框架且自身算法中包含加权算法的设计与使用,本文提出的 CECWOC 算法是在 WoCE 算法上的改进,添加了分组与连接的思想. 不难看出在这 8 个数据集中,除了 Iris 数据集外,在其余数据集中本文提出的 CECWOC 算法都能得到最佳的聚类结果. 对 Iris 数据集,CECWOC 算法与最佳结果相差不到 2%. 在实验结果图中的 Sonar 数据集上,WoCE 算法没有跑赢 HCSS 算法,但是改进后的 CECWOC 算法展现出很好的性能. 显然将群体智慧理论应用到聚类集成中能产生更高的性能,在群体智慧框架下添加连接三元组算法进行聚类集成得到的聚类结果要优于直接使用群体智慧框架得到的聚类结果.

Table 2 UCI Data Sets Used in the Experiments Without Supervising Information

表 2 无监督信息实验中所用的 UCI 数据集

Datasets	Categories	Characteristic	Number
CNAE-9	9	857	1 080
Glass	6	10	214
Iris	3	4	150
Letters	26	16	20 000
Pendigits	10	16	10 992
Sonar	2	60	208
Statlog	7	36	6 435
Wine	2	13	178
Yeast	10	8	1 484

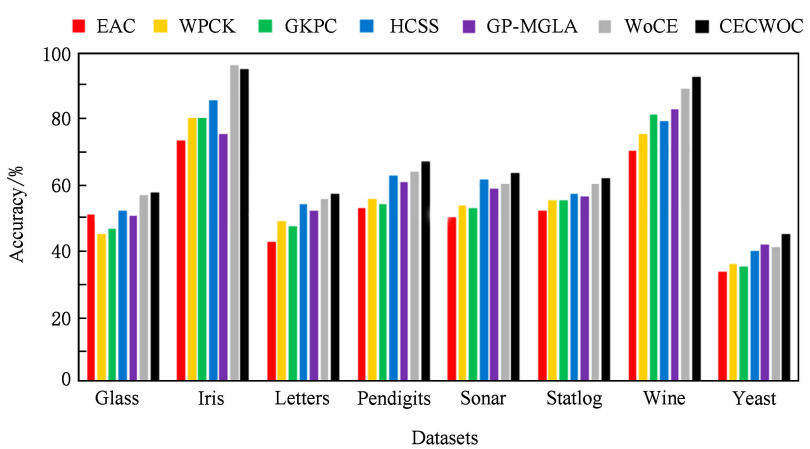

Fig. 3 Accuracy of clustering integration results for each data set

图 3 各数据集聚类集成结果的准确率

3.2 添加监督信息的聚类实验

由于大多数半监督聚类集成算法使用基于监督信息的特征选择,本文在高维度和大规模的数据集上比较半监督聚类集成算法的性能,选择表 3 中的数据集进行实验.

本文中随机选择 1%～5%样本的类标签生成监督信息(一半为 must-link,另一半为 cannot-link);例如 1%的样本有 500 个,我们就选择 250 个作为 must-link,250 个作为 cannot-link. 通过计算准确率[12](最终聚类结果的标签与数据集真实标签的正确率)来评价聚类集成算法的性能.

Table 3 Data Set Used in the Experiments with Supervising Information

表 3 有监督信息实验中所用数据集

Datasets	Categories	Features	Number
CNAE-9	9	857	1 080
Letters	26	16	20 000
Sonar	2	60	208

实验结果如图 4～6 所示. 图 4～6 分别表示在 CNAE-9,Letters 和 Sonar 数据集上使用本文提出的 CECWOC 算法与 RP[20],BGCM[21],SKMS[22],NBF[23]和 WoCE[13]算法进行比较的结果. 其中,RP 算法是一种经典的半监督聚类集成算法;BGCM 算法是一种新的基于图的半监督聚类集成算法,但是它有 2 种版本,一种是无监督的,另一种是半监督的,这里我们使用的是半监督的算法;SKMS 算法是一种基于核的半监督聚类集成算法;NBF 和 WoCE 算法都属于启发式的半监督聚类集成算法. 与不同种类的半监督聚类集成算法的比较结果显示 CECWOC 算法具有更好的性能.

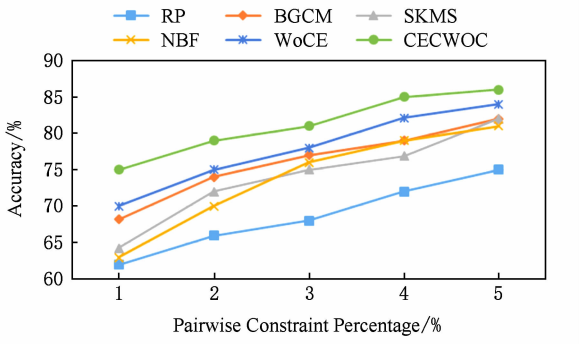

Fig. 4 Relationship between pairwise constraints and algorithms in CNAE-9

图 4 CNAE-9 中成对约束与各算法的关系

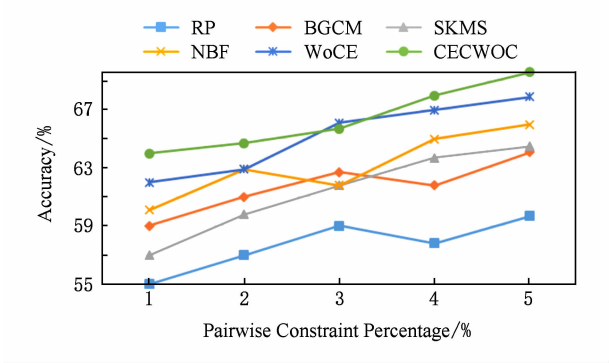

Fig. 5 Relationship between pairwise constraints and algorithms in Letters

图 5 Letters 中成对约束与各算法的关系

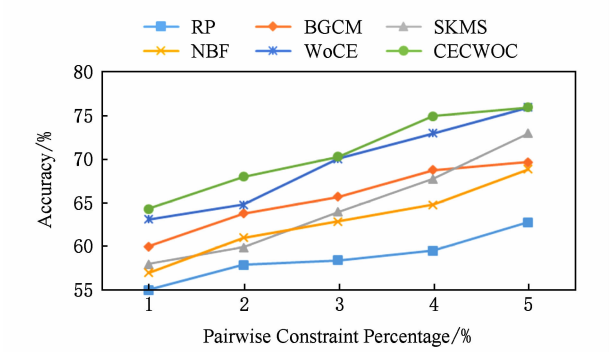

Fig. 6 Relationship between pairwise constraints and algorithms in Sonar

图 6 Sonar 中成对约束与各算法的关系

如图 5 所示,在 Letters 数据集的实验中,不难看出除了 WoCE 算法和本文提出的 CECWOC 算法外,其余算法在成对约束增加的条件下性能很不稳定,有时会下降;事实上,成对约束往往是导致高度不稳定聚类性能的原因,以前的一些算法不能处理额外的监督信息,监督信息使个体聚类结果不稳定,显著降低了上述算法的性能. 在这些情况下,本文提出采用群体智慧理论处理监督信息的算法,拥有更好的数据表示(独立性和分散性)、强大的个体聚类评价(均匀性度量)、有效集聚机制,得到了更加稳定的性能和更好的结果. 在图 4～6 中,很容易观察到,WoCE 算法和本文所提出的 CECWOC 算法在性能上是很稳定的,在群体智慧框架的基础上分组并加入连接三元组算法能有效提高准确率.

4 总 结

本文将群体智慧理论应用于聚类集成,并在此基础上提出连接三元组算法,显著提高聚类集成结

果的准确性.本文提出的算法具有的优点是:1)引入独立性和分散性的概念,用于提高基聚类结果的准确率.2)提出了一个新的框架,在满足4个标准的条件下,采用了一系列生成基聚类结果和获得最终结果的新机制.3)在映射函数中使用期望值和协方差的概念最小化特征之间的相关性来满足独立性,提出了基于成对约束的高维到低维数据的分散准则.此外,文中用一种称为均匀性的新度量来评估基聚类结果的多样性.4)提出了基聚类结果的聚集算法,通过连接三元组的方式增加关联矩阵中样本间的潜在信息,使最终结果更加准确.在未来的研究工作中,我们将结合工业大数据中数据需要聚类分析的实际应用,不断改进提出的算法,尤其需要研究本文提出的算法的并行执行效率,争取将其运用到实际的工业数据分析中去.

参 考 文 献

[1] Dong Hongbin, Teng Xuyang, Yang Xue. Feature selection based on the measurement of correlation information entropy [J]. Journal of Computer Research and Development, 2016, 53(8): 1684-1695 (in Chinese)
(董红斌, 滕旭阳, 杨雪. 一种基于关联信息熵度量的特征选择方法[J]. 计算机研究与发展, 2016, 53(8): 1684-1695)

[2] Strehl A, Ghosh J. Cluster ensembles: A knowledge reuse framework for combining partitions [J]. Journal of Machine Learning Research, 2003, 3(3): 583-617

[3] Fern X Z, Lin Wei. Cluster ensemble selection [J]. Statistical Analysis & Data Mining, 2008, 1(3): 128-141

[4] Alizadeh H, Minaei-Bidgoli B, Parvin H. Cluster ensemble selection based on a new cluster stability measure [J]. Intelligent Data Analysis, 2014, 18(3): 389-408

[5] Liu Limin, Fan Xiaoping. A new selective clustering ensemble algorithm [C] //Proc of the 9th IEEE Int Conf on E-Business Engineering. Piscataway, NJ: IEEE, 2013: 45-49

[6] Jia Jianhua, Xiao Xuan, Liu Bingxiang, et al. Bagging-based spectral clustering ensemble selection [J]. Pattern Recognition Letters, 2011, 32(10): 1456-1467

[7] Alizadeh H, Yousefnezhad M, Bidgoli B M. Wisdom of crowds cluster ensemble [J]. Intelligent Data Analysis, 2015, 19(3): 485-503

[8] Surowiecki J. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations [J]. Personnel Psychology, 2006, 59(4): 982-985

[9] Liu Qiao, Zhong Yun, Liu Yao, et al. Consistent collective entity linking algorithm [J]. Journal of Computer Research and Development, 2016, 53(8): 1696-1708 (in Chinese)
(刘峤, 钟云, 刘瑶, 等. 基于语义一致性的集成实体链接算法[J]. 计算机研究与发展, 2016, 53(8): 1696-1708)

[10] Iamon N, Boongoen T, Garrett S. Refining pairwise similarity matrix for cluster ensemble problem with cluster relations [C] //Proc Int Conf on Discovery Science. Berlin: Springer, 2008: 222-233

[11] Baker L, Ellison D. The wisdom of crowds—ensembles and modules in environmental modelling [J]. Geoderma, 2008, 147(1): 1-7

[12] Tan P N, Steinbach M, Kumar V. Introduction to data mining [J]. Intelligent Systems Reference Library, 2006, 22(6): 753-754

[13] Yousefnezhad M, Huang Shengjun, Zhang Daoqiang. WoCE: A framework for clustering ensemble by exploiting the wisdom of crowds theory [J]. IEEE Transactions on Cybernetics, 2018, 48(2): 486-499

[14] Navajas J, Niella T, Garbulsky G, et al. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds [J]. Nature Human Behavior, 2018, 2(2): 126-132

[15] Fred A L N, Jain A K. Combining multiple clusterings using evidence accumulation [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, 27(6): 835-850

[16] Vega-Pons S, Correa-Morris J, Ruiz-Shulcloper J. Weighted partition consensus via kernels [J]. Pattern Recognition, 2010, 43(8): 2712-2724

[17] Vega-Pons S, Ruiz-Shulcloper J, Guerra-Gandón A. Weighted association based methods for the combination of heterogeneous partitions [J]. Pattern Recognition Letters, 2011, 32(16): 2163-2170

[18] Yu Zhiwen, Li Le, Gao Yunjun, et al. Hybrid clustering solution selection strategy [J]. Pattern Recognition, 2014, 47(10): 3362-3375

[19] Huang Dong, Lai Jianhuang, Wang Changdong. Combining multiple clusterings viacrowd agreement estimation and multi-granularity link analysis [J]. Neurocomputing, 2015, 170(C): 240-250

[20] Ho T K. The random subspace method for constructing decision forests [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1998, 20(8): 832-844

[21] Gao Jing, Liang Feng, Fan Wei, et al. A graph-based consensus maximization approach for combining multiple supervised and unsupervised models [J]. IEEE Transactions on Knowledge & Data Engineering, 2013, 25(1): 15-28

[22] Mittal S, Tuzel O, Meer P, et al. Semi-supervised kernel mean shift clustering [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 36(6): 1201-1215

[23] Azimi J, Fern X Z. Active Learning of constraints for semi-supervised clustering [J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(1): 43-54

Zhang Hengshan, born in 1969. PhD. Lecturer at the School of Computer Science & Technology, Xi'an University of Posts & Telecommunications. His main research interests include wisdom of crowds for decision making, machine learning and data mining, and information aggregation.

Gao Yukun, born in 1992. Master candidate. His main research interests include service manning and ensemble cluster, etc.

Chen Yanping, born in 1979. PhD. Professor at the School of Computer Science & Technology, Xi'an University of Posts & Telecommunications. Her main research interests include service computing, Web services, network management, etc.

Wang Zhongmin, born in 1967. PhD. Professor at the School of Computer Science & Technology, Xi'an University of Posts & Telecommunications, and senior member of CCF. His main research interests include embedded intelligent perception, big data technology and application, intelligent information processing, etc.