DOI:10.13718/j.cnki.xsxb.2020.06.015

基于改进近邻传播算法的聚类质量评价模型①

邹臣嵩, 段桂芹, 欧阳明星, 刘 锋

广东松山职业技术学院 电气工程系,广东 韶关 512126

摘要:针对近邻传播(Affinity Propagation,简称 AP)算法在对非团状数据集聚类过程中出现的局部聚类较多、精准度不高等问题,提出了一种基于改进 AP 算法的聚类质量评价模型. 首先,在 AP 算法初步聚类的基础上,通过合并相似度较大的簇,减小聚类上限值 $k_{\max}$,进一步压缩聚类区间范围;其次,给出一个新的内部评价指标,用分属不同簇的样本对的平均距离代表簇间距离,削弱噪声数据的影响,平衡簇间分离度与簇内紧致度的关系. 在 UCI 和 KDD CUP99 数据集上的实验结果表明,新模型可以给出精准的最优聚类数(范围),能够在保持较低漏报率的同时,有效提高样本的检测率和分类正确率.

关 键 词:聚类评价指标;近邻传播;内部评价指标;最优聚类数

中图分类号:TP301.6 **文献标志码**:A **文章编号**:1000-5471(2020)06-0097-10

聚类有效性评价指标是对聚类结果进行优劣判断的依据,通过比较指标值可以确定最佳聚类划分和最优聚类数[1]. 对于无先验知识的样本来说,不同的指标所得到的最优聚类个数可能不同[2-3],哪种评价指标的 K 值更具有参考价值是机器学习领域中的热点问题,许多经典的内部聚类评价指标被先后提出,诸如 CH, IGP, DB, F_r, BWP[4]已被广泛应用,并取得了较好的效果. 但是,这些指标往往倾向于从整体结构上度量聚类结果的紧致性,而对各簇之间可能存在紧致性不一致的情况有所忽略,导致部分指标的应用范围受到一定限制[5]. 因此,如何改进或设计出更为合理的评价指标是聚类评价领域的一个重要研究方向. 此外,聚类结果评价除了和有效性指标本身有关,还与所采用的聚类算法息息相关,与 K-means[6]、K-medoids[7]算法相比,AP 算法[8]无需初始化聚类中心,具有快速、稳定等优点,尤其在处理大规模数据集时,AP 算法的性能是其他传统聚类算法所不能及的. 研究表明:AP 算法对结构复杂的数据集进行聚类时,受参数 p 的影响较大,聚类数目往往高于实际值[9-11],因此有必要进一步分析聚类数的范围及合理性.

鉴于上述指标与 AP 算法在应用中存在的不足,本文对 AP 算法进行了改进,将其聚类区间进一步压缩,同时,在对常用内部评价指标对比分析的基础上,提出了一个新的评价指标,并结合改进后的 AP 算法设计了聚类质量评价模型,采用 UCI 数据集和 KDD CUP99 数据集验证了新模型的有效性.

1 研究现状

为便于对 AP 算法、各内部评价指标和本文算法进行描述,设样本集合 $X=\{x_1, x_2, \cdots, x_i, \cdots, x_N\}$,样本总数为 N, l 表示样本的特征维度, $x_i=\{x_{i1}, x_{i2}, \cdots, x_{il}\}$,将样本集划分为 K 个簇, $X=\{C_1, C_2, \cdots, C_K\}$, n_i 为簇 C_i 的样本个数, c 为样本集的均值中心,簇中心集合 $V=\{v_1, v_2, \cdots, v_K\}$.

① 收稿日期:2018-10-06
基金项目:广东省教育科学规划课题(2018GXJK339).
作者简介:邹臣嵩(1980-),男,讲师,硕士,主要从事数据挖掘、网络安全等方面的研究.

1.1 AP算法

AP算法是一种近邻之间互传信息的聚类方法[12]，其基本思想是：首先通过计算样本对之间的相似度 s 构建整个样本集的相似度矩阵 $\boldsymbol{S}$，再以迭代的方式分别计算出吸引度 r 和归属度 a，通过判断 r 与 a 的和是否满足一定的条件来计算样本成为类代表点的可能性，以迭代次数达到预设上限或类代表点经多次迭代后不发生变化为算法终止条件[13-14]，AP算法的具体执行步骤如下所示：

步骤1：构建相似度矩阵 $\boldsymbol{S}$. 使用式(1)计算2个样本 x_i 和 x_k 间的相似度值 $s(i,k)$，进而完成样本集相似度矩阵 $\boldsymbol{S}$ 的构建.

$$s(i,k)=\begin{cases}-\|x_i-x_k\| & i\neq k\\ p(k) & i=k\end{cases} \tag{1}$$

其中，$p(k)$ 表示 x_k 被选作簇中心的倾向性，该值越大，意味着 x_k 成为簇中心的可能性越大.

步骤2：传递信息，更新吸引度 r 与归属度 a. 通过迭代循环不断地进行信息传递，交替更新吸引度与归属度值，以产生高质量的类代表. 其中，$r(i,k)$ 表示 x_k 对 x_i 的吸引度，可以理解为 x_k 适合作为 x_i 的类代表程度，该值越大，说明 x_k 作为 x_i 簇中心的几率越大；$a(i,k)$ 表示 x_i 对 x_k 的归属度，用来描述 x_i 选择 x_k 作为其类代表的适合程度，该值越大，说明 x_i 将 x_k 视为簇中心的几率越大. $r(i,k)$ 和 $a(i,k)$ 的具体更新方法如式(2)、(3)所示：

$$r(i,k)=s(i,k)-\max_{k'\neq k}\{a(i,k')+s(i,k')\} \tag{2}$$

$$a(i,k)=\begin{cases}\min\{0,\ r(k,k)+\sum\limits_{i'\notin\{i,k\}}\max[0,\ r(i',k)]\} & i\neq k\\ a(k,k)=\sum\limits_{i'\neq k}\max[0,\ r(i',k)] & i=k\end{cases} \tag{3}$$

由于在上述信息传递过程中，存在一定的振荡，因此需要引入阻尼因子 λ，通常 $\lambda\in[0,1)$，其作用是校正前后2次迭代的 $r(i,k)$ 和 $a(i,k)$，λ 越大矩阵更新速度越慢，迭代过程愈加平稳，消除震荡的效果越好，设迭代次数为 t，经校正的信息传递过程如式(4)、(5)所示：

$$r(i,k)^{t+1}=(1-\lambda)\times r(i,k)^{t+1}+\lambda\times r(i,k)^{t} \tag{4}$$

$$a(i,k)^{t+1}=(1-\lambda)\times a(i,k)^{t+1}+\lambda\times a(i,k)^{t} \tag{5}$$

其中，$r(i,k)^t$ 和 $a(i,k)^t$ 分别表示第 t 次迭代的吸引度和归属度，$r(i,k)^{t+1}$ 和 $a(i,k)^{t+1}$ 分别表示第 $t+1$ 次迭代的吸引度和归属度.

步骤3：确定类代表点. 选择满足吸引度 $r(i,k)$ 和归属度 $a(i,k)$ 之和最大的样本 x_k 作为 x_i 的类代表点，k 所满足的条件如式(6)所示：

$$k=\arg\max\{a(i,k)+r(i,k)\} \tag{6}$$

从上述公式可知，偏向参数 p 将会在式(2)计算吸引度 r 中出现，当 p 值增大时，r 和 a 也随之增大，因此候选点成为簇中心的可能性增大. 当候选点数量较多且其 p 值较大时，将有更多的候选点倾向于成为真正的簇中心，而 p 值的选取目前尚无成熟的理论依据，这样就导致了该算法大多以局部最优或近似全局最优作为最终结果[15]. 取相似度的中位数作为 p 值得到的聚类数往往大于正确类数，这虽然具备一定的参考价值，但精度依旧不足，因此需要进一步分析聚类数的合理性，以达到更理想的效果. 此外，当样本数量过大时，AP算法易出现存储空间不足或者聚类时间过长的问题，尤其对于非团状的数据集，AP算法往往产生较多的局部聚类. 为了解决这一问题，文献[16]使用 BWP 指标对AP算法得到的全部聚类结果进行评价，得出最佳聚类数，但是由于聚类上限值过大，算法耗时长的问题依旧未得到有效解决. 文献[17]将merge过程融入AP算法中，将簇间距离最小值小于整个数据集平均距离的2个簇进行合并，提高了算法的运算速度，解决了对非团状数据聚类效果不佳的问题. 但是由于该算法采用的是簇间最小距离，因此，在合并过程中存在将相似度较低的样本数据聚在一类的隐患.

1.2 内部评价指标

内部评价指标是指在不涉及任何外部信息，仅依赖数据集自身特征和度量值的条件下，通过计算簇

内、簇间或整体相似度来评价聚类效果的优劣. 理想的聚类效果是簇内紧密且簇间分离，已有内部评价指标[18-22] 的主要思想是通过簇内距离、簇间距离或样本集的平均距离的某种形式的比值来度量的，常见的内部评价指标及其特点分析如下所示：

1) *DB* 指标(Davies-Bouldin Index)[18]

$$DB(K)=\frac{1}{K}\sum_{i=1}^{K}\max_{i,\,j\neq i}\frac{\frac{1}{n_i}\sum_{x\in C_i}d(x,\,v_i)+\frac{1}{n_j}\sum_{x\in C_j}d(x,\,v_j)}{d(v_i,\,v_j)} \tag{7}$$

DB 指标首先将相邻 2 簇的簇中各样本与簇中心的平均距离之和作为簇内距离，将相邻 2 个簇中心的距离作为簇间距离，然后取二者比值的最大值作为该簇的相似度，最后对所有簇的相似度取平均值得到样本集的 *DB* 指标. 可以看出，该指标越小说明各簇间的相似度越低，对应的聚类结果越理想，该指标常用于评价“簇内紧凑，簇间远离”的数据集. 但当数据集的重叠度较大，如遇到环状分布数据时，由于各簇中心重叠，因此 *DB* 指标很难对聚类结果形成有效评价.

2) *CH* 指标(Calinski-Harabasz)[19]

$$CH(K)=\frac{\sum_{i=1}^{K}n_i d^2(v_i,\,c)/(K-1)}{\sum_{i=1}^{K}\sum_{x\in C_i}d^2(x,\,v_i)/(N-K)} \tag{8}$$

CH 指标将各簇中心点与样本集的均值中心的距离平方和作为数据集的分离度，将簇中各点与簇中心的距离平方和作为簇内的紧密度，将分离度与紧密度的比值视为最终指标. 该指标越大表示各簇之间分散程度越高，簇内越紧密，聚类结果越优. 从 *CH* 指标的函数表达式可以看出，当聚类数趋近于样本容量 N 时，各样本自成一簇，簇中心即为各样本自身，此时簇内距离和约等于 0，分母为极小值，*CH* 指标将趋于最大，此时的聚类评价结果无实际意义.

3) F_r 指标[20]

$$F_r=CH\cdot\lg K \tag{9}$$

针对 *CH* 指标中簇内距离和有可能趋于 0 的问题，文献[10] 对其进行了改进，在原指标的基础上乘以因子 $\lg K$，来调节当 K 值趋于 N 时 *CH* 出现极值的情况. 一般情况下 $K\geqslant 2$，$\lg K$ 大于 0，因此 F_r 指标的变化趋势与 *CH* 指标相类似，即该指标值越大，聚类质量越好.

4) *Dunn* 指标(Dunn's Indices)[21]

$$DVI=\min_{1\leqslant i\leqslant K}\left(\min_{1\leqslant j\leqslant K,\ i\neq j}\left(\frac{d(C_i,\,C_j)}{\max\limits_{1\leqslant t\leqslant K}(\delta(C_t))}\right)\right) \tag{10}$$

Dunn 指标用簇间距离与簇内距离的比值表示，簇间距离是任意 2 个不同簇的样本间的最小距离，簇内距离用最大簇的“直径”表示(簇内两个样本间的最大距离)，可以看出 *Dunn* 指标越大，意味着簇间距离越大，同时簇内距离越小，说明聚类质量越好. 事实上，*Dunn* 指标跟 *CH* 指标一样，都不适用于散点状数据，即当簇内样本趋于 1 时，簇内距离接近于 0，指标值最大，意味着聚类结果最佳，但此时的聚类结果明显与真实分布有较大偏差. 此外，当数据成环状或长条状分布时，由于簇间距离很小，而簇内最大距离却很大时，也会出现聚类结果理想，但评价指标很低的情况. 另一方面，由于 *Dunn* 采用了聚类结果中所有簇类的最大直径，而忽略了不同簇间直径的差异性，因此，在一定程度上降低了聚类评价的准确度.

5) *IGP* 指标(In-Group Proportion)[22]

$$IGP(K)=\frac{1}{K}\sum_{i=1}^{K}igp(i,\,X) \tag{11}$$

IGP 将 2 个距离最近的样本划分到同一簇的比值作为评判聚类质量的标准，其依据是在对某样本进行聚类划分时，与该样本所属同一簇的其它对象应该与该样本的相似度最高. 即

$$igp(i,\,X)=\frac{\{j\mid \mathrm{Class}_X(j)=\mathrm{Class}_X(j^N)=i\}}{\{j\mid \mathrm{Class}_X(j)=i\}} \tag{12}$$

其中 $igp(i, X)$ 表示数据集 X 中的第 i 类的指标值，j^N 是距离样本 j 最近的样本，$\mathrm{Class}_X(j)$ 表示数据集 X 中的第 j 个样本所属的类别. 由于 IGP 仅关注最近邻的一致性，当 K 值逐渐增加时，该指标会不断减少，在实际使用中，使用该指标得到的聚类个数往往少于真实个数[23].

2 聚类质量评价模型

聚类质量评价模型由改进的 AP 算法和新内部评价指标两部分构成，鉴于 AP 算法对非团状数据的聚类个数远大于真实的聚类数目，以及上述常用内部评价指标的不足，本文对二者分别进行了改进.

2.1 改进 AP 算法

本文在文献[24]的基础上，对 AP 算法得到的初步聚类结果按相似度进行合并，通过减小聚类上限 $k_{\max}$，将聚类区间压缩至合理范围，以提高聚类精度. 基本思路是：首先使用 AP 算法对样本集聚类，再计算任意两簇间最远边界点的距离与样本集平均距离的比值 α，α 表示相邻两簇与样本集在空间结构上的相对关系，该值越小意味着 2 个簇的相对相似度越高，在遍历完全部簇之后，若最小 α 在指定阈值范围内，则将两簇合二为一，否则，保持不变，新算法的定义和公式如下所示：

定义 1 空间任意两点间的欧氏距离定义为

$$d(x_i, x_j) = \sqrt{\sum_{p=1}^{l} (x_i^p - x_j^p)^2} \tag{13}$$

其中，$i = 1, 2, \cdots, N$；$j = 1, 2, \cdots, N$；l 表示样本的特征维度.

定义 2 样本集的平均距离定义为：各数据对象间的距离总和与样本对数量的比值.

$$\overline{Dist} = \sum_{i=1}^{N} \sum_{j=1}^{N} d(x_i, x_j) / A_N^2 \tag{14}$$

其中，A_N^2 表示从样本集 X 中任意选取 2 个样本的排列次数.

定义 3 任意两簇之间的距离定义为：2 个簇间最远边界点的距离.

$$Dist(C_i, C_j) = \max_{x_t \in C_i,\, x_u \in C_j} (d(x_t, x_u)) \tag{15}$$

其中，x_t 和 x_u 表示簇 C_i 和 C_j 相距最远的 2 个样本.

定义 4 簇间相似度定义为：簇间距离与样本集平均距离的比值.

$$\alpha_{i,j} = Dist(C_i, C_j) / \overline{Dist} \tag{16}$$

定义 5 如果簇间相似度 α 在给定阈值范围 w 内，则将两簇合并，否则保持不变.

$$C_i = \begin{cases} C_i \cup C_j (\alpha_{i,j} \in w) \\ C_i (\alpha_{i,j} \notin w) \end{cases} \tag{17}$$

其中，阈值范围 w 可由用户自行设定，其默认值为[1, 1.5].

2.2 聚类有效性指标 Improve-XB

在对无先验知识样本的聚类结果进行评价时，通常将簇内紧密，簇间分离作为内部评价的重要标准，若将各簇中心点间的距离作为簇间距离，可能会出现因簇中心重叠而导致聚类评价结果失效等问题. 本文在 XB 指标[25]的基础上进行了优化，用分属 2 个不同簇的全部样本对的距离之和的最小值表示簇间距离，通过增强簇间的平均相似性，削弱噪声数据的影响，避免数据成环状或长条状分布时指标失效，新指标 Improve-XB(以下简称 IXB)的定义及公式如下：

定义 6 簇内紧密度(Compactness) 定义为：簇内全部样本与所属簇中心的距离之和.

$$Com = \sum_{i=1}^{K} \sum_{x \in C_i} d(x, v_i) \tag{18}$$

其中，x 表示簇 C_i 内的样本，v_i 是簇 C_i 的中心，K 表示样本集的聚类个数.

定义 7 簇间分离度(Separation) 定义为：两簇间全部样本对的距离之和的最小值.

$$Sep = \min \sum_{i=1,\, j=1 x_t \in C_i,\, x_u \in C_j,\, i \neq j}^{K} d(x_t, x_u) \tag{19}$$

其中，x_t 和 x_u 分别表示簇 C_i 和簇 C_j 中的任意 2 个样本.

定义 8　*IXB* 指标定义为：簇内紧密度与簇间分离度的比值与其倒数之和.

$$IXB(K)=\frac{Sep}{Com}+\frac{Com}{Sep} \tag{20}$$

定义 9　最优聚类数 K_{opt} 定义为：$IXB(K)$ 取最大值时的聚类数目.

$$K_{opt}=\operatorname{argmax}\{IXB(K)\} \tag{21}$$

其中，$K\in[2,K_{\max}]$，$K_{\max}$ 由改进的 AP 算法给出.

从定义 8 可以看出：*IXB* 指标中的 Sep/Com 随着聚类数 K 的增加而递增，而 Com/Sep 则反之，*IXB* 通过制衡簇内紧密度和簇间分离度之间的关系，确保了最优聚类的划分，该值越大，说明聚类质量越好.

2.3　聚类质量评价模型描述

将改进的 AP 算法与 *IXB* 指标相结合，构建的聚类质量评价(Cluster Quality Evaluation，以下简称 CQE)模型如图 1 所示，模型的实施分为 3 个环节，具体描述如下：

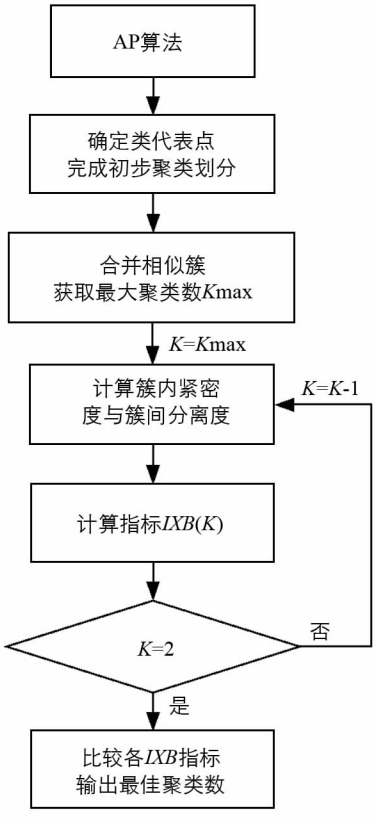

图 1　聚类质量评价模型

1. 确定类代表点

a) 使用式(1)计算各样本对间的相似度 s，进而得到样本集的相似度矩阵 $\boldsymbol{S}$.

b) 设置阻尼因子 λ，迭代次数 t，使用式(2)、式(3)迭代更新吸引度 r 与归属度 a，使用式(4)、式(5)对吸引度和归属度进行校正，消除震荡.

c) 根据式(6)将吸引度与归属度之和最大的候选点作为类代表点，重复执行步骤 b，完成全部簇代表点的选取，最后将非代表点样本划分至相应簇中.

2. 获取最大聚类数 $K_{\max}$

a) 根据式(13)、式(14)计算样本集的平均距离.

b) 根据式(13)、式(15)计算任意两簇的最远边界点间的距离，并将该值作为这两簇的簇间距离，重复执行本步骤，得到各簇间的距离矩阵.

c) 使用式(16)计算各簇间的相似度，设置阈值范围 w，根据式(17)对满足条件的簇进行合并，实现簇更新.

d) 重复执行步骤 b，c，更新各簇间的相似度，直至相似度值超出给定的阈值范围，得到最大聚类数 $K_{\max}$. 至此，簇更新结束.

3. 确定最优聚类数 K_{opt}

a) 将最大聚类数 $K_{\max}$ 作为 K 的初始值.

b) 使用式(18)、(19)分别计算各簇的簇内紧密度 Com 与簇间分离度 Sep，使用式(20)计算 *IXB* 指标.

c) 令 $K=K-1$，重复步骤 b，完成 $K\in[2,K_{\max}]$ 范围内的 *IXB* 指标的计算.

d) 根据式(21)，将 *IXB* 取最大值时的 K 值作为最优聚类数 K_{opt} 输出.

3　实验结果与分析

实验分为有效性测试和模型应用 2 个部分：首先，使用改进的 AP 算法依次结合 *IXB* 和其他 5 个评价指标对 UCI 数据集(表 1)进行聚类数对比测试，目的是验证模型的有效性；其次，将 CQE 模型应用于数据集 KDD CUP99，从检测率、分类正确率、漏报率 3 个方面验证其实用性. 本文实验环境：Intel(R) Core (TM) i3－3240，CPU @3.40GHz，8G 内存，Win10 专业版，实验平台 Matlab 2011b.

表1 UCI实验数据

数据集	样本个数	属性个数	标准聚类个数
bupa	345	6	2
diabetes	768	8	2
breast	699	9	2
wine	178	13	3
wdbc	569	30	2

3.1 CQE模型的有效性测试

在该测试环节中，使用 DB,CH,Fr,DVI 和 IGP 5个内部评价指标与本文的 IXB 指标进行对比，对比结果如表2所示.

表2 各内部评价指标聚类数对比

	标准类数	DB	CH	Fr	DVI	IGP	IXB
bupa	2	14	2	2	2	2	2
diabetes	2	20	5	3	2	2	2
breast	2	18	2	2	3	3	2
wine	3	3	2	2	4	5	3
wdbc	2	3	2	2	3	3	3

观察表2可知，IXB 在5个UCI数据集上的聚类数正确率为80%，而 DB，CH，Fr，DVI 和 IGP 指标依次为20%，60%，60%，40%，40%，IXB 指标的聚类数正确率明显高于其他5种指标.

3.2 CQE模型在KDD CUP99中的应用

本实验环节选取KDD CUP99[26-27]训练集中的12 320条记录作为训练数据，从corrected数据集中随机选取10 520条数据分成4组，作为测试集用于检验模型的性能. 在数据预处理方面，首先使用One-hot编码完成字符数据的格式转换，再将数据集的41个特征降维至14个[28]，经归一化处理后的样本集描述如表3.

表3 KDD CUP99数据集

	训练集	测试集 T1	测试集 T2	测试集 T3	测试集 T4
Normal	9 000	2 000	1 850	1 600	1 550
DOS	1 400	420	340	370	470
PROBE	1 500	330	440	380	450
U2R	20	5	5	5	5
R2L	400	80	90	80	50

3.2.1 IXB 指标与 K 值的关系

设定阈值范围 $w=[1, 1.5]$，阻尼因子为0.9，最大迭代次数为1 000次，执行改进AP算法后得出训练集的最大聚类数 $K_{max}=32$，使用CQE模型得出的 IXB 指标结果如图2所示. 可以看出，随着 K 值的不断增大，IXB 呈现上升趋势，当 $23\leqslant K\leqslant 25$ 时，IXB 逐渐趋于平稳；当 $K=26$ 时，IXB 达到极大值，此后随着 K 值的再次增大，IXB 缓慢下降；当 $K=29$ 时，下降幅度明显增加. 需要特别指出的是，当 $K=18$ 时，IXB 指标发生了异常变化，通过观察表4中的 Com、Sep 以及 IXB 指标值可知，$Com_{18}<Com_{19}$，查询 $K=18$ 和 $K=19$ 时训练集的聚类结果发现，前者对异常数据的分类正确率为78%，后者为76%，因此，前者的簇内紧致程度优于后者，而当 $K=19$ 时，虽然簇间分离度有所增加($Sep_{18}=1\ 524\ 019\ 748$，$Sep_{19}=1\ 524\ 738\ 937$)，但其增幅较小，所以 $IXB(18)>IXB(19)$.

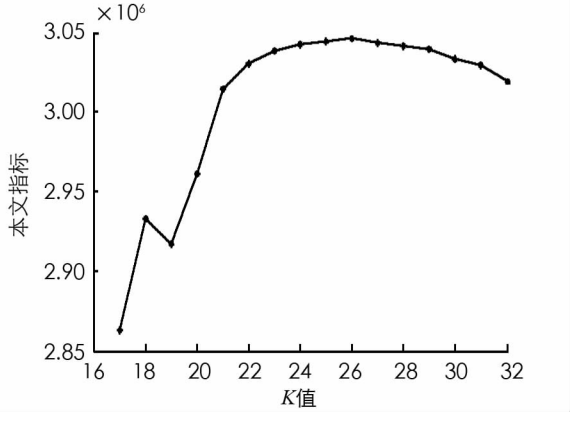

图2 训练集的 IXB-K 的关系图

表 4　$K\in[17, 32]$的 IXB 指标值(单位：$Com\times10^2$，$Sep\times10^9$，$IXB\times10^6$)

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Com	5.26	5.21	5.22	5.22	5.22	5.19	5.10	5.09	5.08	4.94	4.94	4.89	4.88	4.87	4.86	4.85
Sep	1.51	1.52	1.52	1.55	1.57	1.57	1.55	1.55	1.55	1.51	1.50	1.49	1.48	1.48	1.47	1.47
IXB	2.86	2.93	2.92	2.96	3.01	3.03	3.04	3.04	3.04	3.05	3.04	3.04	3.04	3.03	3.03	3.02

3.2.2　最优聚类数范围内的各检测指标

为了验证通过 IXB 指标得出的最优 K 值是否有效，即 K 取最优值时的各项入侵检测指标是否有效，本文将图 2 中 IXB 缓慢上升至峰值，再从峰值缓慢下降这一阶段所对应的多个连续 K 值定义为最优聚类数范围，即 $K_{opt}\in[23, 29]$，使用 4 组测试集对该范围内的入侵检测指标分别进行了验证，结果如表 5 所示. 取平均值后的折线图如图 3、图 4 所示，可以看出，当聚类数为 27 时，入侵检测率和正确分类率同时达到最大值，分别是 93.62%，95.17%；当聚类数为 26 时，漏报率达到最小，为 3.37%；当聚类数为 25 时，误报率达到最小，为 4.14%.

表 5　测试集各指标检测结果($K\in[23, 29]$)

K	检测率				分类正确率				漏报率				误报率			
	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4
23	92.69	92.92	92.88	92.83	92.61	92.71	92.62	93.1	6.44	6.28	6.36	6.41	5.24	5.31	5.31	5.22
24	93.21	92.61	92.82	92.75	93.87	93.92	93.84	93.92	5.58	5.51	5.61	5.48	4.62	4.63	4.68	4.62
25	93.24	93.21	93.24	93.22	94.02	93.75	93.91	93.82	3.89	3.94	3.98	3.86	4.11	4.21	4.11	4.14
26	93.42	93.41	93.38	93.41	94.35	94.21	94.34	94.18	3.49	3.39	3.41	3.18	4.63	4.69	4.62	4.61
27	93.61	93.54	93.64	93.67	95.06	95.21	95.19	95.21	3.29	3.61	3.66	3.33	5.35	5.26	5.41	5.43
28	93.29	93.16	92.28	93.31	94.23	94.29	94.32	94.19	3.88	3.76	3.88	3.89	5.92	5.81	5.83	5.89
29	93.08	93.2	93.19	93.18	94.23	94.33	94.16	94.23	4.29	4.31	4.41	4.46	4.88	4.92	5.02	5.01

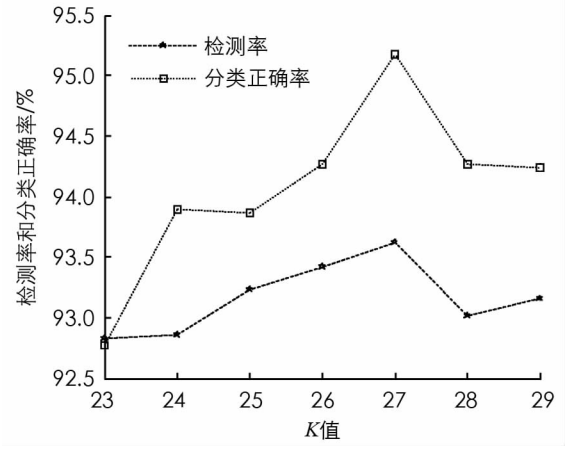

图 3　不同 K 值的平均检测率和分类正确率

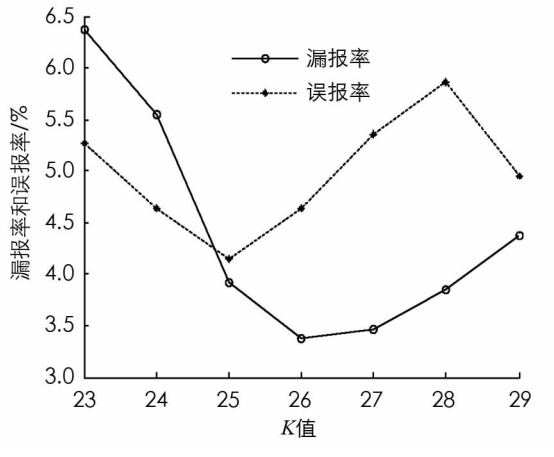

图 4　不同 K 值的平均漏报率和误报率

3.2.3　CQE 模型和其他算法的对比测试

为进一步验证 IXB 指标的普适性，同时也为了将 CQE 模型和其他算法的聚类质量进行横向比较，本环节将 K-means、文献[18]和文献[19]算法与 IXB 指标相结合，在 $K_{opt}\in[23, 29]$这一区间内对入侵检测指标进行了对比测试. 观察表 6 和表 7 的对比结果可知，结合了 IXB 指标的其他 3 种算法的检测率和分类正确率在 $K=\{25, 26, 27\}$时可得到最大值，CQE 模型的检测率和分类正确率在整个区间内全部优于其他 3 种算法，其最大检测率为 93.62%，最大分类正确率为 95.17%.

表6 检测率对比结果 %

K	K-means	文献[17]	文献[18]	CQE
23	90.34	90.69	90.71	92.83
24	91.13	90.83	91.95	92.85
25	91.29	91.60	92.34	93.23
26	90.47	92.18	92.39	93.41
27	91.72	91.60	92.68	93.62
28	86.19	90.96	90.39	93.01
29	91.03	90.13	91.60	93.16

表7 分类正确率对比结果 %

K	K-means	文献[17]	文献[18]	CQE
23	90.46	91.16	92.11	92.76
24	90.47	92.75	92.79	93.89
25	91.46	92.47	93.10	93.87
26	92.43	93.25	93.79	94.27
27	93.24	94.28	94.22	95.17
28	92.64	92.66	92.84	94.26
29	92.29	92.31	93.24	94.24

由表8可看出，在漏报率方面，除了个别结果，本文算法均优于其他3种算法，其最小漏报率为3.37%.

表8 漏报率对比结果 %

K	K-means	文献[17]	文献[18]	CQE
23	5.93	5.11	6.11	6.37
24	5.61	5.35	6.07	5.55
25	4.93	4.81	5.86	3.92
26	4.42	4.38	5.27	3.37
27	4.89	4.91	4.96	3.47
28	4.71	5.17	5.64	3.85
29	5.03	5.91	5.75	4.36

由表9的误报率可看出，CQE模型在$K=\{24, 25, 26\}$时的误报率低于其他3种算法，其他范围内CQE模型的误报率略高，但总体上低于其他3种算法的平均值.

表9 误报率对比结果 %

K	K-means	文献[17]	文献[18]	CQE
23	5.21	5.23	5.19	5.27
24	5.01	5.02	5.12	4.64
25	4.92	4.79	4.88	4.14
26	4.85	4.87	4.85	4.64
27	4.88	5.02	5.13	5.36
28	5.34	5.14	5.23	5.86
29	4.91	4.94	5.29	4.96

从本环节的对比测试可以看出，IXB指标具有良好的普适性，相比于其他3种算法，CQE模型的聚类结果更为理想，能够以较小的误报率代价，提高入侵检测率和分类正确率，降低了漏报率，在整体性能上取得了较好的效果.

4 结 语

本文提出了一种改进的AP聚类算法，以簇间相似度为参考依据，通过循环合并相似簇的方式，将聚

类区间压缩至较小范围，解决了 AP 算法对结构复杂的数据集进行聚类时出现的局部聚类过多、精准度低的问题；改进了 XB 评价指标，通过计算不同簇间样本对平均距离的方法削弱噪声数据影响，增强簇间的平均相似性，解决了因中心点过于紧密，而导致指标无效的问题. 实验结果表明，由改进 AP 算法和 IXB 指标构建的 CQE 模型可以给出精准的最优聚类数范围，但是在如何设定阈值方面，本文暂时无法给出较为成熟的理论依据. 下一步工作将深入研究阈值范围的设定与样本集的关系，使得 CQE 模型的功能更加完善，具有更广的应用范围.

参考文献：

[1] 张辉荣，唐　雁，何　荧，等. 面向分类数据的重叠子空间聚类算法 SCCAT [J]. 西南大学学报(自然科学版)，2016，38(3)：171-176.

[2] 刘光华，杨沛霖，赵　鹏. 基于模糊聚类分析—BP 神经网络法的平缓硬质岩斜坡卸荷带宽度评价模型 [J]. 西南大学学报(自然科学版)，2016，38(8)：167-173.

[3] YUE S，WANG J，WANG J，et al. A New Validity Index for Evaluating the Clustering Results by Partitional Clustering Algorithms [J]. Soft Computing，2016，20(3)：1-12.

[4] 周世兵，徐振源，唐旭清. 基于近邻传播算法的最佳聚类数确定方法比较研究 [J]. 计算机科学，2011，38(2)：225-228.

[5] 朴尚哲，超木日力格，于剑. 模糊 C 均值算法的聚类有效性评价 [J]. 模式识别与人工智能，2015，28(5)：452-461.

[6] HARTIGAN J A，WONG M A. Algorithm AS 136：A K-Means Clustering Algorithm [J]. Journal of the Royal Statistical Society，1979，28(1)：100-108.

[7] LUCASIUS C B，DANE A D，KATEMAN G. On K-medoid Clustering of Large Data Sets with the Aid of a Genetic Algorithm：Background，Feasiblity and Comparison [J]. Analytica Chimica Acta，1993，282(3)：647-669.

[8] FREY B J，DUECK D. Clustering by Passing Messages Between Data Points [J]. Science，2007，315(5814)：972-976.

[9] 刘自豪，张　斌，祝　宁，等. 基于改进 AP 聚类算法的自学习应用层 DDoS 检测方法 [J]. 计算机研究与发展，2018，55(6)：1236-1246.

[10] 王卫涛，钱雪忠，曹文彬. 自适应参数调整的近邻传播聚类算法 [J]. 小型微型计算机系统，2018，39(6)：1305-1311.

[11] 倪志伟，荆婷婷，倪丽萍. 一种近邻传播的层次优化算法 [J]. 计算机科学，2015，42(3)：195-200.

[12] LI P，JI H F，WANG B L，et al. Adjustable Preference Affinity Propagation Clustering [J]. Pattern Recognition Letters，2017，85：72-78.

[13] 郭秀娟，曹　东，陈　莹. 改进的 AP 聚类算法研究 [J]. 吉林建筑大学学报，2015，32(1)：72-75.

[14] 梁修荣，杨正益. 基于聚类和 SVM 的数据分类方法与实验研究 [J]. 西南师范大学学报(自然科学版)，2018，43(3)：91-96.

[15] 赵延龙，滑　楠. 基于初始偏向度的 AP 算法聚类性能优化研究 [J]. 计算机应用研究，2018，35(2)：372-374，399.

[16] 周世兵，徐振源，唐旭清. 一种基于近邻传播算法的最佳聚类数确定方法 [J]. 控制与决策，2011，26(8)：1147-1152，1157.

[17] 甘月松，陈秀宏，陈晓晖. 一种 AP 算法的改进：M-AP 聚类算法 [J]. 计算机科学，2015，42(1)：232-235，267.

[18] DAVIES D L，BOULDIN D W. A Cluster Separation Measure [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，1979，1(2)：224-227.

[19] CALINSKI T，HARABASZ J. A Dendrite Method for Cluster Analysis [J]. Communications in Statistics-Simulation and Computation，1974，3(1)：1-27.

[20] 谢娟英，周　颖. 一种新聚类评价指标 [J]. 陕西师范大学学报(自然科学版)，2015，43(6)：1-8.

[21] DUNNÂ J C. Well-Separated Clusters and Optimal Fuzzy Partitions [J]. Journal of Cybernetics，1974，4(1)：95-104.

[22] LIN T. Data Mining and Machine Oriented Modeling：a Granular Computing Approach [J]. Applied Intelligence，2000，13(2)：113-124.

[23] 冯柳伟，常冬霞，邓　勇，等. 最近最远得分的聚类性能评价指标 [J]. 智能系统学报，2017，12(1)：67-74.

[24] 王开军，张军英，李　丹，等. 自适应仿射传播聚类 [J]. 自动化学报，2007，33(12)：1242-1246.

[25] XIE X L，BENI G. A Validity Measure for Fuzzy Clustering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，1991，13(8)：841-847.

[26] BOLÓN-CANEDO V, SÁNCHEZ-MAROÑO N, ALONSO-BETANZOS A. Feature Selection and Classification in Multiple Class Datasets: an Application to KDD Cup 99 Dataset [J]. Expert Systems With Applications, 2011, 38(5): 5947-5957.

[27] 文 华, 王斐玉. 利用SSO加速最佳路径森林聚类的网络入侵检测 [J]. 西南师范大学学报(自然科学版), 2017, 42(5): 34-40.

[28] 吴建胜, 张文鹏, 马 垣. KDD CUP99数据集的数据分析研究 [J]. 计算机应用与软件, 2014, 31(11): 321-325.

Cluster Quality Evaluation Model Based on Improved Affinity Propagation Algorithm

ZOU Chen-song, DUAN Gui-qin,
OUYANG Ming-xing, LIU Feng

Department of Electrical Engineering, Guangdong Songshan Polytechnic College, Shaoguan Guangdong 512126, China

Abstract: In order to solve the problems of more local clustering and low precision of non-spherical data sets in the clustering process for Affinity Propagation algorithm, a clustering quality evaluation model based on improved AP algorithm has been proposed. Firstly, based on the initial clustering of AP algorithm, the upper limit value k_{max} of clustering has been reduced by merging clusters with larger similarity, and the range of clustering interval been further compressed. Secondly, a new internal evaluation index has been given with the average distance of sample pairs belonging to different clusters represents the distance between clusters, which has weakened the influence of noise data, balanced the relationship between cluster separation and cluster compactness. The experimental results on UCI and KDD CUP99 datasets show that the new model can give accurate optimal clustering number (range), and can effectively improve the detection rate and classification accuracy of samples while maintaining a low false alarm rate.

Key words: cluster evaluation index; affinity propagation; internal evaluation index; optimal clustering number

责任编辑 崔玉洁