计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(6):235−242 [doi: 10.15888/j.cnki.csa.006946]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于改进 K-medoids 的聚类质量评价指标研究①

邹臣嵩[1], 段桂芹[2]

[1](广东松山职业技术学院 电气工程系, 韶关 512126)
[2](广东松山职业技术学院 计算机系, 韶关 512126)
通讯作者: 邹臣嵩, E-mail: 190352915@qq.com

摘　要: 为了更好地评价无监督聚类算法的聚类质量, 解决因簇中心重叠而导致的聚类评价结果失效等问题, 对常用聚类评价指标进行了分析, 提出一个新的内部评价指标, 将簇间邻近边界点的最小距离平方和与簇内样本个数的乘积作为整个样本集的分离度, 平衡了簇间分离度与簇内紧致度的关系;提出一种新的密度计算方法, 将样本集与各样本的平均距离比值较大的对象作为高密度点, 使用最大乘积法选取相对分散且具有较高密度的数据对象作为初始聚类中心, 增强了 K-medoids 算法初始中心点的代表性和算法的稳定性, 在此基础上, 结合新提出的内部评价指标设计了聚类质量评价模型, 在 UCI 和 KDD CUP 99 数据集上的实验结果表明, 新模型能够对无先验知识样本进行有效聚类和合理评价, 能够给出最优聚类数目或最优聚类范围.

关键词: 聚类评价指标; K-medoids; 无监督聚类; 最优聚类数

引用格式: 邹臣嵩,段桂芹.基于改进 K-medoids 的聚类质量评价指标研究.计算机系统应用,2019,28(6):235–242. http://www.c-s-a.org.cn/1003-3254/6946.html

Cluster Quality Evaluation Index Based on K-medoids Algorithm

ZOU Chen-Song[1], DUAN Gui-Qin[2]

[1](Department of Electrical Engineering, Guangdong Songshan Polytechnic, Shaoguan 512126, China)
[2](Department of Computer Science, Guangdong Songshan Polytechnic, Shaoguan 512126, China)

Abstract: In order to better evaluate the clustering quality of unsupervised clustering algorithm and solve the problem of invalidation of clustering evaluation results caused by overlapping cluster centers, the commonly used cluster evaluation index is analyzed and a new internal evaluation index is proposed, the product of the minimum square of the distance between the adjacent boundary points and the number of samples in the cluster is taken as the separation degree of the whole sample set, the relation between the degree of separation between clusters and the degree of compactness within clusters is balanced; a new density calculation method is proposed, which takes the object with a larger average distance ratio between the sample set and each sample as a high-density point, and uses the maximum product method to select the relatively dispersed data object with a higher density as the initial cluster center, thus enhancing the representativeness of the initial center of K-medoids algorithm and the stability of the algorithm. On this basis, the cluster quality evaluation model is designed with the newly proposed internal evaluation index. The experimental results on UCI and KDD CUP 99 data sets show that the new model can effectively cluster and reasonably evaluate non-prior knowledge samples, and can give the optimal number or range of clustering.

① 基金项目: 韶关市科技计划项目 (2017CX/K055); 广东松山职业技术学院重点科技项目 (2018KJZD001)
Foundation item: Science and Technology Plan of Shaoguan City (2017CX/K055); Key Science and Technology Project of Guangdong Songshan Polytechnic (2018KJZD001)
收稿时间: 2018-12-19; 修改时间: 2019-01-10; 采用时间: 2019-01-15; csa 在线出版时间: 2019-05-25

Key words: cluster evaluation index; K-medoids; unsupervised clustering; optimum clustering number

聚类是在没有任何先验知识的指导下, 从样本集合中挖掘出潜在的相似模式, 并将其划分成多个组或簇的过程, 其目的是使得簇内相似度高, 而簇间相似度低, 数据对象的簇可以看做隐含的类, 聚类可以自动地发现这些类.由于在聚类过程中并没有提供类的标号信息, 因此, 聚类又被称做无监督学习, 对于无先验知识的样本来说, 如何对其聚类结果进行有效评价是国内外的研究热点, 许多经典的内部聚类评价指标先后被提出, 如 CH, I, DB, SD, BWP 等, 然而这些指标可能会受噪声等异常的影响, 因此, 如何改进或设计出更为科学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密不可分, 研究表明, 没有任何一种聚类算法可以得到所有数据集的最优划分[1], 而应用范围较广的 K-means、K-medoids 及其衍生算法[2]在实际应用中又存在一定的不足.为此, 本文对常用聚类评价指标进行了对比分析, 提出了一个新的评价指标, 并对 K-medoids 算法进行了改进, 在此基础上, 设计了聚类质量评价模型, 先后采用 UCI 数据集和 KDD CUP99 数据集对新模型进行了验证, 实验结果表明, 新评价指标的聚类数正确率明显高于其他四种常用指标, 聚类质量评价模型可以给出精准的聚类数范围.

1 研究现状

1.1 划分式聚类算法

划分式聚类算法[3]在运算前需要人工预定义聚类数 k, 再通过反复迭代更新各簇中心, 不断优化 (降低) 目标函数值, 逐渐逼近最优解, 完成最终聚类, K-means 和 K-medoids 是两种典型的划分式聚类算法.K-means 算法的簇中心是通过计算一个簇中对象的平均值来获取, 它根据数据对象与簇中心的距离完成"粗聚类", 再通过反复迭代, 将样本从当前簇划分至另一个更合适的簇来逐步提高聚类质量, 其核心思想是找出 k 个簇中心, 使得每个数据点到其最近的簇中心的平方距离和被最小化.该方法描述容易、实现简单快速, 是目前研究最多的聚类方法, 文献[4–6]从初始簇中心的选择、对象的划分、相似度的计算方法、簇中心的计算方法等方面对该经典算法进行了改进, 使其适用于不同的聚类任务, 但在使用中存在一些不足: 簇的个数难以确定; 聚类结果对初始值的选择较敏感; 算法容易陷入局部最优值; 对噪音和异常数据敏感; 不能用于发现非凸形状的簇, 或具有各种不同规模的簇.

当样本中存在一个或多个极值对象时, 采用均值算法会显著地扭曲数据的分布, 而平方误差函数的使用会进一步地扩大这一影响, 针对这一问题, K-medoids 通过试图最小化所有对象与其所属簇的中心点之间的绝对误差之和的方式找出簇中心点.典型基于中心的划分方法有 PAM、CLARA 和 CLARANS[7], 虽然 K-medoids[8]算法对噪声和异常数据的敏感程度有所改善, 但仍依赖于初始类簇中心的随机选择, 且更新类簇中心时采用全局最优搜索, 故耗时较多. 文献[9]提出一种快速 K-medoids 算法, 从初始聚类中心选择、类簇中心更新方法两方面对 PAM 算法进行改进, 基本思想是: 首先计算数据集中每个样本的密度, 选择前 k 个位于样本分布密集区域的样本为初始聚类中心, 再将其余样本分配到距离最近的类簇中心, 产生初始聚类结果; 然后, 在每个类簇内部找一个新中心, 使该簇非中心样本到中心样本距离之和最小, 进而得到 k 个新中心; 最后按距离最近原则, 重新分配所有非中心样本到最近类簇中心, 如果本次迭代所得聚类结果与前一次不同, 则转至下一次迭代, 否则算法停止.在实际应用中, 该算法的初始中心在一定程度上存在过于集中的可能, 因此, 从样本的空间结构来看, 各中心点的分散程度不高, 代表性依然不足.

1.2 内部评价指标

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, 在对聚类结果进行评估时, 内部评价指标在不涉及任何外部信息的条件下, 仅依赖数据集自身的特征和度量值, 通过计算簇内部平均相似度、簇间平均相似度或整体相似度来评价聚类效果的优劣和判断簇的最优个数.理想的聚类效果是簇内紧密且簇间分离, 因此, 常用内部评价指标的主要思想是通过簇内距离和簇间距离的某种形式的比值来度量的.

为便于对各指标和本文算法进行描述, 设样本集合 $\boldsymbol{X}=\{x_1, x_2, \cdots, x_i, \cdots, x_N\}$, $|\boldsymbol{X}|=N$, 各样本特征数为 p, 该样本集由 k 个簇构成, 即 $\boldsymbol{X}=\{C_1, C_2, \cdots, C_k\}$, 每簇样本数为 n, c 为样本集的均值中心, 簇中心集合 $\boldsymbol{V}=\{v_1, v_2, \cdots, v_k\}(k< N)$, 常见的内部评价指标及其特点如下:

(1) DB 指标 (Davies-Bouldin Index)[10]

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

DB 指标将相邻两簇的簇中各样本与簇中心的平均距离之和作为簇内距离, 将相邻两簇的簇中心间的距离作为簇间距离, 取二者比值最大者作为该簇的相似度, 再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, 该指标越小意味着各簇间的相似度越低, 从而对应更佳的聚类结果. DB 指标适合评价"簇内紧凑, 簇间远离"的数据集, 但当数据集的重叠度较大 (如:当遇到环状分布数据时), 由于各簇的中心点重叠, 因此 DB 指标很难对其形成正确的聚类评价.

(2) CH 指标 (Calinski-Harabasz)[11]

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

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 CH 的最终指标.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. Milligan 在文献[12]中, 对 CH 等评价指标的性能进行了深入探讨, 实验结果表明:CH 指标在多数情况下, 都要优于其它的指标, 但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) XB 指标 (Xie-Beni)[13]

$$XB(k)=\frac{\sum_{i=1}^{k}\sum_{x\in C_i} d^2(x,v_i)}{n\times \min_{i,j\neq i} d^2\left(v_i,v_j\right)} \tag{3}$$

和 CH 指标一样, XB 也是将簇内紧密度与簇间分离度的比值作为指标的表达式, 期望在簇内紧密度与簇间分离度之间寻找一个新的平衡点, 使其达到一个理想化的极值, 从而得到最优的聚类结果. 与 CH 指标不同是: XB 指标使用最小簇中心距离的平方作为整个样本集的分离度.

(4) Sil 指标 (Silhouette-Coefficient)[14]

$$Sil(k)=\frac{1}{k}\sum_{i=1}^{k}\left\{\frac{1}{n_i}\sum_{x\in C_i}\frac{b(x)-a(x)}{\max[b(x),a(x)]}\right\} \tag{4}$$

式中, $a(x)$ 是样本 x 所属簇的簇内凝聚程度, 用 x 与其同一个簇内的所有元素距离的平均值来表示, 凝聚度 $a(x)$ 定义为:

$$a(x)=\frac{1}{n_i-1}\sum_{x,y\in C_i, y\neq x} d(x,y) \tag{5}$$

式中, $b(x)$ 是样本 x 与其他簇的簇间分离程度, 用 x 与其它簇之间平均距离的最小值来表示, 分离度 $b(x)$ 定义为:

$$b(x)=\min_{j,j\neq i}\left[\frac{1}{n_j}\sum_{x\in c_i, y\in c_j} d(x,y)\right] \tag{6}$$

从式 (4) 可以看出, 当簇内距离减小, 簇间距离增大时, Sil 指标值增大, 聚类结果趋于理想, 因此, 要取得最佳聚类, 就需要减少簇内各点之间的距离, 同时增大簇间的距离.然而, 与 CH 指标类似, 当数据接近散点状分布时, 聚类结果并不理想, 此外, 当数据成条状分布的时, 各簇中心非常接近, 且聚类结果非常理想, 但 Sil 指标此时最小, 并不能对聚类结果做出客观的评价.

2 聚类质量评价模型

聚类质量评价模型由聚类算法和内部评价指标两部分构成, 鉴于 K-medoids 算法及其改进算法在初始聚类中心选择阶段存在的问题, 以及上述常用内部评价指标存在的缺陷, 本文对二者分别进行了改进, 具体如下.

2.1 改进 K 中心点聚类算法

本文在文献[7,8]的基础上, 选取被其他样本紧密围绕且相对分散的数据对象作为初始聚类中心, 使得中心点在确保自身密度较大的同时还具备良好的独立性.基本思路是: 首先通过计算样本集中各样本间的距离得到样本集的距离矩阵; 将样本集与各样本的平均

距离的比值作为样本的密度, 选取密度值最高的 α 个样本存入高密度点集合 $\boldsymbol{H}$ 中, α 表示候选代表点在样本集中所占的比例, 该值可由用户指定, 本文实验环节中的 α 值为 30%; 从集合 $\boldsymbol{H}$ 中选取相对分散且具有较高密度的初始中心存入集合 $\boldsymbol{V}$, 即 $\boldsymbol{V}=\{v_1, v_2, \cdots, v_k\}$. 最后, 借鉴文献[7]的算法, 将各样本按最小距离分配至相应簇中, 重复这一过程, 直至准则函数收敛, 本文算法的定义和公式如下.

定义 1. 空间任意两点间的欧氏距离定义为

$$d\left(x_i, x_j\right)=\sqrt{\sum_{w=1}^{p}\left(x_i^w-x_j^w\right)^2} \tag{7}$$

其中, i=1, 2, ⋯, N; j=1, 2, ⋯, N

定义 2. 数据对象 x_i 的平均距离定义为:x_t 与全部样本的距离之和除以样本集的总个数.

$$Dist(x_i)=\sum_{j=1}^{N} d\left(x_i, x_j\right)/N \tag{8}$$

定义 3. 样本集的平均距离定义为:各数据对象间的距离总和除以从样本集中任选两个对象的所有排列次数.

$$DistMean=\sum_{i=1}^{N}\sum_{j=1}^{N} d\left(x_i, x_j\right)/A_N^2 \tag{9}$$

定义 4. 数据对象 x_i 的密度定义为:样本集的平均距离与数据对象 x_i 的平均距离的比值.

$$Density(x_i)=DistMean/Dist(x_i) \tag{10}$$

定义 5. 数据对象 x_i 与所属簇的各数据对象的距离之和为:

$$DistSum(x_i)=\sum_{j=1}^{n_i} d\left(x_i, x_j\right) \tag{11}$$

其中, $x_i, x_j \in C_t$, t=1, 2, ⋯, k

定义 6. 簇 C_t 的簇内距离和矩阵为:

$$DistSum(C_t)=\begin{bmatrix} DistSum(x_1) \\ DistSum(x_2) \\ \cdots \\ DistSum(x_n) \end{bmatrix} \tag{12}$$

其中, t=1, 2, ⋯, k

定义 7. 数据对象 x_i 在簇更新过程中被视为簇中心的条件为:

$$DistSum(x_i)=\min(DistSum(C_t)) \tag{13}$$

其中, $x_i \in C_t$, t=1, 2, ⋯, k

定义 8. 聚类误差平方和 E 的定义为

$$E=\sum_{t=1}^{k}\sum_{j=1}^{n_t}\left|x_{tj}-v_t\right|^2 \tag{14}$$

其中, x_{tj} 是第 t 簇的第 j 个数据对象, v_t 是第 t 簇的中心.

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, 文献[11]和文献[13]将各簇中心之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, 为此本文提出: 使用两个簇的最近边界点间的距离取代簇中心之间的距离, 为了便于分析, 以图 1 所示的人工数据集进行说明, 该数据集由三个环形结构的簇构成, 各簇中心极为接近, 从 DB、XB 指标公式可知, 在计算该数据集的簇间距离时, 由于簇中心点趋于重叠, 必然会出现簇间距离近似于 0 的情况, 当以"簇内紧密, 簇间分离"作为评判依据时, 意味着此时簇间的相似度极高, 从聚类划分的角度来看, 应将二者合并为一簇以提高聚类质量, 而事实上, 这种错误的划分将导致图 1 的最终聚类全部合并为一个簇, 这显然与真实的结果有着明显偏差.而本文使用簇间最近边界点间的距离表示簇间距离, 则可以从几何特征上确保各簇的结构差异性, 进而避免簇间距离为 0 现象的发生, 最大程度地反映出簇间的相似程度, 克服了环状中心点因重叠而导致的聚类结果合并等问题, 新指标 IXB 定义及公式如下.

定义 9. 簇内紧密度 (Compactness) 定义为:各样本与所属簇的中心点的距离平方和. 实质为:

$$Comp(k)=\sum_{i=1}^{k}\sum_{x\in C_i} d^2(x, c_i) \tag{15}$$

定义 10. 簇间分离度 (Separation) 定义为:簇间最邻近边界点的距离平方和与簇内样本个数的乘积.

$$Sep(k)=n\cdot\min_{i,j\neq i} d^2\left(x_i, x_j\right) \tag{16}$$

其中, x_i 和 x_j 是簇 C_i 和 C_j 之间距离最近的边界点.

定义 11. 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, 即实质为:

$$IXB(k)=\frac{Sep}{Comp}+\frac{Comp}{Sep} \tag{17}$$

定义 12. 最优聚类数 k 定义为 $IXB(k)$ 取最大值时

的聚类数目, 即:

$$k_{opt} = \arg\max_{k_{\min}\leqslant k\leqslant k_{\max}} \{IXB(k)\} \quad (18)$$

其中, $k_{min}=2$, $k_{\max}$ 采用文献[15]的 AP 算法得到.

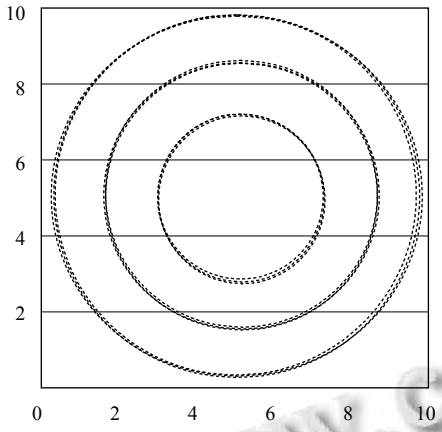

图 1 环状分布数据集

IXB 指标由两部分组成: *Sep*/*Comp* 随着聚类数 k 的增加而递增, *Comp*/*Sep* 随着聚类数 k 的增加而递减, 可以看出, IXB 指标通过制衡 *Sep*/*Comp* 和 *Comp*/*Sep* 之间的关系, 确保了最优聚类划分, IXB 越大, 意味着聚类质量越好.

2.3 聚类质量评价模型描述

将改进的 K 中心点算法与 IXB 指标相结合, 构建聚类质量评价模型如图 2 所示, 模型描述如下:

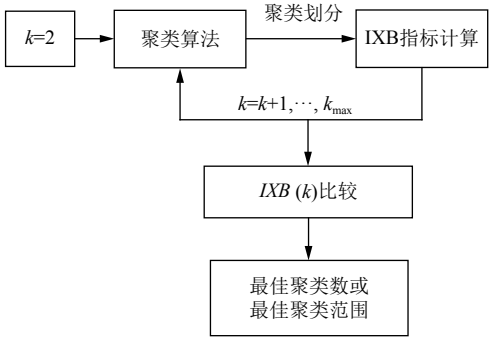

图 2 聚类质量评价模型

(1) 根据式 (6)~(8) 依次计算整个样本集的任意两点间的欧氏距离、各数据对象的平均距离和样本集的平均距离.

(2) 根据式 (10) 计算各数据对象的密度, 选取密度值最高的 α 个样本存入候选初始中心集合 $\boldsymbol{H}$ 中, 令 k=2.

(3) 根据式 (6) 将集合 $\boldsymbol{H}$ 中相距最远的两个高密度点 v_1 和 v_2 存入初始中心集合 $\boldsymbol{V}$ 中.

(4) 从 $\boldsymbol{H}$ 中选择满足与 v_1、v_2 距离乘积最大的 v_3, 将其存储至 $\boldsymbol{V}$ 中.依次类推, 得到相对分散且具有较高密度的初始中心集合 $\boldsymbol{V}$, $\boldsymbol{V}=\{v_1, v_2, \cdots, v_k\}$.

(5) 在更新簇中心阶段, 根据式 (11)、式 (12) 得到簇内距离和矩阵, 根据式 (13) 选出新的簇中心, 沿用文献[7]的算法, 将各样本按最小距离分配至相应簇中, 重复这一过程, 直至准则函数式 (14) 收敛.

(6) 使用式 (17) 计算 IXB 指标, 对当前聚类结果进行评价, 令 k=k+1, 重复步骤 (4), 直至 k=$k_{\max}$.

(7) 使用式 (18), 将 IXB 取最大值时的 k 值作为最优聚类数.

2.4 模型的性能分析

本文算法将样本集与各样本的平均距离比值作为样本的密度, 将高密度点视为候选簇中心, 确保了聚类中心的代表性, 使用最大乘积法对高密度点进行二次筛选, 增强了候选簇中心在空间上的离散程度, 使得聚类中心兼具代表性和分散性, 提高了逼近全局最优解的概率, 这样选择的初始聚类中心更符合样本的分布特征, 甚至有可能位于真实的簇中心, 因此, 所得到的初步聚类划分与样本的真实分布更加接近.在聚类结果评价方面, 通过将簇间最邻近边界点的距离平方和与簇内样本个数的乘积作为簇间分离度, 在簇内紧密度与簇间分离度之间寻找一个新的平衡点, 从而得到一个理想化的极值, 通过搜索该极值, 即可有效地完成最优聚类划分, 确定最优聚类数目 k, 从而得到更好的聚类结果, 从整体上提升无先验知识样本的检测率和分类正确率等评价指标.

3 实验结果与分析

实验分为两个部分, 第一部分对聚类质量评价模型的有效性进行了验证: 首先选用表 1 中的 UCI 数据集对本文算法和其他改进算法的准确率、迭代次数、总耗时进行了对比测试, 然后使用本文算法依次结合 IXB 及其它 4 个评价指标完成了聚类数对比测试; 第二部分将模型应用于数据集 KDD CUP99, 从检测率、分类正确率、漏报率三个方面验证模型的实用性.本文实验环境: Intel(R) Core(TM) i3-3240 CPU @3.40 GHz, 8 GB 内存, Win10 专业版, 实验平台 Matlab 2011b.

表 1 实验数据

数据集	样本个数	属性个数	标准聚类个数
wine	178	13	3
iris	150	4	3
wdbc	569	30	2
heart	270	13	2
ionosphere	351	34	2

3.1 聚类质量评价模型的有效性测试

(1) 改进聚类算法的对比测试

从表 2~表 4 的实验对比结果可以看出, 改进聚类算法的聚类准确率、迭代次数和总耗时全部优于其他四种算法, 主要原因在于 K 中心点的初始中心随机分布令迭代次数与总耗时同时增加, 而文献[7]的初始中心点过于集中, 文献[8]的初始中心虽然具有一定的分散性, 但每次迭代都将簇平均值作为簇中心, 个别维度存在受异常数据影响的隐患, 因此, 本文算法的整体耗时要低于文献[7,8]算法.需要特别指出的是, 由于本文算法的初始聚类中心同时兼具代表性和分散性, 不仅提高了聚类准确率, 同时还降低了算法后期的迭代次数, 因此, 对算法的运算效率的提升产生了正向的推动作用.

表 2 聚类准确率比较

	K 中心点	文献[7]	文献[8]	本文
iris	0.84	0.89	0.89	0.92
heart	0.51	0.53	0.52	0.61
ionosphere	0.59	0.61	0.72	0.71
wine	0.66	0.71	0.71	0.74
wdbc	0.81	0.85	0.86	0.87

表 3 迭代次数

	K 中心点	文献[7]	文献[8]	本文
iris	6.1	3	3	4
heart	9.2	2	5	1
ionosphere	10.2	9	12	3
wine	6.8	5	4	3
wdbc	9.5	8	10	6

表 4 聚类总耗时 (单位: s)

	K 中心点	文献[7]	文献[8]	本文
iris	0.201	0.135	0.146	0.112
heart	0.195	0.158	0.159	0.142
ionosphere	0.184	0.16	0.152	0.138
wine	0.211	0.182	0.149	0.135
wdbc	0.483	0.365	0.378	0.313

(2) IXB 指标的有效性对比测试

观察表 5 可知, IXB 在 5 个 UCI 数据集上的聚类数正确率为 80%, 而 DB、CH、XB 和 Sil 指标依次为 40%, 60%, 40%, 60%, IXB 指标的聚类数正确率明显高于其他四种指标.

3.2 聚类质量评价模型在 KDD CUP99 中的应用

本实验环节选取 KDD CUP99[16]训练集中的 17330 条记录作为训练数据, 从 corrected 数据集中随机抽取 11420 条数据作为测试集用于检验模型的性能.为提高整体运行效率, 在数据预处理方面, 首先使用独热编码完成字符数据的格式转换, 再通过属性简约法将数据集的 41 个特征约简为 15 个[17], 最后将数据集归一化处理, 形成新样本集, KDD CUP99 数据描述如表 6.

表 5 各内部评价指标聚类数对比

	标准类数	DB	CH	XB	Sil	IXB
iris	3	2	3	2	3	3
heart	2	3	2	2	2	3
ionosphere	2	2	4	3	2	3
wine	3	3	2	2	2	3
wdbc	2	3	2	2	4	2

表 6 KDD CUP99 数据集

	训练集	测试集
Normal	13 000	7000
DOS	2000	2000
PROBE	1800	2000
U2R	30	20
R2L	500	400

(1) 最优 k 值的获取

借鉴文献[15], 使用 AP 算法对样本集完成"粗聚类"得到训练集的最大聚类数 k_{max}=29, 如图 3 所示, 在训练过程中, 当 15≤k≤20 时, K 中心点算法的 IXB 增幅较大, 当 k=20 时, IXB 达到峰值, 此后, 随着 k 值的不断增大, IXB 总体呈下降趋势; 文献[8], 文献[9]和本文算法的 IXB 随着 k 值的不断增大而缓慢增加, 当 25≤k≤28 时, 三种算法的 IXB 依次达到峰值, 此后随着 k 值的再次增大, IXB 缓慢下降. 由定义 11 可知, 当 IXB 最大时 k 值即为最优, 因此, 三种算法使用 IXB 得到训练集的最优 k 值分别为: $k_{文献[8]}$=26, $k_{文献[9]}$=28, $k_{本文算法}$=28.

(2) IXB 对入侵检测指标的影响

在验证 IXB 是否有助于提高检测精度的环节中, 将图 3 中 IXB 缓慢上升至峰值, 再从峰值缓慢下降的这一阶段所对应的多个连续 k 值定义为最优聚类数范围, 并对该范围内的各入侵检测指标进行对比, 由于 K 中心点算法的随机性较强, 各项指标与 k 值之间无明显规律可循, 因此, 这里仅对其他三种算法的结果进行统计与分析.

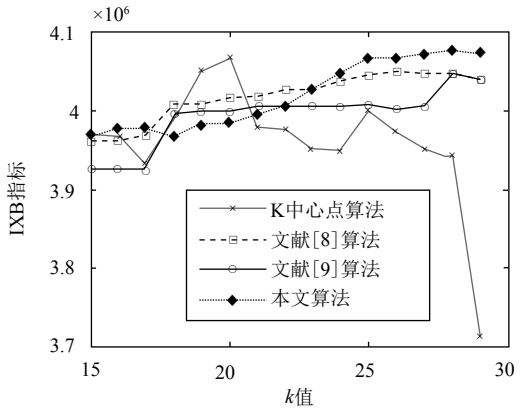

图 3 训练集的 IXB-K 的关系图

如图 4 所示, 当聚类数在[26, 28]范围内时, 3 种算法的检测率达到了最大, 分别是: 92.68%、91.69%、93.2%.

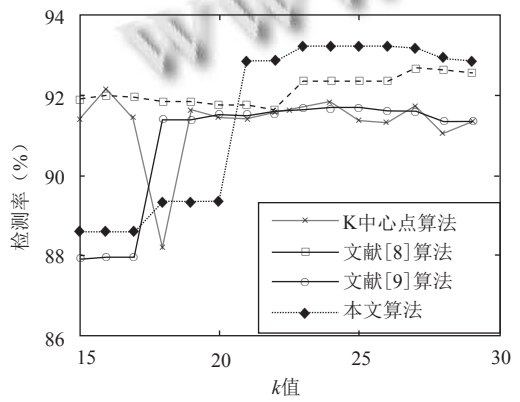

图 4 不同 k 值的检测率

从图 5 的漏报率对比结果中可以看出: 当聚类数在[26, 27]范围内时, 文献[8]和本文算法的漏报率最小, 分别是: 3.81%、2.86%.

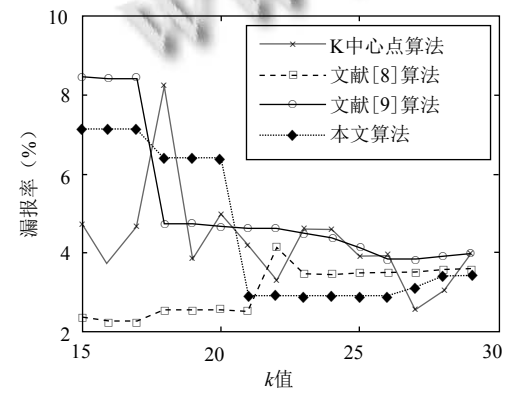

图 5 不同 k 值的漏报率

图 6 的正确分类率结果表明: 当聚类数在[27, 28]范围内时, 3 种算法的正确分类率达到最大, 分别是: 94.27%、94.38%、94.78%. 从图 4~6 可以看出, IXB 越大, 入侵检测指标越优. 综上所述, 本文提出的 IXB 能够合理、客观地评价聚类结果, 能够准确地反映出聚类质量, 可以为无先验知识样本集的有效聚类提供重要参考依据.

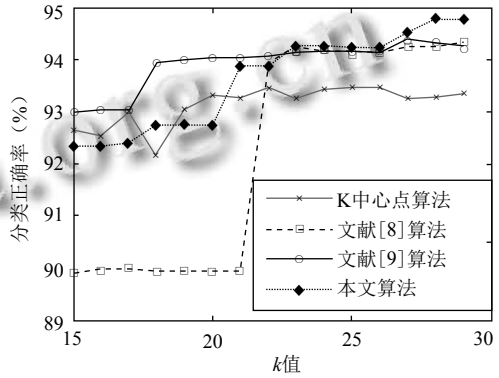

图 6 不同 k 值的分类正确率

5 结语

针对 K 中心点算法的初始聚类中心代表性不足, 稳定性差等问题, 提出了一种改进的 K 中心点算法, 将样本集与样本的各自平均距离比值作为样本的密度参数, 采用最大距离乘积法选择密度较大且距离较远的 k 个样本作为初始聚类中心, 兼顾聚类中心的代表性和分散性.针对常用内部评价指标在聚类评价中的局限性, 提出将簇间边界最近点的距离平方作为整个样本集的分离度, 定义了以簇内紧密度与簇间分离度的比值与其倒数之和为度量值的内部评价指标 IXB, 结合改进的 K 中心点算法设计了聚类质量评价模型.在 UCI 数据集的测试表明, IXB 指标的聚类数正确率明显高于其他四种常用指标, 在 KDD CUP99 数据集的实验结果表明, 本文提出的聚类质量评价模型可以给出精准的聚类数范围, 能够在保持较低漏报率的同时, 有效提高入侵检测率和分类正确率.

参考文献

1 唐益明, 丰刚永, 任福继, 等. 面向结构复杂数据集的模糊聚类有效性指标. 电子测量与仪器学报, 2018, 32(4): 119–127. [doi: 10.13382/j.jemi.2018.04.017]

2 邹臣嵩, 杨宇. 基于最大距离积与最小距离和协同 K 聚类

算法. 计算机应用与软件, 2018, 35(5): 297–301, 327. [doi: 10.3969/j.issn.1000-386x.2018.05.053]
3 邵东恒, 杨文元, 赵红. 应用 k-means 算法实现标记分布学习. 智能系统学报, 2017, 12(3): 325–332. [doi: 10.11992/tis.201704024]
4 刘美玲, 黄名选, 汤卫东. 基于离散量优化初始聚类中心的 k-means 算法. 计算机工程与科学, 2017, 39(6): 1164–1170. [doi: 10.3969/j.issn.1007-130X.2017.06.021]
5 徐鹏程, 王诚. K-Means 算法改进及基于 Spark 计算模型的实现. 南京邮电大学学报 (自然科学版), 2017, 37(4): 113–118. [doi: 10.14132/j.cnki.1673-5439.2017.04.018]
6 王国辉, 林果园. 基于图聚类的入侵检测算法. 计算机应用, 2011, 31(7): 1898–1900. [doi: 10.3724/SP.J.1087.2011.01898]
7 谢娟英, 周颖. 一种新聚类评价指标. 陕西师范大学学报 (自然科学版), 2015, 43(6): 1–8. [doi: 10.15983/j.cnki.jsnu.2015.06.161]
8 Van Der LaanM, Pollard K, Bryan J. A new partitioning around medoids algorithm. Journal of Statistical Computation and Simulation, 2003, 73(8): 575–584. [doi: 10.1080/0094965031000136012]
9 Park HS, Jun CH. A simple and fast algorithm for K-medoids clustering. Expert Systems with Applications, 2009, 36(2): 3336–3341. [doi: 10.1016/j.eswa.2008.01.039]
10 Davies DL, Bouldin DW. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, 2(2): 224–227. [doi: 10.1109/TPAMI.1979.4766909]
11 Caliński T, Harabasz J. A dendrite method for cluster analysis. Communications in Statistics, 1974, 3(1): 1–27. [doi: 10.1080/03610927408827101]
12 Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. Psychometrika, 1985, 50(2): 159–179. [doi: 10.1007/BF02294245]
13 Xie XL, Beni G. A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(8): 841–847. [doi: 10.1109/34.85677]
14 Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 1987, 20: 53–65. [doi: 10.1016/0377-0427(87)90125-7]
15 王开军, 张军英, 李丹, 等. 自适应仿射传播聚类. 自动化学报, 2007, 33(12): 1242–1246. [doi: 10.16383/j.aas.2007.12.017]
16 Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset. Expert Systems with Applications, 2011, 38(5): 5947–5957. [doi: 10.1016/j.eswa.2010.11.028]
17 吴建胜, 张文鹏, 马垣. KDDCUP99 数据集的数据分析研究. 计算机应用与软件, 2014, 31(11): 321–325. [doi: 10.3969/j.issn.1000-386x.2014.11.081]