基于交点的新层次聚类算法

李青旭,陈天鹰,胡 波
(华北计算机系统工程研究所,北京 100083)

摘 要:介绍了一种新的分层聚类算法,该聚类算法的主要目的是利用交点提供更好的聚类质量和更高的准确性。为了验证该聚类算法,对基准数据集进行了几次实验,并与其他五种广泛使用的聚类算法进行对比。使用纯度作为外部标准来评估聚类算法的性能,并计算了由聚类算法得出的每个聚类的紧密度,以评估聚类算法的有效性。实验结果表明,在大多数情况下,该算法的错误率低于研究中使用的其他聚类算法。

关键词:数据挖掘;无监督学习;聚类分析;聚类算法;分层聚类
中图分类号:TP393 **文献标识码**:A DOI:10.19358/j.issn.2096-5133.2020.10.004
引用格式:李青旭,陈天鹰,胡波. 基于交点的新层次聚类算法[J].信息技术与网络安全,2020,39(10):18-22.

New hierarchical clustering algorithm based on intersection

Li Qingxu, Chen Tianying, Hu Bo
(National Computer System Engineering Research Institute of China, Beijing 100083, China)

Abstract: This paper introduces a new hierarchical clustering algorithm. The main purpose of this clustering algorithm is to provide better clustering quality and higher accuracy by using intersections. In order to verify this clustering algorithm, we conducted several experiments on the benchmark data set. In addition to the algorithm we proposed, five well-known clustering algorithms were also used. The purity was used as an external standard to evaluate the performance of the clustering algorithm, and the tightness of each cluster obtained by the clustering algorithm was also calculated to evaluate the effectiveness of the clustering algorithm. Finally, the experimental results show that in most cases, the error rate of the proposed algorithm is lower than other clustering algorithms used in this study.

Key words: data mining; unsupervised learning; cluster analysis; clustering algorithm; hierarchical clustering

0 引言

由于处理的数据量每天都在增加,因此能够检测数据结构并识别数据集中的子集的方法变得越来越重要。聚类是这些方法中的一种。聚类或聚类分析是一项无监督的归纳学习任务,它基于各个点之间的相似性将数据组织到同质的组中。聚类是机器学习,是数据挖掘和统计中已研究的基本问题之一[1-3]。聚类方法可以产生与分类方法相同的结果,但是不存在预定义的类,因此也可以视为无监督分类[4-5]。

聚类算法的性能可以通过其发现数据集中某些或所有隐藏模式的能力来衡量,可以通过测量数据点之间的相似性(不相似性)来发现隐藏的模式。相似度表示在明确定义的意义上测得的数学相似度,通常使用距离函数进行定义,根据聚类算法的规则,可以测量数据点本身之间或数据点与某个特殊点之间的距离。同时,随着数据的划分,同一群集中的数据点应尽可能相似,而不同群集中的数据点应尽可能不相似[6-7]。多年来,已经开发出多种不同的聚类方法。1998 年,Fraley C 和 RAFTERY A E 将聚类算法分为层次结构和分区两组。Han 和 Kamber 在 2006 年将聚类算法分为 5 类:分层、分区、基于密度、基于网格和基于模型[8]。

JOHNSON S 定义的分层方法将点安排到一个基础层次结构中,该层次结构随后确定各种聚类[9]。层次聚类分为聚集和分裂两种类型。聚集方法具有自下而上的过程,首先将每个数据点放置在其自己的聚类中,然后将聚类连续合并为更大的聚类,或者直到满足给定的终止条件(例如特定数量的聚类)为止。分裂方法与聚集法相反,并且以自顶向下的方

式执行。分区方法将数据集划分为 K 个分区,每个分区代表一个聚类,它有两种类型的分区,即清晰分区和模糊分区。如果数据集的每个数据点仅属于一个簇,则称为“清晰”,但如果允许数据点成为多个具有不同程度的簇的成员,则称为“模糊”[10]。K-means 和 K-mediods 方法是两种常用的聚类方法。在 K-means 算法中,每个聚类由数据点的平均值表示,而在 K-mediods 中,一个聚类由聚类中位于最中心的数据点表示。

在基于密度的方法中,簇是数据空间中最密集的区域,被较低密度的区域隔开。ESTER M 等人1996年提出的空间聚类是基于密度的方法的一个示例,只要邻域中的密度超过某个阈值,该方法就会不断地增长聚类效果[11]。基于网格的方法将数据空间量化为有限数量的单元,这些单元形成一个网格结构,在该网格结构上执行所有用于聚类的操作,它与数据点无关,但与围绕数据点的值空间有关。基于统计信息网格是 WANG W 等人 1997 年提出的基于网格的方法对空间数据集进行聚类的典型示例,在这种方法中,将空间区域划分为由分层结构表示的矩形单元[12]。基于模型的聚类方法假定数据是由模型生成的,并尝试从数据中发现原始模型,统计方法和神经网络方法是基于模型的两种主要方法[13]。

本文的目的是在分层聚类的基础上优化分层算法,并使用更多的验证措施来证明提出算法的强度。该算法使用交点作为链接标准,以合理的计算复杂度提供更有效、更准确的聚类结果。该算法的第一步是为每个数据点找出最接近的邻居(NN),以形成对,然后找出对之间的交点以形成主聚类。本文以二维示例介绍了新的层次聚类算法,解释了聚类评估,并介绍了新层次聚类算法与某些现有聚类算法进行比较的实验结果。

1 新的聚类算法

与在各领域中使用的其他自下而上的层次聚类方法相比,本文提出的聚类算法的目的是提供具有相同计算复杂度 $O(n^2)$ 的、更准确和更有效的聚类,该算法利用神经网络和交叉点的优势进行分层聚类。交集是此算法中使用的基本概念。新算法的步骤如下:

(1)找到每个数据点的 NN 并将它们配对,因此,对于 n 个数据点,将有 n 对配对(其中一些会重复)。欧几里德距离公式(1)将用于测量数据点之间的距离(相似度)以形成 $n\times n$ 的距离矩阵,然后每个数据点将与其 NN 成对,数据点 IoP 的索引及其 NN IoNN 将用于配对。

$$\text{Pairs}=\{(\text{IoP}_1,\text{IoNN}_1),(\text{IoP}_2,\text{IoNN}_2),\cdots,(\text{IoP}_n,\text{IoNN}_n)\}$$

$$\text{NN}=\min_{v}\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{1}$$

其中,min 代表数据点的所有邻居,IoP_1 是第一个数据点的索引,IoNN_1 是第一个数据点的 NN 的索引。

(2)将每对视为包含一个数据点的索引和其 NN 的索引的集合,找到点对之间的交点并将其连接起来。通过找到集合之间的交点,在每一步都可以将所有最接近的数据点放置在同一簇中。

假设有 5 个数据点,并且在上一步中已经为每个数据点找到了 NN。Pairs={(1,2)(2,3)(3,2)(4,5)(5,4)}。第一对(1,2)显示第二个数据点是第一个数据点的 NN。请注意,应删除或忽略诸如(3,2)之类的重复对。

为了找到线对之间的交点,每次考虑两对点对,例如对于前两对$(\text{IoP}_1,\text{IoNN}_1)$和$(\text{IoP}_2,\text{IoNN}_2)$交点是$(1,2)\cap(2,3)=(2)$,可以通过以下计算得出:

第一对$=(\text{IoP}_1,\text{IoNN}_1)=(1,2)$

第二对$=(\text{IoP}_2,\text{IoNN}_2)=(2,3)$

交点$=(\text{IoP}_1,\text{IoP}_2)=(1-2)=-1$

$(\text{IoP}_1-\text{IoNN}_2)=(1-3)=-2$

$(\text{IoNN}_1-\text{IoNN}_2)=(2-3)=-1$

$(\text{IoNN}_1-\text{IoP}_2)=(2-2)=0$

$(\text{IoNN}_1-\text{IoP}_2)=(2-2)=0$ 表示 2 是第一和第二对之间的交点,通过考虑点对 1 与其他点对,可以找到所有具有交点的点对并将它们连接起来以构成一个聚类。

(3)计算上一步得出的每个聚类的平均值(中心)。

(4)连续对平均值重复步骤(1)和(2),以达到所需的簇数。这意味着需要找到主要聚类的每个平均值的 NN,然后制作成点对的平均值,再在下文中寻找成对的平均值之间的交点以形成聚类。重复此过程以获得所需的簇数,或者直到所有数据点都在一个簇中。该算法的流程图如图 1 所示。

假设有一个二维数据集,如图 2 所示。图 2 中的(a)部分显示每个数据点都与它的最近邻居配对,并且其中一些对于一个以上的数据点是最近邻居,这些交点用白色圆圈表示。在(b)部分中,将具有交点的那些对合并,并使其成为主要簇,计算每个主

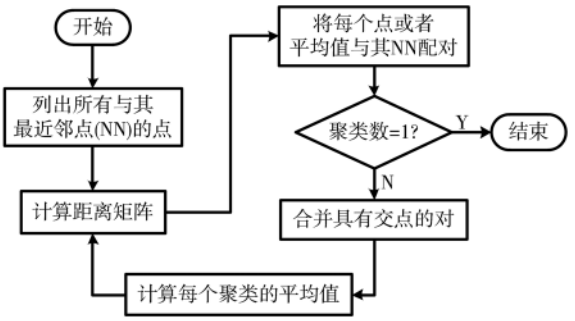


图 1 具有交点的新聚类算法流程图

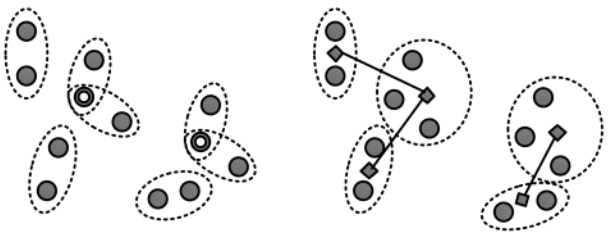


图 2 使用新的聚类算法进行数据聚类的图示(◆为平均值)

要簇的平均值。随后,求平均值和最接近均值来加入最接近的类集。重复此过程以达到所需的簇数,或者直到所有数据点都在一个簇中。图 3 和图 4 分别显示了新算法和现有算法的步骤。

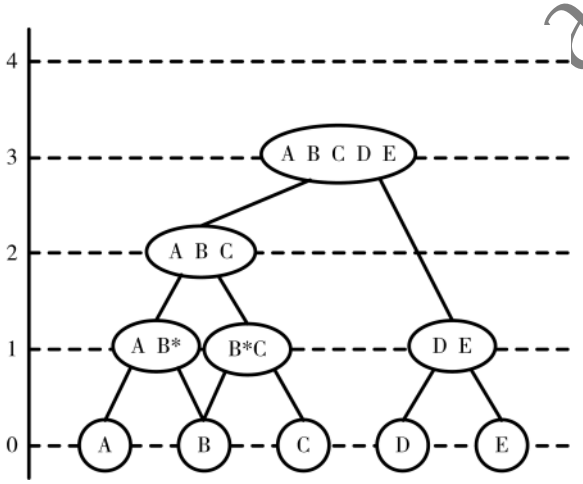


图 3 新算法的聚类步骤

2 聚类评估

在对数据进行聚类之后,用户希望检查聚类结构的有效性。聚类评估或聚类有效性检验是一项必要但具有挑战性的任务,已成为聚类分析中的核心任务。通常,验证措施分为内部和外部标准,内部标准基于数据的固有信息并分析聚类结构的优劣,而外部标准则基于有关数据的先前知识并分析聚类

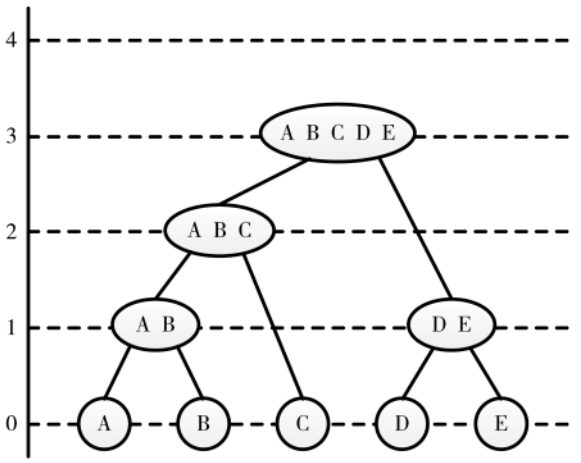


图 4 现有聚类算法聚类的步骤

与参考的标准。例如预定义的类标签[14]、边缘相关性、结构分析、邓恩指数和预期密度是内部标准的一些示例;兰德指数、归一化的互信息、F 度量和纯度是聚类质量的外部标准的一些示例。内部聚类评估措施经常对聚类结构的形成做出潜在假设,并且通常具有很高的计算复杂性。因此,当研究目的仅是评估聚类算法和可用的类别标签时,研究人员倾向于使用外部标准。由于本研究的目的是介绍和评估一种新的聚类算法,并且类别标签也可用于本文的数据集,因此本研究使用外部标准。纯度是一种流行的外部标准,本文使用纯度来评估本研究中使用的聚类算法。要计算纯度,应将每个簇分配给该簇中重复次数最多的类别,接下来,将通过对正确分配的数据点数除以总数据点数[15-16]来测量此分配的准确性。

3 实验

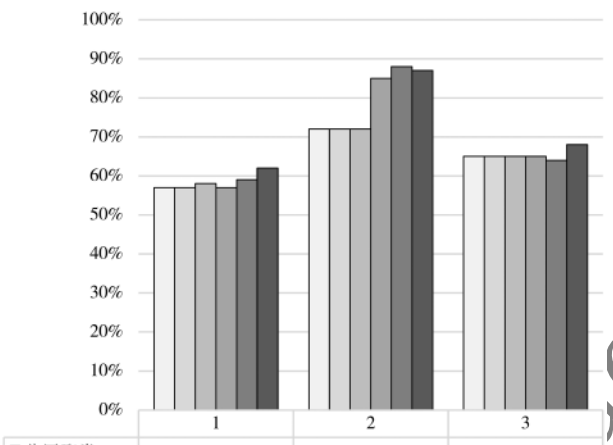
为了验证新的层次聚类算法,将其应用于三个数据集[17],数据集的详细信息如表 1 所示。所有提到的数据集都有预定义的标签,在不使用其类标签的情况下使用它们来形成聚类。将分层、分区、基于密度、基于网络和基于模型的层次聚类用于数据聚类,并将其结果与新聚类算法的结果进行比较。如前所述,在这些实验中,所有数据集都预定义了标签,因此可以使用它们来计算聚类算法的纯度,并

表 1 数据集特征

数据集名称	实例数	属性数量	类别数
肝病	345	7	2
甲状腺病	215	6	3
糖尿病	768	9	3

根据其准确性对结果进行排名。聚类方法的准确性如图 5 所示,毫无疑问,聚类结果总是需要最低的聚类间距和最高的聚类相似度。换句话说,每个聚类的成员应尽可能地彼此靠近,这被定义为紧密度,而方差是紧密度的常用度量,因此,本文还计算了数据集的每个聚类中的属性值之间的方差。通过比较每个聚类的方差值,可以测量每个聚类算法如何在一个聚类中设置相似的数据点。除了纯度之外,方差分析还可以帮助评估聚类的准确性。肝病数据集、甲状腺病数据集和糖尿病数据集的方差值如图6~图 8 所示。

图 5 介绍了应用于基准数据集的六种聚类算法的准确性,在肝病数据集(柱状图 1)和甲状腺数据集(柱状图 3)中,新的层次聚类算法的准确性高于其他聚类算法。图 6~图 8 介绍了每种聚类算法提供的总方差值,那些具有零密度簇的算法的总值以斜体的形式标注,它们不参与排名。较小的总方差值表示较高的聚类相似度,这是聚类算法所需的。对于肝病数据集,新算法相似度较高,对于甲状腺病数据集,分区聚类和新算法方差较小。对于糖尿病数据集,模型聚类和新聚类算法提供的方差小于其他。由以上实验结果可以看出,在本研究中使用的所有聚类算法中,本文提出的算法在准确率方面始终位于前 2 位。



	1	2	3
分层聚类	57%	72%	65%
分区聚类	57%	72%	65%
基于密度聚类	58%	72%	65%
基于网格聚类	57%	85%	65%
基于模型聚类	59%	88%	64%
新聚类算法	62%	87%	68%

图 5 聚类方法的准确性

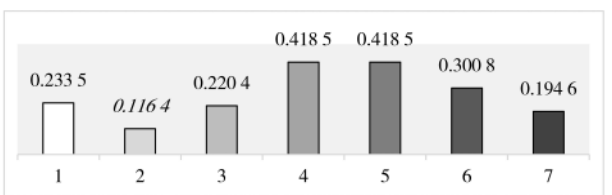


图 6 肝病数据集方差(1 为预定义标签)

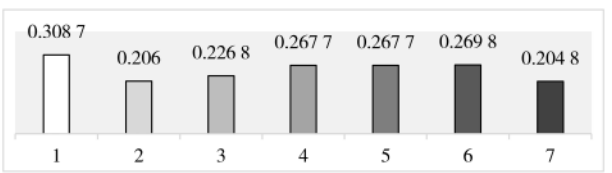


图 7 甲状腺病数据集方差(1 为预定义标签)

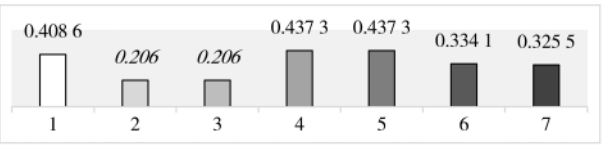


图 8 糖尿病数据集方差(1 为预定义标签)

4 结论

本文提出了一种新的基于交点的层次聚类算法,并进行了一些基准数据集实验,以验证提出算法的有效性。实验中还使用了其他五种层次聚类算法,使用纯度作为外部标准来评估所有聚类算法的聚类结果。计算每个聚类的属性方差值,并将其用于评估聚类质量。实验结果表明,在大多数情况下,无论数据点的大小和数量如何,具有计算复杂度为 $O(n^2)$ 的基于交叉点的新层次聚类算法都可以提供良好的结果,并且错误率较低。而具有相同计算复杂度的其他聚类算法仅在少数情况下才能很好地执行。

参考文献

[1] NAZARI Z, KANG D, ASHARIF M R, et al. A new hierarchical clustering algorithm[C]. 2015 International Conference on Intelligent Informatics and Biomedical Sciences, 2015, 8(2): 148-152.

[2] HAN J W, KAMBER M, PEI J. Data mining: concepts and techniques[M]. Massachusetts: Elsevier Inc., 2006.

[3] EVERITT B S, LANDAU S, LEESE M, et al. Cluster analysis[M]. John Wiley & Sons, Ltd., 2011.

[4] MOOI E, SARSTEDT M. A concise guide to market research[C]. Springer-Verlag, New York, 2014.

[5] TAN P N, STEINBACH M, KUMAR V. Introduction to data mining[M]. Addison Wesley, 2005.

[6] Liu Fasheng, Xiong Lu. Survey on text clustering algorithm-research present situation of text clustering algorithm[C]. 2011 IEEE 2nd International Conference on Software Engineering and Service Science, Beijing, 2011: 196-199.

[7] PATEL K M A, THAKRAL P. The best clustering algorithms in data mining[C]. 2016 International Conference on Communication and Signal Processing(ICCSP), Melmaruvathur, 2016: 2042-2046.

[8] FRALEY C, RAFTERY A E. How many clusters? which clustering methods? answers via model-based cluster analysis[J]. Computer Journal, 1998, 41(8): 578-588.

[9] JOHNSON S. Hierarchical clustering scheme[J]. Psychometrika, 1967, 32(3): 241-254.

[10] DUNN J C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters[J]. Journal of Cybernetics, 1973, 3(3): 32-57.

[11] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, Portland, 1996: 226-233.

[12] WANG W, YANG J, MUNTZ R R. STING: a statistical information grid approach to spatial data mining[C]. Proceedings of the 23rd VLDB Conference, Athens, Greece, 1997: 186-195.

[13] LIAO T W. Clustering of time series data-a survey[J]. The Journal of Pattern Recognition Society, 2005, 38(11): 1857-1874.

[14] WALDE S S I. Experiments on the automatic induction of German semantic verb classes[J]. Computational Linguistics, 2006, 32(2): 159-194.

[15] ROMESBURG H C. Cluster analysis for researchers[M]. LuLu Press, North Carolina, 2004.

[16] KOGAN J. Introduction to clustering large and high dimensional data[M]. New York: Cambridge University Press, 2007.

[17] LICHMAN M. UCI machine learning repository[EB/OL]. [2020-05-01]. http://archive.ics.uci.edu/ml/index.php.

(**收稿日期**:2020-06-12)

作者简介:

李青旭(1993-),**男,硕士研究生,主要研究方向:网络安全、自然语言处理。**

陈天鹰(1963-),**男,硕士,高级工程师,主要研究方向:网络安全、轨道交通一体化和自动控制。**

版权声明