基于内聚度和耦合度的二分 K 均值方法

郁 湧[1,2] 康庆怡[1] 陈长赓[1] 阚世林[1] 骆永军[1]
(云南大学软件学院 昆明 650504)[1] (云南省软件工程重点实验室 昆明 650504)[2]

摘 要 聚类分析是数据挖掘中最重要的技术之一,它在社会经济的各个领域都具有重要作用,并被广泛应用。K 均值算法是最经典、应用最广泛的聚类方法之一,但其缺点是过度依赖初始条件和聚类数目难以确定,这制约了其应用范围。引入簇的内聚度和耦合度的定义与度量方法,基于"高内聚低耦合"的原理,在二分 K 均值聚类过程中对得到的簇进行不断的分裂和合并,并判断聚类结果是否满足要求以确定聚类的次数和簇的个数,从而实现对二分 K 均值聚类过程的改进。在 Iris 数据集上的实验测试与分析表明该算法不仅更加稳定,而且其聚类结果的正确率也较高。

关键词 聚类,二分 k 均值,内聚度,耦合度
中图法分类号 TP391 **文献标识码** A

Bisecting K-means Clustering Method Based on Cohesion and Coupling

YU Yong[1,2] KANG Qing-yi[1] CHEN Chang-geng[1] KAN Shi-lin[1] LUO Yong-jun[1]
(School of Software, Yunnan University, Kunming 650504, China)[1]
(Key Laboratory for Software Engineering of Yunnan Province, Kunming 650504, China)[2]

Abstract Clustering analysis is one of the most important techniques in data mining. It has important role and wide application in every field of social economy. K-means is one kind of the simple and widely used clustering methods, but its disadvantage is that it depends on the initial conditions and the number of clusters is difficult to determine. This paper introduced the cohesion and coupling of cluster, and presented the measurement of cohesion and coupling. Based on the principle of "high cohesion and low coupling", the clusters are constantly divided and merged in the process of bisecting K-Means clustering algorithm. By judging whether the clustering results meet the requirements, it can determine the number of clusters, thus improving the bisecting K-Means clustering algorithm. The experimental results on Iris data show that the algorithm is not only more stable, but also has higher clustering accuracy.

Keywords Clustering, Bisecting K-means, Cohesion, Coupling

1 引言

随着社会经济技术的发展和大量数据的出现,数据挖掘在各个产业领域显得越来越重要。数据挖掘是从数据中发现有趣模式与知识的过程[1]。聚类分析是数据挖掘中最重要的技术之一,在数据的组织、分析和挖掘中具有重要作用,广泛应用于处理模式识别、信息检索、图像处理、机器学习等领域。K 均值算法是一种基于原型进行不断迭代的聚类技术,可用于处理许多类型的数据。在众多聚类方法中,K 均值算法是最经典的、应用最广泛的聚类方法之一,但其缺点是过度依赖初始条件,如聚类数目 K 值的确定、初始聚类质心的选取以及数据的输入次序的变化等都会影响聚类结果,制约了其应用范围。

二分 K 均值聚类算法是对基本的 K 均值算法的一个扩充。在二分 K 均值聚类过程中,为了得到 K 个簇,先将所有点分成两个簇,再从这些簇中选取一个簇继续进行分裂,直到产生 K 个簇[2-3]。二分 K 均值聚类算法能够产生划分聚类算法或层次聚类算法,具有受初始质心选择影响较小的优点[4]。文献[2]从效率、效果和可扩展性等方面对三层次方法、二分 K 均值、K 均值和后缀树聚类方法进行了比较,并证明二分 K 均值算法的性能在总体上优于其他聚类算法。文献[3]对层次聚类算法、K 均值聚类算法和二分 K 均值聚类算法进行了性能对比研究,结果表明二分 K 均值聚类算法的性能优于 K 均值聚类算法,具有与层次聚类算法相当的聚类质量,但其时间复杂度低于层次聚类的时间复杂度。在二分 K 均值聚类算法的研究方面,文献[5]采用层次聚类对二分法进行改进,解决了二分 K 均值算法受用户指定的聚类个数的影响的问题。文献[6]对二分 K 均值算法和 K 均值聚类算法进行比较分析,结果表明二分 K 均值算法的效果优于 K 均值聚类算法。文献[7]采用数据并行的思想和均匀划分的策略,提出以极大距离点作为二分聚类初始质心并对算法进行并行化处理,以提升二分 K 均值算法的运行速度。文献[8]将二分 K 均值聚类和 SVM 决策树相结合,提出一种可适用于高维数据聚类的自适应方法。文献[9]通过实验证明二分 K 均值算法的计算效率比顺序执行多次 K 均值算法的效率更高。文献[10]提出了核二分 K 均值聚类算法,用以减少 SVM 训练集样本,改善了 SVM 的可扩展性,加快了 SVM 训练算法的速度。文献[11]在基于整体和局部相似性的序列聚类算法

本文受国家自然科学基金项目(61462091),云南大学数据驱动的软件工程省科技创新团队项目(2017HC012)资助。
郁 湧(1980—),男,博士,副教授,CCF 会员,主要研究方向为软件工程、数据分析,E-mail:yuy1219@163.com。

中,利用带剪枝策略的二分 K 均值聚类算法对基于整体相似性的序列聚类通过启发式方法获得二分 K 均值聚类算法的质心,使整个算法关于序列数获得多项式时间复杂度。

由上述分析可知,二分 K 均值算法和经典的 K 均值算法类似,仍然要求用户指定聚类个数,其聚类效果往往受聚类个数的影响,并且传统二分 K 均值算法可能会产生分簇过细的问题。

二分 K 均值算法虽然有效,但是也存在一些不足之处,因此本文提出了簇的内聚度和簇间耦合度的概念和度量方法,基于"高内聚低耦合"的原理,在二分 K 均值聚类过程中通过不断判断聚类结果是否符合要求来确定聚类的次数和簇的个数,并对聚类过细子簇进行合并以解决过细分簇的问题,从而实现对二分 K 均值聚类过程的改进。

2　簇的内聚度和耦合度的定义与度量

聚类分析是一种研究"物以类聚"的科学有效的方法,聚类后同一类的数据尽可能被聚集到一起,不同数据被尽量分离。为了解决二分 K 均值聚类过程中需要用户指定聚类个数和可能会产生分簇过细的问题,本文提出簇的内聚度和簇间耦合度的概念,"物以类聚"的原理也就等同于聚类后各个簇的内聚度尽可能高,而簇之间的耦合度尽可能低,因此软件工程中的"高内聚低耦合"原理就可以作为二分 K 均值聚类中的评判标准。

2.1　簇的内聚度

给定 d 维数据集 $DB=\{X_1,X_2,\cdots,X_n\}$,其中 $X_i=\langle X_{i1},X_{i2},\cdots,X_{id}\rangle(i=1,2,\cdots,n)$ 为一个数据点,n 为数据点数目。将数据集 DB 中的数据划分为 k 个不同的集合 $S=\{C_1,C_2,\cdots,C_k\}$,且任取 $i\neq j,(C_i\cap C_j)=\varnothing$,则 $C_i(i=1,2,\cdots,k)$ 称为划分后的簇集 S 中的一个簇。

对于数据集 DB 中的两个 d 维数据 $X_i=\langle X_{i1},X_{i2},\cdots,X_{id}\rangle$ 和 $X_j=\langle X_{j1},X_{j2},\cdots,X_{jd}\rangle$,它们的欧氏距离定义为:

$$\| X_i,X_j \|=\sqrt{\sum_{t=1}^{d}(X_{it}-X_{jt})^2}$$

数据集 DB 的均方差为:

$$SD(DB)=\sqrt{\frac{\sum_{i=1}^{n}\| X_i-V_0 \|^2}{n}}$$

其中,V_0 为数据集 DB 的中心点。

$$V_0=\langle\frac{\sum_{i=1}^{n}X_{i1}}{n},\frac{\sum_{i=1}^{n}X_{i2}}{n},\cdots,\frac{\sum_{i=1}^{n}X_{id}}{n}\rangle$$

数据集 DB 中数据 X_i 到中心点 V_0 的距离为:

$$\| X_i-V_0 \|=\sqrt{\sum_{j=1}^{d}(X_{ij}-\frac{\sum_{i=1}^{n}X_{ij}}{n})^2}$$

簇的内聚度是指一个簇内各个元素彼此相似的紧密程度的度量。若一个簇内各元素之间越相似,则它的内聚性就越高。

对于一个簇 C_i 来说,簇 C_i 的内聚度为:

$$Cohesion(C_i)=\frac{1}{1+\frac{SD(C_i)}{SD(DB)}}$$

其中,$SD(C_i)$ 为簇 C_i 的均方差。

此处,簇的内聚度的度量是一个概率相对值,即假设原来数据集 DB 的内聚度为 0.5,介于内聚度高与低之间,新划分的簇的内聚度的高低由簇和数据集 DB 的均方差来共同确定,新簇的均方差大于数据集 DB 的均方差,则说明新簇的元素紧密程度低于数据集 DB 的紧密程度,即:

$$Cohesion(C_i)=\frac{1}{1+\frac{SD(C_i)}{SD(DB)}}<0.5$$

新簇的均方差小于数据集 DB 的均方差,则说明新簇的元素紧密程度高于数据集 DB 的紧密程度,即:

$$Cohesion(C_i)=\frac{1}{1+\frac{SD(C_i)}{SD(DB)}}>0.5$$

如果簇只有一个元素,则其均方差为 0,此时内聚度 $Cohesion(C_i)=\frac{1}{1+\frac{0}{SD(DB)}}=1$,取最大值。

数据集 DB 分成 k 个簇的集合 $S=\{C_1,C_2,\cdots,C_K\}$(C_i 是一个簇,$i=1,2,\cdots,k$)时的平均内聚度为:

$$MCohesion(S)=MCohesion(C_1,C_2,\cdots,C_k)=\frac{\sum_{i=1}^{k}Cohesion(C_i)}{k}$$

2.2　簇之间的耦合度

簇之间的耦合度是指一个数据集 DB 分为多个簇时,簇之间元素相似程度的一种度量。簇之间元素越相似,其耦合性越高,簇之间的差异性越低。

对于两个不同的簇 C_i 和 C_j,簇 C_i 与 C_j 之间的耦合度为:

$$Coupling(C_i,C_j)=\frac{1}{1+\frac{\| V_i,V_j \|}{SD(DB)+\frac{SD(C_i)+SD(C_j)}{2}}}$$

其中,V_i 和 V_j 分别表示两个簇 C_i 和 C_j 的中心,$\| V_i,V_j \|$ 表示两个簇 C_i 和 C_j 中心之间的距离。

一个簇在任何内聚度情况下都不具有耦合性;而对于不同的两个簇,簇之间的耦合性由簇中心点之间的距离、每个簇和数据集 DB 的均方差共同确定,在数据集 DB 的均方差一定的情况下,耦合度与簇中心点之间的距离成反比,与簇的均方差成正比。如果一个簇只有一个元素,则其均方差为 0。两个点之间的距离等于数据集 DB 的均方差时,它们之间的耦合度为 0.5。

两个簇之间的耦合度具有对称性,即 $Coupling(C_i,C_j)=Coupling(C_j,C_i)$

多个簇之间的耦合度度量:设数据集 DB 可分成 k 个簇的集合 $S=\{C_1,C_2,\cdots,C_K\}$,则簇集 S 中簇之间的耦合度度量为:

$$MCoupling(S)=MCoupling(C_1,C_2,\cdots,C_k)=\frac{1}{1+\frac{\| V_1,V_2,\cdots,V_k \|}{SD(DB)+\frac{SD(C_1)+SD(C_2)+\cdots+SD(C_k)}{k}}}$$

其中,$\| V_1,V_2,\cdots,V_k \|$ 为 k 个簇 $C_1,C_2,\cdots,C_K$ 中每两个簇的中心的平均距离。

3　二分 k 均值逐步聚类原理和判断准则

基于簇的内聚度和簇间耦合度的二分 K 均值逐步聚类的

思想是:为了得到 k 个簇,选择距离最远的两个数据作为初始聚类中心;先将原数据集分裂成两个簇,再从这些簇中选择合适的簇继续进行分裂,如此循环直到产生需要的 k 个簇为止。

对于二分 K 均值逐步聚类,需要解决两个问题:

1)在分裂后的簇集中,如何选择一个合适的簇来继续进行分裂;

2)二分 K 均值逐步聚类终止的判断标准是什么,即需要进行二分 K 均值聚类的次数是多少。

在子簇的选择方面,本文以簇的内聚度为基础,在聚类后的簇集中选择内聚度最小的一个簇作为合适的簇进行二分 K 均值聚类。根据 K 均值聚类的原理和簇内聚度的度量方法可知,聚类后的两个簇的均方差比原簇的均方差小,而初始数据集的均方差保存不变,因此通过二分 K 均值聚类得到的簇的内聚度都会增大。

但是,如果只考虑簇的内聚度平均值最大化,那么将会导致过度聚类,即每个簇只含一个数据时每个簇的类聚度取最大值 1,此时的结果毫无意义。为此,如果数据集 DB 能够分成 k 个簇的集合 $S=\{C_1,C_2,\cdots,C_K\}$,根据"高内聚低耦合"的原理,本文提出一个判断二分 K 均值逐步聚类细分过程终止的判断标准:

$$MAX\{\alpha MCohesion(S)-\beta MCoupling(S)\}$$

其中,$\alpha+\beta=1$。

在二分 K 均值逐步聚类的过程中,需要使得簇集 S 的平均内聚度尽量高而簇之间的平均耦合度尽量低,此时函数 $\alpha MCohesion(S)-\beta MCoupling(S)$ 应该取最大值。因此,在二分 K 均值逐步聚类中,使 $\alpha MCohesion(S)-\beta MCoupling(S)$ 值增大的一次二分 K 均值聚类是合理的聚类,否则是不合理的聚类,需要选择其他簇来进行二分 K 均值聚类。如此循环,直到 $\alpha MCohesion(S)-\beta MCoupling(S)$ 不再增大为止。

4 簇的合并原理和判断准则

在二分 K 均值聚类的过程中,可能会存在不合理的聚类,使得本该属于一个簇的数据元素被划分成两个耦合度很高的簇,此时就需要对簇集 S 中的一些簇进行合并。

对于簇的合并,也需要解决两个问题:

1)在聚类后的簇集 S 中,哪些簇需要进行合并;

2)合并过程终止的判断标准是什么,即需要进行几次簇的合并才能满足要求。

在需要合并的簇的选择方面,以簇之间的耦合度为基础,在聚类后的簇集 S 中选择耦合度最大的两个簇作为待合并的簇,把这两个待合并的簇放在一起得到一个更大的新簇。此时,再计算合并后的簇集的 $\alpha MCohesion(S)-\beta MCoupling(S)$ 值,如果值增大,则说明合并是合适的,需要进行合并;否则认为合并是不合适的,应该保持原来的簇不变。如此循环,直到 $\alpha MCohesion(S)-\beta MCoupling(S)$ 不再增大为止。

具体算法如下:

输入:d 维数据集 DB=$\{X_1,X_2,\cdots,X_n\}$,其中 $X_i=\langle X_{i1},X_{i2},\cdots,X_{id}\rangle$ (i=1,2,…,n)为一个数据点,n 为数据点数目

输出:簇集 S

1)设定 α 和 β 的值,使得 α+β=1;

2)选择距离最大的两个数据作为中心点,对所有数据进行二分 K 均值聚类,将产生的两个簇 C_1 和 C_2 加入到簇集 S 中,S=$\{C_1,C_2\}$;

3)Repeat//簇的分裂

计算簇集 S 中各个簇的内聚度、簇间耦合度和 αMCohesion(S)−βMCoupling(S)值;

从簇集 S 中选择出一个内聚度最小的簇;

在选出的簇中选择距离最大的两个数据作为中心点对簇中数据进行二分 k 均值聚类产生的两个簇,从簇集 S 中删除选择的簇,把新获得的两个簇加入到簇集 S 中以获得新簇集 S′;

计算簇集 S′的 αMCohesion(S′)−βMCoupling(S′)值,如果值 αMCohesion(S′)−βMCoupling(S′)>αMCohesion(S)−βMCoupling(S),则使 S=S′;

until αMCohesion(S)−βMCoupling(S)值不再增大为止;

Repeat //簇的合并

从簇集 S 中选择出两个耦合度最大的簇;

把选出的两个簇合并成一个新簇;

计算各个簇的内聚度、簇间耦合度和 αMCohesion(S)−βMCoupling(S)值,如果 αMCohesion(S)−βMCoupling(S)值增大,则从簇集 S 删除选定的两个簇,把新获得的簇加入到簇集 S 中;直到 αMCohesion(S)−βMCoupling(S)值不再增大为止或者满足簇的合并次数。

在聚类的过程中,如果不知道最终的聚类簇的数目 K,则 K 的取值由聚类的分裂和合并的次数以及 $\alpha MCohesion(S)-\beta MCoupling(S)$ 的值不再增大来决定。因此,在不知道聚类数量时也可以根据分裂和合并的情况以及 $\alpha MCohesion(S)-\beta MCoupling(S)$ 来获得聚类应该得到的簇的数目。

当然,如果能够知道最终聚类的簇数目 K,则簇的合并次数 M 与簇的分裂次数 D 之间的关系为:$M=D-K+1$。

5 过度聚类和聚类不足的问题及解决方法

在聚类过程中,由于数据和参数的不同会导致聚类过程中出现过度聚类和聚类不足的现象。过度聚类是指在聚类和二分过程中聚类会产生大量的无实用意义和效果的簇。根据簇类聚度的定义可知,如果一个簇中只有一个元素,则其均方差为 0,内聚度取最大值 1,因此在聚类和二分过程中,当参数 $\alpha=1,\beta=0$ 时,为了获得判断标准 $\alpha MCohesion(S)-\beta MCoupling(S)$ 的最大化,二分就会一直持续,直到每个簇中都只含一个元素为止,故极端情况就是 n 个数据的数据集获得 n 个簇。过度聚类主要是由于 α 的取值过大且 β 的取值过小而导致的。但是,如果少量的簇中只包含一个数据,则这些簇中的数据可能就是异常点或者孤立点,此时只要进行删除或者根据数据分析的目的进行特殊处理即可。对于过度聚类问题,可以适当减少 α 的取值来进行调整。

聚类不足则与过度聚类相反,是指聚类的结果中包含的簇太少,没有达到对数据集细分的目的。对不同数据进行实验分析可知,如果原始数据集为部分或者完全对称数据且 α 的取值过小时,可能会产生聚类不足的问题。如果在聚类和二分过程中得到的簇数远小于预期数目,则以调整 α 的取值和在选择簇进行分裂时选择多个簇来进行分裂,这样会得到更好的效果。

此外,在聚类过程中也会出现 $\alpha MCohesion(S)-\beta MCoupling(S)$ 值可能无法获得最优解的问题。对于聚类不足和 $\alpha MCohesion(S)-\beta MCoupling(S)$ 值无法获得最优解的情况,可以在从簇集 S 中选择簇进行分裂时选择内聚度最小的两个簇分别进行分裂,从而对这些情况进行改善。

6　实验分析

本文采用鸢尾花卉数据集 Iris 为基础，以基于簇的内聚度和耦合度的二分 K 均值方法对该数据集进行聚类和分析。在数据图中，不同颜色的数据点代表不同的簇，每次二分均值的过程使用随机取初始中心点的方式进行，并设定 $\alpha=0.4$，$\beta=0.6$，过程如下：

(1)初始状态如图 1 所示，此时 Iris 数据集全为一类。

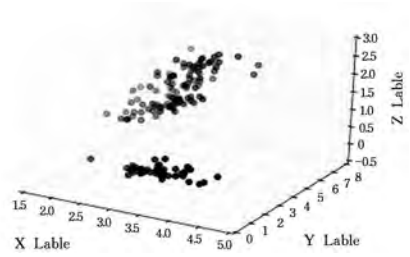

图 1　Iris 数据集

(2)二分过程。4 次二分的结果如图 2－图 5 所示。

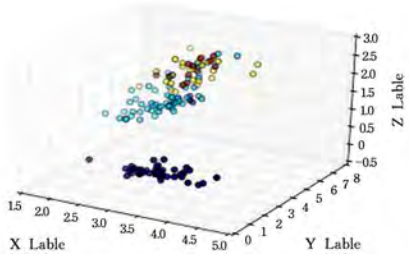

图 2　第一次二分

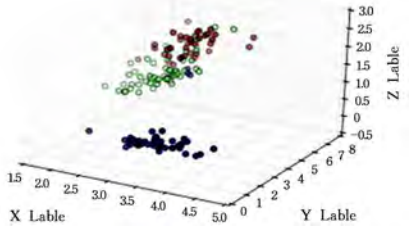

图 3　第二次二分

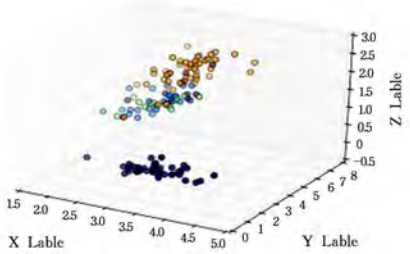

图 4　第三次二分

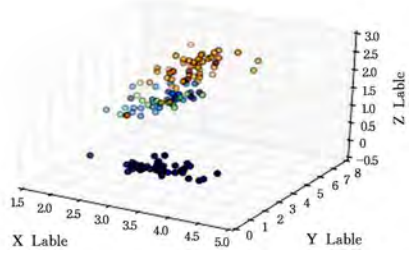

图 5　第四次二分

经过 4 次二分后满足二分过程结束的条件，二分过程结束，共产生 5 个簇。

(3)合并过程。已知 Iris 数据集的分类数 $K=3$，而一共进行的二分次数 $D=4$，因此合并次数 $M=4-3+1=2$，合并过程如图 6、图 7 所示。

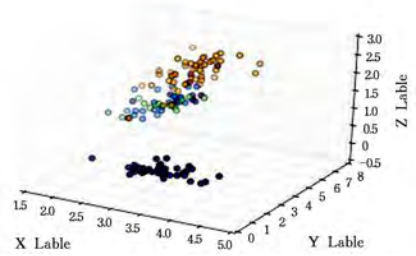

图 6　第一次合并

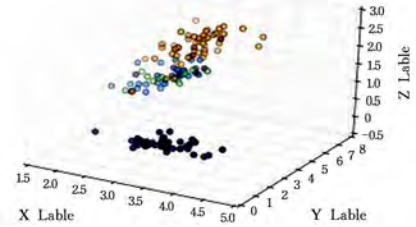

图 7　第二次合并

此时，合并过程结束，经正确率判别 Classification accuracy 算法验证，得到的结果的最终正确率为 89.8%。

在其他 α 和 β 下的正确率如表 1 所列。

表 1　不同 α 和 β 下的正确率

α	0.2	0.3	0.5	0.6	0.7	0.8	0.9
β	0.8	0.7	0.5	0.4	0.3	0.2	0.1
正确率/%	50	50	89.8	89.8	89.8	43.4	43.4

在相同正确率判别算法的情况和环境下，对于 DBSCAN 算法，当 $Eps=0.5$，$minpoints=3$ 时，运行结果的正确率为 50.7%；当 $Eps=0.6$，$minpoints=3$ 时，运行结果的正确率为 62.14%。对于 K 均值算法，$K=3$ 时，运行结果的正确率为 83.8%。

由上述分析可知，相较 DBSCAN 算法和 K 均值算法而言，本算法更加稳定，正确率也较高。

结束语　随着社会经济技术的发展和大量数据的出现，数据挖掘在各个产业领域显得越来越重要。聚类分析是数据挖掘中最重要的技术之一，具有广泛的应用领域。K 均值算法是一种基于原型进行不断迭代的聚类技术，可用于处理许多类型的数据。在众多聚类方法中，K 均值算法是最经典且应用最广泛的聚类方法之一。二分 K 均值算法和经典的 K 均值算法类似，仍然要求用户指定聚类个数，聚类效果往往受到聚类个数的影响，并且传统二分 K 均值算法可能会产生分簇过细的问题。

二分 K 均值算法虽然有效，但是存在一些不足之处，因此本文提出了簇的内聚度和簇间耦合度的概念和度量方法，将“高内聚低耦合”的原理作为二分 K 均值聚类中的评判标准。在二分 K 均值聚类过程中通过不断判断聚类结果是否符合要求来确定聚类的次数和簇的个数，并对聚类中得到的簇进

行分裂和合并,以解决过细分簇的问题,从而实现了对二分 K 均值聚类过程的改进。

参 考 文 献

[1] HAN J W,KAMBER M,PEI J. Data mining:concepts and techniques(3rd ed)[M]. Burlington:Elsevier Science,2011.

[2] ILLHOI Y,HU X H. A comprehensive comparison study of document clustering for a biomedical digital library MEDLINE [C]// Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries. New York,USA:ACM,2006:220-229.

[3] SILVA J D A,HRUSCHKA E R. Extending k-Means-Based Algorithms for Evolving Data Streams with Variable Number of Clusters[C]// International Conference on Machine Learning and Applications and Workshops. 2011:14-19.

[4] SAVARESI S M,BOLEY D. On the Performance of Bisecting K-Means and PDDP[C]// Proc. of the 1st SIAM International Conference on Data Mining. Chicago,USA:2001:1-14.

[5] 刘广聪,黄婷婷,陈海南. 改进的二分 K 均值聚类算法[J]. 计算机应用与软件,2015,32(2):261-263.

[6] VAMSI K B S,SATHEESH P,SUNEEL K R. Comparative Study of K-means and Bisecting K-means Techniques in Wordnet Based Document Clustering[J]. International Journal of Engineering and Advanced Technology,2012,1(6):119-234.

[7] 张军伟,王念滨,黄少滨,等. 二分 K 均值聚类算法优化及并行研究[J]. 计算机工程,2011,37(17):23-25.

[8] 裘国永,张娇. 基于二分 K-均值的 SVM 决策树自适应分类方法[J]. 计算机应用研究,2012,29(10):3685-3709.

[9] STEINBACH M,KARYPIS G,KUMAR V. A Comparison of Document Clustering Techniques[C]// Proc. of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston,USA,2000:525-526.

[10] LIU X Z,FENG G C. Kernel Bisecting K-Means Cluster- ing for SVM Training Sample Reduction[C]// Proc. of the 19th International Conference on Pattern Recognition. Tampa, USA, 2008:1-4.

[11] 戴东波,汤春蕾,熊赟. 基于整体和局部相似性的序列聚类算法[J]. 软件学报,2010,21(4):702-717.

(上接第 446 页)

[9] HIMMELSTEIN D S,BARANZINI S E. Heterogeneous Network Edge Prediction:A Data Integration Approach to Prioritize Disease-Associated GenesJ]. Plos Computational Biology,2015, 11(7):e1004259.

[10] IBRAHIM N M A,CHEN L. Link prediction in dynamic social networks by integrating different types of information[J]. Applied Intelligence,2015,42(4):738-750.

[11] 王祯骏,王树徽,张维刚,等. 基于社交内容的潜在影响力传播模型[J]. 计算机学报,2016,39(8):1528-1539.

[12] GAO F,MUSIAL K,COOPER C,et al. Link prediction methods and their accuracy for different social networks and network metrics[J]. Scientific Programming,2015,2015:1-13.

[13] CHEN G,WANG Y. Community detection in complex networks using extremal optimization modularity density[J]. Journal of Huazhong University of Science & Technology,2011,39(4): 82-85.

[14] LI Z,ZHANG S,ZHANG X. Modularity and community detection in bipartite networks[J]. American Journal of Operations Research,2015,5(5):421-434.

[15] BAKER A. Complexity,Networks,and Non-Uniqueness [J]. Foundations of Science,2013,18(4):687-705.

[16] KAYA B,POYRAZ M. Age-series based link prediction in evolving disease networks[J]. Computers in Biology and Medicine, 2015,63:1-10.

[17] GUIMERA R,SALES-PARDO M,AMARAL L A. Module identification in bipartite and directed networks[J]. Physical Review E,2007,76(2):066102.

[18] MICHAEL J. BARBER. Modularity and community detection in bipartite networks[J]. Physical Review E,2007,76(2):066102.

[19] EWMAN M E J. The Structure and Function of Complex Networks [J]. Siam Review,2003,45(2):167-256.

[20] MURATA T. Detecting communities from bipartite networks based on bipartite modularities[C]// 2009 International Conference on Computational Science and Engineering. 2009:50-57.

[21] LIU X,MURATA T. How does label propagation algorithm work in bipartite networks[C]// 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology(WI-IAT'09). 2009:5-8.

[22] RAGHAVAN U N,ALBERT R,KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics,2007,76(32):036106.

[23] FUJITA S,FUJINO A. Word Sense Disambiguation by Combining Labeled Data Expansion and Semi-Supervised Learning Method[J]. Acm Transactions on Asian Language Information Processing,2013,12(2):1-26.

[24] LIU X,MURATA T. Community Detection in Large-scale Bipartite Networks[C]// IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. 2009:50-57.

[25] DAVIS A,GARDNER B B,GARDNER M R. Deep South[M]. University of Chicago Press,1941.

[26] ERGUN G. Human sexual contact network as a bipartite graph [J]. Physica A,2002,308:483-488.

[27] ZHANG P,WANG D,XIAO J. Improving the recommender algorithms with the detectedcommunities in bipartite networks [J]. Physica A,2017,471:147-153.

[28] SCOTT J,HUGHES M. The Anatomy of Scottish Capital:Scottish Companies and Scottish Capital[J]. Economic History Review,1980,3(4):1900-1979.

[29] KARUMUR R P,NGUYEN T T,KONSTAN J A. Exploring the Value of Personality in Predicting Rating Behaviors:A Study of Category Preferences on MovieLens[C]// ACM Conference on Recommender Systems. 2016:139-142.