基于有效性函数的聚类算法*)

范 明 戴冠中 覃 森
(西北工业大学自动化学院 西安 710072)

摘 要 聚类作为一种无监督的学习方法,通常需要人为地提供聚类的簇数。在先验知识缺乏的情况下,通过人为指定聚类参数是不合实际的。近年来研究的聚类有效性函数(Cluster Validity Index)用于估计簇的数目及聚类效果的优劣。本文提出了一种新的基于有效性指数的聚类算法,无需提供聚类的参数。算法每步合并两个簇,使有效性指数值增加最大或减小最少。本文运用引力模型度量相似度,对可能出现的异常点情况作均匀化的处理。实验表明,本文的算法能正确发现特定数据的簇个数,和其它聚类方法比较,聚类结果具有较低的错误率,并在效率上优于一般的基于有效性指数的聚类算法。

关键词 聚类,有效性函数,无监督学习,引力

A Clustering Algorithm Based on Cluster Validity Indices

FAN Ming DAI Guan-Zhong QIN Seng
(College of Automation, Northwestern Polytechnical University, Xi'an 710072)

Abstract Being an unsupervised learning, clustering algorithm suffering from the limitation that the number of clusters is specified by a human user. However, it is not suitable for human to specify the cluster parameters while lacking of prior knowledge. The main purpose of cluster validity index studied recently is to estimate the number of clusters and the goodness or validity of the clusters. In this paper, we propose a new clustering algorithm based on cluster validity indices, which obviates the needs for cluster parameters. We choose at each step the merging of clusters that results in the greatest increase (or smallest decrease) in cluster validity index. And we take gravitation model as calculating similarity distance; dispose the abnormal points that could be appeared. Experimental results show that our methods can detect the number of clusters based on given dataset exactly, and the error rates of clustering results are lowerthan the other clustering algorithm. We outperform the other cluster algorithms based on cluster validity indices with higher efficiency.

Keywords Clustering, Cluster validity indices, Unsupervised learning, Gravitation

1 引言

聚类问题,就是将给定的数据集 $X=\{x_1,x_2,\cdots,x_N\}$,$x_i\in R^d$,$i=1,2,\cdots,N$,(N 是数据集 X 的维数,d 是 x_i 的维数)分成 c 个簇。通常的聚类算法,是将 c 的数目作为先验知识[1,2]。但是,当数据量非常大或者先验知识难以获得时,人为指定簇个数 c 是不合实际的。处理这类问题常用的方法是定义一个准则函数,随着类数的增加,准则函数下降,通过寻找下降的拐点来确定最佳初始分类。例如,Stan Salvador 和 Philip Chan 提出了 L 方法(L-method)来计算拐点[3],从而得到最优的聚类个数。更直接的方法是计算聚类有效性函数。该函数主要作用是估计最优的簇个数,以及衡量聚类效果的优劣。若数据集 X 有 m 种划分方式 $U_1,U_2,\cdots,U_m$,每种划分方式 $U_1,U_2,\cdots,U_m$ 的有效性函数分别为 $V_1,V_2,V_3,\cdots,V_m$,若 $V_{k1}\geqslant V_{k2}\cdots\geqslant V_{km}$ 则 $U_{k1}\uparrow U_{k2}\uparrow U_k\cdots\uparrow U_{km}$。$U_{k1}\uparrow U_{k2}$ 表示 U_{k1} 优于 U_{k2}。一般来说,有效性函数的最优值对应的聚类结果,总是希望簇内的距离尽可能小,而簇间距离尽可能大。近年来提出的有效性函数在一定程度上能提供较为合理的簇数 c,但如何有效得到正确的簇数仍然是开放的问题。文[4]用层次聚类的单连接法(Single linkage)、K-means 法以及模拟退火(SA)法对计算机产生的数据和真实数据聚类,计算不同簇数对应的有效性函数值,最优的值就是最佳聚类。但是类数的确定在一定程度上和特定聚类算法有关,不同的聚类方法得到的最佳类数可能会有差异。层次聚类的单连接法可发现链状、环状的簇,而 k-means 法仅对紧凑的球形簇有较好的聚类效果[1],即类数的确定随簇间距离度量方式的不同而不同,且这些聚类算法每执行一步就要额外计算有效性函数,对算法效率影响较大。Kim 和 Ramakrishna 提出了一种基于有效性函数的聚类方法,使用聚类有效性函数度量簇间距离[5]。如果两个簇合并之后有效性函数增加,则这两个簇认为是相似的。由于该方法每次计算有效性函数都要重算聚类中心和平方误差,因此时间复杂性较高。

本文提出了一种物理的方法度量元组之间的相似性,并对可能出现异常点的情况采取均匀化的方法。同时提出了一种新的有效性函数 CVI,为簇内的相似性和簇间的相似性的之差。本文的聚类方法,采取层次聚类算法的思想,初始时所有元组自成一类,每次迭代合并两个最相似的簇,合并的两个簇遵循使有效性函数值增加最大或减少最少的原则,经过迭代得到最终结果。此算法依靠寻求最大的有效性函数,这比人为给定聚类的簇数目要合理、简单和有效。经实验验证,本文的方法能较准确地发现最优的簇个数,对有重叠的数据、界线模糊的数据,以及多维数据有较好的处理效果,且本文的算法有较低的时间复杂度,与 k-means,S-link 算法相比,均具有较小的错误率。

*)国家 863 计划资助项目(项目批准号 2005AA147030)。范 明 硕士生,主要研究领域为信息处理,数据挖掘;戴冠中 博士生导师,教授,主要研究领域为自动控制,信息安全;覃 森 博士生,主要研究领域为复杂网络。

2 基于有效性函数的聚类算法

2.1 聚类有效性函数

近些年来提出的聚类有效性函数主要用来评价聚类效果的优劣和判断簇的最优个数[6,7]。而理想的聚类效果是具有最小的簇内距离和最大的簇间距离。

聚类有效性函数可分成两类：第一类为基于密度的有效性函数。这类有分裂函数(PC)[8]。PC 计算每个簇的密度。第二类有效性函数的主要思想是计算簇内距离和簇外距离的比例。这类包括 DB 函数，Dunn 函数(DI)，I 函数，Calinski Harabasz(CH)[4] 函数。以下为效果较为显著的有效性函数：

CH 函数定义如下：

$$CH=\frac{traceB/(k-1)}{traceW/(n-k)} \tag{1}$$

其中，$traceB=\sum_{j=1}^{k} n_j \| \mu_j-\mu \|^2$，$n_j$ 是第 j 个簇的元组数目，μ 是整个数据集的均值，μ_j 是第 j 个簇的均值。$traceW=\sum_{j=1}^{k}\sum_{i=1}^{n_j} \| x_i-\mu_j \|^2$。$n$ 是整个数据集元组的个数，k 是簇的个数。CH 函数计算簇间距离和簇内距离的比例，CH 值越大，代表聚类效果越好。

I 函数定义如下：

$$I(k)=\left(\frac{E_1\times D_k}{k\times E_k}\right)^p \tag{2}$$

$D_k=\max_{i,j=1}^{k} \| \mu_i-\mu_j \|$，$E_k=\sum_{i=1}^{n}\sum_{j=1}^{k}\mu_{ij} \| x_i-\mu_j \|$。其中，$k$ 是簇的数目，n 是数据点的总数。对于一个给定的数据集，E_1 是常数。p 用来调整不同的簇结构的对比。在本文中，p 取 2。随着 k 值增大，$1/(k\times E_k)$ 减小，而 D_k 值增大。这两项一个增加，一个减少，而 D_k 增加的幅度大于 $1/(k\times E_k)$ 减少的幅度。使聚类有效性函数 $I(k)$ 最大的 k 值，就是最优的簇个数。

2.2 相似度距离

假设 N 维空间有 M 个数据点，一些点彼此“距离”远，而另一些点彼此“距离”近。此处的距离指数据点间的相似性距离，从而如何度量数据点之间的相似性距离就成为区分数据点的关键。从拟物的思路[11]，将这个样本视作 N 维空间中的一些质点，为每个数据点赋予一个质量。由万有引力定律可知，任何两个物体之间都有相互作用的引力，引力的大小和两个物体的质量乘积成正比，和两者距离的平方成反比。将物体质量都简化为 1，我们得到：

$$F_{ij}=\frac{k}{\varepsilon+\| x-y \|_2^2} \tag{3}$$

其中，k 是常数，$\| x-y \|_2$ 为 N 维向量 x 与 y 的 2-范数(欧式距离)。ε 为小正数，若两个点位置完全相同，即欧式距离为 0，可调节使它们之间引力不至为 ∞。我们取 $\varepsilon=\min_{i,j=1,i\neq j}^{M}(\| x_i-y_j \|_2^2)$，$\varepsilon\neq 0$。

由于质点之间的引力与距离的平方成反比，质点之间的作用力对距离的变化比较敏感，呈非线性关系。簇内质点相距较近，彼此之间的作用力很大，即簇内相似性很大。而簇间质点相距较远，作用力较小，即簇间相似性很小。这样，我们就可以根据明显的作用关系，判断质点之间的相似性。

M 个质点两两计算之间的引力，得到矩阵 $(F_{ij})_M$。$(F_{ij})_M$ 是一个 $M\times M$ 的对角线元素为 0 的对称矩阵，即：

$$\begin{bmatrix} 0 & & & & \\ F_{21} & 0 & & & \\ F_{31} & F_{32} & 0 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ F_{M1} & F_{M2} & \cdots & \cdots & 0 \end{bmatrix} \tag{4}$$

此引力模型也有不足之处。当向量 x,y 之间的距离与其它点之间相比非常小，导致它们之间的引力特别大，有可能淹没其它点的作用。为弥补这一缺陷，我们采取均匀化的手段，对数据平均化处理。具体实现如下：对 $(F_{ij})_M$ 矩阵每行元素按从大到小排序，然后从最大的数值开始，与它相临的元素做几何平均：$F_{ij}=\sqrt{F_{ij}\times F_{i(j+1)}\times\cdots F_{i(j+m)}}$。如果处于聚类中心的某个点，仅和少数点距离非常近，那么该点和这少数几个点之间特别大的引力就可以通过和其它点引力的几何平均来抵消。m 的取值，根据质点的数目而定。在本文中，m 取 2。

2.3 基于有效性函数的聚类算法

前面提到的第二类有效性函数计算的是簇间距离和簇内距离的比例。I，CH 等聚类有效性函数用一个点(即聚类簇的重心)来代替整个簇。计算簇间距离时，若对于长方形的簇，它的各个子聚类簇的重心可能会距离相当远，就可能导致这个聚类簇被分开。这些有效性函数也不能计算在初始状态每个点都为单独的簇的情况。初始时簇内距离全为 0，那么有效性函数的初始值为 ∞，则有效性函数的比较就无任何意义[5]。

Newman 提出了一种快速分析复杂网络社团结构的方法，成功地应用于大型的网络社团划分[10]。通过最优化评价函数 Q(为社团内的连边数与社团外部连边数之差)的方法得到最佳的社团结构。本文借助于这种思想，针对质点的聚类，提出了一种新的有效性函数，为簇内相似度与簇间相似度之差。如果聚类使得簇间相似度之和尽可能的大，而簇内相似度之和尽可能小，则结果使有效性函数的值最大。

我们定义有效性函数：

$$CVI=\sum_i(d_{ii}-s_i^2) \tag{5}$$

即：$CVI=traced_{ij}-\sum_i s_i^2$。

其中，矩阵 $(d_{ij})_M$ 的元素 d_{ij} 是簇 i 的所有点与簇 j 所有点之间相似度之和。S_i 为所有的簇到簇 i 的相似度之和。

$S_i=\sum_j d_{ij}$。

既然使 CVI 值最大的聚类结果是最优的，那么一种直接方法是穷尽搜索，对所有可能的情况计算 CVI 值，从中选择最优值。但是这种方法的时间复杂性是很大的，不利于实际应用。本文采用层次聚类算法思想，选择使 CVI 增加最大或减少最小的两个簇合并，直到所有质点聚为一类时停止。具体算法如下：

Step1：初始化为 M 个簇，即每个节点为独立的簇。CVI 初始为 0。

初始的 d_{ij} 满足：$d_{ij}=\begin{cases}\frac{1}{2}\times\frac{F_{ij}}{\sum_i\sum_j F_{ij}}, & i\neq j\\ 0, & i=j\end{cases}$

Step2：合并 CVI 增加最大的两个簇 i,j，并计算合并后 CVI 的增量：

$\Delta CVI=d_{ij}+d_{ji}-2S_iS_j=2(d_{ij}-s_is_j)$。

对相应的 d_{ij} 更新，并将 $(d_{ij})_M$ 对应的行、列合并相加。则其它的簇与合并之后簇的距离可通过矩阵 $(d_{ij})_M$ 的合并得到。

Step3：重复执行步骤 2，不断合并最相似的两个簇，直到所有质点都合并为一个簇。最多执行 $M-1$ 次合并。

Step4：所有质点合并之后得到不同簇数 k 的聚类结果。在这些结果中，选择对应 CVI 值最大的，即为最佳聚类。同

时聚类的最优簇数 k 也随之确定。

当簇之间距离较大从而引力(d_{ij})很小时，由于 $\Delta CVI=2(d_{ij}-S_iS_j)$，计算出来的 ΔCVI 往往是负值。我们设定一个阈值，仅对簇间距离 d_{ij} 高于此阈值的计算簇 i 和 j 的 ΔCVI。对簇间引力从大到小排序，即 $d_1,d_2,\cdots,d_n$，则阈值取 $d=d_{s\%\times n}$，S 为 20 或更小。

在时间复杂性上，合并两个簇后，相似度距离矩阵对应的行和列相加，只需常数时间。每次迭代的时间复杂度是 $O(l)$，(l 为高于阈值 d 的簇数)。由于需要最多合并 $M-1$ 个簇，因此总的时间复杂度是 $O(M\times l)$。而其它的基于有效性函数的聚类算法[5]每一步都要重算聚类中心和平方误差，总的复杂度为 $O(M^3)$，复杂性较大。

3 实验结果

本文随机产生 3 组聚类数据，以及一组 UCI 机器学习数据库提供的数据集[9]。表 1 为实验数据集的描述。

表 1 实验所用的数据集

数据集	总的数据数目	簇数目	维数	单个簇的数据点数
数据集 1	500	10	2	50
数据集 2	750	15	2	50
数据集 3	400	4	3	100
Iris	150	3	4	50

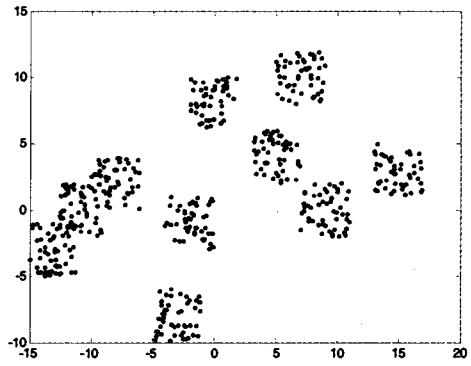

图 1 均匀分布的 10 个重叠的簇

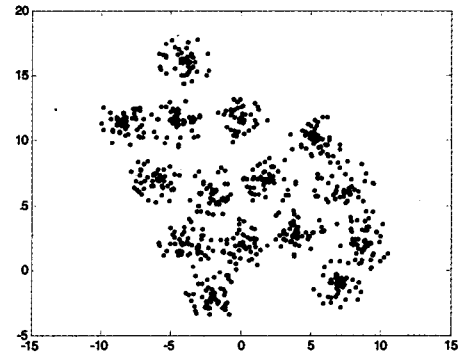

图 2 均匀分布的 15 个重叠的簇

图 1 为数据集 1，为均匀分布的 10 个方形簇，有 7 个簇彼此距离较大，分得较开。另外 3 个簇之间有重叠交叉部分。图 2 为数据集 2，为均匀分布的球形簇。有 15 个簇，彼此边缘比较模糊，边界交叉重叠的部分较为显著。图 3 是 3 维的数据集，共有 4 个簇。

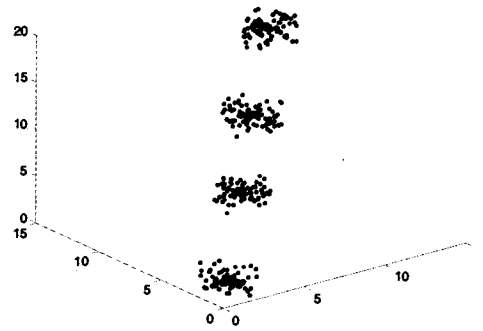

图 3 均匀分布的 4 个簇 3 维数据

表 2 对不同数据估计簇的数目

数据集	有效性函数	簇的数目
数据集 1	*I*	10
	CH	3
	CVI	10
数据集 2	*I*	9
	CH	4
	CVI	15
数据集 3	*I*	4
	CH	4
	CVI	4
Iris	*I*	3
	CH	2
	CVI	3

图 4 中横轴为不同的簇数 c，纵轴为 3 个数据集的 *CVI* 的变化值。随着横轴的簇数 c 由大变小，*CVI* 值缓慢的增加。到达峰值后，随着簇数的变小，*CVI* 值有一个陡峭的下降，簇数 c 为 0 时 *CVI* 接近于 0。从表 2 中可以看到，对于数据集 1，只有 I，*CVI* 能正确发现簇的数目，对于数据集 2，只有 *CVI* 能正确发现簇的数目。数据集 3 中，I，CH，*CVI* 都能正确发现簇的数目。对于真实数据 Iris，I，*CVI* 能发现簇的数目。所以，从表 2 可以得出，*CVI* 能正确发现四个数据集的个数，而其它的有效性函数则有不同程度的偏差。

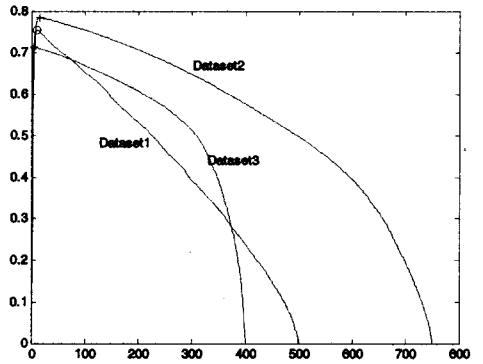

图 4 *CVI* 有效性函数变化曲线

在表 3 中，本文的方法和著名的 K-mens，S-Link 方法作比较。K-means 输入的 k 为正确的簇个数，S-Link 取正确簇个数的聚类结果。K-means 的聚类结果和初始值的选取有较大关系，我们初始随机选取 k 个点作为聚类中心。运行 K-means 方法多次，误差取平均值。我们取运行次数 20。虽然 K-means 方法输入的 k 是准确的簇个数，但对方形的簇处理

(下转第 207 页)

数据的一元线性回归方程。

故障模型最终由网络和主机两部分的每周故障数目曲线图以及系统总体每周故障数目的一元线性回归方程构成。

2.5 模型评价阶段

从数据分析的角度来说，进入此阶段，表明一个或者多个高质量的模型已经建立。在进入模型的最终部署之前，为了保证模型真正体现了商业/研究目标，应对模型进行全面而细致的评估，并且审查构造模型的每一个具体步骤。其中，一个重要的目的就是确保所有的目标和需求都得到了足够的重视。在该阶段的末尾，应该对数据挖掘结果的使用完成决策。

在该项目中，通过负载模型，数据挖掘人员得出以下结论：

1)网络设备的负载，包括其 CPU 利用率和缓存利用率，普遍高于主机类的所有负载指标，因此，应酌情对网络设备进行升级；

2)系统整体负载呈缓慢上升趋势，与系统维护人员的预期相符。

通过故障模型，数据挖掘人员得出以下结论：

1)网络部分的故障数量明显多于主机部分的故障数量，因此，在系统维护人员数量配置上，应向网络维护部分倾斜；

2)利用一元线性回归方程，预测系统还可以健康运行 3～4 年，与系统维护人员的预期相符。

通过系统维护人员的验证，表明上述结论具有相当的准确性和较高的参考价值。同时，也提出了修改意见，比如，增加负载模型的指标类型，以加大趋势的全面性；将故障模型的一元线性回归该为高阶多项式回归，以提高预测的准确性。

2.6 实际部署阶段

根据需求，该阶段可能只是简单的生成报告，也可能是复杂到在整个组织内实施一个可重复的数据挖掘流程，最重要的是，如何将建立的模型应用到实际环境中去。

该项目的实际部署阶段，就是交付给系统维护人员和高层决策人员的一份对指标数据和事件数据的数据挖掘报告，该报告的框架如下：

1)描述了数据挖掘的基础概念和将其技术应用到 IT 基础设施监控系统中的可行性；

2)介绍了数据挖掘流程标准 CRISP-DM 以及采用该标准的优势；

3)阐述了按照标准的数据挖掘流程，将具体的数据挖掘技术应用到 IT 基础设施监控系统中的过程；

4)分析了上述过程中可以改进的方法和步骤；

5)探讨了将其它数据挖掘方法应用到该系统中的可行性。

结论 数据挖掘是跨学科的前沿领域，能够从各个学科汲取营养，博采众长，为我所用；它也是一个通用的方法学，在将抽象的方法和具体的应用相结合的过程中发生矛盾，解决矛盾，不断进步。将数据挖掘技术应用到 IT 基础设计监控系统中，就是一个鲜活的例子。本文采用了数据挖掘中的时序数据挖掘技术和统计学中的一元线性回归技术，论述从大量的监控历史数据中发现有趣模式的过程，再次证明了数据挖掘的魅力所在。

参 考 文 献

1 Han J, Kamber M. Data Mining: Concepts and Techniques. Beijing: High Education Press, 2001

2 Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms. Washington: John Wiley & Sons, 2003

3 Daniel T. Larose. Discovering knowledge in data: an introduction to data mining. Washington: John Wiley & Sons, 2005

4 Hand D, Mannila H, Smyth P. Principles of Data Mining. Cambridge: The MIT Press, 2001

5 Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. Silicon Valley: Morgan Kaufmann, 2005

6 罗可，林睦纲，郗东妹. 数据挖掘中分类算法综述. 计算机工程，2005,31(1)：3～5,11

7 吴志勇，吴跃. 数据挖掘在电信业中的应用研究. 计算机应用，2005,25(S1)：213～214

8 王曙燕，周明全，耿国华. 医学图像的关联规则挖掘方法研究. 计算机应用，2005，25(6)：1408～1409

9 刘蓉，陈晓红. 利用数据挖掘技术提高电信企业管理决策水平. 计算机应用与软件，2005，22(12)：66～68

(上接第 199 页)

效果较差，劣于 CVI,S-Link，而对球状数据效果优于 S-Link，但劣于 CVI。实验表明，本文的方法有较小的错误率，优于 K-means,S-Link 方法。

表 3 本文方法和其它聚类方法错误率之比较

有效性函数	数据集 1	数据集 2	数据集 3	Iris
CVI	0.028	0.010	0.000	0.020
K-means	0.200	0.136	0.000	0.100
S-Link	0.072	0.092	0.002	0.300

结束语 本文提出了一种元组间新的相似度计算方式，对计算中可能出现异常点的情况做了平均化的处理，并提出了一种新的有效性函数。实验表明，该方法对有重叠交叉的簇、方形的簇有较好的计算效果，能正确发现簇的个数。本文的方法也能处理多维的数据集。由于有效性函数增量 ΔCVI 计算简单，算法复杂性较低，效率较高。本文的方法与 k-means,S-Link 聚类方法比较，具有较小的错误率。

参 考 文 献

1 Jain A, Murty M, Flynn P. Data clustering: a review. ACM Computing Surveys, 1999,31(3):264～323

2 Ka L,ufman, Rousseeuw P. Finding Groups in Data: An Introduction To Cluster analysis, New York: Wiley, 1990

3 Salvador S, Chan P. Determining the number of Clusters/segments in Hierarchical Clustering/Segmentation Algorithms. Tools with Artificial Intelligence, ICTAI 2004. In: 16th IEEE International Conference, 2004. 576～584

4 Maulik U, Bandyopadhyay S. Performance evaluation of some clustering algorithms and validity indices. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), 20002, 2004(12): 1650～1654

5 Kim M, Ramakrishna R S. A New Clustering Algorithm Based on Cluster Validity Indices. In: Proc. 7th International Conference, DS 2004, Padova, Italy, October 2004. 2～5

6 Halkidi M, Vazirgiannis M. Quality scheme assessment in the clustering process. In: European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD). Lecture Notes in Artificial Intelligence, 2000, 1910: 265～276

7 Schwarz G. Estimation the dimension of a model. Annals of Statistics, 1978, 6(1): 461～464

8 Bezdek J. Mathematical taxonomy with fuzzy sets. Journal of Mathematical Biology, 1974,1:57～71

9 http://www.ics.uci.edu/～mlearn/MLRepository.html

10 Newman M E J. Fast algorithm for detecting community structure in networks. Phys. Rev. E 69,066133,2004

11 卜东波. 聚类/分类理论研究及其文本挖掘中的应用:[中科院计算所博士论文]. 2000.10