聚类趋势问题的研究综述 *

褚 娜[1], 马利庄[1,2], 王 彦[1]
(1. 上海交通大学 电子信息与电气工程学院 计算机科学与工程系, 上海 200240; 2. 上海中医药大学 信息科学与技术中心, 上海 201203)

摘 要: 聚类算法的性能与数据集的结构是密切相关的,虽然目前已经研究出了很多聚类算法,但没有普遍适用的万能聚类算法,欠缺对数据集结构的有效解释。对聚类分析过程中重要的关键性问题,即聚类趋势问题进行了系统性的研究,从统计检验、可视化分析等角度给予了讨论,为数据集的无监督聚类分析提供了合理和有效的前期分析工具。

关键词: 聚类趋势; 聚类分析; 统计检验; 可视化评估
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2009)03-0801-03

Research for clustering tendency

CHU Na[1], MA Li-zhuang[1,2], WANG Yan[1]
(1. *Dept. of Computer Science & Engineering, School of Electronic Information & Electrical Engineering, Shanghai Jiaotong University, Shanghai* 200240, *China*; 2. *Center of Information Science & Technology, Shanghai University of Traditional Chinese Medicine, Shanghai* 201203, *China*)

Abstract: It is closely related between the performance of clustering algorithms and the structure of data sets. No methods were good enough for all types of data, nor were all methods equally applicable to all problems, and were short of reasonable interpretation to data structure. And systematic researched the clustering tendency, which was one of key problem about clustering analysis. This paper discussed it based on statistic tests, visual analysis and so on, and proved that it could present reasonable and effective analysis tools for unsupervised clustering analysis of data sets.

Key words: clustering tendency; clustering analysis; statistic tests; visual assessment

由于数据库技术和传感器技术的飞速发展、数据收集和数据存储技术的快速进步,使得各组织及研究机构积累了海量数据,另一方面,网络技术的发展也使获得大量数据变得较容易。然而在很多实际应用中,这些海量数据由于缺少形成模式类过程的知识,都是没有类别标签的。聚类分析技术能解决这一类问题帮助分析数据的分布、了解各数据类的特征、确定所感兴趣的数据类以便作进一步分析[1],寻找隐藏在数据中的结构。它将一组分布未知的数据进行分类,以尽可能地使得在同一类中的数据具有相同性质,而在不同类中的数据其性质各异[2]。目前,聚类分析技术已经在很多领域得到了成功的应用,如模式识别、图像处理、商业数据分析、市场研究、生物基因、信息安全、计算机视觉等,涉及面非常广泛,并且提出了各种聚类算法[3~7],如比较典型的 K-means 均值算法、FCM 等等。然而,大多数聚类分析算法对输入参数是敏感的,且都存在一个不合理的假设:待分析的数据集是可聚的[8]。事实上,现有的大多数聚类算法并不分析数据集的可聚性,只要对数据集进行聚类操作就能得到一个聚类结果。因此,这一不合理假设的存在会产生两个问题:a)如果数据集在空间中是均匀分布的,即自然簇不存在,对数据集进行聚类操作,显然得到的聚类结果是不合理的,且是不可解释的;b)因某种聚类算法并不是对所有的数据类型或者数据结构的分析是适用的[9],若不先对数据集进行聚类趋势分析,盲目地选择聚类算法,相当于是硬性地对数据施加一定的结构,不仅浪费了时间和精力,反而得到错误的结果。

为了避免聚类算法的不适当应用,也为了帮助更有效地解释聚类结果和了解数据本身的自然属性,并且降低将错误的结构强加给数据集的可能性,在应用聚类算法之前进行聚类趋势分析是很重要的[10]。

1 聚类趋势的定义

聚类趋势或簇聚倾向分析就是试图评估数据集中是否存在自然簇,即检验数据集是否具有类分性,而不是进行聚类[4,11]。从统计学的角度,就是检测数据集是否是随机或者均匀分布的[12]。在一些文献中,也有将这个问题称为关于模式在特征空间中分布类型的预测[4,13]或者聚类能力的检测[14]。

2 聚类趋势分析的原理

近年来,在聚类趋势分析理论方面的研究已经做了很多工作,提出了许多方法,并在许多领域内得到了应用。例如,在信息检索中应用聚类趋势分析检测词汇在自然语言中的分布情况,帮助提高检索的效率[15~17];在化学领域,聚类分析用于寻

收稿日期: 2008-07-02; **修回日期**: 2008-09-29 **基金项目**: 国家"973"计划资助项目(2006CB504801); 上海市科委世博科技专项资助项目(06dz05815)

作者简介: 褚娜(1976-),女,辽宁锦州人,博士研究生,主要研究方向为智能信息处理、数据挖掘、计算机图形图像学等(cina19@163.com); 马利庄(1963-),男,教授,博士,主要研究方向为计算机图形学、智能信息处理; 王彦(1973-),女,讲师,博士研究生,主要研究方向为数据挖掘.

找药品结构之间的关系[18,19]；对绿茶中抗氧化剂成分进行采样分析时，利用聚类趋势分析进行孤立点的检测[20]等。

2.1　基于统计检验方法的聚类趋势分析

基于统计检验方法的聚类趋势分析依赖于零假设，即数据集合在测量空间中是均匀分布的。早期的工作主要局限于低维空间中，在假设抽样窗口已知的条件下，应用空间抽样原理，基于最近邻模式或最小生成树的方法[21,22]建立统计量，如 Hopkins、Holgate、Cox-Lewis[23,24]、*T*-平方统计量[10]和 Eberhardt 统计量[25]，辨析数据集的空间结构，研究聚类趋势问题。其中 Hopkins 统计量在应用稀疏数据集的情况时效果欠佳[26,27]，而且在扩充到多维空间中时也不如 Cox-Lewis 统计量检测性能优良[28]；Cox-Lewis 统计量虽然在高维空间检测中存在优势，但有时得到的检测结果是片面的、不稳定的[29]；*T*-平方统计量是在实践中最经常被使用的，并表现良好。根据 *T*-平方采样原理[30,31]和 *k*-近邻模式距离，推广到 $d(d\geqslant 3)$ 维空间中的 *k*-近邻 *T*-平方统计量如下：

$$T_k=(1/M)\sum_{i=1}^{M}U_i^d(k)/[U_i^d(k)+0.5V_i^d(k)] \tag{1}$$

其中：*M* 为原始采样点 *O* 的个数；*U* 和 *V* 分别是原始 *O* 点到 P_1 点以及 *T*-平方采样 P_2 点的 *k*-近邻距离，如图 1 所示。

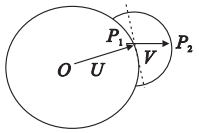

图1　*T*-平方采样原理

T-平方统计量的定义中，*T*-平方统计量不仅依赖于数据集中的模式向量与采样窗口内选择的抽样始点向量之间的距离，还依赖于数据集以及采样窗口的个数[32]，而且对抽样窗口的设置有要求。曾广周人等虽然在文献[23,25]中对抽样的矩形窗口、球形窗口以及球—矩形窗口的设置进行了分析探讨，并在低维空间中取得了良好的效果，然而样本窗的推导需要计算数据集的凸包，其计算代价很高；后来曾广周又提出半框框架制约条件，将样本窗定义为一个超球体，以数据集的均值点为中心，包含原来向量的一半[8]，拓宽了零假设范围，在一定程度上减弱了抽样窗口的设置问题。改进后的 *T*-平方统计量定义如下：

$$T_k=\{(1/M)\sum_{i=1}^{M}U_i^d(k)/[U_i^d(k)+0.5V_i^d(k)]-0.5\}\times 12M \tag{2}$$

然而总体来说，这些抽样窗口在高维空间中的设置仍然是困难的，直接影响了检验统计量的分布，并不能很好地发挥作用。综合以上讨论，基于统计检验方法的聚类趋势分析，对于选择正确的模型、窗口参数等是非常关键的，具有挑战性。

2.2　基于图论方法的聚类趋势分析

如果把数据看做是多维空间中的点，整个数据集就可以看做是一个无向图 *G*，由一组顶点 *X* 和连接顶点的边 *E* 构成。

$$G=[X,E],X(G)=\{x_1,x_2,\cdots,x_n\}$$
$$E(G)=\{e_{ij}=(x_i,x_j)\mid x_i,x_j\in X\} \tag{3}$$

定义 1　完全 *k*-部图。图 *G* 中的顶点 *X* 被划分为 *k* 个完全不相交的子集，如果来自不同子集中的任意一对顶点在图 *G* 中都是相邻的，那么就称这个图是 *k*-部完备的，即 *k*-部完全图。

根据着色原理，对于任意一个数据集，总是存在着一个对应着的 *k*-部图。根据定义 1 可以得到聚类趋势指数的定义：

定义 2　聚类趋势指数 *IC*[33,34]。一个数据集的聚类趋势指数是指使得该数据集成为 *k*-部完全图而需要补充的边数，即

$$IC=(1/2)\sum_{i=1}^{k}\sum_{x\in X}[d_G^{\max}(X)-d_G(X)]=(n^2-\sum_{i=1}^{k}n_i^2)/2-m \tag{4}$$

归一化处理：

$$IG=1-m/E_{\max} \tag{5}$$

其中：*m* 为图 *G* 的总边数；n_i 为第 *i* 个顶点子集包含的顶点个数；$E_{\max}$ 为该数据集 *k*-部完全图的边数。

如果 $E_{\max}=0$，表示所有的顶点都是不相邻的，完全 *k*-部图中没有边存在，此时分配聚类趋势指数为 −1，着色方法得到的所有顶点颜色将是相同的，即数据集中不存在聚类；反之，如果所有的顶点都是相邻的，即该数据集是一个完全 *k*-部图，着色方法需要使用 *n* 种颜色来标记图中的顶点，即数据集中每个顶点自成一类。如果上述两种情况不存在，则该数据集具有可聚类趋势。

2.3　基于可视化方法的聚类趋势分析

簇数的初始化问题一直是传统聚类方法的难点。可视化操作的目的不仅是为了显示聚类的结果，而且也能显式地观察数据集是否具有可聚性，优化传统聚类分析方法的输入参数，如 FCM、K-means 方法的预先类别数 *K* 值的初始化。如果数据集的样本量很大或者特征属性很多，或者样本之间具有强关联性等，在这种情况下，用简单的散点绘图方法观察数据的空间分布情况是不可行的。

2002 年，Bezdek 等人[35]提出了聚类趋势分析的可视化评估算法 VAT(visual assessment of tendency)。该方法提出基于最小支撑树的剪枝原理对相异度矩阵 ***R*** 进行重排序，使得具有小相异度的数据样本被分配相对邻近的索引指数，再将其元素值转换成灰度信息进行可视化，就可以明显判断出该数据集所具有的潜在聚类结构，即聚类趋势信息。实验证明[36]，VAT 对低于 500 个小样本的数据集有很好的效果；对大样本数据集，其重排序操作将造成严重的计算负担，具有高计算时间复杂度 $O(n^2)$，并且相异度图像矩阵的维数也有可能超过显示器可显示的分辨率。为了解决对大样本或关联数据集进行聚类趋势分析时遇到的问题，Huband[36]和 Hathaway 等人[33]分别提出了 reVAT 算法(revised VAT)及 sVAT(scalable VAT)算法。利用伪排序和采样技术对重排序操作过程进行改进，用相异度图的二维剖面图集合代替了 VAT 中的相异度图矩阵，其迭代过程仅运行 *c* 次。其中 *c* 是数据集包含的类别个数，大大缩减了排序的操作次数，降低了计算时间复杂度 $O(cn)$，具有更强的适应特征维数增加的能力和计算效率。剖面图 p_i 和阈值 δ 计算如下：

$$p_i=[p_{ij}]:\ p_{ij}=1-R_{ij}/g_M;\ j=1,2,\cdots,n \tag{6}$$
$$\bar{p}_i=\sum_{j=1}^{n}p_{ij}/n \tag{7}$$
$$\delta=(1+\bar{p}_i)/2 \tag{8}$$

其中：g_M 是表示相异度矩阵 ***R*** 中的最大值，表示数字灰度方图里的最高灰度级。然而，当数据集自然簇数($c>5$)变得很大时，或者当数据集的各簇没有明显簇内紧凑和簇间分离的特点时，使用 reVAT 方法得到的剖面图是混杂的，具有部分重叠区域，于是对剖面图的解释将变得很困难。针对 VAT 以及 reVAT 存在的问题，Huband 等人[34]提出 bigVAT(VAT for large datasets)算法，其中保留了 VAT 以及 reVAT 算法具有的优势，在建立剖面图的基础上进行 *N* 采样，得到一个 $N\times N$ 类 VAT 相异度图 I_{bV}，*N* 值上限限制为 500 个采样样本。采样过程如下：

$$I_{bV} = \lfloor R_{tiij}/g_M \rfloor 或者 \lceil R_{tiij}/g_M \rceil \tag{9}$$

$$\begin{cases} t_i \in T_1, 如果\ 1 \leqslant i \leqslant n_1 \\ t_i \in T_2, 如果\ 1+n_1 \leqslant i \leqslant n_1+n_2 \\ \cdots \\ ti \in T_k, 如果\ 1+\sum_{j=1}^{k-1} n_j \leqslant i \leqslant \sum_{j=1}^{k} n_j \end{cases} \tag{10}$$

其中:t_i 表示在剖面图中采样得到的相异度矩阵 R 中元素的下标;T_i 表示剖面图中大于阈值 δ 的下标集合。图 2[34] 给出了 40 000个样本的集合 S_8 的实验结果,使用 reVAT 算法处理后,在第 8 个剖面图出现了重叠区域,不能明确地得到聚类数是 7 个还是 8 个,还需要其他的辅助方式进行确定。而 bigVAT 算法对 S_8 的处理,沿主对角线方向的黑块显示了聚类趋势,可为传统的聚类分析算法提供良好的先验知识。

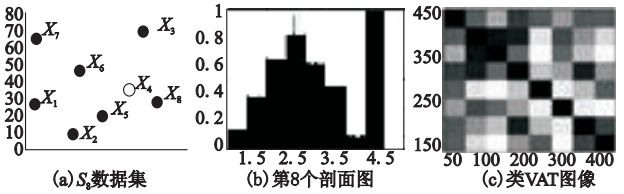

图2　bigVAT算法处理结果

综上所述,VAT 在现有软/硬件条件的限制下能有效地对小样本数据集进行可视化的聚类趋势分析,但对大样本和关联性数据集的处理是困难的,其改进方法 reVAT、bigVAT、sVAT 基本解决了这些问题。同时,bigVAT 算法亦解决了随着样本集聚类数目增加所带来的剖面图难以解释的问题。但是上述方法都只针对相异度矩阵是方阵的情况,随后 Bezdek 等人[37] 对此技术进行了拓展,提出基于矩形相异度矩阵的可视化聚类趋势评估方法,提高了聚类趋势可视化评估方法的应用范围和能力,并有良好的应用效果,但是计算复杂度很高。

3　结束语

为避免聚类算法的不合理应用,以及为传统聚类算法提供更多的先验知识,聚类趋势问题的研究已引起很多研究者的重视。本文主要讨论了基于欧几里德距离方法的聚类趋势检验和可视化评估问题,对于避免聚类算法对单一类或均匀分布的数据集的不适当应用起到了良好的效果,有助于聚类结果的正确解释。但是对于实际数据来说,其空间分布形状和大小一般是不知道的,所以常用的距离方法未必能提供良好的模式空间分布状态信息,因此必须注意这个问题;除此之外,对于不同领域的数据集,应根据实际应用作具体的聚类趋势问题分析[38]。下一步笔者将致力于寻找恰当的、更低计算时间复杂度的方法,进行聚类趋势问题的研究。

参考文献:

[1] TAN Pang-ning, STEINBACH M, KUMAR V. 数据挖掘导论[M]. 范明,范宏建,译. 北京:人民邮电出版社, 2006.

[2] HAN Jia-wei, KAMBER M. Data mining: concepts and techniques [M]. San Fransisco: Morgan Kaufmann Publishers, 2000.

[3] JAIN A K, MUTRY M N, FLYNN P J. Data clustering: a review [J]. ACM Computing Surveys, 1999, 31(3):264-323.

[4] JAIN A K, DUBES R C. Algorithms for clustering data[M]. New Jersey: Prentice Hall, 1988.

[5] 蔡元苹,陈立潮. 聚类算法研究综述[J]. 科技情报开发与经济, 2007,17(1):145-146.

[6] ANDERBERG M R. Cluster analysis for application[M]. New York: Academic Press, 1973.

[7] 行小帅,焦李成. 数据挖掘的聚类方法[J]. 电路与系统学报, 2003,8(1):59-67.

[8] 高新波. 模糊聚类分析及其应用[M]. 西安:西安电子科技大学出版社, 2004.

[9] HALKIDI M, VAZIRGIANNIS M. A data set oriented approach for clustering algorithm selection[C]//Proc of the 5th European Conference on Principles of Data Mining and Knowledge Discovery. London: Springer-Verlag, 2001:165-179.

[10] 曾广周. 随机分布模式的统计检验[J]. 山东工业大学学报, 1986,16(4):12-18.

[11] THEODORIDIS S, KOUTROUMBAS K. Pattern recognition[M]. 2nd ed. New York: Academic Press, 2003.

[12] CROSS G R, JAIN A K. Measurement of clustering tendency[C]// IFAC Symposium on Digital Control. New Delhi:[s. n.], 1982:24-29.

[13] DUBES R C, JAIN A K. Validity studies in clustering methodologies [J]. Pattern Recognition, 1979,12(11):235-254.

[14] EPTER S, KRISHNAMOORTHY M, ZAKI M. Clusterability detection and initial seed selection in large data sets[EB/OL]. 1999-06 (2004). http://libra.msra.cn/papercited.aspx? id=108767.

[15] DUBIN D. Document analysis for visualization[C]//Proc of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1995:199-204.

[16] DUBIN D. Structure in document browsing spaces[D]. Pittsburgh: University of Pittsburgh, 1996.

[17] DUBIN D. Clustering tendency and the cluster hypothesis in information retrieval[C]//Proc of Annual Meeting of the Classification Society of North America. 1996.

[18] LAWSON R G, JURS P C. New index for clustering tendency and its application to chemical problems[J]. Journal of Chemical Information and Computer Sciences, 1990,30(1):36-41.

[19] WILLETT J. Similarity and clustering in chemical information systems [M]. New York: Wiley, 1987:138-142.

[20] ZHANG M H, LUYPAERT J, FERNÁNDEZ PIERNA J A, *et al*. Determination of total antioxidant capacity in green tea by near-infrared spectroscopy and multivariate calibration[J]. Talanta, 2004,62(1): 25-35.

[21] ZENG G, DUBES R C. A test for spatial randomness based on *k*-NN distance[J]. Pattern Recognition Letters, 1995,3(2):85-91.

[22] SMITH S P, JAIN A K. Testing for uniformity in multidimensional data[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1984,6(1):73-81.

[23] PANAYIRCI E, DUBES R C. Extension of the Cox-Lewis method for testing multidimensional data [J]. Pattern Recognition, 1988, 7 (1):1-8.

[24] VINAY V, COX I J, MILIC-FRAYLING N, *et al*. On ranking the effectiveness of searches[C]//Proc of the 29th Annual International ACM SIGIR Conference on Research and Development. New York: ACM Press, 2006:398-404.

[25] 曾广周. 聚类趋势的 Monte-Carlo 检验[J]. 计算机应用与软件, 1989,6(1):35-50.

[26] LAWSON R G, JURS P C. Cluster analysis of acrylates to guide sampling for toxicity testing[J]. Journal of Chemical Information and Modeling, 1990,30(2):137-144.

[27] FERNANDEZ P J A, MASSART D L. Improved algorithm for clustering tendency[J]. Analytica Chimica Acta, 2000,408(1-2):13-20.

[28] BANERJEE A, DAVE R N. Validating clusters using the Hopkins statistic[C]//Proc of IEEE International Conference on Fuzzy Systems. Piscataway: IEEE Press, 2004:1749.

(下转第 822 页)

得这个权重向量制作调查表,多次反馈修正后获取到准确的权重向量,然后直接将这个向量运用到 AHP 方法中进行组合权重的计算。这样做也消除了专家调查法主观因素过大的弊病,利用这种方法的综合后,专家调查法得到的结果只是整个评估过程中权重向量这一部分,再放到 AHP 方法中计算时,可以通过模型的完整及严密性抵消人为主观性带来的误差。

2)人工神经网络方法与模糊分析法的综合运用 模糊分析法中也需要使用指标体系的权重向量,这个向量一般也可以通过专家调查法获取直接使用。这个主观判断的权重向量比较粗糙,且模糊分析法的模糊评价等级集已经是粒度比较粗的判断标准。为了提高模糊分析法的精度,可以将专家调查法获得的权重指标先运用人工神经网络进行训练,通过人工神经网络的自主学习功能,多次计算待权重向量值稳定下来后,权重向量精度会得到很大提高,再利用这个训练过的权重向量进行模糊综合评价,可以提高效能评估的精确度。

4 结束语

效能评估是现在电子信息系统日趋复杂的情况下很有必要做的一个分析优化工作,本文介绍的六种方法在各自适用的情况下都能够很好地评估出系统当前的效能状况,为系统的进一步改进提供方向性的指导。同时六种方法又有着各自的优缺点,在作效能评估时可以根据需要平衡这些优缺点选择合适的方法,并且根据系统的实际情况采用多种方法综合的方式取得更加准确有效的系统效能。而且文中还指出了运用各种方法时需要注意到的关键点,这些关键点对该种模型计算结果的正确性起着至关重要的影响。正确应用了合适的效能评估方法,采用了合理的综合运用方式,就可以客观地得到准确、定量的系统效能值,为后续的工作提供量化依据。

参考文献:

[1] 潘高田,周电杰,王远立,等. 系统效能评估 ADC 模型研究和应用[J]. 装甲兵工程学院学报,2007,21(2):5-7.

[2] 刘果靓. 综合航空电子系统效能评估研究[D]. 西安:西北工业大学,2006.

[3] 于开峰,吴德伟,王晓薇. 卫星导航系统作战效能评估[J]. 航天电子对抗,2007, 23(3):5-8.

[4] 缪小鹏. 防空信息作战效能的评估技术[D]. 武汉:华中科技大学,2005.

[5] 王宇. ADC 方法在通信星座效能评估中的应用[J]. 科教文汇,2007(3):195-196.

[6] 谢振华,倪成敏. 基于层次分析和模糊数学的电解铝生产安全评价[J]. 安全科学技术,2008(1):8-11.

[7] 陈根忠,薛业飞. 基于层次分析法的引导式电子对抗系统效能评估[J]. 电子对抗,2006(1):26-29.

[8] 梁权,石丽美. 基于模糊分析法的第三方物流顾客满意度综合评价[J]. 中国市场,2007(49):46-47.

[9] 李斌,刘建强,董奎义. 基于神经网络的舰载直升机反潜武器系统作战效能评估[J]. 舰船电子工程,2007,27(5):50-52.

[10] 白松浩,马锦永,吕善伟. 基于神经网络对管制中心系统的效能评估[J]. 电讯技术,2004(6):125-128.

[11] 张梅荣,姜玉英. 多属性决策方法及其应用[J]. 北京印刷学院学报,2007,15(2):72-75.

[12] 吴智辉,张多林. AHP 法评估地空导弹武器系统效能[J]. 战术导弹技术,2003(4):8-12.

[13] 张文胜. 基于模糊综合评判的动态路径行程时间预测模型[J]. 地理与地理信息科学,2006,22(4):25-28.

[14] 李士勇. 工程模糊数学及应用[M]. 哈尔滨:哈尔滨工业大学出版社,2004:96-103.

[15] 刘树林. 多属性决策理论方法与应用研究[D]. 北京:北京航空航天大学,1997.

[16] LIBERATORE M J, NYDICK ROBERT L. Group decision making in higher education using the analytic hierarchy process[J]. Research in Higher Education, 1997,38(5):593-614.

[17] LIN Yong-huang, LEE P C, CHANG Ta-peng, *et al*. Multi-attribute group decision making model under the condition of uncertain information [J]. IEEE Automation in Construction, 2008,17(6):792-797.

[18] SEAVER B, TRIANTIS K, HOOPES B J. Efficiency performance and dominance in influential subsets: an evaluation using fuzzy clustering and pair-wise dominance[J]. Journal of Productivity Analysis, 2004,21(2):201-220.

[19] SIMON HAYKIN. Neural networks:a comprehensive foundation[M]. 2nd ed. New Jersey: Prentice Hall: 1998:60-75.

(上接第 803 页)

[29] PANAYIRCI E, DUBES R C. Test for multidimensional clustering tendency[J]. Pattern Recognition, 1983,16(4):433-444.

[30] SAKKATAT P. Estimation of number and density, and random distribution testing of important plant species in Ban Pong forest Sansai district, Chiang Mai province, Thailand using *T*-square sampling[J]. Maejo International Journal of Science and Technology, 2007,1(1):64-72.

[31] SAKKATAT P. *T*-square sampling[J]. Bang Phra Center Journal, 1997(4):27-30.

[32] FORINA M, LANTERI S, ESTEBAN D I. New index for clustering tendency[J]. Analytica Chimica Acta, 2001,446(1-2):59-70.

[33] HATHAWAY R J, BEZDEK J C, HUBAND J M. Scalable visual assessment of cluster tendency for large data sets[J]. Pattern Recognition,2006,39(7):1315-1324.

[34] HUBAND J M, BEZDEK J C, HATHAWAY R J. bigVAT: visual assessment of cluster tendency for large data sets[J]. Pattern Recognition, 2005,38(1):1875-1886.

[35] BEZDEK J C, HATHAWAY R J. VAT: a tool for visual assessment of (cluster) tendency[C]//Proc of International Joint Conference on Neural Networks. Piscataway:IEEE Press, 2002:2225-2230.

[36] HUBAND J M, BEZDEK J C, HATHAWAY R J. Revised visual assessment of (cluster) tendency (reVAT)[C]//Proc of International Conference of North American Fuzzy Information Processing Society. Piscataway:IEEE Press, 2004:101-104.

[37] BEZDEK J C, HATHAWAY R J, HUBAND J M. Visual assessment of clustering tendency for rectangular dissimilarity matrices[J]. IEEE Trans on Fuzzy Systems, 2007,15(5):890-903.

[38] 包志强,韩冰,吴顺君. 空间相关噪声下信源个数的聚类检测算法[J]. 测试技术学报, 2006,20(5):444-450.

[39] 曾广周. 聚类趋势检验中的窗口设置[J]. 山东工业大学学报,1985,15(4):29-36.

[40] 曾广周. 球—矩形窗口在聚类趋势检验中的效[J]. 山东工业大学学报:工学版, 1987,17(1):39-45.

[41] SILVA H B, BRITO P, da COSTA J P. A partitional clustering algorithm validated by a clustering tendency index based on graph theory [J]. Pattern Recognition, 2005,39(5):776-788.