改进的共享型最近邻居聚类算法

耿 技[1]，印 鉴[2]

(1. 电子科技大学计算机科学与工程学院 成都 610054; 2. 中山大学信息科学与技术学院 广州 510275)

【摘要】聚类效果往往依赖于密度和相似度的定义，并且当数据的维增加时，其复杂度也随之增加。该文基于共享型最近邻居聚类算法SNN，提出了一种改进的共享型最近邻居聚类算法RSNN，并将RSNN应用于高速公路交通数据集上，解决了SNN算法在“去噪”、孤立点和代表点的判断、聚类效果等方面的不足之处。实验结果表明，RSNN算法比SNN算法在时空数据集上具有更好的聚类效果。

关 键 词 聚类分析; 共享型最近邻居; 孤立点; 相似度
中图分类号 TP391 文献标识码 A

Refined Shared Nearest Neighbor Clustering Algorithm

GENG Ji[1]，YIN Jian[2]

(1. School of Computer Science and Engineering, UEST of China Chengdu 610054;
2. School of Information Science and Technology, SUN Yat-sen University Guangzhou 510275)

Abstract Clustering results often depend on density and similarity critically, and its complexity often changes along with the augment of sample dimensionality. This paper refers to classical shared nearest neighbor clustering algorithm (SNN) and refined shared nearest neighbor clustering algorithm (RSNN). By applying this RSNN algorithm on freeway traffic data set, we settled several problems existed in SNN algorithm, such as outliers, statistic, core points, computation complexity and so on. Experiment results prove that this refined algorithm has better clustering results on multi-dimensional data set than SNN algorithm.

Key words cluster analysis; shared nearest neighbor; outlier; similarity

聚类分析源于许多研究领域，如数据挖掘、统计学、生物学以及机器学习等[1]。在数据挖掘领域，研究工作主要集中在聚类方法的可伸缩性，对复杂形状和类型数据的聚类有效性，高维聚类分析技术，以及针对大型数据库中混合数值和分类数据的聚类方法等几方面。

聚类是一个颇具挑战性的研究领域，不同的应用有各自特殊的要求，如可伸缩性、用于决定输入参数的领域知识最小化、处理噪声数据的能力、对于输入记录的顺序不敏感、高维性、可解释性和可用性等。本文将描述一个改进的共享型最近邻居聚类算法RSNN，并将其应用于时空数据集。

1 共享型邻居聚类算法

Levent Ertoz等人提出了一种基于共享型邻居聚类算法SNN[2-3]。该算法的基本思想为：先构造相似度矩阵，再进行最近k邻居的稀疏处理，并以此构造出最近邻居图，使得具有较强联系的样本间才有链接。然后统计出所有样本点的链接力度，以此确立聚类中心和噪声数据，将噪声数据从样本点中排除出来，并再次对图中的链接进行一次过滤。最后依据确定的聚类中心和剩下的最近邻居图来进行聚类处理。

该算法有效地实现了对交通时空数据集的聚类，并具有很高的可伸缩性和处理噪音的能力，同时也具有对输入样本的顺序不敏感、输入参数的领域知识最小化等特点，但也存在以下不足之处：

(1) 孤立点的预处理不够，导致计算增多。

SNN算法对于孤立点的处理非常有限，必须直到对所有样本点建立了SNN图，并计算了所有样本点的

收稿日期：2005 – 09 – 06

作者简介：耿 技(1963 –)，男，在职博士生，副教授，主要从事数据处理与数据挖掘、信息安全方面的研究；印 鉴(1968 –)，男，博士，教授，博士生导师，主要从事数据处理与数据挖掘、人工智能方面的研究.

链接力度之后才开始判断是否孤立点。经过算法的复杂度分析，计算相似度矩阵和构造SNN图所需的复杂度最大，都是$O(M^2)$。因此，对孤立点的预处理不足，直接导致了过多的无谓计算。

(2) 用来确定代表点、孤立点、以及用于过滤链接力度的阈值没有明确定义。

虽然经过对样本数据的统计，能得出用于确定代表点、孤立点等的阈值，但由于统计步骤本身具有较大的时空复杂度，这无疑额外增加了整个算法的复杂度。

(3) 代表点的确定过程不够全面。

直接通过阈值来确定代表点的方法，存在一个问题：确定出来的代表点极有可能同属于或者部分同属于同一聚类中。也就是说，几个代表点可能处于同一稠密区域内。但对后期聚类来说，代表点最好能分散一点。

2　改进的共享型最近邻居聚类算法RSNN

针对经典SNN算法中的不足之处，不难发现，问题的出现主要集中于：相似度、密度的确定，孤立点的识别和去除方法等。只要能顺利解决这些问题，算法的改进就能逐步实现。

2.1　相似度

在Jarvis-Patrick模式中，共享型最近邻居图(snn图)是利用相似度矩阵构建出来的。假定p和q是两个样本点，那么它们之间的相似度可如下定义：

$$\text{similarity}\ (p, q) = \text{size}(NN(p) \cap NN(q))$$

式中　$NN(p)$、$NN(q)$分别对应p和q的最近邻居。基于该定义，可以通过移除低于某一用户给定阈值的所有边，并将所有相连的样本点作为聚类而获得，这被称为Jarvis-Patrick聚类[2-3]。

2.2　密度

考虑到SNN相似度健壮性好，能很好地处理高维数据，本文采用SNN相似度、k个最近邻居度量方法来进行密度度量。如果一样本点的第k个最近邻居与它本身很近，就意味着有较高的SNN相似度，也可断定该点处于高密度区域内。SNN相似度反映数据空间内局部样本点构造情况，它相对来说对密度的变化和空间维都并不太敏感，因此可以作为新密度的度量。

2.3　孤立点

文献[4]中展示了一个孤立点去除的算法，效果显著。它是基于分包循环，并合成了随机选择的特性。所以，本文采用该算法在进行相似度计算之前执行，实现对孤立点的初步去噪。

算法中距离的度量采用基于连续特征的欧式距离或者离散特征的Hamming距离。而score函数则必须满足关于最近邻居距离单调递减，比如到第k个最近邻居的距离，或者到k个最近邻居的平均距离。

该算法的主要思想是，对于初始样本集中的每个样本点，都视为候选孤立点，并尽可能地找到它的最近邻居。当发现某个样本的最近邻居获得了一个低于cutoff值的score时，就将它从候选孤立点集中删除出去，因为它不可能为孤立点，同时不断调整cutoff值。此过程不断循环，直到得到了需要的n个孤立点为止。

2.4　改进的共享型最近邻居聚类算法

针对经典SNN算法中的不足之处，并重点讨论了相关改进，本文提出了下述改进的共享型最近邻居聚类算法RSNN。算法步骤如下：

(1) 调用孤立点去除算法，对初始样本集进行初步“去噪”处理，得到经过粗略筛选后的样本集。

(2) 基于该样本集，构建相似度矩阵。

(3) 采用k个最近邻居稀疏方法对相似度矩阵进行稀疏操作。

(4) 用上述稀疏后的相似度矩阵构造共享型最近邻居图。

(5) 对图中的每个样本点，计算它的全局链接力度linkstrength。

(6) 将具有低linkstrength值的样本点，定义为孤立点，也即噪声，并同时将其从初始样本点集中删除出去，实现对样本集的“瘦身”，得到求精后的样本集。

(7-10) 基于求精后的样本集，重新执行前面的步骤(2)～(5)。

(11) 将具有较高linkstrength值的样本点确定为“粗糙”代表点集。

(12) 对求得的代表点集进行求精，也即将“相对距离”较近的点去除，将“相对距离”较远的点保留下来，作为聚类过程的中心点。

(13) 根据此时的共享型最近邻居图和确定的代表点进行聚类，使得聚类中的每个样本点，或者是代表点，或者对于各代表点而言，与某类的中心点(代表点)最近。

3 实 验

实验数据集源自美国高速公路网络，并记录为数据矩阵模式，其中，行坐标表示各检测点，列坐标表示各时间记录点。故对于该数据矩阵中的每个元素$D_{i,j}$，它记录着第i个检测点在第j个时间记录点所检测到的交通状况。该数据已经过初步量化处理，以0-4共五个自然数来记录交通状况，0表示交通闲置，4表示交通拥塞。图1和图2展示的是对600个样本分别采用SNN算法和新算法所得到的最大聚类的数据图。通过分别计算它们的类内相似度，发现采用新算法所得的聚类具有更高的类内相似度值，也就是说，新算法获得了更为精密的聚类。

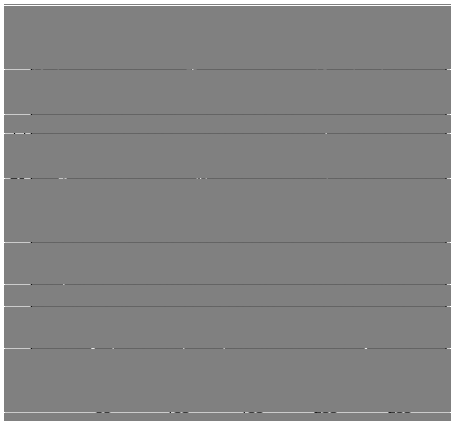

图1 SNN算法的聚类结果

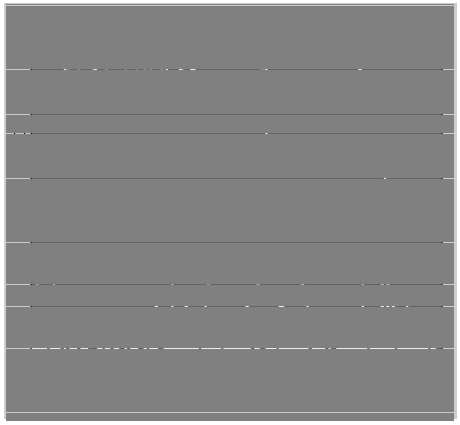

图2 RSNN算法的聚类结果

4 结 论

本文首先分析了SNN算法，然后针对其不足之处，提出了改进的共享型最近邻居聚类算法RSNN。RSNN算法有效缩减了时空复杂性，并且对孤立点、代表点、链接过滤等问题都提出了相应的改进。通过对美国高速公路数据集的实验，表明RSNN能在维持时空复杂度的前提下，有效地降低计算代价；同时经过更准确地判断代表点和孤立点，得到了比SNN算法更为优良的聚类性能。RSNN算法仍有需要改进之处，如：如何扩大该算法的应用领域、如何进一步提高聚类的准确性，这些将是我们下一步的研究工作。

参 考 文 献

[1] Guha S, Rastogi R, Shim K. Cure: An efficient clustering algorithm for large databases[C] //1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD’98), seattle WA. USA: 1998: 73-84.

[2] Ertoz L, Michael, S, Vipin Kumar. A new shared nearest neighbor clustering algorithm and its applications[C] // Workshop on Clustering High Dimensional Data and its Applications, Second SIAM International Conference on Data Mining, Arlington, VA, USA: 2002.

[3] Ertoz L, Michael S, Vipin Kumar. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data[C]. // Proceedings of Third SIAM International Conference on Data Mining, San Francisco, CA, USA: 2003.

[4] Stephen D B, Mark S. Mining Distance-Based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule[C] // Conference on Knowledge Discovery in Data archive Proceedings of the ninth ACM SIGKDD International Conference (KDD), 29-38, Washington, USA: 2003:29-38.

编 辑 徐安玉