基于最近邻相似度的孤立点检测及半监督聚类算法①

郑灵芝，黄德才

(浙江工业大学 计算机应用技术，杭州 310023)

摘 要：传统的聚类算法是一种无监督的学习过程，聚类的精度受到相似性度量方式以及数据集中孤立点的影响，并且算法也没有很好的利用先验知识，无法体现用户的需求。因此提出了基于共享最近邻的孤立点检测及半监督聚类算法。该算法采用共享最近邻为相似度，根据数据点的最近邻居数目来判断是否为孤立点，并在删除孤立点的数据集上进行半监督聚类。在半监督聚类过程中加入了经过扩展的先验知识，同时根据图形分割原理对数据集进行聚类。文中使用真实的数据集进行仿真，其仿真结果表明，本文所提出的算法能有效的检测出孤立点，并具有很好的聚类效果。

关键词：孤立点；共享最近邻；半监督聚类；先验知识

Outlier Detection and Semi-Supervised Clustering Algorithm Based on Shared Nearest Neighbors

ZHENG Ling-Zhi, HUANG De-Cai

(Department of Computer Application Technology, Zhejiang University of Technology, Hangzhou 31100, China)

Abstract: Traditional clustering analysis is unsupervised. Its precision is affected by similarity measures and outlier in the dataset and the algorithm don't take advantage of prior knowledge which can reflect the demands of users, therefore this article proposes the outlier detection and semi-supervised clustering algorithm which based on shared nearest neighbors. The algorithm according to the number of the nearest neighbors of the data in the dataset to detect the outliers in data dataset, then deal with the dataset which be operated by detecting the outliers by using semi-clustering. And during the clustering process, it adds some prior knowledge which was expanded and cluster the dataset based on the principle of graph segmentation. And the article uses some UCI datasets to make simulation experiments. The results show that the algorithm can detect the outliers effectively, and have good performance of the clustering effect.

Key words: outliers; shared nearest neighbors; semi-supervised clustering; prior knowledge

聚类方法是研究数据间逻辑上或物理上的相互关系的技术，其分析结果不仅可以揭示数据间的内在联系与区别，还可以为进一步的数据分析与知识发现提供重要依据，它是数据挖掘技术中的重要组成部分[1]。而传统的无监督聚类方法存在很多的不足，它忽略了先验知识的指导作用，使得聚类结果并不能满足用户的真正需求，具有一定的盲目性。其次，由于数据结构的不同以及数据点分布的复杂性，只有得到聚类结果后才能对聚类结果进行评价，可伸缩性小。

由于人们需求个性化的趋势，在聚类过程中加入相关领域的知识或者用户的主观因素，显得越来越重要。半监督聚类就是这样的一种方法，它可使得聚类过程有章可循，能有效地避免错误的聚类倾向，使其向好的方向收敛，同时充分反映了用户的需求，使得聚类结果更加合理。

而在数据挖掘领域中，孤立点的检测同样可以提高聚类的精度，并且具有重要的实际意义。因为孤立点点包含了很多不易被挖掘的潜在信息，例如网站检

① 收稿时间:2011-06-15;收到修改稿时间:2011-08-09

测数据中的一个孤立点可能代表了一个黑客的入侵；运动员成绩记录中的孤立点可能是一种作弊行为等。它可以应用于电子商务、贷款信誉、天气预报等领域。

1 相关研究

高纬空间中的数据是稀疏的，并且存在密度变化不一的簇，那么数据点之间的距离函数或相似性度量就会变得趋于一致，因此直接使用距离函数来定义相似性度量很难判断两个数据点是否相似[2]。一些学者提出了一种相似性的间接度量方法，即共享最近邻相似度（Shared Nearest Neighbors，SNN）[3]，它基于如下原理：如果两个数据点与相同的数据点中的大部分都相似，则即使无法使用直接的相似性方法来度量，它们也是相似的。

另外，数据集中的孤立点数据严重影响聚类算法的精度，并且它可能代表了某些应用领域中的新知识，或者反映了少数人群真实的兴趣意向，由此可见，对孤立点的检测研究具有重要的意义[4]。目前，许多研究者根据假设孤立点存在的不同情况，提出了许多孤立点检测算法，主要分为基于统计的方法、基于偏离的方法、基于距离的方法和基于密度的方法等[5]。基于统计的方法主要针对数据的单一属性进行检测，检测结果受到数据分布以及参数选取的影响；基于偏离的方法需要事先获取数据集的主要特征，在描述相异性时较难抉择，因此该方法很少应用于实际中；基于距离的方法则很难选取到最好的最小距离，并且检测出来的孤立点无法区别是全局孤立点还是局部孤立点；基于密度的孤立点检测方法可以很好的度量孤立点的孤立程度，但是局部异常因子 LOF 的计算量非常大，并且其参数的选取也是一个难题。

其次由于在实际应用中人们可以获取少量的带类标号的数据，因此本文考虑在传统聚类算法过程中加入这些先验信息，提高聚类的精度。本文将这两部分信息作为一种先验知识加入到聚类过程中，先验知识的形式采用 Wagstaff 等人在文献[6]中提出了两种类型的成对约束限制，即 must-link 约束和 cannot-link 约束。并且根据约束集自身的传递性以及数据集的特点对约束集进行扩展，最大程度上利用了先验知识。

基于以上的研究，本文提出了一种基于最近邻相似度的孤立点检测及半监督聚类算法（Outlier Detection and Semi-supervised Clustering Algorithm Based on SharedNearest Neighbors，ODSCA-SNN）。它采用共享最近邻为相似度，通过检测排除孤立点以及在聚类过程中加入先验知识来提高聚类的精度，更好的体现用户的需求。

2 算法描述

2.1 相关定义

首先给出两个相关的定义，假设本文训练数据集为 $X=\{x_i\}_{i=1}^{n}$。

定义 1 数据点的最近邻：数据点 x_i 和 x_j 若满足 dist（x_i,x_j）< ζ ,1≤i,j≤n，则有 $x_i \in P_j$。

定义 2 数据点的共享最近邻：若 $x_k \in Pi$ 且 $x_k \in P^j$，则 $xk \in P_{ij}$。

其中 P_i 表示数据点 x^i 的最近邻集，P_j 表示数据点 x_j 的最近邻集，P^{ij} 表示数据点 x_i 和 x_j 的共享最近邻集，Pij=Pi∩Pj。ζ 为一给定距离阈值，dist（xi,xj）表示数据点 xi 和 xj 的距离，这里指的距离不仅限于通常的欧氏距离，是一种广义上的距离定义。

本文采用 K-means 算法来确定数据点的最近邻。先选定一初始 K 值，然后根据需要对给定样本空间使用 r 次 K-means 算法。假设 C 表示类别，σ 为阈值，|P|表示最近邻集的个数，Voteij 则是记录了 xi 和 xj 同属于一个类别的次数，则最近邻的判断方法可以描述为：

如果 xi∈C，xj∈C，则 Voteij=Voteij+1。

如果（Voteij/r）>σ，则|Pi|=|Pi|+1，|Pj|=|Pj|+1。

而半监督聚类则是通过共享最近邻来构造 SNN 相似度图，并在聚类过程中加入了扩展的成对约束 M 和 C（must-link 和 cannot-link），最后使用分割图形的方法来得到聚类结果。

2.2 算法步骤

本文提出的基于最近邻相似度的孤立点检测及半监督聚类算法主要分为两大部分，算法的前四步为孤立点检测（Outlier Detection based on Shared Nearest Neighbors，OD-SNN），后五步为基于最近邻的半监督聚类算法（Semi-supervised clustering algorithm based on Shared Nearest Neighbors，SCA-SNN）。算法的输入为：数据集、K 值、初始成对约束集 M 和 C，算法的输出为聚类结果，算法步骤描述如下：

Step one：输入 K 值。假设数据集的真实类别数为 C，则从 K=C 开始逐步增加。

Step two：使用 K-means 算法训练数据集，每个 K 值可以训练 r 次，并根据训练结果计算每个数据点的最近邻集，判断方法如上所述。

Step three：确定孤立点。方法为：在每一 K 值训练完得到最近邻集之后，如果某个数据点拥有极少量的邻居数据点甚至最近邻集为空，这样的数据点就可以直接判定为孤立点数据，本文称为直接孤立点。

Step four：当确定了一个孤立点后，就将此孤立点作为依据，然后分析其他数据点是否为孤立点。即如果存在一数据点是该孤立点的最近邻，则该数据点也为孤立点，本文称为衍生孤立点。如果第一个孤立点没有确定或者孤立点未被全部检测出来，则转到第一步，增加 K 的值，继续训练，直至所有的孤立点都被检测到。

Step five：确定每个数据点的共享最近邻。

① 在第 Step two 得到的 K-means 训练结果中，计算数据点 xi 和 xj 被分到同一个类中的次数，记为 Voteij。

② 然后计算数据点 xi 和 xj 被分到同一个类中的概率 co-assoc(xi ,xj)，计算公式如下：

$$co-assoc(x_i,x_j)=\frac{votes_{ij}}{r} \qquad (1)$$

③ 如果 co-assoc(xi ,xj)>0.5，则将数据点 xj 归属到数据点 xi 的最近邻集 Pi 中，即 $x_j \in P_i$。

④ 重复步骤②-③，直到每个数据点的共享最近邻集都确定。

Step six：根据得到的共享最近邻，构造 SNN 相似度图，具体计算步骤如下：

① 根据每个数据点 x_i 的最近邻集 P_i 来计算 SNN 相似度图中每条边的权值。如果数据点 xj 在数据点 x_i 的最近邻集 P_i 中，则连接数据点 x_i 和 x_j，并将连接数据点 x_i 和 x_j 的边的权值记为 σ_{ij}，计算公式如下：

$$\sigma_{ij}=2\times\frac{|P_{ij}|}{|P_i|+|P_j|} \qquad (2)$$

② 重复第①步直到每个数据点的最近邻集中的元素都被遍历一次。由于许多数据点之间的 SNN 相似度为 0，因此会形成一个稀疏的 WSnnG。

Step seven： 扩展约束集：假设初始约束集为 M（Must-link 集）和 C（Cannot-link 集），约束集的扩展方法如下：

① if（xi,xj）∈M and（xj,xk）∈M，then 扩展 M 为 M∪{（xi,xk）}。if（xi,xj）∈M and（xj,xk）∈C，then 扩展 C 为 C∪{（xi,xk）}。

② if（xi,xj）∈M and（xl,xk）∈M and（xi,xl）∈C，then 扩展 C 为 C∪（xi , xk）∪（xj , xk）。

③ if（xi , xj）∈M，对于 Pi 中的任一点 xd ，if $\sigma_{d,j}\geq\sigma_{i,j}$ 且 Md∩Cj=ϕ，Cd∩Mj=ϕ，then 扩展 M 为 M∪{（xd , xj）}。

用此方法扩展 Pi 中所有满足条件的点，重复上面的步骤，直至 M 和 C 规模不再增加。

Step eight ：在上述运算之后，得到了数据点的最近邻图以及扩展过的约束集，为了使聚类结果尽量满足用户需求，并且简化最近邻图，方便计算，我们使用扩展约束集对 WSnnG 进行修改，修改方法如下：

If（xi , xj）∈must-link，then $\sigma_{ij}=1.0$

If（xi , xj）∈cannot-link，then $\sigma_{ij}=0$

Step nine：以上四步的运算得到一个更加稀疏的 WSnnG，这时的聚类就已经转化为对 WSnnG 图的分割了，使用 METIS[7-9]分割技术对其进行分割，得到的每个子图对应的就是一个类，子图中的点就是类中的元素。

3 实验分析

本文的实验数据集都来自于 UCI 真实数据集(http://archive.ics.uci.edu/ml/)，实验结果为多次实验所得数据的平均值。孤立点检测的性能判断主要通过分析检测出来的正确孤立点在所有孤立点中所占的比例，而对半监督聚类算法的评价才采用文献[10]提出的评价函数：

$$RI=\frac{\text{正确的判别数目}-\text{已知的成对约束}}{\text{总的判别数目}-\text{已知的成对约束}} \qquad (3)$$

正确的判别数目是指算法对样本空间中的所有数据点两两之间是否属于同一个类的判别结果，假设样本空间中数据点的个数为 n，则总的判别数目等于 n(n-1)/2。已知的成对约束数就是系统随机产生的初始约束集，在评价指标中减去已知约束，因为半监督聚类算法中，已知的监督信息是不能反映聚类算法的效果的。实验使用 Lypmphography 数据集、以及 Glass 数据集(http://archive.ics.uci.edu/ml/)来做对比实验。数据集的对象分布情况如表 1、2 所示。

表 1　Lypmphography 数据集的数据对象分布

类别	类别	所占百分比
普通类	类 2、3	95.9%
孤立点类	类 1、4	4.1%

表 2　Glass 数据集的数据对象分布

类别	类别	所占百分比
普通类	类 1、2、3、7	89.8%
孤立点类	类 5、6	10.2%

孤立点的检测的实验结果如下表所示，表中第一列为 K 值，第二列表示通过分析数据点的最近邻而得到的孤立点数目，即拥有极少最近邻的数据点直接被判断为孤立点，第三列表示根据孤立点的最近邻集得到的孤立点个数，称为衍生孤立点，第四列是指在第三列得到的孤立点中，是真正孤立点的个数，最后一列为孤立点检测的正确率。

表 3　Lypmphography 数据集上的孤立点检测结果

K 值	直接孤立点	衍生孤立点	正确孤立点数	正确率
8	3	12	4	66.7%（4/6）
12	5	10	4	66.7%（4/6）
16	8	15	6	66.7%（4/6）

Lypmphography 数据集的实验结果如表 3 所示。由于数据集的真实类别数为 4，因此实验从 K=4 开始训练。当 K=8 时，我们得到一个数据点的最近邻集包含了极少的对象，因此将其确定为孤立点，并分析该孤立点的最近邻集中的数据点。所以在 K 值为 8 的时候，我们得到了 12 个孤立点，其中包含了 4 个正确的孤立点。当 K=12 时，虽然孤立点检测的正确率从表面上没有提高，但是从分析类间特性而得到的孤立点数目变少了，去除掉部分判断有误的数据点，从这个角度来说，检测率提高了 7%。当 K=16 时，6 个孤立点全部被检测出来，检测率达到了 100%。算法在 Glass 数据集上也有明显的效果，如表 4 所示。

表 4　Glass 数据集上的孤立点检测结果

K 值	直接孤立点	衍生孤立点	正确孤立点数	正确率
8	10	24	16	72%（16/22）
10	12	28	18	82%（18/22）
16	16	33	22	100%（22/22）

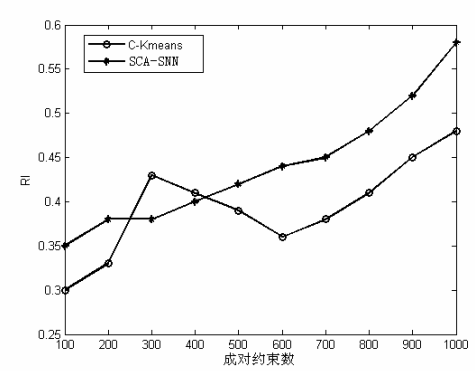

图 1　Lypmphography 数据集上的实验结果

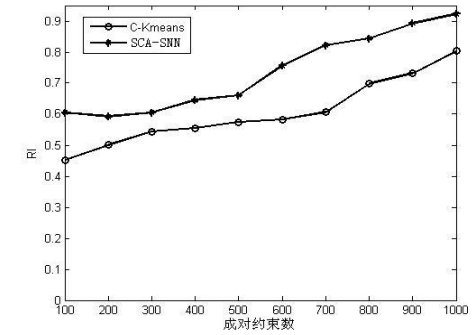

图 2　“去噪”的 Lypmphography 数据集实验结果

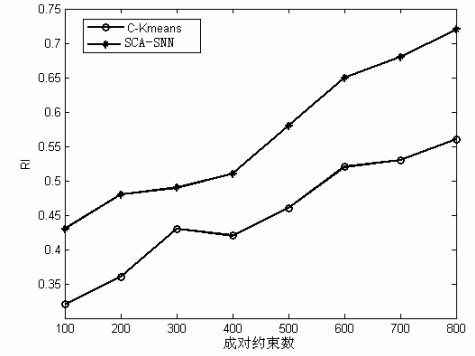

图 3　Glass 数据集上的实验结果

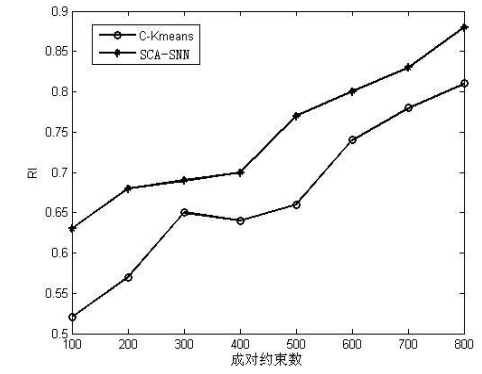

图 4　“去噪”的 Glass 数据集实验结果

在孤立点检测的结果上，删除孤立点，然后对“去噪”数据集进行半监督聚类。与新算法对比的算法是C-Kmeans 算法。实验结果如图 2 和图 4 所示。为了充分验证孤立点检测的重要性以及半监督聚类的性能，本文在原始数据集上作了对比实验，实验结果如图 1 和图 3 所示。

图 1 是 C-Kmeans 算法和 SCA-SNN 算法在没有删除孤立点的原始数据集 Lypmphography 上进行的实验结果，对比的是没有孤立点的 Lypmphography 数据集，实验结果如图 2 所示。两种算法在原始 Lypmphography 数据集上都没有很好的效果。虽然随着成对约束数的增加，实验结果有所提高，但是当约束数达到 1000 时，C-Kmeans 算法的正确判断率只有 0.48，而 SCA-SNN 算法也只达到了 0.58，这表明数据集中的孤立点数据对聚类结果造成了很大的影响，而且减弱了成对约束的指导作用，从而导致整个聚类算法都没有很好的结果。从图 3 可以明显的发现，由于存在“噪声”数据，C-Kmeans 算法表现了它的不稳定性。从聚类结果的整体来看，SCA-SNN 算法的聚类性能始终优于 C-Kmeans 算法，进一步证明了半监督聚类的有效性。图 2 的实验数据集是“去噪”Lypmphography 数据集，只包含原始数据集中的第 2 类和第 3 类。在此数据集上进行实验对比，从图 2 中可以发现随着成对约束数的增加，SCA-SNN 算法效果呈稳步逐渐上升状态，而删除了孤立点之后，C-Kmeans 算法也表现的相对较稳定，没有出现聚类结果的大幅度波动。但从整体聚类结果来看，SCA-SNN 算法的性能明显比 C-Kmeans 算法要好。

图 3 和图 4 则是在 Glass 数据集的实验结果，不管数据集中是否存在孤立点，SCA-SNN 算法的聚类效果都比 C-Kmeans 算法的效果好，尤其在删除了孤立点之后的数据集上，SCA-SNN 算法有着更好的实验效果。

从上面四个实验结果图来看，ODSCA-SNN 算法在没有孤立点的数据集上的实验效果为最优，这不仅表明孤立点的检测是至关重要的一个过程，而且充分地验证了 ODSCA-SNN 算法的聚类性能。在很多的实际应用中，数据集往往都包含一些孤立点，这些孤立点也许含有潜在的有价值的信息，因此挖掘出孤立点既可以有效的提高聚类的性能，得到正确的分类情况，而且还可以帮助人们获取更有价值的信息。

4 总结

本文提出了基于最近邻相似度的孤立点检测及半监督聚类算法。本算法使用 K-means 算法来训练数据集，可以得到合理而又准确的数据共享最近邻集，并根据所得结果快速准确的检测出全局孤立点，对局部孤立点也有很有的效果。算法有效的避免了噪声点的预处理不足以及输入不准确参数对结果的影响，也克服了如 Jarvis-Patrick 算法的计算量大的问题。在半监督聚类的过程中，对已经获取的成对先验知识进行扩展，使得先验知识的指导作用达到最大。算法不仅检测出了孤立点，并且有效地避免了对参数的依赖性，很好了排除了孤立点对聚类的影响。算法结合先验知识并扩展，使得聚类过程“有章可循”。通过对真实数据集的实验，表明孤立点检测算法结合半监督聚类得到的聚类结果为最好。

参考文献

1 Barnett V, Lewis T. Outliers in statistical data. New York,1994.

2 Knorr EM, Ng RT. Algorithms for mining distance-based outliers in large dataset. Proc.of the 24th VLDB Con New York,USA 1998, 392–403.

3 李订芳,胡文超,何炎祥.基于共享最近邻聚类和模糊集理论的分类器.控制与决策,2006,21(10):1103–1108.

4 李强,李振东.数据挖掘中孤立点的分析研究在实践中应用.微计算机应用,2006,27(3):323–327.

5 Levent E, Michael S, Vipin K. A new shared nearest neighbor clustering algorithm and its applications,Workshop on Clustering High Dimensional Data and its Applications Second SIAM International Conference on Data Mining,Arlington, 2002:105–115.

6 Wagstaff K, Cardie C. Clustering with instance-level constraints. The 17th International Conference on Machine Learning,2000,1103–1110.

7 George K. and Vipin K. A fast and high quality multilevel scheme for partitioning irregular graphs, SIAM Journal on Scientific Computing,1998,59–392.

8 George K, Vipin K. Multilevel k–way partitioning scheme for irregular graphs, Parallel & Distributed Computing, 1998,l: 48,96–129.

9 George K, Vipin K. Multilevel algorithms for multi-constraint graph partitioning,Proceedings of the 1998 ACM/IEEE conference on Supercomputing,San Jose,1998,1–13.

10 Xing EP, Ng AY, Jordan MI. Distance metric learning, with application to clustering with side-information, Advances in Neural Information Processing Systems 15 (NIPS), Cambridge, MA, 2003, 505–512.