基于不完备标签数据的半监督聚类算法①

袁利永

(浙江师范大学 数理与信息工程学院，金华 321004)

摘 要：针对 seeded-K-means 和 constrained-K-means 算法要求标签数据类别完备的限制，本文提出了基于不完备标签数据的半监督 K-means 聚类算法，重点讨论了未标签类别初始聚类中心的选取问题。首先给出了未标签类别聚类中心最优候选集的定义，然后提出了一种新的未标签类别初始聚类中心选取方法，即采用 K-means 算法从最优候选集中选取初始聚类中心,最后给出了基于新方法的半监督聚类算法的完整描述,并通过实验测试对新算法的有效性进行了验证。实验结果表明本文所提算法在执行速度和聚类效果上都优于现有算法。

关键词：半监督聚类；K-means；不完备先验知识；初始聚类中心；标签数据

Semi-Supervised Clustering Algorithm Based on Incomplete Labeled Data

YUAN Li-Yong

(College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua 321004, China)

Abstract: For the seeded-K-means and constrained-K-means algorithm limitations that complete category information in labeled data is required, this paper put forword an semi-supervised K-means clustering algorithm based on incomplete labeled data, focused on selection of the initial cluster center of unlabeled category. We gave a definition of the Best Candidate Set of cluster center of unlabeled category, proposed a new method that selecting initial cluster center of unlabeled category from the Best Candidate Set using K-means. Finally, a complete description of semi-supervised clustering algorithm based on the new method is given, the validity of the new algorithm is verified by experiment. Experimental results show that the proposed algorithm is superior to existing algorithms not only in clustering effect and in execution speed.

Key word: semi-supervised clustering; K-means; incomplete prior knowledge; initial cluster center; labeled data

随着数据采集技术和存储技术的发展，获取大量的无标签数据比较容易，而获取标签数据通常需要付出较大的代价。传统的监督学习只能利用少量的标签数据进行学习，而无监督学习只利用无标签数据进行学习。半监督学习是近年来机器学习领域的一个研究热点[1-6]，它的优越性体现在能同时利用标签数据和无标签数据进行学习。目前，半监督学习的研究主要包括半监督分类、半监督聚类及半监督回归，半监督聚类是当前研究的热点。

聚类分析试图将一组未标签数据按照一定的相似度准则分到几个类中去，使得在同一类别中数据的相似度较大，不同类别间数据的相似度较小[7]。聚类分析作为一种非监督学习方法，是机器学习领域中的一个重要的研究方向，同时也是数据挖掘中进行数据处理的重要分析工具和方法[8]。现有的半监督聚类算法很多是在传统聚类算法基础上引入监督信息发展而来，代表算法是基于经典 K-means 算法[9]的各种半监督 K-means 算法。如 K.Wagstaff 等人提出将 must-link 和 cannot-link 成对约束引用到半监督聚类方法中[2]，Basu 等人在 K-means 算法的基础上，提出了利用少量已标签数据的 seeded-K-means 和 constrained-K-means 算法[3]。

seeded-K-means 和 constrained-K-means 算法假设所有类别都至少有 1 个已标签数据，并以此为基础产生初始聚类种子，同时利用标签数据的约束指导聚类过程。针对上述两种半监督聚类算法要求标签数据类

① 收稿时间:2010-05-23;收到修改稿时间:2010-07-04

别完备的限制，本文提出了一种基于不完备标签数据的半监督聚类算法，重点讨论了聚类中心初始化过程中未标签类别初始聚类中心的选取问题。

1 研究背景

1.1 seeded-K-means 和 constrained-K-means 算法

Basu 等人在 K-means 算法的基础上，引入由少量标签数据形成的 seed 集，并假设 seed 集中包含所有 K 个聚类，且每个类最少包含一个数据。将 seed 集划分为 K 个聚类，并在此基础上进行算法的初始化，形成两种半监督 K-means 算法[3]：seeded-K-means 和 constrained-K-means。

seeded-K-means 算法：

输入：数据集 $X=\{x_i\}_{i=1}^{N}, x_i \in R^d$，簇的数量为 K，已标签数据集 $S=\cup_{h=1}^{K} S_h$。

输出：数据集 X 的 K 个划分 $\{X_h\}_{h=1}^{K}$，使得目标函数最小化。

算法过程：

第一步：用已标签数据求得 K 个初始聚类中心 $\{\mu_h^{(0)}\}_{h=1}^{K}$，其中 $\mu_h^{(0)}=\frac{1}{|S_h|}\sum_{x\in S_h} x, h=1,\ldots K, t\leftarrow 0$；

第二步：重复如下过程直至收敛；

a) 分配聚类：重新分配数据点 x 到类 h*，使得 h*满足下列条件：$h^*=\arg_h \min\{d(x,\mu_h^{(t)})^2\}$；

b)重新计算每个簇的中心点：$\mu_h^{(t+1)}=\frac{1}{|X_h^{(t+1)}|}\sum_{x\in X_h^{(t+1)}} x$；

c)更新迭代次数：$t\leftarrow t+1$；

第三步：输出数据集 X 的 K 簇划分。

constrained-K-means 算法与 seeded-K-means 类似，只是对数据点 x 分配簇时，采用了如下的策略：若 x 属于 S_h，则分配 x 到 h 簇；否则分配 x 到 h*，使得 h* 满足下列条件：$h^*=\arg_h \min\{d(x,\mu_h^{(t)})^2\}$

在 constrained-K-means 的优化过程中，seed 集中数据标签不发生改变；而在 seeded-K-means 算法中，seed 集中数据标签可以发生改变。实验表明[3]，上述两种方法的聚类性能比起经典的 K-means 有明显改善。

1.2 基于不完备标签数据的聚类问题

seeded-K-means 和 constrained-K-means 算法假设所有类别都至少有 1 个已标签数据，并以此为基础产生初始聚类种子，同时利用标签数据的约束指导聚类过程。上述两种半监督的 K-means 聚类算法只适用于标签数据类别完备的情况，即数据需要满足下列条件：每类都至少存在 1 个已标签数据，显然这种假设带有很强的限制性。在实际的应用中(比如 web 文档聚类)上述条件很难被满足，更为实际的情况往往是用户无法提供所有类别的已标签数据，即标签数据类别不完备的聚类问题。

不完备标签数据的聚类问题可描述为：假设数据集包含 N 个数据点，表示为 $X=\{x_1,x_2,x_3,\cdots,x_N\}, x_i\in R_d$，假设其中包含 K 个类别，即聚类算法最终应生成 K 簇。又假设已标签数据集表示为,其中 L(L<K)为 S 中包含的类别数，即在这 K 个类别中，有 L 个类别有标签数据，而另外的 K-L 个类别没有标签数据。要求将数据集 X 划分为 K 簇，使得准则函数收敛，μ_i 是簇 C_i 的均值。

2 基于不完备标签数据的半监督聚类算法

2.1 随机选取候选点法和最优候选搜索法

基于不完备标签数据的半监聚 K-means 聚类问题的关键是初始聚类中心的选取。L 个类别有标签数据,可以用标签数据求均值得到簇的初始聚类中心:。但是还有 K-L 个簇没有标签数据，因此必须解决的问题就是怎样为这K-L个簇选择初始聚类中心,使得初始化步骤能够完成。

文献[10]提到的随机选取候选点法是从无标签数据中随机选取 K-L 个数据作为 K-L 个无标签数据簇的初始聚类中心。这个算法虽然可以工作，但是对于那些缺少先验信息的类别，随机选取的初始聚类中心容易导致算法聚类效果不理想。

文献[10]提出使用最优候选集搜索法来选择 K-L 个无标签数据簇的初始聚类中心，其初始化过程描述如下：

a) 用标签数据求得 L 个中心点集 $\{\mu_h^{(0)}\}_{h=1}^{L}$，其中 $\mu_h^{(0)}=\frac{1}{|S_h|}\sum_{x\in S_h} x, h=1,\ldots L, t\leftarrow 0$；

b) 分配所有未标签数据到距离最近的类别中，同时选取 N/K 个到最近类别距离最大的数据，构成候选集合 X_C；

c) 对于无标签数据的类别 l_{L+1}, l_{L+2}, l_K,在 X_C 中依次随机选取 m(m≤N/K)个数据，求其均值作为无标签数据簇的初始化中心：$\mu_h^{(0)}=\frac{1}{|S_r|}\sum_{x\in S_r} x, h=L+1, L+2,\ldots K, S_r\subseteq X_C$；

d) 合并得到完整的初始化中心集合 $\{\mu_h^{(0)}\}_{h=1}^{K}$。

基于最优候选搜索法的聚类中心初始化过程利用了已标签数据所提供的相关信息，能够搜索得到比较好的未标签类别的初始聚类中心。但上述算法存在两个问题：第一、最优候选集合 XC 的大小采用 N/K 缺乏理论依据或有违直观经验，如图 1 所示的情况，按上述方法选取的XC绝大多数元素将来自被画圈的簇；第二、K-L 个无标签数据簇的初始化中心的选取方法还存在缺陷，从概率角度考虑，按上述方法得到的 K-L 个初始聚类中心会十分集中(或接近)，这将影响 K-means 聚类准确性，同时也会导致算法收敛速度的下降。

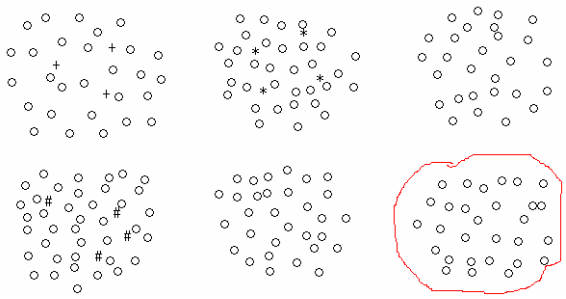

图 1　不完备标签数据的聚类问题

2.2 基于最优候选集的 K-means 选取法

针对文献[10]提出的最优候选搜索法的不足，提出用基于最优候选集的 K-means 选取法来选择 K-L 个无标签数据簇的初始聚类中心，其基本思想如下：首先利用已标签数据采用 K-means 求出 L 个已标签类别的初始聚类中心；然后按一定比率选出若干个离已标签数据簇中心距离最大的数据，构成最优候选集，显然这些数据点成为无标签数据簇中心的可能性更大；然后从最优候选集中选取部分或全部进行多次参数为 K-L 的 K-means 聚类，取准则函数值最小的聚类中心点作为未标签类别的初始聚类中心；最后合并两类初始聚类中心，再执行 constrained-K-means 算法。

定义 1(点的聚类距离). 设数据集 $X=\{x_i\}_{i=1}^{N}, x_i\in R^d$，已标签数据集 $S=\cup_{h=1}^{K}S_h$，$\{\mu_h^{(0)}\}_{h=1}^{L}$ 为已标签类别的初始聚类中心集，$x\in X-S$，则数据点 x 的聚类距离定义为：$dc(x)=\min\{x-\mu_h\}, h=1,2...L$

定义 2(未标签类别聚类中心最优候选集). 设 x_m 为 X-S 中聚类距离第 m 大的数据点，则未标签类别聚类中心最优候选集定义为：

$X_C(m)=\{x\mid dc(x)\geq dc(x_m), x\in X-S\}$。

改进的基于不完备标签数据的半监督聚类算法描述如下：

输入：数据集 $X=\{x_i\}_{i=1}^{N}, x_i\in R^d$，簇的数量为 K，已标签数据集 $S=\cup_{h=1}^{L}S_h$,L 为已标签数据类别数。

输出：数据集 X 的 K 个划分 $\{X_h\}_{h=1}^{K}$，使得目标函数最小化。

第一步：初始化簇聚类中心。

a) 用标签数据求得 L 个中心点集 $\{\mu_h^*\}_{h=1}^{L}$,其中$\mu_h^*=\frac{1}{|S_h|}\sum_{x\in S_h}x, h=1,\ldots L$；

b) 计算 dc(x)，$x\in X-S$,并构建参数 m=N*(K-L)/K 的未标签类别聚类中心聚选集 X_C；

c) 对于无标签数据的类别 l_{L+1},l_{L+2},l_K，选取 XC 部分或全部执行多次参数为 K-L 的 K-means 算法，取准则函数 $E=\sum_{i=1}^{K}\sum_{p\in C_i}|p-\mu_i|^2$ 最小的聚类中心作为未标签类别的初始聚类中心，记为 $\{\mu_h^+\}_{h=1}^{K-L}$；

d) 合并 $\{\mu_h^*\}_{h=1}^{L}$和$\{\mu_h^+\}_{h=1}^{K-L}$ 得到完整的初始化中心集合 $\{\mu_h^{(0)}\}_{h=1}^{K}, t\leftarrow 0$。

第二步：重复如下过程直至收敛；

a) 分配聚类：重新分配数据点到类 h*，使得 h* 满足下列条件：$h^*=\arg_h\min\{\|x,\mu_h^{(t)}\|^2\}$；

b) 重新计算每个簇的中心点：

$\mu_h^{(t+1)}=\frac{1}{|X_h^{(t+1)}|}\sum_{x\in X_h^{(t+1)}}x$；

c) 更新迭代次数：$t\leftarrow t+1$；

第三步：输出数据集 X 的 K 簇划分。

选取 N*(K-L)/K 个未标签数据作为聚类中心最优候选集是基于这样的思想：考虑 K 簇划分样本数相等的情况，属于 K-L 个未标签类别的样本数量应为 N*(K-L)/K。对 X_C 部分或全部执行多次参数为 K-L 的 K-means 算法，能够获取更优的 K-L 个未标签类别的初始聚类中心，从而提高聚类的准确性和算法的收敛速度。

3　实验结果与分析

为了验证本文所提算法的有效性，随机构造了两种有些类别无标签数据的实验测试案例，如图 2 和图 3 所示，其中小圆圈代表未标签数据，“+”、“*”、“#”、“x”表示已标签数据以及相应类别。

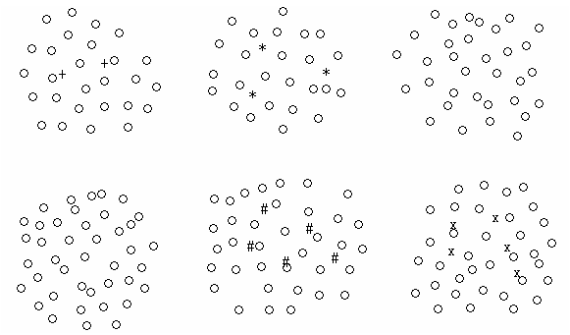

图 2 实验数据一

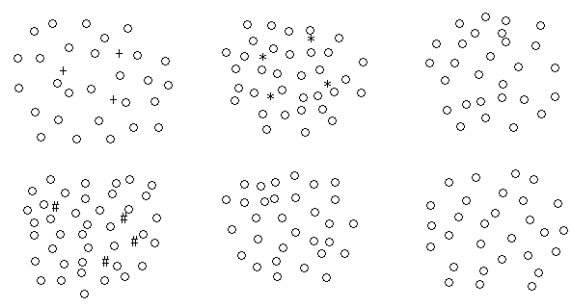

图 3 实验数据二

采用本文所提算法和文献[10]所提的另外两种算法分别进行了 100 次实验，对聚类效果的准确性和算法执行时间进行了统计，结果如表 1 所示。

实验结果显示，本文提出的算法无论在算法执行速度还是在聚类效果上都优于另外两种算法。尤其当多个未标签类别簇比较分散的情况，本文所提算法的优势更为明显。

表 1 算法性能比较

算法名称	执行速度比较		聚类准确率比较	
	实验数据一	实验数据二	实验数据一	实验数据二
随机选取法	0.21秒	0.19秒	21%	48%
最优候选搜索法	0.18秒	0.12秒	82%	96%
本文提出的算法	0.06秒	0.09秒	100%	100%

4 结束语

提出基于不完备标签数据的半监督聚类算法，消除了 seeded-K-means 和 constrained-K-means 算法要求标签数据类别完备的限制，扩展了半监督 K-means 算法的应用范围。实验结果表明，本文所提出的未标签类别的初始聚类中心选取方法与现有的其它方法相比，在聚类准确性和算法收敛速度上都有所提高。

参考文献

1 Pedrycz W, Vukovich G. Fuzzy clustering with supervision. Pattern Recognition, 2004,37(7):1339－1349.

2 Wagstaff K, Cardie C, Rogers S, et al. Constrained K-Means Clustering with Background Knowledge. In: Brodley CE, Danyluk AP, eds. Proc. of the 18th lnt'l Conf. on Machine Learning. Williamstown: Morgan Kaufmann Publishers, 2001. 577－584.

3 Basu S, Banerjee A, Mooney RJ. Semi-Supervised clustering by seeding. In: Claude S, Achim GH, eds. Proc. of the 19th Int'l Conf. on Machine Learning(ICML2002). San Fransisco: Morgan Kaufmann Publishers, 2002. 19－26.

4 李志圣,孙越恒,何丕廉,等.基于 K-Means 和半监督机制的单类中心学习算法.计算机应用,2008,28(10):2513－2517.

5 高滢,刘大有,齐红,等.一种半监督 K-均值多关系数据聚类算法.软件学报,2008,19(11):2814－2821.

6 Kulis B, Basu S, Dhillon I, et al. Semi-Supervised Graph Clustering: A Kernel Approach. Machine Learning, 2009, 1(74):1－22.

7 Jain AK, Dubes RC. Algorithms for clustering data. Englewood Cliffs, NJ: Prentice Hall, 1988.

8 Han J, Kamber M. Data mining: Concepts and techniques. SanFrancisco: Morgan Kaufmann, 2001.

9 MacQueen J. Some methods for classification and analysis of multivariate observations. Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, CA, 1967,1:281－297.

10 高云天,王学辉,郭涛.基于不完整信息的半监督聚类算法. 北华大学学报(自然科学版),2009,10(5):457－463.