一种用于处理高维稀疏数据的半监督聚类算法

崔 鹏[1,2] 张汝波[1]
(哈尔滨工程大学计算机与技术学院 哈尔滨 150001)[1]
(哈尔滨理工大学计算机与技术学院 哈尔滨 150080)[2]

摘 要 半监督聚类是近年来研究的热点,传统的方法是在无监督算法的基础上加入有限的背景知识来提高聚类性能。然而大多数半监督聚类技术都基于邻近或密度,难以处理高维数据,因此必须将约减的特征加入到半监督聚类过程中。为解决此问题,提出了一种新的半监督聚类算法框架。该算法利用样本约束传递性进行预处理,然后将特征投影到低维空间实现降维,最终用半监督算法对约减后的样本进行聚类。通过实验同现行主要降维方法进行了比较,说明此方法能有效地处理高维数据,聚类效果良好。

关键词 降维,半监督聚类,特征选择,约束
中图法分类号 TP181 **文献标识码** A

Novel Semi-supervised Clustering for High Dimensional Data

CUI Peng[1,2] ZHANG Ru-bo[1]
(School of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)[1]
(College of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China)[2]

Abstract Semi-supervised clustering is a popular clustering method in recent year, which usually incorporates limited background knowledge to improve the clustering performance. However, most of existing methods based on neighbors or density can't be used for processing high dimensionality data. So it is critical of merging the reduced feature with semi-supervised clustering process. To solve the problem, we proposed a framework for semi-supervised clustering. The framework firstly preprocesses instances with transmissibility of constraints; then reduced dimensionality by projecting feature into low dimensional space; finally it clustered instances with reduced features. To evaluate the effectiveness of the method, we implemented experiments on datasets, the results show the method has good clustering performance for handling data of high dimension.

Keywords Dimensionality reduction, Semi-supervised clustering, Feature selection, Constraints

半监督聚类近来在数据挖掘和机器学习领域受到极大的关注。大多数现有的半监督方法都不是为处理高维数据而设计的。传统的欧式密度概念在高维数据集下是没有意义的。由于大多数半监督聚类[1-4]技术都基于邻近或密度,难于处理高维数据,因此必须将约减的特征加入到半监督聚类过程中。基于此,我们提出了一种基于约束特征投影的球面 k 均值(Constrained Spherical Kmeans based on feature Projection, CSKFP)半监督聚类算法。本文第1节和第2节介绍相关工作与CSKFP半监督聚类;第3节进行实验,并讨论了实验结果;最后给出结论。

1 相关工作

1.1 数据对象间的两类约束

在聚类分析时,可能需要对聚类结果做一些限制:某两个数据对象应该分入一个集合,或某两个数据对象不应分入一个集合。为描述数据对象间的这类关系,何振峰[5]引入了Must-link和Cannot-1ink两类约束,并分析了两类约束的一些性质。

Must-link和Cannot-1ink都是布尔函数。给定数据对象集 S,已知 $P,Q\in S$,若 P 和 Q 应当在同一类中,则 Must-link(P,Q)=True;若 P 和 Q 不应在同一类中,则 Cannot-link(P,Q)=True。显然,

(1) Must-link和Cannot-link具有对称性,对 $P,Q\in S$:

Must-link$(P,Q)\Leftrightarrow$Must-link(Q,P)

Cannot-link$(P,Q)\Leftrightarrow$Cannot-link(Q,P)

(2) Must-link和Cannot-link具有有限的传递性,对 $P,Q,R\in S$:

Must-link(P,Q) && Must-link$(Q,R)\Rightarrow$Must-link(P,R)

Must-link(P,Q) && Cannot-link$(Q,R)\Rightarrow$Cannot-link(P,R)

结合两类约束的现有算法在利用约束之前,均依据以上性质来扩充限制集。这样,算法在执行时所加的约束数常会

到稿日期:2009-08-28 返修日期:2009-10-18 本文受863国家重点基金项目(2009AA04Z215)资助。
崔 鹏(1971-),男,博士生,副教授,主要研究方向为模式识别、机器学习等,E-mail:cuipeng83@163.com;张汝波(1963-),男,博士,教授,博士生导师。

多于最初提供的约束数。

1.2 球面 k 均值聚类算法

在某些高维数据如文本中，欧式距离不是一种好的相似性度量方式。其它的相似性度量，如概率文本重叠，已成功地用于文本聚类，但 cosine 相似性仍是关注的重点。

球面 k 均值[2]（SP-KMeans）是 KMeans 的一种解释，它使用 cosine 相似性作为它的基本相似性度量。在 SP-KMeans 算法中，中心矢量 $\{\mu_h\}_{i=1}^{k}$ 被限定在单位球面上。假定 $\| x_i \| = \| \mu_h \| = 1, \forall i,h$ 则有 $\| x_i - \mu_h \|^2 = 2 - 2x_i^T \mu_h$；那么聚类问题能等价于最大化目标函数，其公式为

$$J_{SP\text{-}KMeans} = \sum_{h=1}^{k} \sum_{x_i \in X_h} x_i^T \mu_h \tag{1}$$

第 h 个聚类中心 μ_h 是那个聚类中所有点的均值。SP-KMeans 算法对于高维稀疏数据矢量的计算是高效的，其在文本聚类域是非常常见的。

2 CSKFP 半监督聚类算法

CSKFP 半监督聚类算法利用约束特征投影（Constrained Feature Projection，CFP）来降低原始数据集维数，并采用约束的球面 k 均值算法（Constrained Sp-KMeans，CSK 算法）来进行低维投影数据聚类。

图 1 为 CSKFP 法的结构。给定样本集与监督集，其中 Must-link 约束 $C_{ML} = \{(x_i, x_j)\}$，(x_i, x_j) 必须分在相同的聚类，而 Cannot-link 约束 $C_{CL} = \{(x_i, x_j)\}$，(x_i, x_j) 必须分在不同的聚类。CSKFP 由 3 步组成：首先，根据 Must-link 约束的传递性，利用预处理法减少无标记样本和成对约束；然后，用 CFP 法，将原始数据投影到低维空间；最后，利用 CSK 算法产生聚类结果。

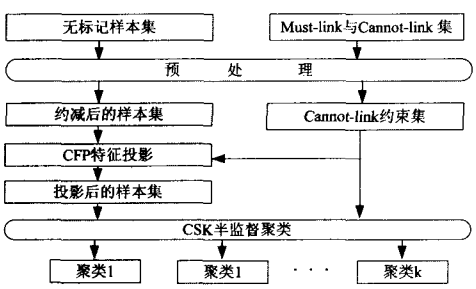

图 1 CSKFP 算法结构示意图

2.1 预处理

若约束中对没有错误，则用 Must-link 约束很容易表明一种对等关系。可用 Must-link 约束的传递闭包替换图 2 中的样本，其中实线表示 Must-link，点线代表 Cannot-link 约束。图 2 中，原始样本用白色结点表示，平均样本用黑色结点表示，集合 $\{x_1, x_2, x_3, x_4, x_5\}$，$\{y_1, y_2, y_3, y_4\}$，$\{z_1, z_2, z_3\}$ 表示由 Must-link 约束的传递闭包。预处理后，可以删去 Must-link 约束，用每个闭包的平均样本 x, y, z 的 Cannot-link 约束表示。在转换数据集中，每个闭包的大小代表样本的权。如果有一些大的约束，则需要在预处理步中对它们识别和移除。预处理步的好处在于可以简化引导聚类的约束，我们只需关注优化步中的 Cannot-link 约束。另一个好处是能够减少无标记样本的大小和 Cannot-link 约束，这有助于处理一些大数据集。

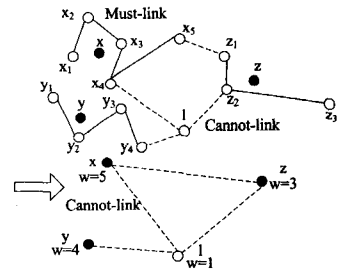

图 2 算法预处理步示意图

2.2 CFP 特征投影

在文献[1,2,4]中，采用成对约束学习一种样本原型间的自适应度量。然而，学习高维样本间的测度度量是非常费时的。更重要的是，最近关于高维空间研究表明，在高维空间中测度的概念是没有意义的。本文不是采用基于约束的度量学习，而是采用 CFP（基于约束的特征投影法）来进一步提高高维数据集的聚类[6,7]性能。目的是学习投影阵 $F_{d\times k} = \{F_1, \cdots, F_k\}$，包含 k 个正交单位长度的 d 维向量，能将原始数据集投影到一个低维空间，从而使任意 Cannot-link 约束的样本对间的测度最大，而任意 Must-link 约束的样本对间的测度最小。目标函数可最大化为

$$f = \sum_{(x_1,x_2)\in C_{CL}} \| F^T(x_1 - x_2) \|^2 - \sum_{(x_1,x_2)\in C_{ML}} \| F^T(x_1 - x_2) \|^2 \tag{2}$$

满足约束

$$F^T F_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{3}$$

式中，$\| \cdot \|$ 表示 L_2 的范数，F 为投影阵，其列向量彼此正交。如 2.1 节所述，在图 2 中初始化，可删除每个 Must-link 约束传递闭包，用它的等价平均样本来代替，然后方程(1)中的目标函数可进一步减化为

$$f = \sum_{(x_1',x_2')\in C_{CL}'} \| w_1 w_2 \cdot F^T(x_1' - x_2') \|^2 \tag{4}$$

式中，$\{w_i\}_{i=1}^{N'}$ 为预处理 Must-link 后的样本权值集（度量每个传递闭包的样本数），N' 为预处理后的样本大小（$N' \leqslant N$）。在方程(4)中，采用了欧式距离，未采用 cosine 相似性度量，这是因为单位长度样本需满足下述方程的属性。

$$\| x - \mu \|^2 = \| x \|^2 + \| \mu \|^2 - 2x^T\mu = 2 - 2x^T\mu \tag{5}$$

由方程(5)可以看出，当用于计算单位长度样本时，采用 cosine 相似性与采用欧式测度是等价的。存在一个方程(4)中最优投影阵的解。下面的定理表明，最优投影阵 F 由差值阵 $C_{d\times m}$ 的协方差阵的前 k 个特征向量给定，其中 C 的每列都为一个在 C'_{CL} 中对 $(x_1' - x_2')$ 的带权差向量 $w_1 w_2 \cdot (x_1' - x_2') \in R^d$，$m$ 为 Cannot-link 约束对数。

定理 1 给定想要的维数 $k (k < d)$，Cannot-link 约束 C'_{CL}，协方差阵 $M = cov(C)$，其中 C 在前面已定义，最优投影阵 $F_{d\times m}$ 为 M 对应于 k 个最大特征值的前 k 个特征向量。

证明：考虑到目标函数

$$\begin{aligned} f &= \sum_{(x_1',x_2')\in C_{CL}'} \| w_1 w_2 \cdot F^T(x_1' - x_2') \|^2 \\ &= \sum_{(x_1',x_2')\in C_{CL}'} w_1^2 w_2^2 \sum_l F_l^T (x_1' - x_2')(x_1' - x_2')^T F_l \\ &= \sum_l F_l^T \Big[\sum_{(x_1',x_2')\in C_{CL}'} w_1^2 w_2^2 (x_1' - x_2')(x_1' - \end{aligned}$$

$$x_2')^T]F_l = \sum_l F_l^T (CC^T) F_l = \sum_l F_l^T M F_l \quad (6)$$

式中，F_l^T 服从 $l=h$，$F_l^T F_h=1$，否则为 0。可得到

$$\frac{\partial L}{\partial F} = 2MF_l - 2\varepsilon_1 F_l = 0 \quad \forall l=1,\cdots,k$$
$$\Rightarrow MF_l = \varepsilon_1 F_l \quad \forall l=1,\cdots,k \quad (7)$$

从方程(7)可以看出，F_l 为 M 的特征向量，ε_1 为 M 相应的特征值。为最大化 f，F 必须是 M 的前 k 个特征向量，使 f 为 M 的 k 个最大特征向量和。

经过上述的基于约束特征投影，可在低维空间下表示原始样本，使其符合在成对约束下给定的类别信息。

2.3 CSK 半监督聚类

由于删除了 Must-link 约束，在初始化步骤中缩减了数据集，我们提出的基于约束的球面 k 均值半监督聚类法(CSK)与文献[2,8]中描述的不同。

给定缩减了的样本集 $X'=\{x_1',\cdots,x_{N'}'\}$，对应的权值为 $W'=\{w_1,\cdots,w_{N'}\}$，Cannot-link 约束集，预先指定的聚类数 K，为找到 K 个户不相交的划分，我们采用了一种局部贪婪启发法来更新聚类中心。

给定每个 Cannot-link$(x_i',x_j')\in C_{CL}$，找到两个不同的聚类中心 $\mu_{x_i'}$ 和 $\mu_{x_j'}$，使得 $w_i \cdot x_i'^T \mu_{x_i'} + w_j \cdot x_j'^T \mu_{x_j'}$ 最大，并将 x_i' 和 x_j' 分配到两个不同的中心以避免违反 Cannot-link 约束。

3 实验结果与分析

我们在不同应用领域的数据集上进行了实验，并根据实验结果评价了算法的有效性。

3.1 数据集的选择与评价方法

(1)数据集的选择。选择了 3 个 TREC 高维数据集 tr11，tr12，tr31，用于比较 PCA 和 RCA[9] 在高维数据集上的性能。

(2)评价方法：本文采用修正的 Rand 指数，CRI(Corrected Rand Index)作为聚类验证方法。

CRI 作为一种外部验证度量，基于给定数据集的真实标签来估计聚类质量。CRI 值越大，聚类效果越好。比较聚类结果与真实划分情况，对每一个样本对，存在 4 种可能：

①它们应归入一类，在结果中确实将其归入一类；

②它们应归入一类，在结果中却将它们分入不同类；

③它们应属于不同类，在结果中却将它们归入了一类；

④它们应属于不同类，在结果中确实将它们归入不同类。

设满足以上条件的样本对数分别为 a,b,c,d，总样本数为 n，则总样本对数为 $n\times(n-1)/2$。CRI 可表示为

$$CRI = \frac{a+d-C}{n\times(n-1)/2-C} \quad (8)$$

式中，C 表示所有限制的数目，包括根据用户提供的限制直接传播所得的所有限制。

对每个数据集，重复每个实验 20 次。在每次测试时，随机产生 100 对约束，而后在整个数据集上进行测试。实验结果为 20 次测试的平均值。

3.2 CSKFP 法的有效性

为将 CFP 与其它的降维方法(如 PCA 和 RCA)进行比较，从 Trec 中选择了高维数据集。首先利用 PCA 把原始数据投影到一个 100 维的空间，尔后应用不同的算法进一步将维数降低到 30。

图 3 为 TREC 数据集的实验结果。从图中可以看到 CSKFP 在 3 个数据集上几乎都获得了最好的性能。相比而言，虽然我们利用 PCA 先将维数降低到 100，但 RCA 的性能仍在所有的算法中最差。

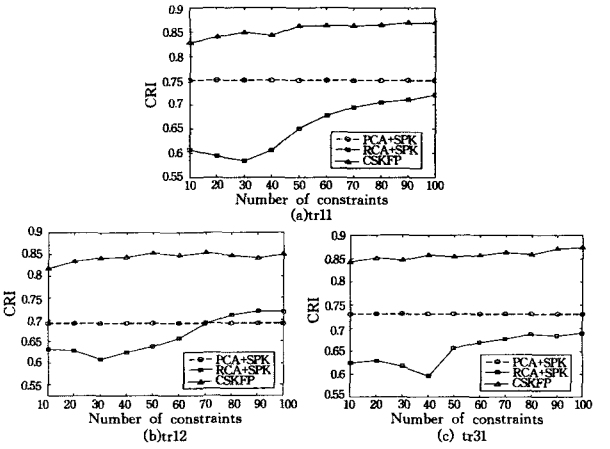

图 3 3 个 TREC 数据集上的聚类性能

结束语 本文提出了一种利用特征投影的基于球面 k 均值的半监督聚类算法，用于处理低维与稀疏高维数据。此算法利用基于约束的特征投影来降低维数，并应用约束的球面 k 均值算法聚类约减后的低维数据。

在算法的设计过程中，我们将约束特征降维作为一种优化问题，目标是找到一种基于约束样本对的特征投影阵；并给出了方便执行的解决方案，该方案能够应用到更广泛的应用领域。

参 考 文 献

[1] Wagstaff K, Cardie C. Clustering with instance-level constraints [C]// Proc. of the 17th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2000

[2] Basu S. Semi-supervised Clustering: Probabilistic models, algorithms and experiments[D]. Austin: The University of Texas, 2005

[3] 易星. 半监督学习若干问题的研究[D]. 北京：清华大学，2004

[4] Zhu X. Semi-supervised learning literature survey[R]. 1530. Computer Sciences, University of Wisconsin-Madison, 2005

[5] 何振峰，熊范纶. 结合限制的分隔模型及 K-Means 算法[J]. 软件学报，2005，16(5)：799-809

[6] Gosselin P H, Cord M. Feature-based approach to semi-supervised similarity learning [J]. Pattern Recognition, 2006, 39: 1839-1851

[7] Liu Huawen, Sun Jigui. Feature selection with dynamic mutual information[J]. Pattern Recognition, 2009, 42: 1330-1339

[8] Wang Fei, Zhang Changshui. Robust self-tuning semi-supervised learning[J]. Neurocomputing, 2007, 70: 2931-2939

[9] Roweis S. EM algorithms for PCA and SPCA[J]. Advances in Neural Information Processing Systems, 2007, 16: 626-632