基于判别分析的半监督聚类方法

陈小冬[1,2],尹学松[2],林焕祥[3]
CHEN Xiao-dong[1,2],YIN Xue-song[2],LIN Huan-xiang[3]

1.浙江大学 计算机科学与技术学院,杭州 310027
2.浙江广播电视大学 信息与工程学院,杭州 310012
3.浙江科技学院 信息学院,杭州 310012
1.College of Computer Science,Zhejiang University,Hangzhou 310027,China
2.College of Information and Engineering,Zhejiang Radio & TV University,Hangzhou 310012,China
3.College of Information & Electronic Engineering,Zhejiang University of Science & Technology,Hangzhou 310012,China

CHEN Xiao-dong,YIN Xue-song,LIN Huan-xiang.Semi-supervised clustering approach with discriminant analysis.Computer Engineering and Applications,2010,46(6):139-143.

Abstract: The semi-supervised clustering is to mine and help to understand better the structure of unlabeled data and to more closely conform to the user's preferences using those supervised data,in comparison with unsupervised clustering.Most existing semi-supervised clustering methods are designed for handling low-dimensional data.In this paper,a novel Semi-supervised Clustering Approach with Discriminant Analysis(SCADA) is presented for clustering the high-dimensional data.Specifically,the data are first mapped onto the low-dimensional space by principal component analysis such that constrained spherical K-means algorithm is used to cluster those transformed data.Secondly,linear discriminant analysis is used to reduce the number of the dimensionality of the data in terms of the clustering results.Finally,the data in the embedded space are clustered.Indeed,the experimental results on several real-world data sets show the SCADA method can effectively deal with the high-dimensional data and provides an appealing clustering performance.

Key words: semi-supervised clustering;pairwise constraint;principal component analysis;linear discriminant analysis

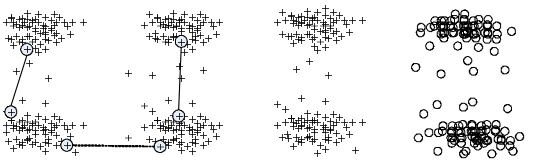
摘 要:与无监督聚类相比,半监督聚类是利用一部分先验信息来更好地挖掘和理解数据的内在结构,并紧密遵从用户的偏好。现有的典型半监督聚类算法仅仅适合于低维数据,文中提出一种新颖的基于判别分析的半监督聚类算法来解决高维数据聚类问题。新算法首先使用主成分分析来投影高维数据,进一步在投影空间中,使用基于球形 K 均值聚类算法对数据聚类;然后利用聚类结果,使用线性判别分析降维输入空间数据;最后在投影空间中对数据再次聚类。在一组真实数据集上的实验表明,所提出的算法不仅可以有效地处理高维数据,还提高了聚类性能。

关键词:半监督聚类;成对约束;主成分分析;线性判别分析
DOI:10.3778/j.issn.1002-8331.2010.06.040 **文章编号:**1002-8331(2010)06-0139-05 **文献标识码:**A **中图分类号:**TP311

1 引言

在机器学习和数据挖掘领域中,人们经常遇到大量的无类标号数据。对这些无类标号数据进行标号时,不仅费时费力,有时甚至要付出相当大的代价,如会谈中说话人语音的分割与识别[1],GPS 数据中的道路检测[2]和电影片段中不同男演员或者女演员的分组问题[3]等。因此,利用样本的先验信息或者背景知识,来解决这一问题,已成为机器学习领域的研究热点。半监督聚类正是利用样本的先验信息或者背景知识,并结合无标号数据,来完成对样本数据聚类。它能自然地应用到无监督聚类中,提高无监督聚类性能和质量,故成为最近机器学习和数据挖掘领域中具有重要意义的研究课题。目前,半监督聚类正受到越来越多的学者和研究人员的广泛关注[4-10]。

现有的半监督聚类算法可以分为三类。首先,基于约束的半监督聚类算法。这类算法一般使用两类叫 must-link 和 cannot-link 成对约束来引导聚类过程[2,4],如图 1 所示。这两类约束



(a)成对约束关系 (b)聚类结果

图 1 成对约束关系图例和它对聚类结果的影响

(在图(a)中,实线连接的两个样本点表示的是一个 must-link 成对约束,虚线连接的两个样本点表示的是一个 cannot-link 成对约束。在图(b)中,由成对约束关系得到的聚类结果)

作者简介:陈小冬(1978-),女,讲师,主要研究领域为模式识别、智能计算;尹学松(1975-),男,讲师,主要研究领域为模式识别、智能计算;林焕祥(1975-),男,讲师,主要研究领域为模式识别、智能计算。

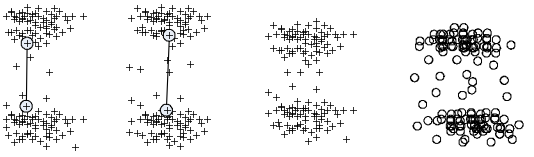
收稿日期:2008-09-08 **修回日期:**2008-12-15

首先是由 Wagstaff 等人提出的[2],旨在提高无监督聚类算法的性能。must-link 和cannot-link约束含义如下:

must-link 约束规定:如果两个样本属于must-link 约束,那么这两个样本在聚类时必须被分配的同一个聚类中。

cannot-link约束规定:如果两个样本属于cannot-link约束,那么这两个样本在聚类时必须被分配的不同聚类中。

其次,基于距离的半监督聚类算法。这类算法通过对成对约束进行学习,得到一个好的距离度量,从而改进样本之间的距离,提高聚类性能[3,5],如图 2 所示。现有的基于距离的半监督聚类算法,一般是利用凸优化技术,得到一个马氏距离。



(a)辅助信息 (b)聚类结果

图 2 辅助信息图例和它对聚类结果的影响

(在图(a)中,实线连接的两个样本点表示的是属于同一类别。在图(b)中,通过学习这些辅助信息得到一个好的距离度量后,所得到的聚类结果)

最后,集成成对约束与距离的半监督聚类算法。这类算法就是将must-link、cannot-link约束与距离度量同时应用到聚类中,来提高聚类的质量[6-7]。

尽管上述三类算法都能在某种程度上提高了聚类的性能,但它们只能适用于维数较低的样本数据中。在遇到维数较高的数据时,这些算法便显得力不从心,因为在高维数据空间中,对于不同数据分布和距离函数的样本点对之间的距离几乎是相同的[11]。在一些实际应用中,如信息检索、计算生物和图像处理等,所处理的数据都是高维的,研究有效的聚类算法既克服这些数据的维数灾难,又能提高聚类质量是一个严峻的挑战。为了解决这一问题,Wei Tang[8]等人提出基于特征投影的半监督聚类算法。该方法可以解决高维数据聚类问题,但它的缺点是仅仅使用成对约束来完成对高维数据的降维,没有考虑到大量无标号样本数据,以及降维以后样本之间的距离度量,因此限制了聚类性能的提高。

最近,一些学者针对高维数据的无监督聚类,又提出了新的方法。De la Torre[12]等人提出判别聚类分析算法。该算法集成了降维和聚类,即首先对高维数据降维,其次在低维空间中对数据聚类。Chris Ding[13]等人提出了一种自适应无监督降维迭代算法。该算法使用均值聚类来产生数据的类标号,然后用线性判别分析方法对高维数据降维。在降维空间中,再用均值方法对数据聚类。Jeiping Ye[14-15]提出自适应距离学习聚类算法。该算法同样是集成降维和聚类,不同的是,算法是最优化同一个目标函数来得到聚类和降维。以上这三种方法的目的是通过降维来提高聚类性能,再利用聚类结果来指导降维,但它们都面临一个问题,即算法在高维空间里是先执行降维还是先执行聚类。

文中提出一种新颖的基于判别分析的半监督聚类算法(Semi-Supervised Clustering Approach with Discriminant Analysis,SCADA)。详细地说,新算法首先利用主成分分析(Principal Components Analysis,PCA)得到初始投影矩阵,在投影空间中对数据聚类;其次,根据聚类标号,利用线性判别分析(Linear Discriminant Analysis,LDA)投影原空间样本;最后,使用基于球形的均值算法对新投影空间中的数据聚类。一方面,新算法将动态聚类方法引入半监督聚类中,即聚类和降维同时进行。现有的半监督聚类方法要么只关注辅助信息对聚类的帮助[4-7],忽略了对数据的降维,要么分离了聚类与降维[8]。新算法利用聚类结果来指导降维,然后又利用得到投影矩阵来提高聚类,两者迭代进行,有效地提高了聚类性能。另一方面,新算法将 LDA 应用到半监督聚类中。LDA 是有监督的维数约减方法,通常需要样本的类标号来完成对样本的降维。新算法通过利用成对约束得到较好的聚类结果为 LDA 提供聚类标号,从而得到线性嵌入空间。

文中提出的算法,不仅集成投影空间选择与数据聚类,还努力架起一座连接原空间中样本和投影空间中样本的桥梁。通过这个桥梁,可以在全局最优的投影空间中对数据聚类,避免了维数灾难。

2 提出的方法

对于给定的样本集合 $X=[x_1,x_2,\cdots,x_n]$,其中 $x_i\in\Re^m$,其中样本已经被归一化。must-link 成对约束集合 $M=\{(x_i,x_j)\}$、cannot-link 成对约束集合 $C=\{(x_i,x_j)\}$,新的半监督算法由以下三步组成。首先是利用 PCA 得到一个初始投影矩阵,在投影空间中对数据聚类;其次,根据聚类结果,利用 LDA 将数据投影到一个低维空间中;最后,基于球形的 K 矩阵聚类。

2.1 PCA

PCA 的基本思想是将数据沿着最大变化的方向投影,因此能最小化重建误差。令 $\boldsymbol{w}$ 是投影矩阵,$y_i=\boldsymbol{W}^{\mathrm{T}}x_i$,则 PCA 的目标函数如下:

$$W_{opt}=\arg\max_{W}\sum_{i=1}^{n}\|y_i-\bar{y}\|^2=\arg\max tr(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{W}) \tag{1}$$

其中$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$,$\boldsymbol{S}$ 是输入数据的协方差矩阵。最优的投影矩阵 $\boldsymbol{W}$ 是由 $\boldsymbol{S}$ 的 $d(d<m)$个最大特征值对应的特征向量组成。

2.2 LDA

LDA 是利用有监督的信息来得到最佳判别投影方向。具体地说,LDA 是最大化类间距离,而最小化类内聚类。LDA 的目标函数是:

$$W_{opt}=\arg\max tr\left(\frac{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_B\boldsymbol{W}}{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_W\boldsymbol{W}}\right) \tag{2}$$

$$\boldsymbol{S}_B=\sum_{i=1}^{C}n_i(m_i-m)(m_i-m)^{\mathrm{T}} \tag{3}$$

$$\boldsymbol{S}_W=\sum_{j=1}^{C}\sum_{i=1}^{n_j}(x_i-m_j)(x_i-m_j)^{\mathrm{T}} \tag{4}$$

其中,C 是类数,m 是整个样本均值,m_i 是第 i 类样本均值,n_i 是第 i 类样本数。$\boldsymbol{S}_B$ 是类间散布矩阵,$\boldsymbol{S}_W$ 是类内散布矩阵。

2.3 基于球形的 K 矩阵聚类算法

must-link 约束可以表示点对之间的等价关系,cannot-link 约束表示点对之间不等价关系。这样,借助于传递闭包,对 must-link 约束的点对合并,如图 3 所示,实线是表示 must-link 约束,虚线表示 cannot-link 约束,白点表示原始样本,黑点代表传递闭包中的均值样本。其中,$\{a_1,a_2,a_3\}$,$\{b_1,b_2,b_3,b_4,b_5\}$,$\{d_1,d_2,d_3,d_4\}$和$\{e_1,e_2,e_3\}$分别表示不同的传递闭包,a,b,d 和 e 分别代表它们的均值样本。原数据的 cannot-link 约束,就可以用 a,b,d 和 e 表示,为集合 C',如图 3 中的右图所示。

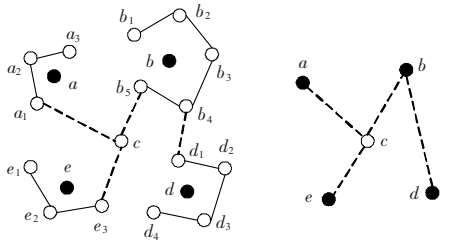


图 3 合并 must-link 约束构成新的 cannot-link 约束

借助于上面的操作,数据 X 可简化为 $X'=[x_1',x_2',\cdots,x_{n'}']$ $(n'<n)$,cannot-link 成对约束的集合为 $C'=\{(x_i',x_j')\}$。因此,对于 cannot-link 约束,$(x_i',x_j')\in C'$, 要找到两个不同的聚类中 u_{c_i} 和 u_{c_j},使

$$\| x_i'-u_{c_i} \|^2+\| x_j'-u_{c_j} \|^2 \tag{5}$$

最大化,就分配 x_i'、x_j'分别到 u_{c_i} 和 u_{c_j} 聚类中。基于球形的 K 均值算法(PCBKM)如图 4 所示。

算法:基于球形的 K 均值算法
输入:样本数据 X'、cannot-link 约束集合 C'和聚类数目 K
输出:K 个不相交的样本分割
步骤 1 初始化聚类中心
步骤 2 重复执行下面步骤,直到收敛
for i=1 to n'
(a)对每一个不在任何 cannot-link 成对约束中的样本 x_i',找到一个聚类中心使 $c_i=\text{argmax}_i \| x_i'-u_{c_i} \|^2$
(b)对于每个属于 cannot-link 约束$(x_i'-x_j')\in C'$的点对,找到两个聚类中心,使 $\max(\| x_i'-u_{c_i} \|^2+\| x_j'-u_{c_j} \|^2)$
(c)对某个聚类 c_i,更新其聚类中心 $u_{c_i}=\sum_{x'\in X_{c_i}'}/\| \sum_{x'\in X_{c_i}'} \|$
步骤 3 K 个不相交的样本分割

图 4 基于球形的 K 均值算法

2.4 基于判别分析的半监督聚类算法

基于上面的描述,提出一个同时执行降维和聚类的迭代算法,称之为判别型的半监督聚类算法。该算法的步骤如下:

算法:基于判别分析的半监督聚类算法(SCADA)

输入:样本数据 X、must-link 和cannot-link 约束、聚类数目 K

输出:K 个不相交的样本分割

步骤 1 (1)使用 PCA,求解初始矩阵 $\boldsymbol{W}$。

(2)在投影空间中使用基于球形的 K 均值算法对样本聚类。

步骤 2 重复执行下面步骤,直到收敛。

(1)利用聚类结果,使用 LDA 更新投影矩阵 $\boldsymbol{W}$。

(2)利用投影矩阵 $\boldsymbol{W}$ 投影样本,在投影空间中使用基于球形的 K 均值算法对样本聚类。

步骤 3 K 个不相交的样本分割。

类似于算法[12-15],判别型的半监督聚类算法收敛于有限的步骤。在实验中观察到,算法迭代 4 到 5 次即收敛。

基于球形的 K 均值算法的计算复杂度是 $O(tnl^2)$,SCADA 的计算复杂度为 $O(ptnl^2)$,l 是样本维数,n 是样本数,t 是 K 均值算法迭代次数,p 是自适应度量学习半监督聚类算法迭代次数。

3 实验

3.1 实验搭建

为了评价所提出算法的有效性, 使用 K 均值聚类算法作为基线,与其他三个相关的算法进行对比。这三个算法是:

(1)相关成分分析算法(RCA)[3],是一种半监督度量学习算法。该方法的实验结果已经证明其聚类性能优于基于度量的半监督聚类算法[5]。使用该算法得到度量,然后利用 K 均值聚类算法对变换后的样本聚类;

(2)集成度量与成对约束的半监督聚类算法(MPCKM)[7]是度量学习和成对约束集成到均值聚类。在聚类高维数据时,首先使用 PCA 对数据降维,然后使用该方法对数据聚类;

(3)特征投影的半监督聚类算法(SCREEN)[8],是 SIGKDD'2007 上提出的一种聚类效果较好的半监督聚类算法。该算法即可用于低维数据聚类,也可用于高维数据聚类。

几个算法的性能在 6 个真实数据集上进行测试,其中三个文本数据,两个人脸数据和一个 UCI 数据。这 6 个数据集的特性如表 1 所示。

表 1 一组数据集的属性概要

数据集	样本数	维数	聚类数
DOC1	3 970	3 759	5
DOC2	4 500	2 887	4
DOC3	4 300	7 455	7
YALEB	110	2 500	10
ORL	100	1 024	10
Vehicle	846	18	4

为了全面客观地评价新算法与其他算法的性能,文中使用两种评价度量。第一种聚类算法评价测度是 Rand Index[5,9]。该评价度量通过测试由聚类算法得到的聚类标号和原样本的类标号差异,反映聚类算法的有效性。其具体描述为:

$$Acc=(N_S+N_D)/N_P \tag{6}$$

其中,N_S 是一组被聚类到相同类的样本对的数目,即两个样本原来属于同一类,在聚类后,被聚类到同一类;N_D 是一组被聚类到不同类的样本对的数目, 即两个样本原来属于不同类,在聚类后, 被聚类到不同类;N_P 是所有样本组成的样本对数目,其值为 $n(n-1)/2$。

第二种聚类算法评价测度是规范化的互信息(NMI)[14-15]。如果 C 是样本聚类以后的类标号,Y 是样本原有的类标号,则规范化的互信息表示为:

$$NMI=\frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \tag{7}$$

其中 $I(C;Y)=H(Y)-H(Y|C)$是 C 和 Y 之间的互信息,$H(Y)$是 Y 的香农熵,$H(Y|C)$是给定 C 的条件下,Y 的条件熵。NMI 值范围在 0 和 1 之间,NMI 值越大,聚类的性能就越好。

最后, 五个算法运行在 Intel Pentium 3.00 GHz CUP,1G 内存 Windows 环境下的机器上。每个算法在每个数据集上测试 20 次,取它们的均值作为最终的聚类结果。

3.2 实验结果

图 5 和图 6 分别展示了五个算法的 Acc 和 NMI 值, 使用的成对约束分别从 50 到 500 对,学习曲线上的每个点是 20 次不同成对约束的平均值。

从图 5 和图 6 可以看出如下几点:

(1)SCADA 无论是在少量约束数量还是大量约束数量条件下,Acc 值和 NMI 值都要高于其他几个算法,这说明 SCADA 在大多数数据集上,性能都优于其他几个算法。

(2)SCREEN 在 Vehicle 数据集上性能最优,在其他几个数据集上,其性能虽然没有 SCADA 好,但也高于另几个算法的性

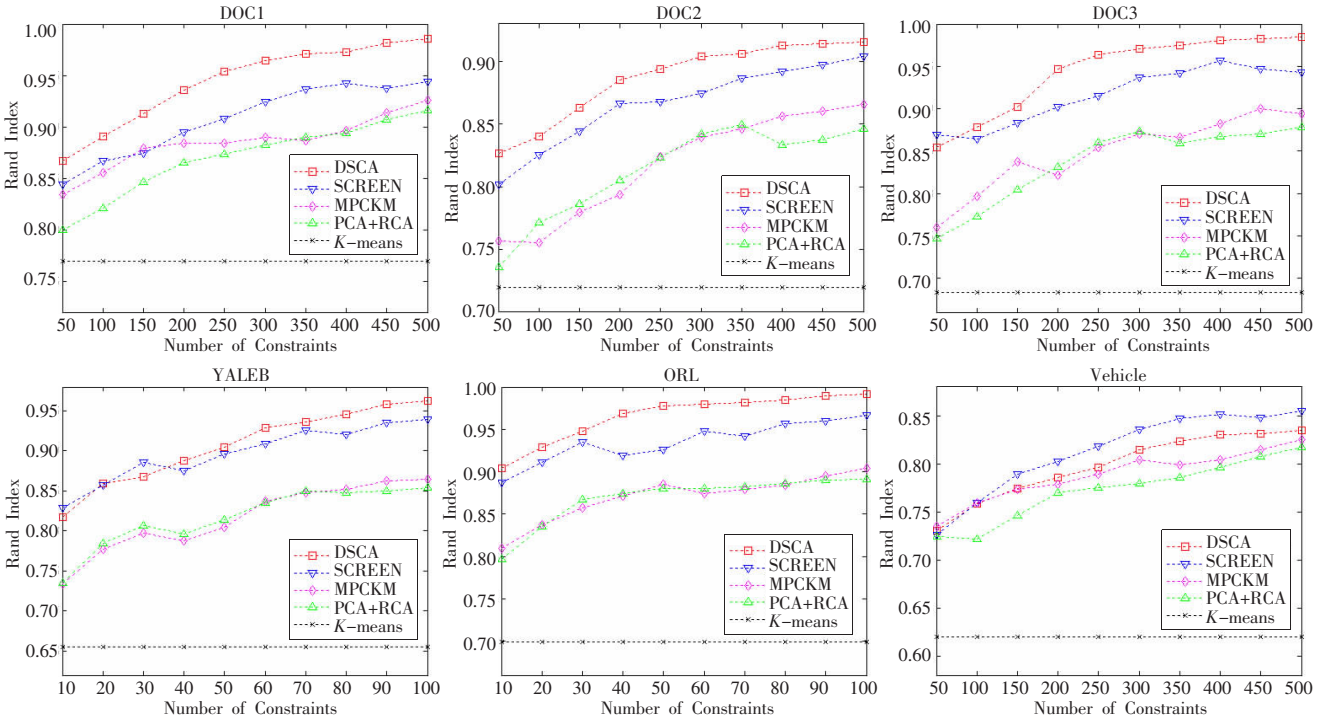


图 5　在 6 个数据集上由 *Acc* 评价度量得到的五个算法的聚类性能

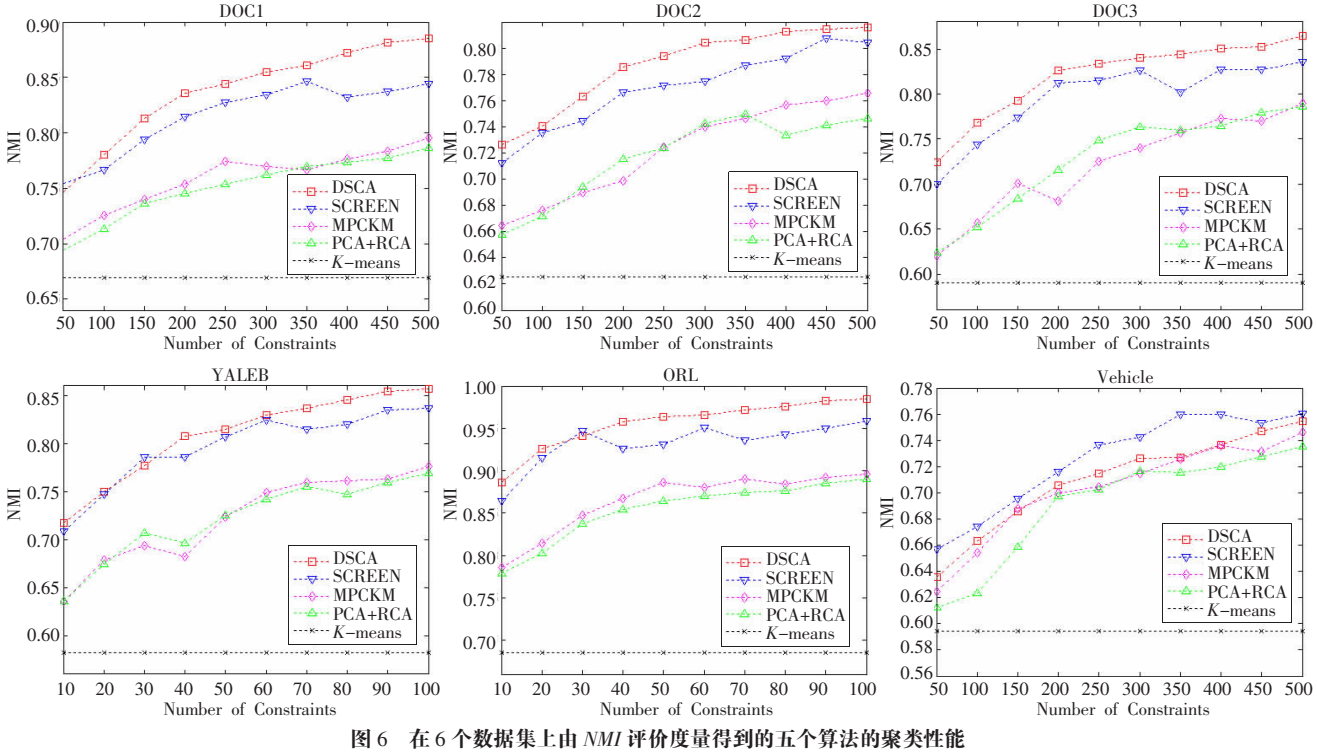


图 6　在 6 个数据集上由 *NMI* 评价度量得到的五个算法的聚类性能

能。因此,SCREEN 也是一个有效地高维数据的半监督聚类算法。

(3)从图 5 和图 6 可以观察到,在成对约束逐渐增多时,SCADA 的学习曲线平稳上升,因此,不能发现,SCADA 比较稳定,而 SCREEN 的学习曲线随着成对约束的增多,变化较大,有时甚至出现随之成对约束增加,性能下降的情况。因此,SCADA 比 SCREEN 稳定。

(4)MPCKM 和 RCA 在三个文本数据集和两个人脸数据集上聚类效果都不理想。因此,这两个算法都不适合处理高维数据。相反,SCADA 和 SCREEN 都适合对高维数据聚类。

4　结论

文中提出了一种新颖的基于判别分析的半监督聚类算法。新算法首先利用 PCA 得到初始投影矩阵,对数据降维,得到初始聚类;其次,利用得到的聚类结果指导 LDA 对数据降维;最后,用基于球形的 K 均值聚类算法对新投影空间中的数据聚类。新算法是迭代执行降维和聚类,即用聚类的结果得到降维矩阵,然后在投影空间中,再对数据聚类。实验结果表明,新算法不仅很好地处理高维数据,还有效地提高了聚类性能。

参考文献:

[1] Bar-Hillel A,Hertz T,Shental N,et al.Learning a mahalanobis metric from equivalence constraints[J].Journal of Machine Learning Research,2005,6:937-965.

[2] Wagstaff K,Cardie C,Rogers S,et al.Constrained *K*-means clustering with background knowledge[C]//Proceedings of the 18th International Conference on Machine Learning (ICML),San Francisco,2001:577-584.

[3] Bar-Hillel A,Hertz T,Shental N,et al.Learning distance functions using equivalence relations[C]//Proceedings of the 20th International Conference on Machine Learning (ICML),Washington DC,USA,2003:11-18.

[4] Basu S,Banerjee A,Mooney R J.Semi-supervised clustering by seeding [C]//Proceedings of the 19th International Conference on Machine Learning(ICML),Sydney,Australia,2002:19-26.

[5] Xing E P,Ng A Y,Jordan M I,et al.Distance metric learning,with application to clustering with side-information[C]//Advances in Neural Information Processing Systems 15 (NIPS),Cambridge,MA,2003:505-512.

[6] Basu S,Bilenko M,Mooney R J.A probabilistic framework for semi-supervised clustering[C]//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD),Seattle,WA,2004:59-68.

[7] Bilenko M,Basu S,Mooney R J.Integrating constraints and metric learning in semi-supervised clustering[C]//Proceedings of the 21st International Conference on Machine Learning(ICML),Banff,Canada,2004:81-88.

[8] Tang W,Xiong H,Zhong S,et al.Enhancing semi-supervised clustering:A feature projection perspective[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(SIGKDD),San Jose,California,USA,2007:707-716.

[9] Yan B,Domeniconi C.An adaptive kernel method for semi-supervised clustering[C]//Proceedings of the 17th European Conference on Machine Learning(ECML),Berlin,Germany,2006:18-22.

[10] Liu Y,Jin R,Jain A K.BoostCluster:Boosting clustering by pairwise constraints[C]//Proceedings of the The Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(SIGKDD),San Jose,California,USA,2007:450-459.

[11] Beyer K,Goldstein J,Ramakrishnan R,et al.When is nearest neighbors meaningful?[C]//Proceedings of International Conference on Database Theory(ICDT),Jerusalem,Israel,1999:217-235.

[12] dela Torre F,Kanade F.Discriminative cluster analysis[C]//Proceedings of the 19th International Conference on Machine Learning(ICML),Pittsburgh,Pennsylvania,USA,2006:241-248.

[13] Ding C H Q,Li T.Adaptive dimension reduction using discriminant analysis and *K*-means clustering[C]//Proceedings of the 19th International Conference on Machine Learning(ICML),Corvalis,Oregon,USA,2007:521-528.

[14] Ye Jie-ping,Zhao Zheng,Liu Huan.Adaptive distance metric learning for clustering[C]//Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR),Minneapolis,Minnesota,USA,2007.

[15] Chen J,Zhao Z,Ye J,et al.Nonlinear adaptive distance metric learning for clustering[C]//Proceedings of the The Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(SIGKDD),San Jose,California,USA,2007:123-132.

[16] Basu S,Banerjee A,Mooney R J.Active semi-supervision for pairwise constrained clustering[C]//Proceedings of the SIAM International Conference on Data Mining (SDM),Lake Buena Vista,FL,2004:333-344.

[17] Yeung D Y,Chang H.Extending the relevant component analysis algorithm for metric learning using both positive and negative equivalence constraints[J].Pattern Recognition,2006,39(5):1007-1010.

[18] Zhang D,Zhou Z,Chen S.Semi-supervised dimensionality reduction[C]//Proceedings of the 7th SIAM International Conference on Data Mining(SDM),Minneapolis,MN,2007.

(上接 138 页)

据集上的实验表明,FCluStream 算法是一种简单、高效的数据流聚类算法。下一步工作是进一步改进算法,进而提高算法的计算效率。

参考文献:

[1] Barbará D,Chen P.Using the fractal dimension to cluster datasets[C]//Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2000.

[2] Guha S,Mishra N,Motwani R,et al.Clustering data streams[C]//IEEE Symposium on Foundations of Computer Science,2000.

[3] O'Callaghan L,Mishra N,Guha S.Streaming-data algorithms for high-quality clustering[C]//ICDE Conf,2002:685-704.

[4] Aggarwal C C,Han Jia-wei,Wang Jian-yong,et al.A framewrok for clustering evolving data streams[C]//Proceedings of VLDB 2003,2003:81-92.

[5] Cao F,Estery M,Qian W,et al.Density-based clustering over an evolving data stream with Noise[C]//Proceedings of the 2006 SIAM Conference on Data Mining(SDM'2006),2006.

[6] 朱蔚恒,印鉴,谢益煌.基于数据流的任意形状聚类算法[J].软件学报,2006,17(3):379-386.

[7] 刘青宝,戴超凡,邓苏,等.基于网格的数据流聚类算法[J].计算机科学,2007,34(3):159-161.

[8] Elaine P M,de Sousa A,Traina J M,et al.SID:Calculating the intrinsic dimension of data streams[C]//Proceedings of the 2006 ACM Symposium on Applied Computing,2006.

[9] 颜晓龙,沈鸿.一种适用于高维数据流的字空间聚类算法[J].计算机应用,2007,27(7):1680-1684.

[10] Aggarwal C C,Han Jia-wei,Wang Jian-yong,et al.A framewrok for projected clustering for high dimensional data stream[C]//Proceedings of VLDB,2004.

[11] Meyerson A,Mishra N,Motwani R,et al.Clustering data streams:Theory and practice[J].IEEE Transaction on Knowledge and Data Engineering,2003,15(3):505-528.

[12] Gaber M M,Zaslavsky A B,Krishnaswamy S.Mining data streams:A review[J].SIGMOD Record,2005,34(2):18-26.

[13] Barbará D.Chaotic mining:Knowledge discovery using the fractal dimension[C]//1999 ACM SIGKDD Workshop on Research Issues in Data Mining and Knowledge Discovery,Philadelphia USA,1999.

[14] Barbará D.Requirements for clustering data streams [J].ACM SIGKDD Explorations Newsletter,2003,3(2):23-27.

[15] Traina C,Traina A,Wu L,et al.Fast feature selection using fractal dimension[C]//Proc XV Brazilian Symposium on Databases,2000.

[16] 孙霞,吴自勤,黄韵.分形原理及应用[M].合肥:中国科学技术大学出版社,2003:23-24.