基于流形距离的人工免疫半监督聚类算法

李岩波[1] 宋 琼[2] 郭新辰[2]
(吉林大学数学学院 长春 130012)[1] (东北电力大学理学院 吉林 132012)[2]

摘 要 将流形距离作为样本间相似性的基本度量测度,加入成对约束信息,通过近邻传播得出新的度量矩阵。把聚类问题转化为一优化数学模型。采用克隆选择算法求解这个优化模型,得出最后的聚类结果,通过人工数据集和UCI标准数据集验证了这种方法具有较高的准确性。

关键词 流形距离,半监督聚类,人工免疫算法
中图法分类号 TN915 **文献标识码** A

Artificial Immune Clustering Semi-supervised Algorithm Based on Manifold Distance

LI Yan-bo[1] SONG Qiong[2] GUO Xin-chen[2]
(School of Mathematics, Jilin University, Changchun 130012, China)[1] (College of Science, Northeast Dianli University, Jilin 132012, China)[2]

Abstract Manifold distance was used as the basic measure of the sample similarity between samples. The pair-wise constrains prior information was introduced, then the measure matrix was obtained through affinity propagation. So the clustering problem was transformed as one optimal model. Clonal selection algorithm was employed to solve this model, and the clustering results were given. Experiments on artificial data sets and UCI benchmark data set show that the proposed method can give the better accuracy.

Keywords Manifold distance, Semi-supervised clustering, Artificial immune algorithm

半监督聚类主要是将少量先验信息加入到原本无监督的聚类过程中,以提高聚类性能。它结合了无监督学习和监督学习的特点,现已成为机器学习、数据挖掘等领域的研究热点。

现有的半监督聚类方法按照监督信息使用方式的不同可以分为以下几种:

1. 基于约束的方法。主要是利用先验信息(如成对约束)来指导聚类过程。

文献[1]采用修改聚类目标函数的方法,在原始目标函数中增加了一个包含样本标记信息的成分,函数形式如下:

$$\min_{m_k,k=1,\cdots,K} \beta\times Cluster\_Dispersion+\alpha\times Cluster\_Impurity$$

式中,$\alpha>0$,$\beta>0$。$Cluster\_Dispersion$ 是指聚类散度,$Cluster\_Impurity$ 是指类内混杂度,文中采用遗传算法求解。

2. 基于距离的方法。这类方法是训练样本间的相似性度量,使其满足给定的先验信息,聚类过程在新的样本间相似性度量上进行。如 Xing 等人[2]学习度量矩阵 A,定义两个样本 x,y 之间的距离为:

$$d(x,y)=\|x-y\|_A=\sqrt{(x-y)^{\mathrm{T}}A(x-y)}$$

构造优化函数:

$$\min_{A}\sum_{(x_i,x_j)\in S}\|x_i-x_j\|_A^2$$

$$\text{s.t.}\sum_{(x_i,x_j)\in D}\|x_i-x_j\|_A^2\geqslant 1$$

$$A\geqslant 0$$

求解度量矩阵 A 使得属于同一类的样本点间的距离较近,不同类中的样本间距离较远。

3. 基于约束和距离融合的方法

Basu 等人提出一个基于 K-means 的半监督算法,该算法融合了基于约束和基于距离的方法,即在 K-means 算法基础上,引入成对约束作为监督信息,体现在目标函数上就是增加了不满足约束的惩罚项[3]。Basu 等人随后提出一种基于隐马尔可夫随机域模型的半监督聚类方法,将算法度量从欧式距离推广到任意的 Bregman 散度[4]。

还有基于核的方法,如文献[5-7]将数据点映射到核空间,通过在核空间构造新的距离进行聚类,取得了比较好的效果。

本文提出一种基于流形距离的人工免疫半监督聚类算法。即将成对约束信息加入到样本度量矩阵中,通过近邻流形距离得到的新的度量矩阵,把聚类问题转化为以各类心为变量的优化模型,在采用克隆选择算法求得各最优类心后,将各样本归类。

1 近邻流形距离与克隆选择算法介绍

1.1 近邻流形距离

在聚类过程中,一般做如下假设:

到稿日期:2012-02-05 返修日期:2012-04-23 本文受吉林省自然科学基金项目(201215165),符号计算与知识工程教育部重点实验室开放基金项目(93K-17-2010-K05)资助。

李岩波(1972—),女,博士,副教授,主要研究方向为智能计算及应用,E-mail:57458030@qq.com。

局部相似性：空间上距离比较近的样本点具有较高的相似性。

全局相似性：在同一流形上的样本点具有较高的相似性。

欧氏距离是人们最为熟悉的一种度量距离的方法，然而在聚类过程中，它有一定的不足，如图 1 所示。

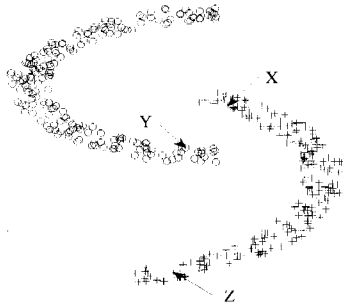

图 1　流形样本

图 1 有两个流形，数据点 X 和 Z 位于同一流形上，数据点 Y 位于另一流形上，在欧氏距离度量下，点 X 和 Y 的距离要小于 X 和 Z 的距离。可见，若用距离来度量两个样本的相似性，欧氏距离有所不足，流形距离[8,9]的引入就很好地解决了上述困难。

定义 1（两样本间的 k 近邻距离）　若样本点 x_j 是点 x_i 的 k 近邻，其距离表示为 $\rho^{\|x_i-x_j\|}-1$，否则为 0，表示如下：

$$L(x_i,x_j)=\begin{cases}\rho^{\|x_i-x_j\|}-1, & x_j \text{ 是 } x_i \text{ 的 } k \text{ 近邻}\\ \infty, & \text{else}\end{cases} \tag{1}$$

$\rho>1$ 为调节因子，用来放大或缩小两点的长度，$\|x_i-x_j\|$ 一般取与 x_i 和 x_j 的欧氏距离。

定义 2　将数据点看作是图 $G(V,E)$ 的顶点，每对顶点的边的权重为其 k 近邻距离，令 $p\in V^l$ 表示连接点 p_1 和 p_l 的长度为 l 的路径，p_{ij} 表示连接 x_i 与 x_j 的所有路径，则两个数据点 x_i 与 x_j 的流形距离 $D(x_i,x_j)$ 为连接两点间的所有路径长度的最小值。

$$D(x_i,x_j)=\min_{p\in p_{ij}}\sum_{k=1}^{|p|-1}L(p_k,p_{k+1}) \tag{2}$$

1.2　克隆选择算法

焦等人受免疫系统克隆选择机理启发，构造了一种克隆选择算子[10]，机体在抗原的刺激作用下产生抗体群，抗体群循环进行克隆操作、免疫基因操作（变异操作）、克隆选择操作，最终达到抗体抗原亲合力成熟。

一般地，考虑求解优化问题：

$\max f(X), X\in R^n$

在克隆选择算法中，优化问题本身视为抗原，将变量 X 的编码串 $A=a_1a_2\cdots a_l$ 视为抗体，$A\in I$，编码函数 $A=e(X)$；解码 $X=e^{-1}(A)$；$f(e^{-1}(A))$ 为抗体 A 的抗体抗原亲合度函数；n 个抗体的抗体种群空间表示为 $\tilde{A}=[A_1,A_2,\cdots,A_n]$，$A_k\in I$，$1\leqslant k\leqslant n$。则主要操作可表述为：

克隆操作：抗体群 $\tilde{A}(k)$ 经克隆操作 T_c^C 变为 $Y(k)=[T_c^C(A_1(k))\ T_c^C(A_2(k))\ \cdots\ T_c^C(A_n(k))]$，$k$ 为迭代的次数，其中 $T_c^C(A_i(k))=I_i\times A_i(k)$，$i=1,2,\cdots n$，$I_i$ 是元素为 1 的 q_i 维行向量，$T_c^C(A_i(k))$ 表示把抗体 $A_i(k)$ 克隆了 q_i 次。克隆个数 q_i 为该抗体与抗原的亲合度 $f(A_i(k))$、抗体与抗体的亲合力 θ_i 的函数，定义为：

$$q_i(k)=g(n_c,f(A_i(k)),\theta_i)$$

式中，$\theta_i=\min\{D_{ij}\}=\min\{\exp\|A_i-A_j\|\}$，$i\neq j,i,j=1,2,\cdots,n$。

一般取

$$q_i(k)=Int\left(n_c\cdot\frac{f(A_i(k))}{\sum_{j=1}^{n}f(A_j(k))}\cdot\theta_i\right)$$

$n_c>n$ 是设定值，与克隆规模有关；$Int(\cdot)$ 是向上取整函数。克隆后的抗体群记为 $Y(k)=\{Y_1(k),Y_2(k),\cdots,Y_n(k)\}$，$Y_i(k)=T_c^C(A_i(k))$，$i=1,2,\cdots n$。这样定义之后，当某个抗体的亲合力较大且该抗体周围的抗体浓度较小时，其克隆规模较大。

免疫基因操作：主要是变异操作。依事先设定的变异概率 p 对克隆后的群体进行变异，得抗体群 $Z(k)=T_g^C(Y(k))$，记为 $Z(k)=\{Z_1(k),Z_2(k),\cdots,Z_n(k)\}$。

克隆选择操作：从抗体群 $Z(k)$ 中选择 n 个抗体，得新抗体群 $\tilde{A}(k+1)$，选择方法如下：记 $B_i(k)=\max\{Z_i(k)\}$，则选择概率 p_s：

$$p_s(\tilde{A}(k+1)=B_i(k))=$$

$$\begin{cases}1, & f(A_i(k))<f(B_i(k))\\ \exp\left(-\dfrac{f(A_i(k))-f(B_i(k))}{\alpha}\right), & f(A_i(k))\geqslant f(B_i(k))\text{且 }A_i(k)\text{不是目前种群最优抗体}\\ 0, & f(A_i(k))\geqslant f(B)\text{且 }A_i(k)\text{是目前种群最优抗体}\end{cases}$$

$$p_s(\tilde{A}(k+1)=A_i(k))=$$

$$\begin{cases}0, & f(A_i(k))<f(B_i(k))\\ 1-\exp\left(-\dfrac{f(A_i(k))-f(B_i(k))}{\alpha}\right), & f(A_i(k))\geqslant f(B_i(k))\text{且 }A_i(k)\text{不是目前种群最优抗体}\\ 1, & f(A_i(k))\geqslant f(B)\text{且 }A_i(k)\text{是目前种群最优抗体}\end{cases}$$

抗体群演化过程可表述为：

$$\tilde{A}(k)\xrightarrow{\text{克隆操作}}Y(k)\xrightarrow{\text{免疫基因操作}}Z(k)\xrightarrow{\text{克隆选择操作}}\tilde{A}(k+1)$$

文献[9]中用马尔可夫链的知识已证明克隆选择算法是依概率 1 收敛于全局最优解的。

2　基于流形距离的人工免疫半监督聚类算法

本文将成对约束先验信息加入到聚类过程中，成对约束表述为一对样本点 x_i 与 x_j，若 $(x_i,x_j)\in must\text{-}link$，规定点 x_i 与点 x_j 在同一类中，它们具有很强的相似性；若 $(x_i,x_j)\in cannot\text{-}link$，规定点 x_i 与点 x_j 不在同一类中，它们具有较大的差异性。

算法具体步骤如下：

（1）构造度量矩阵。首先计算样本间的欧氏距离矩阵，将成对约束加入到距离阵中，若两样本 $(x_i,x_j)\in must\text{-}link$，则令它们的距离为 0；若两样本 $(x_i,x_j)\in cannot\text{-}link$，则令它们的距离为 ∞。根据 1.1 节的介绍构造样本点间的流形 k 近邻度量矩阵。

（2）生成目标准则函数。设样本集各类的中心 $u=(u_1,$

$u_2,\cdots,u_K$)为样本集中的 K 个样本。将样本点 $x_i(i=1,2,\cdots,n)$依流形距离最短划分到类 C_j 中：

$$j=\min_{j=1,2,\cdots K}(D(x_i,u_j))$$

$D(x_i,u_j)$表示样本 x_i 到类心 u_j 的流形距离。

则最优类心 u 可以表示为使式(3)达到最大值的 K 个样本点：

$$F(u)=\frac{1}{1+\sum_{k=1}^{K}\sum_{x_i\in C_k}D(x_i,u_k)} \tag{3}$$

(3) 用克隆选择算法求解目标准则函数(3)的最大值。将优化函数本身作为抗原输入，类中心作为抗体。

(4) 求得最优类心后，依流形距离最短原则将样本点进行归类。

3 实验结果与分析

为了验证所提算法的可行性，将分别在人工数据集和真实数据集上进行实验。所有相关计算在 PC 机(Pentium(R) 4CPU，2.40GHz，256MB 内存)上用 Matlab 7.0.4 编程实现。

为了评价算法的有效性，设置误差率指标 ER(Error Rate)：

$$ER=\frac{a}{n}$$

式中，n 代表数据的个数，a 表示分类错误的数据的个数。

3.1 人工数据集实验与分析

人工数据集如图 2、图 3 所示。数据集 1(见图 2)的聚类目标是将大圆环上数据点聚为一类，小圆环上的点聚为一类；数据集 2(见图 3)的聚类目标是把对角的两组团状数据聚为一类，分为两类。

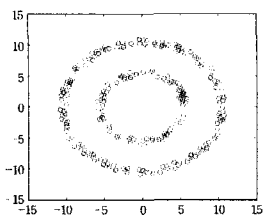

图 2 人工数据集 1

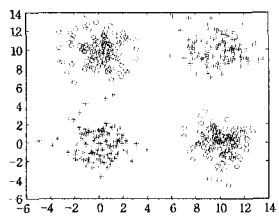

图 3 人工数据集 2

首先用 K-means 算法进行聚类，聚类结果如图 4、图 5 所示，相同颜色的样本点属于同一类，聚类的误差率 ER 都约为 0.5。

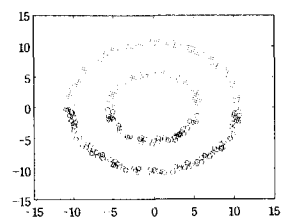

图 4 人工数据集 1 K-means 聚类结果

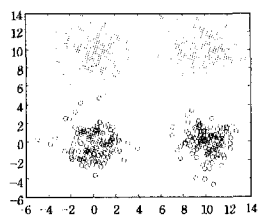

图 5 人工数据集 2 K-means 聚类结果

然后用本文提出的算法进行聚类，实验中近邻数 $KN=10$，调节因子 $\rho=2$，聚类结果如图 6、图 7 所示，聚类误差 ER 都约为 0。

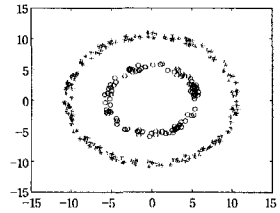

图 6 人工数据集 1 本文算法结果

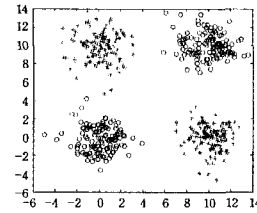

图 7 人工数据集 2 本文算法结果

可见，K-means 算法对于不是团状的数据集聚类误差率高，而本文所提算法使得位于空间同一流形上的样本在近邻流形距离的度量下距离变小。引入少量的先验信息，就会提高聚类的性能，得到很好的聚类效果。

3.2 直觉模糊最小二乘支持向量机

在真实的 UCI 标准数据集上进行类似的试验，具体数据集的信息如表 1 所列。

表 1 数据集信息

数据集	属性数	数据个数	类别数
Iris	4	150	3
Wine	13	178	3
Glass	9	214	6

聚类之前各样本进行如下归一化处理：

$$x'_{ij}=\frac{x_{ij}-\min(x_i)}{\max(x_i)-\min(x_i)},i=1,2,\cdots,n;j=1,2,\cdots,p$$

式中，x_{ij} 表示归一化前第 i 个样本第 j 个属性的值，x'_{ij} 表示归一化之后的数据，$\min(x_i)$表示所有样本点第 i 个属性的最小值，$\max(x_i)$表示所有样本点第 i 个属性的最大值，归一化后样本点的各属性值分布在[0,1]之间。

将本文中所提出的算法(记为 AISS)与文献[11]中的 Constrained complete-link clustering algorithm (CCL)和文献[12]中的 MPCK-means algorithm 进行比较。CCL 算法是在层次聚类算法的基础上改进的，实验时让它停止的条件是聚类数目为真实类数。AISS 的近邻数 KN 取为整个数据集样本总数的 1/10。每次实验的约束对是从数据集中随机选取的，诊断正确率计算是取 30 次实验的平均值。3 种算法对 3 个数据集的聚类结果如图 8—图 10 所示。

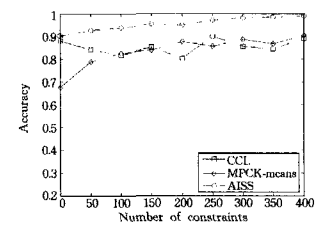

图 8 Iris 上 3 种方法聚类结果的比较

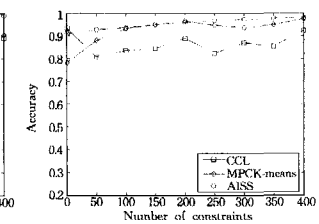

图 9 Wine 上 3 种方法聚类结果的比较

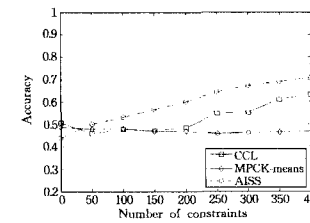

图 10 Glass 上 3 种方法聚类结果的比较

从图8—图10中可以看出，对于UCI上的3个数据集，本文提出的AISS算法的聚类准确性都是比较高的。聚类性能都可以随着约束对的增加而逐渐增强，CCL算法则有一定的波动性。对于IRIS数据集和Wine数据集，当约束对较少时，AISS算法的聚类正确率就可以达到90%，这可能是由于这两个数据集的各类样本点在较大程度上分布在某一流形上，使用流形距离度量时，能让同类中的样本点距离更近，这也符合大多数聚类样本的性质。

综合以上实验可以看到，本文所提出的算法对于人工数据集及真实数据集均有很好的聚类效果。

结束语 本文提出一种基于近邻流形距离的人工免疫半监督聚类算法，其将成对约束信息加入到样本度量矩阵中；通过近邻流形距离得到新的度量矩阵，将聚类问题转化为一个组合优化模型；采用能收敛到全局最优解的克隆选择算法求解模型，以获得最优的聚类划分；在人工数据集和标准数据集上的仿真实验结果表明，本文所提算法具有良好的聚类性能。

参 考 文 献

[1] Demiriz A, Benneit K P. Semi-Supervised clustering using genetic algorithm [C]// ANNIE'99. New York: ASME Press, 1999: 809-814

[2] Xing E P, Ng A Y, Jordan M. Distance metric learning, with application to clustering with side-information [C]// The 16th Annual Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2003: 505-512

[3] Bilenko M, Basu S, Mooney R J. Integrating Constraints and Metric Learning in Semi-Supervised Clus tering [C]// The 21st International Conference on Machine Learning. Banff, Canada, 2004: 81-88

[4] Basu S, Bilenko M, Mooney R J. A Probabilistic Framework for Semi-Supervised Clustering[C]// The 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2004: 59-68

[5] Yin X S, Chen S C, Hu E L, et al. Semi-supervised clustering with metric learning: An adaptive kernel method [J]. Pattern Recognition, 2010, 43(4): 1320-1333

[6] Mahdieh S B, Saeed B S. Kernel-based metric learning for semi-supervised clustering [J]. Neurocomputing, 2010, 73(7-9): 1352-1361

[7] Zhang H X, Lu J. Semi-supervised fuzzy clustering: A kernel-based approach [J]. Knowledge-based systems, 2009, 22(6): 477-481

[8] 冯晓磊，于洪涛. 基于流形距离的半监督近邻传播聚类算法[J]. 计算机应用研究，2011，28(10)：3656-3664

[9] 公茂果，焦李成，马文萍，等. 基于流形距离的人工免疫无监督分类与识别算法[J]. 自动化学报，2008，34(3)：367-375

[10] 焦李成，杜海峰，刘芳，等. 免疫优化计算、学习与识别[M]. 北京：科学出版社，2006：92-99

[11] Klein D, Kamvar S D, Manning C D. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering[C] // Proc of the 19th International Conference on Machine Learning. Sydney, 2002: 307-314

[12] Bilenko M, Basu S, Mooney R J. Integrating constraints and metric learning in semi-supervised clustering[C]//Proc of the 21st International Conference on Machine Learning. Banff: ACM Press, 2004: 81-88

(上接第196页)

[3] Rubio G, Pomares H, Rojas I. A heuristic method for parameter selection in LS_SVM: Application to time series prediction [J]. International Journal of Forecasting, 2011, 27: 725-739

[4] Fung G, Mangasarian O. Proximal support vector machines: knowledge discovery and data mining[C]// Proceedings of the Knowledge Discovery and Data Mining Conference. San Francisco, 2001: 77-86

[5] Lee Y J, Mangasarian O L. RSVM: Reduced support vector machines[C]// Proceedings of the First SIAM International Conference on Data Mining. Chicago, 2001: 5-7

[6] Glenn M, Fung O L. Mangasarian. Multicategory proximal support vector classifiers [J]. Machine Learning, 2005, 59: 77-97

[7] 陶晓燕，姬红兵，董淑福. 改进的PSVM及其在非平衡数据分类中的应用[J]. 计算机工程，2007，33(24)：191-193

[8] Zhu Zhen-feng, Zhu Xing-quan, Guo Yue-fei, et al. Inverse matrix-free incremental proximal support vector machine [J]. Decision Support Systems, 2012, 53(3): 395-405

[9] Jak S, Debrabanter J, Lukas L, et al. Weighted least squares support vector machines: robustness and sparse approximation [J]. Neurocomputing, 2002, 48(1-4): 85-105

[10] Dekruif B J, Devries T J A. Pruning error minimization in least squares support vector machines [J]. IEEE Transactions on Neural Networks, 2003, 14(3): 696-702

[11] Hoegaerts L, Suykens J A K, Vandewalle J, et al. A comparison of pruning algorithms for sparse support vector machines[C]// Proc. of the 11th International Conference on Neural Information Processing (ICONIP). New York: Springer Verlag, 2004: 1247-1253

[12] Renjifo C, Barsic D, Carmen C. Improving radial basis function kernel classification through incremental learning and automatic parameter selection [J]. Neurocomputing, 2008, 72: 3-14

[13] 王熙照，崔芳芳，鲁淑霞. 密度加权近似支持向量机[J]. 计算机科学，2012，39(1)：182-184

[14] Golub G H, Van Loan C C. Matrix Computations [M]. Baltimore, USA: The John Hopkins University Press, 1996

[15] Murphy M. UCI-Benchmark Repository of Artificial and Real Data Set [DB/OL]. http://www.ics.uci.edu/, 2005-04-01