第35卷 第22期 计 算 机 工 程 2009年11月
Vol.35 No.22 Computer Engineering November 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)22—0179—03 文献标识码：A 中图分类号：TP311

基于多示例的K-means聚类学习算法

谢红薇，李晓亮
(太原理工大学计算机与软件学院，太原 030024)

摘 要：多示例学习是继监督学习、非监督学习、强化学习后的又一机器学习框架。将多示例学习和非监督学习结合起来，在传统非监督聚类算法K-means的基础上提出MI_K-means算法，该算法利用混合Hausdorff距离作为相似测度来实现数据聚类。实验表明，该方法能够有效揭示多示例数据集的内在结构，与K-means算法相比具有更好的聚类效果。

关键词：多示例学习；K-means聚类；包间距；聚类有效性评价

K-means Clustering Learning Algorithm Based on Multi-instance

XIE Hong-wei, LI Xiao-liang
(College of Computer and Software, Taiyuan University of Technology, Taiyuan 030024)

【Abstract】Multi-instance learning is a new machine learning framework following supervised learning, unsupervised learning and reinforcement learning. Multi-instance learning and unsupervised learning are combined. This paper proposes a new multi-instance clustering algorithm MI_K-means based on traditional unsupervised learning algorithm K-means. The algorithm MI_K-means adopts mixed Hausdorff distance as similar measure to carry out clustering. Experimental shows that MI_K-means can effectively reveal inherent structure of a multi-instance data set, and it can get better clustering effect than K-means algorithm.

【Key words】multi-instance learning; K-means clustering; distance between bags; validity measure on clustering

1 概述

20世纪90年代中期，Dietterich等人[1]在对药物活性预测问题的研究中首先提出了多示例学习这个概念，其目的是判断药物分子是否为麝香分子(musky)。在多示例学习问题中，训练集不再是由若干示例组成，而是由一组含有概念标记的包(bag)组成，每个包是若干没有概念标记的示例集合。如果一个包中至少存在一个正例，则该包被标记为正包；如果一个包中不含有任何正例，则该包为反包。学习系统通过对已经标定类别的包进行学习来建立模型，希望尽可能正确地预测不曾遇到过的包的概念标记。

聚类分析是非监督学习的典型方法，它是将物理或抽象对象(示例)的集合分组成为由类似的对象组成的多个类的过程，其目的就是把数据对象按照相似度和连贯性(分布特征)分成若干类，使类间差异尽可能大，类内差异尽可能小。

由于一方面通常很难或很耗费成本去获得每个包的标记，另一方面非监督学习能够有效发现数据集的内在结构，因此有必要研究基于非监督的多示例学习。如果将传统的聚类分析的着眼点从示例的层次上升至包的层次，就可以将传统的非监督聚类学习算法改造为基于多示例的聚类学习算法，本文基于此思想，以包代替示例作为最小的数据单元，利用距离矩阵来度量2个包间的距离，然后结合传统的K-means算法将未标记的包划分成k个不相交包的集合，进而发现隐藏在多示例数据集中的规律或是结构。

2 多示例学习

与监督学习相比，多示例学习中的训练示例是没有概念标记的，这与监督学习中所有训练示例都有概念标记不同；与非监督学习相比，多示例学习中训练包是有概念标记的，这与非监督学习的训练样本中没有任何概念标记也不同。在以往的各种学习框架中，一个样本就是一个示例，即样本和示例是一一对应关系，而在多示例学习中，一个样本(即包)包含了多个示例，即样本和示例是一对多的对应关系。

学习算法需要生成一个分类器，能对未知的包(unseen bags)进行正确分类。多示例学习的问题描述见图1。学习算法的目标是要找出unknown process f(·)的最佳逼近方法。

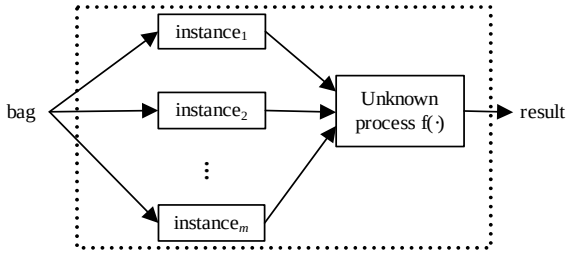

图1 多示例问题描述

Dietterich等人研究了标准的多示例预测问题(STDMIP)，在研究中一个包如果至少含有一个正示例，则被主动标记为正包，反之则被标记为反包。它的任务是从训练集中学习一些概念来正确标记未知的包。现在，除了STDMIP，Weidmann和Scott等分别提出了2种广义的多示例预测问题(GENMIP)，在该问题中，当包中的示例满足一些特定的限制而不是简单地以至少含有一个正示例为标准时，他们将此包视为正包，

基金项目：山西省自然科学基金资助项目(20051035)
作者简介：谢红薇(1962－)，女，教授，主研方向：人工智能，并行计算；李晓亮，硕士研究生
收稿日期：2009-06-20 **E-mail：**xiehongwei@tyut.edu.cn

反之认为是反包。另外，除了上面以离散值输出的多示例分类问题外，以实值输出的多示例回归问题也已经引起了很多研究者的关注[2]。

3 聚类分析方法

聚类也叫聚簇，是数据挖掘技术中重要的组成部分，它能够在潜在的数据中发现令人感兴趣的数据分布模式。事实上，聚类是一个无监督的分类，没有任何的先验知识可用，它是在无监督的情况下根据一定的相似性或距离计算函数自动地将数据集分成若干类。

聚类的用途是很广泛的。在商业上，它可以帮助市场分析人员从消费者数据库中区分出不同的消费群体来，并且概括出每一类消费者的消费模式或者说习惯；在生物学中，它可以被用来辅助研究动、植物的分类，可以用来分类具有相似功能的基因，还可以用来发现人群中的一些潜在的结构等。

没有任何一种聚类技术(聚类算法)可以普遍适用于揭示各种多维数据集所呈现出来的多种多样的结构。根据数据在聚类中的积聚规则以及应用这些规则的方法，有多种聚类算法：划分式聚类算法，层次聚类算法，基于密度和网格的聚类算法和其他聚类算法。不同的聚类方法有其不同的特点，为了减少复杂度，采用划分式聚类算法中简单高效的K-means方法。该方法的数学描述如下：

设 R^p 是 p 维实数集，$X=\{x_1, x_2,\cdots,x_n\}$是任一有限数据集，K-means 聚类算法的目标函数是 $J_m = \sum_{j=1}^{n}\sum_{i=1}^{k} u_{ij} d_{ij}^2$ ，其中，d_{ij} = $\|x_j - c_i\|$为第 j 个数据点到第 i 个聚类中心的距离；$c_1,c_2,\cdots,c_k$ 是 k 个聚类中心，$c_i \in R^p$, $u_{ij} \in \{0,1\}$，表示第 j 个数据属于(u_{ij} = 1)或不属于(u_{ij} =0)第 i 类。

该算法聚类步骤为：

(1)任意选定一组 k 个初始聚类中心 $c_1,c_2,\cdots,c_k$。

(2)将数据点 x_j(j=1,2,⋯,n)置于类 c_i(i=1,2,⋯,k)中，当且仅当 $\| x_j - c_i \| \leqslant \| x_j - c_p \|$ ，p=1,2,⋯,k 且 $i \neq p$ ，$\|\cdot\|$为相似度的距离度量，此处为欧式距离。

(3)用下式计算新的聚类中心：$c_i^* = \frac{1}{n_i}\sum_{x_j \in c_i} x_j$ ，i= 1,2,⋯,k，其中，n_i 为属于类 c_i 的数据点数。

(4)若 $c_i^* = c_i$ ，$\forall i$ =1,2,⋯,k，则停止，否则继续步骤(2)。

4 基于多示例的 K-means 聚类方法

聚类分析中的数据集是由一个个的示例组成，要通过计算示例之间的相似性来度量示例之间的紧密程度，对于多示例数据集来说，最小的数据单元是包，这就需要计算各个包之间的距离。

4.1 包间距离的测量

常用的包间距离测量有 2 种方法：最大 Hausdorff 距离和最小 Hausdorff 距离[3]，公式如下：

$$\max H(A,B) = \max\{\max_{a\in A}\min_{b\in B}\| a-b\|, \max_{b\in B}\min_{a\in A}\| b-a\|\}$$

$$\min H(A,B) = \min_{a\in A, b\in B}\| a-b\|$$

其中，给定 2 个包 $A=\{a_1,a_2,\cdots,a_m\}$，$B=\{b_1,b_2,\cdots,b_n\}$。由于最大 Hausdorff 距离对孤立点很敏感，而最小 Hausdorff 距离只考虑到包 A 和包 B 中距离最近的 2 个示例。鉴于从均衡性考虑，采用一种混合 Hausdorff 距离 $mixH(A,B)$，公式如下：

$$mixH(A,B) = \frac{\sum_{a\in A}\min_{b\in B}\| a-b\| + \sum_{b\in B}\min_{a\in A}\| b-a\|}{|A|+|B|}$$

其中，$|\cdot|$为测量集合的势。事实上 $mixH(A,B)$平均了一个包中示例与另一个包中距它最近的示例之间的的距离。从概念上来说，$mixH(A,B)$将包间示例的几何关系考虑了进来。

4.2 MI_K-means 聚类算法描述

算法描述如下：

输入：

U：没有标记的训练包集合$\{X_1,X_2,\cdots,X_n\}$

k：聚类的数目

输出：

G：聚类的结果(即每个包落入各个聚簇的情况)

Centers：每个聚簇的类中心

算法过程：

(1)从 U 中随机选择 k 个训练包作为初始的聚类中心 $C_1,C_2,\cdots,C_k$，将各个初始聚类中心作为一个聚簇集合 G_j，即 $G_j=\{C_j\}$，其中，j=1,2,⋯,k。

(2)将数据包 X_i(i=1,2,⋯,n)加入聚簇 G_j(j=1,2,⋯,k)中，当且仅当 $mixH(X_i,C_j) \leqslant mixH(X_i,C_p)$, p=1,2,⋯,k，且 $j \neq p$ 。

聚簇 G_j 在加入包 X_i 后形成新的聚簇 G_j^* ，即 $G_j^* = G_j \cup \{X_i\}$，然后令 $G_j = G_j^*$ 。

(3)用下式计算新的聚类中心

$$C_j^* = \frac{1}{n_j}\sum_{X_j \in G_j} X_j \text{ , } j=1,2,\cdots,k$$

其中，n_j 为聚簇 G_j 中包的数目。

(4)若 $\forall j$ =1,2,⋯,k, $C_j^* = C_j$，则停止，输出最终的结果，即 $G=\{G_j|\ 1 \leqslant j \leqslant k\}$，Centers=$\{C_j|\ 1 \leqslant j \leqslant k\}$，否则令 $C_j = C_j^*$，继续步骤(2)。

从表面上来看，MI_K-means 方法仅仅是对 K-means 算法在聚类对象上的扩展，然而，用多示例的思想来聚类包的任务有其自身的特点。为了聚类数据集中的对象，最直观的方法是让数据集中的每个示例在聚类过程中起到同等作用，但是在多示例聚类中，由于包中的示例往往起到不同的作用，因此只有少数的示例就决定了整个包的功用。这就使得尽管多示例聚类中没有包的标记，但这些包也不能被简单地认为是独立同分布的示例的集合，而应该仔细研究包中各个示例的特性及其之间的关系。

5 实验测试

5.1 评价标准

聚类有效性是判断聚类结构优劣的重要指标，不同的评价标准会得出不同的聚类结果。MI_K-means 方法将没有标记的包的集合$\{X_1,X_2,\cdots,X_n\}$分成 k 个不相交聚簇 G_j(j=1,2,⋯,k)，每个聚簇中包含一定数量的包。然而 K-means 方法划分数据是在示例的层次而不是在包的层次，这样的话，用 K-means 方法对多示例数据集进行划分时，包 X_i(i=1,2,⋯,k)就会被分割成若干部分，每个部分分别属于不同的聚簇。为了说明包的各部分对聚类的影响，给包 X_i 中的每一个示例分配一个权值 $\frac{1}{|X_i|}$，这样，小包中的示例就会比大包中的示例拥有更高的权值，而各个包的权值是相同的(均为 1)。文献[4]表明，这种权重策略对多实例学习是可行的。

对每个聚簇 G_j(j=1,2,⋯,k)，设 W_j^t ($t \in \{0,1\}$)表示聚簇 G_j 中从标记为 t 的包中来的示例的权重和。W_j 表示聚簇 G_j 中所有示例的权重和(即 $W_j = W_j^0 + W_j^1$)。这样，可以看出，$\sum_{j=1}^{k} W_j = n$ 。基于此提出了一个有效性评价标准，公式如下：

$$avgprecise(\{G_1,G_2,\cdots,G_k\})=\sum_{j=1}^{k}\frac{W_j}{n}\frac{\max\{W_j^0,W_j^1\}}{W_j}$$

其中，$avgprecise(\cdot)$为测度聚类结果的平均权重。聚簇 G_j 的准确度代表了在 G_j 中起决定作用的包的权重占总的包权重的比率，然后通过 $\frac{W_j}{n}$ 来求其平均权重。可以看出，这个平均标准的值越大，表示这种聚类方法的效果越好，当这个值达到 1 时，聚类的效果最好。

实验采用 2 个数据集 MUSK1 和 MUSK2[5]，数据集的数据构成见表 1。

表 1　麝香分子数据集

数据集	总包数	正包数	负包数	示例数
MUSK1	92	47	45	476
MUSK2	102	39	63	6 598

5.2　实验结果及其分析

分别利用 MI_K-means 方法和传统的 K-means 方法对多示例数据集 Musk1 和 Musk2 进行聚类，以平均准确度来衡量聚类的效果，得出 2 种方法在给定不同的聚类数目 k 时的聚类结果见表 2。

表 2　2 种方法的聚类效果

数据集	聚类方法	聚簇数目			
		k=5	k=20	k=35	k=50
Musk1	MI_K-means	0.638	0.817	0.858	0.914
	K-means	0.593	0.722	0.756	0.825
Musk2	MI_K-means	0.646	0.795	0.832	0.902
	K-means	0.613	0.687	0.714	0.738

表 2 显示了以平均准确率为标准的 2 种聚类方法随聚簇数目增加时的聚类效果，其中 MI_K-means 方法采用了上面提出的混合 Hausdorff 距离 $mixH(A,B)$来测量 2 个包之间的相似度。当聚簇数目固定时，实验重复了 5 次后将结果进行平均得出了最后的实验数据(图 2 和图 3 中的各个点)。

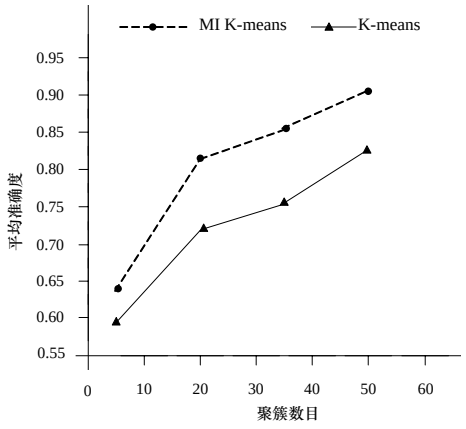

图 2　Musk1 上 2 种聚类方法的效果

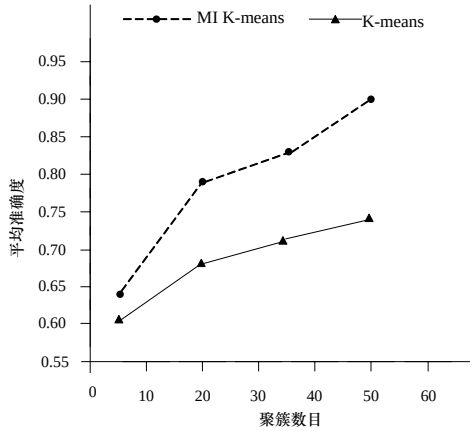

图 3　Musk2 上 2 种聚类方法的效果

从表 2 和图 2、图 3 可以看出，以平均准确率为评价标准时，不论是数据集 Musk1 还是 Musk2，MI_K-means 方法都比传统的 K-means 方法有更好的聚类效果。

6　结束语

本文将多示例学习和非监督学习结合起来，从训练有标记的示例转变成训练无标记的包，通过文中提出的混合 Hausdorff 距离 $mixH(A,B)$来测量 2 个包之间的距离，然后采用简单高效的 MI_K-means 聚类方法将无标记的包分成不相交的若干个包的集合，从而发现多示例数据集的内部结构，实验结果表明，将多示例学习和聚类结合起来是可行的，能够得到比较满意的聚类效果，但由于基于多示例的公开数据源比较少，本文采用最通用的 Musk 数据集，下一步的研究是将其应用到实际的数据集上，以达到实际应用多示例聚类的目的。

参考文献

[1] 蔡自兴，李枚毅. 多示例学习及其研究现状[J]. 控制与决策, 2004, 19(6): 607-610.

[2] 詹德川，周志华. 基于流形学习的多示例回归算法[J]. 计算机学报, 2006, 29(11): 1948-1955.

[3] 徐遵义，晏　磊. 基于 Hausdorff 距离的海底地形匹配算法仿真研究[J]. 计算机工程, 2007, 33(9): 7-9.

[4] Blockeel H, Page D, Srinivasan A. Multi-instance Tree Learning[C]//Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany: [s. n.], 2005.

[5] Blake C, Keohg E, Merz C L. UCI Respositroy of Machine Learning Databases[EB/OL]. (2008-10-20). http://www.ics.uci.edu/~mlearn/mlrepository.html.

编辑　索书志

(上接第 169 页)

4　结束语

阈值选择算法通过在不同模式下设定相应阈值，兼顾视频质量和运算复杂度选取最佳阈值，对最佳运动矢量块进行预测，以确定运动估计是否提前结束，在保证视频质量的前提下，极大地缩短了编码时间，提高了编码效率，对实时应用场合极为有利。

参考文献

[1] Wiegand T, Sullivan G J, Bjontegaard G, et al. Overview of the H.264/AVC Video Coding Standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2003, 13(7): 560-576.

[2] ITU-T. JVT-G050-2003 Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification[S]. 2003.

[3] Richardson I E G. H.264/MPEG-4 Part 10 White Paper[EB/OL]. (2002-10-15). http://www.vcodex.com/h264.html.

[4] Tourapis A M, Au O C, Liou M L. Highly Efficient Predictive Zonal Algorithms for Fast Block-matching Motion Estimation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2002, 12(10): 934-947.

[5] Wedi T. Motion Compensation in H.264/AVC[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2003, 13(7): 577-586.

编辑　任吉慧