基于EMD距离的多示例聚类

李 展[1] 彭进业[1,2] 温 超[1]
(西北大学信息科学与技术学院 西安 710069)[1] (西北工业大学电子信息学院 西安 710072)[2]

摘 要 多示例学习中,包由多个示例组成,有明确标记,而示例标记却不确定。已有聚类研究都针对单示例、单标记,因而无法直接应用于多示例问题。基于推土机距离(earth mover's distance,EMD)提出了一种新的多示例聚类算法ECMIL。该方法首先利用欧式距离计算包内示例相似度,将相似示例合并;然后将需要度量距离相似性的包内示例分别看作供货者和消费者,计算货物拥有量和货物需求量;对推土机距离无法供货问题,通过增大满足条件供货者的权值加以解决;最后使用k-medoids算法进行聚类。在基准数据集MUSK,Corel和SIVAL上进行实验,表明ECMIL算法是有效的。

关键词 多示例聚类,推土机距离,k-medoids

Multi-instance Clustering Based on EMD

LI Zhan[1] PENG Jin-ye[1,2] WHEN Chao[1]
(School of Information Science and Technology, Northwest University, Xi'an 710069, China)[1]
(School of Electronics Information, Northwestern Polytechnical University, Xi'an 710072, China)[2]

Abstract In the setting of multi-instance learning, each sample is represented by a bag composed of multiple instances. Previous studies on clustering mainly deal with the single instance in traditional learning setting, so it can't be applied to multi-instance problem directly. In this paper, based on earth mover's distance, a novel multiple-instance clustering algothrim named ECMKIL was presented. Firstly we calculated the bag's instances' similarity, emerged the similarity ones, then regarded the two bags' instances as suppliers and consumers, calculated the goods and capacity. To deal with the supplier-consumer imbalance problem, we solved it by multiplying the goods. Finally, used k-medoids to cluster the multi-instance data. Experimental results on MUSK, Corel and SIVAL data set indicate that the ECMKIL method is effective.

Keywords Multi-instance clustering, Earth mover's distance, K-medoids

1 引言

Dietterich[1]在1997年研究药物活性预测工作中,第一次提出了多示例学习(Multiple Instance Learning,MIL)的概念。研究中,他们发现分子存在多种低能形状,而药物实验数据只能辨别分子参与制药能力,无法确定其具体形状。针对此问题,Dietterich将分子看成包(bag),分子的低能形状作为包中的一个示例(instance),由此提出了MIL学习问题。在该问题中,包具有明确的标记,而示例标记则依赖于包标记。如果一个包的标记是负(negative),则它的所有示例标记都为负。如果包的标记为正(positive),则包中必然存在正示例,但是不能确定是哪个或哪些。

由于MIL问题具有独特的性质,适合表达许多应用领域的潜在结构,因此MIL框架已应用到图像分类[2]、图像检索[3]、文本分类[4]、计算机安全[5]、人脸识别[6]、医学辅助诊断[7]等多个领域。

聚类是一种无监督学习方法,它从数据对象的角度出发,通过计算它们之间的相似程度(距离)或数据对象生成方式(密度),形成合理的对象簇[8]。在无监督情况下,聚类MIL数据以发现其内部潜在结构引发了越来越多研究者的兴趣[9-12]。但由于MIL包的集合性特点,因此无法直接使用已有的聚类算法。

本文以推土机距离(Earth Mover's Distance,EMD)为基础,提出了一种新的MIL聚类算法,称之为ECMIL(Multiple Instance Learning Clustering based on EMD)。该算法首先计算包内示例相似度,将相似示例合并;然后针对EMD无法供货问题,增大特定满足条件供货者的权值;最后使用k-medoids进行聚类。在基准测试集MUSK,Corel和SIVAL上实验,验证了算法的有效性。

2 相关工作

自提出MIL问题后,近10年很多科研人员对此进行了

到稿日期:2010-08-26 返修日期:2010-12-15 本文受教育部新世纪优秀人才支持计划项目(NCET-07-0693),陕西省教育厅科研项目(10JK852)资助。

李 展(1973—),男,博士,讲师,主要研究方向为机器学习、图像检索,E-mail:lizhan@nwu.edu.cn;彭进业(1964—),男,教授,博士生导师,主要研究方向为图像处理。

研究。

针对 MIL 分类问题，一种方法是是基于示例的空间聚集性假设，利用轴平行矩形（axis-parallel rectangles，APR）算法[1]，或定义多样性密度函数（diverse density，DD）[13]；一种方法是利用决策树或神经网络模型，重新定义熵、覆盖函数[14]，或神经网络的误差函数[15]；一种方法使用修正的 Hausdorff 距离，利用 K 近邻（k-nearest neighbor，KNN）算法[16]；研究最多的是基于支持向量机（Support Vector Machine，SVM）的方法，或修改 SVM 目标函数[17,18]，或定义针对 MIL 包特点的核函数[19,20]，或将包非线性投影并嵌入到构造空间中[21]。最近，出现了利用高斯模型[22]和条件随机场方法[23]研究 MIL 问题的趋势。

在 2010 年，Sun 和 Ren 首次研究了 MIL 降维问题。Sun 定义降维目标优化函数，并对其进行连续变形，最后利用梯度下降法求解[24]；Ren 则从 MIL 样本非独立同分布出发，构造包级图结构，然后基于类间和类内离散度矩阵定义优化目标函数，从而解决 MIL 降维问题[25]。

Kriegel 于 2006 年首次开始研究 MIL 聚类问题，他使用高斯模型描述示例集合，用高斯模型的多项式分布来描述包的聚类情况，基于期望最大化（Expectation Maxi-mization，EM），利用 MI-EM 算法聚类[9]。Lu 提出了两种 MIL 聚类的框架和 6 种算法（pn-sum，mi-max，mi-mean，mi-fcm，mi-gmm，mi-svm），一种针对 Kriegel 的模型，采用了与模拟退火相结合的模糊 C-均值算法；一种从包质心出发启发，迭代优化进行聚类[10]。Zhang 在最大间隔聚类模型基础上，针对 MIL 聚类问题给出了 M3IC 优化问题和相应算法 M3IC-MBM。对 M3IC 问题连续条件放松变形，最后利用 CCCP（Constrained Concave-Convex Procedure）求解[11]。Zhang 使用平均 Hausdorff 距离度量包间距离，在此基础上使用 k-medoids 聚类，提出了 BAMIC 聚类算法[12]。

本文第 3 节首先给出 MIL 聚类问题和 EMD 距离定义，然后针对 MIL 聚类问题，修改 EMD 距离，并提出了 ECMIL 聚类算法；第 4 节在基准测试集 MUSK，Corel 和 SIVAL 上实验，将本文方法和其它方法进行对比，并给出实验结果；最后总结了本文的工作。

3 基于 EMD 距离的 MIL 聚类方法

3.1 问题描述

假定 B 为 MIL 问题内所有包组成的集合，$B=\{B_i, i=1,\cdots,n\}$，n 为 MIL 中包总数。包 B_i 内示例表为 $B_i=\{x_{ij}, j=1,\cdots,m\}$，$m$ 为包 B_i 内示例总数，且有 $x_{ij}\in R^d$。MIL 聚类就是将集合 B 分成 k 个不相交的子集合 $B=\{C_1, C_2, \cdots, C_k\}$，有 $C_i \bigcap_{i\neq j} C_j=\phi$，$B=\bigcup_{i=1\sim k} C_k$。

受 BAMIC 聚类算法的启示，本文方法也从寻找适合包的距离度量方法出发，然后利用 k-medoids 算法聚类。相对于其它方法，本方法充分考虑了包内元素的集合性，没有复杂的数学模型，计算简单，因此具有很好的实际应用价值。

3.2 EMD 距离

BAMIC 算法基本思想是寻找合适的包距离度量方法，并应用 k-medoids 算法进行多示例聚类。设 $A=\{a_1, a_2, \cdots, a_n\}$ 和 $B=\{b_1, b_2, \cdots, b_m\}$ 是两个包，$a_i\in R^d$，$i=1,\cdots,n$，$b_j\in R^d$，$j=1,\cdots,m$ 为包中的示例。

最大和最小 Hausdorff 距离公式为式(1)和式(2)，其中 $\|a-b\|$ 为示例 a 和 b 的距离。

$$\max H(A,B)=\max\{\max_{a\in A}\min_{b\in B}\|a-b\|, \max_{b\in B}\min_{a\in A}\|b-a\|\} \tag{1}$$

$$\min H(A,B)=\min_{a\in A, b\in B}\|a-b\| \tag{2}$$

由于最大 Hausdorff 距离对噪声比较敏感，因此 Zhang 对最小 Hausdorff 距离修正，并定义式(3)为平均 Hausdorff 距离。$|\cdot|$ 表示集合的势。

$$aveH(A,B)=\frac{\sum_{a\in A}\min_{b\in B}\|a-b\|+\sum_{b\in B}\min_{a\in A}\|b-a\|}{|A|+|B|} \tag{3}$$

$aveH(\cdot,\cdot)$ 计算一个包中每个示例到另一个包的最近示例的求和平均。平均 Hausdorff 距离取决于两个包在特征空间位置最近的示例距离，因此并不能正确反映包的整体距离。如对图像进行分割，则将图像与分割区域分别看作包（Bag）和示例（Instance），图像的聚类问题就变成了 MIL 聚类问题。如图 1 所示，$aveH(A,B)=2.65$，$aveH(A,C)=1.34$。虽然图像 A，C 属于不同场景，但它们的 $aveH$ 小于相同场景图像 A，B 之间的 $aveH$，这是因为它们都包含非常相似的 sky 和 mountain 区域，平均 Hausdorff 距离主要受这个区域的影响。

(A) Beach 1　(B) Beach 2　(C) Mountain

图 1　图像示例

EMD 距离用于度量集合相似性，能实现包内示例的多对多最佳匹配，更适合表示多示例聚类的集合整体相似性要求。

EMD 距离源自运输问题。假设某个空间中存在着若干已知质量的土堆以及若干已知容量的土坑，运输问题需要解决从土堆到土坑总搬运耗费最小问题。假定 $P=\{(p_1, w_{p1}), \cdots, (p_m, w_{pm})\}$，其中 $p_i\in R^d$ 为土堆所在位置，w_{pi} 为土堆质量；$Q=\{(q_1, w_{q1}), \cdots, (q_n, w_{qn})\}$，其中 $q_i\in R^d$ 为土坑所在位置，w_{qi} 为土坑大小，则 EMD 距离就为如下线性优化问题。

$$\begin{cases} \min\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij} \\ \text{s. t. } f_{ij}\geqslant 0 \\ \sum_{j=1}^{n} f_{ij}\leqslant w_{pi}, 1\leqslant i\leqslant m \\ \sum_{i=1}^{m} f_{ij}\leqslant w_{qj}, 1\leqslant j\leqslant n \\ \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}=\min(\sum_{i=1}^{m} w_{pi}, \sum_{j=1}^{n} w_{qi}) \\ d_{ij}=\sqrt{\sum(p_{ik}-q_{ik})^2} \end{cases} \tag{4}$$

若将 p_i 当作供货者，w_i 为其货物拥有量；q_j 当作消费者，w_j 为其对货物的需求量；d_{ij} 为从供货者 p_i 到消费者 q_j 运输单位货物的所需的运输成本。式(4)就是寻找满足条件的最优流量 $F=[f_{ij}]$，使总体运输成本 $\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}$ 达到最小，其中 f_{ij} 表示 p_i 到 q_j 的搬运流量。EMD 规格化定义如下。

$$EMD(P,Q)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \tag{5}$$

3.3 改进 EMD 距离

针对 MIL 聚类问题，将两个包内的示例，一个看作供货

者，另外一个看作消费者，假定两包分别为 $B_i=\{x_{il},l=1,\cdots,m\}$ 和 $B_j=\{x_{jk},k=1,\cdots,n\}$；$x_{il}\in R^d$，$x_{jk}\in R^d$。两个包的EMD距离受权值 w_l，w_k 和运输成本 d_{lk} 的影响。d_{lk} 度量包间示例的距离相似性，使用式(6)进行计算。

$$d_{lk}=\sqrt{\sum_{t=1}^{d}(x_{il}-x_{jkt})^2} \tag{6}$$

已有的基准测试集 MUSK，Corel 和 SIVAL，都需要定义 w_l 和 w_k。其中，MUSK 是药物活性预测时获取的数据，而 Corel 和 SIVAL 分别用于图像分类和检索。受 Rubner[26] 工作的启发，我们对 Corel 和 SIVAL 图像数据使用“分割区域的像素数与整个图像的总像素数的比值”作为相应区域的权值。而针对 Musk 数据库，则采用示例到其质心的距离计算，其计算公式如下。

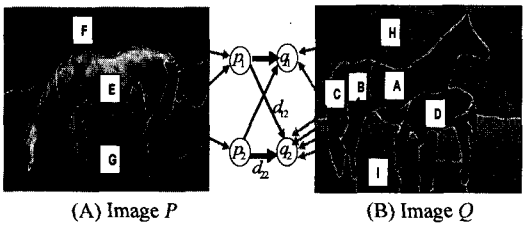
$$O_i=\frac{\sum_{j=1}^{|B_i|}x_{ij}}{|B_i|} \tag{7}$$

$$w_l=\frac{\|x_{il}-O_i\|_2}{\sum_{j=1}^{|B_i|}\|x_{ij}-O_i\|_2} \tag{8}$$

式中，O_i 为包 B_i 的质心，w_l 为示例 x_{il} 权重。

直接将 EMD 距离用于 MIL 图像聚类，存在如下 3 个问题：1)存在图像过分割问题；2)MIL 的正包中，可能存在多个正例；3)存在供货者无法供货问题。以上 3 个问题都会导致 EMD 距离不能正确反映集合相似距离，从而影响 MIL 聚类性能。

如图 2 所示，两图像 P，Q 经分割后，均包含相同的图像区域，“马与草地”显然属于相同的图像类型，希望 EMD(P，Q)应尽可能地小。图像 Q 中的马过分割生成了 A，B，C 3 部分，且图像 Q 中存在两个正示例马，这些部分很可能从图像 P 中的草地 F 和 G 运输，从而导致 EMD 变大。图像 P 和图像 Q 中的马所占区域面积分别为 0.48 和 0.63，从而出现供货者无法满足消费者的情况，这样图像 Q 中马除了从 d_{22} 运输，还会从 d_{12} 运输，从而也出现 EMD 变大的情况。



(A) Image P　　(B) Image Q

图 2　图像 EMD 距离示例

针对问题 1 和 2，计算 MIL 中包 B_i 的示例相似性，合并小于某个阈值 δ 的示例，其过程如下。首先用式(9)计算示例相似性。

$$\mathrm{sim}(x_{ik},x_{il})=\|x_{ik}-x_{il}\|_2 \tag{9}$$

如果示例相似性计算结果 $\mathrm{sim}(x_{ik},x_{il})<\delta$，则将相应示例合并成一个新示例。其合并示例计算公式如式(10)和式(11)，l 表示 B_i 中和 x_{ik} 相似的示例个数。

$$x_{im}=\frac{\sum_{t=1,x_{it}\in \mathrm{sim}(x_{ik},x_{il})<\delta}^{|B_i|}x_{it}}{l} \tag{10}$$

$$w_{im}=\sum_{t=1,x_{it}\in \mathrm{sim}(x_{ik},x_{il})<\delta}^{|B_i|}w_t \tag{11}$$

针对问题 3，通过设置相似阈值 β，对满足相似条件的供货者增大权值，使运输尽可能多地发生在运输成本低的供求者之间。以上问题具体解决方法如算法 1 所示。

算法 1　改进的 EMD 距离

输入：图像包 $P=\{(p_1,w_1),\cdots,(p_n,w_n)\}$，$Q=\{(q_1,w_1),\cdots,(q_m,w_m)\}$，相似阈值 δ 和 β；

输出：图像包 P 和 Q 之间的 EMD 距离，即 EMD(P，Q)；

Step1　使用式(9)，计算图像包 P(供货者)和图像包 Q(消费者)每个区域对应的示例间的相似性，将小于阈值 δ 的示例按式(10)、式(11)进行合并；

Step2　图像包 P(供货者)每个区域对应的权值进行如下调整：对 $\forall p_i\in P$，若 $\exists q_j\in Q$，满足 $d_{ij}<\beta$，则 $w_i^*=\gamma\cdot w_i$(其中 γ 为调整系数)；否则权值不变，即 $w_i^*=w_i$；

Step3　求解式(4)的线性优化问题，用式(5)计算 EMD(P，Q)。

示例合并和增大权值过程，设置阈值 δ，其目的是确保一个图像内确实存在足够的相似区域，才进行合并；设置阈值 β，确保两幅不同图像中确实存在相同的区域(或者说是足够相似的区域)时，才对第一个图像相应区域权值放大。实验中调整系数 γ 设置为 5，原因是当两个区域的相对面积之差大于 5 倍时，这样的差异在 EMD 距离中应得到体现，而不能完全被忽略。

相似阈值 δ 和 β 的确定：在图像库中，选择与示例 x_k 最近的 10 个邻居示例 x_i，$i=1,\cdots,10$，取其平均值作为 δ，$\delta=\frac{\sum_{i=1}^{10}x_i}{10}$。由于相似阈值 δ 就代表了与示例 x_k 运输成本低的供求者，因此令 $\beta=\delta$。

3.4　ECMIL 算法步骤

最后，本文提出的 ECMIL 多示例学习算法具体步骤如下。

算法 2　ECMIL 算法

输入：MIL 聚类数据集合 $B=\{B_i,i=1,\cdots,n\}$，聚类个数 K；

输出：MIL 聚类结果集合 $I=\{I_i,i=1,\cdots,n\}$；

初始化：$I=\varnothing$；

Step1　从 n 个包中任意选取 K 个对象作为初始的聚类中心；

Step2　利用算法 1 计算每个包到各个聚类中心的距离，把包分配到距离最近的聚类中，生成集合 I；

Step3　所有包完成分配后，利用 k-medoids 算法计算 k 个聚类的中心；

Step4　与前一次计算得到的 k 个聚类中心比较，如果聚类中心发生变化，转 Step2，否则 Step5；

Step5　输出聚类结果 I。

4　实验结果与分析

为了验证 MCMIL 算法的有效性，采用 MIL 的基准数据集 MUSK，Corel 和 SIVAL 进行对比实验。本文采用与 Kriegel 等人相同的指标来评价聚类算法的性能，分别称为准确度(precision)与平均熵(average entropy)。准确度衡量聚类结果与手工标注的类别标记之间的符合程度，平均熵表示了类别中含有噪声数据程度，其具体公式参阅文献[9]。

4.1　MUSK 数据集

MUSK 数据集是研究药物活性时提出的，它由 Musk1 和 Musk2 组成，来自麝香分子样本。一个包表示一个分子，包中的样例对应着该分子的一个低能形态，每个形态对应一个 166 维的向量。Musk1 包括 47 个正包、45 个负包，其中正示例数为 207、负示例数为 269。每个包中示例数从 2 到 40 不等。Musk2 包含 39 个正包、63 个负包，其中正示例数为

1017、负示例数为 5581。包中示例数从 1 到 1044 不等。

为了试验公平性，采取和文献[10]同样的方法进行实验。由于 MUSK 数据集由正、负包组成，其正包集合具有潜在概念，而负包集合则属于背景噪声，因此采用聚类个数 $K=2$。图 3、图 4 分别为 $K=2$ 情况下的聚类结果。可见，ECMIL 在 MUSK 上的准确度性能远远高于 BAMIC 和 ori-svm，性能和 mi-svm 相当。平均熵的实验结果与此类似，ECMIL 的平均熵远低于 BAMIC 和 ori-svm 算法，表明 ECMIL 聚类的结果中噪声数据较少。其原因在于：由于采用了 EMD 距离，它采用包内示例的多对多最佳匹配，可以更好地度量 MUSK 这种一个潜在概念的聚类情况，具有更好的鲁棒性，因而取得了和 mi-svm(one-class SVM)相当的性能。

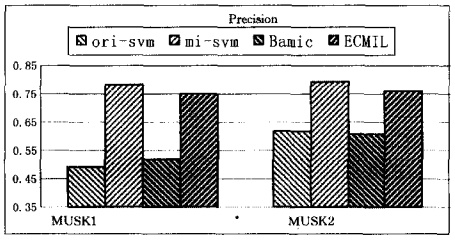


图 3　MUSK 数据集上准确度比较

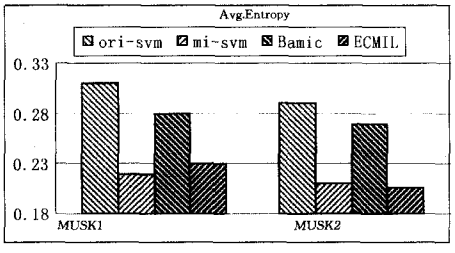


图 4　MUSK 数据集上平均熵比较

4.2　Corel 和 SIVAL 数据集

由于聚类在于发现数据中的潜在模式，而负包则通常由背景图像组成，通常不含有潜在模式。因此采取和 Deng 相同的方法，构造两个 MIL 聚类数据集。将 Corel 中的 3 类图像 elephant、fox 和 tiger 中的正包合并，忽略负包，构造数据集 1。数据集 2 采用 SIVAL 数据集，图像集由 1500 幅图像组成，包含 25 类不同物体，每类 60 副图像。随机将 25 类分成 5 组，每组中包括 5 类图像。将 5 组命名为 SIVAL1，SIVAL2，SIVAL3，SIVAL4 和 SIVAL5。具体细节请参阅文献[11]。实验结果如表 1 所列。

表 1　Corel 和 SIVAL 数据集聚类准确度比较

	ECMIL	M3IC-MBM	BAMIC
Corel	55.8%	53.6%	37.8%
SIVAL1	47.8%	46.1%	39.6%
SIVAL2	44.8%	42.2%	40.5%
SIVAL3	43.6%	41.3%	39.8%
SIVAL4	41.3%	38.6%	33.9%
SIVAL5	42.2%	40.3%	36.2%

如表 1 所列，ECMIL 方法性能最好，相对于最大间隔分类方法 M3IC-MBM 性能提高了 2%左右，而比同类的算法 BAMIC 性能提高了 6%以上。其主要原因在于 EMD 距离是一种多对多的集合间距离相似性计算方法，而不像 BAMIC 采取一对一方法。由于采用了包内相似示例合并以及针对无法供货问题的权重调节方法，它能够更好地度量集合间距离，从而大大提高了聚类性能。

结束语　针对 MIL 的聚类问题研究，本文主要的创新点在于提出了一种结合 EMD 距离和 k-medoids 的 ECMIL 聚类算法，为求解 MIL 聚类问题提供了一种新思路。相对于 BAMIC 算法来说，EMD 采用多对多的集合度量距离，可以更好地获取集合的相似距离，从而性能更优。基于多个 MIL 基准数据集 MUSK，Corel 和 SIVAL 的试验表明，ECMIL 相对其它 MIL 聚类算法，具有更高的聚类准确度，而且对于一个潜在概念的 MUSK 聚类也取得了不错的性能，表明其具有很好的鲁棒性。

本文的不足在于，EMD 作为一种基于线性优化的距离计算方法，其时间复杂度为多项式时间，其算法效率并不是很好。因此，如何提高 ECMIL 的算法效率，是一个值得进一步研究的课题。

参 考 文 献

[1] Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles[J]. Artificial Intelligence, 1997, 89(12): 31-71

[2] Chen Y, Bi J, Wang J Z. MILES: Multiple-instance learning via embedded instance selection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28: 1931-1947

[3] Zhang Q, Yu W, Goldman S A, et al. Content-based image retrieval using multiple-instance learning[C]// Proceeding of 19th Intl. Conf. Mach. Learn. 2002: 682-689

[4] Settles B, Craven M, Ray S. Multiple instance active learning[C]// Proceeding of Adv. Neural Inf. Process. Syst. 2008: 1289-1296

[5] Ruffo G. Learning single and multiple instance decision trees for computer security applications Doctoral dissertation[D]. Torino, Italy: CS Dept., Univ. Turin, 2000

[6] Zhang C, Viola P. Multiple-instance pruning for learning efficient cascade detectors[C]// Proceeding of Adv. Neural Inf. Process. Syst. 2008: 1681-1688

[7] Fung G, Dundar M, Krishnappuram B, et al. Multiple instance learning for computer aided diagnosis[C]// Proceeding of Adv. Neural Inf. Process. Syst. 2007: 425-432

[8] 张霞，王素贞，尹怡欣，等. 基于模糊粒度计算的 K-means 文本聚类算法研究[J]. 计算机科学，2010，37(2)：209-211

[9] 路晶，马少平. 基于多例学习的 Web 图像聚类[J]. 计算机研究与发展，2009，46(9)：1462-1470

[10] Lu Jing, Ma Shao-ping. Web Image Clustering Based-on Multiple Instance[J]. Journal of Computer Research and Development, 2009, 46(9), 1462-1470

[11] Zhang M-L, Zhou Z-H. Multi-instance clustering with applications to multi-instance prediction[J]. Applied Intelligence, 2009, 31(1): 47-68

[12] Zhang Dan, Wang Fei, Si Luo, et al. M3IC: Maximum Margin Multiple Instance Clustering[C]// Proceeding of the Twenty-first International Joint Conference on Artificial Intelligence. 2009: 1339-1344

[13] Maron O, Lozano-Pérez T. A framework for multiple-instance learning[C]// Proceeding of Adv. Neural Inf. Process. Syst., 1998: 570-576

[14] Chevaleyre Y, Zucker J-D. Solving multiple-instance and multiple-part learning problems with decision trees and decision rules. Application to the mutagenesis problem[C]// Proceeding

of the 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence. 2001:204-214

[15] Zhang M-L, Zhou Z-H. Adapting RBF Neural Networks to Multi-Instance Learning[J]. Neural Processing Letters, 2006, 23(1):1-26

[16] Wang Jun, Zucker J-D. Solving the multiple-instance problem: a lazy learning approach[C]// Proceeding of the 17th Intl. Conf. Mach. Learn. 2000:1119-1125

[17] Stuart A, Thomas H, Ioannis T. Multiple instance learning with generalized support vector machines [C] // Proceeding of the 18th National Conference on Artificial Intelligence. 2002:943-944

[18] Zhou Zhi-hua, Xu Jun-ming. On the relation between multi-instance On the relation between multi-instance learning and semi-supervised learning [C] // Proceeding of the 24th Intl. Conf. Mach. Learn. 2007:1167-1174

[19] Kwok J T, Cheung P-M. Marginalized multi-instance kernels[C]// Proceeding of the 20th International Joint Conference on Artificial Intelligence. Hydrabad, India, 2007:901-906

[20] Zhou Z-H, Sun Y-Y, Li Y-F. Multi-instance learning by treating instances as non-I. I. D. samples[C]// Proceeding of the the 26th Intl. Conf. Mach. Learn. 2009:1249-1256

[21] Chen Yi-xin, Wang J Z. Image categorization by learning and reasoning with regions [J]. Journal of Machine Learning Research, 2004, 5(8):913-939

[22] Kim M, De la Torre F. Gaussian Processes Multiple Instance Learning[C]// Proceeding of the 27th Intl. Conf. Mach. Learn. 2010:535-542

[23] Deselaers T, Ferrari V. A Conditional Random Field for Multiple-Instance Learning[C]// Proceeding of the 27th Intl. Conf. Mach. Learn. 2010:287-296

[24] Wei Ping, Ye Xu, Non-I. I. D. Multi-Instance Dimensionality Reduction by Learning a Maximum Bag Margin Subspace[C]// Proceeding of 24th AAAI Conference on Artificial Intelligence. 2010:551-556

[25] Sun Yu-yin, Ng M K, Zhou Zhi-hua. Multi-Instance Dimensionality Reduction[C] // Proceeding of 24th AAAI Conference on Artificial Intelligence. 2010:587-592

[26] Rubner Y, Tomas C, Guibas L J. The earthmover's distance as a metric for image retrieval [J]. International Journal of Computer Vision, 2000, 40(2):99-121

(上接第 193 页)

粒子群算法在不同粒子数目时，得到的极值不同，表明算法的求解过程存在着差异，而多样性的观测要能够正确地表征这种信息。

采用不同的粒子个数，f_7 函数的最优值处在同样的数量级，50 个粒子的情况稍好于 100 个粒子，反映到多样性的变化上，如图 4(a) 所示。算法在不同粒子个数上，认知和速度多样性变化趋势基本相同，50 个粒子时多样性的最终数值略小于 100 个粒子的情况。而对于函数，100 个粒子时的最优值稍小于 50 个粒子的情况，对应到多样性的变化上，如图 4(b)、(c) 所示，100 个粒子时位置和速度的多样性收敛也稍早于 50 个粒子。

通过以上讨论可以得出，新的基于 L_1 范式的群体多样性观测可以正确地反映粒子群优化在不同粒子数目上的运行信息。

结束语 在本文中，分析了不同的群体多样性定义方式的优缺点，提出了一种新的观测方式；并通过实验观测粒子群算法在不同解空间维数、不同粒子群拓扑结构和不同种群大小上的多样性变化情况，验证了新的群体多样性观测表征粒子群算法运行状态的准确性和新的观测方式的通用性。新的观测方式减少了已有方法的计算量，并提供了变化范围更广泛的观测值，从而可以更清晰地观测粒子群算法执行过程中粒子的运动。

通过多样性的观测，分析粒子群体在算法执行过程中位置、速度和认知变化信息，进而找出有效地间接或者直接的控制多样性的方法，就可以动态地改变粒子群的“探索”或“开发”的阶段和能力，从而取得更好的优化效果。

参 考 文 献

[1] Eberhart R, Kennedy J. A new optimizer using particle swarm theory [C] // Proc of the Sixth International Symposium on Micro Machine and Human Science. 1995:39-43

[2] Kennedy J, Eberhart R. Particle swarm optimization [C]// Proc of IEEE International Conference on Neural Networks(ICNN). 1995:1942-1948

[3] Eberhart R, Kennedy J, Shi Yu-hui. Swarm Intelligence [M]. Morgan Kaufmann Publisher, 2001

[4] Eberhart R, Shi Yu-Hui. Computational Intelligence: Concepts to Implementations[M]. Morgan Kaufmann Publisher, 2007

[5] Engelbrecht A. Foundamentals of Computational Swarm Intelligence [M]. John Wiley & Sons Ltd, 2005

[6] 谭营. 计算群体智能基础 [M]. 北京：清华大学出版社，2009

[7] 张军，詹志辉，等. 计算智能 [M]. 北京：清华大学出版社，2009

[8] Zhan Zhi-hui, Zhang Jun, Li Yun, et al. Adaptive particle swarm optimization [J]. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 2009, 139(6):1362-1381

[9] Golub G, Van Loan C. Matrix Computations(third edition)[M]. The Johns Hopkins University Press, 1996

[10] Shi Yu-hui, Eberhart R. Population diversity of particle swarms [C]// Proc of the 2008 Congress on Evolutionary Computation. 2008:1063-1067

[11] Shi Yu-hui, Eberhart R. Monitoring of particle swarm optimization [J]. Frontiers of Computer Science, 2009, 3(1):31-37

[12] Zhan Zhi-hui, Zhang Jun, Shi Yu-hui. Experimental study on PSO diversity [C]// Third International Workshop on Advanced Computational Intelligence. Jiangsu, china, August 2010:310-317

[13] Yao Xin, Liu Yong, Lin Guang-ming. Evolutionary programming made faster [J]. IEEE Transactions on Evolutionary Computation, 1999, 3(2):82-102

[14] Shang Yun-wei, Qiu Yu-huang. A note on the extended rosenbrock function [J]. Evolutionary Computation, 2006, 14(1):119-126

[15] Bratton D, Kennedy J. Defining a standard for particle swarm optimization [C]// Proc of the 2007 IEEE Swarm Intelligence Symposium(SIS 2007). 2007:120-127

[16] Spears W, Green D, Spears D. Biases in particle swarm optimization [J]. International Journal of Swarm Intelligence Research, 2010, 1(2):34-57