ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.3, March 2008, pp.663−673
DOI: 10.3724/SP.J.1001.2008.00663

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于 Tri-Training 和数据剪辑的半监督聚类算法*

邓 超+, 郭茂祖

(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

Tri-Training and Data Editing Based Semi-Supervised Clustering Algorithm

DENG Chao+, GUO Mao-Zu

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: Phn: +86-451-86402407, E-mail: dengchao@hit.edu.cn, http://nclab.hit.edu.cn/dc.html

Deng C, Guo MZ. Tri-Training and data editing based semi-supervised clustering algorithm. *Journal of Software*, 2008,19(3):663−673. http://www.jos.org.cn/1000-9825/19/663.htm

Abstract: In this paper, a algorithm named DE-Tri-training semi-supervised K-means is proposed, which could get a seeds set of larger scale and less noise. In detail, prior to using the seeds set to initialize cluster centroids, the training process of a semi-supervised classification approach named Tri-training is used to label unlabeled data and add them into the initial seeds set to enlarge the scale. Meanwhile, to improve the quality of the enlarged seeds set, a nearest neighbor rule based data editing technique named Depuration is introduced into Tri-training process to eliminate and correct the mislabeled noise data in the enlarged seeds. Experimental results show that the novel semi-supervised clustering algorithm could effectively improve the cluster centroids initialization and enhance clustering performance.

Key words: semi-supervised clustering; semi-supervised classification; K-means; seeds set; Tri-training; depuration data editing

摘 要: 提出一种半监督聚类算法,该算法在用 seeds 集初始化聚类中心前,利用半监督分类方法 Tri-training 的迭代训练过程对无标记数据进行标记,并加入 seeds 集以扩大规模;同时,在 Tri-training 训练过程中结合基于最近邻规则的 Depuration 数据剪辑技术对 seeds 集扩大过程中产生的误标记噪声数据进行修正、净化,以提高 seeds 集质量.实验结果表明,所提出的基于 Tri-training 和数据剪辑的 DE-Tri-training 半监督聚类新算法能够有效改善 seeds 集对聚类中心的初始化效果,提高聚类性能.

关键词: 半监督聚类;半监督分类;K-均值;seeds 集;Tri-Training;Depuration 数据剪辑

中图法分类号: TP18 **文献标识码**: A

* Supported by the National Natural Science Foundation of China under Grant Nos.60702033, 60772076 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2007AA01Z171 (国家高技术研究发展计划(863)); the Science Fund for Distinguished Young Scholars of Heilongjiang Province of China under Grant No.JC200611 (黑龙江省杰出青年科学基金); the Natural Science Foundation of Heilongjiang Province of China under Grant No.ZJG0705 (黑龙江省自然科学重点基金); the Foundation of Harbin Institute of Technology of China under Grant No.HIT.2003.53 (哈尔滨工业大学校基金)

Received 2006-06-21; Accepted 2007-03-07

监督学习需要大量带标记数据作训练集以保证泛化能力.但在文本处理、生物信息学和网页分类等实际应用中,对数据进行人工标记的代价很高,容易获取的是大量无标记数据.无监督学习属于无任何监督信息的自动学习,虽然不需要带标记数据,但所得模型却不够精确.因此,将少量带标记数据和大量无标记数据结合的半监督学习成为机器学习的研究热点.按照学习任务的不同,半监督学习分为半监督聚类和半监督分类[1].

半监督聚类算法研究无监督学习中如何利用少量的监督信息来提高聚类性能[2].少量的监督信息可以是数据的类别标记或一对数据是否属于同一类的连接约束关系.按照监督信息的使用方式不同,分为基于距离(distance-based)的方法和基于约束条件(constraint-based)的方法[3].基于约束条件的方法目前研究应用较多[2−4],它用监督信息约束最优聚类的搜索过程,典型算法包括将约束条件加入聚类目标函数的算法[5]、强制满足连接约束条件的 COP-K-均值算法[6]、基于隐马尔可夫随机域模型的 HMRF-K-均值算法[7]以及 Generative 模型结合 EM 理论支持的 Seeded-K-均值和 Constrained-K-均值算法[8].其中,Seeded-K-均值和 Constrained-K-均值算法是基于 seeds 集的,它们用少量带标记数据形成 seeds 集来改善 K-均值聚类的初始化效果,进而提高整个数据集上的聚类性能.然而,Basu 等人[4]通过实验表明,这两种基于 seeds 集的半监督聚类算法对 seeds 集的规模和噪声都十分敏感,规模大、噪声小的 seeds 集将显著地提高聚类性能.

半监督分类是从监督学习的角度出发,考虑带标记训练样本不足时如何利用大量无标记样本信息辅助分类器的训练[9],已有方法包括基于 EM 算法的 Generative 模型参数估计法[10−12]、基于转导推理(transductive inference)的支持向量机方法[13]及 Graph-cut 方法[14].Blum 和 Mitchell[15]提出的 Co-training 算法是另一种著名的半监督学习范型.该算法独立地训练两个分类器,然后采用互助方式迭代地扩充带标记数据集并重新训练.标准 Co-training 算法要求属性集能够分为两个不相交的子集,且每个属性子集都能独立训练出分类器,这在实际问题中很难得到满足.为此,Goldman 等人[16]提出了一种改进的 Co-training 算法,不受属性集划分的约束,可用整个属性集训练两个分类器,但要求两个分类器所用的监督学习算法能够将实例空间划分为等价类集合,而且训练过程需要频繁使用耗时的交叉验证作决定.Tri-training 算法是 Zhou 等人[17]提出的一种新的 Co-training 模式半监督分类算法,它使用 3 个分类器进行训练.Tri-training 对属性集和 3 个分类器所用监督学习算法都没有约束,而且不使用交叉验证,因此适用范围更广、效率更高,而且比 Goldman 的改进算法更适合解决带标记样本少的分类问题.

实际聚类应用中带标记数据非常少,而基于 seeds 集的半监督聚类算法受 seeds 集规模和质量的影响显著.针对这个矛盾,本文借鉴 Tri-training 训练过程能够用少量带标记数据指导对无标记数据进行标记,使带标记训练集增大的优点,在由 seeds 集初始化聚类中心前,用 Tri-training 训练过程增大初始 seeds 集的规模.然而,与文献[2,18]所指出的,当少量带标记数据不足以反映大量无标记数据所蕴含的完整聚类结构时,半监督分类方法无法取代半监督聚类算法完成聚类任务的情形一样,当初始 seeds 集比例小时,Tri-training 扩大的 seeds 集会包含大量误标记和噪声数据,甚至出现类别缺失的现象,因此,Tri-training 训练过程只适合在聚类中心初始化前辅助扩大 seeds 集规模.为弥补 Tri-training 过程中不可避免的误标记所导致的 seeds 集噪声增大和类别缺失这样的负面影响,本文同时考虑将基于最近邻规则的数据剪辑技术结合到 Tri-training 对 seeds 集规模的扩大过程中,进行消噪、修正误标记等净化操作,以提高 seeds 集质量,为聚类提供大规模、低噪声的 seeds 集.

本文第 1 节简单介绍 Seeded-K-均值和 Constrained-K-均值两种基于 seeds 集的半监督聚类算法.第 2 节阐述用 Tri-training 扩大 seeds 集过程和 Depuration 数据剪辑技术净化 seeds 集操作,提出基于 Tri-training 和 Depuration 数据剪辑的 DE-Tri-training 半监督 K-均值聚类算法,并给出时间复杂度分析.第 3 节通过实验对算法进行性能测试并对结果进行比较分析.第 4 节对本文工作进行总结与展望.

1 Seeded-K-均值和 Constrained-K-均值算法

MacQueen[19]提出的 K-均值算法是著名的无监督聚类算法.它随机初始化 k 个聚类中心,按距离测度分配数据点到最近的聚类中心,使数据集聚类中心迭代地更新,直至不再变化.最终聚类结果是局部最小化式(1)所示的数据点到聚类中心距离平方和的目标函数.

$$J_{KMeans}=\sum_{h=1}^{k}\sum_{x_i\in X_h}\|x_i-\mu_h\|^2 \tag{1}$$

其中,$\{x_i\}_{i=1}^n$ 是数据集,$\{X_h\}_{h=1}^k$ 是 k 个聚类,$\{\mu_h\}_{h=1}^k$ 是 k 个聚类中心.

与 K-均值算法随机初始化 k 个聚类中心不同,Basu 等人[8]提出的 Seeded-K-均值和 Constrained-K-均值算法用少量带标记的数据作 seeds 集,并按标记将 seeds 集划分为 k 个聚类,由此计算 k 个聚类初始中心.另外,在数据点分配过程中,Constrained-K-均值算法还强制约束 seeds 集中数据点所属聚类不变,只对无标记数据点重新分配.

Basu 通过实验表明[4]:当无噪声 seeds 集所占比例增大时,两种算法的性能都显著提高;相反地,当 seeds 集中噪声比例增大时,Constrained-K-均值性能显著降低.尽管 Seeded-K-均值分配数据点时不受 seeds 集中初始标记的约束,因而相对于 Constrained-K-均值抗噪性稍强,但与无噪声时的 Seeded-K-均值相比,性能下降仍十分显著.

2 基于 Tri-Training 和数据剪辑的半监督聚类算法

Seeded-K-均值与 Constrained-K-均值算法的性能受 seeds 集规模和质量的影响显著,如果能在增大 seeds 集规模的同时提高 seeds 集质量,则其性能的提高也将十分显著.为此,在初始化聚类中心之前,本文利用 Tri-training 的迭代训练过程增大 seeds 集规模,同时,对扩大的 seeds 集进行 Depuration 数据剪辑,以提高 seeds 集质量.

2.1 Tri-Training扩大Seeds集

Tri-Training 分类算法的详细伪代码见文献[17],本文借鉴 Tri-training 执行的 Co-training 式训练过程来实现 seeds 集扩大:假设初始少量带标记的数据集为 L(即初始 seeds 集),由 L 训练得到 3 个不同分类器 H_1,H_2,H_3,x 是无标记数据集 U 内任意一点,如果 H_2 和 H_3 对 x 的分类结果 $H_2(x)$和 $H_3(x)$一致,那么,可将 x 标记为 $H_2(x)$并加入 H_1 的训练集,如此形成 H_1 的新训练集 $S_1'=L\cup\{x\mid x\in U且H_2(x)=H_3(x)\}$.类似地,$H_2$ 和 H_3 的训练集分别扩充为 S_2' 和 S_3',然后,3 个分类器重新训练,如此重复迭代,直至 H_1,H_2,H_3 都没有变化,训练过程结束.Tri-training 训练过程中,3 个分类器相应训练集的迭代扩充过程就是实现初始 seeds 集逐步扩大的过程.本文的算法正是以训练结束时 3 个分类器相应训练集的并集 $S_1'\cup S_2'\cup S_3'$ 作为扩充后的 seeds 集,为初始化聚类中心做准备.

显然,seeds 集扩大过程中,H_2 和 H_3 共同标记 x 为 $H_2(x)$.给 H_1 作训练数据时,如果准确性足够高,会优化 H_1 的训练结果;否则,会在 H_1 的训练集中加入噪声,影响训练效果.为此,Zhou 等人[17]证明:在 PAC 可学习框架下,如果新标记的训练样本足够多且式(2)定义的约束条件满足,则 H_1 重新训练所得假设的分类性能会迭代提高.

$$|L\cup L^t|\left(1-2\frac{\eta_L|L|+\tilde{e}_1^t|L^t|}{|L\cup L^t|}\right)>|L\cup L^{t-1}|\left(1-2\frac{\eta_L|L|+\tilde{e}_1^{t-1}|L^{t-1}|}{|L\cup L^{t-1}|}\right) \tag{2}$$

其中,L^t 是第 t 次迭代 H_2 和 H_3 为 H_1 新标记的训练集,$\tilde{e}_1^t$ 是 H_2 和 H_3 第 t 次迭代的错误率上限,η_L 是初始训练集 L 的噪声率.η_L 通常很小,假定 $0\le\tilde{e}_1^t,\tilde{e}_1^{t-1}<0.5$,则当$|L^t|>|L^{t-1}|$时,由公式(2)可推出 $\tilde{e}_1^t|L^t|<\tilde{e}_1^{t-1}|L^{t-1}|$,即等价于公式(3):

$$0<\frac{\tilde{e}_1^t}{\tilde{e}_1^{t-1}}<\frac{|L^{t-1}|}{|L^t|}<1 \tag{3}$$

其中,$\tilde{e}_1^t$ 和 $\tilde{e}_1^{t-1}$ 可用 H_2 和 H_3 在初始带标记数据集 L 上的错误率近似.在 Tri-training 训练过程中,用公式(3)来判断由 H_2 和 H_3 给出相同标记的数据点 x 组成的数据集 L_1^t 是否应该被加入到 H_1 的新训练集中.

2.2 Depuration剪辑seeds集

如上所述,我们可以用 Tri-training 训练过程对无标记数据作标记,实现 seeds 集的扩充.尽管有约束条件保证各分类器在新标记训练样本足够大时分类精度迭代提高,但实际应用中,初始带标记数据集规模很小,不足以训练出高精度分类器,所以,Tri-training 训练过程中误标记相当数量的数据是难免的,这使得在进行下次迭代时,

另一个分类器的训练集包含许多噪声[20].如果迭代过程中(尤其迭代的早期)能够找到这些误标记数据并修剪,seeds 集质量会得到提高,算法会有更好的聚类效果.各种数据剪辑技术能够有效地改善训练集的质量,本文在 Tri-training 迭代训练中结合了数据剪辑技术,在每个分类器获得新的训练集,在重新训练之前对新训练集进行数据剪辑操作.如此改进的 Tri-training 训练过程称为带数据剪辑的 Tri-training,为方便起见,记作 DE-Tri-training.

考虑到 K-均值聚类算法“最小化类内距离”[4]的目标与最近邻规则的思想一致,所以将最近邻规则下的数据剪辑技术用于聚类算法中 seeds 集的精化是合理的.基于最近邻规则的 Depuration 技术是第 1 种应用原型选择策略的剪辑技术,它从有标记数据集中移除“可疑”数据并修改其他数据的误标记,达到消除训练集噪声、误标记和偏离数据的目的[21].具体操作为:按最近邻规则选取一个数据的 k 个近邻,观察其中是否有 k'个近邻标记相同,以决定移除数据或修改标记.k 和 k'按 generalised editing 规定的条件$(k+1)/2\le k'\le k$ 给定.文献[21]指出,当 k 和 k'设为 3 和 2 时,Depuration 实际效果最好.本文的算法采用这个设定.

2.3 基于Tri-Training和数据剪辑的DE-Tri-Training半监督K-均值算法

图 1 是新的半监督聚类算法在初始化聚类中心前,由 DE-Tri-training 训练过程对初始少量带标记数据集经过扩充和数据剪辑形成大规模、低噪声 seeds 集的详细过程.seeds 集规模的扩充通过 DE-Tri-training 迭代训练过程对无标记数据不断标记,形成新的训练集来实现.逐步扩大的 seeds 集中噪声和误标记数据的剪辑操作分布在两个阶段:第 1 阶段是 DE-Tri-training 训练过程每次迭代标记后,对新训练集的 Depuration 数据剪辑操作,主要消除误分类引起的噪声,同时,对初始 seeds 集中可能存在的误标记和偏离数据也能起到净化作用;第 2 阶段是在训练结束时,DE-Tri-training 所得联合分类器$\{H_1,H_2,H_3\}$对训练集的并集重新分类标记后进行的 Depuration 数据剪辑操作.这样,在联合分类器对 seeds 集重新标记的基础上进一步消除噪声和误标记,提高用于聚类中心初始化的 seeds 集质量.

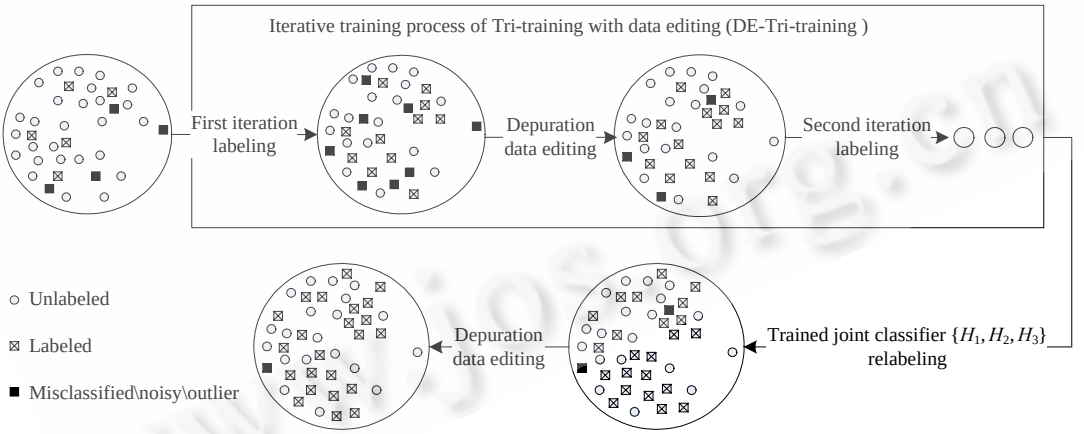


Fig.1 The process of enlargement and data editing on seeds set

图 1 Seeds 集的扩充和数据剪辑过程

算法. 基于 Tri-training 和数据剪辑的 DE-Tri-training 半监督 K-均值聚类算法.

输入:数据集 $X=\{x_i\}_{i=1}^n$,$x_i\in\Re^d$,聚类数目 k,带标记 seeds 集 $S=\cup_{h=1}^k S_h$ (S_h 非空),数据点所属聚类的重新分配方式(seeded 方式或 constrained 方式),未训练的分类器 H_1,H_2,H_3;

输出:数据集 X 上 k 个不相交的聚类划分 $\{X_h\}_{h=1}^k$.

Step 1. DE-Tri-training 训练过程对 S 迭代扩充、剪辑

(a) $L\leftarrow S,U\leftarrow X-L$;对 L 进行 Bootstrap 采样,产生 3 个训练集 S_1', S_2', S_3' 分别训练 H_1,H_2,H_3.

(b) 对每个 $H_i(i=1,2,3)$,从无标记数据集 U 中由 H_j 和 $H_k(j,k\ne i)$选择满足式(3)判别条件的子集

$L_i=\{x|x\in U$ 且 $H_j(x)=H_k(x)\}$进行标记,并生成 H_i 新训练集 $S_i'=L\cup L_i$.

(c) 对每个 S_i' 中的新标记数据子集 L_i 进行 Depuration 数据剪辑操作:

$S'\leftarrow S_i'$;

遍历考察非空 L_i 中每个新标记元素 x,在 $S_i'-\{x\}$ 中找到 x 的 3 个最近邻点.如果某个标记 c 在 x 的 3 个最近邻点中至少出现 2 次,就把 S'中的 x 标记为 c;否则,如果不存在这样的标记 c,就从 S'移除 x;

$S_i'\leftarrow S'$.

(d) 对每个 $H_i(i=1,2,3)$,若 $|S_i'|>|S'|$,则用 S_i' 重新训练.

(e) 如果 H_1,H_2,H_3 中至少有 1 个发生变化,则转(b).

(f) $S\leftarrow S_1'\cup S_2'\cup S_3'$;训练所得联合分类器$\{H_1,H_2,H_3\}$按加权投票规则对 S 中新标记数据重新分类标记.

(g) 对 S 中由$\{H_1,H_2,H_3\}$重新标记的数据子集进行 Depuration 数据剪辑操作.

Step 2. 初始化 k 个聚类中心

将扩充的 seeds 集 S 中数据点按标记划分为 k 个聚类(即 $S=\cup_{h=1}^{k}S_h$,S_h 非空),并计算 k 个聚类初始中心 $\mu_h=\frac{1}{|S_h|}\sum_{x\in S_h}x$,$h=1,...,k$.

Step 3. 重新分配数据点

若是 Seeded 方式,则 X 中每个数据点 x 都重新分配到距离最近的聚类中;

若是 Constrained 方式,对 X 中每个数据点 x,如果 $x\in S_h$,就保留分配到聚类 X_h;否则,分配到最近的聚类

Step 4. 重新计算聚类中心: $\mu_h=\frac{1}{|X_h|}\sum_{x\in X_h}x$,$h=1,\ldots,k$.

Step 5. 如果 k 个聚类中心都未改变,则算法结束;否则,转 Step 3.

需要指出的是,为保证由初始数据集 L 训练得到 3 个不同的分类器,算法 Step 1(a)步沿用文献[17]采用的 Bootstrap 采样技术.在文献[17]中,Tri-training 训练结束对新数据分类时,训练所得联合分类器$\{H_1,H_2,H_3\}$采用多数投票规则.考虑到实际聚类数目通常大于 2,新算法中 DE-Tri-training 训练所得联合分类器采用式(4)所示加权投票规则进行类别标记(算法 Step 1(f)),权重由 3 个分类器在初始带标记数据集 L 上的分类准确率 $A_i(L)$所决定.

$$H_{\{1,2,3\}}(x)=\arg\max_{c\in label}\frac{\sum_{i=1}^{3}\delta(c,H_i(x))\times A_i(L)}{\sum_{i=1}^{3}A_i(L)} \tag{4}$$

其中,$\delta(c,H_i(x))=\begin{cases}1, & H_i(x)=c\\ 0, & H_i(x)\neq c\end{cases}$.

另外,为便于考察新算法中 Tri-training 和 Depuration 的作用,本文考虑初始 seeds 集无噪声的情况,所以,算法 Step 1(c)和 Step 1(g)只对新标记的数据子集进行 Depuration 数据剪辑操作.

2.4 算法复杂性分析

算法 Step 1(a)对基分类器训练若采用基于迭代的学习算法(如神经网络)则可设定初始带标记数据集 L 上的分类错误率阈值为迭代结束条件;Step 1(e)Tri-training 迭代可以新训练集 S_i' 不再变化为结束条件;Step 5 聚类迭代则可用相应聚类中心的距离误差阈值作为结束条件.

假设数据集规模为 n,维数为 d,新算法时间复杂度主要包括两个阶段的时间代价.第 1 阶段,DE-Tri-training 对初始 seeds 集扩大和剪辑共 m_1 次迭代,时间代价为 $O(m_1\times[2\times T_c(H;n-|L|)+|L^t|\times(|L|+|L^t|)\times d+T_t(H;|L|+|L^t|)])$,其中,$T_c(H;n-|L|)$是分类器对无标记数据标记的代价,$T_t(H;|L|+|L^t|)$是用新标记训练集重新训练分类器的代价,二者都由具体选用的分类器 H 所决定,$|L^t|\times(|L|+|L^t|)\times d$ 是 Depuration 按最近邻规则对新标记数据集 L^t 剪辑的代价.因此,第 1 阶段最坏情况下,$(|L|\to 0;|L^t|\to n)$代价为 $O(m_1\times(T_c(H;n)+T_t(H;n)+dn^2))$.第 2 阶段进行半监督 K-均值聚类的代价主要集中在重新分配数据点和重新计算聚类中心的 m_2 次迭代过程上,代价为 $O(m_2\times kdn)$.所以,算法最坏情况

下的时间复杂度为 $\max\{O(m_1\times T_c(H;n)),O(m_1\times T_t(H;n)),O(m_1\times dn^2),O(m_2\times kdn)\}$.

3 实验分析

本文采用 UCI 机器学习数据库[22]中的 6 个数据集进行实验,测试新算法的聚类性能,表 1 列出了这些数据集的信息,其中,Letters 数据子集是从手写字符集中随机抽取{*I*,*J*,*L*}这 3 个最难识别字符的 10%所组成.类似地,Digits 数据子集是从{3,8,9}中随机抽取 10%所组成.Image segmentation 中 1 个常数属性对聚类过程帮助不大,所以被删除.本文实验数据集、程序和实验过程详细记录可从 http://nclab.hit.edu.cn/dc.html 获得.

Table 1 Information of UCI datasets in the experiments

表 1 实验所用 UCI 数据集

Data set	No. of instances	No. of attributes	No. of classes
Iris	150	4	3
Wine	178	13	3
Letters	227	16	3
Digits	318	16	3
Ionosphere	351	34	2
Image segmentation	2310	18	7

为便于比较,我们测试随机初始化的 K-均值算法、Seeded-K-均值算法、Constrained-K-均值算法及本文提出的 DE-Tri-training 半监督 K-均值算法(DE-Tri-training Seeded-K-均值和 DE-Tri-training Constrained-K-均值).为明确分析 Tri-training 和数据剪辑的作用,对只用 Tri-training 而无数据剪辑的半监督 K-均值算法(Tri-training Seeded-K-均值和 Tri-training Constrained-K-均值)也进行了实验比较.

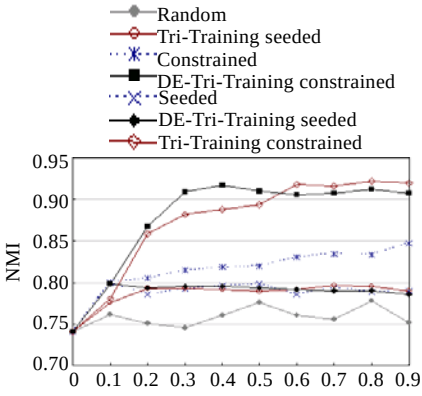
聚类性能评价采用目前广泛使用的正则化互信息(normalized mutual information,简称 NMI)指标[2,4,23].NMI 可确定聚类算法在测试集上的类别标记结果与实际结果之间的共有信息量,NMI 越大,则聚类效果越好.NMI 定义方式有多种,本文中 NMI 计算采用文献[23]给出的定义.

每个数据集上不同聚类算法的学习曲线和相应 seeds 集规模的比较如图 2~图 5 所示,曲线上每个点是每种算法独立运行 10 次 10-folds 交叉验证的结果.具体方案为:数据集等分为 10 个 fold,每次取 1 个 fold 作测试集(整个数据集的 10%),其余 9 个 fold(占 90%)按预定的比例划分为有标记 seeds 集和无标记数据子集(子集中不同类别数据的比例与整个数据集中比例一致),然后运行被测算法,10 个 fold 依次被选作测试集,相应 10 个结果的平均值作为一次 10-folds 交叉验证结果.如此独立运行 10 次 10-folds 交叉验证实验,取平均值作为学习曲线上的一点.每种聚类算法都是在整个数据集上运行,但只在测试集上计算 NMI 值.

在 Tri-training 中,3 个分类器 H_1,H_2,H_3 采用 3 层 BP 神经网络(BPNN).K-均值聚类和 Depuration 数据剪辑中距离测度均采用欧氏距离,最大迭代步数为 200,误差停止阈值为 1e–5.实验分为两组:第 1 组主要考察 Tri-training 的作用,在 Tri-training 训练过程中为每个 BPNN 分类器设定充足的训练迭代次数,以保证当初始 seeds 集的比例为 0.1 时,Tri-training 训练所得联合分类器中至少有两个分类器对初始 seeds 集的分类准确率大于 0.9;第 2 组主要考察 Depuration 的作用,每个 BPNN 分类器训练次数不足,当初始 seeds 集的比例为 0.1 时,联合分类器中至少有 1 个分类器对初始 seeds 集的分类准确率小于 0.7.

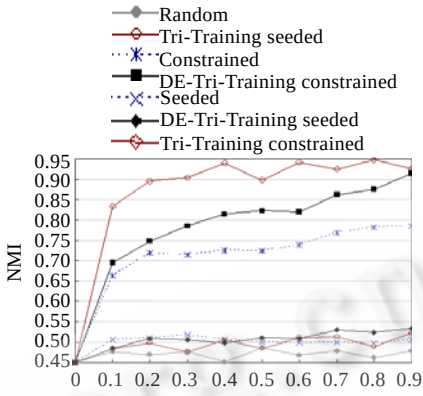
3.1 BPNN分类器训练充分时Tri-training作用的分析

第 1 组实验,6 个数据集上 7 种不同聚类算法的 NMI 学习曲线的比较如图 2 所示.图中 Constrained-K-均值和 Seeded-K-均值优于 Random K-均值,这与文献[4]的结果基本一致,但所有 Seeded 模式的 K-均值算法比 Constrained 模式要差,这可由文献[2]中的结果来解释.本文算法相关的 DE-Tri-training Constrained-K-均值和 Tri-training Constrained-K-均值的 NMI 曲线明显优于 Random K-均值、Seeded-K-均值和 Constrained-K-均值,这是因为每个 BPNN 分类器泛化能力得到保证时,seeds 集的规模扩大且 Tri-training 训练过程标记新数据和联合分类器重新标记 seeds 集的高准确率都保证了用于聚类中心初始化的 seeds 集是大规模低噪声的.



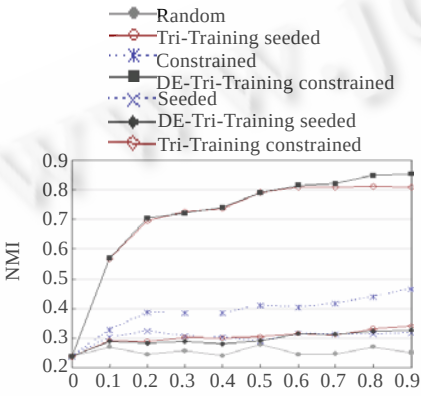
(a) Iris: 8 nodes in hidden layer, 500 iterations, learning rate=0.1

(a) Iris: 隐层结点 8 个,迭代次数 500 次,学习速率=0.1



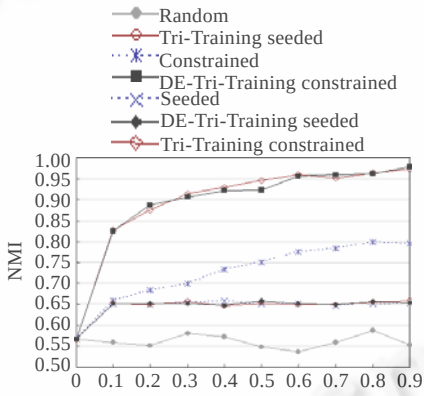
(b) Wine: 8 nodes in hidden layer, 300 iterations, learning rate=0.1

(b) Wine: 隐层结点 8 个,迭代次数 300 次,学习速率=0.1



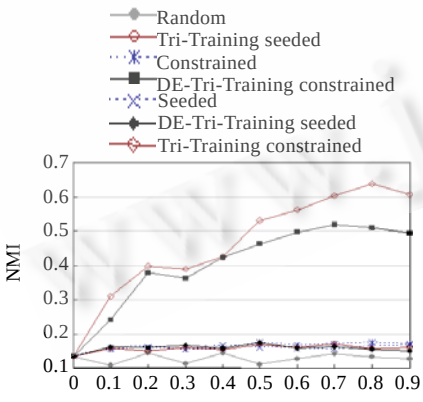
(c) Letters: 8 nodes in hidden layer, 300 iterations, learning rate=0.1

(c) Letters: 隐层结点 8 个,迭代次数 300 次,学习速率=0.1



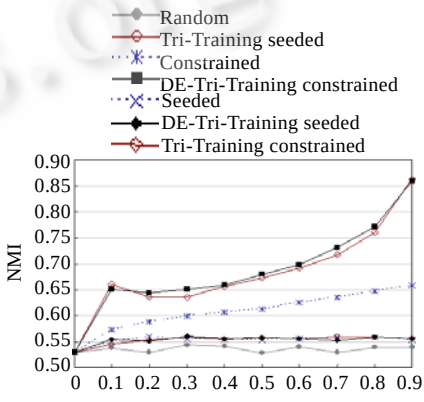
(d) Digits: 10 nodes in hidden layer, 300 iterations, learning rate=0.1

(d) Digits: 隐层结点 10 个,迭代次数 300 次,学习速率=0.1



(e) Ionosphere: 12 nodes in hidden layer, 500 iterations, learning rate=0.1

(e) Ionosphere: 隐层结点 12 个,迭代次数 500 次,学习速率=0.1



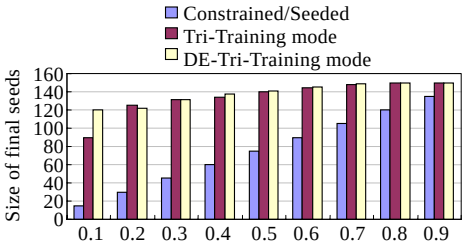
(f) Image segmentation: 12 nodes in hidden layer, 1000 iterations, learning rate=0.3

(f) Image segmentation: 隐层结点 12 个,迭代次数 1000 次,学习速率=0.3

Fig.2 NMI comparison on six datasets when BPNN trained sufficiently

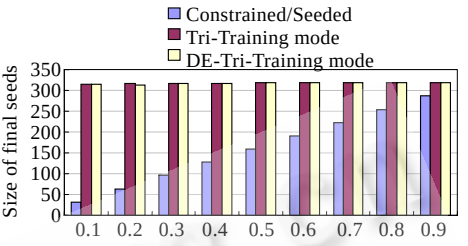
图 2 BPNN 分类器训练充分时 6 个数据集上 NMI 的比较

图 3(a)和图 3(b)是 Iris 和 Digits 数据集上 3 种模式算法用于聚类中心初始化的最终 seeds 集规模的比较.显然,Tri-training 模式和 DE-Tri-training 模式的算法中,Tri-training 训练过程对 seeds 集规模的扩大有十分显著的作用.



(a) Iris: 8 nodes in hidden layer, 500 iterations, learning rate=0.1 (Corresponding to NMI curves in Fig.2(a))

(a) Iris: 隐层结点 8 个,迭代次数 500 次,学习速率=0.1 (对应图 2(a)中 NMI 曲线)



(b) Digits: 10 nodes in hidden layer, 300 iterations, learning rate=0.1 (Corresponding to NMI curves in Fig.2(d))

(b) Digits: 隐层结点 10 个,迭代次数 300 次,学习速率=0.1 (对应图 2(d)中 NMI 曲线)

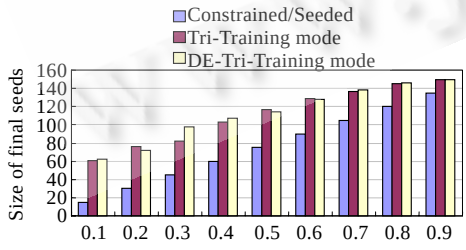
Fig.3 Size comparison of final seeds used to cluster centroids initialization when BPNN trained sufficiently

图 3 BPNN 分类器训练充分时用于聚类中心初始化的 seeds 集规模的比较

此外我们发现,在 Wine,Ionosphere 以及 Iris(初始 seeds 集比例大于 0.5 后)3 个数据集上,Tri-training Constrained-K-均值优于 DE-Tri-training Constrained-K-均值,这说明每个分类器训练充分时 Tri-training 的作用大于 Depuration.然而,图 2(a)Iiris 数据集上的 NMI 学习曲线表明,当初始 seeds 集比例小于 0.1 时,无数据剪辑的 Tri-training Constrained-K-均值的 NMI 值不仅低于带数据剪辑的 DE-Tri-training Constrained-K-均值,而且低于 Constrained-K-均值.随着初始 seeds 集比例的增大,Tri-training Constrained-K 均值由于训练样本增大使新样本的标记准确性提高,所以 NMI 值迅速超过 Constrained-K-均值,但与 DE-Tri-training Constrained-K-均值相比,由于后者的 Depuration 对 seeds 集不断剪辑,在规模增大的同时又保证了质量的提高,因而直至初始 seeds 集比例达 0.55 时,后者性能一直优于前者;但初始 seeds 集比例大于 0.55 后,前者由于训练集充分而保证新标记数据集准确性大幅度提高,后者由于 Depuration 固有的误差率使性能趋于稳定,但即使前者在 0.9 时达到的 NMI 最大值也仅与后者在 0.4 时就已达到的最大值相当,这说明在 Iris 数据集上 DE-Tri-training Constrained-K-均值用相对小的初始带标记 seeds 集就能获得较好的聚类性能.

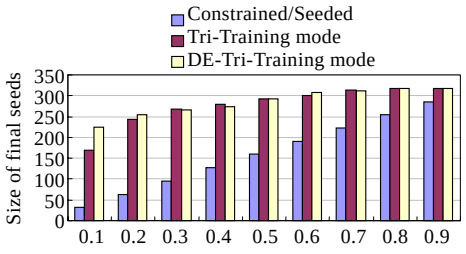
3.2 BPNN分类器训练不充分时Depuration作用分析

图 4、图 5 是第 2 组实验,由于训练不充分,使 Tri-training 训练过程标记新数据准确率降低时,Iris 和 Digits 两个典型数据集上 7 种算法的性能比较.与图 3(a)和图 3(b)相比,图 4(a)和图 4(b)表明,Tri-training 模式和 DE-Tri-training 模式中 seeds 集规模的扩大程度受到极大影响,尤其当初始 seeds 集比例小时扩大规模明显降低.



(a) Iris: 8 nodes in hidden layer, 50 iterations, learning rate=0.1 (Corresponding to NMI curves in Fig.5(a))

(a) Iris: 隐层结点 8 个,迭代次数 50 次,学习速率=0.1 (对应图 5(a)中 NMI 曲线)



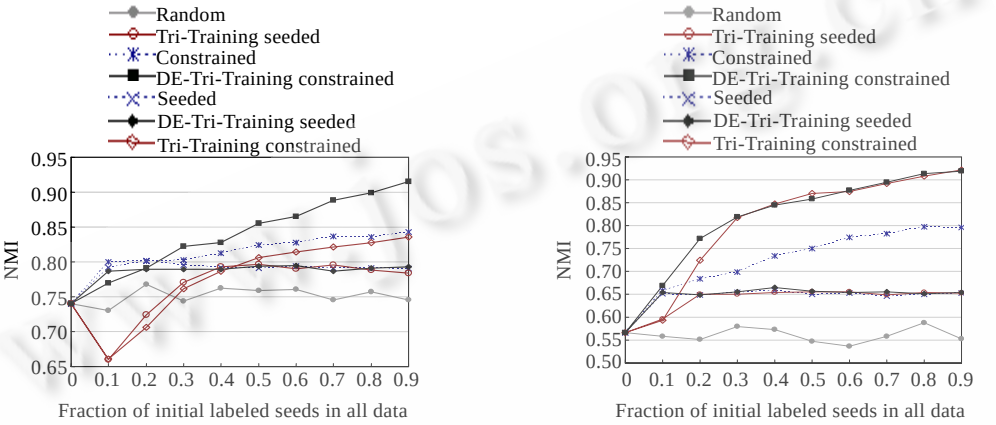
(b) Digits: 10 nodes in hidden layer, 20 iterations, learning rate=0.1 (Corresponding to NMI curves in Fig.5(b))

(b) Digits: 隐层结点 10 个,迭代次数 20 次,学习速率=0.1 (对应图 5(b)中 NMI 曲线)

Fig.4 Size comparison of final seeds used to cluster centroids initialization when BPNN trained insufficiently

图 4 BPNN 分类器训练不充分时用于聚类中心初始化的 seeds 集规模的比较

图 5(a)表明,无数据剪辑的 Tri-training Constrained-K-均值和 Tri-training Seeded-K-均值在初始 seeds 集比例小于 0.3 时,性能远差于其他所有算法,即使初始比例增至 0.9,NMI 值都一直小于 Constrained-K-均值;而在带数据剪辑的 DE-Tri-training Constrained-K-均值和 DE-Tri-training Seeded-K-均值中,Depuration 数据剪辑操作能够对误标记引起的噪声数据加以纠正,有效弥补了训练不充分带来的负面影响,在初始 seeds 集比例小于 0.2 时,其性能就远好于无数据剪辑的算法,并接近 Constrained-K-均值和 Seeded-K-均值的性能,尤其是 DE-Tri-training Constrained-K-均值从初始比例大于 0.2 时,NMI 值就一直大于 Constrained-K-均值.同样,图 5(b)表明,Digits 上初始 seeds 集比例小时,DE-Tri-training Constrained-K-均值和 DE-Tri-training Seeded-K-均值的优势更明显.因而,该组实验进一步证明了新算法中 Depuration 操作对增大的 seeds 集所含噪声进行有效净化的关键作用.



(a) Iris: 8 nodes in hidden layer, 50 iterations, learning rate=0.1 (b) Digits: 10 nodes in hidden layer, 20 iterations, learning rate=0.1

(a) Iris: 隐层结点 8 个,迭代次数 50 次,学习速率=0.1 (b) Digits: 隐层结点 10 个,迭代次数 20 次,学习速率=0.1

Fig.5 NMI comparison on two typical datasets when BPNN trained insufficiently

图 5 BPNN 分类器训练不充分时 2 个典型数据集上 NMI 的比较

3.3 纵向比较分析

纵向比较两组实验而发现,第 1 组中 Tri-training 模式与 DE-Tri-training 模式算法性能的差异程度远不如第 2 组显著.由图 4(a)和图 4(b)可以看出,最终 seeds 集规模的差异并不是造成第 2 组中二者性能明显差异的主要原因,而应该是 seeds 集质量的差异.第 2 组各分类器训练不充分时,互标记准确率低,Tri-training 迭代出现严重误标记传播现象.尽管 DE-Tri-training 可通过剪辑操作减少误标记噪声降低误标记传播,但与第 1 组训练充分时 Tri-training 模式性能的差距仍十分明显,这说明提高 Tri-training 中各分类器标记精度是避免误标记数据传播的关键.

两组实验中,Constrained-K-均值性能曲线表明,在保证 seeds 集质量的前提下,聚类性能随着 seeds 数量的增加而提高.新算法中,seeds 集质量由 Tri-training 中各分类器互标记准确率保证.当标记准确率低时,Depuration 剪辑可辅助提高 seeds 集质量.在实际应用中,seeds 集质量可通过 Tri-training 当前迭代所得联合分类器对初始 seeds 集 L 的标记准确率进行评价,当 seeds 集数量增大但 Tri-training 在 L 上出现过拟合情形时,即可认为种子数量达到平衡.

如前所述,新算法假定初始 seeds 集无噪声,便于考察 Depuration 对 Tri-training 迭代扩大 seeds 集时所产生误标记噪声的有效净化作用.实际应用中考虑初始 seeds 集噪声,依据文献[21]中 Depuration 能够获得高质量训练集的结论,可在 Step 1 用 DE-Tri-training 扩充、剪辑 seeds 集之前,先用 Depuration 剪辑操作对初始 seeds 集消噪.

4 结论与展望

本文提出了基于 Tri-training 和数据剪辑的 DE-Tri-training 半监督 K 均值聚类算法.实验结果表明,在聚类中心初始化之前,新算法能够有效利用带数据剪辑的 Tri-training 训练过程,对无标记数据标记,并对噪声数据剪辑,使得 seeds 集规模迭代增大的同时质量有所提高,最终达到改善聚类性能的效果.

相比其他基于约束条件的半监督聚类典型算法,如 COP-K-均值算法和利用遗传算法优化含约束条件 K-均值聚类目标函数的算法而言,新算法通过获取大规模低噪声 seeds 集,有效地弥补了因随机初始化聚类中心而易陷入局部极优的不足;此外,对不可避免的噪声标记,COP-K-均值算法会因噪声数据引起约束条件不一致而聚类失败,新算法则通过 Depuration 对噪声数据剪辑而具有更好的鲁棒性和聚类性能.

实验结果启示我们,对 DE-Tri-training 半监督 K.均值算法中 Depuration 数据剪辑误差率和 Tri-training 训练过程中新数据标记准确率这两个指标间的关系进行理论分析,确定不同情形下 Depuration 的最佳启动时机,应是今后新算法的重要研究方向.

值得注意的是,这种从少量带标记数据出发,用半监督分类方法对无标记数据进行标记,增大聚类所需的监督信息,同时结合数据剪辑技术提高监督信息质量的思想,可推广到对其他半监督聚类算法的改进.但是,仍有许多重要问题需要解决,比如数据剪辑技术应如何选择、当监督信息不是类别标记而是其他约束条件形式(如数据间 must- link 和 cannot-link 连接关系)时如何实现数据剪辑操作,都需要深入的研究和探讨.

References:

[1] Olivier C, Bernhard S, Alexander Z. Semi-Supervised Learning. Cambridge: MIT Press, 2006.

[2] Zhong S. Semi-Supervised model-based document clustering: A comparative study. Machine Learning, 2006,65(1):3−29.

[3] Bilenko M, Basu S, Mooney RJ. Integrating constraints and metric learning in semi-supervised clustering. In: Carla EB, ed. Proc. of the 21st Int'l Conf. on Machine Learning (ICML 2004). Banff: ACM Press, 2004. 81−88.

[4] Basu S. Semi-Supervised clustering: Probabilistic models, algorithms and experiments [Ph.D. Thesis]. Austin: University of Texas at Austin, 2005.

[5] Demiriz A, Bennett KP, Embrechts MJ. Semi-Supervised clustering using genetic algorithms. In: Dagli CH, ed. Proc. of the Intelligent Engineering Systems Through Artificial Neural Networks 9 (ANNIE'99). New York: ASME Press, 1999. 809−814.

[6] Wagstaff K, Cardie C, Rogers S, Schroedl S. Constrained K-means clustering with background knowledge. In: Carla EB, Andrea PD, eds. Proc. of the 18th Int'l Conf. on Machine Learning (ICML 2001). San Fransisco: Morgan Kaufmann Publishers, 2001. 577−584.

[7] Basu S, Bilenko M, Mooney RJ. A probabilistic framework for semi-supervised clustering. In: Won K, Ron K, Johannes G, William D, eds. Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2004). Seattle: ACM Press, 2004. 59−68.

[8] Basu S, Banerjee A, Mooney RJ. Semi-Supervised clustering by seeding. In: Claude S, Achim GH, eds. Proc. of the 19th Int'l Conf. on Machine Learning (ICML 2002). San Fransisco: Morgan Kaufmann Publishers, 2002. 19−26.

[9] Seeger M. Learning with labeled and unlabeled data. Technical Report, Edinburgh: University of Edinburgh, 2002.

[10] Nigam K, McCallum AK, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. Machine Learning, 2000,39(2-3):103−134.

[11] Ghahramani Z, Jordan MI. Supervised learning from incomplete data via an EM approach. In: Jack DC, Gerald T, Joshua A, eds. Proc. of the Advances in Neural Information Processing Systems Vol.6 for the 7th NIPS Conf. Denver: Morgan Kaufmann Publishers, 1994. 120−127.

[12] Miller DJ, Browning J. A mixture model and EM-based algorithm for class discovery, robust classification, and outlier rejection in mixed labeled/unlabeled data sets. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003,25(11):1468−1483.

[13] Joachims T. Transductive inference for text classification using support vector machines. In: Ivan B, Saso D, eds. Proc. of the 16th Int'l Conf. on Machine Learning (ICML'99). San Fransisco: Morgan Kaufmann Publishers, 1999. 200−209.

[14] Blum A, Lafferty J, Rwebangira M, Reddy R. Semi-Supervised learning using randomized mincuts. In: Carla EB, ed. Proc. of the 21st Int'l Conf. on Machine Learning (ICML 2004). Banff: ACM Press, 2004. 934−947.

[15] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Peter B, Yishay M, eds. Proc. of the 11th Annual Conf. on Computational Learning theory (COLT'98). Madison: ACM Press, 1998. 92−100.

[16] Goldman S, Zhou Y. Enhancing supervised learning with unlabeled data. In: Pat L, ed. Proc. of the 17th Int'l Conf. on Machine Learning (ICML 2000). San Francisco: Morgan Kaufmann Publishers, 2000. 327−334.

[17] Zhou ZH, Li M. Tri-Training: Exploiting unlabeled data using three classifiers. IEEE Trans. on Knowledge and Data Engineering, 2005,17(11):1529−1541.

[18] Bouchachia A, Pedrycz W. Data clustering with partial supervision. Data Mining and Knowledge Discovery, 2006,12(1):47−78.

[19] MacQueen J. Some methods for classification and analysis of multivariate observations. In: Le L, Neyman J, eds. Proc. of the 5th Berkeley Symp. on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967. 281−297.

[20] Li M, Zhou ZH. SETRED: Self-Training with editing. In: Proc. of the Advances in Knowledge Discovery and Data Mining (PAKDD 2005). LNAI 3518, Heidelberg: Springer-Verlag, 2005. 611−621.

[21] Sánchez JS, Barandela R, Marqués AI, Alejo R, Badenas J. Analysis of new techniques to obtain quality training sets. Pattern Recognition Letters, 2003,24(7):1015−1022.

[22] Blake C, Keogh E, Merz CJ. UCI repository of machine learning databases. Irvine: University of California, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html

[23] Strehl A, Ghosh J, Mooney R. Impact of similarity measures on Web-page clustering. In: Proc. of the AAAI on Workshop on Artificial Intelligence for Web Search (AAAI 2000). Menlo Park: AAAI Press/MIT Press, 2000. 58−64.

邓超(1978－),男,山西长治人,博士生,主要研究领域为机器学习,数据挖掘.

郭茂祖(1966－),男,博士,教授,博士生导师,主要研究领域为机器学习与数据挖掘,计算生物学与生物信息学.