DOI:10.11721/cqnuj20190211

结合密度峰值优化模糊聚类的自训练方法*

罗云松[1]，吕　佳[1,2]
(1. 重庆师范大学 计算机与信息科学学院，重庆 401331；2. 重庆市数字农业服务工程技术研究中心，重庆 401331)

摘要:【目的】为了在迭代自训练之前探索数据集分布情况，挑选出所含信息量较大且置信度较高的无标记样本加入训练集训练，让训练出的初始分类器有较高的准确性，提高自训练方法的泛化性。【方法】以聚类假设为基础，先对无标记样本集进行密度峰值聚类，在人工地选出聚类中心后，将新的聚类中心作为模糊聚类的初始聚类中心进行模糊聚类，从而筛选出有用的无标记样本。【结果】通过使用密度峰值优化模糊聚类算法，筛选出所含信息量大且置信度高的样本加入了训练集，训练出泛化性更强、分类精度更高的分类器。【结论】实验结果表明，改进后的自训练方法能快速发现样本集原始空间结构，筛选出有用无标记样本加入训练集，与结合其他聚类算法的自训练方法相比分类精度有所提高。

关键词:半监督学习；自训练方法；密度峰值优化模糊聚类；聚类假设
中图分类号:TP181　　**文献标志码**:A　　**文章编号**:1672-6693(2019)02-0094-07

自训练方法(Self-training)[1]是一种研究较多且易于实现的半监督学习[2]分类算法。自训练的主要过程就是围绕数据集本身，不断发掘内部有用信息，自我训练，不断提高。该方法因简单有效且训练成本较低而倍受专家学者青睐。

在半监督学习分类中，通常使用聚类假设将无标记样本所揭示的数据分布信息与类别标记相联系[3]。聚类假设是指在样本集中，如果一些样本分布在同一类簇中，则它们的类标签相同的概率会比较大。因此，在对整个样本集进行分类时，应同时考虑有标记样本和无标记样本的分布情况，分类决策平面应尽可能穿过样本分布稀疏的区域，从而避免同一类簇样本被错误划分。基于此聚类假设产生了许多经典的半监督分类算法[4-7]。

也有许多学者在迭代训练之前，通过结合一些聚类算法来探索样本集的数据分布情况，以此来改进自训练方法[8-15]。例如，文献[8]提出将模糊 C-均值聚类算法(Fuzzy C-means，FCM)集成到自训练方法中；文献[9]提出一种结合 K-means 的半监督文本分类方法(CBC)；文献[10-11]提出将密度峰值聚类算法(Density peak clustering，DPC)[12]结合到自训练方法中；文献[13-14]在对无标记样本聚类后，使用了数据剪辑技术对标记错误的无标记样本进行剔除，以此来提高分类器的分类精度。以上文献都是通过先对大量无标记样本使用一种聚类方法，筛选出有用无标记样本加入训练集，从而完成分类器的训练。

DPC 算法是一种方便快捷、简单有效的聚类算法。针对该聚类算法存在的一些缺陷，部分学者提出了一些改进[16-17]。为了通过结合聚类方法挑选出所含信息量较大且置信度高的无标记样本加入训练集训练，让训练出的初始分类器有较高的准确性，提高自训练方法的泛化性，受到上述文献的共同启发，本文提出结合密度峰值优化模糊聚类的半监督自训练方法(Naive Bayes self-training combined DPCFCM，NBSTDPCFCM)。在 UCI 数据集上进行对比仿真实验，证明了本文算法的有效性。

1 密度峰值优化模糊聚类算法

DPCFCM 聚类算法是 DPC 和 FCM 两种算法的结合。该算法和传统 FCM 算法一样是一种软聚类算法，即一个无标记样本并不是固定属于某一类别，而是采用隶属度的概念来度量该样本属于某一类的概率有多大，相比于传统非此即彼的硬划分聚类更为合理。然而，传统 FCM 算法对初始聚类中心较为敏感，该算法的终止条件

* **收稿日期**:2018-07-20　**修回日期**:2019-02-10　**网络出版时间**:2019-03-15 07:00
资助项目:重庆市自然科学基金(No. cstc2014jcyjA40011)；重庆市教育委员会 2016 年人文社会科学研究项目(No. 16SKGH032)；重庆市教育委员会科技项目(No. KJ1600322)；重庆师范大学科研项目(No. YKC18025)
第一作者简介:罗云松，男，研究领域为机器学习、数据挖掘，E-mail:；**通信作者**:吕佳，女，教授，博士，E-mail: 13368431624@qq.com
网络出版地址:http://kns.cnki.net/kcms/detail/50.1165.N.20190315.0056.018.html

为目标函数达到极小值，同时该算法对于非球形分布的数据集来说分类效果并不理想。对于球形分布的数据集，无论初始聚类中心在哪里，它的目标函数往往能迭代下降到全局最优的极值点。而对于许多非球形分布的数据集来说，初始聚类中心的选择并不一定是在全局最优点附近，因而容易陷入局部最优，不能很好地发现数据分布情况。在这种情况下筛选出的无标记样本所暗含的信息量不大，更为严重的是，这些无标记样本可能被错误标记，加入训练后反而会导致分类器性能下降。

针对该问题，DPCFCM 算法通过引入 DPC 算法的密度峰值点思想来进行优化改善。DPC 算法对初始聚类中心的选择基于：1) 自身局部密度应大于周围样本点的局部密度；2) 该聚类中心应与其它密度峰值点(即聚类中心)有较大的距离。

基于此假设条件，给定数据集 $S=\{x_i \mid x_i \in \mathbf{R}^n, i=1,\cdots,N\}$，对每一个样本点 x_i，计算它的局部密度 ρ_i 和它到其他局部密度比它大且距离最近的样本 x_j 之间的相对距离 δ_i，其中局部密度与相对距离的定义如下。

定义 1　样本点 x_i 的局部密度：

$$\rho_i = \sum_j x(d_{ij} - d_c), \tag{1}$$

其中 $\begin{cases}1, x<0\\0, \text{否则}\end{cases}$。$d_{ij}$ 为样本 x_i, x_j 之间的欧氏距离，d_c 为截断距离。

定义 2　密度峰值点 x_i 的相对距离：

$$\delta_i = \begin{cases}\min\limits_{j \in I_s^i}(d_{ij}) I_s^i \neq \varnothing,\\ \max\limits_{j \in I_s^i}(d_{ij}) I_s^i = \varnothing。\end{cases} \tag{2}$$

其中，$I_s^i = \{k \in I_s : \rho_j > \rho_i\}$。

显然，由(1)式知，样本点的局部密度对截断距离 d_c 敏感。当数据集较大时，DPC 算法的聚类结果受到该参数的影响较小，反之则较大。由(2)式可知，对于样本点 x_i，若该点为局部密度峰值点或全局密度峰值点，则局部密度 ρ_i 大于周围样本点的局部密度，相对距离 δ_i 也应远大于近邻样本点的 δ 值。因此，在 DPC 算法中，类簇中心的选择应为 ρ 和 δ 的值都异常大的样本点。DPC 算法通过构造样本局部密度 ρ 和相对距离 δ 的决策图，人为地选择类簇中心。在遍历数据集后，将剩余的样本点 x_j 归入局部密度比 ρ_j 大且距离 x_j 最近的样本点所在的类簇中，一次性完成对剩余样本 x_j 的分配，从而降低该算法的时间复杂度。

这样，改进后的 DPCFCM 算法不再随机选择初始聚类中心，而是通过(1)，(2)式来计算每个样本点的 ρ 和 δ，通过比较它们的大小来确定初始聚类中心。确定初始聚类中心之后，DPCFCM 算法再通过构造目标函数，分别对其中的隶属度和聚类中心进行求导计算，在达到约束条件之后停止迭代。具体的目标函数和约束条件为：

$$J(\boldsymbol{U}, \boldsymbol{V}) = \sum_{j=1}^{C}\sum_{i=1}^{N}(u_{ij})^m \left\| x_i - x_j \right\|^2, \tag{3}$$

$$\text{s.t.}\ \sum_{i=1}^{n} u_{ij} = 1, u_{ij} \in [0,1], \sum_{i=1}^{n} u_{ij} \in (0,n], 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant c。$$

其中 $m \in (1,+\infty)$ 为加权模糊系数，反应控制隶属度在各类之间的共享程度，u_{ij} 为样本 x_i 属于类 j 的隶属程度，$\boldsymbol{U}$ 为 u_{ij} 构成的 $C \times N$ 的隶属度矩阵，$\boldsymbol{V}$ 为聚类中心矩阵，$\| x_i - c_j \|^2$ 为样本点 x_i 到类中心 c_j 的欧氏距离。对于目标函数使用 Langrange 乘数法求解，分别对 u_{ij} 和 c_j 求偏导，得到让(3)式达到最小值的必要条件：

$$u_{ij} = \frac{1}{\sum\limits_{k=1}^{c}\left(\dfrac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|}\right)^{\frac{2}{m-1}}}, \tag{4}$$

$$c_j = \frac{\sum\limits_{i=1}^{n} u_{ij}^m x_i}{\sum\limits_{i=1}^{n} u_{ij}^m}。 \tag{5}$$

综上，改进后的 DPCFCM 算法拥有更好的全局性，能够发现数据的整体结构。同时该聚类算法也是一种改进的软聚类方法，通过采用 FCM 算法的聚类隶属度来改善 DPC 算法的分配策略，以此减少 DPC 算法的二值分配策略对样本集数据空间的改变。DPCFCM 算法既结合了模糊聚类和密度峰值聚类的优点，又互补了对方的缺

点;既能发现样本集的真实数据分布情况,找出样本稀疏的区域,划定分类决策边界,又是采用软聚类的模式,用隶属度来合理地反应样本的类别概率。因此,本文选择该聚类算法来探索样本集的数据分布情况,然后训练贝叶斯分类器。

2 结合密度峰值优化模糊聚类的自训练方法

NBSTDPCFCM算法通过在传统的自训练框架上嵌入DPCFCM聚类算法,将聚类和分类算法有机地统一在自训练框架中。该算法的具体思想为,首先用少量有标记样本训练出一个初始的贝叶斯分类器,然后对大量无标记样本进行DPCFCM聚类,筛选出聚类隶属度高同时所含信息量大的无标记样本给贝叶斯分类器分类,分类后选取出置信度高的有标记样本加入训练集中重新训练贝叶斯分类器,直到无标记样本都作上标记。本文提出的NBSTDPCFCM算法的具体算法步骤如下:

输入:数据集 D,有标记样本集 L,无标记样本集 U,截断距离参数 d_c。

输出:聚类结果 C,聚类中心 V,贝叶斯分类器。

步骤1,对无标记样本集 U 进行数据预处理;

步骤2,输入经过预处理的 U,参数 d_c。根据(1),(2)式计算 U 中每个样本点的局部密度 ρ_i 和相对距离 δ_i,确定初始聚类中心;

步骤3,根据(4)式计算隶属度矩阵 $\boldsymbol{U}^{(1)}$,根据(5)式更新聚类中心 V,输出对 U 的聚类结果 C;

步骤4,利用有标记样本集 L 训练初始贝叶斯分类器;

步骤5,根据聚类结果 C 选取高隶属度的无标记样本集 T_1 给贝叶斯分类器分类;

步骤6,分类后选取高置信度的有标记样本集 T_2 加入 L 中并重新训练新的贝叶斯分类器;

步骤7,更新 $L \leftarrow L \cup T_2$,$U \leftarrow U - T_2$,返回步骤2。

3 实验结果与分析

为了证明本文算法的有效性,选用如下算法进行对比实验。

1) 文献[10]提出的结合密度峰值聚类的朴素贝叶斯自训练算法(NB self-training emsemble density peak clustering,NBSTDPC);

2) 文献[8]提出的结合模糊C均值聚类的朴素贝叶斯自训练算(NB self-training emsemble semi-supervised fuzzy C means,NBSTFCM);

3) 文献[1]提出的半监督朴素贝叶斯自训练方法(Naive Bayes self-training,NBST)。

实验采用9个UCI标准数据集,具体的数据集描述见表1。将每个数据集随机选择80%作为训练集,20%作为测试集,在训练集中随机选择5%作为初始有标记样本集 L,其余的去掉标记作为无标记样本集 U。4个算法在9个数据集上每个重复20次实验,最后取平均值。其中,对数据集Ecoli做了部分处理,将8个类标记归整为4个。对数据集Ionosphere做了部分处理,删除了其中两个离散属性,算上类标记属性一共33个属性。

表1 数据集描述

Tab.1 Descriptions of data sets

数据集	样本个数	属性个数	类别个数
Iris	150	4	3
Wine	178	13	3
Seeds	210	7	3
Vertebral Column	310	6	2
Haberman Survival	306	3	2
Pima Indians	768	8	2
Car-Evaluation	1728	6	4
Ionosphere	351	33	2
Ecoli	336	8	4

如表2所示,有标记样本为10%时4个算法在9个数据集上的分类正确率比较(平均值±标准差)。本文提出的NBSTDPCFCM算法整体上优于NBST算法、NBSTFCM算法和NBSTDPC算法。其中,在Seeds数据集上,NBSTDPCFCM算法的精度最高,这可能是因为通过高斯核函数的映射,该数据集的数据呈现高斯分布的状态,因此该数据集的三分类数据分布均匀,从而分类精度较高。在Vertebral column数据集和Pima indians diabetes数据集上,NBSTFCM算法和NBSTDPC算法精度都没有NBST算法高,而NBSTDPCFCM算法整体上与NBST算法相当。在数据集Ionosphere上,NBSTFCM的分类精度最高,NBSTDPCFCM算法相较NBST算法分类精度提高并不大,这是因为这3个数据集的数据种类不同于前面几个数据集,它们的类标记只有两类。如

果只通过 DPC 算法聚类，大量无标记样本被标注成为离群噪声点，这样不仅在训练集中只添加了少量标记后的无标记样本，而且其置信度也不高，导致后期在迭代训练朴素贝叶斯分类器的过程中错误样本累积，使得样本集原始数据结构发生改变。在数据集 Car-evaluation 和数据集 Haberman survival 中，4 种算法的分类精度相差并不大，这或许是因为以上两个数据集的数据属性为离散性质的缘故。在实验中，使用 DPC 聚类对离散属性的数据进行聚类后，数据会呈现出阶梯状的分布，每一个阶梯上都会存在大量数据点，因此不利于聚类中心的选择，因而 3 种结合聚类的改进算法分类精度提升都不大。在实验中发现，对于数据集 Ionosphere，如果将它的 35 个属性都加入训练，则训练出的贝叶斯分类器分类效果并不好，在计算中很容易出现奇异矩阵。如果将它的两类离散性质的属性删除，则训练出的分类器分类精度将得到大幅度提高。这也说明了结合 DPC 聚类的自训练算法对有离散属性的数据集分类表现并不理想。

表 2　有标记样本为 10%时，4 个算法在 9 个数据集上的性能对比

Tab. 2　Comparison of performance of 4 algorithms on 9 datasets with 10% of labeled samples　　%

数据集	NBST		NBSTDPC		NBSTFCM		NBSTDPCFCM	
	平均分类正确率	标准差	平均分类正确率	标准差	平均分类正确率	标准差	平均分类正确率	标准差
Iris	68.105 590	9.391 505	78.714 285	10.106 574 0	77.142 857	5.453 888	79.255 473	5.236 299
Wine	64.096 385	15.075 490	81.000 000	6.826 533 7	74.698 795	11.919 490	84.125 000	4.647 748
Seeds	71.386 054	12.983 040	82.045 502	5.242 683 5	83.746 355	2.454 439	85.256 084	2.044 395
Vertebral column	72.006 920	4.680 748	67.301 043	3.180 845 5	66.297 578	2.404 785	74.936 042	4.078 734
Haberman survival	69.346 823	3.563 377	65.908 372	2.983 535 8	70.237 258	2.098 427	72.909 327	3.405 662
Pima indians	70.349 932	2.251 959	60.530 075	2.649 945 3	67.057 182	1.910 791	71.584 464	2.761 248
Ionosphere	76.329 156	7.820 356	81.432 566	10.345 801 0	85.132 101	5.026 841	79.369 451	5.301 896
Car-evaluation	73.131 832	2.558 402	75.813 974	5.301 890 2	71.036 012	2.032 014	74.329 851	4.030 129
Ecoli	73.399 558	6.329 125	77.418 941	5.031 207 8	75.265 416	3.069 026	79.418 426	2.078 035

如图 1 所示为 4 个算法在有标记样本数量为 10%时 F-measure 值的对比分析。F-measure 值是一种统计量，它是准确率(Precision)和召回率(Recall)的加权调和平均值，可以评价分类结果的准确性。F-measure 指标越大，则该分类器的分类准确性越高。F-measure 值是 IR(信息检索)领域的常用的一个评价标准，公式为：

$$F=\frac{2\times PR}{P+R}。\tag{6}$$

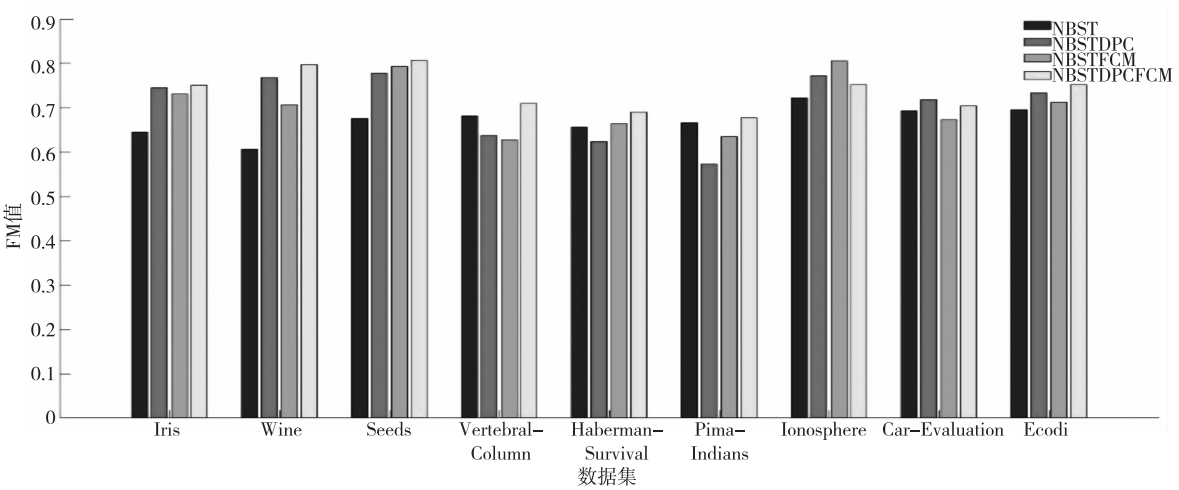

图 1　4 种算法在有标记样本为 10%时的 FM 值对比

Fig. 1　Comparison of the FM values of the four algorithms when the labeled sample is 10%

可以看出，NBSTDPCFCM 算法的 F-measure 值除了在数据集 Ionosphere 和数据集 Car-evaluation 上不是最高之外，在其他几个数据集上都比另外 3 个算法要高。同时，在这两个不是最高的数据集上，NBSTDPCFCM

算法的F-measure值都在0.70以上,这已经是一个相对较高的F-measure值了。

图2为NBSTDPC算法、NBST算法、NBSTFCM算法和NBSTDPCFCM算法在9个数据集上,随着有标记样本数量的增加算法分类精度的变化示意图。可以看出,在数据集Pima indians diabetes,Wine,Vertebral column,Haberman survival,Ecoli和Iris上,NBSTDPCFCM算法整体性能优于NBST算法、NBSTFCM算法和NBSTDPC算法。其中,在数据集Haberman survival上,虽然NBSTDPCFCM算法整体上优于其他3种算法,但分类精度提高却不多,而在Car-evaluation上更是只比NB算法分类精度大一些。这可以说明,NBSTDPCFCM算法在离散型数据的数据集上提高不大。同时在数据集Seeds上,当有标记样本比率为20%和30%时,NBSTDPC算法的分类精度要优于NBSTDPCFCM算法,但从整体上看,本文提出的NBSTDPCFCM算法要优于其他3种算法。在数据集Pima indians diabetes,Ionosphere,Vertebral column和Haberman survival上,随着有标记样本数量的增加,4种算法的分类精度都没有特别大的提高,这可能是因为这些数据集都是二分类的数据集,进行聚类分析时,很多无标记样本被判定为离群噪声点,导致训练集的质量有所下降。而其他如Iris,Wine,Seeds和Ecoli这4个数据集,本身数据集所含的样本数量并不多,在训练中容易出现过拟合现象。因此在训练初期所含有标记样本较少的情况下,通过用少量的有标记样本进行聚类分析并不能筛选出信息量大的有用无标记样本,4种算法的分类精度都较低,而在后面的迭代训练中,随着无标记样本的大量被标记加入训练,其他3种结合聚类的算法能够筛选出足够多的有用无标记样本加入训练,从而能够实现分类精度的大幅度提高。同时这也反映出,这4个数据集内在数据结构较为清晰,离群噪声点较少。对于数据集Ionosphere,虽然它是二分类的数据集,但是它所含的样本并不多。在这个数据集上,尽管结合聚类分析的改进算法相比于原始的NBST算法提高并不大,但3种算法也都达到了将近80%的分类精度。

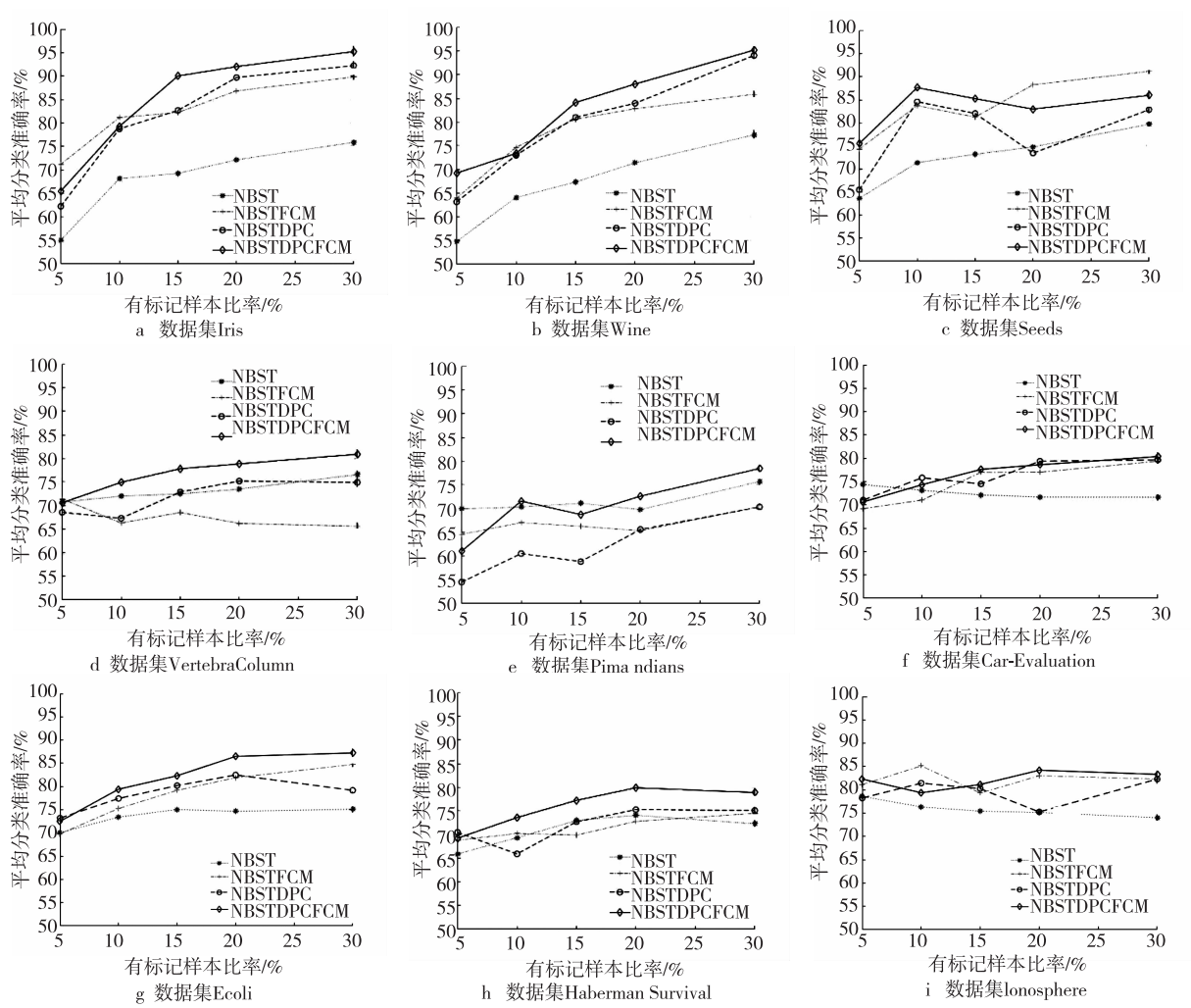

图2 4种算法在9个数据集上的平均分类正确率对比

Fig. 2 Comparison of average classification accuracy of 4 algorithms on 9 data sets

综上所述，本文考虑用 DPCFCM 聚类算法来做聚类分析，在迭代训练之前先通过计算样本点的局部密度和相对距离来确定样本集的整体数据结构，反映样本集的原始数据空间，再通过聚类隶属度来筛选有用的无标记样本加入训练集。这样，当所有无标记样本都根据其聚类隶属度作上相应标记后加入训练集训练时，训练出的贝叶斯分类器具有最好的泛化性。同时相比于结合其他聚类方法的自训练算法，本文算法具有更高分类精度。

4 结语

本文针对半监督自训练方法在迭代自训练过程中容易选出所含信息不大，置信度不高的无标记样本加入训练的问题，提出了 NBSTDPCFCM 算法。该算法在自训练过程中先对大量无标记样本进行 DPCFCM 聚类，筛选出聚类隶属度高，所含信息量较大的无标记样本，并做好相应标记后加入训练集中，同少量有标记样本共同训练朴素贝叶斯分类器。相对于传统半监督自训练方法，该算法能够很好地发现数据集中样本的内在数据结构，随着大量无标记样本被标记后加入训练，分类精度也比传统方法有所提升。同时该算法在实验时需要注意各项参数的设置，合理的参数设置会让该算法的性能有所提高。本文算法主要针对自训练方法如何选择高置信度的无标记样本加入训练集这个问题，而对于如何在迭代过程中剔除被错误标记的样本并没有涉及到。本文算法在时间复杂度以及在离散数据集上的分类精度还存在很大的提升空间，下一步的研究工作将主要针对此展开。

参考文献：

[1] ROSENBERG C, HEBERT M, SCHNEIDERMAN H. Semi-supervised self-training of object detection models [C]// IEEE Workshops on Application of Computer Vision. [S. l.]: IEEE Computer Society, 2005: 29-36.

[2] 刘建伟，刘媛，罗雄麟. 半监督学习方法[J]. 计算机学报，2015，38(8)：1592-1617.
LIU J W, LIU Y, LUO X L. Semi supervised learning method [J]. Chinese Journal of Computers, 2015, 38(8): 1592-1617.

[3] 周志华. 机器学习[M]. 北京：清华大学出版社，2016：293-294.
ZHOU Z H. Machine learning [M]. Beijing: Tsinghua University Press, 2016: 293-294.

[4] JOACHIMS T. Transductive inference for text classification using support vector machines[C]//Sixteenth International Conference on Machine Learning. [S. l.]: Morgan Kaufmann Publishers Inc, 1999: 200-209.

[5] 汪西莉，蔺洪帅. 最小代价路径标签传播算法[J]. 计算机学报，2016，39(7)：1407-1418.
WANG X L, LIN H S. Minimum cost path label propagation algorithm[J]. Chinese Journal of Computers, 2016, 39 (7): 1407-1418.

[6] 李南. 基于聚类假设的数据流分类算法[J]. 模式识别与人工智能，2017，30(1)：1-10.
LI N. Data flow classification algorithm based on clustering assumption[J]. Pattern Recognition and Artificial Intelligence, 2017, 30(1): 1-10.

[7] GAN H, SANG N, CHEN X, et al. An improved self-training for face recognition [C]//International Conference on Image & Graphics. [S. l.]: IEEE, 2013: 489-492.

[8] GAN H, SANG N, HUANG R, et al. Using clustering analysis to improve semi-supervised classification [J]. Neurocomputing, 2013, 101(3): 290-298.

[9] ZENG H J, WANG X H, CHEN Z, et al. CBC: Clustering based text classification requiring minimal labeled data [C]//IEEE International Conference on Data Mining. [S. l.]: IEEE, 2003: 443-450.

[10] 艾震鹏，王振友. 基于数据密度的半监督自训练分类算法[J]. 计算机应用研究，2019，21(5)：1-5.
AI Z P, WANG Z Y. Semi supervised self-training classification algorithm based on data density[J]. Application Research of Computers, 2019, 21(5): 1-5.

[11] WU D, SHANG M S, LUO X, et al. Self-training semi-supervised classification based on density peaks of data[J]. Neurocomputing, 2018, 275(1): 180-191.

[12] RODRIGUEZ A, ALESSANDRO L. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.

[13] 吕佳，黎隽男. 结合半监督聚类和数据剪辑的自训练方法[J]. 计算机应用，2018，38(1)：110-115.
LÜ J, LI J N. Self-training method combining semi supervised clustering and data editing[J]. Application Research of Computers, 2018, 38(1): 110-115.

[14] 刘伟涛，许信顺. 一种使用未标记样本聚类信息的自训练方法[J]. 计算机应用研究，2010，27(9)：3341-3344.
LIU W T, XU X S. A self-training method using clustering information of unlabeled samples[J]. Application Research of Computers, 2010, 27(9): 3341-3344.

[15] 赵芳，马玉磊. 自训练半监督加权球结构支持向量机多分类方法[J]. 重庆邮电大学学报(自然科学版)，2014，26(3)：404-408.
ZHAO F, MA Y L. Multi-Class classification based on seif-training semi-supervised weighted sphere structuied support vector machine[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edi-

tion),2014,26(3):404-408.

[16] 谢娟英,高红超,谢维信. K 近邻优化的密度峰值快速搜索聚类算法[J]. 中国科学:信息科学,2016,46(2):258-280.
XIE J Y,GAO H C,XIE W X. K nearest neighbor optimization density peak fast search clustering algorithm[J]. Scientia Sinica Information:Informations Science,2016,46(2):258-280.

[17] 刘沧生,许青林. 基于密度峰值优化的模糊 C 均值聚类算法[J]. 计算机工程与应用,2018,21(5):1-6.
LIU C S,XU Q L. Fuzzy C means clustering algorithm based on density peak value [J]. Computer Engineering and Applications,2018,21(5):1-6.

Self-Training Algorithm Combined with Density Peak Optimization Fuzzy Clustering

LUO Yunsong[1], LÜ Jia[1,2]

(1. College of Computer Science and Information Sciences, Chongqing Normal University, Chongqing 401331;
2. The Engineering & Technology Research Center of Digital Agriculture Service, Chongqing 401331, China)

Abstract: [**Purposes**]In order to explore the distribution of data sets before iterative self-training, the unlabeled samples with large amount of information and high confidence should be taken into the training set, and the initial classifiers are given higher accuracy and the generalization of self-training method is improved. [**Methods**]Basing on the clustering hypothesis, it first clusters the unlabeled sample set with the density peak clustering. After the clustering centers are selected out artificially, the new cluster centers are used as the initial cluster centers for fuzzy clustering. Hence the useful unlabeled samples are selected out. [**Findings**]By using the density peak optimization fuzzy clustering algorithm, the samples with large amount of information and high confidence are selected out and added into the training set, so that a classifier with stronger generalization and higher classification accuracy is obtained. [**Conclusions**]The experimental results show that the improved self-training method can quickly find the original spatial structure of the data sets, and find out the useful unlabeled samples to join the training set. Compared with the self-training method combined with other clustering algorithms, our algorithm can obtain better accuracy.

Keywords: semi-supervised; self-training; density peak optimization fuzzy clustering; clustering hypothesis

(责任编辑 黄 颖)