基于BvSBHC的主动学习多类分类算法

曹永锋[1] 陈 荣[2] 孙 洪[2]
(贵州师范大学数学与计算机科学学院 贵阳 550001)[1] (武汉大学电子信息学院 武汉 430079)[2]

摘 要 对尽量少的样本进行人工标注并获得较好的分类性能是图像分类应用的一个关键问题。针对标注样本选择,提出了一种综合样本不确定性度量和代表性度量的主动学习样本选择准则。基于最优标号和次优标号(Best vs. second-best,BvSB)的主动学习方法构建不确定性度量,利用分层聚类(Hierarchical Clustering,HC)方法得到数据集的分层聚类树,然后依据聚类树结构和已标注样本在其中的分布信息定义每个未标注样本的代表性度量。将新方法与随机样本选择以及BvSB主动学习方法进行了比较,对1个光学图像集和1个全极化SAR数据集分类问题的实验结果显示,新方法性能稳定,优于其他两种方法。

关键词 主动学习,分层聚类,图像分类
中图分类号 TP181 **文献标识码** A

Multi-class Image Classification with Best vs. Second-best Active Learning and Hierarchical Clustering

CAO Yong-feng[1] CHEN Rong[2] SUN Hong[2]
(School of Mathematics and Computer Science, Guizhou Normal University, Guiyang 550001, China)[1]
(School of Electronic Information, Wuhan University, Wuhan 430079, China)[2]

Abstract Using the least manually labeled samples to train a good classifier is a key problem in image classification. Aiming at selecting these samples for labeling, this paper proposed a criterion combining two different measures of samples: uncertainty of classification and representativeness. The best vs. second-best (BvSB) method is used to get the measure of uncertainty. The dataset is first hierarchically clustered and then the measure of representativeness of each unlabeled sample is defined based on the structural information of clusters and the distribution information of those labeled samples. The proposed method was compared with the random-selection method and BvSB method on an optical image dataset and a fully-polarimetric synthetic aperture radar (SAR) image dataset. The results show that it has stably better performance.

Keywords Active learning, Hierarchical clustering, Image classification

1 引言

多数图像监督分类算法需要充足的带有标号的训练样本来学习模式/模型。在实际中,对大量图像样本进行标注非常困难[1]。首先,标注工作非常枯燥、耗时。其次,对于一些特殊图像如合成孔径雷达(Synthetic aperture radar,SAR)图像,普通用户对内容进行判读非常困难,需要借助同一场景的高分辨率光学图像或者由专家来完成标注。近年来,自动/半自动标注方法得到了一定的发展[2-5],这些方法虽然用计算机代替了人工进行标注,释放了部分标注的劳动负担,但其实现前提仍然是基于部分已标注样本学习分类器,而且其准确率往往有限,若将含有错误标注号样本作为训练样本,可能会对分类器性能带来一些意想不到的影响。因此,如何对尽量少的样本进行人工标注,并获得较好的分类性能也成为图像分类应用的一个关键问题。

为了解决标注困难带来的有限样本情况下的分类问题,主动学习(Active learning)[1]已经成为机器学习和模式识别领域的研究热点。虽然到目前为止主动学习方法在理论上没有进行严密论证,但是从一些实际应用中,我们可以看到其有效性。例如,在某些具有稀疏解的学习方法中,以SVM为例,SVM的分类面的确定只与支持向量有关,因此,向现有训练集中增加非支持向量样本,将不会改变当前的分类面模型。从这个例子可看出,标注某些样本可能只是在做无用功,因为其对分类模型的改变不起作用或者作用微弱。这为在学习过程中选择采用主动学习样本提供了一个直观动机。

在基于SVM的机器学习问题中,Tong等[6]提出的主动学习方法已经得到了广泛的认可和应用。在该主动学习方法中,那些最靠近当前分类面的样本被认为是类别最模糊、最具信息性的样本,因此这些样本被要求标注并加入到当前训练样本集中。该方法的主动学习准则存在两个问题:1)每轮迭

到稿日期:2012-10-17 返修日期:2013-02-25 本文受国家自然科学基金(41161065,40901207)资助。
曹永锋(1976-),男,博士,副教授,主要研究方向为图像分析与模式识别、SAR图像解译,E-mail:yongfengcao.cyf@gmail.com;陈 荣(1983-),男,博士生,主要研究方向为图像检索、机器学习;孙 洪(1954-),女,博士,教授,主要研究方向为信号处理、SAR图像处理。

代中只能选择1个新样本，如果选择多个样本，无法保证这些选出的样本是最优的；2）只适合两类SVM，针对多类分类问题（该分类问题需要由多个分类器组合完成，因此，存在多个分类面）不再适用。基于熵的准则可以较好用于多类分类问题，但是当类别数量较多时，熵往往不能很好地代表样本的分类的不确定性[2,7]。Joshi等提出的基于最优标号和次优标号（Best vs. second-best，BvSB）的主动学习准则[7]进一步改善了这个问题。以上方法都仅考虑了待标注样本分类不确定性，而没有考虑该样本的代表性。针对以上问题，本文将主动学习准则分为两部分，一部分为分类不确定性度量，另外一部分为代表性度量。不确定性度量用来度量样本对于当前分类模型的信息性；代表性度量主要用来度量已选择的样本之间的冗余性。基于此，提出了一种新的BvSB和分层聚类（Hierarchical Clustering）的主动学习准则BvSBHC，并构造了相应的主动学习算法。

2 基于BvSBC的主动学习算法

2.1 BvSB准则

在多类分类问题中，常用熵来作为不确定性度量选取样本：

$$ENT^* = \arg\max_{x_i \in U} -\sum_{y_i \in Y} p(y_i|x_i)\log p(y_i|x_i) \quad (1)$$

其中，U为未标注样本集，Y代表所有可能取得的类别标号。样本的熵越高，说明样本的类别不确定性越高。当分类问题中的类别数量为2时，基于熵的主动学习准则退化为选择属于正类的概率最接近0.5的样本选择准则。基于熵的主动学习样本选择准则在多类学习问题中存在一个问题：熵的值很容易受到那些不重要类别的概率的影响[2,7]。Joshi等提出的主动学习样本选择准则BvSB[7]很好地解决了这个问题。在BvSB准则中只考虑样本分类可能性最大的两个类别，忽略其他对该样本的分类结果影响较小的类别。将样本x_i的最优标号和次优标号的概率分别记为$p(y_{Best}|x_i)$和$p(y_{Second_Best}|x_i)$，该准则可以表示如下：

$$BvSB = \arg\min_{x_i \in U}(p(y_{Best}|x_i) - p(y_{Second\_Best}|x_i)) \quad (2)$$

当分类问题的类别数量等于2时，BvSB准则退化为选择属于正类的概率最接近0.5的样本选择准则。

在BvSB准则中存在一个问题，即样本选择的过程中只考虑了样本的分类不确定性，而忽视了样本的代表性。以图1中例子进行说明。在图1中，给出了SVM分类器的分类面，红色的圆点和蓝色的圆点分别代表两类训练样本，黄色的圆点a和b为两个新样本。根据BvSB准则，a离当前分类面的距离近一些，因此a属于正类和负类的概率差相对于b要小一些，应该选择样本a加入到训练样本集。但是在a附近已有一个训练样本，所以即使样本a加入到训练样本集，也不会给分类面带来明显影响。从这个角度上说，样本a的信息量很低；样本b的概率差虽然比样本a小，但是样本b一旦加入到训练集中，对分类面的可能影响将比样本a大，因此，相对来说样本b对当前分类模型的信息量更大一些。

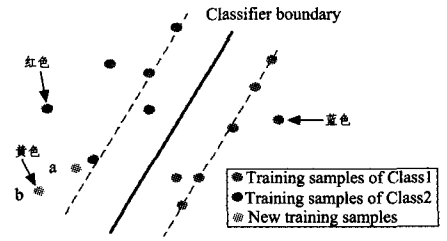

图1 BvSB准则存在的问题

2.2 BvSBHC度量

从以上对BvSB准则的分析可看出，在样本选择的过程中，仅考虑样本的分类不确定性还不够，同时也应考虑样本的代表性。代表性可从两个方面来体现：

1）从与已选出的训练样本集之间的关系来讲，新选出的样本与已有样本之间应该具有低的信息冗余。

2）从与整个未标注样本集之间的关系来讲，新选出的样本应具有较高的代表性，即该样本一旦被加入到训练样本集中，能够使训练样本集更好地代表整个未标注样本集。

基于这个考虑，本文将主动学习样本选择准则划分为两个部分，一部分是不确定性度量，另外一部分是代表性度量。这样定义一个新的样本选择准则：

$$BvSBHC = \arg\min_{x_i \in U}(S(x)) \quad (3)$$

$$S(x) = \alpha \cdot m_u(x) + (1-\alpha)m_r(x)$$

其中确定性度量$m_u(x)$基于BvSB准则得到，即

$$m_u(x) = p(y_{Best}|x) - p(y_{Second\_Best}|x) \quad (4)$$

代表性度量$m_r(x)$使用2.3节中式(5)、式(6)得到。参数α是一个用户定义的参数，取值在0和1之间，用来调整对两种不同度量的侧重。用户可以根据数据情况自行设定。

2.3 代表性度量

这里用样本在特征空间的距离关系衡量样本间的相关性，即如两样本在特征空间中彼此接近，则认为它们彼此相似且两者之间的信息冗余很大。如已选择了其中一个作为训练样本，则另外一个被认为代表性不高。从此角度出发，借鉴Dasgupta等[8]提出的基于分层聚类的主动学习思想，这里采用基于average linkage的分层聚类方法[9]获取代表性度量。

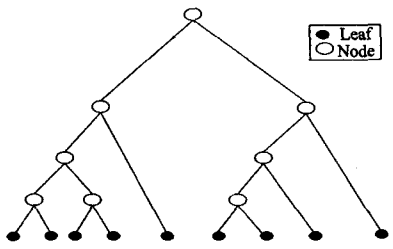

图2 分层聚类树结构

对未标注样本集进行分层聚类：数据集中的所有独立样本作为叶子，设共有n个叶子，将每个叶子看作一个聚类；计算n个聚类的两两间距离，将距离最小的2个聚类合并成1个新的聚类，这时，总的聚类数变为$n-1$个。这里average linkage是指在计算2个聚类之间的距离时，采用2个聚类中包含的所有样本两两之间的距离的平均值；计算$n-1$个聚类

的两两间距离，从中选择 2 个距离最小的聚类合并，这时，总聚类数变为 $n-2$ 个；重复合并过程，直到所有聚类都合并成一个聚类，这样就生成了一个分层结构的二叉聚类树(倒立的树，见图 2)。

定义任意节点所包含的训练样本的比例为

$$r_{select}(i)=\frac{N_{select}(i)}{N_{total}(i)} \tag{5}$$

式中，$N_{select}(i)$为节点 i 中当前包含的训练样本的数量，$N_{total}(i)$为节点 i 中包含的所有样本的数量。由于同一个父节点下的样本可以看作是比较相似的，一个样本所在的父节点的 r_{select}越大，说明该节点下包含的训练样本的比例越高，那么如果再从该节点下选择训练样本，则会与现有训练样本集有较高的冗余，该样本被认为具有低代表性。因此，用样本 x 所在的父节点的 r_{select}作为该样本的代表性度量，以 $fa(x)$代表取样本 x 的父节点，则有

$$m_r(x)=r_{select}(fa(x)) \tag{6}$$

理论上讲，非线性 SVM 降低了数据流形的复杂度，因此，在构建分层聚类树结构时，我们在由该函数重构得到的希尔伯特空间中计算样本之间的距离。

2.4 基于 BvSBHC 的主动学习算法

基于 BvSBHC 准则，建立主动学习算法，其伪码流程见图 3。算法迭代结束后的标注样本集可以用来训练最终分类器对未知标号样本进行分类。由于主动学习算法是基于 SVM 分类器的，因此最终分类器的最佳选择是 SVM。

```
Inputs:
U: unlabeled dataset; L: labeled training dataset;
T: Tree structure of hierarchical clustering;
Batch_size: number of samples selected in each iteration;
Max_ite: number of iteration; Seed: size of initial training dataset.
%Training dataset initialization
L=∅
for i=1 to Seed do
  Pick a random point z from U, query z's label l
  L←L ∪ z,  U←U\z
end for
Update r_select for all notes in T
%Sample selection iteration
for i=1 to Max_ite do
  train SVMs using L
  for each sample z ∈ U do
    Calculate labeling probability p(y_i|z) for all possible labels
    Calculate m_u(z)= p(y_Best|z)-p(y_SecondBest|z)
    Calculate m_r(z)= r_select(fa(z))
    Calculate S(z)=α · m_u(z)+(1-α) · m_r(z)
  endfor
  Sort the S of all samples in U, and query the labels of bach_size
samples {z_new} with the smallest S
  L←L ∪ {z_new}, U←U\{z_new}
  Update r_select for all nodes in T
end for
```

图 3　BvSBHC 主动学习算法伪码流程

3　分类实验与分析

为验证本文提出的 BvSBHC 主动学习算法的有效性，我们在美国邮政 USPS 手写体光学图像集[10]和一个全极化 SAR 图像上进行了分类实验，分别从总体分类准确率和各个类别各自的分类准确率两个方面进行评价，并与随机样本选择(Random)、BvSB 样本选择这两种方法的性能进行比较。采用 LIBSVM[11]作为实验中 SVM 的实现。

3.1　USPS 光学图像集上的实验

在 USPS 手写体数字图像集中，共包含有阿拉伯数字 0－9这 10 个类别，每个类别中的图像有 1100 幅，共 11000 幅，每张图像大小为 16 * 16，为 8 位灰度图像。从每类图像中选 600 幅组成一个大小为 6000 的未标注样本集，每类剩下的 500 幅图像组成大小为 5000 的测试图像集。直接将图像中每个像素的灰度值排列起来，组成一个 256 维的特征向量。实验中采用 RBF 核($\gamma=0.01$)，$Seed=100$(每类 10 个)，$Batch\_size=10$，$Max\_ite=41$，α 取值可变(文中仅给出了取值为 0.1 和 0.5 的结果)。分别统计了在每次迭代中各种样本选择方法下的整体分类准确率，以及在迭代停止之后，各种方法在各个类别的分类准确率上的差别(见图 4、图 5)。

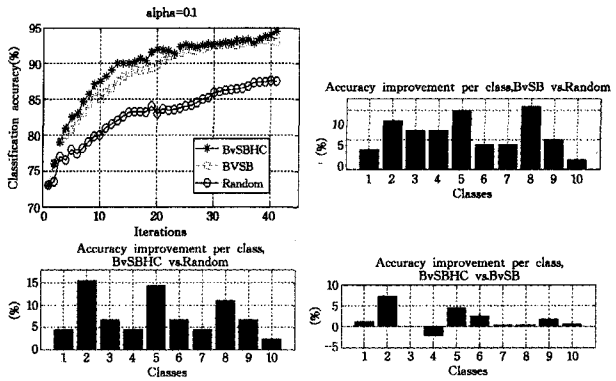

图 4　USPS 上分类性能(Alpha=0.1)

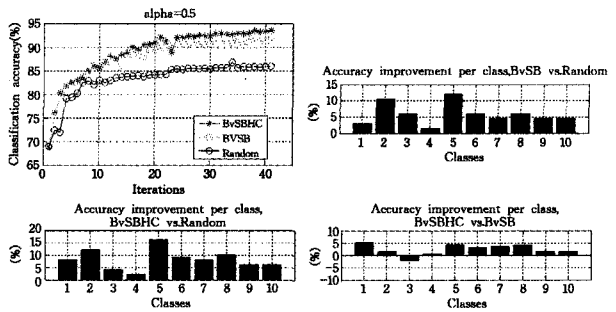

图 5　USPS 上分类性能(Alpha=0.5)

BvSB 和 BvSBHC 都比随机选择的分类性能好。在文中只列出了几组实验结果，在我们已经进行的所有实验中，BvSB 的性能都要比随机选择好(性能高出 6%左右)，这与文献[7]中的结果是吻合的。

对比 BvSB 和 BvSBHC 的性能曲线，我们发现，在实验中，在曲线的前端部分，无论 α 如何取值，两种方法的性能都比较相似(这可能因为在前几次迭代中，已有的训练样本数目相对数据集内总体样本数目来说非常小，致使代表性度量没能发挥作用)。在曲线中后端，BvSBHC 的性能较稳定，优于 BvSB：当 α 取值较大时(如 0.9～1，此时 BvSBHC 主要考虑不确定性度量)，两种方法的性能相差不多；当 α 取值中等时(如 0.5 附近)，BvSBHC 的性能相对 BvSB 有较明显的提升(2%左右)；当 α 取值较小时(如 0～0.1，此时 BvSBHC 主要考虑代表性度量)，两者的性能差距也比较小。BvSBHC 的整体分类性能差于 BvSB 的情况仅在很少的迭代中发生。

在 *Batch_size* 参数的选择上，暂时无法从当前的实验结果中给出结论，需要进一步进行大量实验。从直观上说，

Batch_size 越小，BvSBHC 的性能应该越好。这是因为本文方法在每次迭代内选择多个样本的时候，并没有考虑各样本之间信息的冗余性，所以 *Batch_size* 越小，每次迭代内选择的样本就越少，这样造成的样本冗余也应该会越小。

3.2 全极化 SAR 图像上的实验

采用由 ALOS PALSAR 获取的北美地区北卡罗莱纳州华盛顿县全极化数据以及美国地质局 USGS 提供的相应的地物覆盖图，选取了两幅子图(两幅子图对应的 ground truth 见图 6)。

其中，子图(a)的大小为 1070 * 1236，子图(b)的大小为 989 * 1036。出于简化考虑，我们仅考虑水、湿地、林地、耕地这 4 类，而不考虑除此以外的其他类别。实验中所使用的特征包括：Sinclair 散射矩阵(6 维)，协方差矩阵(9 维)，相干矩阵(9 维)，各通道极化强度比率、相位差及相关特征(7 维)，Pauli 分解(6 维)，Freeman 分解(6 维)，Huynen 分解(9 维)，H/Alpha/A 分解(6 维)，所有特征组成一个 52 维向量。实验基于图像小块进行分类。将图像分为不重叠的子块，子块的大小为 5 * 5，由于上面的特征提取方法都是针对像素的，因此这里用子块内包含的所有像素的特征平均值作为该子块的特征。将子图(b)作为未标注样本集(34534 个样本)，子图(a)作为测试样本集(47288 个样本)。实验中采用 RBF 核(γ =0.01)，*Seed*=800(每类 200 个)，*Batch_size*=40，*Max_ite*=41，α 取值可变。

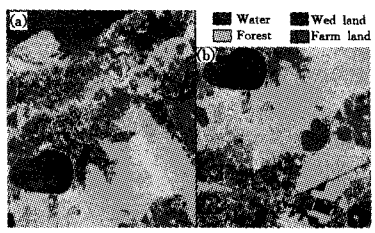

图 6 全极化 SAR 数据的 ground truth 图

由全极化 SAR 数据集实验结果得出的结论与由光学数据集实验结果得到的结论相似。不同的是在 SAR 数据集合上的分类性能曲线波动变化更大(相比光学图像，SAR 图像中类别具有更大的复杂性)，BvSBHC 和 BvSB 相对于 Random 的性能提升了 10%左右。图 7 给出了 α 取值为 0.4 时的 SAR 图像分类结果。

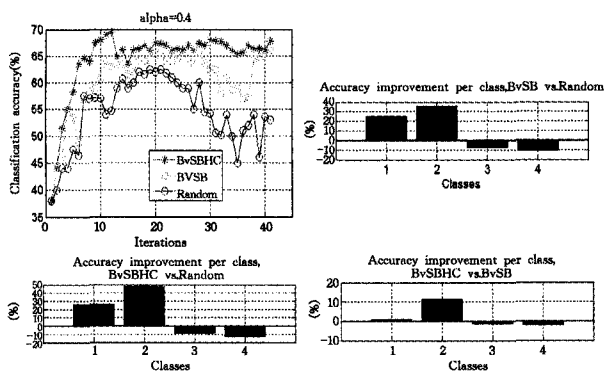

图 7 全极化数据上的分类性能(Alpha=0.4)

结束语 主动学习期望在较少样本的情况下获得较好的学习性能。在我们的实验中，相对随机地选取样本进行标注，BvSBHC 和 BvSB 都有很大的分类性能提升(7%～10%)，说明了主动学习的有效性和实际应用价值。通过引入样本的代表性度量，在绝大多数迭代次数上，本文方法 BvSBHC 的稳定性都优于 BvSB。我们认为样本代表性的引入固然应该提高分类的性能，但并不可能使分类器的性能发生质的飞跃，所以，从提升 1%～2%左右正确率的实验结果来看，本文提出的主动学习方法是有效的。

不确定性度量和代表性度量间的权重系数 α 是本文方法中重要的参数，目前主要基于经验和不断试验来择优选择。寻找一个根据数据集内在结构信息来确定参数，或者在迭代过程中自适应调整参数的方法是未来的研究目标之一。此外，采用不同策略组合、不确定性度量和代表性度量，以及采用可以得到更能代表数据集内在本质结构的聚类方法，可能会进一步提高性能。

参 考 文 献

[1] Settles B. Active Learning Literature Survey, Computer Science Technical Report 1648[R]. University of Wisconsin-Madison, USA, 2008: 3-4

[2] 陈荣，曹永锋，孙洪. 基于主动学习和半监督学习的多类图像分类[J]. 自动化学报，2011，37(8)：954-962

[3] 李志欣，施智平，李志清，等. 融合语义主题的图像自动标注[J]. 软件学报，2011，22(4)：801-812

[4] Carneiro G, Chan A B, Moreno P J, et al. Supervised learning of semantic classes for image annotation and retrieval[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2007, 29(3): 394-410

[5] Lienou M, Maitre H, Datcu M. Semantic Annotation of Satellite Images Using Latent Dirichlet Allocation[J]. IEEE Geoscience and remote sensing letters, 2010, 7(1): 28-32

[6] Tong S, Chang E. Support vector machine active learning for image retrieval[C]// Proceedings of ACM Multimedia. Ottawa, Canada, 2001

[7] Joshi A J, Porikli F, Papanikolopoulos N. Multi-class active learning for image classification[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2009: 2372-2379

[8] Dasgupta S, Hsu D. Hierarchical Sampling for Active Learning [C]// Proceedings of the 25th international conference on Machine Learning (ICML). 2008, 307: 208-215

[9] Webb A R. Statistical Pattern Recognition (Second Edition) [M]. John Wiley & Sons Ltd., 2002

[10] Asuncion A, Newman D J. UCI machine learning repository [OL]. http://achive.ics.uci.edu/ml/datasets.html

[11] Chang C C, Lin C J. LIBSVM: a library for support vector machine [OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/