基于主动学习和半监督学习的多类图像分类

陈 荣 [1] 曹永锋 [2] 孙 洪 [1]

摘 要 多数图像分类算法需要大量的训练样本对分类器模型进行训练. 在实际应用中, 对大量样本进行标注非常枯燥、耗时. 对于一些特殊图像, 如合成孔径雷达 (Synthetic aperture radar, SAR) 图像, 对其内容判读非常困难, 因此能够获得的标注样本数量非常有限. 本文将基于最优标号和次优标号 (Best vs second-best, BvSB) 的主动学习和带约束条件的自学习 (Constrained self-training, CST) 引入到基于支持向量机 (Support vector machine, SVM) 分类器的图像分类算法中, 提出了一种新的图像分类方法. 通过 BvSB 主动学习去挖掘那些对当前分类器模型最有价值的样本进行人工标注, 并借助 CST 半监督学习进一步利用样本集中大量的未标注样本, 使得在花费较小标注代价情况下, 能够获得良好的分类性能. 将新方法与随机样本选择、基于熵的不确定性采样主动学习算法以及 BvSB 主动学习方法进行了性能比较. 对 3 个光学图像集及 1 个 SAR 图像集分类问题的实验结果显示, 新方法能够有效地减少分类器训练时所需的人工标注样本的数量, 并获得较高的准确率和较好的鲁棒性.

关键词 主动学习, 半监督学习, 支持向量机, 图像分类

DOI 10.3724/SP.J.1004.2011.00954

Multi-class Image Classification with Active Learning and Semi-supervised Learning

CHEN Rong[1] CAO Yong-Feng[2] SUN Hong[1]

Abstract Most image classification methods require adequate labeled training samples to train classifier models. In real world applications, labelling samples are often very time consuming and expensive, especially for some special images, e.g. synthetic aperture radar (SAR) images. So the number of labeled samples is usually limited. In this study, we propose a novel image classification method based on SVMs, incorporating best vs second-best (BvSB) active learning and constrained self-training (CST). In this method, BvSB active learning is used to explore examples that are the most valuable to current classifier model for manual labelling. And CST is used to exploit useful information from examples that remain in the unlabeled dataset. With this new method, satisfying classification performance can be achieved while the human labelling load is low. We demonstrate results on 3 optical image datasets and a SAR image dataset. The proposed method gives large reduction in the number of human labeled samples as compared with random selection, entropy based active learning and BvSB active learning to achieve similar classification accuracy, and has little computational overhead and good robustness.

Key words Active learning, semi-supervised learning, support vector machines (SVM), image classification

图像分类是图像处理中的一个非常重要的应用. 大多数图像监督分类算法都是建立在统计模型的基础上, 用户需要对大量图像样本进行人工标注, 然后由带有类别标号的训练样本训练得到该模型. 在实际应用中, 对大量图像进行标注是比较困难的[1]. 首先, 对整个图像集进行标注需要耗费大量的时间, 用户往往没有足够的耐性来完成整个样本集的标注; 其次, 对于某些比较复杂的图像, 例如合成孔径雷达 (Synthetic aperture radar, SAR) 图像, 普通用户对其内容进行判读是比较难的, 通常需要借助同一场景的高分辨率光学遥感图像或者通过有经验的专家来完成. 正是由于标注上的困难, 使得在图像分类中能够获得的训练样本是比较有限的. 然而, 在小训练样本情况下, 分类器的性能可能受到很大影响. 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中的一个关键问题. 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) 已经成为机器学习和模式识别领域的研究热点. 在主动学习中, 学习器不再是被动地接受由用户提供的训练样本, 而是主动要求用

收稿日期 2010-04-01 录用日期 2010-11-08
Manuscript received April 1, 2010; accepted November 8, 2010
国家高技术研究发展计划 (863 计划) (2007AA12Z155), 国家自然科学基金 (40901207), 测绘遥感信息工程国家重点实验室专项科研经费, 中央高校基本科研业务费专项资金资助
Supported by National High Technology Research and Development Program of China (863 Program) (2007AA12Z155), National Natural Science Foundation of China (40901207), Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS) Special Research Funding, and the Fundamental Research Funds for the Central Universities
1. 武汉大学电子信息学院信号处理实验室 武汉 430079 2. 贵州师范大学数学与计算机科学学院 贵阳 550001
1. Signal Processing Laboratory, Electronic Information School, Wuhan University, Wuhan 430079 2. School of Mathematics and Computer Science, Guizhou Normal University, Guiyang 550001

户对那些对于当前分类器模型最有价值的样本进行标注, 并将这些带有类别标号的样本添加到训练样本集, 对分类模型进行重新训练. 通过迭代的方式, 对分类器模型进行更新. 理论上的结果表明, 在获得相似的分类准确率的情况下, 主动样本选择相对于随机选择可以显著地减少所需的样本数[2]. 典型的主动学习图像分类框架如图 1 所示.

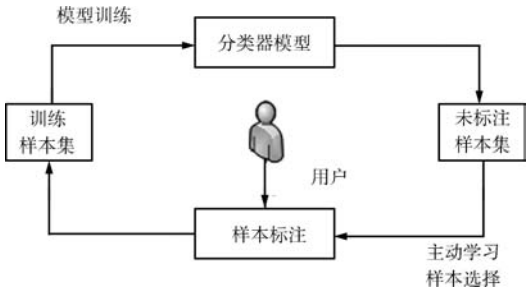


图 1　典型的主动学习图像分类框架

Fig. 1　A typical framework of image classification with active learning

近年来, 研究者对主动学习进行了大量的研究工作, 也提出了很多主动学习方法. Tong 等[3−4] 在基于支持向量机 (Support vector machines, SVMs) 的分类和检索中, 根据样本到当前 SVM 分类面的距离对样本进行采样 (Margin sampling, MS), 最靠近分类面的样本被认为是最具信息量的, 在下一次迭代中选择最靠近当前分类面的样本添加到现有训练样本集中; 基于委员会的采样 (Query by committee)[5−6] 也是比较常用的主动学习方法, 在这种方法中, 采用多个学习器对样本的类别进行判断, 选择那些分类结果最不一致的样本加入到当前的训练样本集; 在基于熵的不确定性采样 (Entropy based uncertainty sampling) 方法中, 首先计算样本属于每个可能的类别的概率, 根据得到的概率计算每个样本的熵, 熵越大表示该样本的分类不确定性越高, 在每次迭代中选择那些具有最大熵的样本添加到当前训练样本集. 此外, 还有基于 Fisher 信息矩阵 (Fisher information matrix) 的主动学习方法[7] 等. 其中, MS 方法是使用最广泛、具有较好性能的方法之一, 但是该方法只适用于二类 (Two-class) SVM 分类问题, 对于多类 (Multi-class) SVM 分类问题, 由于分类器是由多个二分类器联合而成, 每个二分类器都有各自的分类面, 这时, MS 主动学习方法将不再适用. 基于熵的主动学习方法虽然可以较好地用于多类分类问题中, 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不确定性. 鉴于这个问题, Joshi 等提出了一种基于最优标号和次优标号 (Best vs second-best, BvSB) 的主动学习方法[8], 该方法可以看成是 MS 主动学习方法在多类分类问题中的扩展, 在多类分类问题中获得了较好的性能. 为了尽可能地减少分类器训练过程中所需人工标注的数量, 同时获得较好的分类性能, 本文对半监督学习 (Semi-supervised learning, SSL) 中的自学习 (Self-training) 技术加以改进, 提出一种带约束条件的自学习方法 (Constrained self-training, CST), 并且在 BvSB 主动学习方法的基础上, 结合 CST 进一步对样本集里大量的未标注样本加以利用, 从而进一步提高分类器的分类性能. 最后, 将包含 BvSB 主动学习和 CST 半监督学习的新方法 (BvSB + CST), 结合 SVM 分类器进行图像分类实验, 并与 BvSB、基于熵的主动学习以及随机选择等 3 种样本选择方法的性能进行了比较. 通过对 3 个常用的光学图像集和 1 个 SAR 图像集上的分类结果进行分析, 我们的新方法能够有效地减少在训练过程中所需的人工标注的负担, 并取得较好的分类准确率.

1　BvSB 主动学习及带约束条件的自学习

1.1　BvSB 主动学习

首先简单介绍基于熵的主动学习方法. 设未标注样本集为 $\boldsymbol{U} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, $\boldsymbol{Y} = \{1, 2, \cdots\}$ 为所有可能的类别标号, 在由当前已经获得的训练样本集训练得到的分类模型下, 样本 $\boldsymbol{x}_i$ 属于各个类别的概率为 $p(y_i|\boldsymbol{x}_i)$, 则基于熵的主动学习样本选择准则可以表示为

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

熵越大的样本被认为是对于当前分类器来说分类结果最不确定, 最具信息量的样本, 用户对熵最大的这一部分样本进行人工标注, 然后将其添加到现有的训练样本集, 用更新后的训练样本集重新训练分类器模型. 在多类分类问题中, 熵往往不能很好地代表样本的不确定性. 有些具有较小熵的样本的分类不确定性相对于有些熵稍大的样本可能更高, 如图 2 所示.

从图 2 所示的例子可以看到, 在图 2 (a) 中, 样本属于类别 4 和类别 5 的概率都比较高, 并且比较接近, 这说明分类器无法对该样本属于类别 4 或者类别 5 作出明确的判断, 即该样本的分类不确定性较高, 在图 2 (b) 中, 样本仅仅在类别 4 上具有较高的概率, 这说明分类器对该样本的分类结果比较明确. 通过计算图 2 (a) 和图 2 (b) 中两个样本的熵, 结果却发现分类不确定性高的样本图 (图 2 (a)) 的熵小于分类结果较确定的样本图 (图 2 (b)) 的熵. 上面的这个问题是由于在多类分类问题中, 样本的熵会受到那些不重要的类别的影响 (值较小的那些 $p(y_i|\boldsymbol{x}_i)$). Joshi 等[8] 提出了一种更为直接的主动

学习样本选择准则 BvSB, 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.

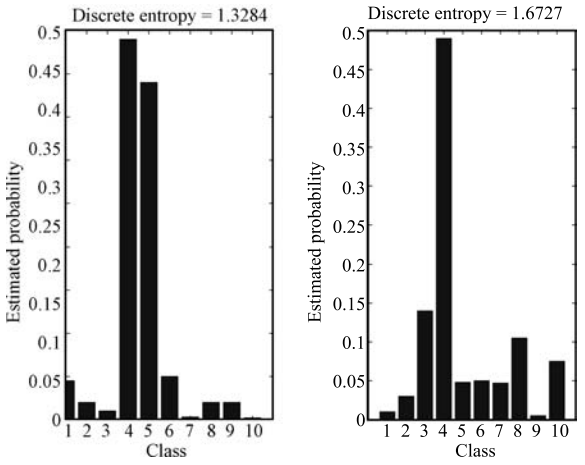


图 2 一个样本的分类不确定性与其对应的熵相背离的例子 (图中给出了在一个 10 类的分类问题中, 两个未标注样本通过当前分类器估计出的类别的概率分布)

Fig. 2 An illustration of why entropy can be a poor estimate of classification uncertainty (The plots show estimated probability distributions for two unlabeled samples in a 10-class classification problem.)

将样本 $\boldsymbol{x}_i$ 的最优标号和次优标号的概率分别记为 $p\left(y_{\text{Best}}|\boldsymbol{x}_i\right)$ 和 $p\left(y_{\text{Second-Best}}|\boldsymbol{x}_i\right)$, 该准则可以表示如下:

$$BvSB^* = \arg\min_{\boldsymbol{x}_i \in \boldsymbol{U}} \left(p\left(y_{\text{Best}}|\boldsymbol{x}_i\right) - p\left(y_{\text{Second-Best}}|\boldsymbol{x}_i\right)\right) \tag{2}$$

作为对样本的分类不确定性估计的一种贪婪近似, BvSB 准则也可以从另外一个角度来进行解释. 我们以一对一 (One-against-one) 形式的 SVM 分类器组为例, 设 $C_{i,j}\,(i,j \in \boldsymbol{Y})$ 为区分第 i 类和第 j 类之间的分类器. 如果一个未标注样本 $\boldsymbol{x}$ 的真实标号为 l, 那么一旦 $\boldsymbol{x}$ 经过标注, 并加入训练样本集, 将会影响那些用来区分第 l 类和其他类的分类器的分类面, 我们将这些分类器记为 $C_l = \{C_{(y,l)} | y \in \boldsymbol{Y}, y \neq l\}$. 由于并不知道 $\boldsymbol{x}$ 的真实标号, 我们只能用该样本的最优标号 y_{Best} 作为对其真实标号的估计, 这样, 样本 $\boldsymbol{x}$ 会影响的分类器为 $C_{y_{\text{Best}}} = \{C_{(y,y_{\text{Best}})} | y \in \boldsymbol{Y}, y \neq y_{\text{Best}}\}$, 对于集合 $C_{y_{\text{Best}}}$ 中的每个分类器 $C_{(y,y_{\text{Best}})}$ 而言, $\boldsymbol{x}$ 的分类不确定程度可以通过它属于该二分类器的正负类别的概率差 $p_{y_{\text{Best}}} - p_y$ 来表示, 这个差值可以作为 $\boldsymbol{x}$ 对于特定的分类器 $C_{(y,y_{\text{Best}})}$ 所具有的信息量高低的一个度量指标[8]. 通过最小化 $p_{y_{\text{Best}}} - p_y$, 即最大化分类不确定度, 我们可以得到 BvSB 准则:

$$\begin{aligned} BvSB^* = & \\ & \arg\min_{\boldsymbol{x}_i \in \boldsymbol{U}} \left(\min_{y \in \boldsymbol{Y}, y \neq y_{\text{Best}}} \left(p\left(y_{\text{Best}}|\boldsymbol{x}\right) - p\left(y|\boldsymbol{x}\right)\right)\right) = \\ & \arg\min_{\boldsymbol{x}_i \in \boldsymbol{U}} \left(p\left(y_{\text{Best}}|\boldsymbol{x}\right) - p\left(y_{\text{Second-Best}}|\boldsymbol{x}\right)\right) \end{aligned} \tag{3}$$

从分类边界的改变这个角度来说, BvSB 准则是一个有效的度量来选择那些对分类器的分类边界影响最大的样本. 当分类问题的类别数量等于 2 时, BvSB 准则退化为 MS 准则.

1.2 带约束条件的自学习

自学习是半监督学习中一个常用的技术. 在自学习中, 首先由少量的已经人工标注过的样本训练得到分类器, 然后通过这个分类器对未标注样本的标号进行判断. 通常, 那些分类结果最确定的未标注样本, 连同它们对应的由分类器预测得到的类别标号一起, 加入到当前的训练样本集. 用扩充后的训练样本集重新训练分类器, 对分类结果进行更新[9]. 典型的自学习框架如图 3 所示.

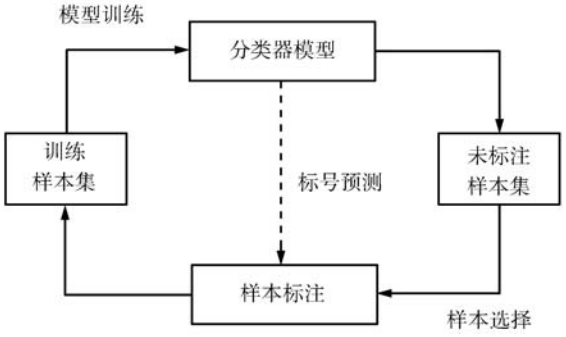


图 3 自学习典型框架

Fig. 3 A typical framework of self-training

自学习方法在模式识别问题中已经得到广泛应用, Riloff 等[10] 用自学习方法来识别语言中的主观名词, Rosenberg 等[11] 在图像的目标识别系统中采用自学习, 并证明了该半监督学习方法相对于现有的其他检测方法具有更好的性能.

在自学习中, 添加到训练样本集里的样本的标号不是由用户进行人工标注, 而是由当前的分类器预测得到的. 因此, 如果预测得到的标号与样本的真实类别标号不一致, 即预测错误, 那么该错误会在迭代过程中不断积累加强. 因此, 如何尽量减少在自学习中引入的标号错误, 也成为算法设计中的一个重要问题.

从直观上说, 如果选择那些在当前分类器下分类结果最明确的样本进行自学习, 引入错误标号的概率是最小的. 但是从样本所包含的信息量这个角

度来说, 这些分类结果最明确的样本所包含的信息量是非常低的, 对于当前分类面的影响极小. 因此, 将这些样本加入到训练样本集, 对分类模型的影响很小, 同时反而增加了分类器训练时的计算负担.

为了在样本的信息性和预测标号的准确性两者之间获得较好的平衡. 我们在自学习的基础上, 提出了带约束条件的自学习 (CST). 通过阈值的设置和空间位置关系这两个约束条件, 来提高自学习选出的样本的预测标号的准确性. 对于当前的每个未标注样本 $\boldsymbol{x}_u$, 具体如下:

1) 计算 $\boldsymbol{x}_u$ 属于各个类别的概率的最大值 (最优标号的概率)

$$p\left(y_{\text{Best}}|\boldsymbol{x}_u\right)=\max_{y_i\in\boldsymbol{Y}}p\left(y_i|\boldsymbol{x}_u\right)$$

2) 将训练样本集中所有由用户人工标注的样本子集记为 $\boldsymbol{S}_m$, 计算 $\boldsymbol{x}_u$ 到 $\boldsymbol{S}_m$ 中所有样本的距离, 寻找 $\boldsymbol{x}_u$ 的最近邻点 $NN\left(\boldsymbol{x}_u\right)=\arg\min_{\boldsymbol{x}\in\boldsymbol{S}_m}\text{dis}\left(\boldsymbol{x},\boldsymbol{x}_u\right)$, 并将该最近邻点的类别标号记为 y_{nn}

3) 约束条件

阈值约束:

$$p\left(y_{\text{Best}}|\boldsymbol{x}_u\right)\geq threshold \tag{4}$$

其中, $0\leq threshold\leq 1$ 为设置的阈值.

空间位置关系约束:

$$y_{\text{Best}}=y_{nn} \tag{5}$$

4) 设当前未标注样本集中满足约束条件 (4) 和 (5) 的子集为 $\boldsymbol{S}_{\text{satisfied}}$, 将 $\boldsymbol{S}_{\text{satisfied}}$ 中所有样本按照各自的最优标号的概率 $p\left(y_{\text{Best}}|\boldsymbol{x}_u\right)$ 进行排序, 选择 $p\left(y_{\text{Best}}|\boldsymbol{x}_u\right)$ 最小的 k 个样本, 连同其各自对应的最优预测标号 y_{Best} 一起, 加入到当前的训练样本集里.

从 CST 的具体步骤中可以看到, 上面的两个约束条件保证了通过自学习添加到训练集里的样本标号具有较高的正确率, 同时, 在选择样本的时候, 没有选择那些分类结果最确定的样本, 因此使得选出的样本对于当前的分类模型也具有一定的信息量, 在预测标号准确率和样本的信息量之间达到了一个较好的平衡.

2　基于 BvSB 主动学习和 CST 半监督学习样本选择的 SVM 多类图像分类算法

2.1　创新点及算法设计动机

考虑到构造训练样本集的标注负担, 我们在算法设计时主要从两个方面出发: 1) 对于选出的用于人工标注的样本必须是对于当前的分类模型而言最具信息量的, 以最大化人工标注的效率; 2) 对于剩余的大量未标注样本所包含的信息, 在不增加人工标注负担的情况下, 要进一步加以利用.

基于以上两点考虑, 我们提出了一种基于 BvSB + CST 样本选择的 SVM 多类图像分类算法. 其中, BvSB 主动学习被用来寻找那些最具信息量的样本, 提供给用户进行手工标注, CST 半监督学习用来对样本集中剩下的未标注样本中分类结果相对较确定且具有一定信息量的一部分样本进行自动标注, 进一步对训练样本集进行补充更新.

2.2　算法具体步骤

在本文提出的图像分类算法中, 主要包括初始样本选择及分类器模型训练、BvSB 主动学习、CST 半监督学习、分类器模型更新等几个关键步骤. 完整的算法框架如图 4 所示.

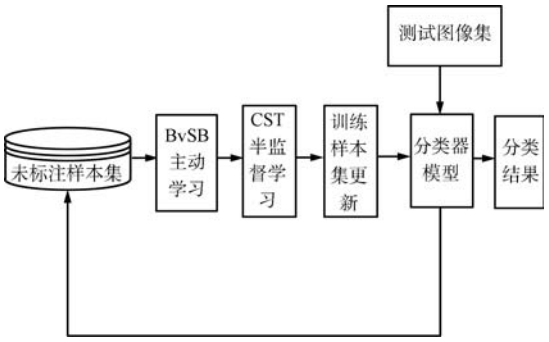


图 4　BvSB + CST 图像分类框图

Fig. 4　The framework of image classification with BvSB + CST

2.2.1　初始样本选择及分类器模型训练

将训练样本集和未标注样本集分别记为 $\boldsymbol{L}$ 和 $\boldsymbol{U}$, 在初始分类时, 从 $\boldsymbol{U}$ 中随机选择 K_{ini} 个样本, 由用户对其进行人工标注, 将该样本集合记为 $\boldsymbol{S}_{\text{ini}}$. 对训练样本集 $\boldsymbol{L}$ 和未标注样本集 $\boldsymbol{U}$ 进行更新: $\boldsymbol{L}=\boldsymbol{S}_{\text{ini}}$, $\boldsymbol{U}\rightarrow\boldsymbol{U}\backslash\boldsymbol{S}_{\text{ini}}$. 用训练集 $\boldsymbol{L}$ 训练 SVM 分类器, 同时, 对未标注样本集 $\boldsymbol{U}$ 中的样本的类别进行预测, 并计算其属于各个类别的概率 $p\left(y_i|\boldsymbol{x}\right)$, $y_i\in\boldsymbol{Y}$, $\boldsymbol{x}\in\boldsymbol{U}$.

2.2.2　BvSB 主动学习

BvSB 主动学习被用来寻找那些对当前分类模型最有价值的样本. 根据前面得到的未标注样本属于各个类别的概率 $p\left(y_i|\boldsymbol{x}\right)$, 通过 BvSB 度量准则从 $\boldsymbol{U}$ 中选出 K_{BvSB} 个样本, 记为 $\boldsymbol{S}_{\text{BvSB}}$, 由用户对 $\boldsymbol{S}_{\text{BvSB}}$ 中的样本进行人工标注.

2.2.3　CST 半监督学习

除了通过 BvSB 主动学习选择那些最有价值的样本外, 我们还通过 CST 半监督学习进一步挖掘剩下的未标注样本中的有用信息. 首先, 通过当前的分

类器, 计算未标注样本集 $\boldsymbol{U}$ 中所有样本的最优标号的概率 $p(y_{\text{Best}}|\boldsymbol{x})$, 然后, 计算 $\boldsymbol{U}$ 中样本在用户手工标注样本集 $\boldsymbol{S}_m$ 的最近邻点, 并记录其类别标号. 对 $\boldsymbol{U}$ 中符合约束条件 (4) 和 (5) 的样本按照其最优标号的概率大小从小到大进行排列, 选择前 K_{CST} 个样本, 记为 $\boldsymbol{S}_{\text{CST}}$, 将 $\boldsymbol{S}_{\text{CST}}$ 中的样本的最优标号作为各自的类别标号.

2.2.4 训练样本集及分类器模型更新

用新选出的样本对训练样本集和未标注样本集进行更新: $\boldsymbol{L} = \boldsymbol{L} \cup (\boldsymbol{S}_{\text{BvSB}} \cup \boldsymbol{S}_{\text{CST}})$, $\boldsymbol{U} \rightarrow \boldsymbol{U} \backslash (\boldsymbol{S}_{\text{BvSB}} \cup \boldsymbol{S}_{\text{CST}})$. 用更新后的训练样本集重新训练 SVM 分类器, 对整个分类过程进行迭代.

3 实验及相关分析

为了验证本文提出的基于 BvSB + CST 的图像分类算法的有效性和鲁棒性, 我们分别在 3 个光学图像集和 1 个 SAR 图像集上进行分类实验, 分别从测试集的总体分类准确率、每个类别各自的分类准确率、标注负担等 3 个方面对算法的性能进行评价, 并将我们的方法与基于随机选择 (Random)、基于熵的主动学习 (Entropy based active learning) 以及基于 BvSB 的主动学习 (BvSB) 这三种方法的性能进行比较.

采用 LIBSVM[12] 作为实验中 SVM 的实现. LIBSVM 采用一对一的方式来处理多类分类问题, 并且能够输出测试样本属于各个可能的类别的概率.

3.1 标准光学图像集上的分类实验

3.1.1 实验设置

我们从 UCI 数据库[13] 中选择了 3 个比较适合进行分类的图像集来进行实验. 其中包括英文字母数据集 (Letters)、美国邮政手写体数字图像集 (USPS), 以及另外一个手写数字图像集 (Pendigits). 3 个光学图像数据集的基本情况如表 1 所示.

我们在每个图像集上进行了 10 次分类实验, 实验的具体参数设置如表 2 所示. 表 2 中各个参数具体含义如下:

K_{ini}: 初始随机选择训练样本数量;

K_{BvSB}: 每次迭代中通过 BvSB 主动学习选出的样本数量;

K_{CST}: 每次迭代中通过 CST 选出的样本数;

$threshold$: 约束条件 (4) 中的阈值;

$Kernel$: SVM 分类器中使用的核函数;

Max_ite: 最大迭代次数 (初始样本选择作为第 1 次迭代).

表 1 3 个光学图像集的基本情况

Table 1 General information of 3 optical image datasets

	类别数量	特征维数	未标注样本集大小	测试集大小
USPS	10	256 (通过 PCA 降至 65)	6 000	5 000
Pendigits	10	16	2 000	5 100
Letters	26	16	10 000	10 000

需要说明的是, 在对实验中的参数设置时, 我们保证了在该参数下, 每次迭代中满足 CST 约束条件的未标注样本的数量大于等于 K_{CST}, 使得 CST 能够顺利进行.

3.1.2 分类准确率

我们对每个图像集上的分类准确率进行计算, 得到分类准确率随迭代次数的变化曲线, 并将 10 次实验得到的分类准确率变化曲线进行平均, 得到一个统计上的平均准确率变化曲线, 如图 5 所示. 在图 5 中, x 轴代表迭代次数, y 轴代表分类准确率. 通过对分类准确率变化曲线进行观察, 在迭代初期, 采用各种样本选择方法的分类性能相差不大, 这是由于在迭代初期, 训练样本的数量较少, 训练得到的分类器不是很准确, 在这种情况下, 各种样本选择方法都近似于随机选择. 随着迭代不断进行, BvSB + CST 的作用逐渐体现出来, 采用该方法的分类性能要优于其他三种方法. 当迭代次数相同时 (固定 x 轴), 采用 BvSB + CST 方法可以获得最高的分类准确率. 由于我们在每次迭代中, 各种不同的方法所需的人工标注的数量都是相同的, 这说明在相同的人工标注负担的情况下, BvSB + CST 方法相对于其他方法能够有效地提高分类准确率. 当获得相同的分类准确率时 (固定 y 轴), BvSB + CST 方法所需的迭代次数最少, 说明获得相同的分类准确率时, 该方法所需的人工标注数量更少. 需注意的是, 在 USPS

表 2 实验设置

Table 2 Experimental setup

	K_{ini}	K_{BvSB}	K_{CST}	$threshold$	$Kernel$	Max_ite
USPS	100	5	10	0.7	RBF ($\gamma = 0.01$)	51
Pendigits	50	5	15	0.6	RBF ($\gamma = 0.0001$)	51
Letters	260	10	10	0.6	RBF ($\gamma = 0.05$)	101

和 Letters 两个数据集上, 基于熵的主动学习和随机选择的性能差别不大, 这说明在这两个数据集上, 熵并不能很好地代表样本的分类不确定性, 这也从实验上说明了第 1.1 节中提出的问题. 在 3 个数据集上, 采用 BvSB + CST 方法的分类准确率都要高于采用 BvSB 主动学习方法的分类准确率, 这说明 CST 中选出的样本经过自动标注, 加入到训练样本集之后能够较有效地提高分类器的分类性能. 在 Letters 数据集上, BvSB + CST 相对于 BvSB 方法在分类准确率上提高不大, 这说明在该数据集上, 那些符合 CST 中的约束条件的样本包含的信息量相对比较有限, 对分类器的性能提升并不大.

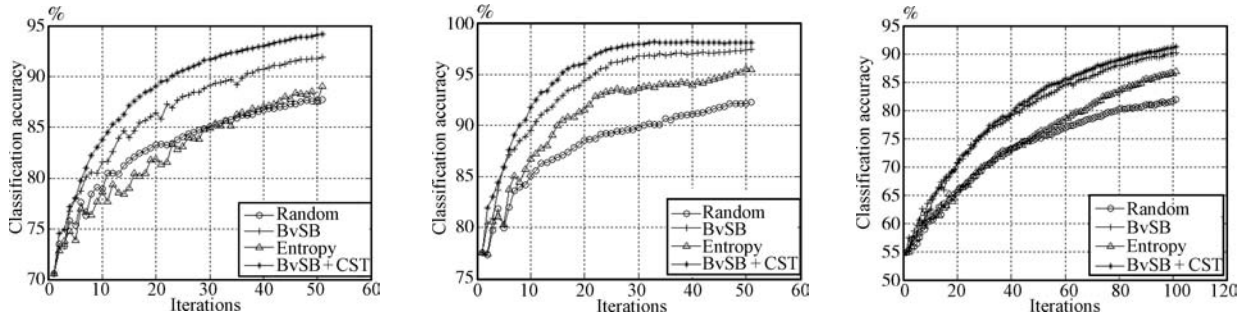
图 6～8 分别给出了在 3 个数据集上, 经过实验中设置的最大迭代次数后, 采用本文提出的 BvSB + CST 样本选择方法和其他 3 种方法在各个类别上分类准确率的比较. 图中的柱状代表采用 BvSB + CST 方法与各种对比方法在各个类别上的分类准确率的差, 正数代表 BvSB + CST 在该类别上的分类准确率高于该对比方法. 从图中可以看到, BvSB + CST 方法相对于其他三种方法在绝大多数类别上的分类准确率都有一定程度的提高. 其中, BvSB + CST 相对于 Random 和 Entropy 两种方法的性能提升幅度较大, 相对于 BvSB 方法的性能提升相对较小. 这说明在 BvSB + CST 方法中, BvSB 主动学习能够有效地选择那些最具信息量、对当前的分类器最优价值的样本, 而 CST 半监督学习选出的样本所包含的信息量相对较小, 但由于 CST 选出的样本是自动标注的, 没有增加人工标注负担, 因此 CST 仍然可以认为是有用的.

3.1.3　人工标注负担

为了说明 BvSB + CST 方法在减少人工标注负担上所起的作用, 我们分别比较了在达到相同的分类准确率时, 各种方法所需的迭代次数. 表 3 给出了对 USPS 数据集的分类实验中, 采用各种不同的样本选择方法达到相同的分类准确率时所需的迭代次数. 我们以 BvSB + CST 方法作为基准进行比较, 例如, 采用 BvSB + CST 方法迭代 14 次所能达到的分类准确率, 采用 Random 方法需要迭代 39 次, 采用 Entropy 方法需要迭代 36 次, 采用 BvSB 方法需要迭代 20 次才能获得. 表 3 中的 "–" 表示该方法即使达到实验中设定的最大迭代次数时, 仍不能达到相应的分类准确率.

3.1.4　关键参数 *threshold* 对分类性能的影响

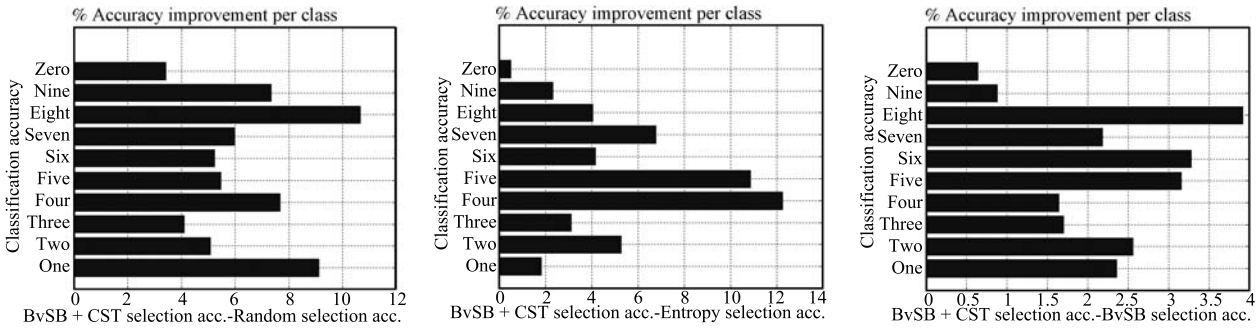
在 BvSB + CST 方法中, 包含一个重要的参数 *threshold*. 它控制着在 CST 中用来自动标注的样本



(a) USPS　　(b) Pendigits　　(c) Letters

图 5　分类准确率

Fig. 5　Classification accuracies on USPS, Pendigits, and Letters datasets

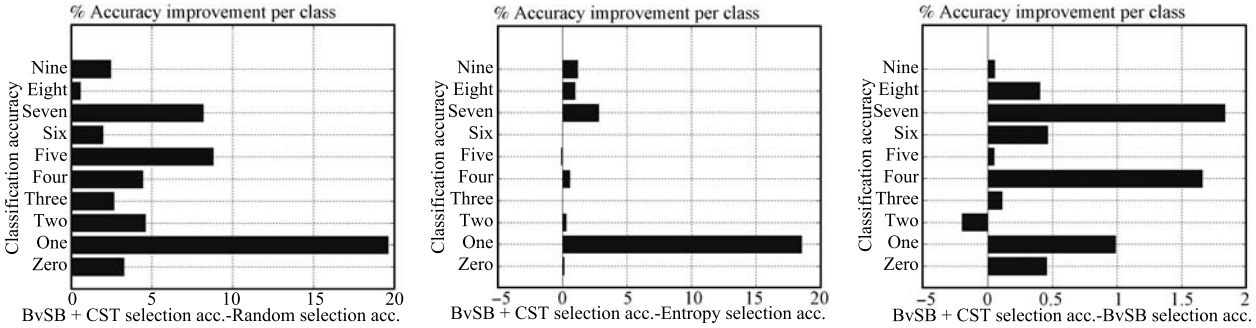


(a) BvSB + CST 与 Random 性能比较　　(b) BvSB + CST 与 Entropy 性能比较　　(c) BvSB + CST 与 BvSB 性能比较

(a) BvSB + CST vs. Random　　(b) BvSB + CST vs. Entropy　　(c) BvSB + CST vs. BvSB

图 6　BvSB + CST 方法与其他三种方法在 USPS 图像集各个类别上的分类准确率比较

Fig. 6　Classification accuracy comparison for each class in USPS dataset

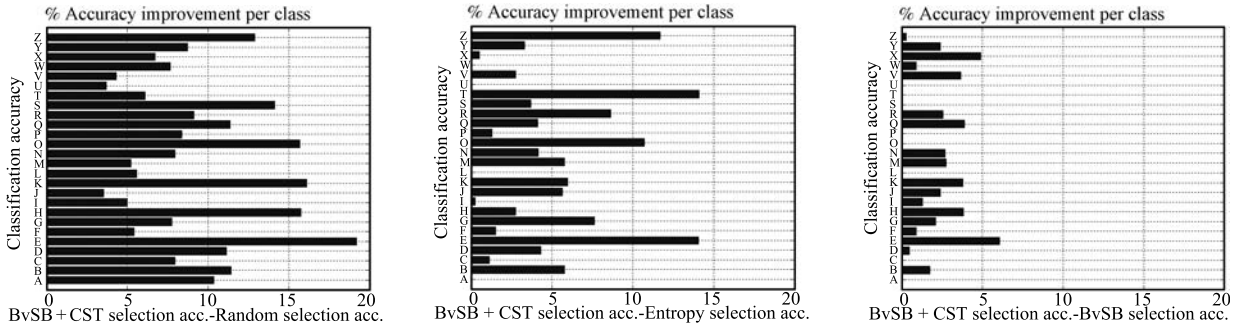


(a) BvSB + CST 与 Random 性能比较 (b) BvSB + CST 与 Entropy 性能比较 (c) BvSB + CST 与 BvSB 性能比较

(a) BvSB + CST vs. Random (b) BvSB + CST vs. Entropy (c) BvSB + CST vs. BvSB

图 7 BvSB + CST 方法与其他三种方法在 Pendigits 图像集各个类别上的分类准确率比较

Fig. 7 Classification accuracy comparison for each class in Pendigits dataset



(a) BvSB + CST 与 Random 性能比较 (b) BvSB + CST 与 Entropy 性能比较 (c) BvSB + CST 与 BvSB 性能比较

(a) BvSB + CST vs. Random (b) BvSB + CST vs. Entropy (c) BvSB + CST vs. BvSB

图 8 BvSB + CST 方法与其他三种方法在 Letters 图像集各个类别上的分类准确率比较

Fig. 8 Classification accuracy comparison for each class in Letters dataset

表 3 在 USPS 数据集上, 各种不同的样本选择方法在达到相同分类准确率所需的标注负担

Table 3 The numbers of iteration rounds required of different methods when achieving the same classification accuracy on USPS dataset

BvSB + CST selection rounds	Random selection rounds	Entropy selection rounds	BvSB selection rounds
4	6	6	5
6	11	12	7
8	17	23	12
10	24	26	13
12	31	31	17
16	48	43	24
18	–	51	27

的选择, 同时也影响着整个分类系统的性能. 在前面我们已经定性地讨论了 *threshold* 的设置问题, *threshold* 的值过小, 会使得对未标注样本进行自动标注的时候引入过多的错误标号, 从而对分类器模型的更新产生误导; *threshold* 的值过大, 会使得经过自动标注的样本所包含的信息量很低, 对分类器模型的更新作用很小. 我们通过实验来说明上述分析的有效性. 在 USPS 数据集上分别取 *threshold* = 0.2 ∼ 0.8, 并计算各种不同的参数值下经过 51 次迭代后系统的分类准确率, 如图 9 所示.

从图 9 可以看出, 当 *threshold* 由小变大, BvSB + CST 方法的分类准确率随着 *threshold* 的变化

曲线是先上升后下降的. 在 *threshold* 小于 0.7 时, 分类准确率随着 *threshold* 的增大而提高, 在 $threshold = 0.7$ 的时候, 分类准确率达到最大值, 之后随着 *threshold* 的增大, 分类准确率开始下降. 实验结果证明了前面分析的准确性, 也说明我们在 USPS 数据集上设置 $threshold = 0.7$ 是合理的. 依照上面的分析, 我们对其他几个图像集的分类实验中的 *threshold* 进行设置, 并根据各个图像集的复杂程度不同进行相应的调整, 对于较复杂的图像集, 适当降低 *threshold*.

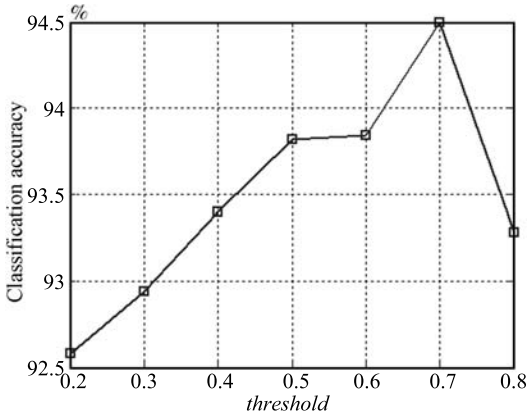


图 9　不同 *threshold* 值下, BvSB + CST 方法的分类准确率

Fig. 9　Classification accuracy of BvSB + CST with different values of *threshold*

3.2　SAR 图像集分类实验

我们从 TerraSAR-X 拍摄的中国广东地区的 SAR 图像中, 截取了 8 类地物样本, 包括 Forest, Hill, Industrial area, Land, Pool, Residential area, River, Woodland, 其中每类地物包含 160 幅样本图像, 每张图像的大小为 64 像素 × 64 像素, 16 位 raw 格式. 我们将原始的 16 位图像量化到 8 位, 用长度为 256 的灰度直方图作为图像特征. 从每类图像中挑选 50 % 作为测试集 (共 $80 \times 8 = 640$ 幅), 剩下的 50 % 作为待选的未标注样本集. 图 10 给出了图像集中各个类别所包含的部分典型样本.

在实验中, 我们在初始样本选择时选择 16 个样本进行训练, 在之后的每次迭代中, 通过 BvSB 主动学习选择 5 个样本由用户进行人工标注, 通过 CST 选择 10 个样本并对其进行自动标注, 其中 CST 约束条件 (4) 中的阈值 *threshold* 设置为 0.5, SVM 分类器采用 RBF 核函数 ($\gamma = 128$), 最大迭代次数 $Max\_ite = 31$.

图 11 给出了在 SAR 图像集上的分类结果, 在绝大多数迭代次数中, 采用 BvSB + CST 方法的分类性能都要高于其他的几种方法. 经过 5 次迭代以后, 分类性能得到明显提升. 在经过 15 次迭代以后, 采用 BvSB + CST 方法的分类准确率随迭代次数变化较为平坦, 说明这时训练样本集中的样本已经比较充分, 能较好地代表样本在特征空间中的分布, 在之后的迭代中新添加的训练样本对分类器模型几乎没有影响. 我们将整个未标注样本集 (640 个样本) 连同其对应的真实类别标号, 作为训练样本训练分类器, 在测试集上得到的分类准确率为 71.25 %. 采用我们的方法在第 15 次迭代时, 在测试集上的分类准确率为 71.22 %, 所需的人工标注样本数量为 86, 仅占整个未标注样本集的 13.44 %, 说明采用我们的方法可以很快收敛到一个较好的性能.

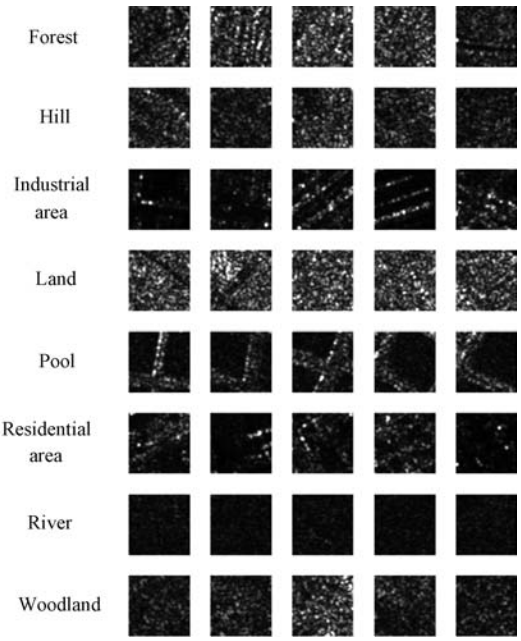


图 10　SAR 图像集中的典型样本

Fig. 10　Examples in SAR image dataset

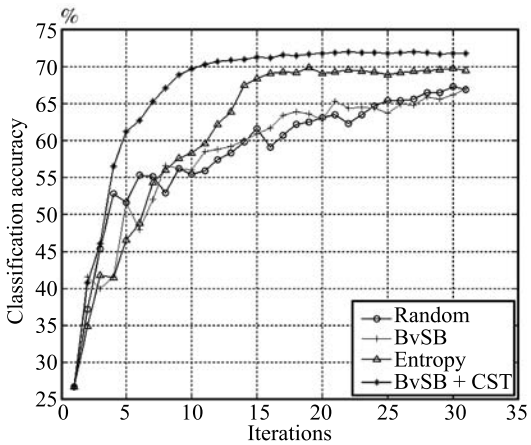


图 11　SAR 图像集上分类准确率

Fig. 11　Classification accuracies on SAR image dataset

4 结论

本文提出了一种基于 BvSB + CST 样本选择和 SVM 分类器的图像分类算法. 在该方法中, 通过 BvSB 主动学习将对当前分类器最具信息量、最有价值的样本选出来, 提交给用户进行人工标注; 同时, 利用 CST 半监督学习, 从剩余的大量未标注样本中, 选择一部分兼具信息性和分类确信度的样本, 由当前分类器进行自动标注, 有效地提高了图像分类问题中训练样本选择的效率, 减少了训练分类器的过程中所需要的人工标注量. 在 3 个光学图像数据集和 1 个 SAR 图像数据集上的分类实验结果表明, 新算法能够有效地减少训练过程中的人工标注负担, 并获得较好的分类性能.

在 CST 的约束条件 (4) 中, 阈值的确定是通过经验选取, 可能并不是最优的, 如何根据数据集及当前分类模型来对该阈值进行自适应地调整是下一步工作中需要考虑的重要问题. 此外, 约束条件 (5) 中, 最近邻点的计算是在欧氏空间完成的, 在一些复杂的数据分布下, 欧氏距离往往不能够准确地描述样本之间的相似关系, 将流形 (Manifold) 等非线性空间映射引入来更准确地刻画样本之间的相似关系也是今后工作中的一个重要方向.

References

1 Settles B. Active Learning Literature Survey, Computer Science Technical Report 1648, University of Wisconsin-Madison, USA, 2009. 3−4

2 Dasgupta S. Coarse sample complexity bounds for active learning. *Advances in Neural Information Processing Systems*. Cambridge: The MIT Press, 2006. 235−242

3 Tong S, Chang E. Support vector machine active learning for image retrieval. In: Proceedings of the 9th ACM International Conference on Multimedia. New York, USA: ACM, 2001. 107−118

4 Tong S, Koller D. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2002, **2**: 45−66

5 Seung H S, Opper M, Sompolinsky H. Query by committee. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory. New York, USA: ACM, 1992. 287−294

6 Dagan I, Engelson S P. Committee-based sampling for training probabilistic classifiers. In: Proceedings of the 12th International Conference on Machine Learning. California, USA: Morgan Kaufmann, 1995. 150−157

7 Hoi S C H, Jin R, Lyu M R. Batch mode active learning with applications to text categorization and image retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 2009, **21**(9): 1233−1248

8 Joshi A J, Porikli F, Papanikolopoulos N. Multi-class active learning for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE, 2009. 2372−2379

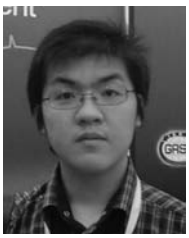
9 Zhu X J. Semi-supervised Learning Literature Survey, Computer Sciences Technical Report 1530, University of Wisconsin-Madison, USA, 2008. 11−13

10 Riloff E, Wiebe J, Wilson T. Learning subjective nouns using extraction pattern bootstrapping. In: Proceedings of the 7th Conference on Natural Language Learning. Stroudsburg, USA: Association for Computational Linguistics, 2003. 25−32

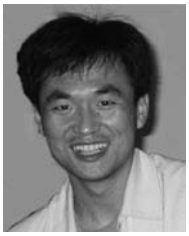
11 Rosenberg C, Hebert M, Schneiderman H. Semi-supervised self-training of object detection models. In: Proceedings of the 7th IEEE Workshop on Applications of Computer Vision. Breckenridge, USA: IEEE, 2005. 29−36

12 Chang C C, Lin C J. LIBSVM: a library for support vector machines [Online], available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/, December 1, 2009

13 Asuncion A, Newman D J. UCI machine learning repository [Online], available: http://archive.ics.uci.edu/ml/datasets.html, January 10, 2010



陈 荣 武汉大学电子信息学院博士研究生. 主要研究方向为图像处理, 模式识别和机器学习.
E-mail: chenrong0707@gmail.com
(**CHEN Rong** Ph. D. candidate at the Signal Processing Laboratory, Electronic Information School, Wuhan University. His research interest covers image processing, pattern recognition, and machine learning.)



曹永锋 贵州师范大学数学与计算机科学学院副教授. 主要研究方向为图像处理和模式识别.
E-mail: yongfengcao.cyf@gmail.com
(**CAO Yong-Feng** Associate professor at the School of Mathematics and Computer Science, Guizhou Normal University. His research interest covers image processing and pattern recognition.)

孙 洪 武汉大学电子信息学院教授. 主要研究方向为信号与图像处理. 本文通信作者. E-mail: hongsun@whu.edu.cn
(**SUN Hong** Professor at the Electronic Information School, Wuhan University. Her research interest covers signal and image processing. Corresponding author of this paper.)