基于动态可行域划分的SVM主动学习

张晓宇
(中国科学技术信息研究所战略研究中心 北京 100038)

摘 要 针对传统SVM主动学习中批量采样方法的不足,提出了动态可行域划分算法。从特征空间与参数空间的对偶关系入手,深入分析SVM主动学习的本质,将特征空间中对样本的标注视为参数空间中对可行域的划分;通过综合利用当前分类模型和先前标注样本两方面信息,动态地优化可行域划分方案,以确保选取的样本对模型改进的价值,最终实现更为高效的选择性采样。实验结果表明,基于动态可行域划分的SVM主动学习算法能够显著提高所选样本的信息量,从而能够在有限的标注代价下大幅提高其分类性能。

关键词 半监督学习,主动学习,选择性采样,支持向量机,可行域
中图法分类号 TP37 **文献标识码** A

SVM Active Learning via Dynamic Version Space Division

ZHANG Xiao-yu
(Research Center for Strategic Science and Technology Issues, Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract This paper presented a dynamic version space division algorithm for SVM active learning, in view of the drawbacks of traditional batch sampling methods. Based on the duality property of feature space and parameter space, we discussed SVM active learning in dual space and concluded that example labeling in feature space corresponds to version space division in parameter space. Taking both the existing classification model and the previously labeled examples into consideration, we optimized the version space division process and maximized the value of selected examples for model refinement. In this way, a more effective selective sampling was achieved. Experimental results demonstrate the effectiveness of the dynamic version space division algorithm, which remarkably improves the classification performance under the cost of limited labeling effort.

Keywords Semi-supervised learning, Active learning, Selective sampling, Support vector machine, Version space

1 引言

传统的机器学习可分为监督学习(supervised learning)和非监督学习(unsupervised learning)两类。前者完全基于已标注样本训练分类模型,而后者则只利用未标注样本的信息。近年来,半监督学习(Semi-Supervised Learning, SSL)[1]作为一种介于监督学习和非监督学习之间的方法,旨在综合利用标注样本和未标注样本两者的信息,以提高分类模型的性能,其受到越来越多的关注。在实际分类问题中,已标注样本由于往往需要通过人工标注获得,因此代价较高,数量也非常有限;另一方面,未标注样本大量存在于样本库中,且易于获取。因此,如何有效利用有限的已标注样本和大量易于获得的未标注样本,是研究半监督学习的关键所在。主动学习(active learning)[2,3]是解决上述问题的一种行之有效的方法。其主要思想是:仅仅选取对模型改进最有价值的样本进行标注,使得根据这些样本的标注信息所训练出来的新模型性能得到尽可能大的提升。分类模型的选择是决定分类效果的又一关键因素。支持向量机(Support Vector Machine, SVM)[4]作为一种性能优越的分类器,在许多分类问题中得到广泛应用。因此,本文将重点研究SVM分类模型下的主动学习,提出优化改进算法。

本文首先对SVM主动学习方法进行深入分析;然后针对现有方法的不足,提出动态可行域划分算法;最后通过实际分类问题对算法的有效性进行验证。

2 SVM主动学习

SVM的基本思想是学习一个最优的超平面,以最大的分类间隔(margin)将训练数据分开。给定一个训练集 $L=\{(x_1,y_1),\cdots,(x_n,y_n)\}$ $(x_i\in X, y_i\in\{+1,-1\})$,通过一个适当的Mercer核函数 K 所得到的隐式映射 $\Phi: X\rightarrow F$[4],SVM分类器可以表示为

$$f(x)=w\cdot\Phi(x)+b \tag{1}$$

相应的分类面即

$$w\cdot\Phi(x)+b=0 \tag{2}$$

对于样本 $x_i\in X$, $f(x_i)$ 可以看成是一种"带符号距离",其绝对值反映了样本到分类面的距离,其符号表示样本所属

到稿日期:2011-08-20 返修日期:2011-11-22 本文受中央级公益性科研院所基本科研业务费专项资金项目(ZD2011-7-3),中国科学技术信息研究所科研项目预研资金(YY-201114),国家自然科学基金项目(60475010)资助。

张晓宇(1983—),男,博士,助理研究员,主要研究方向为模式识别与智能系统、科技情报与知识管理,E-mail:zhangxy@istic.ac.cn。

的类别。

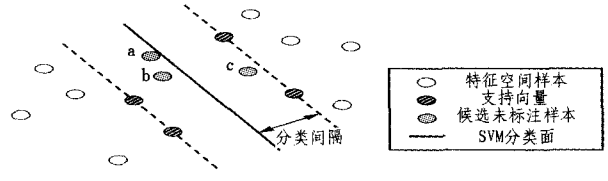
对于给定的训练集 L 和 Mercer 核 K，所有能够对训练集样本正确分类的超平面所组成的集合称为可行域(version space)[5]。换言之，分类面 f 在可行域中，当且仅当对每一个训练样本 x_i，若 $y_i=+1$，则 $f(x_i)>0$，若 $y_i=-1$，则 $f(x_i)<0$。可行域可以表示为

$$V=\{f\mid y_i f(x_i)>0, i=1,\cdots,n\} \tag{3}$$

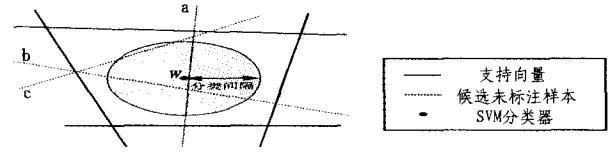
由式(1)，分类面 f 与参数 w 是一一对应的，因而可行域的表示方式可以改写成

$$V=\{w\in W\mid \|w\|=1, y_i(w\cdot\Phi(x_i)+b)>0, i=1,\cdots,n\} \tag{4}$$

根据文献[5]的分析，特征空间 F 和参数空间 W 之间存在着一种对偶关系：参数空间中的点对应于特征空间中的超平面；反之，特征空间中的点对应于参数空间中的超平面。当一个样本 x_i 被标注之后，就对现有可行域提出了新的约束，即要求新可行域中的分类面满足 $y_i f(x_i)=y_i(w\cdot\Phi(x_i)+b)>0$。这样，原可行域中只有与超平面 $y_i(w\cdot\Phi(x_i)+b)=0$ 正方向一侧相交的部分被保留下来。从这个意义上说，在特征空间中对样本的标注相当于在参数空间中对可行域的划分。随着越来越多的样本被标注，可行域被划分得越来越小，分类模型也随之越来越趋近于最优。图 1 描述了特征空间与参数空间之间的对偶关系。



(a) 特征空间中样本和分类器的表示形式



(b) 参数空间中样本和分类器的表示形式

图 1　特征空间与参数空间对偶关系示意图

主动学习的目标是选取最有信息的样本进行标注，从而最有效地改进分类效果；从对偶空间的角度看，则是要选择能够最有效地划分可行域的样本，从而利用有限的标注量将可行域划分得尽可能小。

传统的 SVM 主动学习(如 $\text{SVM}_{\text{Active}}$[6,7])采用的是批量采样模式，即批量地选择距离当前分类面最近的 k 个样本进行标注，然后根据这 k 个样本的标注信息对分类模型进行训练更新。根据对偶原理，特征空间中距离分类面最近的样本对应于参数空间中距离可行域中心最近的超平面，因此无论样本标注后保留下的是超平面的哪一侧，其都可以达到近似平分当前可行域的效果。但是，这种方法中样本的选取完全取决于当前分类模型，忽略了同一批被选取的样本之间的相互关系，导致先前标注样本的信息不能用来为后续样本的选取提供有效的指导，因此其大大限制了所选取样本的信息量，从而无法达到比平分可行域更优的效果，最终影响了分类模型的改进。例如，在图 1 中，若能预测样本 c 在标注后其所对应的超平面的左上方一侧被保留的可能性较大，则不妨选取样本 c 进行标注，这样可以获得比标注样本 a 更好的可行域划分效果。

3　动态可行域划分算法

为了克服传统 SVM 主动学习的缺陷，提出了“动态可行域划分”(dynamic version space division)算法。

算法采用动态批量采样模式，即每次选取一个最有信息的样本进行标注，并根据该样本的标注信息指导下一个样本的选取，如此循环；只有当一定数目的样本标注完成后，才通过一次训练对分类模型进行更新。在这种情况下，样本的选取不再仅仅取决于当前分类面，还在新标注样本信息的指导下动态地进行调整，从而在充分利用样本之间相互关系的同时有效地弥补了当前分类模型的不足。

算法的主要思想是：每当一个样本被标注之后，就利用其标注信息对预测结果相近的样本进行预测，并在此基础上动态地选取能够将当前可行域划分得更小的样本进行标注。重复上述过程，直到一定批量的样本标注完成。最后，利用这些标注样本对分类模型进行更新。

动态可行域划分算法通过综合利用当前分类模型和先前标注样本的信息，来保证所选取的各样本在对偶空间中划分当前可行域所保留的范围呈现近似递减趋势，即所选取样本的信息量不断增大，从而利用有限的标注样本获得尽可能大的分类性能提升。

3.1　算法流程

为了选取信息量大的样本，需要对不同样本改进当前分类模型的能力(即划分当前可行域的能力)进行量化度量。为此，本文提出如下命题。

命题 1　给定由 n 个训练样本所确定的如式(4)所示的可行域 V，在分别标注样本 (x_{n1},y_{n1}) 和 (x_{n2},y_{n2}) 之后所获得的新可行域记为 V_{n1}^{new} 和 V_{n2}^{new}。如果 $y_{n1}f(x_{n1})>y_{n2}f(x_{n2})$，则 $Area(V_{n1}^{new})>Area(V_{n2}^{new})$，其中 $Area(V)$ 表示可行域 V 的大小。

证明：在分别标注样本 (x_{n1},y_{n1}) 和 (x_{n2},y_{n2}) 之后，新的可行域 V_{n1}^{new} 和 V_{n2}^{new} 可以表示为

$$V_{n1}^{new}=\left\{w\in W\left|\begin{array}{l}\|w\|=1,\\ y_i(w\cdot\Phi(x_i)+b)>0, i=1,\cdots,n\\ y_{n1}(w\cdot\Phi(x_{n1})+b)>0\end{array}\right.\right\} \tag{5}$$

$$V_{n2}^{new}=\left\{w\in W\left|\begin{array}{l}\|w\|=1,\\ y_i(w\cdot\Phi(x_i)+b)>0, i=1,\cdots,n\\ y_{n2}(w\cdot\Phi(x_{n2})+b)>0\end{array}\right.\right\} \tag{6}$$

如果 $y_{n1}f(x_{n1})>y_{n2}f(x_{n2})$，则有

$$y_{n2}(w\cdot\Phi(x_{n2})+b)>0\Rightarrow y_{n1}(w\cdot\Phi(x_{n1})+b)>0 \tag{7}$$

根据式(5)、式(6)中新可行域 V_{n1}^{new} 和 V_{n2}^{new} 的形式，并结合式(7)，可得

$$V_{n2}^{new}\subset V_{n1}^{new} \tag{8}$$

所以

$$Area(V_{n1}^{new})>Area(V_{n2}^{new}) \tag{9}$$

从命题 1 可知，对于样本 (x_i,y_i)，其 $y_i f(x_i)$ 的值越小，则该样本标注后所对应的超平面划分当前可行域保留的部分越小，因而该样本对于分类模型的训练就越有价值。

动态可行域划分算法的思路是：动态地按照 $yf(x)$ 值递

减的原则选取一系列样本，以保证后选取的样本比先前选取的样本更有标注价值，从而使得最终选取的一批样本具有尽可能高的信息量。

对于一个未标注样本(x_i, y_i)，其类标 y_i 是未知的，因而无法直接根据 $y_i f(x_i)$ 值选取最有信息的样本，但可以利用已标注样本的信息来预测其它样本的类标。使用当前分类器 f 可以对未标注样本 $x_i \in U$ 进行分类，并按照分类结果 $f(x_i)$ 进行升序排序，得到序列 S。由于在排序结果 S 中，相邻的样本（即具有相近的分类结果 $f(x_i)$ 的样本）在大多数情况下具有一致的真实类标，因此可以利用新标注样本的信息来预测其相邻样本的类标。

在每轮标注的开始，由于之前没有新样本被标注，因此第一个样本的选取完全取决于当前分类面，即与传统方法一样，选取距离当前分类面最近的样本进行标注。在此之后，每当一个样本(x_i, y_i)被标注，如果 $y_i = +1$，为了使得下一个样本具有更小的 $yf(x)$ 值，则在序列 S 中沿着 $f(x)$ 减小的方向搜索下一个样本；如果 $y_i = -1$，则在序列 S 中沿着 $f(x)$ 增大的方向搜索下一个样本。

整个动态可行域划分算法流程如图 2 所示。

输入：当前分类器 f，已标注样本集 L，未标注样本集 U，在每轮交互中需要标注的样本数 k

步骤 1　用 f 对 U 进行分类。按照 f(x) 递增的顺序将未标注样本排成序列 S。

步骤 2　(1)选择距离当前分类面 f 最近的一个样本 x_i 进行标注：

$$l = \arg\min_i |f(x_i)|, x_i \in U$$

记录类标 $y_{previons} = y_i$

(2)从 U 和 S 中删除 x_i，并将 x_i 加入 L：

$U = U - \{x_i\}, S = S - \{x_i\}, L = L \cup \{x_i\}$

步骤 3　对 r=2,…,k，重复执行下述操作：

(1)如果 $y_{previons} = +1$，选取序列 S 反方向上的相邻样本 $x_{i'}$ 进行标注，记录类标 $y_{previons} = y_{i'}$

如果 $y_{previons} = -1$，选取序列 S 正方向上的相邻样本 $x_{i'}$ 进行标注，记录类标 $y_{previons} = y_{i'}$

(2)从 U 和 S 中删除 $x_{i'}$，并将 $x_{i'}$ 加入 L：

$U = U - \{x_{i'}\}, S = S - \{x_{i'}\}, L = L \cup \{x_{i'}\}$

步骤 4　利用 L 对 f 进行更新。

图 2　动态可行域划分算法流程

3.2　算法讨论

如上所述，传统 SVM 主动学习选取距离当前分类面最近的样本（也即具有最小 $|f(x)|$ 值的样本）进行标注，而

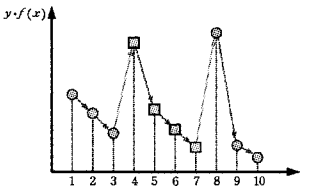
$$|f(x)| = \mathrm{sgn}[f(x)] \cdot f(x) \tag{10}$$

因此，不难发现其本质上是在假设样本的真实类标 y 与分类器判定类标 $\mathrm{sgn}[f(x)]$ 相等的前提下对命题 1 的一种近似。由于当前分类器的准确度往往不高，因此上述假设往往是不成立的（事实上，如果 $y = \mathrm{sgn}[f(x)]$ 对所有样本都成立，那么说明当前分类模型已经非常完善，从而也就没有必要再通过选取样本进行标注的方法对其进行改进了）。由此可见，简单地用 $\mathrm{sgn}[f(x)]$ 来代替 y 是非常不可靠的，因而也影响了传统方法的效果。

在动态可行域划分算法中，实际上是使用 $y_{previous} f(x)$ 来代替 $yf(x)$ 作为样本信息量的度量，其中 $y_{previous}$ 并不是样本的真实类标，而是利用先前标注的相邻样本的类标所进行的预测。由于这种近似并不直接依赖于当前分类面的准确度，而是基于分类结果相似性和类标一致性的相对关系，因此动态可行域划分算法较传统 SVM 主动学习更加接近命题 1 所述的信息量度量。动态可行域划分算法对于传统方法的另一个改进在于，传统的选择性采样完全由当前分类器 $f(x)$ 决定，而在动态可行域划分算法中一个样本的选取由两个因素决定：$f(x)$ 和 $y_{previous}$，即除了 $f(x)$ 之外还受到先前标注样本的类标 $y_{previous}$ 的指导，从而使得样本的选取过程更加具有针对性。

在图 3(a)中，我们模拟了根据动态可行域划分算法进行样本选取的过程。所选取样本的相应 $yf(x)$ 值显示在图 3(b)中，由此可见，按照我们的方法选取出来的样本的 $yf(x)$ 值是（非严格）单调递减的，只有在样本的真实类标与前一个样本不一致时才发生跳变。

(a) 样本选取过程



(b) 所选取样本对应的 $yf(x)$ 值

图 3　动态可行域划分算法示例

值得一提的是，在文献[8]中，作者给出直观结论：被当前分类器分错的样本比分对的样本更加有价值。这种现象也可以用命题 1 进行解释：对于被分错的样本(x_{false}, y_{false})，其真实类标与分类器判定类标异号，因而 $y_{false} f(x_{false}) < 0$；相反，对于被分对的样本$(x_{true}, y_{true})$，有 $y_{true} f(x_{true}) > 0$，因此

$$y_{false} f(x_{false}) < 0 < y_{true} f(x_{true}) \tag{11}$$

根据命题 1，被分错的样本标注以后将比被分对的样本更有效地划分当前可行域，从而对分类器的改进更有价值。

4　实验

为了验证动态可行域划分算法（表示为 $\mathrm{SVM}^{D}_{Active}$）的有效性，将其应用到图像检索[9,10]的相关反馈[11,12]中，与传统的 SVM_{Active} 及非主动学习方法 $\mathrm{SVM}_{Passive}$ 进行比较。实验流程为：根据用户提交的查询，系统通过匹配给出初始检索结果，同时提供一批图像由用户进行“相关”或“不相关”的标注，最后利用标注信息对检索模型进行重新训练，以提高检索性能。非主动学习方法中，供用户标注的图像是随机选取的；传统 SVM 主动学习方法批量地选定所有供标注的图像；动态可行域划分算法则按照本文所述流程动态地选取供标注的图像。

实验集来自 Coral 图像库。其中总共有 10200 幅图像，分为 102 个不同语义类别（如老虎、汽车、人等），每类包含 100 幅图像。实验中，每类图像的前 10 幅（共 1020 幅）作为查询图像测试图像检索的性能，最终的准确率是这 1020 幅查询图像准确率的平均值。我们采用颜色和纹理特征来描述图

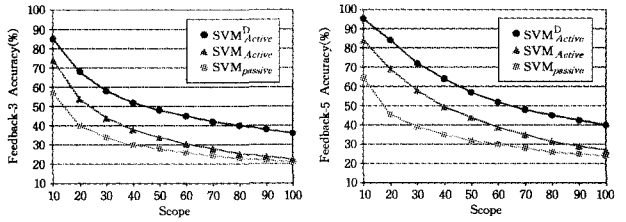
（下转第 189 页）

bined use between DEA and AHP [J]. European Journal of Operational Research, 2009, 199(1): 219-231
[11] 陈伟. 一种基于等级法的联网审计绩效评价方法[J]. 计算机科学, 2010, 37(11): 111-116
[12] 陈伟. 一种基于 AHP 的联网审计绩效评价方法[J]. 审计与经济研究, 2011, 26(5): 47-52
[13] 陈伟, Smieliauskas W. 联网审计技术方法与绩效评价[M]. 北京: 清华大学出版社, 2012
[14] 刘思峰, 党耀国, 方志耕, 等. 灰色系统理论及其应用(第五版)[M]. 北京: 科学出版社, 2010
[15] Satty T L. The Analytic Hierarchy Process [M]. New York: McGraw-Hill, 1980
[16] 陈伟, Smieliauskas W, 刘思峰. 联网审计绩效评价影响因素的灰色关联分析[J]. 计算机应用研究, 2012, 29(2): 435-437

(上接第 177 页)

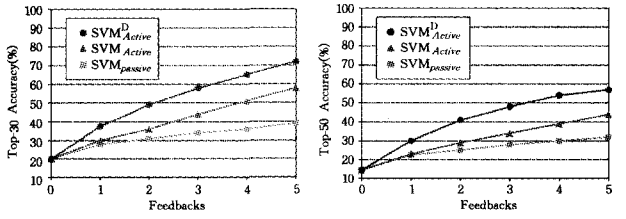
像，其中颜色特征由 125 维的颜色直方图和 6 维的颜色矩构成，而纹理特征则利用 3 层的离散小波变换之后在 10 个子带上分别取均值和方差构成一个 20 维的特征。参与比较的各种方法均使用 RBF 核的 SVM 分类器，相关参数由交叉验证决定。在每一轮相关反馈中，各种方法均选取相同数目的图像进行标注。

图 4 给出了分别在 3 轮和 5 轮相关反馈之后返回的前 10～100 个结果的准确率，横坐标 x 表示前 x 个返回结果，纵坐标表示准确率；图 5 给出了前 30 和前 50 个返回结果范围内准确率随相关反馈轮次变化的趋势，横坐标 x 对应于第 x 次相关反馈，纵坐标表示准确率。



(a) 3 轮相关反馈之后的准确率　　(b) 5 轮相关反馈之后的准确率

图 4　前 n (n=10,20,…,100)个返回结果的准确率



(a) 前 30 个返回结果的准确率　　(b) 前 50 个返回结果的准确率

图 5　经过 k (k=0,1,…,5)轮相关反馈之后的准确率

实验结果表明，主动学习(曲线●、▲)优于非主动学习(曲线■)，说明通过有针对性地主动选取样本进行标注，能够更加有效地改进分类性能；主动学习中，动态可行域划分算法(曲线●)优于传统批量采样方法(曲线▲)，说明在主动学习中选择性采样方法起到了至关重要的作用，动态可行域划分算法通过综合利用当前分类模型和先前标注样本两方面信息对采样过程进行动态优化，进一步提高了所选取样本的信息量，从而显著地提高分类精度。

结束语　本文深入分析了 SVM 分类模型下的主动学习的特点，针对传统批量采样方法的不足，提出了动态可行域划分算法。根据特征空间与参数空间的对偶关系，特征空间中对样本的标注等价于参数空间中对可行域的划分。算法采用动态批量采样模式，通过综合利用当前分类模型和先前标注样本两方面信息，动态选取具有更优的可行域划分能力的样本进行标注，从而保证了所选取样本对于模型改进的价值，最终利用有限的样本标注量获得尽可能大的分类性能提升。实验结果表明，相对于传统方法，动态可行域划分算法能够显著提高所选取样本的信息量，对于减轻人工标注负担、改善分类性能具有重要意义。

参 考 文 献

[1] Zhu X. Semi-supervised learning literature survey [R]. Wisconsin: Computer Sciences, University of Wisconsin-Madison, 2008
[2] Cohn A, Ghahramani Z, Jordan M I. Active learning with statistical models [J]. Journal of Artificial Intelligence Research, 1996, 4(1): 129-145
[3] Mccallum A, Nigam K. Employing EM in pool-based active learning for text classification[C]// Proceeding of the 15th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 1998: 350-358
[4] Burges J C. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121-167
[5] Mitchell T. Generalization as search [J]. Artificial Intelligence, 1982, 18(2): 203-226
[6] Tong S, Koller D. Support vector machine active learning with applications to text classification [J]. The Journal of Machine Learning Research, 2000, 2(1): 45-66
[7] Tong S, Chang E. Support vector machine active learning for image retrieval [C]// Proceedings of the 9th ACM International Conference on Multimedia. New York: ACM, 2001: 107-118
[8] Zhang X, Cheng J, Lu H, et al. Weighted co-SVM for image retrieval with MVB strategy[C]// Proceedings of 2009 IEEE International Conference on Image Processing. San Antonio, TX: Signal Processing Society, 2007: 517-520
[9] Smeulders A W M, Worring M, Santini S, et al. Content-based image retrieval at the end of the early years[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(12): 1349-1380
[10] Lew M S, Sebe N, Djeraba C, et al. Content-based multimedia information retrieval: state of the art and challenges[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2006, 2(1): 1-19
[11] Rui Y, Huang T S, Ortega M, et al. Relevance feedback: a power tool for interactive content-based image retrieval [J]. IEEE Transactions on Circuits and Systems for Video Technology, 1998, 8(5): 644-655
[12] Zhou X S, Huang T S. Relevance feedback in image retrieval: a comprehensive review[J]. Multimedia Systems, 2003, 8(6): 536-544