综合学习方法 AdaBoost 在暴雨预测中的应用①

An Application of AdaBoost in Rainstorm Predicting

杨　艳　燕东渭　（成都信息工程学院　四川成都　610041）
赵奎锋　魏　亭　（陕西省气象局）

摘要：机器学习技术在气象预报中的应用仅仅局限于 ANN 和一些核方法，本文将综合学习方法引入预报中，并为 AdaBoost 方法融入减抽样的思想。通过暴雨预测的试验，表明减抽样的确可以提高 AdaBoost 暴雨预测的效果。

关键词：综合学习　AdaBoost　暴雨预测

人工智能科学中机器学习方法由于具有自动学习的功能，近年来在许多领域中发挥了重要的作用。在气象预报这一复杂的领域中，主要的应用和研究有人工神经网络和支持向量机等核方法。而另外一类重要的机器学习算法：综合学习（Ensemble Learning）[1]，目前还没有在气象预报中得到应用。以往机器学习方法在气象预报中应用时常遇到的一个重要的难题，就是对不平衡类别问题。就分类问题而言，天气预报中诸多的灾害性天气的预测，均属于不平衡类别问题。对训练数据减抽样处理可以使不平衡问题"平衡化"，是解决不平衡学习问题的有效办法。本文把减抽样的思想用于重要的综合学习算法：AdaBoost。进而用此方法进行了铜川暴雨的预测试验。实验结果表明对暴雨预测有较好的效果，相对传统的 AdaBoost，准确率明显提高。

1　综合学习及 AdaBoost 简介

综合学习指利用多个为同一任务而训练出的基分类器的输出，以得到更好的分类器。综合多个分类器往往能够改善性能 [2,3]。AdaBoost 是目前最重要的综合学习算法[4,5]。

Kearns 和 Valiant 指出[6]，在 PAC 学习模型中[7]，若存在一个多项式级的学习算法来识别一组概念，并且识别正确率很高，那么这组概念是强可学习的；而如果学习算法识别一组概念的正确率仅比随机猜测略好，那么这组概念是弱可学习的。Kearns 和 Valiant 提出了弱学习算法与强学习算法的等价性问题，即是否可以将弱学习算法提升成强学习算法。如果两者等价，那么在学习概念时，只要找到一个比随机猜测略好的弱学习算法，就可以通过综合将其提升为强学习算法，而不必直接去找通常情况下很难获得的强学习算法。

1990 年，Schapire 通过一个构造性方法对该问题作出了肯定的证明[8]，其构造过程就是最初的 Boosting 算法。1999 年 Schapire 和 Singer 提出了 AdaBoost 更一般的形式[4]，并引入"自信率预测"以改善 Boosting 的性能，本文提到的 AdaBoost 就是这种算法。

设样本空间为 $\aleph$，类别空间为 $\Re = \{1, -1\}$，只考虑二分类问题。训练样本集为 $S = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$，其中 $x_k \in \aleph$，$y_k \in \Re$，S 中共有样本 m 个。算法 1 给出了 AdaBoost[9,10] 的详细过程。AdaBoost 的目的是为了能够根据给定的例子 x 预测类别标签 y，而找出假设 H(x)。AdaBoost 在训练样本上维护一套概率分布，在每一轮迭代中 AdaBoost 在每个样本上调整这套分布，成员分类器在训练样本上的错误率被计算出来，并以此在训练样本上调整概率分布。设第 t 轮样本 x_i 的权重为 $D_t(i)$，调整训练样本概率分布的作用是在被误分的样本上设置更多

① 基金项目：国家科技部公益项目（2001DIB20095）

的权重，在分类正确的样本上减少其权重，初始时所有样本的权重相等（为 $D_1(k)=\frac{1}{m}$）。通过 T 个单个分类器的加权投票建立最终分类器，每个分类器按其在训练集上的精度而加权。

算法 1 AdaBoost（Schapire）

初始化 $D_1(k)=\frac{1}{m}$

For t=1..T 执行以下三步

（1）用分布 D_t 训练出一个弱分类器 $h_t: \aleph \to \Re$；

（2）选择 $a_t=\frac{1}{2}\ln(\frac{1+r_t}{1-r_t})$，其中 $r_t=\sum_{k=1}^{m}D_t(k)h_t(x_k)y_k$；

图 1　SadaBoost 的基本思想

（3）更新 D_t 为：$D_{t+1}(k)=\frac{D_t(k)\exp(-a_t y_k h_t(x_k))}{Z_t}$，其中 Z_t 为归一化因子，$Z_t=\sum_k D_t(k)\exp(-a_t y_k h_t(x_k))$。

最后得到的分类器为：

$H(x)=\mathrm{sign}(f(x))$，其中 $f(x)=\sum_{t=1}^{T}a_t h_t(x)$。

2　减抽样的思想

从上面可以看出，AdaBoost 没有考虑问题类别的不平衡性，但在基分类器中都需要对于训练样本集做了抽样，这为改造这种方法提供了条件。

不平衡类别最基本的方法是处理样本，而处理样本最简单的方法是抽样。基本的抽样法有增抽样和减抽样两种，哪种抽样策略更有优势，目前尚没有定论。由于 AdaBoost 是有放回的重复抽样，所以不太可能丢失有用的反类样本。因此对于这种综合学习算法而言，减抽样是比较简单可行的选择。

AdaBoost 为样本维持着权重，根据权重大小进行抽样，不容易在迭代时改进，所以我们在迭代之前为 AdaBoost 增加一个减抽样过程，将所有样本在欧氏空间中，距离少数类样本点较近的点通常是噪声的可能较大，而较远的点则认为是较可靠的多数类样本。同时规定少数类为正类，多数类是反类。通过减抽样得到可靠的与正类点数目相当的反类，这样也会使样本集变得平衡，接下来使用标准 AdaBoost 过程。把这个方法称为减抽样的 AdaBoost：SadaBoost（Sub－sampled Adaboost）

3　SadaBoost 算法

对于不平衡类别问题，反类样本比较多，会过多地侵入正类的空间，从而导致分类面向正类偏移，造成的结果是把更多的样本分为反类。所以为了对正类较为有利，一个可行的办法是去掉离正类比较近的反类。图 1 给出了一个示例，图中“＋”表示正类，“－”表示反类，左图中反类很多，几乎包围了正类，右图中的虚线表示可以去掉的反类，去掉之后就会形成有利于正类的分类面。

设正类集为 P，其中的元素数为 n_+。反类集为 N，其中的元素数为 n_-。x，y 表示任意两个样本。首先定义 x，y 之间的距离为 d(x，y)。d(x，y)可以取为欧氏距离：

$$d(x,y)=\|x-y\|^2 \qquad (1)$$

接着定义反例 y 到正类集 P 的距离为 d(y，P)，一般地，d(y，P)可取为：

$$d(y,P)=\min_{x\in P}d(x,y) \qquad (2)$$

SadaBoost 分两步，第一步是从 N 中选出 n_+ 个和正类集 P 距离最大的反类，这些反类是比较可靠的样本，用这些反类和正类组成新的训练集，这样可使样本集变得平衡；第二步是对该训练集再使用传统的 AdaBoost 算法。算法 2 给出了 SadaBoost 的过程。

算法 2 SadaBoost

（1）$\forall y \in N$，求出 d(y，P)。找出 d(y，P)最大的 n_+ 个反类，并与 P 组成新的训练集；

(2) 对新的训练集使用 AdaBoost 算法。

4 SadaBoost 暴雨预测的试验

4.1 试验评分标准

预测科学研究中对于预测结果的衡量标准一般会采用总体准确率。但为了能够更加客观、公证和科学地对于暴雨等突发的灾害性天气预报结果进行考核,气象科学领域通常采用 TS 评分[11] (TS: Threat Score)作为参考指标。TS 的计算如式(3):

$$TS = \frac{tp}{tp + fp + fn} \quad (3)$$

TS 是 1975 年美国科学家 Donaldson 提出的,最初被叫做临界成功指数(CSI: Critical Success Index),旨在对预测严重气象事件的技巧进行客观的评估。定义式中 tp 的被称作命中(hits),即预报正确的灾害天气次数,fp 被称作空报或虚警(false alarm),即没有灾害天气出现而空报的次数,fn 被称作漏报(misses),即出现灾害天气但没有预测出的次数。TS 的意义在于不忽视虚警的前提下对命中更多的关注。

4.2 试验数据

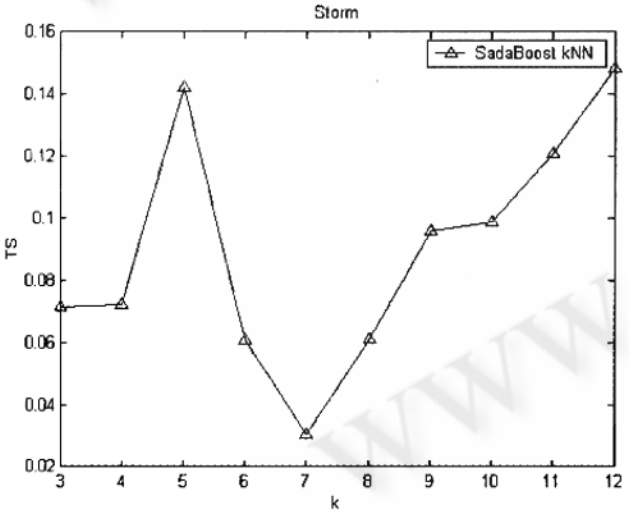

图 2　SadaBoost kNN 对于"暴雨"的实验结果

盛夏暴雨对于植被覆盖率很低的黄土高原边缘地带,常常诱发山体滑坡和泥石流,给人民生活和生命财产安全造成严重威胁,对其准确预测很有现实意义。本文研究的是预报未来 24 小时内"铜川盛夏暴雨"。预报对象正是铜川市所属三个县(区)中,是否有一个或以上站在当日 20 点至次日 20 点之间出现暴雨(即日降水超过 38mm),而使用的是当日 8 点得到的实时数据。模型训练数据中的预报因子是从 1980 年到 2002 年间每年 7 至 8 月兰州、西安等 20 个高空站的 850hPa、700hPa、500hPa 三层当日 08 点天气要素中选取。经过对历史资料的整理,挑选了六个与铜川盛夏西南气流型暴雨发生非常密切且物理意义比较清晰的因子,分别是:西安与武汉 500 hPa 高度差、平凉 500 hPa 24 小时温度变化、平凉 700 hPa 24 小时相对湿度变化、酒泉与西安 500 hPa 温度差、甘孜 500 hPa 南风分量、格尔木 500 hPa 24 小时变高。经过入型条件和无暴雨指标处理,最后形成 94 个日数据,每个数据有 6 维,其中共有暴雨个例 23 个,显然是一个不平衡类别问题。在本文的后续试验中,均采用前 60 个数据(1980 年到 1989 年的数据)作为训练样本,用后 34 个数据(1990 年至 2002 年的数据)进行测试。

4.3 试验结果

SadaBoost 需要一个基分类算法,很多应用中都选 Stumps。我们用 Stumps 进行了试验,发现得到的分类器把所有样本都分为反类,因而 TS 全部变为 0,没有实际价值。kNN 是一个古老而有效的分类器,我们选择 kNN 作为基分类器再次进行试验。发现基分类器的个数不足够大时(实际选用 100 个基分类器),结果很不稳定,分析认为由于有抽样的过程导致训练样本很不稳定而造成的。将其提高到 1000 后,才比较稳定。图 1 给出了 kNN、AdaBoost kNN 以及 SadaBoost kNN 对于"暴雨"的试验结果。kNN 中的 k 可以取不同值,而且经常对实验结果有很大影响,我们给出了 k 从 3 取到 12 时预测结果的 TS 值,图中的横轴表示 k,纵轴表示 TS。由于 AdaBoost 抽样过程,同样的训练样本不一定能得到完全相同的结果,我们对每个 k 进行 5 次试验,图中给出的 TS 是这 5 次的均值。

因为训练数据的不平衡,所以暴雨预测是很困难的任务。对于这个数据,kNN、AdaBoost kNN 都失效了,没有预测出一次暴雨,所有的结果 TS 评分都是 0。而 SadaBoost 对于"暴雨"的则预测出不少目标类别,TS 评分比前者有很大提高。

4.4 试验结果展望

从试验结果看,SadaBoost 的确在处理暴雨预测

这样的不平衡问题时显示了明显的优势。但其 TS 评分结果还不是很理想,方法本身还有许多改进的地方:第一步用欧氏距离度量样本之间的距离(1),这实质上限制了样本空间为欧氏空间,可以进一步采用核函数来表示样本间的距离,得到更灵活的距离定义,并解除对样本空间为欧氏空间的限制;SadaBoost 中反类到正类集的最小距离作为该反类到正类集的距离(2),这两个距离也可以改进,比如使用到正类的平均距离等。

参考文献

1 T. Dietterich. Machine learning research: four current directions. Artificial Intelligence, 1997, 18(4): 97 – 136.

2 罗雪晖、李霞、张基宏,支持向量机及其应用研究,深圳大学学报(理工版),2003,20(3):40 – 46.

3 D. M. J. Tax, R. P. W. Duin. Data Description in subspaces. In: A. Sanfeliu, J. J. Villanueva, M. Vanrell, R. Alquezar, A. K. Jain, J. Kittler (eds.), Proc. 15th Int. Conference on Pattern Recognition and Neural Networks (ICPR15). Los Alamitos: IEEE Computer Society Press, 2000. vol. 2672 – 675.

4 R. Schapire, Y. Singer. Improved boosting algorithms using confidence – rated predictions. Machine Learning, 1999,37(3):297 – 336.

5 L. Breiman. Bagging predictors. Machine Learning,1996,24(2):123 – 140.

6 M. Kearns, L. G. Valiant. Learning boolean formulae or factoring. Tech. Rep. TR 14 – 88. Aiken Computation Laboratory, Harvard University. 1988.

7 M. Anthony. Probabilistic analysis of learning in artificial neural networks: the PAC model and its variants. Neural Computing Surveys, 1997,1:1 – 47.

8 R. E. Schapire. The strength of weak learnability. Machine Learning, 1990,5(2):197 – 227.

9 涂承胜、刀力力、鲁明羽等, Boosting 家族 AdaBoost 系列代表算法, 计算机科学,2003. 30(3). 30 – 34,145.

10 沈学华、周志华、吴建鑫等, Boosting 和 Bagging 综述, 计算机工程与应用,2000. 36(12). 31 – 32,40.

11 丁金才,天气预报评分方法评述,南京气象学院学报,1995(1). 143 – 148.