doi:10.3772/j.issn.1002-0470.2015.05.001

多分类 BP-AdaBoost 算法研究与应用①

吕雁飞②* 侯子骄③** *** 张 凯**

(*国家计算机网络应急技术处理协调中心 北京 100029)

(**中国科学院信息工程研究所 北京 100093)

(***北京航空航天大学软件学院 北京 100191)

摘 要 研究了多类别样本数据集的分类,针对传统的"一对一"或"一对多"BP-AdaBoost 算法,训练时间开销随着训练样本数以及训练样本种类的增加急剧增加,使其实际应用十分受限,尤其不适用于大规模数据分类的问题,提出了将多分类 BP 神经网络与使用多类分类指数损失函数的逐步叠加建模(SAMME)算法相结合以构造 AdaBoost 强发类的 Multi-BP AdaBoost 算法,实现模型信息的有效利用与融合增强。对传统"一对多"BP-AdaBoost 算法和 Multi-BP AdaBoost 算法进行了对比试验,结果表明,在相同测试情况下,后者有效降低了 BP-AdaBoost 训练过程中的时间开销。

关键词 AdaBoost, BP 神经网络, 二分类, 多分类

0 引 言

神经网络因具有广泛的适用性而受到诸多关注与研究。反向传播(back propagation, BP)神经网络是目前应用最广泛的神经网络模型之一,它是一种按误差反向传播算法训练的多层前馈网络,具有广泛的适应性和有效性,其主要应用在模式识别、分类等方面。分类是机器学习领域的一个重要分支,被广泛应用于各领域[1-11]。Boosting 方法是一类用来提高弱分类算法准确度的方法,其中之一是最流行的 AdaBoost 算法。"AdaBoost"是"adaptive boosting"的缩写。AdaBoost 算法通过改变训练样本的权重训练多个弱分类器,并将多个弱分类器进行线性组合,以提高分类准确率。将 BP 神经网络与 AdaBoost 算法相结合,则可充分利用各自优点,从而能进一步提高分类算法的性能。传统多分类 AdaBoost 方法是基于二分类的"一对多"[12,13]和"一对一"[12,14]的方法,但这两种方法存在一些缺点,对于 K 类问题,"一对多"方法需要构造 K 个分类器进行分类,而"一对一"方法则需要构造 $K(K-1)/2$ 个分类器,随着训练样本数以及训练样本种类的增加,二者的分类速度急剧减慢,因而这两种方法需要更长的训练时间。针对这种情况,本文提出了将多分类 BP 神经网络与专门处理多类分类问题的算法—使用多类分类指数损失函数的逐步叠加建模(Sagewise Additive Modeling using a Multi-class Exponential loss function, SAMME)算法[15]相结合,构造 AdaBoost 强分类器的多分类方法—Multi-BP-AdaBoost 算法,该方法有效降低了训练时间,并在 UCI 数据集(常用的标准测试数据集)和恶意代码应用程序编程接口(API)特征数据集上验证了算法的有效性。

① 国家自然科学基金(61271275,61202067),863 计划(2013AA013205,2013AA013204)和北京市科技计划基金(Z131100001113034,Z131100001113034612020 67)资助项目。

② 男,1984 年生,博士;研究方向:大数据处理技术,闪存数据库,机器学习;E-mail: lyf@cert.org.cn

③ 通讯作者,E-mail: houzj@buaa.edu.cn

(收稿日期:2015-02-18)

1 相关工作

1.1 Boosting 方法

Boosting 是一种提升学习方法,它的思想起源于 Valiant 提出的可能近似正确(probably approximately correct,PAC)学习模型[16]。Valiant 和 Kearns 提出了强可学习(strongly learnable)和弱可学习(weakly learnable)的概念,以及它们是否等价的问题。1990 年,Schapire[17]最先构造出一种算法,证明强学习与弱可学习是等价的,即,在 PAC 学习的模型下,一个概念是强可学习的充要条件是这个概念是弱可学习的。将弱学习方法提升为强学习算法引起了很大关注,很多 Boosting 算法被提出,AdaBoost[18]算法是 Boosting 算法家族中最具代表性的算法,被评为数据挖掘十大算法之一[19]。AdaBoost 在修改训练样本权值方面,提高前一轮被弱分类器错误分类的样本权值,降低被正确分类的的样本权值,使得后一轮更加关注前一轮被错分的样本;在弱分类器组合方面,提高分类误差率小的弱分类器的权值,提升其在表决中的作用,降低分类误差率大的弱分类器的权值,减弱其在表决中起的作用。

AdaBoost 算法具有很多优点,它的参数设置简单,只需指定迭代次数,不需要任何先验知识,一切迭代过程中的参数算法可自适应地调整。

1.2 BP 神经网络

神经网络是一种模仿动物神经网络行为特征,进行分布式并行信息处理的算法数学模型。它具有自学习与自适应的能力,可以通过输入的标注数据,分析掌握两者之间签字的规律,最终根据这些规律,预测新的输入数据。其中,目前应用最广泛的神经网络模型之一就是反向传播(BP)神经网络,它是一种按误差逆传播算法训练的多层前馈网络,具有广泛的适应性和有效性,主要应用在模式识别、分类和函数逼近等方面。

Bryson 等人在 1969 年提出了误差反向传播(error back-propagation)思想。1986 年 Rumelhart 和 McCelland 及其研究小组在 Nature 上发表其研究成果[20],从此反向传播学习算法(简称 BP 算法)得到人们的关注。采用 BP 算法的前馈型神经网络称为 BP 神经网络,简 称 BP 网络。

BP 网络拥有前向型神经网络的体系结构。图 1所示为一个具有两个隐藏层,一个输出层的多层前向神经网络结构图。

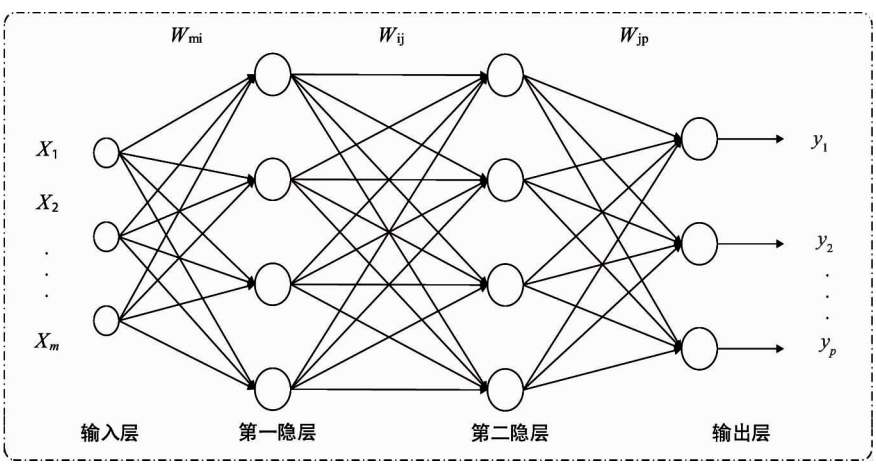

图 1 BP 网络结构图

BP 算法的主要思想是把学习过程分为信号的正向传播与误差的反向传播两个阶段:

(1) 正向传播阶段

输入信息从输入层经隐含层传向输出层,在输出端产生输出信号。在信号的向前传递过程中网络的权值固定不变,每一层神经元的状态只影响下一层神经元的状态。如果在输出层不能得到期望的输出,则转入误差信号反向传播。

(2) 反向传播阶段

未能满足精度要求的误差信号由输出端开始,以某种方式逐层向前传播,并将误差分摊给各层的所有单元,依据误差信号动态地调整各单元层的连接权重。

通过周而复始的正向传播与反向调节,神经元间的权值不断地被修正。当输出信号的误差满足精度要求时,停止学习。

2 BP-AdaBoost

BP-AdaBoost 模型即把 BP 网络作为弱分类器,反复训练 BP 网络预测样本输出,通过 AdaBoost 算法得到多个 BP 网络弱分类器组成的强分类器。

BP-AdaBoost 算法流程图如图 2 所示。

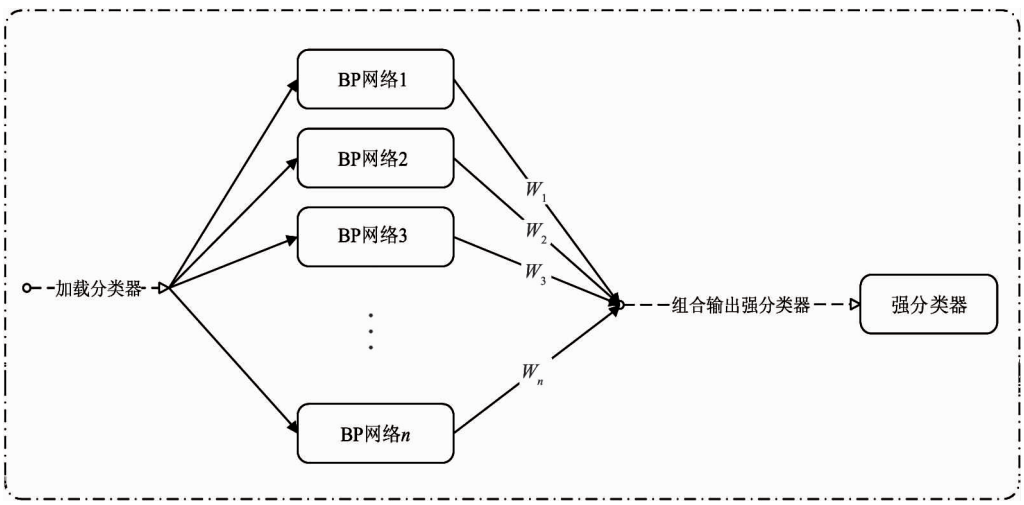

图 2 BP-AdaBoost 强分类器原理框图

BP-AdaBoost 算法的主要思想是:假设给定二分类训练数据集 $T = \{(x_i, y_i), \cdots, (x_N, y_N)\}$,其中输入数据 $x_i \in X \subseteq \mathbb{R}^n$,标签 $y_i \in \{-1, +1\}$。开始时,初始化权值分布,将各权值置为 $1/N$。之后开始迭代,每轮迭代将提高被前一轮弱分类器错误分类样本的权重值,降低被正确分类样本的权重值。最后组合各弱分类器,提高分类误差率小的弱分类器的权值,降低分类误差率大的弱分类器的权值,将一系列弱分类器构成一个强分类器。

BP-AdaBoost 算法见算法 1。

算法 1 BP-AdaBoost 算法

输入:训练数据集 $T = \{(x_i, y_i), \cdots, (x_N, y_N)\}$,其中输入数据 $x_i \in X \subseteq \mathbb{R}^n$,标签 $y_i \in \{-1, +1\}$。

输出:强分类器 $G(x)$.

1) 初始化训练数据集 T 的权值分布

$$\omega_{1i} = \frac{1}{N},\ i = 1,2,3,\cdots,N$$

2) for $m = 1:M$

① 对有权值分布的训练数据集进行训练,得到 BP 弱分类器:

$$G_m(x): X \to \{-1, +1\}$$

② 计算 $G_m(x)$ 的分类误差率

$$\mathrm{err}_m = \sum_{i=1}^{N} \omega_{mi} \| (G_m(x_i) \neq y_i)$$

③ 计算 $G_m(x)$ 的系数

$$\alpha_m = \frac{1}{2} \log \frac{1 - \mathrm{err}_m}{\mathrm{err}_m}$$

其中对数是自然对数。

④ 更新训练数据集权值分布

$$\omega_{m+1,i} = \frac{\omega_{mi}}{Z_m} \exp(-\alpha_m y_i G_m(x_i)),\ i = 1,2,\cdots,N$$

⑤ 其中 Z_m 是标准化因子

$$Z_m = \sum_{i=1}^{N} \omega_{mi} \exp(-\alpha_m y_i G_m(x_i))$$

end

3) 得到最终分类器

$$G(x) = \mathrm{sign}(\sum_{m=1}^{M} \alpha_m G_m(x))$$

3 多分类 BP-AdaBoost

本文中,BP-AdaBoost 实现多分类方法有两种:

第一种。通过经典的“一对多”方法(即构造一系列二分类器,每一个分类器都将其中一类与其他类划分开),对输入的样本进行预测。具体地,对于具有 K 个类别的数据集,构造 K 个 BP-AdaBoost 强分类器,其中,在构造第 n 个强分类器时,将属于第 n 类的样本集标注为正类,除此之外的其他类标注为负类。在进行数据预测时,对被测试数据分别计算其属于各个类别的决策值,选取最大决策值所对应的类别作为其所属类别。其算法见算法 2。

算法 2 “一对多”BP-AdaBoost 算法

输入:训练数据集 $T = \{(x_1, y_1), \cdots, (x_N, y_N)\}$,其中输入数据 $x_i \in X \subseteq \mathbb{R}^n$,标签 $y_i \in \{-1, +1\}$。

输出:强分类器 $G(x)$,预测分类结果.

(1) for $i = 1:K$

1) 将 $y_i = n$ 的样本设置为证样本,$y_i \neq n$ 的样本设置为负样本。(其中 n 为数据的类别)

2) 初始化训练数据集 T 的权值分布

$$\omega_{1i} = \frac{1}{N},\ i = 1,2,3,\cdots,N$$

3) for $m = 1:M$

①对有权值分布的训练数据集进行训练,得到 BP 弱分类器:

$$G_m(x): X \to \{-1, +1\}$$

②计算 $G_m(x)$ 的分类误差率

$$\mathrm{err}_m = \sum_{i=1}^{N} \omega_{mi} \parallel (G_m(x_i) \neq y_i)$$

③计算 $G_m(x)$ 的系数

$$\alpha_m = \frac{1}{2}\log\frac{1 - \mathrm{err}_m}{\mathrm{err}_m}$$

其中对数是自然对数。

④更新训练数据集权值分布

$$\omega_{m+1,i} = \frac{\omega_{mi}}{Z_m}\exp(-\alpha_m y_i G_m(x_i)),\ i = 1,2,\cdots,N$$

⑤其中 Z_m 是标准化因子

$$Z_m = \sum_{i=1}^{N} \omega_{mi}\exp(-\alpha_m y_i G_m(x_i))$$

end

4) 得到最终分类器

$$G(x) = \mathrm{sign}\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right)$$

输出测试数据的所属类别的决策值

end

(2)根据输出决策值的大小觉得测试数据所属类别

第二种。通过构造多类分类器,对输入的样本进行预测,结合使用多类分类指数损失函数的逐步叠加建模(SAMME)算法构造强分类器,最后输出多分类的结果。具体地,对于具有 K 个类别的数据集,使 BP 网络弱分类器直接输出多分类结果的决策值,选取最大决策值所对应的类别作为其所属类别。其算法见算法 3。

算法 3 Multi-BP-AdaBoost 算法

输入:训练数据集 $T = \{(x_1, y_1), \cdots, (x_N, y_N)\}$,其中输入数据 $x_i \in X \subseteq \mathbb{R}^n$,标签 $y_i \in \{-1, +1\}$。

输出:强分类器 $G(x)$,预测分类结果.

1) 初始化训练数据集 T 的权值分布

$$\omega_{1i} = \frac{1}{N},\ i = 1,2,3,\cdots,N$$

2) for $m = 1:M$

① 对有权值分布的训练数据集进行训练,得到 BP 弱分类器:

$$G_m(x): X \to \{1,2,\cdots,K\}$$

② 计算 $G_m(x)$ 的分类误差率

$$\mathrm{err}_m = \sum_{i=1}^{N} \omega_{mi} \parallel (G_m(x_i) \neq y_i)$$

③ 计算 $G_m(x)$ 的系数

$$\alpha_m = \frac{1}{2}\log\frac{1 - \mathrm{err}_m}{\mathrm{err}_m} + \log(K-1)$$

其中对数是自然对数。

④ 更新训练数据集权值分布

$$\omega_{m+1,i} = \frac{\omega_{mi}}{Z_m}\exp(-\alpha_m y_i G_m(x_i)),\ i = 1,2,\cdots,N$$

⑤ 其中 Z_m 是标准化因子

$$Z_m = \sum_{i=1}^{N} \omega_{mi}\exp(-\alpha_m y_i G_m(x_i))$$

end

3) 得到最终分类器

$$G(x) = \mathrm{sign}\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right)$$

4) 输出测试数据的所属类别的决策值,并根据输出决策值的大小觉得测试数据所属类别

4 试 验

本节通过试验比较传统“一对多”BP-AdaBoost 算法和 Multi-BP-AdaBoost 算法,来验证 Multi-BP-AdaBoost 算法的有效性。试验数据采用 UCI 的 Wine 公开数据集和恶意代码 API 特征数据集。为保证试验的公正性并不失一般性,试验中已标注集均进行随机选取,且所有结果均为相同参数下无人为干涉 10 次试验的平均值。试验用机的 CPU 型号为 Intel Xeon E5-2630,内存为 32GB。

4.1 Wine 数据集试验条件及结果

Wine 数据集[21]的数据是生长在意大利同一地区不同品种的葡萄酒,通过化学分析的结果,分析确定了在 3 种葡萄酒中发现的 13 种成分的数量。其中,类 1 数量为 59 个,类 2 数量为 71 个,类 3 个数为 48 个,总个数为 178 个。

试验通过传统“一对多”BP-AdaBoost 算法和 Multi-BP-AdaBoost 算法实现多分类,并分别记录计算耗时,以及计算算法误差率。对所有算法,弱分类器个数均设置为 10 个。

最初设定训练集个数为 10 个,之后逐渐增加训练集的规模,研究两种算法的算法性能差别与训练集个数的关系。试验结果如图 3 和图 4 所示。图 3 为两种算法的效率对比图,图 4 为两种算法的误差率对比图。

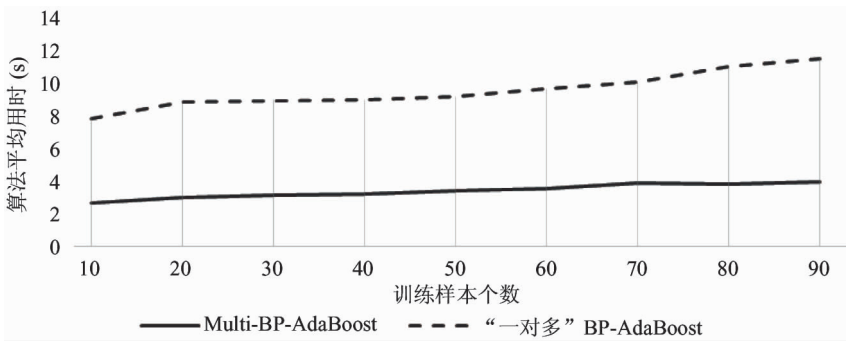

图 **3** 两种多分类实现算法应用于 **wine** 数据集的效率对比图

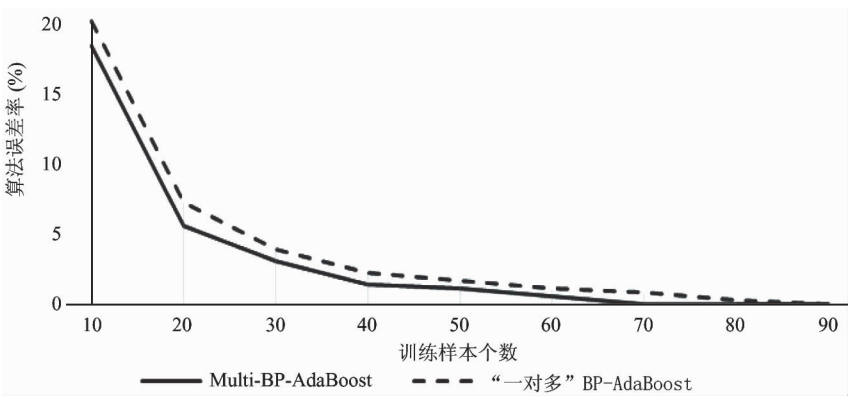

图 **4** 两种多分类实现算法应用于 **wine** 数据集的误差率对比图

结合图 3、图 4 可见,总体上,Multi-BP-AdaBoost 算法的性能要优于“一对多”BP-AdaBoost 算法的性能。具体而言,Multi-BP-AdaBoost 算法的运算效率是“一对多”BP-AdaBoost 算法运算效率的 3 倍左右,且相比于“一对多”BP-AdaBoost 算法,Multi-BP-AdaBoost 算法的计算误差率也有提升。

4.2 恶意代码数据集实验条件及结果

恶意代码(Malware)通常是指故意创建用来执行未经授权并通常是有害行为的软件程序。通过动态分析恶意代码对 Win32 API 函数的调用情况,将

其归纳记录到对应的 144 个动态行为特征项中，得到特征矩阵，构成该数据库。数据中类 1 为非恶意代码样本，数量为 479 个。2 – 5 类为恶意代码样本，共有 361 个，其中类 2 为 68 个，类 3 为 96 个，类 4 为 157 个，类 5 为 40 个。

试验通过传统“一对多”BP-AdaBoost 算法和 Multi-BP-AdaBoost 算法实现恶意代码多分类，并分别记录计算耗时，以及计算算法误差率。所有算法中弱分类器个数设置为 10 个。

最初设定训练集个数为 200 个，之后逐渐增加训练集的规模，研究两种算法的算法性能差别与训练集个数的关系。实验结果如图 5 和图 6 所示。其中图 5 为两种算法的效率对比图，图 6 为两种算法的误差率对比图。

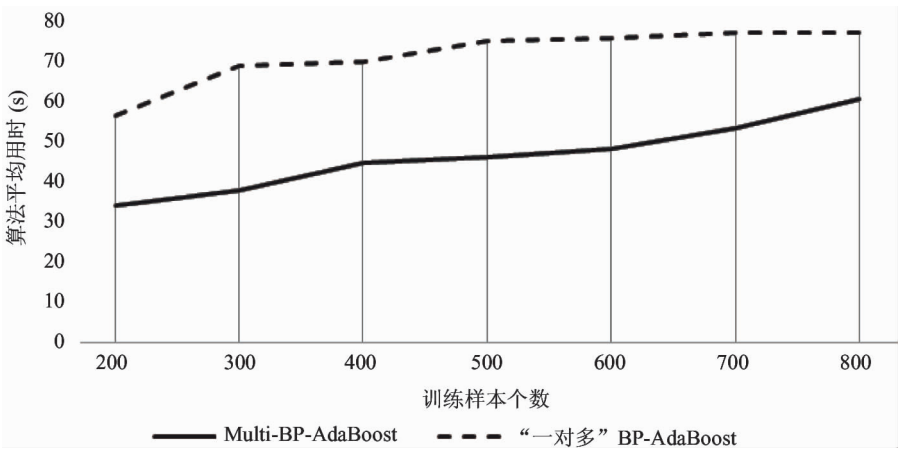

图 **5** 两种多分类实现算法应用于恶意代码数据集的效率对比图

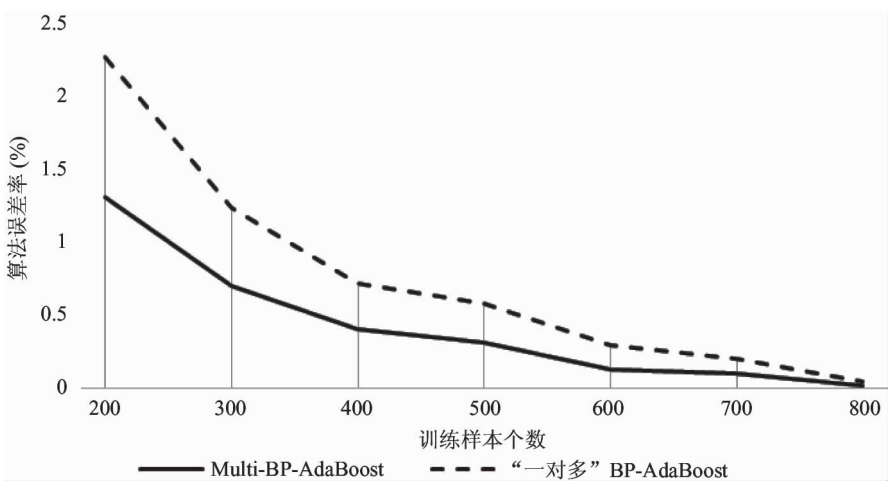

图 **6** 两种多分类实现算法应用于恶意代码数据集的误差率对比图

结合图 3、图 4 可见，总体上，Multi-BP-AdaBoost 算法的性能要优于“一对多”BP-AdaBoost 算法的性能。具体而言，Multi-BP-AdaBoost 算法的运算效率是“一对多”BP-AdaBoost 算法运算效率的 1.5 倍左右，且相比于“一对多”BP-AdaBoost 算法，Multi-BP-AdaBoost 算法的计算误差率有较大提升。

4.3 试验小结

综合以上试验结果，本文提出的 Multi-BP-AdaBoost 算法将多分类 BP 算法与 SAMME 算法相结合，在多分类问题上，与“一对多”BP-AdaBoost 算法相比，有效提升了运算效率及分类准确率。

5 结 论

本文针对传统“一对多”、“一对一”多分类算法运算效率低下,以及二分类造成训练种类不对称的问题,提出了将多分类 BP 算法与 SAMME 算法相结合的方法,该方法有效提升了算法的运算效率及分类准确率。通过在 Wine 数据集与恶意代码数据集上开展分类试验,充分验证了在相同测试条件下,Multi-BP-AdaBoost 算法的有效性。

参考文献

[1] Zhang XY, Wang S, Yun X. Bidirectional active learning, a two-way exploration into unlabeled and labeled dataset. *IEEE Transactions on Neural Networks and Learning Systems*, 2015: 1-11. doi: 10.1109/TNNLS.2015.2401595

[2] Zhang X Y. Interactive patent classification based on multi-classifier fusion and active learning. *Neurocomputing*, 2014, 127: 200-205

[3] Zhang X Y, Wang S, Zhu X, et al. Update vs. upgrade: modeling with indeterminate multi-class active learning. *Neurocomputing*, 10.1016/j.neucom.2015.03.056

[4] Zhang X Y, Cheng J, Xu C, et al. Multi-view multi-label active learning for image classification. In: Proceedings of the IEEE International Conference on Multimedia and Expo, Cancun, Mexico, 2009. 258-261

[5] Zhang X Y, Cheng J, Lu H, et al. Selective sampling based on dynamic certainty propagation for image retrieval. In: Proceedings of the Advances in Multimedia Modeling, Kyoto, Japan, January, 2008. 425-435

[6] Zhang X Y, Xu C, Cheng J, et al. Automatic semantic annotation for video blogs. In: Proceedings of the IEEE International Conference on Multimedia and Expo, Hannover, Germany, 2008. 121-124

[7] Zhang X Y, Cheng J, Lu H, et al. Weighted co-SVM for image retrieval with MVB strategy. In: Proceedings of the IEEE International Conference on Image Processing, San Antonio, USA, 2007. 517-520

[8] Zhang X Y. Preference modeling for personalized retrieval based on browsing history analysis. *IEEE Transactions on Electrical and Electronic Engineering*, 2013, 8 (S1): 81-87

[9] Zhang X Y. Effective search with saliency-based matching and cluster-based browsing. *High Technology Letters*, 2013, 19(1):105-109

[10] Zhang X Y, Cheng J, Xu C, et al. Effective annotation and search for video blogs with integration of context and content analysis. *IEEE Transactions on Multimedia*, 2009,11(2): 272-285

[11] Zhang X Y, Cheng J, Xu C, et al. Effective annotation and search for video blogs with integration of context and content analysis. *IEEE Transactions on Multimedia*, 2009,11(2): 272-285

[12] Hsu C W, Lin C J. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 2002,13(2):415-425

[13] B1anz V, Schuolkopf B, Bulthoff H, et al. Comparison of View-based Object Recognition Algorithms Using Realistic 3D Models. In: International Conference on Artificial Neural Networks, Springer Verlag, Berlin, Germany, 1996. 375-381

[14] Schuolkopf B, Burges C J C, SmolaA J, et al. Advances in Kernel Methods: Support Vector Learning. Cambridge, MA, MIT Press, 1999. 255-268

[15] Zhu J, Zou H. Multi-class AdaBoost, *Statistics and Its Interface*, 2009,(2): 349-360

[16] Valiant L G. A Theory of the Learnable. *Communications of the ACM*, 1984, 27(11): 1134-1142

[17] Schapire R E. The strength of weak learnability. *Machine Learning*, 1990,5(2): 197-227

[18] Freund Y, Schapire R E. A Decision-theoretic generalization of online learning and an application to Boosting. *Journal of Computer and System Sciences*, 1997, 55 (01): 119-139

[19] Wu X D, Kumar V, Zhou Z H, et al. The top ten algorithms in data mining. New York: CRC Press, 2009. 127-149

[20] Rumelhart D E, Hlinton G E, Williams R J. Learning representations of back-propagation errors, *Nature*, 1986, 1(323):533-536

[21] UC Irvine Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Wine

Study of Multi-class BP-AdaBoost and its application

Lv Yanfei[*], Hou Zijiao[**,***], Zhang Kai[**]

([*] National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029)

([**] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093)

([***] School of Software, BeiHang University, Beijing 100191)

Abstract

The study focused on the classification of the dataset referred to multi-class samples, and paid attention to the problem that the time cost of traditional "one-against-one" or "one-against-all". BP-AdaBoost algorithm increases rapidly with the increase of the sample amount and the sample class number, thus leading to the hindrance to its practical application, especially when dealing with large-scale datasets. Then, to solve this problem, the multi-BP-AdaBoost algorithm was proposed by combinig multi-class BP neural networks with the algorithm of Stagewise Additive Modeling using a Multi-class Exponential loss function (SAMME) to construct a strong AdaBoost classifier. The algorithm can effectively use and fuse model information to improve its performance. The test on the traditional "one-against-all" BP-AdaBoost algorithm and the proposed multi-BP-AdaBoost algorithm was performed, and the results showed that the latter had the better abiligy in reducing the time cost than the former under the same testing conditions.

Key words: AdaBoost, BP neural network, binary classification, multi-class classification