

文章编号:1001-9081(2016)02-0291-04

DOI:10.11772/j.issn.1001-9081.2016.02.0291

基于概率校准的集成学习

姜正申¹, 刘宏志^{2*}

(1. 北京大学 信息科学技术学院, 北京 100871; 2. 北京大学 软件与微电子学院, 北京 102600)

(* 通信作者电子邮箱 liuhz@pku.edu.cn)

摘要:针对原有集成学习多样性不足而导致的集成效果不够显著的问题,提出一种基于概率校准的集成学习方法以及两种降低多重共线性影响的方法。首先,通过使用不同的概率校准方法对原始分类器给出的概率进行校准;然后使用前一步生成的若干校准后的概率进行学习,从而预测最终结果。第一步中使用的不同概率校准方法为第二步的集成学习提供了更强的多样性。接下来,针对校准概率与原始概率之间的多重共线性问题,提出了选择最优(choose-best)和有放回抽样(bootstrap)的方法。选择最优方法对每个基分类器,从原始分类器和若干校准分类器之间选择最优的进行集成;有放回抽样方法则从整个基分类器集合中进行有放回的抽样,然后对抽样出来的分类器进行集成。实验表明,简单的概率校准集成学习对学习效果的提高有限,而使用了选择最优和有放回抽样方法后,学习效果得到了较大的提高。此结果说明,概率校准为集成学习提供了更强的多样性,其伴随的多重共线性问题可以通过抽样等方法有效地解决。

关键词:集成学习;概率校准;多重共线性;有放回抽样;随机子空间

中图分类号: TP391 **文献标志码:**A

Ensemble learning based on probability calibration

JIANG Zhengshen¹, LIU Hongzhi^{2*}

(1. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China;

2. School of Software and Microelectronics, Peking University, Beijing 102600, China)

Abstract: Since the lackness of diversity may lead to bad performance in ensemble learning, a new two-phase ensemble learning method based on probability calibration was proposed, as well as two methods to reduce the impact of multiple collinearity. In the first phase, the probabilities given by the original classifiers were calibrated using different calibration methods. In the second phase, another classifier was trained using the calibrated probabilities and the final result was predicted. The different calibration methods used in the first phase provided diversity for the second phase, which has been shown to be an important factor to enhance ensemble learning. In order to address the limited improvement due to the correlation between base classifiers, two methods to reduce the multiple collinearity were also proposed, that is, choose-best and bootstrap sampling method. The choose-best method just selected the best base classifier among original and calibrated classifiers; the bootstrap method combined a set of classifiers, which were chosen from the base classifiers with replacement. The experimental results showed that the use of different calibrated probabilities indeed improved the effectiveness of the ensemble; after using the choose-best and bootstrap sampling methods, further improvement was also achieved. It means that probability calibration provides a new way to produce diversity, and the multiple collinearity caused by it can be solved by sampling method.

Key words: ensemble learning; probability calibration; multiple collinearity; bootstrap sampling; random subspace

0 引言

集成学习是一种通过组合多种不同分类器来提高学习效果的机器学习方法,被组合的分类器通常也被称为基分类器。许多研究表明,这种通过组合基分类器的学习方法可以有效地提高学习精度^[1]。正因为如此,集成学习模型已经被应用到许多领域,例如时间序列预测^[2]、协同过滤^[3]、信用卡欺诈检测^[4]等。

已有的对集成学习的理论分析表明,基分类器之间的多样性(diversity)是提高集成学习效果的一个重要因素^[5]。基

于此,本文提出了一种新的产生多样性的方法,即概率校准(probability calibration)。对于二分类问题,简单地将基分类器的分类结果进行多数投票并不是最佳的组合方法,一种更好的方法是让基分类器给出样本属于每个类别的概率,然后对这些概率进行组合。然而,许多分类器给出的概率往往是不准确且缺乏合理解释的。鉴于此,概率校准的方法被提出来以对这些概率进行修正。

本文使用不同的概率校准技术对基分类器输出的概率进行校准,从而得到具有多样性的基分类器,然后使用逻辑回归进行概率的集成。由于采用了不同的概率校准方法,算法就

收稿日期:2015-08-29;修回日期:2015-09-11。 基金项目:国家自然科学基金资助项目(61232005);CCF-腾讯科研基金资助项目。

作者简介:姜正申(1990-),男,吉林松原人,博士研究生,主要研究方向:模式识别; 刘宏志(1982-),男,湖北武汉人,副教授,博士,CCF会员,主要研究方向:信息融合、模式识别。

可以从不同的角度去校准概率,这样就提高了基分类器的多样性。实验表明,这种多样性确实提高了学习精度,与未加入概率校准技术相比,提高了 0.6% 以上。

需要指出的是,本文提出的概率校准是提高基分类器多样性的一个新的角度,完全可以混合其他增加多样性的技术使用,例如概率校准混合 bootstrap 抽样,即成为有概率校准的 Bagging 算法;混合特征空间的抽样方法,即成为有概率校准的随机子空间算法,等等。

1 相关工作

1.1 概率校准

概率校准是一种常被社会科学和心理学使用的研究方法。对于某个带有不确定性的问题,每个人会给出自己的判断,并同时给出对自己判断的把握程度,这个把握程度就称为主观概率。而事物实际发生的概率就称为客观概率。而概率校准就是衡量主观概率与客观概率之间一致程度的方法^[6]。如果一个人的主观概率判断经常比客观概率大,那么就说这个人是过分自信的;相反的就称为过低自信。

在机器学习领域,尽管许多分类器都可以给出样本归属各个类别的概率,例如决策树、K 最近邻、支持向量机等,但是这些概率往往缺乏合理的解释。例如支持向量机中,常常是根据样本点与分类超平面的距离来计算概率,然而这样的计算是缺乏依据的。事实上,多数分类器所给出的概率分布与样本的真实概率分布常常相差甚远,所以许多文献称这些计算结果为分数(score)而不是概率。经验上,朴素贝叶斯是一个过分自信的分类器,而支持向量机是过低自信的分类器,逻辑回归往往能给出较好的概率预测。

为了修正这些概率的偏差,已有学者提出了一些方法,这些方法称为概率校准。这种技术可以部分修正上面提到的概率偏差,来尽可能地得到既不过分自信也不过低自信的概率预测。然而目前并没有一个可以完全修正概率偏差的模型,使用不同的概率校准方法将从不同的方面去修正概率。

目前,机器学习领域较为成熟的两个概率校准方法为 Isotonic 回归^[7] 和 Sigmoid 回归^[8]。

Isotonic 回归的目标是找到一个能够拟合数据点的单调递增函数,同时最小化均方误差。使用 Isotonic 回归进行概率校准时,就是以分类器输出的每个样本的概率值作为自变量,以样本的真实值作为因变量,来拟合一个单调递增函数。对样本进行测试时,先根据分类器给出原始概率,然后使用拟合出来的函数计算校准后的概率。

Sigmoid 回归就是令数据点拟合一个 sigmoid 函数。以此方法校准概率,就是使原始概率拟合为一个 sigmoid 函数,然后以此校准概率。这种方法实际上是 Isotonic 校准的特殊情形。

1.2 集成学习

集成学习常常又被称为模型组合、模型融合等。但是通常情况下,集成学习中的基分类器是采用同种算法训练出来的,例如 Boosting^[10] 中的基分类器为决策树;而模型组合中,基分类器可以使用多种不同算法,例如将决策树、K 最近邻等进行混合,从而得到新的集成模型。

目前,学者已经提出众多的集成学习方法,其中效果最好、应用最为广泛的两个算法为 Bagging^[9] 和 Boosting^[10]。在

这两种算法的基础上,学界提出了大量的变体,例如 AdaBoost^[11]、随机森林^[12]、随机子空间^[13]等。

本文使用了多种不同算法进行组合,所以属于模型组合。但为了叙述上的方便,统一称为集成学习而未加区分。

多样性是影响集成学习效果的一个关键因素。不同的集成学习算法采用不同的方式来产生多样性。例如 Bagging 通过对样本进行有放回的抽样来得到不同的样本集,进而训练出一系列基分类器;而 Boosting 则是每轮迭代都根据前一轮训练的误差来调整样本的权重,这样每轮训练使用的样本权重都不相同,使得训练出来的基分类器能够从不同的角度对样本建模,后续分类器更加注重那些被分类错误的样本。最后,这些具有多样性的基分类器会根据一定的方式进行组合,并得到一个最终的预测值。

集成学习可以分为有监督的和无监督的两种。有监督集成学习又分为集成回归和集成分类。无监督集成学习又被称为共识聚类(consensus clustering)。本文主要研究集成分类问题,但推广到集成回归问题并不困难。

1.2.1 集成学习的泛化误差

目前,已有一些针对集成学习理论的研究。这其中,对泛化误差进行成分分解具有较强的实践指导意义。

1996 年,Ueda 等^[14] 提出了集成学习泛化误差的偏差-方差-协方差分解,这种方法假定:

$$\hat{f}_F(x) = \frac{1}{K} \sum_{i=1}^K [\hat{f}_i(x)] \quad (1)$$

其中 K 为参与集成的基分类器数目。式(1)的意义为集成学习的输出是基分类器输出的简单平均。

这样,集成学习的均方误差可以分解为如下三个部分:

$$E[(\hat{f}_F - f)^2] = \overline{\text{bias}}^2 + K^{-1} \overline{\text{var}} + (1 - K^{-1}) \overline{\text{covar}} \quad (2)$$

其中:

$$\overline{\text{bias}} = \frac{1}{K} \sum_{i=1}^K [E_i(f_i) - f] \quad (3)$$

$$\overline{\text{var}} = \frac{1}{K} \sum_{i=1}^K \{E_i\{[\hat{f}_i - E_i(\hat{f}_i)]^2\}\} \quad (4)$$

$$\begin{aligned} \overline{\text{covar}} &= [K(K-1)]^{-1} \cdot \\ &\sum_{i=1}^K \sum_{j=1, j \neq i}^K E_{i,j} \{[\hat{f}_i - E_i(\hat{f}_i)][\hat{f}_j - E_j(\hat{f}_j)]\} \end{aligned} \quad (5)$$

其中期望的下标 i 和 j 表示在对应的训练集上进行求期望。

上面的泛化误差分解说明,要提高集成学习的效果,不仅应该提高基分类器的准确性,即降低基分类器的偏差,而且应该提高基分类器之间的差异性,即降低基分类器间的协方差。

1.2.2 集成学习的多样性

上面提到的差异性在集成学习领域常被称为多样性(diversity),即基分类器之间的对立(disagreement)程度。不同类别的基分类器,可以从不同侧面对数据建模,将这些分类器组合起来,就可能达到单一分类器不可能达到的精度。通常,基分类器之间相差越大,集成效果就越好。

2 使用概率校准的两层集成模型

本文基于概率校准这一新的产生多样性的途径,提出了一种新的两层集成学习模型,如图 1 所示。

本文的叙述以二分类问题为例进行说明,但本文提出的算法可以轻易地扩展到多类的分类问题上。

为了训练模型,需要将样本划分为训练集、验证集和测试集。

在模型的第一层,将不同的基分类器与不同的概率校准方法进行组合,然后使用训练集进行训练,就可以得到一系列的基分类器,这些基分类器共同组成了模型池(Base Model Pool)。

接下来,使用模型池中的模型对训练集进行预测。这样,对训练集中的每个样本,模型池中的模型会给出一系列的预测概率,这些概率构成了样本的新特征集合。所有样本的预测概率最终组合为概率池(Probability Pool),作为下一层的训练集。

模型的第二层以上一层给出的校准概率作为特征,以原始样本的类别标记为预测目标,使用逻辑回归进行学习,从而训练出第二层模型。最后,使用测试集来测试整个算法的学习效果。

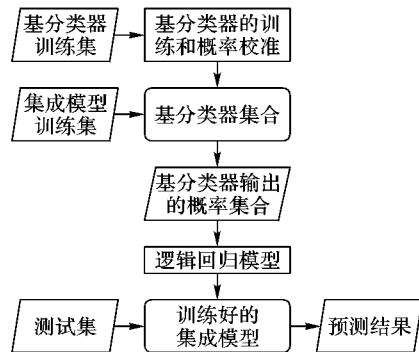


图1 实验流程

3 概率校准对多样性的影响的检验和分析

3.1 数据说明

本文使用 P2P 贷款网站 Prosper.com 公开的数据集进行实验,根据用户的个人信息对用户发起的贷款是否违约进行预测。采用的是 2006—2008 年的数据,因为这段时间发生的用户贷款,其是否最终违约可以在数据集中得到,从而得到完整的有标注数据。

在 Prosper 原始数据集的基础上还加入了一些美国宏观经济的数据,因为这些数据可能会影响用户的违约行为。

另外,考虑到 Prosper.com 网站还具有交友、群组等社交功能,也对这些社交关系进行了简单的分析处理,并将每笔贷款对应的社交关系指标也作为特征加入了数据集。

处理后的数据集共有 158 个特征,29 106 个样本。

3.2 实验结果

本文选择了如下几种机器学习算法作为基分类器:朴素贝叶斯、支持向量机、K 最近邻、Bagging 的决策树,以及 AdaBoost 算法的单层决策树(decision stumps)。这五大类算法各自使用不同的参数或设定,可以产生许多模型,本文使用了类似于网格搜索(grid search)的方法,对这五类算法的参数进行了枚举,在枚举出的模型中,去掉了个别效果极端的模型后,得到共 123 个模型。

对于概率校准,本文共使用了三种方法:无校准(None)、Isotonic 校准以及 Sigmoid 校准。这三种校准方法分别与上面 123 个模型进行组合,共产生 369 个基分类器。

首先分别对无校准、Isotonic 校准以及 Sigmoid 校准后的

基分类器独立进行了集成,然后将三种概率校准方法进行两两组合集成,结果如图 2 所示。图 2 中:横坐标为算法第二层采用的逻辑回归所使用的正则化参数,越往后参数值越大;纵坐标为模型的曲线下面积(Area Under Curve, AUC) 分数,这里的曲线为受试者工作特征(Receiver Operating Characteristic, ROC) 曲线。

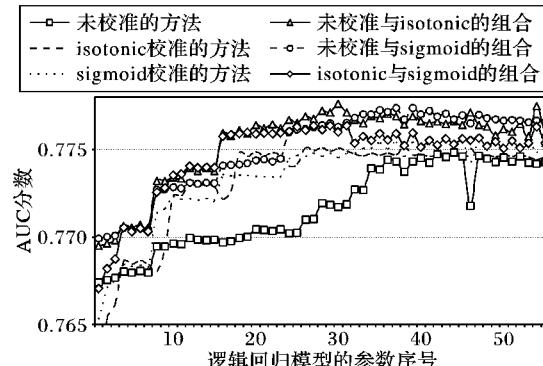


图2 三种校准方法各自集成与两两组合进行集成的比较

可以看出,Isotonic 校准和 Sigmoid 校准后进行集成均比未校准的集成要好。在正则化参数过小的情况下,校准的效果并不好,但随着正则化参数的调整,校准后的学习效果开始优于未校准的学习算法。Isotonic 校准平均比未校准高出 0.075%, Sigmoid 校准比未校准高出 0.145%;如果将前面正则化参数过小的部分排除,那么 Isotonic 校准比未校准高出 0.260%,而 Sigmoid 校准也比未校准高出 0.266%。

通过图 2 还可以清楚地看出,两两组合后的结果要好于各自集成的结果。无校准与 Isotonic 校准集成后,比无校准单独集成的效果平均高出 0.432%,排除前面参数过小的数据后,这一数值变为 0.494%;无校准与 Sigmoid 校准集成的结果比无校准单独集成平均高出 0.407%,排除前面数据后为 0.449%;Isotonic 校准与 Sigmoid 校准集成后,由于参数过小时效果影响较大,平均效果仅提高了 0.299%,排除参数过小数据后,平均提高了 0.405%。

三种校准方法共 369 个分类器全部集成的结果如图 3 所示。可以看出,全部集成的结果优于未校准的集成。整体上,全部集成的结果比未校准的集成方法高 0.423%;如果排除前面参数过小的部分,那么这一数值为 0.429%。

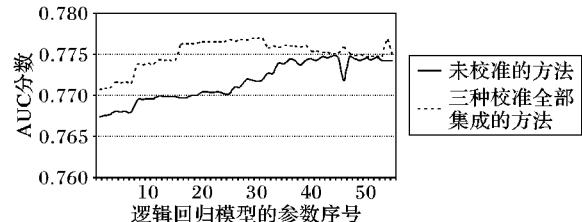


图3 三种校准方法全部参与集成的结果

3.3 结果分析

经过校准的分类器在集成时效果是好于未校准的分类器的。这个结果是意料之中的,因为概率校准以后的基分类器,其精确度比未校准的概率要好。基分类器精度的提高,也使集成学习的效果得到提高。

两两组合进行集成时,集成效果并没有被未校准的分类器拉低,反而是比任何一种校准方法单独集成的效果都要好。这种情况反映了多样性在集成学习中的重要意义。组合两种

概率校准方法以后,提高整个基分类器集合的多样性,从而提高了集成学习的性能。

第三个实验中将三种校准方法全部集成了起来。尽管在第二层集成模型的正则化参数较小时,集成效果有了进一步的提高,然而在后面正则化参数过大时,全部集成的效果开始下降,且降低到与未校准的集成差不多的程度。这表明,集成模型在这里发生了过拟合。如果排除掉参数过小以及过大的部分,那么全部集成的效果比未校准的基分类器单独集成好 0.804%。

上述过拟合情况可能是基分类器之间的相关性导致的。由于基分类器是根据五类机器学习算法,使用不同参数后,再配合以三种不同的校准方法得到的,这样使得基分类器之间难以避免地有较大的相关性,第二层使用逻辑回归时,这种相关性就会对学习效果造成不良影响。

这种相关性在线性回归中又被称为多重共线性。逻辑回归虽然是一种非线性回归,但依然会受到自变量之间相关性的影响。为此,本文使用了两种办法来降低这种影响。

4 降低相关性影响的两种方法

从上面的结果看,尽管三种校准方法组合起来的效果最好,但是这却带来了相关性的问题,导致学习效果未能持续提高。为了降低这种相关性的影响,本文使用了两种办法:选择最优(choose-best)和有放回抽样(bootstrap)方法。

选择最优方法中,对每一种机器学习模型,从无校准、Isotonic 校准和 Sigmoid 校准中选择一个效果最好的基分类器,这样会挑选出 123 个基分类器进行集成。这样将基分类器的数量降低为原来的 1/3,从而降低了多重共线性,而且最大可能地保留了学习效果。

选择最优方法的结果如图 4 所示。可以看出,这种方法虽然在参数过小的情况下较差,但是在参数较大的情况下,取得了很好的结果。这种方法好于三种校准方法的各自集成,已经与两两组合和全部集成效果相当,尤其是在参数很大的一侧,取得了比全部集成更好的效果。这说明,此方法部分解决了相关性的问题。这种方法的缺点是,它只利用了 1/3 的基分类器,导致其他的分类器被浪费,而这些分类器却未必是没用的。

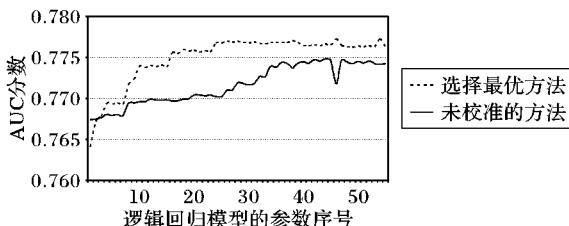


图 4 使用选择最优方法降低相关性的实验结果

本文接下来使用了另一种可以较为充分利用基分类器的方法——有放回抽样方法。这是一种类似于 Bagging 的算法,通过从 369 个基分类器中,有放回地抽取若干次,从而得到一个基分类器的集合,再进行集成。与 Bagging 类似的,算法会进行多次抽取,从而得到多个集成模型,最后这些集成模型进行投票来得到最终的分类结果。本质上,这种算法是在第二层运行了一次随机子空间模型。显然,这种抽样的方法,其效果受到抽样比例的影响会比较大。图 5 显示了抽样比例与学习效果的关系。可以看出,抽样比例在 0.4~0.6 时集成

学习的效果最好,本文下面的实验将以 0.6 的比例进行抽样。

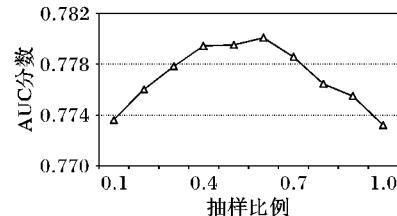


图 5 抽样比例对学习效果的影响

有放回抽样方法的结果如图 6。可以看到,抽样法的结果比前面的所有集成方法都好,在参数过大的一侧,其学习效果依然未出现回落,这说明此方法有效地避免了相关性的问题。不仅如此,使用抽样方法后,集成的最好结果的 AUC 分数达到了 0.780。整体上,有放回抽样法比未校准的集成好 0.540%,排除参数过小的部分后,达到了 0.665%。

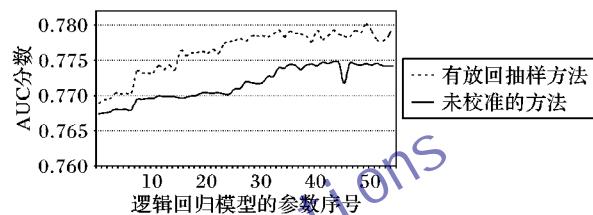


图 6 使用有放回抽样方法降低相关性的实验结果

5 结语

本文提出了一种基于概率校准的集成学习方法,并通过实验证明了概率校准对学习效果的影响。本文使用了三种不同的概率校准方法对 369 个基分类器进行了校准,从而得到了 369 个不同的基分类器。实验第一部分是使用三种不同概率校准方法得到的基分类器分别进行集成、两两组合进行集成,以及全部用来集成。实验结果显示,两两组合的效果明显好于单独集成的效果,而全部集成的效果略好于两两组合,但在正则化参数过高的情况下开始下降。本文将这种情况归因为基分类器间的相关性,所以接着做了第二部分的实验:使用两种方法来降低相关性的影响。一种方法是对每种算法的无校准、Isotonic 校准和 Sigmoid 校准中选择一个最好的进行集成。结果显示,这种方法明显提高了正则化参数过高时的学习效果,说明相关性问题得到了部分的解决。另一种方法是使用了类似于随机子空间的模型,对 369 个基分类器进行有放回的随机抽样。经过实验,此方法再次提高了学习效果,AUC 分数达到了 0.780,是所有实验中的最好结果。这表明此方法有效解决了相关性的问题。

综合本文实验,可得以下结论:概率校准作为一种提高集成多样性的新途径,可以有效提高集成学习效果,但是同时必须解决相关性的问题。随机抽样作为一种简单而有效的方法,降低了基分类器之间的相关性,同时提高了集成学习的效果。

尽管本文取得了较好的结果,但依然有改进的空间:

首先,本文只采用了两种校准方法,这对于集成学习来讲还是远远不够的。然而直到目前,依然没有足够的校准方法可供使用。考虑到本文的方法实际上相当于对第一层输出的概率作了一定的变换,那么将概率校准替换为其他变换方法也应该是可行的。所以进一步的工作中,需要继续寻求有效的概率变换方法。

(下转第 407 页)

- (1): 71–83. (FENG D G, ZHANG M, ZHANG Y, et al. Cloud computing security research [J]. *Journal of Software*, 2011, 22(1): 71–83.)
- [10] 李红霞. 云计算中身份认证与访问控制管理系统的实现策略研究[D]. 北京: 北京邮电大学, 2011: 8–12. (LI H X. Research on strategy of implementation of identity authentication and access control management system in cloud computing [D]. Beijing: Beijing University of Posts and Telecommunications, 2011: 8–12.)
- [11] 曹源. 面向跨域联邦环境的身份管理关键技术研究[D]. 长沙: 国防科学技术大学, 2013: 21–23. (CAO Y. Research of key techniques for identity management in across domains federal environment [D]. Changsha: National University of Defense Technology, 2013: 21–23.)
- [12] ZHANG W, LI Y. Federation access control model based on Web-service [C]// ICEE '10: Proceedings of the 2010 International Conference on E-Business and E-Government. Washington, DC: IEEE Computer Society, 2010: 38–41.
- [13] STANDARDWORKING O, CAHILL C P, AOL J, et al. Assertions and protocols for the OASIS Security Assertion Markup Language (SAML) V2. 0 — errata composite [S/OL]. [S. l.]: OASIS, 2006: 243–247. (2014-03-15) [2015-07-22]. <https://lists.oasis-open.org/archives/security-services/200404/pdf00002.pdf>.
- [14] TELNONI P, MUNIR R, ROSMANSYAH Y. SAML single sign-on protocol development using combination of speech and speaker recognition [C]// Proceedings of the 2014 International Conference of Advanced Informatics: Concept, Theory and Application. Piscataway, NJ: IEEE, 2014: 299–304.
- [15] 张进铎, 毛承国, 李硕, 等. OpenStack 开源云平台主模块的架

(上接第 294 页)

其次,本文使用了 Prosper.com 提供的公开数据进行实验,实验结果是否具有普遍的意义仍需验证。

最后,后面两个降低相关性影响的方法类似于集成学习领域中的剪枝技术,在众多剪枝方法中,本文使用的选择最优和有放回抽样显得比较简单。所以,进一步的工作需要寻找更好的剪枝方法,来进一步地降低相关性,提高集成准确性。

参考文献:

- [1] POLIKAR R. Ensemble learning [M]// Ensemble Machine Learning. Berlin: Springer-Verlag, 2012: 1–34.
- [2] GNEITING T, RAFTERY A E. Weather forecasting with ensemble methods [J]. *Science*, 2005, 310(5746): 248–249.
- [3] KOREN Y, BELL R. Advances in collaborative filtering [M]// Recommender Systems Handbook. Berlin: Springer-Verlag, 2011: 145–186.
- [4] BHATTACHARYYA S, JHA S, THARAKUNNEL K, et al. Data mining for credit card fraud: a comparative study [J]. *Decision Support Systems*, 2011, 50(3): 602–613.
- [5] BROWN G, KUNCHEVA L I. “Good” and “bad” diversity in majority vote ensembles [C]// MCS '10: Proceedings of the 9th International Conference on Multiple Classifier Systems, LNCS 5997. Berlin: Springer-Verlag, 2010: 124–133.
- [6] NAEINI M P, COOPER G F, HAUSKRECHT M. Obtaining well calibrated probabilities using Bayesian binning [C]// Proceedings of the 2015 Twenty-Ninth AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2015: 2901–2907.
- [7] QIN J, GARCIA T P, MA Y, et al. Combining isotonic regression and EM algorithm to predict genetic risk under monotonicity constraint [J]. *Information Technology and Informationization*, 2014(4): 17–21. (ZHANG J D, MAO C G, LI S, et al. Main module architecture analysis of OpenStack cloud platform [J]. *Information Technology and Informationization*, 2014(4): 17–21.)
- [16] 黄凯,毛伟杰,顾骏杰. OpenStack 实战指南[M].北京:机械工业出版社,2014: 23–39. (HUANG K, MAO W J, GU J J. OpenStack practical guide [M]. Beijing: China Machine Press, 2014: 23–39.)
- [17] 汪奕君,张兴元. OpenLDAP 中访问控制机制的分析[J]. 中国科技信息,2005(21):41. (WANG Y J, ZHANG X Y. Analysis of access control mechanism in the OpenLDAP [J]. *Journal of Information Science and Technology of China*, 2005(21): 41.)
- [18] OpenStack. OpenStack installation guide for Ubuntu 14.04 [EB/OL]. [2015-03-29]. <http://docs.openstack.org/kilo/install-guide/install/apt/content/>.

Background

This work is partially supported by the Fundamental Research Funds for the Central Universities (YZDJ1202), the Fundamental Research Funds for the Central Universities (328201537).

CHI Yaping, born in 1969, M. S., professor. Her research interests include network security.

WANG Yan, born in 1988, M. S. candidate. Her research interests include cloud computing access control.

WANG Huili, born in 1991, M. S. candidate. Her research interests include authentication, cloud computing security.

LI Xin, born in 1989, M. S. candidate. Her research interests include cloud computing authentication, trusted computing.

- [J]. *Annals of Applied Statistics*, 2014, 8(2): 1182–1208.
- [8] ZADROZNY B, ELKAN C. Transforming classifier scores into accurate multiclass probability estimates [C]// KDD '02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2002: 694–699.
- [9] BREIMAN L. Bagging predictors [J]. *Machine Learning*, 1996, 24(2): 123–140.
- [10] SCHAPIRE R E. The strength of weak learnability [J]. *Machine Learning*, 1990, 5(2): 197–227.
- [11] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting [J]. *Journal of Computer and System Sciences*, 1997, 55(1): 119–139.
- [12] BREIMAN L. Random forests [J]. *Machine Learning*, 2001, 45(1): 5–32.
- [13] HO T K. The random subspace method for constructing decision forests [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(8): 832–844.
- [14] UEDA N, NAKANO R. Generalization error of ensemble estimators [C]// Proceedings of the 1996 IEEE International Conference on Neural Networks. Piscataway, NJ: IEEE, 1996, 1: 90–95.

Background

This work is partially supported by the National Natural Science Foundation of China (61232005), the CCF-Tencent Open Research Fund.

JIANG Zhengshen, born in 1990, Ph. D. candidate. His research interests include pattern recognition.

LIU Hongzhi, born in 1982, Ph. D., associate professor. His research interests include information fusion, pattern recognition.