计算机系统应用 ISSN 1003-3254, CODEN CSAOBN E-mail: csa@iscas.ac.cn
Computer Systems & Applications,xxxx,xx(x):x−x [doi: 10.15888/j.cnki.csa.00xxxx]http://www.c-s-a.org.cn

基于朴素 Bayes 组合的简易集成分类器①

宋丛威 [1]

[1](浙江工业大学之江学院理学院, 绍兴 312030)
通讯作者: 宋丛威, E-mail: songcwzjut@163.com

摘 要: 朴素 Bayes 分类器是一种简单有效的机器学习工具. 本文用朴素 Bayes 分类器的原理推导出“朴素 Bayes 组合”公式, 并构造相应的分类器. 经过测试, 该分类器有较好的分类性能和实用性, 克服了朴素 Bayes 分类器精确度差的缺点, 并且比其他分类器更加快速而不会显著丧失精确度.

关键词: 朴素 Bayes 分类器; 朴素 Bayes 组合; 集成学习; 条件独立性

引用格式: 宋丛威.基于朴素 Bayes 组合的简易集成分类器,xxxx,xx(x):x–x. http://www.c-s-a.org.cn/1003-3254/xxxx.html

An Simply Constructed Ensemble Classifier Based on Naive Bayes Combination

SONG Cong-Wei[1]

[1](Zhijiang College of Zhejiang University of Technology, Science College, Shaoxing 312030, China)

Abstract: Naive Bayes classifier is a simple and effective machine learning tool. Based on the principle of naive Bayes classifier, this paper deduces the formula of "naive Bayes combination" and constructs the corresponding classifier. Through testing, the classifier has better classification performanceandpracticality, overcomingthe shortcoming of poor accuracy of naive Bayes classifier, and faster than other classifiers without significant loss of accuracy.

Key words: Naive Bayes classifier; Naive Bayes combination; Ensemble learning; Conditional independence

朴素 Bayes 分类器是一种简单易用的分类器, 在文本分类方面表现出色[1-4]. 垃圾邮件过滤是它最为成功的商业应用[5]. 到现在一直有人尝试把它应用于各种领域[6-9].

朴素 Bayes 分类器建立在条件独立假设的基础上,

$$c(x)=\arg\max_c \arg\max_c p(c\mid x), p(c\mid x)\sim\prod_i p(x_i\mid c)p(c) \quad (1)$$

其中一个只和 x 有关的系数被省略了. 而这个假设比较强, 通常无法被满足; 计算出的后验概率和实际值也相差较大. 不过, 朴素 Bayes 分类器却不会因此而受太大影响[10]. 实际上, 朴素 Bayes 分类器是一种可加模型[1], 即有下述分解

$$f(x)=\sum_i f_i(x_i), x=(x_1,x_2,\cdots,x_n). \quad (2)$$

历史上人们提出了不少改进方案了[3,8]. 本文提出的改进方法解决了朴素 Bayes 分类器的两个问题.

1. 通常朴素 Bayes 分类器要么解决连续型的分类问题, 要么解决离散型的分类问题, 总之 $p(x_i\mid c)$ 的分布类型是同一的[11]. 而本文的方法不受此限制.

2. 在很多方面神经网络等机器学习算法和 Bayes 分类器是互补的. 本文方法可以以非常简单的方式将两者结合起来.

符号约定: 若求和范围不会引起歧义, 则累加符号被简单写作 $\sum_i a_i$, 累乘符号写作 $\prod_i a_i$; $p(x_i\mid c)$ 表示 X 的第 i 个分量 X_i 取 x_i 时的条件概率, 严格的写法应是 $p_{X_i}(x_i\mid c)$.

1 朴素Bayes组合分类器

本节主要推导朴素 Bayes 组合公式, 并简述分类器的构造.

1.1 朴素 Bayes 组合公式

设 $x=(x_1,x_2,\cdots,x_m)$, 即输入变量被分解成 m 部分, 在条件独立假设的基础上, 通过简单变形可得

①基金项目:浙江省自然科学基金(编号 LQ19F050004)
收稿时间:2020-06-11;收到修改稿时间:2020-07-xx

$$p(c|x)\sim\prod_i p(x_i|c)\,p(c)\sim\prod_i p(c|x_i)\,p(c)^{1-m}. \quad (3)$$

在算法设计上, 下面的等价公式会比较好用

$$\ln p(c|x)\sim\sum_i \ln p(c|x_i)+(1-m)\ln p(c). \quad (4)$$

作为分量, x_i 不必是 1 维的; $p(c|x_i)$ 都是独立计算的, 互不干扰, 而且也不是每一个都必须用 Bayes 估计. 如果第 i 项是用分类器 f_i 进行估计的, 那么

$$\ln p(c|x)\sim\sum_i f_{i,c}(x_i)+(1-m)\ln p(c), \quad (5)$$

其中 $f_{i,c}$ 表示 f_i 在 c 上的分量, 代表 $\ln p(c|x_i)$ 的估计. 这就是说, 只要用不同部分的数据独立训练多个训练分类器, 然后简单求和就可以得到一个不错的分类器. 这些分类器被称为基分类器（相当于线性代数中的基向量）. 这是一种特殊的可加模型, 也可以看成一种简单的集成机器学习, 即把 $f_i(x_i)$ 看成是 $f_i(P_i x)$, 其中 P_i 是 x 到 x_i 的投影. 我们把公式(5)叫做**朴素 Bayes 组合公式**, 对应的分类器为朴素 Bayes 组合分类器.

本文的方法最初是为了改进朴素 Bayes 分类器而提出的, 允许任意组合不同的朴素 Bayes 分类器, 如当面对包含连续变量和连续变量的机器学习问题时, 可以组合基于 Gauss 分布和基于多项式的朴素 Bayes 分类器. 但 (4) 式确实不是非要用朴素 Bayes 分类器计算 $p(c|x_i)$ 不可, 而且实验也支持用其他分类器能大大提高精确度. 此时, 严格地说它不再是朴素 Bayes 分类器.

1.2 分类器的推广

作为加性模型的特殊形式, (5) 的一种简单推广是增加系数:

$$\ln p(c|x)\sim\sum_i \alpha_i(c) f_{i,c}(x_i)+\beta(c). \quad (6)$$

这些系数可以通过遗传算法获得, 而初始种群可根据 (5) 合理设置. 这个推广将在以后的研究中实现.

本文的分类器还可以对缺失型数据进行分类, 比如, 只知道 $x=(x_1,x_,,\cdots,x_l)$, 则只需计算

$$\ln p(c|x)\sim\sum_{i=1}^{l} f_{i,c}(x_i)+(1-l)\ln p(c), \quad (7)$$

即只用其中 l 个基分类器.

输入变量的每个分量的分布通常是很不相同的, 如果单纯采用单一分布下的朴素 Bayes 分类器, 效果会很差.本文的分类器允许人们根据每个变量的分布情况设计更有效的朴素 Bayes 分类器, 从而提高分类精度

2 算法设计与实现

2.1 算法

算法基于公式 (5). 输入变量 X 会被分解成 m 部分, 第 i 部分作为第 i 个基分类器的输入; 这些分类器的输出则是共同的. 根据分割后的样本, 分类器被独立训练. 具体的流程如下.

算法 1 朴素 Bayes 组合分类算法（准备数据集 X,Y）

(1) 选择一组基分类器, 构造朴素 Bayes 组合分类器;
(2)将输入数据 X 分割为 $(X_1,X_2,\cdots,X_m)$;
(3) 第 i 个基分类器拟合 (X_i,Y);
(4) 利用朴素 Bayes 组合公式(5), 对任意输入 x 计算概率值 $p(c|x)$;
(5) 根据概率值给出预测值.

其中第 (3) 步根据属性的数据类型进行分割, 基本原则是分离离散与连续变量; 第 (4) 步可以并行计算, 获得较快的速度.

注 离散型和连续型的分别通常是相对的. 一般多数观测值的频率都比较小时, 该变量就应被看作连续变量.

2.2 实现

本文采用 Python 实现, 主要依赖 scikit-learn 机器学习库[12]. 本文算法基于和朴素 Bayes 分类器一样的公式, 因此它的实现只需继承 scikit-learn 提供的实现朴素 Bayes 算法的抽象类即可.

表 1 不同算法比较

	算法	测试精确度	耗时(s)	评价
本文算法	只用朴素 Bayes 基分类器	0.6720	0.07523	中等精度、快速
	包含决策树	0.7903	0.2250	较高精度、较慢
	包含决策树与神经网络	0.7929	5.581	较高精度、慢速
其他算法	高斯朴素 Bayes	0.3856	0.02743	低精度、快速
	多项式朴素 Bayes	0.5345	0.01116	低精度、快速
	决策树	0.8241	0.4796	高精度、较慢
	神经网络	0.8317	6.496	高精度、慢速

程序的运行环境为 macOS10.15, Python3.7, scikit-learn0.23.1. 源代码、数据和实验结果已上传至 GitHub(https://github.com/Freakwill/nb-ensemble).

3 实验分析

实验数据来自 CCF 人工智能竞赛平台 https://www.datafountain.cn/competitions/337. 为了使它成为一个分类问题, 根据数值大小已经把输出变量分为三个等级, 即分 3 类. 总共 50000 条数据, 抽出 30%作为测试数据.

根据数据, 输入变量被大致分为三个部分: 0-1 型, 整数型, 实数型. 关键的原则依然是看数据的频数. 选取适合的基分类器. 集成的分类器将和这些基分类器（单独使用）进行比较.

所有模型都会被重复运行 10 遍, 计算两项指标（精确度与耗时）的均值; 每次运行可能有微小的偏差. 每种模型确实可以设置各项参数调整性能, 但并不显著. 除了神经网络设置了 2000 次迭代, 其隐层大小为 8（朴素 Bayes 组合中为 5）, 其他都采用默认值.

实验结果（见表 1）符合预期. 无论耗时还是精确度, 朴素 Bayes 组合分类器都是介于朴素 Bayes 分类器和其他分类器之间. 该算法适用于那些允许牺牲一定精确度来节省时间的分类问题.如果对数据的分布做更深入的观察, 设计更有针对性的基分类器,是可以获得更好的结果的.

4 结论与展望

本文利用本朴素 Bayes 组合公式设计出一种新的分类器, 它是一种非常简便的集成机器学习方法. 实验结果表明, 算法在不要求高精确度的情况下, 可以提高算法性能. 如果需要在精确度和计算时间之间权衡, 那么可以使用本算法.

如果基分类器都是朴素 Bayes 分类器, 那么朴素 Bayes 组合的结果当然也是朴素 Bayes 分类器. 这样就可以轻易组合出能处理混合不同分布类型的数据集, 如本文中的实验数据, 存在至少三种类型的分布. 实验结果表明这种处理是成功的.因为基分类器并不限于朴素 Bayes 分类器, 所以对条件独立性的假设的依赖也减轻了, 从而提高了精确度. 实验还表明基分类器采用决策树、神经网络等分类器能获得更好的结果.

由于, 本分类器以依然保留了不完全的条件独立假设, 因此精确度的提升也是有限的, 不能和某些成熟的算法竞争. 但是, 它的设计灵活简单, 并具有并行性, 和那些“成熟算法”相比, 适当的设置可以大大缩减计算时间, 而不会显著降低精确度. 因此, 这个算法适合那些对计算时间有较高要求的领域.

未来的研究主要沿着两个方向发展: 设计更好的基分类器, 如为那些含 0 较多的数据选择更合理的分布进而设计相应的朴素 Bayes 分类器; 优化朴素 Bayes 组合公式突破现有的性能瓶颈.

参考文献

1 Hastie T, Tibshirani R, Friedman J. Theelements of statistical learning: Data mining, inference, and prediction, second edition. Springer, 2001.

2 李航. 统计学习方法. 北京: 清华大学出版社, 2012.

3 Chiong R, Theng LB. A hybrid naive Bayes approach for information filtering, 2008 3rd IEEE Conference on Industrial Electronics and Applications, Singapore, 2008:1003-1007.

4 Rish I. An empirical study of the naive Bayes classifier, Journal of Universal Computer Science, 2001,1(2):127.

5 Bin N, Wu JW, Hu F. Spam message classification based on the naive Bayes classification algorithm. International journal of computer science(IAENG), 2019, 46(1):46-53.

6 Liu SY, Xiao J, Xu XK. Sign prediction by motif naive Bayes model in social networks. Information Sciences, 2020:316-331.

7 Manuel J SF, Antonio NG, Francisco J RC. A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services. Journal of Business Research, 2019, 101:499-506.

8 Ma H, Yan W, Yang Z, Liu H. Real-time foot-ground contact detection for inertial motion capture based on an adaptive weighted naive Bayes model. IEEE Access. 2019:1-1.

9 Zeng F, Yao L, Wu B, Meng L. Dynamic human contact prediction based on naive Bayes algorithm in mobile social networks. Software Practice and Experience, 2019(2).

10 Zhang H. The optimality of naive Bayes. Seventeenth International Florida Artificial Intelligence Research Society Conference. 2004.

11 Singh G, Kumar B, Gaur L, Tyagi A. Comparison between multinomial and Bernoulli naive Bayes for text classification, 2019 International Conference on Automation, Computational and Technology Management (ICACTM), London, United Kingdom, 2019:593-596.

12 刘长龙. 从机器学习到深度学习——基于 scikit-learn 与 TensorFlow 的高效开发实战. 北京: 电子工业出版社,2019:101-106.