文章编号：2096-1472(2016)-12-30-03

基于朴素贝叶斯分类器的校园信息智能推荐算法

贾志鹏
（辽宁师范大学计算机与信息技术，辽宁 大连 116081）

摘 要：本文结合对原始朴素贝叶斯分类器原理的分析，论述智能助理软件的设计过程中，所需推荐算法与其之间存在的差异性。并针对在校园收集和整合信息的特点和所需推荐方式，对原始朴素贝叶斯文本分类器算法加以修改。将得到的校园信息智能推荐算法实现在智能助理软件中。经测试，算法具有较好的准确性。

关键词：朴素贝叶斯分类器；校园信息提示；智能推荐算法
中图分类号：TP181 **文献标识码**：A

Intelligent Recommendation Algorithm of Campus Message Based on the Naive Bayes Classifier

JIA Zhipeng
(*College of Computer and Information Technology*,*Liaoning Normal University Computer and Information*,*Dalian* 116081,*China*)

Abstract: Based on the analysis of the naive Bayes classifier,this paper discusses the differences between the recommendation algorithm and the naive Bayes classifier in the design process of intelligent assistant software.According to the characteristics of campus information collection and integration,and the required recommendation methods,the original naive Bayes text classifier algorithm has been modified.The campus information intelligent recommendation algorithm is implemented in the intelligent assistant software.The experimental results show that the algorithm has good prediction accuracy.

Keywords:naive Bayes classifier;campus information prompt;intelligent recommendation algorithm

1 引言(Introduction)

随着知识社会的到来及“互联网+”行动计划的制定，互联网上的海量数据逐渐被有效地收集和整合。国内的一些互联网企业在针对用户的个性化服务上进行了探索，如豆瓣网提供了推荐书籍、音乐等服务，百度旅游在假期提供推荐旅游路线，自动匹配低价机票和酒店等服务。这些创新取得了很好的效果，大大提高了企业的竞争力。目前的智能个人助理软件都没有针对特定群体进行优化，而是面向所有用户进行开发。这样的软件涉及的信息过于分散，缺乏解决实际问题的能力。此外，由于朴素贝叶斯方法在预测和分类中被广泛应用，如在预测项目交付率[1]、互联网流量分类[2]、云检测和估计算法[3]等。因此本文提出了针对校园实时信息进行推荐的基于朴素贝叶斯方法的智能推荐算法研究。

2 朴素贝叶斯分类器(Naive Bayes classifier)

2.1 朴素贝叶斯分类器概述

贝叶斯学习方法中的朴素贝叶斯学习器，常被称为朴素贝叶斯分类器。在某些领域其性能可与神经网络和决策树学习能力相当[4]。分类问题一直是机器学习、模式分类和数据挖掘的核心问题[5]。

贝叶斯方法的新实例分类目标是在给定描述实例的属性值$<a_1,a_2,\ldots,a_n>$下，得到最可能的目标值vmap。朴素贝叶斯分类器所使用的方法：

$$v_{NB}=\underset{v_j\in V}{\operatorname{argmax}}P(v_j)\prod_i P(a_i\mid v_j)$$

其中，VNB表示朴素贝叶斯分类器输出的目标值。

2.2 朴素贝叶斯分类器分析

假设给定了如下表1所示的训练样本数据，学习的目标是

基金项目：2015中国大学生创业创新训练项目，项目编号201510165006.

根据给定天气的结果判断是否打网球。

表1 训练样本数据1

Tab.1 Training sample data1

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

样本数据集提供了14个训练样本，使用此表的数据，并以朴素贝叶斯分类器来分类下面的新实例：(Outlook=sunny，Temperature=cool，Humidity=high，Wind=strong)对于新实例预测目PlayTennis的目标值(Yes或No)，由上面的公式可以得到：

$$P_{(\text{PlayTennsi}=\text{yes})}=9/14=0.64$$

$$P_{(\text{PlayTennsi}=\text{no})}=5/14=0.36$$

$$P_{(\text{Wind}=\text{Strong}\mid\text{PlayTennis}=\text{yes})}=3/9=0.33$$

$$P_{(\text{Wind}=\text{Strong}\mid\text{PlayTennis}=\text{no})}=3/5=0.6$$

其他数据同理代入后得到：

$$P_{(\text{yes})}P_{(\text{Sunny}\mid\text{yes})}P_{(\text{Cool}\mid\text{yes})}P_{(\text{high}\mid\text{yes})}P_{(\text{Strong}\mid\text{yes})}=0.0053$$

$$P_{(\text{no})}P_{(\text{Sunny}\mid\text{no})}P_{(\text{Cool}\mid\text{no})}P_{(\text{high}\mid\text{no})}P_{(\text{Strong}\mid\text{no})}=0.0206$$

故应分类到no中。

3 校园信息智能推荐算法(Campus recommendation algorithm)

3.1 算法说明

与上述例子有所区别，在校园信息智能推荐算法中，所面对情况中的新实例的属性值范围不是仅限于数据库中记录，而是所有可能的输入值。在新实例中可能存在记录中没有涉及到的属性值。算法需要根据新实例与数据记录的匹配程度推测新实例的目标值，将其所对应关键字返回，将新实例记录于数据库中，以此来达到对新实例学习的目的。

给定了如表2所示的训练样本数据，学习的目标是以当前的时间节点为条件，根据用户历史查询记录，推测客户当前可能最需要获取的信息，即返回通过算法计算得出的概率值最大的记录所对应的关键字（时间仅以上下午进行分类）。

表2 训练样本数据2

Tab.2 Training sample data2

Id	Day	Time	Keyword
1	Mon	10	明日首节课程
2	Fri	11	学术杂志
3	Tues	8	休闲杂志
4	Mon	13	今日次节课程
5	Mon	15	明日首节课程
6	Fri	8	语言书籍

根据表2训练样本（二），当在case1（Day=Sun，Time=10）时发出请求。以如下方法计算。

$$Rate_i=P(v_j)\prod_i P(a_i\mid v_j)$$

得到的每组数据概率值将都为0，无法用于比较，故对以上公式进行修改，定义如下公式。

$$Rate=\underset{v_j\in V}{\operatorname{argmax}}[P(v_j)\prod_i P(a_i\mid v_j)-k]\,(a_i\neq 0)$$

其中，k为请求发出时的时间条件与数据库匹配数为0的数量（如上述样本数据，数据记录中没有符合Day=Sun条件的记录，则k值为1）。则首条记录的概率值为

$$P_{(\text{id}=1)}=Rate_1=\frac{1}{3}\times\frac{1}{2}-1=-0.833$$

当在case2（Day=Mon，Time=10）时发出请求。

$$P_{(Id=1)}=1\times\frac{1}{2}\times\frac{1}{3}=0.167$$

其他数据同理可得，训练样本计算结果如表3所示。

表3 数据计算结果

Tab.3 Data computing result

Id	Day	Time	Keyword	Rate
1	Mon	9	明日首节课程	0.167
2	Fri	11	学术杂志	−0.833
3	Tues	8	休闲杂志	−0.833
4	Mon	13	今日次节课程	−0.833
5	Mon	15	明日首节课程	0.167
6	Fri	8	语言书籍	−0.833

返回Rate值最大的关键字，作为查询的输入、输出用户可能需要的数据。

3.2 算法的过程

如图1程序框图所示，通过二层循环依次计算每条记录与其他记录匹配程度的概率值并保存概率最高的一组记录。最后，将保存的记录关键字返回。算法包含二层循环，时间复杂度为$O(n^2)$。

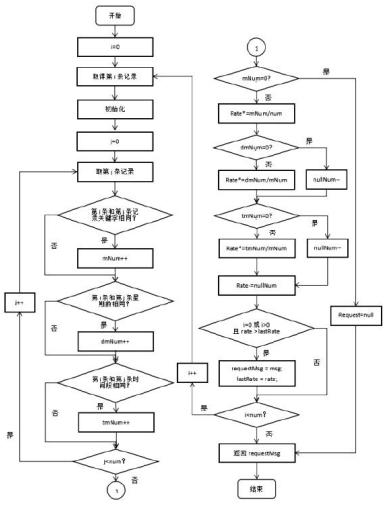

图1 算法程序框图

Fig.1 Algorithm flow diagram

3.3 算法测试

算法测试所采用的数据库为MySQL，通过Android客户端在特定时间进行相应内容的查询，由Java Web端的Servlet将数据记录在服务器中，即生成如图2所示的数据表。

Id	Day	Time	Record
1	星期一	10	明日首节课程
2	星期五	11	学术杂志
3	星期四	8	休闲杂志
4	星期一	13	今日次节课程
5	星期一	15	明日首节课程
6	星期五	8	书籍语言

图2 MySQL测试数据表

Fig.2 MySQL test data table

当数据库完成对基础数据的收集后，用户再通过客户端要求推荐信息时，算法会根据以上信息进行计算，分析数据表记录。将得出关键字返回，查询相应关键字，得到所需信息，结果和表2计算结果一致，如图3所示。

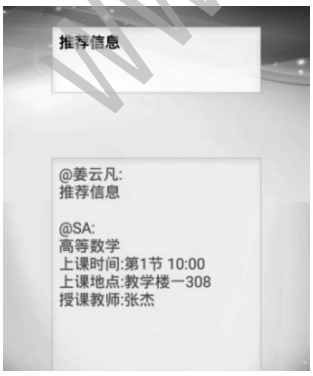

图3 安卓客户端测试结果

Fig.3 Android client test result

4 结论(Conclusion)

本文阐述了原始朴素贝叶斯分类器的基本原理，给出了对其进行变换后的形式化定义。在此基础上，进一步对基于此理论的校园信息智能推荐算法进行描述，给出了算法的过程、时间复杂度，并将算法实际应用于基于JavaWeb服务器的项目中，得到测试结果。测试结果与计算结果一致，达到了预期目的，能够返回所期望数据。

参考文献(References)

[1] Stewart,B.Predicting Project Delivery Rates Using the Naive-Bayes Classifier[J].Journal of Software Maintenance and Evolution Research and Practice,2002,14(3):161-79.

[2] Jun Zhang,Chao Chen.Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions[J]. IEEE Transactions on Information Forensics and Security,2013,8(1):5-15.

[3] Islam,T.CLOUDET:a Cloud Detection and Estimation Algorithm for Passive Microwave Imagers and Sounders Aided by NaiveBayes Classifier and Multilayer Perceptron[J].IEEE Journal of Selected Topics In Applied Earth Observations And Remote Sensing,2015,8(9):296-301.

[4] Tom M.Mitchell.Machine Learning[M].The McGraw-Hill Companies,Inc.1997.

[5] LI Xu-Sheng,GU0 Yao-Huang.Extended Tree Augmented Naive Bayesian Classifier[J].Pattern Recognition and Artificial Intelligence,2006,19(4):470-474.

[6] Zhong Hong-rui,Zhang Yong,Yu Jing-wen.Data Classification On Naive Bayes In Cloud Computing Environment[J]. Computer Applications and Software,2005,32(3):28-30.

[7] Yazdami,M.Artificial intelligence:Principles and Applications[M]. Chapman and Hall,Ltd.,New York,NY,1995(1):1.

[8] Langley P,Iba W,Thompson K.An Analysis of Bayesian Classifiers[J].In:Proc of the 10th National Conference on Artificial Intelligence.San Jose,USA;AAAI Press,1992:223-228.

[9] Thomas G.Dietterieh.Ensemble Methods in Machine Learning [C]. Lecture Notes in Computer Science,2000.

[10] Phinecos.Text Classification Algorithm Based on Naive Bayes Classifier(http://phinecos.cnblogs.com/2008,10,21).

作者简介：

贾志鹏(1994-)，男，本科生.研究领域：机器学习.