# 基于 Naive Bayes 算法的大豆病害诊断研究

时 雷,虎晓红,席 磊　(河南农业大学信息与管理科学学院,河南郑州 450002)

**摘要**　介绍了 Naive Bayes 算法的基本理论。以 UCI 数据库中的大豆数据集为实例,研究了 Naive Bayes 算法在大豆病害诊断中的应用。试验结果表明,Naive Bayes 算法的预测精度优于决策树 C4.5 算法和最近邻 INN 算法。

**关键词**　Naive Bayes;大豆;病害诊断
**中图分类号**　S126　　**文献标识码**　A　　**文章编号**　0517 - 6611(2009)11 - 05320 - 01

**Research on the Diagnosis of Soybean Diseases Based on Naive Bayes Algorithm**

**SHI Lei et al**　(College of Information and Management Science, Henan Agricultural University, Zhengzhou, Henan 450002)

**Abstract**　The basic theory of Naive Bayes algorithm was firstly introduced. Taking the soybean dataset in UCI database as an example, the applications of Naive Bayes algorithm in soybean disease diagnosis were studied. The research results indicated the prediction accuracy of Naive Bayes algorithm was better than that of decision tree C4.5 algorithm and nearest neighbor INN algorithm.

**Key words**　Naive Bayes; Soybean; Disease diagnosis

随着我国农业信息化建设的蓬勃发展,需要进行分类和预测的各类农业数据和信息日益增多。针对这种需求,一种有效的方法就是应用机器学习和模式识别领域中的分类算法对采集的农作物数据进行分类和预测,自动诊断出农作物所患的是何种病害,从而可以节约大量的人力和物力。目前,已经有相关文献研究了决策树、神经网络等算法在农业数据分类和病害诊断中的应用[1-2]。

Naive Bayes 算法是一种简单而有效的分类技术,它也是模式识别领域中广泛应用的分类算法之一[3]。笔者介绍了 Naive Bayes 算法的基本原理,并在 UCI 数据库的大豆数据集上进行了大豆的病害诊断试验。试验结果表明,Naive Bayes 算法对大豆数据集的病害诊断预测能力优于决策树 C4.5 算法和最近邻 INN 算法。

## 1　Naive Bayes 算法基本原理

假设 $d_j$ 为一任意样本,它的特征为 $(a_1,a_2,\cdots,a_m)$,其中 $a_i$ 表示该样本中出现的第 $i$ 个特征项。预定义的样本类别为 $C=\{c_1,\cdots,c_k\}$。假设在给定的条件下,特征项之间都是相互独立的,不存在任何依赖关系。则根据 Naive Bayes 算法,样本 $d_j$ 与已知各类的条件概率 $P(c_i|d_j)$ 定义为:

$$P(c_i|d_j)=\frac{P(c_i)P(d_j|c_i)}{P(d_j)} \tag{1}$$

因为 $P(d_j)$ 对计算结果没有影响,所以可以省略。而得到:

$$P(d_j|c_j)=\prod_{k=1}^{m}P(a_k|c_i) \tag{2}$$

其中,$P(c_i)$ 和 $P(a_k|c_i)$ 可以通过如下的公式来估计:

$$\hat{P}(C=c_i)=\frac{N_i}{N} \tag{3}$$

$$\hat{P}(a_k|c_i)=\frac{1+N_{ki}}{m+\sum_{k=1}^{m}N_{ki}} \tag{4}$$

式中,$N_i$ 表示类 $c_i$ 中的样本数目,$N_{ki}$ 为特征项 $a_k$ 在类 $c_i$ 中出现的频率总数。

对样本 $d_j$ 进行分类,就是按公式(1)计算所有样本类在给定 $d_j$ 情况下的概率,概率值最大的那个类就是 $d_j$ 所在的类,即:

$$d_j\in c_i \text{ if } P(c_i|d_j)=\max_{y=1}^{k}[P(c_y|d_j)] \tag{5}$$

## 2　应用实例

**2.1　数据集**　在 UCI 数据库[4]中,大豆数据集是一个关于大豆疾病分类和诊断的农业数据集。大豆数据集包括了 683 个样本,共有 35 个特征,它将样本分到了 19 个类别中,其中每个类别分别表示了一种大豆所患的疾病。

**2.2　评价指标**　笔者采用精度来衡量分类算法对大豆病害诊断的性能。分类器对样本的预测结果有 4 种情况[5],如表 1 所示。

**表 1　分类器对于一个类别的分类情况**

**Table 1　Cases of the classification for one class by using classifier**

| 真实类别 Actual sorts | 分类器的分类结果 Classification results of classifier | |
|---|---|---|
| | 属于 Belong to | 不属于 Not belong to |
| 属于 Belong to | TP | FN |
| 不属于 Not belong to | FP | TN |

表 1 中,TP 表示被正确地分类为属于此类别的样本数量。TN 表示被正确地分类为不属于此类别的样本数量。FP 表示被错误地分类为属于此类别的样本数量。FN 表示被错误地分类为不属于此类别的样本数量。

根据以上 4 种情况,分类性能可以按照精度来评价,精度的计算公式如下:

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

**2.3　试验结果**　为了比较 Naive Bayes 算法对大豆病害诊断的性能,笔者也在试验中使用决策树 C4.5 算法和最近邻 1NN 算法对大豆数据集进行了病害的预测。试验中,分类性能的评价方法采用的是十折交叉验证法。

在大豆数据集上,Naive Bayes 算法、C4.5 决策树算法和最近邻 INN 算法的分类精度如图 1 所示。从图 1 中可以发现,Naive Bayes 算法的预测精度高于 C4.5 算法和 1NN 算法的预测精度。在试验中,使用 Naive Bayes 算法在大豆数据集上进行病害诊断得到的精度是 92.97%,分别比 C4.5 决策

(下转第 5323 页)

作者简介　时雷(1979 - ),女,河南遂平人,硕士,助教,从事机器学习、模式识别研究。
收稿日期　2009-02-05

成:①数据采集模块。主要是对外部传感器数据的读取,同时引进软件抗干扰措施[4],利用数字滤波器过滤掉不需要的信号,避免外界各种干扰影响系统的准确性。②数据分析模块。主要是对多个指标加权评级给出综合测试结果,分析过程中还采用了平行双样相对偏差计算方法验证仪器测定数据结果的精密度;采用测定加标回收率法进行准确度的校验。通过与其他方法的分析比较,结果表明仪器数据分析结果准确度和精密度都比较高。③外部仪表数据通讯模块。通过串口 RS-232 直接与 WS-1040 仪表通讯,实现串口初始化和水温、水位参数的的读取。外部仪表的使用,节省了硬件电路的搭建,也减轻了软件开发工作。④数据查询模块。主要实现了按照数据采集时间对历史数据进行查询。⑤硬盘驱动模块。主要是小键盘的扫描处理和 LCD 液晶显示。⑥其他模块。系统除了完成对主要任务的处理工作,还必须有对自身维护的功能。如系统设置,主要是外部通道设置、数据循环采集周期设置,系统串口测试、系统复位以及用户操作提示和帮助也是系统重要的组成部分。

**3.4　图形界面开发**　该系统的另一个关键技术就是移植了一个体积、功能理想的 GUI 图形库,并在此基础上开发该软件的图形界面,针对嵌入式 Linux 系统的特点选择了 GTK + 运行在 X-Windows 系统上,然后让 X-Windows 系统运行在嵌入式系统的 frame buffer 上,完成了了 GTK + 的定制与移植,使 GTK + 与 X-Windows 完美结合,开发出人性化的嵌入式 GUI 图形界面。开发界面见图 3。

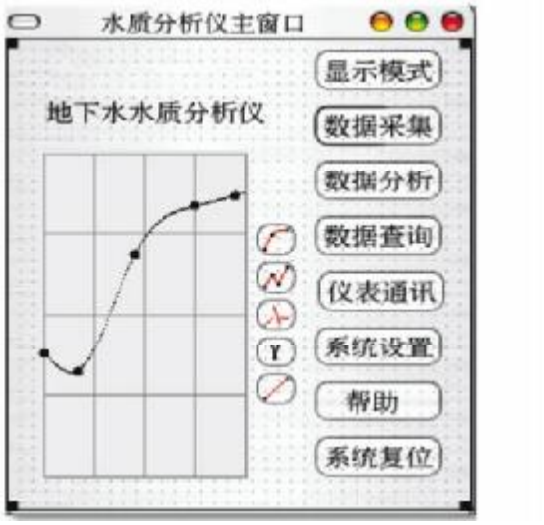


**图 3　系统运行界面**

**Fig. 3　The operation interface of the system**

## 4　结论

根据预先提出的目标与需求分析,结合嵌入式系统软硬件平台资源,该系统研究设计的现场地下水水质分析仪,可以对地下水水质进行定性分析(目前可以实现 5 个参数的采集、8 个参数分析)。笔者把嵌入式 Linux 与 ARM 硬件平台引入水质分析仪研究领域,并尝试把 GTK 库移植到嵌入式平台。分析仪的设计充分发挥 Linux 开源的特性,软件设计模块化;对于硬件资源,引入嵌入式操作系统来管理,利用 ARM 平台功耗低、运算控制功能强大的特点,使得现场分析仪的功能得以实现。

## 参考文献

[1] 徐宝成,王 磊. 基于 ARM 和 Linux 的嵌入式信息终端系统的设计与实现[J]. 微电子学与计算机,2007,24(6):51 – 55.
[2] 陈渝,秦贵和,徐华中,等. ARM9 嵌入式技术及 Linux 高级实践教程[M]. 2 版. 北京:航空航天大学出版社,2005:124 – 129.
[3] 王蕾,陈功新,陆玲,等. 基于 ARM – Linux 的嵌入式系统 GUI 开发研究[J]. 微计算机信息,2007(29):122 – 124.
[4] 陆军. 基于 ARM 的水电站多参数远程动态测量系统[J]. 机械与电气,2007,10(3):53 – 54.

(上接第 5320 页)

树算法高出 1.47 个百分点,比 INN 算法高出 1.76 个百分点。

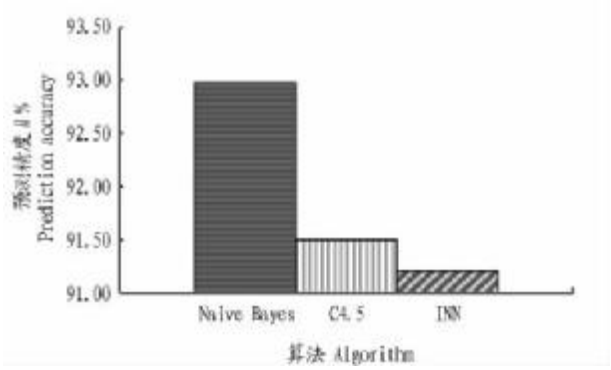


**图 1　不同算法在大豆数据集上的预测精度比较**

**Fig. 1　Comparison of the prediction accuracy on soybean dataset by different algorithms**

## 3　结论

随着我国农业信息化进程的不断推进,对农业数据进行分类和预测的需求会越来越多,而农作物病害的自动诊断和预测就是其中的一个应用热点。笔者研究了 Naive Bayes 算法在大豆病害诊断中的应用,试验表明了 Naive Bayes 算法的诊断效果优于决策树 C4.5 算法和最近邻 INN 算法。

## 参考文献

[1] 方惠敏,张守涛,丁文珂. 基于 BP 神经网络的玉米区试产量预测研究[J]. 安徽农业科学,2007,35(34):10969 – 10970.
[2] 金海月,宋凯. 决策树算法在农业病害诊断中的应用[J]. 当代农机,2007(5):76 – 77.
[3] LEWIS D D. Naive (Bayes) at forty: the independence assumption in information retrieval [C]//The 10th european conference on machine learning. New York: Springer, 1998: 4 – 15.
[4] BLAKE C L, MERZ C J. UCI repository of machine learning databases [EB/OL]. http://www.ics.uci.edu/~mlearn/MLRepository.html. 1998.
[5] YANG Y. An evaluation of statistical app-roaches to text categorization [J]. Journal of Information Retrieval, 1999, 1(1/2): 67 – 88.