文章编号：2096-1472(2019)-02-29-03
DOI:10.19644/j.cnki.issn2096-1472.2019.02.009

基于支持向量机的食品安全抽检数据分析方法

游清顺[1]，王建新[1]，张秀宇[2]，罗 曦[1]
(1.北京林业大学信息学院，北京 100083；
2.贵州省分析测试研究院，贵州 贵阳 550000)

摘 要：本文通过实验对比了逻辑回归、决策树、人工神经网络和支持向量机等方法在分析抽检数据中的表现。结果表明，支持向量机的预测准确率最高。该方法及相应软件系统能够为未来的抽检计划制定提供决策依据，在有限的抽检成本和时间花费的条件限制下，能更多地暴露食品安全中的问题，从而提升食品安全和食品质量。

关键词：食品安全；抽检数据；支持向量机；数据挖掘
中图分类号：TP311.5 **文献标识码**：A

SVM-Based Analysis on Food Safety Sampling and Inspection Data

YOU Qingshun[1],WANG Jianxin[1],ZHANG Xiuyu[2],LUO Xi[1]
(1.*School of Information*,*Beijing Forestry University*,*Beijing* 100083,*China*;
2.*Guizhou Academy of Testing and Analysis*,*Guiyang* 550000,*China*)

Abstract: By means of experiments,this paper compares the performances of logistic regression,decision tree,artificial neural network,and support vector machine in the analysis of sampling data.The results show that support vector machine has the highest prediction accuracy.This method and the corresponding software system can provide decision-making basis for the future sampling plan,expose more problems in food safety,and thus promote food safety and quality.

Keywords:food safety;sampling and inspection data;support vector machine;data mining

1 引言(Introduction)

食品安全是政府、行业机构和人民群众特别关注的问题，只有保证食品安全，才可以从根本上保证人民身体健康和生命安全，这是关系国计民生的重大问题。中国是食品生产和消费大国。为了保障食品安全，每年都要进行规模庞大的食品安全抽检工作，这个过程积累了大量的抽检数据，数据中的规律和知识，急需加以挖掘和利用。在十三五规划中，国家实施食品安全检测和监测能力建设项目，每年投入超过百亿元经费实施这一项目，以食品抽检为手段保证食品安全、提高食品质量。

中国食品安全抽检分为国抽(国家级抽检)和省抽(省级抽检)两个级别，每年可以产生几千万条记录[1]，但这些数据背后的价值却并没有完全被挖掘出来。而利用传统的人工方式对这些数据进行分析，则会由于数据量过于庞大而使得工作难以进行下去。因此智能化的分析处理方法势在必行，而数据挖掘就是其中的典型技术手段[2]。

数据挖掘是指从海量的、信息不完全的大型数据库存放的数据集中自动地发现有用规律或者先前未知的潜在有用的模式[3]。通过数据挖掘获取的规律、模式和知识还可以用来预测未来的观测结果。数据挖掘技术借助聚类、分析、预测、统计等技术手段，在海量数据资源中快速分辨、自动寻找符合人们应用需求的模式，并且提炼出对人有用的信息[4]。

文献[5]利用数据挖掘的手段处理抽检数据。该研究把众多影响因素归一化，并通过多层神经元网络建立抽检数据中原因变量和结果变量之间的关系，效果良好。

本文实验分析了四种数据挖掘算法的七个版本在抽检数据中的应用，比较了它们的预测准确率，然后基于准确率最高的支持向量机方法实现了一套Web软件系统，自动实现抽检数据的分析、挖掘和预测的操作。这样可以大幅提高挖掘效率，又可以减小数据误差率，同时也能节省政府资金投入，并为未来的决策制定提供参考。

本文的第2部分介绍了数据预处理的方法和我们采取的预处理手段，包括预处理的步骤和数据集的选择；第3部分通过分析和实验对比，从多种数据挖掘算法中选择出最适合的算

基金项目：大学生创新计划(X201810022128)；国家重点研发计划(2018YFC1603302).

法；第4部分是本文的结论。

2 数据预处理(Data preprocessing)

由于数据量的规模庞大，且很多原始数据质量不高，存在着各种缺陷，因此通常要对原始的数据进行预处理，使数据更适合挖掘[6]。数据预处理的目的一方面是为了提高数据的质量，减少冗余信息，另外一方面是为了处理一些由于数据输入、数据库界面设计不当导致的数据描述不完整、数据缺失和数据的不一致的情况。数据的预处理可以提高数据的质量、提升后续的挖掘效果[7]。

2.1 数据预处理的步骤

数据预处理的步骤大致可以分为四步，即数据清理、数据集成、数据规约、数据变换。数据清理技术是对于空缺值等异常进行处理、清除重复的数据，以及对异常数据进行错误纠正和清除等操作。现实中造成数据缺失的原因很多[8]，例如数据采集设备故障导致采集缺失，用户填写时不理解或者不耐心未填入导致数据缺失，数据传输过程中错误造成的缺失，数据录入过程中因为疏忽造成数据缺失，以及存储设备损坏导致的缺失等。处理空缺值的方法通常有手动录入、平均值填充、用最可能的值填充、忽略元组、全局常量填充等方法[9]。

数据集成是将来自不同数据源的数据合并为统一一致的数据存储中,这种数据存储可以是数据库或数据仓库。数据集成主要包括：包含相同字段属性的纵向追加和具有相关属性叠加的横向合并。在进行数据横向合并时,会出现同一对象的一些属性字段在不同数据库中的名称不同或属性值不同,这样就容易造成合成后的数据出现不一致性或者数据的冗余性。

数据变换就是将原始数据进行规格化处理，转换成方便后续数据挖掘处理的形式。数据变换常用的方法有：平滑处理、聚集操作、数据概化与规范化和属性构造等。

数据集约是指在保持数据完整性的前提下，将大容量的数据转换成可高效利用的数据集，即在获得相同或相似挖掘结果的前提下，对数据的容量进行有效的缩减的过程。数据归约常用的方法有数据立方体聚集、维规约、数据压缩等。

2.2 训练集与测试集

本文中的模型采用二类分类器，正例为数据集中的“抽检结果合格”，负例为“抽检结果不合格”。以下用抽检结果“合格”与“不合格”两类标注的数据类型同时举例分析，其输入的样本数据如下所示：

[label，<1:x_1>;<2:x_2>;…;<n:x_n>]

其中的label代表抽检结果，正例代表“合格”类别，其label的值为1；负例代表“不合格”类别，其label值为0。其中<n:x_n>表示抽检样本数据中的第n个属性的值为x_n。

在经过一系列的处理如数据清洗、数据集成、数据变换、数据规约之后，有效数据总共有50225条，由于数据中绝大多数的食品抽检数据都是合格的，而不合格样本较少，正例与负例的比例达到30:1。若将全部数据作为输入将会导致分类器的准确率偏高，因此可以将所有不合格样本作为负例，再从合格的样本中选取一部分作为数据集中的正例。最终数据集中正例与负例的比例在3:2左右。

本文采用最常用的十折交叉验证法来测试分类器的准确性，其具体方法是将数据集分成大致均等的十份，轮流将其中的九份作为训练数据，剩余的一份作为测试的数据进行检验。在本文中每一次十折交叉训练测试使用的数据总共有4112条，其中训练集数据总共3704条，测试集数据有411条。

3 算法选择(Algorithm selection)

本文将使用四种常见的机器学习分类算法，即：Logistic回归分析、C4.5决策树、BP神经网络、支持向量机。依据使用的数据集特点，首先对数据集进行数据预处理，并划分为训练集和测试集。将训练集作为输入，并调整四种分类器中影响准确性的参数，并且比较实验中四种分类器的准确性，确定合适的预测模型。

3.1 相关算法介绍

Logistic回归分析是一种广义的线性回归分析模型，Logistic回归的因变量既可以是二分类的，也可以是多分类的。

决策树(Decision Tree)主要是用于分类和预测的技术，它是一种在实例的基础上进行归纳学习的学习型算法，实际上则是一种采用自上而下递归方式的“贪心”算法[10]。它主要是从一组无序、无规则的实例中通过特定的算法来构造决策树，以达到其表现形式的一种分类规则。基于决策树的预测算法的主要思想都是通过对决策树的构建，确定样本数据中的属性标签在分类中是否起作用或起作用的先后顺序。决策树算法有多种版本，最常见的是ID3算法和C4.5算法。但ID3算法有多值倾向性，也就是如果某个变量包含的值越多，则这个变量就越容易被选为分类标准，而C4.5算法克服了这一缺陷，因此我们选择了C4.5算法进行实验。

人工神经网络的研究在一定程度上受到了生物学的启发，因为生物的学习系统是由相互连接的神经元(neuron)组成的异常复杂的网络[11]。而人工神经网络与此大体相似，它是由一系列简单单元相互密集连接构成，其中每一个单元有一定数量的实值输入(可能是其他单元的输出)，并产生单一的实数值输出。

支持向量机(Support Vector Machine)，也可被简称为SVM，它可以在有限样本下进行统计学习，并且可以研究和解决大数据中的分类问题，支持向量机因其优良的特性而作为一种通用的学习机器[12]。因此也是本文研究和应用的主要方法。支持向量机算法也有多种版本，主要由不同的核函数决定。常见的核函数包括线性核函数、多项式核函数、Sigmoid核函数和径向基核函数，根据它们在数据集上的不同表现，本文选择径向基核函数作为支持向量机的函数。

3.2 实验结果比较

对所选四种训练和预测算法，将使用三种评价指标进行评价，分别为准确率(ac)、精确率(pr)和召回率(re)，具体的方法如下：

$$ac = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

$$pr = \frac{TP}{TP+FP} \quad (2)$$

$$re = \frac{TP}{TP+FN} \quad (3)$$

其中，TP(True Position)为将正类预测为正类数、FN(False Negative)为将正类预测为负类数、TN(True Negative)为将负类预测为负类数、FP(False Position)为将负类预测为正类数。在本文中，准确率代表食品抽检质量被正确分类的比例，精确率代表被预测为合格食品中真正为合格食品的比例，召回率代表所有真正为抽检合格食品中被正确分类为抽检合格食品的比例。

对四种算法进行十折交叉验证测试，以下为四种分类算法的测试实验结果：

表1 分类算法实验结果

Tab.1 Experiment results of classification algorithm

分类算法	准确率(%)	精确率(%)	召回率(%)	参数取值
Logistic回归	70.8	66.2	66.1	R=1E-8,M=-1
C4.5决策树	71.6	63.4	69	C=0.5
BP神经网络	72.3	64.5	66.2	L=0.3,M=0.2
支持向量机	83.3	83.6	72.3	C=5.3,g=9.2

由实验结果可以看出，支持向量机(径向基核函数作为核函数)的准确率为83.3%。在这几种算法中，支持向量机的准确率，以及召回率高于其他的算法，因此是最适合的算法。支持向量机算法在食品抽检数据上的预测较为可靠，具有较高准确率，因此其预测的结果可以作为政策制定的一个参考指标，为各个抽检部门提供决策性的参考意见。

3.3 系统实现与测试

在研究和比较各预测算法的基础上，我们采用基于径向基核函数的支持向量机作为训练和预测算法，并基于该算法，实现了一个Web系统。该系统可以对历史数据进行管理，包括增删改查等操作。

首先，管理员导入.csv格式的历史抽检数据文件，并对支持向量机模型进行训练，训练结果存储在Web系统后台。

普通用户登录后输入食品编码或者输入食品属性，系统调用已经训练好的模型进行预测，把该食品可能合格与否的预测结果返回给用户。

用户也可以批量导入食品编码或食品属性，进行批量预测。

与之前的食品抽检计划方案相比，使用该系统后，能对有食品安全问题的重要食品类别和食品批次进行针对性抽检和排查，因而抽检工作的效率会更高，效果会更好。

4 结论(Conclusions)

本文通过比较Logistic回归分析，决策树算法，人工神经网络和支持向量机算法在食品安全抽检数据的分析效果，发现最适合的分类算法是基于径向基核函数的支持向量机算法。通过使用该算法和人工输入的已知属性，可以较为精确地预测未来某种食品可能的食品质量状况，不但可以进行单一预测，也可以进行批量预测。同时，依据该算法生成的整体报告可以对抽检部门提供全面的参考，支持他们有的放矢地检查重点食品类别、属性和批次，从而能节约部分费用，也能够为保障食品安全和提高食品质量提供智能技术支持。

参考文献(References)

[1] Liu Y,Li X,Wang J,et al.Pattern Discovery from Big Data of Food Sampling Inspections Based on Extreme Learning Machine[C].International Conference on Research and Practical Issues of Enterprise Information Systems. Springer,Cham,2017:132-142.

[2] Singh D,Reddy C K.A survey on platforms for big data analytics[J].Journal of Big Data,2015,2(1):8.

[3] Ma Y,Hou Y,Liu Y,et al.Research of food safety risk assessment methods based on big data[C].2016 IEEE International Conference on Big Data Analysis(ICBDA),2016.

[4] Xiangdong Huang,Lin Yang,Runan Song,et al.Effective pattern recognition and find-density-peaks clustering based blind identification for underdetermined speech mixing systems[J]. Multimedia Tools and Applications,2018(77):22115-22129.

[5] Khosa I,Pasero E.Defect detection in food ingredients using Multilayer Perceptron Neural Network[C].2014 World Symposium on Computer Applications & Research (WSCAR),2014.

[6] Li Xu,Chen Wei,Chan Chingyao,et al.Multi-sensor fusion methodology for enhanced land vehicle positioning[J]. Information Fusion,2019(46):51-62.

[7] 文莎,刑立强,张辉.三维空间中目标跟踪测量数据预处理仿真[J].计算机仿真,2018(5):391-396.

[8] 董师倢.数据挖掘中的数据预处理技术[J].信息与电脑(理论版),2016(19):144-145.

[9] 薛毅.食品质量安全抽检信息分析[J].数学建模及其应用,2013(2):13.

[10] 袁明.基于C4.5决策树的网络入侵检测方法[J].科学技术创新,2018(24):81-82.

[11] 蔡强,王君君,李海生,等.基于神经网络的食品安全评价模型构建研究[J].食品科学技术学报,2014,32(01):69-77.

[12] 袁彦彦,王兴芬.基于SVM的生鲜食品货架期预测[J].物流技术,2015(34):64-67.

作者简介：

游清顺(1998-)，男，本科生.研究领域：软件工程.

王建新(1972-)，男，博士，教授.研究领域：软件测试，软件工程，数据挖掘.本文通讯作者.

张秀宇(1981-)，男，硕士，助理研究员.研究领域：软件测试，数据挖掘.

罗　曦(1994-)，女，硕士生.研究领域：软件工程.