doi:10.3969/j.issn.2095-6002.2016.06.015 文章编号:2095-6002(2016)06-0085-06
引用格式:王星云,左敏,肖克晶,等. 基于BP神经网络的食品安全抽检数据挖掘[J]. 食品科学技术学报,2016,34(6):85-90.
WANG Xingyun, ZUO Min, XIAO Kejing, et al. Data mining on food safety sampling inspection data based on BP neural network[J]. Journal of Food Science and Technology, 2016,34(6):85-90.

基于BP神经网络的食品安全抽检数据挖掘

王星云[1], 左 敏[1,*], 肖克晶[1], 刘 婷[2]
(1.北京工商大学 计算机与信息工程学院, 北京 100048; 2.中国食品药品检定研究院, 北京 100050)

摘 要: 数据挖掘技术在食品安全领域拥有巨大的应用价值和潜力。通过分析逆向传播(BP)神经网络算法,说明使用该方法的可行性和优越性。以抽检数据为对象,阐述了数据预处理过程,设计并实现了数据挖掘实验。最后利用挖掘结果进行食品检验结论预测,验证了方法的实用价值和指导意义。实验表明,基于BP神经网络的数据挖掘方法具有良好的过程健壮性和较高的结果准确性。通过预判不合格食品的出现,可以指导实际食品安全抽检工作,从而杜绝食品安全问题的发生。

关键词: 数据挖掘; 食品安全; 抽检数据; 检验结论; BP神经网络
中图分类号: TS201.6; TP393.01 **文献标志码:** A

在国家经济蓬勃发展,生活水平不断提高的今天,社会大众对身体健康的诉求日益增强。在众多的健康影响因素中,食品的质量和安全对身体有着最直接而重要的影响[1]。确保食品的质量安全是整个食品行业和相关部门的责任与义务,其中食品抽检工作作为保证食品质量安全的最后一关,是最为直接和彻底的手段。抽检工作是通过定期制定计划,对各类食品进行抽样检测,测定各种污染物、微生物、营养指标等的含量来判断食品合格与否[2-3],不合格食品将被严格禁止流入市场,从而避免食品安全问题的发生。

通过对食品抽检工作的了解,各级食品安全监管部门每天都会产生大量的检验报告,详细记录诸如生产企业、食品来源、存储状态等属性信息,以及检验项目、检测结果、检验结论等检测信息。这些信息反映了外界各种因素的综合作用结果,隐含了食品多维属性之间的潜在联系[4]。本文拟采用BP神经网络数据挖掘方法,发掘食品自身属性与检验结论之间的内在规律和模式,进而利用已知的属性信息预测食品的检验结论是否合格。这些预测结果可以作为检测工作的先验知识,为重点检测哪些项目、检验结论是否有误等提供决策指导,有效提高抽检工作的效率和效果,达到预防食品安全问题发生的目的。

1 BP神经网络

人工神经网络受到生物学的启发,通过模仿神经元的结构和功能来模拟人脑,解决一些非线性的复杂问题,具有强大的分类和预测能力[5]。人工神经网络的学习机制是利用大量具有多维属性,并且已知分类的数据作为输入,在初始阈值和权值的作用下,计算输出结果并和实际结果比较获得误差。然后通过不断调整阈值和权值,使误差不断减小到可接受范围,此时的输出结果近似等于实际值[6]。将当前的参数作为已训练好的神经网络模型保存下

收稿日期: 2015-12-02
基金项目: "十二五"国家科技支撑计划项目(2015BAK36B04);北京市属高等学校青年拔尖人才培育计划项目(CIT&TCD201404029);北京工商大学创新团队计划项目(19008001074)。
作者简介: 王星云,男,硕士研究生,研究方向为食品安全数据挖掘;
*左 敏,男,教授,博士,主要从事人工智能方面的研究。通信作者。

来,当再次接收一组输入时,网络就能够准确预测其输出结果。

在众多的神经网络算法中,以 BP 神经网络的应用最为广泛和成功[7],它利用隐含层将误差从输出向输入逐层反向传播,在反向传播的过程中利用最速下降法修改权值和阈值,使误差函数快速收敛,具有很快的训练速度和良好的算法性能。

BP 神经网络可以完成任意 n 维到 m 维的映射[8],对于食品检验结论的预测本质上就是寻找这种映射关系,将具有多维属性且不同取值的食品数据准确分类到合格或不合格的类别中,这与 BP 神经网络数据挖掘方法的功能相吻合。同时,BP 神经网络可以处理具有连续或离散属性的数据,对训练数据集中的空值甚至错误值具有良好的健壮性,适合用于分析数据类型复杂、存在较多缺失值的数据集,这正好符合了食品安全抽检数据的特点。另外,BP 神经网络模仿人脑学习的方式使其具有快速、准确、并行解决复杂问题的能力,当训练数据足够多时,BP 神经网络能够将误差降低至很小的程度,使预测结果足够精确,符合提前控制食品安全问题发生的要求[9-10]。综合考虑,将采用 BP 神经网络方法对食品抽样检测数据进行数据挖掘。

2 数据挖掘实验设计

2.1 任务描述

目前,中国食品药品检定研究院承担了各类食品安全相关的监督检验、委托检验、抽查检验等工作。采样人员根据抽检计划实地获取食品样品,记录样品的属性信息,这些信息被提交到检测检验数据管理平台,检测人员根据系统中的食品分类,确定需要检验的项目,并进行检测得出结论。

以上述抽检工作产生的数据为基础,本文数据挖掘的主要任务是使用已采集的样品属性信息作为输入,已知的检验结论作为输出,训练 BP 神经网络模型反映食品属性和食品检验结论之间的相关关系[11],分析模型的性能和结果的准确度,并利用这种潜在关系来预测在特定的属性取值下,检验结论是否合格。

2.2 数据选择

数据本身的结构和特点直接影响到 BP 神经网络的训练效果,需要充分考虑模型结构和挖掘目的,选择合适的特征属性,确定合适的输出结果,并且选取足够的数据项作为训练样本[12]。

2.2.1 确定结果分类

BP 神经网络的输出节点表示分类结果,假设一项食品抽检数据只包含 2 个检测项(A,B),结论为合格或不合格 2 种,那么分类结果是检测项与结论取值的组合,形如(A 合格,B 合格),(A 合格,B 不合格),(A 不合格,B 合格),(A 不合格,B 不合格)共 2^2 个。实际上一种食品的检验项目平均有 20 种左右,分类结果将达到 2^{20} 个之多,显然这种挖掘结果可读性差,并且训练过程需要耗费大量时间[13],不具有实用价值。一个较好的解决方案是对不同检验项目单独训练 BP 神经网络,使结果分类只包含一种检验项目的两种检验结论共 2 个分类,这样训练时间将大幅缩短并且输出结果清晰明了。对于需要检测多个项目的食品样品,只需使用多个已训练的对应检验项目的 BP 神经网络进行预测,分别得到检验结论即可。

2.2.2 选取特征属性

一条食品抽检数据包含近 70 个属性,事实上并不是所有属性都适合作为训练输入,其中如“抽检编号”、“联系人”等属性不包含分类结果的有用信息,需要去除。另外如“样品规格”、“样品批号”等名词性属性虽然可能含有分类相关的有用信息,但是在填报时没有取值范围限定,对于没有范围限定的名词性属性由于可能存在取值过多的情况,会影响训练效率,并且预测结果会因为新值的出现产生较大误差,所以也需要去除。

综合分析后,选取检测检验数据管理平台数据库中包含污染物“铅”的 10 000 条抽检数据作为训练数据集,选定“生产企业省份”、“抽样地点”、“年销售额”等 13 个属性作为特征属性,检验结论“合格项”、“不合格项”作为输出结果分类。

2.3 数据预处理

数据预处理工作是为了保证输入数据的质量,它是整个数据挖掘过程中唯一需要人工处理的环节,却是极其重要、必不可少的环节[14-15]。在上述选定的 10 000 条食品抽检数据中,存在大量缺失值、不规范项,甚至错误项,所以在开始挖掘之前还需要对数据进行清洗和转换,以保证 BP 神经网络模型能够正确处理这些样本。结合抽检数据的特点,对训练数据集进行如下处理[16-17]。

1)对所有的空值、NULL 值、“未提供”等全部置

为空，BP 神经网络可以处理此类缺失值，不需要做默认值替换或者删除处理[18]。

2)“生产日期”属性的值包含了年月日的完整信息，由于日期的取值在理论上没有范围限定，可以将日期概化为季度，这样既限制了名词性属性的取值范围，同时保留了日期本身的信息。如 2015/3/1 转换为“一季度”，2014/12/20 转换为“四季度”。

3)“年销售额”属性的值包括“万元”、“元”和没有单位的数值，需要统一单位后去除文本信息，将名词性属性转换为数值型属性。如“12 万元”改为数字 12，“2 000 元”改为数字 0.2。

4)“单价”属性的值包括大量不同单位，如“元/个”、“元/500 g”等，同样需要统一单位并去除文本信息，将名词性属性转换为数值型属性。如“3.5 元/盒”改为数字 3.5。

3　实验过程与结果分析

使用怀卡托智能分析环境(Waikato environment for knowledge analysis，WEKA)。WEKA 是一款开源的数据挖掘和机器学习工具，由 Java 语言编写，实现了分类、聚类、回归、关联规则等多种数据挖掘方法[19]，由于其开放源代码，使用者可以方便地调用和改进已有算法。

将经过预处理的训练数据集共 10 000 条导入 WEKA，其结构如图 1。选择分类方法下的 Multilayer Perceptron 算法，进行 BP 神经网络训练，训练过程分为构建网络模型和模型交叉验证 2 步。

No.	生产企业省份 Nominal	抽样地点 Nominal	区域类型 Nominal	年销售额 Numeric	食品大类 Nominal	样品来源 Nominal	生产日期 Nominal	单价 Numeric	样品类型 Nominal	样品属性 Nominal	样品形态 Nominal	包装分类 Nominal	样品存储状态 Nominal	检验结论 Nominal
1	云南	批发市场	城市		调味品	外购	四季度	0.7	工业加...	普通食品	固体	预包装	常温	合格项
2	北京	超市	城市	18000.0	调味品	外购	一季度	4.9	工业加...	普通食品	固体	预包装	避光	不合格项
3	湖南	批发市场	城市	500.0	茶叶及...	外购	一季度	128.0	工业加...	普通食品	固体	预包装	常温	不合格项
4	广东	超市	城市	430.0	茶叶及...	外购	一季度	98.0	工业加...	普通食品	固体	预包装	常温	不合格项
5	湖南	成品库...	乡村	10.0	蔬菜及...	加工/自制	二季度	2.0	工业加...	普通食品	固体	预包装	避光	合格项
...														
9996	河南	超市	城市		调味品	外购	一季度	1.2	工业加...	普通食品	液体	预包装	常温	合格项
9997	河南	其他	城市		饮料	外购	二季度	45.0	工业加...	普通食品	液体	预包装	常温	合格项
9998	重庆	超市	城市	30.0	调味品	外购	一季度	3.8	工业加...	普通食品	液体	预包装	常温	合格项
9999	吉林	半成品库	城市	12000.0	食用油...	加工/自制	一季度	10.0	工业加...	普通食品	液体	散装	避光	合格项
1...	辽宁	超市	城市		调味品	外购	一季度	1.0	工业加...	普通食品	液体	预包装	常温	合格项

图 1　训练数据集

Fig. 1　Structure of training set

在数据量较大且具有较多特征属性、BP 神经网络模型比较复杂的情况下，模型的构建过程耗时 771 秒，训练效果比较理想。训练好的模型确定了各个节点的内部阈值，如表 1，以及节点之间的权值，如表 2。表示在这些确定的取值下，模型的输出结果最为准确。以不合格项输出节点(Node 0)和合格项输出节点(Node 1)为例，其内部阈值分别为 2.54 和 −2.51，隐含层第一个节点 Node 2 与 Node 0 和 Node 1 之间的权值分别为 −0.28 和 0.28。

表 1　各节点内部阈值

Tab. 1　Internal threshold of nodes

节点	阈值
Node 0	2.54
Node 1	−2.51
Node 2	−0.01
Node 3	0.12
Node 4	−0.02
⋮	⋮

表 2　输出节点与隐层节点间权值

Tab. 2　Weight between output nodes and hidden layer nodes

输出节点	隐层节点								
	Node 2	Node 3	Node 4	Node 5	Node 6	Node 7	Node 8	Node 9	…
Node 0	−0.28	2.78	−1.46	−3.37	−1.30	−0.27	−0.38	−0.17	…
Node 1	0.28	−2.80	1.45	3.37	1.29	0.28	0.31	0.16	…

模型建立后采取十折交叉运算进行准确性验证，它将数据集分成 10 份，轮流将其中 9 份作为训

练数据,一份作为测试数据进行试验[20]。在总共10 000 条抽检数据中,被正确分类的数据项有9 943 条,其余有 57 条被错误分类,正确率达到99.4%,模型的一致性检验达到 96.5%,其他误差指标都在较低水平,表明模型在各方面具有较高的准确率(见表 3)。具体分类结果见表 4,总计900 条不合格数据中,有 867 条被正确分类,33 条被错误分类为合格数据,总计 9 100 条合格数据中有 9 076 条被正确分类,24 条被错误分类为不合格数据。

表 3 模型评价指标

Tab. 3 Evaluation indexes of model

指标	数目	比率/%
正确分类实例	9943	99.43
错误分类实例	57	0.57
一致性检验	–	96.5
平均绝对误差	–	0.71
均方根误差	–	7.44
相对绝对误差	–	4.34
相对平方根误差	–	26

表 4 训练结果混淆矩阵

Tab. 4 Confusion matrix of training results

预测分类	实际分类	
	合格项	不合格项
合格项	9 076	33
不合格项	24	867

所有数据项的分类情况见图 2,其中×符号代表分类正确点,方框符号代表分类错误点。由图 2 可见,由于作为训练样本输入的合格项数据基数较大,在不合格项中错误点出现的比例(33/900)远高于合格项中错误点出现的比例(24/9100)。实际

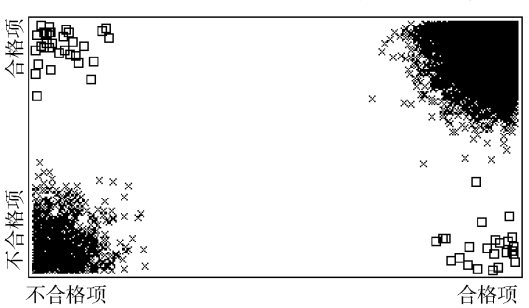

图 2 分类错误图示

Fig. 2 Graph of classifier errors

上,应当尽量避免将实际不合格的项判断为合格项,需要把此类错误限定在较小的范围内,因此实验还存在改进空间,可以考虑从增加不合格项训练样本、提高数据质量等方面进行。

4 预测实验

从检测检验数据管理平台数据库中随机选取“铅”检验项合格与不合格的数据各 100 条,与训练样本做同样的预处理。将数据导入上述训练好的 BP 神经网络模型中进行分类预测(见表 5),给出了预测结果的混淆矩阵。结果显示该模型正确预测了 99 条不合格项的检测数据,1 条被错误预测,其余 100 条合格项检测数据全部被正确预测,正确率达到 99.5%。

表 5 预测结果混淆矩阵

Tab. 5 Confusion matrix of prediction results

预测分类	实际分类	
	合格项	不合格项
合格项	100	1
不合格项	0	99

以图 3 中两条预测结果为例说明该方法对抽检工作的指导意义。某次抽样工作抽取了河南省某调味品样品,获得该样品对应的属性信息为(河南,超市,乡村,10,调味品,外购,二季度,2,工业加工食品,普通食品,固体,预包装,常温),利用已训练好的 BP 神经网络预测出该食品检验项目“铅”的检验结论为“不合格项”,检测人员参考预测结果需对该样品的“铅”检验项进行重点检测。抽取的另一样

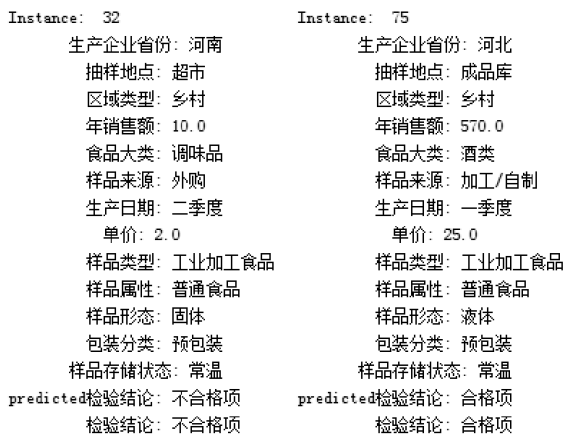

图 3 预测实例

Fig. 3 Instances of prediction

品是河北省某酒类食品，其属性信息为（河北，成品库，乡村，570，酒类，加工/自制，一季度，25，工业加工食品，普通食品，液体，预包装，常温），通过预测给出该食品检验项目“铅”的检验结论为“合格项”，在考虑准确率为 99.5% 的情况下，检测人员可以选择免检该项目。

通过这种方式，监管部门检测人员可以使用待检测食品的属性信息，提前预测出各种检验项目的检验结论，对于预测结论为“不合格项”的食品，需要进行重点检测，避免漏检、错检情况的发生。对于预测结论为“合格项”的食品，可以选择免检，节约大量的时间、设备和人力资源。

5　结论及展望

研究表明，基于 BP 神经网络的数据挖掘方法适用于对食品抽检数据进行分析，在抽检数据本身具有多维属性，且含有较多空值和错误值的情况下，BP 神经网络仍然可以做出准确的判断，其准确率高达 99.4%。预测实验表明，在已经获得了食品基本信息的情况下，该方法可在实际检测工作进行之前，准确预测出对应检验项目的检验结论，对检测工作具有重要的指导意义，不仅可以提高检测工作的效率和效果，还可以节约资源和成本。通过后续的改进、优化和系统性整合，该方法可以实现对不合格食品的准确预测、实时监控和风险预警等功能，为预防食品安全问题的发生提供理论依据和应用参考。

参考文献：

[1] 孙宝国，王静，孙金沅. 中国食品安全问题与思考[J]. 中国食品学报，2013（5）：1－5.
SUN Baoguo，WANG Jing，SUN Jinyuan. Perspectives on China food safety problems[J]. Journal of Food Science and Technology，2013（5）：1－5.

[2] 李月娟，吴霞明，王君，等. 食品分析及安全检测关键技术研究[J]. 中国酿造，2012，31(12)：13－17.
LI Yuejuan，WU Xiaming，WANG Jun，et al. The research of key technology on food analysis and security detection[J]. China Brewing，2012，31(12)：13－17.

[3] 金剑，齐思源，彭亚拉. 神经网络法在食品安全预警中的应用[J]. 食品科技，2011，36(11)：306－309.
JIN Jian，QI Siyuan，PENG Yala. Application of artificial neural networks in food safety pre-warning system[J]. Food Science and Technology，2011，36(11)：306－309.

[4] 田春园. 基于数据挖掘的食品安全风险评价与预警系统[D]. 青岛：青岛理工大学，2012.

[5] 李冰，郭祀远，李琳，等. 人工神经网络在食品工业中的应用[J]. 食品科学，2003，24(6)：161－164.
LI Bing，GUO Siyuan，LI Lin，et al. Application of artificial neural network on food industry[J]. Food Science，2003，24(6)：161－164.

[6] 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(1)：69－76.
CAI Qiang，WANG Junjun，LI Haisheng，et al. Research on establishment of food safety evaluation model based on neural network[J]. Journal of Food Science and Technology，2014，32(1)：69－76.

[7] NAJAH A，EL-SHAFIE A，KARIM O A，et al. Application of artificial neural networks for water quality prediction[J]. Neural Computing and Applications，2013，22(S1)：187－201.

[8] DING Shifen，LI Hui，SU Chunyang，et al. Evolutionary artificial neural networks：a review[J]. Artificial Intelligence Review，2013，39(3)：251－260.

[9] KARAHOCA D，KARAHOCA A，ÖZERK YAVUZ. An early warning system approach for the identification of currency crises with data mining techniques[J]. Neural Computing and Applications，2013，23(7/8)：2471－2479.

[10] LIU Zhen，MENG Lingyong，ZHAO Wei，et al. Application of ANN in food safety early warning[C]// Future Computer and Communication（ICFCC），2010 2nd International Conference. Wuhan：IEEE，2010：677－680.

[11] 章德宾，徐家鹏，许建军，等. 基于监测数据和 BP 神经网络的食品安全预警模型[J]. 农业工程学报，2010，26(1)：221－226.
ZHANG Debin，XU Jiapeng，XU Jianjun，et al. Model for food safety warning based on inspection data and BP neural network[J]. Transactions of the CSAE，2010，26(1)：221－226.

[12] BHARATI M，RAMAGERI M. Data mining techniques and applications[J]. Indian Journal of Computer Science and Engineering，2010，1(2)：25－47.

[13] YANG Jing，ZENG Xiaoqin，ZHONG Shuiming，et al. Effective neural network ensemble approach for improving generalization performance[J]. Neural Networks and Learning Systems，2013，24(6)：878－887.

[14] FAMILI F，SHEN W M，WEBER R，et al. Data pre-

processing and intelligent data analysis[J]. International Journal on Intelligent Data Analysis, 1997, 1(1): 3 – 23.

[15] 刘明吉, 王秀峰, 黄亚楼. 数据挖掘中的数据预处理[J]. 计算机科学, 2000, 27(4): 54 – 57.
LIU Mingji, WANG Xiufeng, HUANG Yalou. Data preprocessing in data mining[J]. Computer Science, 2000, 27(4): 54 – 57.

[16] 晁凤英, 杜树新. 基于关联规则的食品安全数据挖掘方法[J]. 食品与发酵工业, 2007, 33(4): 107 – 109.
CHAO Fengying, DU Shuxin. Data mining technics for food safety based on association rules[J]. Food and Fermentation Industries, 2007, 33(4): 107 – 109.

[17] 徐燕伟. 增量关联规则算法及其在食品安全监管中的应用[D]. 杭州: 浙江大学, 2008.

[18] ENNETT C M, FRIZE M, WALKER C R. Influence of missing values on artificial neural network performance[J]. Studies in Health Technology and Informatics, 2001(1): 449 – 453.

[19] HALL M, FRANK E, HOLMES G, et al. The WEKA data mining software: an update[J]. ACM SIGKDD Explorations Newsletter, 2009, 11(1): 10 – 18.

[20] EFRON B, TIBSHIRANI R. Cross-validation and the bootstrap: estimating the error rate of a prediction rule[M]. Stanford California: Division of Biostatistics, 1995.

Data Mining on Food Safety Sampling Inspection Data Based on BP Neural Network

WANG Xingyun[1], ZUO Min[1,*], XIAO Kejing[1], LIU Ting[2]

(1. *School of Computer and Information Engineering, Beijing Technology and Business University, Beijing* 100048, *China*;
2. *National Institute for Food and Drug Control, Beijing* 100050, *China*)

Abstract: Data mining technology has great application values and potential in the food safety field. The feasibility and advantage of the BP neural network algorithm were explained. The process of data preprocessing was introduced, and the experiment of data mining was designed then realized, focusing on sampling inspection data. Finally, by taking advantage of the mining results, a prediction of food inspection conclusions was put forward which verified the method's practical value and guiding significance. The experiment indicated that data mining method based on BP neural network has favorable robustness and good accuracy. The predictions of unqualified food's appearance can lead food safety sampling and inspection work in practice, which can put an end to the occurrence of food safety problems.

Key words: data mining; food safety; sampling inspection data; inspection conclusion; BP neural network

(责任编辑:檀彩莲)