大数据挖掘在食品安全风险预警领域的应用

王雅洁,杨 冰,罗 艳,何锦林,谭 红*　(贵州省分析测试研究院,贵州贵阳 550002)

摘要　食品安全综合评价与预警是食品安全的重难点。该研究着重介绍了大数据挖掘在食品安全风险预警领域的应用。首先对大数据的基本概念及3种典型的大数据挖掘技术(贝叶斯网络、决策树以及人工神经网络)概念进行分析,并探讨这3种大数据挖掘方式在食品安全行业的应用现状。之后比较3种大数据挖掘方式,提出将其中一种大数据挖掘方式BP神经网络运用于食品安全风险预警的构想。

关键词　食品安全;大数据挖掘;风险预警;贝叶斯网络;决策树;BP神经网络
中图分类号　S126;TP39　　**文献标识码**　A　　**文章编号**　0517 - 6611(2015)08 - 332 - 03

The Application of Big Data Mining in Food Safety Alert Field

WANG Ya-jie, YANG Bing, LUO Yan, TAN Hong* et al　(Guizhou Academy of Testing and Analysis, Guiyang, Guizhou 550002)

Abstract　Comprehensive evaluation and alert in food safety is very important and difficult. This essay mainly focuses on introducing the application of big data mining in food safety alert field. At first, the concept of big data miming and three big data methods(Bayesian network, Decision tree, BP neural network) were analyzed. At the same time, the application status of these three big data miming methods in food safety areas were discussed. Then these big data miming methods were compared, and how to apply BP neural network into food safety risk alert was proposed.

Key words　Food safety; Big data mining; Risk alert; Bayesian network; Decision tree; Back propagation neural network

近年来,信息产业界的新兴服务飞速发展,各行业的数据种类和规模呈现指数级增长,我国大数据的时代已正式开启。大数据概念的兴起为人们提供了一种新的看待世界的方法,为了让存于数据仓库中的大量数据变得有价值,对大数据的挖掘成为普遍关注的话题。与此同时,食品安全事件在我国频频发生,如"瘦肉精"中毒事件、"苏丹红"事件、劣质奶粉事件、"三聚氰胺"事件以及有毒大米事件等,严重威胁群众的身体健康,引起极大的负面社会效应。因此食品安全综合评价与预警越来越成为食品安全的重点。寻找有效的预警方式能够极大程度地提高食品安全水平,大数据挖掘技术正是这样一种有效的预警方式。笔者对大数据的基本概念进行剖析,并分析大数据挖掘中3种典型的挖掘方式及其在食品安全领域的应用,对比分析3种大数据挖掘方式应用于食品安全风险预警领域的适应度,在此基础上,选取其中较优的大数据挖掘方式,探讨将其应用于食品风险预警领域的初步设想。

1　大数据概述

大数据是一个比较抽象的概念,仅从字面上来看,表示数据规模庞大,数据多元化等。对于大数据的定义,目前没有一个比较统一的、公认的定义。"大数据"这个术语最初始的引用可追溯到apache org的开源项目Nutch。那时候,大数据曾被定义为"更新网络搜索所需要进行批量处理或分析的大量数据"[1]。维基百科认为大数据是任何大量复杂的,难以用传统的数据处理方式处理的数据集[2]。Grobelink则定义大数据应具有以下3个特点(3V):Volume(规模性)、Veocity(高速型)和Variety(多样性)[3],这也是目前比较广泛且具有代表性的定义。此外,在"3V"的基础上,某些大型企业提出了"4V"定义,即在已有3V的基础上再添加一个新的特性。目前,关于第四个"V"的定义还未统一,IDC认为大数据还应当具有价值性(Value)[4],而IBM认为大数据必然具有真实性(Veracity)[5]。

2　3种大数据挖掘方式原理概况及在食品安全行业的应用现状

几年来,随着计算机硬件稳定的发展,大量功能强大数据收集设备和存储介质被广泛供应在市场上,与此同时大力促进了数据库的发展,使得大量信息和数据存储与数据库中[6]。在大数据库中大量的数据成了"数据坟墓",如何让这些海量的数据"苏醒"过来,将数据坟墓转变成为有价值的知识"金块",人们需要寻求有效的解决方式。大数据挖掘技术正是这样一种解决数据和知识之间的鸿沟,将数据转变成知识的有效方式。大数据挖掘是将潜在隐含的信息从数据中提取,通过开发计算机程序在数据库中进行自动挖掘,以发现规律或模式的一种有效手段[6]。大数据挖掘,即"从大数据中挖掘知识"[7]。如果能从对海量数据的挖掘中发现明显的模式,这些模式可被人们总结、理解和设计,并可用来对未来大规模的数据做出准确的预测。大数据挖掘方式基于传统的数据挖掘,而数据挖掘技术由众多学科领域技术的集成,比较常见的包括机器学习、统计学、模式识别、高性能计算等。常见的机器学习数据挖掘技术有贝叶斯网络(Bayesian Network)、决策树(Decision Tree)、人工神经网络(Artificial Neuron Network)等。

2.1　贝叶斯网络　贝叶斯网络是由Pearl在1988年提出的。贝叶斯网络是一种不确定的表示模式,实质上是一个赋值的复杂因果关系网络,表现为一种有向无环图(directed acyclic graph,DGA)[8]。每个网络中的结点代表一个变量,即为一个事件。变量之间的弧表示事件发生的直接因果关系。弧的规则使得贝叶斯网络能够很好地表示那些不确定的内

基金项目　贵州省软科学研究项目(黔科合R字[2014]2023号);贵州科学院青年基金重点项目(黔科院J合字[2014]02号)。
作者简介　王雅洁(1990 -),女,贵州贵阳人,助理工程师,硕士,从事大数据挖掘与分析研究。*通讯作者,研究员,教授,从事计算机应用与食品安全研究。
收稿日期　2015-01-29

在概率。贝叶斯网络反映整个数据域中数据间的概率关系，可被用来发现令人信服的概率依赖关系。贝叶斯网络是一个十分简洁，易于理解的模型。基于理解行为、结果及它们之间因果关系的条件下，合理的解释可能出现的结果，从而进行预测和决策[10]。贝叶斯网络能有效处理不完整数据，能和其他技术相结合进行因果分析。同时贝叶斯网络能够使先验知识和数据有机结合，且有效地避免数据的过度拟合。

贝叶斯网络在食品行业中的运用，比较有代表性的是用于食品产品设计[11]。例如，在食品贝叶斯网络建模中，如果知道人们普遍喜欢甜的食品，在样本中也存在既甜又受欢迎的食品，那么贝叶斯网络推理出这个食品的颜色将会影响其受欢迎程度。而传统基于规则的专家推荐系统由于系统是模块划的，其中的一些规则与其他规则或数据源的内容无关，则不能处理类似此类情况的问题，而贝叶斯网络中的条件概率则解决了这一问题。此外，贝叶斯网络模型是风险评价概率统计模型的代表，曾被应用于食品供应链的风险概率估计[12]。通过裁剪食品供应链中物流、信息流和资金流等风险因素，分析初始风险事件，建立贝叶斯网络模型进行风险评价。由于食品供应链对于不同的初始事件响应不同，事件发展过程及结果也是不同的。通过获取贝叶斯网络中每个节点关系的条件概率值，计算联合概率，即可得到食品的风险值。

2.2　决策树　决策树是机器学习中应用相对广泛的归纳推理算法之一，通过逼近离散值函数的方法，以优先选择较小的“树”为原则，将学习到的函数表示为一棵决策树。决策树能够很好地学习噪声数据，从中学习规律，析取表达式[13]。在决策树中，每个节点都代表一个特定的实例，这些实例被决策树从根节点依次排列到叶子结点上。决策树通过判定来分类实例，实例所属的分类最终被表现在叶子节点上。实例的分类方式是从决策树的根节点开始。依次选择某个实例的属性值，然后根据该属性对应的树枝继续向下至另一个节点(实例)。接着以新实例为决策树的根循环以上步骤，最终可得到实例的分类。通过从根到叶子节点的路径选择来生成规则集合，该集合可以高度地概括和归纳样本数据规则，并且精准地判别样本的个体属性，同时也可以应用于预测或判别新的样本属性。

决策树分析法通过树状的逻辑思维方式解决复杂决策问题，是以风险分析为依据的决策方法。决策树在食品行业的运用有基于农产品的食品安全评估研究[14]，其针对影响农产品质量安全的数据特点，结合降维方式进行数据预处理，找出影响质量安全的主要特征值，并构建基于组合优化决策树的农产品质量安全判别模型，选取如地下水重金属含量、土壤 pH、种植规模性等不同的农产品影响因素作为决策树的属性。将数据样本分成训练集和测试集，通过训练，得到规则集合。将测试集中的数据样本输入决策树模型，计算准确率，从而得到决策树方法是否能对农产品质量安全风险进行评估的结论。决策树还被运用于具体检测指标来评价油炸性方便面的品质等[15]。

2.3　人工神经网络　人工神经网络来源于生物学，通过模拟生物学中相互连接神经元组成的复杂网络进行建模，是一种学习精度较高的数据挖掘方式。由于神经网络能够很好地学习数据中的错误，通过训练精准的发现数据中的隐含规律，目前已被成功应用到很多领域。目前，人工神经网络模型有近 10 种，常见人工神经网络为反向传播(BP)神经网络[13]，神经元被分布在不同的层级之中，每一个层级含有一个或多个神经元。每一个神经元里有一定量的输入值(可能为上一层神经元的输出)及输出值(可能将会作为下一层神经元的输入)。每一层级中的每一个神经元，都会跟上一层级及下一层级中的每个神经元进行交互，通过正向传播、权值调整和反向传播，极大程度地学习所给的数据集，从而训练好模型。神经网络拥有健壮性很强的学习能力，其为向量值、离散值或实数值的逼近提供了一种很好的方式。

BP 神经网络具有高度非线性函数映射功能，且其拥有分布式的信息存储能力及大规模的并行处理能力，其良好的自适应性、较强的抗干扰能力使得其拥有较强的学习能力。BP 神经网络是人工智能中对不确定性问题处理具有高度解决能力的方法，其曾与主成分分析结合被用于近红外光谱苹果品种鉴别方法研究[16]，该研究首先使用主成分分析对苹果进行聚类并获取苹果的进红外指纹图谱，即对于苹果品种敏感的特征波段，用特征波段图谱作为神经网络的输入，品种作为输出，建立模型，进行训练，之后对未知的样品进行预测。这样的品种识别准确率达到了 100%。此外，BP 神经网络还被用于冬小麦耗水预测[17]、大米直链淀粉含量预测等[18]。

3　3 种大数据挖掘方式应用于食品安全风险预警领域的适应度对比研究

贝叶斯网络、决策树、BP 神经网络都是数据挖掘中最有效的分类方式。通过建模训练，模型从中学习分类规则，当存在新的未知种类数据时，根据学习经验，模型具有辨识能力，人们称这样的能力为预测。其中，贝叶斯的实现方式是通过依次计算出数据属于某一类的概率值，其中概率最大的类即为对象的所属分类。在贝叶斯分类中，所有的属性都会参与计算及分类。决策树是一棵二叉树或多叉树，针对离散型变量，通过判定的方式，自上而下递归构造，树的各个叶节点都代表一个分类。而 BP 神经网络是基于感知器的分类器，通过训练模式的迭代和学习算法，产生线性或非线性的可分别判别函数。只需给定神经网络大量的输入和目标输出对，BP 神经网络通过正向传播、权值调整及反向传播，进行训练。神经网络把所学到的知识规律记忆在网络的权值中，从而找出数据隐含规则。BP 人工神经网络的权值不是通过计算，而是通过网络自身的训练来完成的。

从准确度来看，数据量越大，训练集则越多，分类器也就越精准。贝叶斯网络和 BP 神经网络的准确度较高，而决策树的精准性很大程度取决于数据的完整程度，某些字段上的缺值会影响其准确性。缺值越多，则决策树越不精准，且决策树存在过拟合现象的几率较高。针对食品安全检测数据

来说,检测指标较多,且很多检测结果值为“不判定”或“未检出”,导致缺值过多,会对决策树的学习造成较大的影响。

从训练速度来看,在大数据环境下,针对某种食品的检测指标繁多,即属性繁多。由于贝叶斯网络依赖于概率计算,属性组合的计算复杂程度会增加,使得预测难度加大,需要的时间也会更长。决策树由于进行深度优先搜索,算法受内存大小限制,难于处理大训练集,所以随着数据量增长,决策树的处理速度也会减慢很多。而基于感知器的神经网络,由于本身对处理不确定问题具有高度的解决能力,大量神经元的围观活动构成了神经网络的总体宏观效应,并且有很好的自适应性,随着数据量的增大,模型会越来越精准。不同于贝叶斯网络和决策树,神经网络是通过自我权值调整进行规则学习,因此从训练速度上,也会优于前 2 种算法。

从健壮性来看,由于食品检测数据常常出现空缺值(e.g. 如不判定),或是噪声(e.g. 如检测不准确),而对于有噪声或空缺值时,由于贝叶斯网络是通过概率计算来实现,无法准确地定义噪声或空缺值概率,会对模型训练造成一定的影响。决策树由于自身容错性较差,数据依赖性过强,数据噪声或不完整性都会对构建决策树模型造成影响。而神经网络本身具有较高的容错性,若一部分数据不完整,则神经网络可以从另一部分数据中学习隐含规律,通过自身权值调整,进行规律学习,从而构造健壮的模型。

综上所述,BP 神经网络以其准确率性高,训练速度快,健壮性强优于其他 2 种数据挖掘方式,且其以并行处理、自学习自适应强,实时性、容错性强等见长。且 BP 神经网络具有较强的灵活度,新的训练数据集可以简便的被用于模型训练当中,从而提高模型的准确性,很适合应用于食品安全风险预警领域。因此,该研究探讨将 BP 神经网络运用到食品安全预警领域的具体设想。

4 BP 神经网络在食品风险预警领域的可能性运用设想

BP 神经网络是人工神经网络中的一种,是人工智能的重要工具,其通过大量样本训练得到模型隐含规律。

在食品检测中,人们往往得到简单的“合格”或“不合格”的判定结果。这样的检测结果虽一目了然,但是对于食品安全风险的控制并无帮助。如果能基于食品各检测指标的具体检测值,对该食品的风险程度进行一个分级评价,有助于为有关风险评价部门提供决策支持。

传统的风险评级方式有专家打分[19]、风险矩阵[20]等。这些的方式虽然较准确,但由于专家打分基于人为评价,风险矩阵计算复杂度高,耗费较高的人力、物力,都不适用于食品安全大数据的风险预警。于是可根据 BP 神经网络潜在的规律,让其学习专家打分方式的风险分级。当存在新的检测数据时,其可根据学习到的规律进行评价。笔者认为,在大数据环境下,BP 神经网络十分适用于基于某类食品的食品安全风险预警。

首先通过筛选影响某类食品检测结果的不同维度,如化学污染、农药残留、兽药残留、重金属情况、致病菌等,采用专家打分法,由专家结合以上不同维度检测项目的检测结果进行一个风险评级。对不同历史数据样本期望得到的评级不同。之后将以上维度的的检测值作为神经网络的输入神经元,神经元的数量由选取的维度决定,并将通过专家打分得到的评级作为目标输出神经元,进行训练。通过将大量的输入、目标输出样本送入神经网络,让其通过正向传播、反向传播和权值调整进行潜在的规则学习。这样当有未知数据时,通过将其输入神经网络,神经网络即可模拟专家进行评级。由于神经网络具有较高的灵活性,新的数据及评级又可以作为神经网络的训练集。这样,随着数据量的增大,神经网络模型将越来越精确,以至于减少人为因素导致的错误及人力成本。

5 结语

该研究首先对大数据的基本概念进行剖析,并分析大数据挖掘中机器学习领域 3 种典型的挖掘方式,探讨其在食品安全风险预警领域的运用。之后对比分析 3 种大数据挖掘方式应用于食品安全风险预警领域的适应度,最后提出将 BP 神经网络应用于食品风险预警领域的方式,并给出了 BP 神经网络优于其他 2 种数据挖掘技术的解释。

参考文献

[1] 孟小峰,慈祥. 大数据管理:概念、技术与挑战[J]. 计算机研究与发展,2013(1):146-169.

[2] Big data[EB/OL]. [2012-10-02] http://en.wikipedia.org/wiki/Big_dat.

[3] GROBELINK M. Big data computing:Creating revolutionary breakthroughs in commerce,science and society[R]. 2012.

[4] BARWICK H. The ‘four Vs’ of Big Data. Implementing Information Infrastructure Symposium[EB/OL]. [2012-10-02]. http://www.computerworld.com.au/article/396198/iiis_four_vs_big_data/.

[5] IBM. What is big data? [EB/OL]. [2012-10-02]. http://www-01.ibm.com/software/data/bigdata/.

[6] WITTEN IAN H,EIBE FRANK. Data Mining:Practical machine learning tools and techniques[M]. Morgan Kaufmann,2005.

[7] 韩家炜,坎伯. 数据挖掘:概念与技术[M]. 北京:机械工业出版社 2001:100-103.

[8] PEARL JUDEA. Probabilistic reasoning in intelligent systems:networks of plausible inference[M]. San Mateo,Calif:Morgan Kaufmann Pub,1988.

[9] 林士敏,田凤占. 贝叶斯网络的建造及其在数据采掘中的应用[J]. 清华大学学报:自然科学版,2001,41(1):49-52.

[10] 冀俊忠,刘椿年,沙志强. 贝叶斯网模型的学习,推理和应用[J]. 计算机工程与应用,2003,39(5):24-27.

[11] CORNEY D. Designing food with bayesian belief networks[C]//ACDM 2000 fourth international conference on adaptive computing in design and manufacture. Springer London,2000:83-94.

[12] 张丽,滕飞,王鹏. 基于贝叶斯网络的食品供应链风险评价研究[J]. 食品研究与开发,2014(18):53.

[13] MITCHELL TOM M. Machine learning[M]. WCB,1997.

[14] 赵静娴. 基于决策树的食品安全评估研究[J]. 安徽农业科学,2012,39(3):20259.

[15] 欧阳一非,薛丹,高海燕,等. 基于决策树方法的油炸型方便面品质评价研究[J]. 食品科学,2009(5):27-31.

[16] 何勇,李晓丽,邵咏妮. 基于主成分分析和神经网络的近红外光谱苹果品种鉴别方法研究[J]. 光谱学与光谱分析,2006,26(5):850-853.

[17] 陈博,欧阳竹. 基于 BP 神经网络的冬小麦耗水预测[J]. 农业工程学报,2010,26(4):81-86.

[18] 刘建学,吴守一. 基于近红外光谱的神经网络预测大米直链淀粉含量[J]. 农业机械学报,2001,32(2):55-57.

[19] 郝书池,姜燕宁. 基于改进型主成分分析法的食品供应商评价模型研究[J]. 物流技术,2010,29(8):62-64.

[20] 刘清珺,陈婷,张经华,等. 基于风险矩阵的食品安全风险监测模型[J]. 食品科学,2010(5):86-90.