食品安全大数据的融合及分类技术综述

张素智[1],陈小妮[1],李鹏辉[1],杨 芮[1],蔡 强[2]
(1. 郑州轻工业大学 计算机与通信工程学院,河南 郑州 450002;
2. 北京工商大学 食品安全大数据技术北京市重点实验室,北京 100048)

摘 要:食品是人们赖以生存和发展的基本物质基础,食品安全不仅仅关乎广大消费者的切身利益,甚至关系到国家经济的稳步发展和社会的繁荣昌盛。食品安全大数据具有数据容量大、来源多样、更新速度快、价值密度低却应用价值大的特点,通过将多源的食品安全大数据进行融合及分类并行处理可以帮助人们实现更多的价值。对食品安全大数据融合及分类技术进行了综述。首先,总结了食品安全大数据的来源特征以及数据处理关键技术,阐述了食品安全大数据预处理过程,分析了食品安全大数据融合三种融合层次以及融合关键技术,介绍了食品安全大数据的并行计算模式;然后,归纳了并行分类算法以及几种常见的分类算法,如朴素贝叶斯、决策树、神经网络等;最后,对食品安全大数据做出总结和展望。

关键词:食品安全大数据;预处理;数据融合;数据挖掘;分类
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2020)02-0159-07
doi:10.3969/j.issn.1673-629X.2020.02.031

Review on Food Safety Big Data Fusion and Classification Technology

ZHANG Su-zhi[1], CHEN Xiao-ni[1], LI Peng-hui[1], YANG Rui[1], CAI Qiang[2]
(1. School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China;
2. Beijing Key Laboratory of Big Data Technology for Food Safety, Beijng Technology and Business University, Beijing 100048, China)

Abstract: Food is the basic material basis for people's survival and development. Food safety is not only related to the vital interests of consumers, and even related to the steady development of the national economy and social prosperity. Food safety big data has the characteristics of large data capacity, diverse sources, fast update speed, low value density but great application value. The fusion and parallel processing of multi-source food safety big data can help people realize more value. The fusion and classification technology of food safety big data is reviewed. Firstly, we summarize the sources and characteristics of food safety big data and the key technologies of data processing, describe the food safety big data pretreatment process, analyze three fusion levels and key fusion technologies of big data fusion of food safety, and introduce the parallel computing mode of food safety big data. Then, we summarize the parallel classification algorithms and some common classification algorithms, such as naive Bayes, decision tree and neural network. Finally, we summarize and look forward to the big data of food safety.

Key words: food safety big data; pretreatment; data fusion; data mining; classification

0 引 言

随着信息时代的到来,大数据迅速发展,逐渐成为科技界和企业界关注的热门话题[1]。互联网和各产业数据的爆炸式增长,使得大数据、云计算等概念越来越广泛。大数据概念的兴起为人们打开了一个新视角,为了更大程度地发挥大数据的价值,大数据挖掘成为了人们的关注热点。与此同时,食品安全相关事件在国内不断发生[2],如“洗衣粉油条”事件、“陈化粮毒米”事件、“铁酱油”事件、“毛发酱油”事件以及牛奶业普遍使用三聚氰胺的事件等,给人民的生命和国家的发展带来严重的威胁。食品安全从原料生产到消费,涉及食品链的各个环节,产生了大量的数据。处理与

收稿日期:2019-03-08 **修回日期**:2019-07-09 **网络出版时间**:2019-09-25
基金项目:国家自然科学基金(61802353);北京市重点实验室开放课题(BKBD-2017KF08)
作者简介:张素智(1965-),男,教授,博士,通讯作者,CCF 会员(07791M),研究方向为 Web 数据库、分布式计算和异构系统集成;陈小妮(1993-),女,硕士研究生,CCF 会员(96585G),研究方向为数据挖掘。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190925.1525.068.html

分析数据量大、数据结构复杂的食品安全大数据,传统的技术手段很难满足要求,因此实现食品安全和大数据产业的融合,增强食品安全大数据的分析,成为了研究的重点方向。

针对食品安全大数据处理关键技术,重点介绍了食品安全大数据预处理、食品安全大数据融合、并行挖掘技术、并行挖掘算法这几方面内容。目前,许多研究人员针对食品安全大数据处理技术进行了大量的研究。例如,孟小峰等[3]详细解析了大数据的基本概念,介绍了大数据处理的基本框架以及大数据的主要应用;王志海等[4]提出了一种懒惰式 shapelets 分类模型,该模型主要依据待分类实例显著局部特征,为各个待分类的实例构建各自的数据驱动懒惰式分类模型,该模型不但具有高准确率,还具有强可解释性;季一木等[5]基于分布式计算平台提出了一种 Storm 的 P-HT 并行化算法,该算法解决了概念漂移问题,同时提高了分类算法的有效性和高效性;宋杰等[6]介绍了 12 个典型的基于 MapReduce 的大数据处理平台的实现原理和适用场景以及基于 MapReduce 的大数据分析算法,并在对外存算法特征进行分析的基础上,提出了适合外存算法性能优化方法的研究思路;程学旗等[1]综述了大数据的应用场景,总结了大数据处理系统的关键技术,梳理了大数据处理所面临的各种挑战,并依次提出了应对措施。

文中对食品安全大数据进行概要性描述,概述食品安全大数据来源、特征以及处理关键技术和挖掘基本流程。总结了食品安全大数据预处理,对食品安全大数据融合的三个层次进行分析和对比,并对已有的食品安全大数据的关键技术进行总结。针对食品安全大数据并行挖掘技术,介绍了并行计算模式。针对食品安全大数据并行挖掘算法的设计,对几种常用分类算法进行总结和比较。最后总结全文并展望未来食品安全大数据面临的挑战和热门研究方向。

1 食品安全大数据概述

食品安全大数据作为大数据的一种,符合大数据的典型 4V 特征,即量大(volume)、多样(varity)、高速(velocity)和价值密度低却应用价值大(value)[7]。食品安全数据作为食品安全大数据处理对象,需要对其进行充分了解,包括:数据来源、数据特征以及处理关键技术,然后才能更加有效地挖掘其信息中的价值。本节介绍了食品安全大数据的来源与特征、食品安全大数据处理关键技术和食品安全大数据挖掘基本流程。

1.1 食品安全大数据来源及其特征

信息时代,食品安全数据来源范围较广,在日常生活中人们能够接触到的与食品相关的数据都在范围之内,主要包括:各种食品安全检测装置的结果;RFID 传感器的食品质量检测数据;企业和监管部门;移动互联网、社交媒体等。食品安全数据涵盖了多种类型,数据量随时间的积累变得越来越大[8]。

食品安全大数据除具有大数据的 4V 特性外,受错综复杂的食品安全环境、消费人群、监测数据飞速增长等因素的影响,还具有如下具体特征[9]:

(1)数据容量大。来自食品安全监测点、哨点的数据,各个地方上报的食品污染物数据,食品安全环境监测数据和其他食品企业自身生产的数据,这些数据聚集在一起就形成了十分庞大的数据库。

(2)更新速度迅速。食品安全信息中包含大量的在线或实时数据分析和处理要求。

(3)种类多。食品安全数据包含各种结构化数据、非(半)结构化数据和其他多种数据存储形式。

(4)成本低、价值大。食品安全大数据中存在着大量无用、冗余的信息,但这些信息具有很大的挖掘和应用价值,与个人生活、食品行业、国民经济息息相关。

1.2 食品安全大数据处理技术

食品安全大数据模型中,层次与层次之间联系紧密,原始的食品安全数据存在很多的冗余和噪音,需要经过数据清洗和提炼、数据融合等预处理的方式转化为规范数据,再经过并行处理、分类等挖掘技术来获取有价值的信息,其采用的关键技术如图 1 所示。

2 食品安全大数据预处理

食品安全大数据预处理的目的主要有:①清除冗余数据;②纠正错误数据;③完善残缺数据;④选出必需的数据进行集成。另外,对食品安全大数据进行预

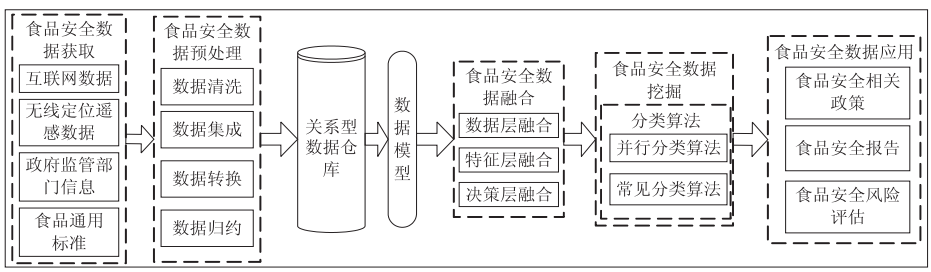

图 1 食品安全大数据处理技术

处理后再挖掘,可以大大提高数据挖掘的质量,缩短实际挖掘所需的时间[10]。食品安全大数据预处理一般包括 4 步:清洗、集成、转换、归约。本节将从这 4 方面介绍食品安全大数据预处理。

2.1　大数据清洗

食品安全大数据的清洗主要是为了检测食品安全数据中的冗余数据、错误数据、不一致数据等噪声数据。一般的清洗内容主要包括:清除重复数据、完善缺失数据、消除噪声数据等[11]。食品安全大数据的清洗技术大致可以分为以下几类:

(1)重复数据的清洗。由于在食品安全数据集中存在重复的记录,为了提高食品安全数据的挖掘效率,对重复数据进行清洗尤为重要。

(2)缺失数据清洗。食品安全大数据清洗需要解决的另外一个重要问题是完善缺失数据。对缺失值清洗的方法有很多,文献[12]提出了一种基于 MapReduce 的大数据缺失值填充算法,用来解决缺失值填充问题,该算法通过 MapReduce 框架中的两种算法实现了大数据处理的并行化。

2.2　大数据集成

由于食品安全大数据具有多源性,因此在对食品安全大数据进行数据处理过程中势必涉及到多个数据库。大量冗余数据可能会影响信息发现过程的性能。因此需要对食品安全大数据进行集成,将多个数据源合并成一致的数据源存储。经过有效的数据集成,能够提高食品安全大数据的挖掘精度和速度。

2.3　大数据转换

食品安全行业在长期的业务实践中累积了大量独立分布异构的数据,这些数据不仅具有不同的数据类型,而且具有不同的存储方式。这些都要求食品安全大数据在集成过程中对数据进行转换。通过转换将食品安全大数据变成适合挖掘的形式。

2.4　大数据归约

食品安全大数据的典型特征是数据规模大,如果直接进行数据挖掘、分析,将消耗大量的时间和精力,并且分析结果也会比较差。而通过归约技术可以将大规模数据集转换为小规模数据集,这样不但保持了原数据的完整性,又为进一步的数据挖掘提供了方便。

3　食品安全大数据融合及关键技术

食品安全大数据融合作为一种技术手段,可以在最大程度上发挥食品安全大数据的价值,它的实现可以使人们对食品安全行业的探索和认识向新的深度和广度拓展。它不同于传统的数据集或知识库技术,需要大跨度、深层次和综合性的研究方法。食品安全大数据的融合层次可以分为数据层融合、特征层融合和决策层融合[13]。文中主要工作是对 3 种层次的融合以及食品安全大数据融合关键技术进行介绍。

3.1　数据融合结构分类

(1)数据层融合。

数据层融合又叫像素级融合,在食品安全大数据中经过数据层融合不仅能够最大程度上保留原始食品安全数据的特征,而且能够提供较多的细节信息[14]。融合过程如图 2 所示。

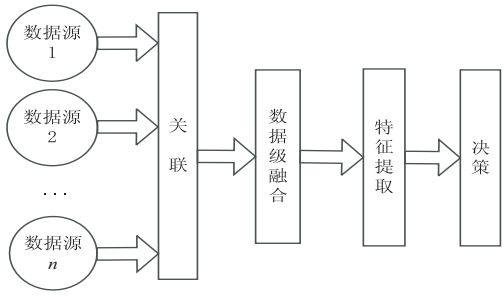

图 2　数据层融合过程

数据层融合作为食品安全大数据融合的最低层次融合,用以消除食品安全数据中的冗余信息,去噪和去异常值。

(2)特征层融合。

特征层融合在食品安全大数据融合过程中属于中间的一个层次。融合过程如图 3 所示。从图中可以看出,特征级融合首先提取特征信息,然后进行融合。特征层融合可以在食品安全大数据融合过程中做到较好的信息压缩,从而减少了数据融合的通信量。相对于数据级融合,特征层融合具有更好的实时性。在食品安全大数据中为了保证数据融合精度,特征层融合常采用的方法有:人工神经网络、特征压缩聚类法、卡尔曼滤波等。

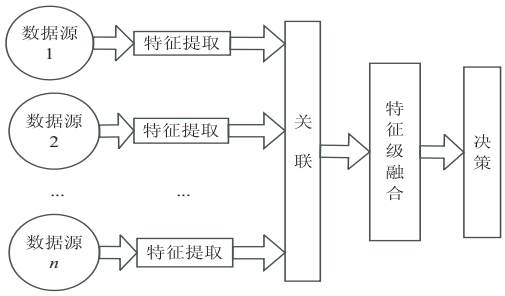

图 3　特征层融合过程

(3)决策层融合。

决策层融合在食品安全大数据融合中属于一种更高层次的融合。融合过程如图 4 所示。通过各传感器的食品安全大数据,在融合之前先完成各自的决策或识别工作,随后将这些决策进行融合,最终获得具有整体一致性的决策结果。

(4)大数据融合层次比较。

总体来说,三个层次的融合在食品安全大数据融

合中各具优势。如表 1 所示,从对传感器的依赖性、数据量、通信量等方面对比分析了几个融合级别的优缺点。

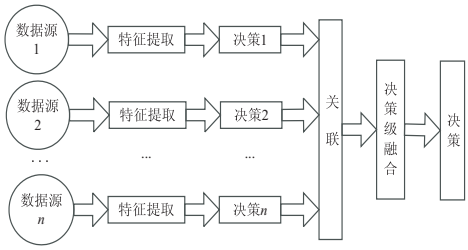

图 4 决策层融合过程

表 1 数据融合级别对比

融合级别	数据级	特征级	决策级
传感器依赖性	同质	不限	不限
数据量	大	中	小
通信量	大	中	小
信息损失	小	中	大
处理代价	大	中	小
实时性	小	中	大
抗干扰性	小	中	大
融合精度	大	中	小

可以看出,由于数据级融合是最基础层次融合,能够在保全尽量多信息的条件下对食品安全大数据进行数据融合,但是对传感器、通信能力、处理代价等要求较高;相反地,决策层融合多源异构食品安全大数据的同时,仅需要较小的数据线路通信,也有较好的通信量,但融合精度低。特征级数据融合各项性能居中,综合了其他两个层次的优缺点。

3.2 数据融合关键技术

食品安全大数据融合方法可以分为经典融合方法和现代融合方法。在经典融合方法中一般采用加权平均数法、卡尔曼滤波法、贝叶斯推理法等方法。在现代融合方法中常常采用神经网络、逻辑模糊法等方法。具体结构如图 5 所示。

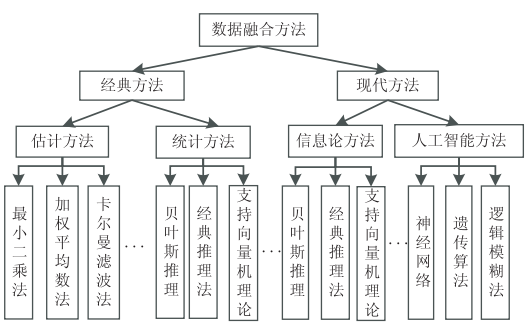

图 5 数据融合算法结构

(1)估计方法。

估计方法主要包括最小二乘、加权平均数、卡尔曼滤波等线性估计方法,以及一些非线性估计方法,主要有高斯滤波、扩展的卡尔曼滤波等。

卡尔曼滤波法一般用于动态环境中多传感器信息的实时融合,其算法核心是计算各传感器数据之间的加权平均值,其中权值与测量方差成反比。在实际应用中,通过调节各传感器的方差值来改变权值,从而得到更可靠的结果。

目前国内外对卡尔曼滤波法进行了大量研究。文献[15]提出一种基于压缩感知的扩展卡尔曼滤波跟踪方法,并将该方法应用到单目标跟踪中,与传统卡尔曼滤波相比,该方法具有更好的精确度和稳定度。文献[16]提出基于模糊卡尔曼算法的姿态误差补偿方法,通过引入模糊卡尔曼滤波数据融合算法对陀螺误差校正,与常规卡尔曼滤波算法相比,精度更高。针对食品安全大数据融合过程,采用卡尔曼滤波器对多传感器采集的食品安全数据进行融合,不仅可显著提高容错性,还可有效降低数据传输运算量。但是由于数据量巨大时,该方法的实时性较差,因此还需要进一步研究。

(2)统计方法。

统计方法一般常用的有贝叶斯推理、支持向量机理论、经典推理等等。

贝叶斯估计提供了一种按概率理论组合多传感器信息的方法,贝叶斯估计理论基础是贝叶斯法则。

文献[17]通过实验证明,利用贝叶斯估计方法对多传感器数据进行融合,可以解决数据的不确定和不一致性。通常来说,在先验概率已知的情况下,贝叶斯估计法是食品安全大数据融合的最佳方法。

(3)信息论方法。

信息论方法在多源数据融合中应用数理统计知识研究信息的处理和传递,其典型算法有:熵方法、模糊理论、模板法、最小描述长度方法等。

模糊理论在数据融合领域应用的实质就是利用一个模糊映射将数据源信息作为输入映射到融合结果的输出空间,其基本思想就是将原本只有两个取值 0 或 1,扩展到一个连续的取值范围:[0,1],用这个区间内的一个值来表示元素对某个模糊集的隶属程度,通过这种度量方法能够很好地描述和表达不确定事件。

模糊理论一定程度上克服了概率论方法的缺点,不需要一个确定的概率表达事情可能性,它对“可能性”的分析更加贴近人的处理方式。多传感器数据融合中,模糊集理论在处理模糊问题和模糊推理上具有显著优势。文献[18]通过实验证明,模糊集理论在多传感器信息融合中计算量小、融合精度较高。在食品安全大数据融合过程中,模糊集理论方法可以实现食品安全数据的简化,去除冗余信息。

(4)人工智能方法。

近年来人工智能方法蓬勃发展,被应用在多个领域,尤其在大数据融合领域应用十分广泛。人工智能

方法一般包括神经网络、遗传算法、逻辑模糊法等。

神经网络可对复杂的非线性映射进行模拟,具有运算速度快、适应能力强、容错率高等特点,使得神经网络能很好地适应多源数据融合的处理要求。BP(back propagation)神经网络是目前使用最普遍的一种神经网络,采用梯度搜索技术对输入的样本进行学习。

基于神经网络方法,文献[19]提出一种粗糙集结合 BP 神经网络的数据融合方法,该方法缩减了 BP 神经网络的规模,提高了数据融合效率,相比于传统的神经网络融合系统,具有较强的有效性。文献[20]提出基于 Mam dani 模糊推理的神经无网络,并应用于通侦信息融合系统。实验证明该方法同时具备模糊集理论和神经网络的优点,相比于贝叶斯、DS,该方法不需要给出先验概率。运用神经网络方法实现食品安全大数据融合,可以仅仅依赖食品安全原始数据样本,从而大大降低了食品安全数据的处理代价。但是,由于网络节点较多,训练需要大量的计算量和时间。另外,由于该方法对食品安全大数据的融合效果不是太理想,因此将神经网络与其他理论相结合还需要进一步的改进。

4　食品安全大数据并行挖掘技术

并行数据挖掘的基础是并行计算。针对食品安全大数据,使用 Hadoop 平台的 MapReduce 可以实现并行挖掘,MapReduce 是 Hadoop 的核心部分之一,主要用于处理大量数据集。

食品安全大数据的并行计算模式一般可以理解为两方面内容。首先将顺序执行的计算任务分成可以同时执行的子任务,然后通过并行执行这些子任务从而完成整个计算任务[21]。并行计算模式的实现可以提高食品安全大数据计算的速度。

在 MapReduce 模型中,程序执行过程主要存在两个核心操作,即:Map 操作和 Reduce 操作,Map 是对数据进行映射,Reduce 是对数据进行规约[22]。目前,运行 MapReduce 的集群往往由数十台、甚至数百上千台服务器组成,用于处理大规模数据。

5　食品安全大数据并行挖掘算法设计

食品安全大数据具有海量、高速变化、噪声、结构复杂等特点,对其进行快速准确的分类,是从食品安全大数据中提取符合需要的、精炼的、可理解信息的重要方法。分类技术是利用已有的训练样本去训练,从而得到一个最佳模型,再利用这个模型对测试数据进行类别判断从而实现分类的目的,也就具有了对未知数据进行分类的能力。本节主要介绍了几种典型的分类算法并对它们的性能进行简单的比较。

5.1　常见分类算法

5.1.1　朴素贝叶斯

朴素贝叶斯分类算法是基于贝叶斯定理,该算法的核心是概率统计知识,属于监督学习的生成模型,算法原理如下:

(1)设 $x=\{a_1,a_2,\cdots,a_m\}$ 为一个待分类的项,而每一个 a 为 x 的一个特征属性;

(2)有类别集合 $C=\{y_1,y_2,\cdots,y_n\}$;

(3)计算 $P(y_1\mid x),P(y_2\mid x),\cdots,P(y_n\mid x)$;

(4)如果 $P(y_k\mid x)=\max\{P(y_1\mid x),P(y_2\mid x),\cdots,P(y_n\mid x)\}$,则 $x\in y_k$。

其中,第 3 步中的每个条件概率的计算,一般采用如下步骤:

(a)找到一个已知分类的待分类项集合,这个集合称为训练样本集。

(b)通过统计得各类别下每个特征属性的条件概率估计值,即:

$P(a_1\mid y_1),P(a_2\mid y_1),\cdots,P(a_m\mid y_1);P(a_1\mid y_2),P(a_2\mid y_2),\cdots,P(a_m\mid y_2),\cdots,P(a_1\mid y_n),P(a_2\mid y_n),\cdots,P(a_m\mid y_n)$

(c)如果特征属性之间是条件独立的,则根据贝叶斯定理可以得出:

$$P(y_i\mid x)=\frac{P(x\mid y_i)P(y_i)}{P(x)}$$

对于所有类通常认为 $P(x)$ 为常数,所以只要将 $P(x\mid y_i)$ 最大化即可。又由于特征属性之间是条件独立的,可以得出:

$$P(x\mid y_i)P(y_i)=P(a_1\mid y_i)P(a_2\mid y_i)\cdots P(a_m\mid y_i)P(y_i)=P(y_i)\prod_{j=1}^{m}P(a_j\mid y_i)$$

对于大数据分类,朴素贝叶斯分类算法的分类效率比较稳定,尤其对于小规模数据。但在另一方面,由于食品安全大数据规模大,属性之间的关联性比较复杂,因此使用朴素贝叶斯分类算法效果不是太好,应该在考虑部分关联性的基础上对贝叶斯算法做进一步改进。文献[23]基于粗糙集的可识别矩阵,提出一种基于属性频率的加权朴素贝叶斯方法;文献[24]结合大样本集的缺点,将泊松分布模型引入到朴素贝叶斯分类算法中,从而提高了分类的精度;文献[25]介绍了代价敏感思想、构造出自适应代价函数,解决了不平衡数据分类问题;文献[16]给出了基于 MapReduce 并行化的朴素贝叶斯算法,该算法的核心处理过程由 MapReduce 完成,Map 函数完成对训练文件的解析,Reduce 函数完成类别属性和特征属性知识库的构建。

5.1.2　决策树

决策树分类算法是一种自顶向下递归建模算法。

该算法可以分为两大部分:构建决策树部分;使用决策树分类部分。

ID3 算法是决策树分类算法的经典算法,其用“信息增益”作为属性选择标准。由于 ID3 算法一般适用于离散型属性,因此提出了一种优化算法 C4.5。C4.5 算法用“信息增益率”进行计算,在运算过程中先将连续型属性转换为离散型,然后再进行属性分类。

针对食品安全大数据,采用决策树分类算法显著提高了食品安全数据的分类效果。另外,研究人员还提出大量的改进算法,例如,文献[26]对生成决策树算法的目标函数进行了改进,且对影响分类结果的约束条件中的特征进行了多方面衡量,从而提高分类节点的精确度;文献[27]提出一种基于粗糙模糊集的容错粗糙模糊决策树算法,与一般决策树相比,该算法具有较快的学习速度和较大的收敛概率;文献[28]提出一种 HAC4.5 决策树算法,该算法与 Hadoop 平台并行,不仅提高了运行速度,而且提高了计算精度。

5.1.3 神经网络

神经网络针对规模大、复杂度高、存在噪声等特点的数据,具有很强的承受力、较高的准确率和较强的分类速率。因此神经网络分类算法可用于食品安全大数据挖掘。但是当食品安全大数据的隐藏节点数量十分大时,实现食品安全大数据的分类将会消耗大量的时间。针对这个问题,文献[29]结合生物神经元学习和记忆形成的特点,提出了一种改进的 BP 算法,解决了网络学习慢的问题;文献[30]又提出了一种基于构造型神经网络的最大密度覆盖分类方法,进一步提高了神经网络的训练速度,同时提高了神经网络分类算法的有效性。基于以上四种算法的原理,综合分类精度、模型效率、非数值型数据处理能力、运行速度、模型结构等几方面给出如表 2 所示的对比情况。

表 2 典型分类算法综合对比情况

方法 标准	分类 精度	模型 效率	非数值型数 据处理能力	运行 速度	模型 结构
决策树	高	高	强	快	简单
贝叶斯	高	高	强	快	复杂
神经网络	高	一般	弱	快	复杂

5.2 并行分类算法

食品安全大数据具有海量、高速变化、噪声、结构复杂等特点,对其进行快速准确的分类,是寻找数据潜在规律的重要方法。传统的数据分类算法处理大数据时存在可行性差、效率低、分类精度不高等问题。而目前基于 MapReduce 模型的分布式并行处理架构成为处理海量数据的新方法。例如,文献[31]提出了一种在分布式环境中执行的决策树分类器构建算法,该算法与传统决策树分类器相比,对多处理器上的流数据具有可伸缩性。文献[32]回顾了分布式支持向量机(DSVMs)的研究现状,并分析现有的分布式支持向量机的优缺点,提出一些支持向量机算法分布的研究和有待解决的问题。文献[33]设计并实现了一种基于 MapReduce 架构的并行决策树分类算法,相比于传统的决策树和 ID3 算法,该算法不仅可以处理规模比较大的数据,还具有较好的可扩展性。因此,从并行计算出发,提高食品安全大数据分类算法的效率和精度是一个重要的研究方向。

6 结束语

食品安全大数据是食品安全科学发展的一种趋势,同样也是大数据研究的重要应用领域之一。随着全国科技水平的不断提高,食品行业积累了大量、来源多样、增长速度快、价值密度低却应用价值大的数据,如何分析、处理和利用这些数据,挖掘其内在信息价值,成为食品安全行业重点关注的问题[34]。大数据作为一门综合性科学,其理论体系不断成熟,随着新的理论和方法的形成,将会催生新的技术,这给研究人员学习利用大数据技术,实现食品安全大数据的更多价值带来了许多挑战。主要从以下几方面展望未来食品安全大数据所面临的挑战。

随着大数据时代的到来,针对当前多源、异构、海量的食品安全大数据,传统单一的处理模式和方法已经不能应对。而提升海量数据处理能力的问题迫在眉睫,同时分布式处理是当下最有效的手段。因此,根据不同的食品安全大数据处理要求,选择合适的分布式处理框架和处理算法,将成为未来食品安全大数据的研究重点。

在大数据和人工智能的不断发展下,深度学习越来越受重视,逐渐成为人工智能领域的研究热点[35]。深度学习被广泛应用于多个领域,目前在图像识别、语音识别、自然语言处理等领域取得了突破性的进展。文献[36]探索了深度学习在手写字符识别中的应用,提出卷积神经网络、深度信念网络两种深度学习算法并在实验中取得了较好的结果。文献[37]将 DBNs 运用到视听语音识别,测试了传统的结合单模态 DBNs 评分的决策融合和基于单模态 DBNs 学习的中级特征的新特征融合两种方法。由此可见,实现深度学习与食品安全大数据的结合,通过建立基于模式融合的深度学习方法,可以有效改善传统食品安全大数据分析处理的缺点,从而更大程度上实现食品安全大数据的信息价值。

参考文献:

[1] 程学旗,靳小龙,王元卓,等. 大数据系统和分析技术综述

[J]. 软件学报,2014,25(9):1889-1908.
[2] LIX,LI G,LIU Z. Mechanism design of generating the risk communication strategies responding food safety incidents [C]//2016 12th IEEE international conference on control & automation. Kathmandu:IEEE,2016:122-126.
[3] 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,2013,50(1):146-169.
[4] 王志海,张　伟,原继东,等. 一种基于 Shapelets 的懒惰式时间序列分类算法[J]. 计算机学报,2019,42(1):29-43.
[5] 季一木,张永潘,郎贤波,等. 面向流数据的决策树分类算法并行化[J]. 计算机研究与发展,2017,54(9):1945-1957.
[6] 宋　杰,孙宗哲,毛克明,等. MapReduce 大数据处理平台与算法研究进展[J]. 软件学报,2017,28(3):514-543.
[7] 张　引,陈　敏,廖小飞. 大数据应用的现状与展望[J]. 计算机研究与发展,2013,50(S):216-233.
[8] 陈　谊,刘　莹,田　帅,等. 食品安全大数据可视分析方法研究[J]. 计算机辅助设计与图形学学报,2017,29(1):8-16.
[9] 周广军,金志刚,王玮健. 食品安全大数据分析的若干思考[J]. 食品安全导刊,2017(36):127-128.
[10] KUHN M, JOHNSON K. Data pre-processing[M]//Applied predictive modeling. Berlin:Springer,2013:27-59.
[11] 郭志懋,周傲英. 数据质量和数据清洗研究综述[J]. 软件学报,2002,13(11):2076-2082.
[12] 金　连,王宏志,黄沈滨,等. 基于 Map-Reduce 的大数据缺失值填充算法[J]. 计算机研究与发展,2013,50(S):312-321.
[13] 周　鹏. 多传感器数据融合技术研究与展望[J]. 物联网技术,2015,5(5):23-25.
[14] 徐雅薇,谢晓竹. 多传感器图像融合方法及应用综述[J]. 四川兵工学报,2015,36(10):116-119.
[15] 常　娟,申晓红,钱　伟,等. 一种基于压缩感知的高精度目标跟踪算法[J]. 科学技术与工程,2019,19(2):101-105.
[16] 章雪挺,许　欢. 基于模糊卡尔曼的 MEMS 陀螺误差校正算法研究[J]. 杭州电子科技大学学报:自然科学版,2019,39(1):1-6.
[17] 孙振东. 面向多源数据融合的贝叶斯估计方法[J]. 齐鲁工业大学学报,2018,32(1):73-76.
[18] 杨永旭,陈旭辉. 模糊集理论在多传感器信息融合中的应用[J]. 计算机应用与软件,2011,28(11):122-124.
[19] GAO W G W,WEN J W J,JIANG N J N,et al. A study of data fusion based on combining rough set with BP neural network[J]//2009 ninth international conference on hybrid intelligent systems. [s. l.]:[s. n.],2006:103-106.
[20] 徐从富,耿卫东,谢　澍,等. 面向通侦信息融合的模糊神经网络方法[J]. 计算机研究与发展,2000,37(10):1212-1217.
[21] 王　彬,雷丽晖. 一种利用大数据分析优化的分布式并行算法[J]. 计算机与数字工程,2013,41(11):1720-1724.
[22] 李成华,张新访,金　海,等. MapReduce:新型的分布式并行计算编程模型[J]. 计算机工程与科学,2011,33(3):129-135.
[23] HE Y,XIE J,XU C. An improved Naive Bayesian algorithm for Web page text classification[C]//2011 eighth international conference on fuzzy system and knowledge discovery. Shanghai:IEEE,2011:1765-1768.
[24] HUANG Y,LI L. Naive Bayes classification algorithm based on small sample set[C]// 2011 IEEE international conference on cloud computing & intelligence systems. Beijing:IEEE,2011:34-39.
[25] 蒋盛益,谢照青,余　雯. 基于代价敏感的朴素贝叶斯不平衡数据分类研究[J]. 计算机研究与发展,2011,48(S):387-390.
[26] 王鹤澎,王宏志,李建中,等. 不一致数据上精确决策树生成算法[J]. 软件学报,2017,28(11):2814-2824.
[27] ZHAI J H,HOU S X,ZHANG S F. Induction of tolerance rough fuzzy decision tree[C]//2015 international conference on machine learning and cybernetics (ICMLC). Guangzhou:IEEE,2015:843-848.
[28] YUAN Z,WANG C. An improved network traffic classification algorithm based on Hadoop decision tree[C]//2016 IEEE international conference of online analysis and computing science (ICOACS). Chongqing:IEEE,2016:53-56.
[29] 刘彩红. BP 神经网络学习算法的研究[J]. 西安工业大学学报,2012,32(9):723-727.
[30] 黄国宏,熊志化,邵惠鹤. 一种新的基于构造型神经网络分类算法[J]. 计算机学报,2005,28(9):1519-1523.
[31] BEN-HAIM Y,TOM-TOV E. A streaming parallel decision tree algorithm[J]. Journal of Machine Learning Research,2008,11:849-872.
[32] STOLPE M,BHADURI K,DAS K. Distributed support vector machines:an overview[M]//Solving large scale learning tasks,challenges and algorithms. [s. l.]:Springer International Publishing,2016:109-138.
[33] 陆　秋,程小辉. 基于 MapReduce 的决策树算法并行化[J]. 计算机应用,2012,32(9):2463-2465.
[34] 肖革新,肖　辉,刘　杨. 食品安全大数据分析思考[J]. 中国数字医学,2014(1):4-7.
[35] AREL I, ROSE D C, KARNOWSKI T P. Deep machine learning - a new frontier in artificial intelligence research [research frontier][J]. IEEE Computational Intelligence Magazine,2010,5(4):13-18.
[36] WU M,CHEN L. Image recognition based on deep learning [C]//2015 Chinese automation congress (CAC). Wuhan:IEEE,2016:542-546.
[37] HUANG J,KINGSBURY B. Audio-visual deep learning for noise robust speech recognition[C]//2013 IEEE international conference on acoustics, speech and signal processing. Vancouver,BC,Canada:IEEE,2013:7596-7599.