分类号： TP391.41 密 级： 公开
U D C _____ 学校： 10127

内蒙古科技大學

硕士学位论文

论 文 题 目：融合气象因子的苏尼特羊肉价格预测的研究

英 文 题 目：Sunite mutton price prediction based on meteorological factors

学 位 类 别 ： 工程硕士
研 究 生 姓 名 ： 喻夏 学号： 2019033058
学科(领域)名称： 计算机技术
指 导 教 师 ： 杜永兴 职称： 教授
协 助 指 导 教 师 ： 周李涌 职称： 副教授

2023 年 6 月 10 日

摘　要

中国作为畜牧业大国，畜牧业在农业的占比越来越重，所以农畜产品对于国家经济影响也就越大。价格预测对于畜牧业的经济发展起到了关键作用，不仅可以将最新的数据信息传达给生产者，及时应对市场波动所产生的影响，还可以为生产者提供重要的决策方向，从而获取最大的经济效益。不同的产品对于影响其价格的因素各不相同。在内蒙古畜牧业中散养类型占比很大，农畜产品作为自然生产环境的产物，很容易受到自然因素的影响。例如，降水、温度、风速等将影响草原的长势，从而影响羊体重的变化，进而对羊肉的价格产生直接影响。因此，迫切需要提出一种更准确并结合气象因子的羊肉价格预测方法，以达到指导牧民科学养殖，指导生产决策，减少广大养殖户的损失，实现增收和利润最大化。

本次课题在内蒙古自治区重大专项的支持下，以苏尼特羊作为研究对象，并结合影响羊肉价格的气象因子进行预测。研究内容主要包含以下几个方面：

（1）搭建畜牧业大数据基础平台，采用实时和离线相结合的方式采集数据，并采用维度建模的思想分层搭建大数据仓库。并且自研热部署接口服务，可以管理接口平台并认证接口，保障了接口的安全性及可靠性。

（2）由于苏尼特羊与育肥羊最大的不同主要是以散养的形式培育，所以气象因子会有较强的影响，因为草场长势影响羊只体重，进而影响羊只产出总量和羊肉价格，所以本次研究过程中选取了气象因子。采用核 PCA 降维，选择累积贡献最大的主要气象因素作为动态贝叶斯网络的输入参数，以苏尼特羊肉价格为输出参数搭建预测模型。然后将所得结果与历年的苏尼特羊肉实际价格进行比较。结果显示，动态贝叶斯网络模型结合了核 PCA 具有更好的拟合性和准确性。

（3）选用支持向量回归模型对苏尼特羊肉价格进行预测，为避免输入变量过多、减小模型复杂度，采用 WOE 经验分箱提取预测能力强的变量并建立价格预测回归模型，其中气象影响因子作为自变量，苏尼特羊肉价格作为因变量，结果表明：采用 WOE 经验分箱并结合支持向量回归模型相比传统的模型，具有更高的预测准确性和更好的拟合性。该模型通过了显著性检验并反映了实际意义，因此可以进行实际预测应用。

关键词：大数据；核PCA；动态贝叶斯网络；价格预测

Abstract

As a big country of animal husbandry, the proportion of animal husbandry in agriculture is more and more heavy in China, so the impact of agricultural and livestock products on the national economy is also greater. Price forecasting plays a key role in the economic development of the livestock industry. It not only communicates the latest data and information to producers, but also responds to the impact of market fluctuations in a timely manner, it can also provide important decision-making directions for producers to maximize economic benefits. Different products have different influences on their prices. In the animal husbandry of Inner Mongolia, the proportion of free-range is very large. As a product of natural production environment, farm and livestock products are easily affected by natural factors. For example, precipitation, temperature, wind speed will affect the growth of grassland, thereby affecting the change in sheep weight, and thus have a direct impact on the price of mutton. Therefore, it is urgent to put forward a more accurate method of mutton price prediction combined with meteorological factors, so as to guide herdsmen scientific breeding, guide production decision-making and reduce the losses of the vast farmers, revenue and profit maximization.

With the support of the Major Inner Mongolia project, this research takes Sunith sheep as the research object, and combines the meteorological factors that affect the price of mutton to predict. The main contents of the study include the following aspects:

(1) The Building the animal husbandry Big Data Foundation platform, using real-time and off-line methods to collect data, and using the idea of dimensional modeling layer to build a big data warehouse. And self-developed hot deployment interface service, can manage the interface platform and authentication interface, to ensure the security and reliability of the interface.

(2) The biggest difference between Sunit sheep and fattening sheep is that they are bred in the form of free-range breeding, so meteorological factors will have a strong influence, because weather is strongly related to pasture growth, and sheep weight will

affect the price of mutton, so meteorological factors are selected in the process of this study. The kernel PCA dimension reduction was used to select the main meteorological factors with the largest cumulative contribution as the input parameters of the dynamic Bayesian network, and the Sunit mutton price as the output parameters to build the prediction model. The results were then compared with the actual price of Sunit lamb over the years. The results show that the dynamic Bayesian network model combined with kernel PCA has better fitting and more accurate accuracy.

(3) Support vector regression model is used to predict the price of Sunit mutton. In order to avoid excessive input variables and reduce the complexity of the model, WOE is applied to extract variables with strong forecasting ability and establish a price forecasting regression model, with meteorological impact factor as the independent variable and Sunit mutton price as the dependent variable. The results show that: Compared with the traditional model, using WOE empirical sorting and support vector regression model can achieve higher forecasting accuracy and better fitting. The model has passed the significance test and reflects the practical significance, so it can be used in practical prediction.

Key Words: *Big Data; kernel PCA; Dynamic Bayesian Network; Prediction*

目　录

1 绪论

1.1 研究的背景与意义

畜牧业作为我国的主要产业之一，对我国的经济发展产生着深远影响。在畜牧业中其农产品的价格也促进了产业的发展，从而也优化了资源配置。价格预测一直都是各大高校和企业的研究热点。一方面，人们可以通过价格走势状况分析目前的市场行情以辅助决策；另一方面可以通过对未来价格进行预测来减少损失，提高利润，及时做出战略调整。商品价格预测的研究价值已被国内外高度重视和评价，但这些研究停留在使用单变量模型对价格进行预测，或者虽然考虑了诸多因素对价格的影响，但未考虑气象因子对价格波动和预测的影响。农畜产品作为自然环境下生长的产物，很容易受到自然因素的影响。价格变化影响着经济的稳定程度，若想提前预知价格的变化趋势，那么价格预测显得尤为重要。价格表示商品的价值，预测即提前测算和推断，将二者合一，价格预测将使经济良好稳定的运行。价格预测可以让售卖者提前预知相关市场行情，可以自由且准确地调整自己的商品存储量，调节售卖时间以及人力，这样可以大大的节约成本，减少由于各种原因导致的生产者收入下降以及价格波动等问题，使决策者能有更准确的判断，以达到利益最大化。

在内蒙古畜牧业中牲畜散养类型占比很大，尤其是苏尼特羊，为保持优质的食用价值和其独特的口感，全年超半年以上都在牧区放养。气候因素是农畜产品面临的最大挑战。例如，降水、温度、风速将影响草原的长势，从而影响羊体重的变化，进而对羊肉的价格产生直接影响。对于传统的分析模型中对于气象因素的影响没有考虑的很全面，这样结果产生一定的误差是必然，所以该模型需要优化和改进，合理且充分地考虑到气象因素对畜牧业的影响，不仅可以及时规避一定的风险，也可以使牧民的利益达到最大，并利用对之后一段时间的预测结果进行评估和推测，所以找到更高效的预测方法，无论是对于牧民来说还是市场经济体系，都是一次史无前例的改革，对于其畜牧业的发展有着深远的影响和意义。

本文以苏尼特羊的相关数据为研究材料，提出由核 PCA 及动态贝叶斯网络相结合的方法以及 WOE 经验分箱并结合支持向量回归模型预测羊肉价格，其结果表示，相比传统的算法具有更高的拟合度和预测精度，从而可以指导牧民生产销售，达到利益最大化。

1.2 国内外研究现状

1.2.1 价格预测研究现状

自 70 年代开始，数据挖掘的思维已经萌芽，开始的阶段还是通过电子邮件进行，但随着通讯量都大量的增加，也开始通过计算机进行大批量处理数据。人工的方式预测产品的价格早已经不能满足当前数据时代的变化。数据挖掘技术在发展，我们需要更加精准的并且快速的预测方法才能满足当前时代。很明显计算机处理相关数据的能力要比人工计算快速且准确的多，处理问题不仅效率大大提高了，而且也大大的降低了人工成本。算法自身的优化也使计算机变得更加智能现代化，根据计算机输入参数进行机器化学习，使其选出最合理的参数，通过标准化处理建立相关预测模型进行预测。这种大批量的数据分析，一般所得结果的可靠可信度也很高[1]。上世纪 90 年代以来，随着计算机技术的不断提高和人工智能领域的迅猛发展，国内外学者对数据挖掘的发展更加深入和系统化[2]。

2000 年，马亮、朱亦斌讨论了数据挖掘计算更新的改进方法，加强了计算机处理数据的操作能力[3]。2015 年，李继娜、杨昱梅应用 AHP 和 BP 神经网络模型 AQZZ 对高等院校毕业学生开展相关探究和讨论，减少了以往传统的问卷调查的方式，依据 MATLAB 得到了相对精准的研究结果[4]。郭志荣、高峰、王其林在 2017 年通过人工智能创建了不同类别的传感器控制的两个方式，并且将这两个方法进行利弊的比对[5]。孟猛基于数据挖掘对畜牧产品的供应链进行深入讨论，得出结论畜牧产品的供应链中含有强相关风险并且提出其解决此风险的办法[6]。靳占新、徐中一两人在 2018 年基于大数据背景，开始预测通过使用线性回归模型进而预测价格，此次操作研究也验证了模型预测结果的准确性[7]。张利、王鹏等人通过使用神经网络模型，利用其对网络模型的预测导致产品价格发生的变化指标，建立相关机制，完成对产品价格预测警告的构建[8]。张喜红和王玉香等人通过 BP 神经网络的技术提出产品价格相关的预测，研究结果显示应用 BP 神经网络模型预测使产品的价格更加准确并且精准度也有所提升，也更加适合产品市场的价格[9]。我们国家在 2019 年已经将神经网络技术应用到人们生产生活的每个领域。张海妮应用神经网络模型提出了机床误差测算模型[10]。胡军研究了基于网络的声速剖面反演技术[11]。闵昌兆针对大数据时代背景下人工智能的未来研究展望作出分析，并解析了其相关特性，使用的优点对神经网络人工智能技术未来展开深远分析[12]。

截至目前，每个行业都在关注机器学习算法的变化，关注其未来发展方向。存在一部分发达国家利用自己国家本身的技术优势，创建人工智能产业链，因此机器

学习技术现已成为提高各个国家之间的竞争力，优化国家发展未来发展的不二之选[13]。美国以及加拿大的学者在 2015 年以生物医学使用神经网络获得了强大的的进展，通过利用并挖掘现存在的基因组，脑图像和临床评估数据库以及蛋白质组学等数据，得到了不在同一渠道的数据来源其间的关联性[14]。为让研究人员能够更加合理且愉快地完成聚合，管理，集成，操作以及建模等许多呈分布式排列的数据，作者创建了全新的功能模块[15]。丁磊等人提出了 ARIMA 与 GARCH 混合模型预测了黄金价格[16]。葡萄牙研究学者于 2016 年提出了一种提高混合的方式，此方法由三部分组成，分别是差分进化粒子群优化（DEEPSO），小波变换（WT）以及自适应神经模糊推理系统（ANFIS）的创新组合组成，以此来推测 EMP 信号[17]。任永富在应用程序以及遥感工具基于神经网络的背景下优化了相关算法的复杂程度[18]。英国研究学者将 SVM 模型与 EMD 方法组合的方法对三七的价格进行了预测[19]。2019 年，陈婷发现使用神经网络方法并结合 NARX 构建的投资组合预测价格的方法精准度高于传统模型[20]。

1.2.2 畜产品预测研究现状

现今经济预测方法被广泛应用在各种经济学领域的各个分支学科中,同样也有相当多的文献应用了现代预测方法对畜牧业商品价格预测。本节简要介绍了畜产品价格预测的一个理论的发展历史，并分析了农畜产品中价格预测所使用的模型方法。

首先是由傅如南系统的论述了经济预测的方法。然而他发现在原则上经济预测方法不太容易实现，其原因是经济预测的方法不是单独存在的，更不是同质的，恰恰概率统计的首要因素是独立性和同质性，因此经济预测不可以只基于概率论的推断[21]。Ethan Cottrill 却觉得大部分就经济并不是基于概率推测的预测方法，而是通过对过去进行推理预测。傅如南以及 Ethan Cottrill 的讨论对于了解经济预测的局限性和可能性产生了巨大的影响[22]。

在 20 世纪 50 年代后，时间序列逐渐成为主导的预测方法，大量模型都在更新迭代出现，因此更加需要研究出适合的算法对不同的模型进行评估。马孝斌指出大部分先验的方式是不可取的，并对预测的模型作出相关解释[23]。Ericsson 考虑到均方误差（MSE）的应用[24]。当使用不一样的方式进行研究预测时，MSE 是一种常用的统计变量，通常用于比较不同的预测结果。经济预测方式的应用相对来说更加广泛，并且属于经济研究的主要领域之一，是为各个行业应用到的当前先进的预测方式。在经济研究文献当中一般定量预测占大多数，但是现在经济的预测方法还包含

两类分别是定量和定性[25]。下面仅对量化的经济预测方法和模型进行了介绍，并将其应用于畜产品价格的研究。

此外还发现有其他相关文献在针对于畜牧业商品的价格当中应用了 ARIMA 模型。以 Oliveira 为例，它的拟合样本范围是从 1972 年起连续 4 年的周价格，运用 ARIMA 模型对 6 种奶牛的现货价格序列以及芝加哥期货交易所的活牛价格序列进行了拟合，并将 1977 年 1 月至 8 月份期间的周价作为预测范围[26]。通过研究成果我们可以了解到，相对于现货的价格预测，ARIMA 模型具备精准的短期预测能力；若想使长期预测加倍准确那么要选取 ARIMA 模型的短期预测。还有林丕源、傅如南、严尚维等人也有一些文献通过 Box-Jenkins 方法来进行畜产品价格的预测[27]。目前，向量自回归（VAR）模型已被广泛地用于动物产品价格的预测。例如，Goodwin 利用逐步跳越 VAR 模型和变参 VAR 模型在美国奶牛（1970-1990 样本范围内）三个月的价格进行估算计量，其发觉在固定的范围之间有清晰的结论数据表示其具有结构性变动[28]。Garcia、 Thraen 和 Zapata 运用 BVAR（Bayesian VAR）模型对美国的牛以及牛奶的价格进行了预计测量。马孝斌等人利用 VAR 模型拟合了北京市 2000-2006 年 6 月的每月猪肉价格，并测算了 2006 年 7-12 月 5 个月的猪肉价格[29]。Box-Jenkins 方法尽管能得出更加有效的短期预测数值，并且完全的拟合单变量时间序列数据，但此方法实际却是一种繁杂的滤波技术并不是经济模型(Naylor et al)，并且也不是建立在经济理论基础上的，因此我们无法找出使用这种方法是达不到理想的预测结果状态[30]。另外由于 Box-Jenkins 方法没有经济理论基础，若最后的目标是为阐述经济系统的行为且不只为获得相应的预测结论，那么就不可以使用这种方式[31]。

有相当一部分的研究学者针对每一个不一样的经济预测开展研究分析以及对照，并且在畜牧业产品价格研究预测之内。史惠婷使用了几种方法对美国内布拉斯加州奥马哈地区的猪肉以及牛肉价格进行了预测，对比了每种不同种预测结果的精准程度[32]。他将预测方法分为三部分：第一类为简约模型（simple models）,包括 5 个模型：使用前一年相同时期的价格当做预测的价格；采取现在的价格用作四个月之后的价格预测；从过去一年中的某个星期中，随机选取一个星期的行情作为预测行情；以上一年度的平均价格为预测值；根据过去八个月的行情来推测未来四个月的行情。第二种模式是以芝加哥期货交易所对应的期货价格做为预测价格[33]。第三类根据以往的经验进行预测，存在两个模型：一是根据以往可靠的农场负责人的预测作为当前预测值；二是通过 USDA 的预测价格。文中讨论了预测值以及真实数

值之间的标准差以及平均相对误差，并对不同预测方式所获得的成果进行对比讨论，最终显示结果为：即便是只凭经验所得到的结果也具有可信度，因此不可以单纯的觉得目前所使用的其中一种方法一定要比另一种方法看起来更加精准[34]。

由于预测技术的发展迅速，针对畜牧业商品价格的预测可以用来选择的方法肉眼可见的增加，并且也会伴随很多不同方式进行对比推论相关模型的预测结果展示。因此我们不可以片面的理解为其中一种方式一定好于另一种，但是可以让我们参考的是，当我们对于其中一个对象进行研究预测时，可以选用两种以及以上的方式进行比对探究，这样可以获得更加可靠且可信的预测成果[35]。

根据学者的研究我们可以发现，在现代的价格预测方法中优缺点大同小异，而与此同时，不同的预测方法也有着不同的对应的优缺点。以下对优缺点进行简单的讨论，都具备的优势是，现代的价格预测方法对可以处理数据的多样性的能力非常强大，一般可直接进行价格预测并且不用清洗数据[36]。都存在的缺点是，与传统的计量经济学、统计分析方法相比，现代的物价预测方法更为繁琐、复杂。总体而言，无论是小波分析还是 BP 神经网络，都可以很好地适应实际情况中的非线性数据，而且人工神经网络具有很强的容错性， Logistic 模型在使用的过程中，并不需要像传统的多元回归模型一样，需要对解释变量和随机误差项进行严格的假定[37]。已有的研究表明，无论是人工神经网络还是小波分析，都能极大地提高预测的准确率。但是这些方式中部分方法也具备各自的缺点，当然共性缺点除外。比如，Logistic 回归模型适合于定性变量的预测，其预测准确率会受到模型随机项的非正态性、异方差性等因素的影响，呈下降的趋势。而对于股票价格的预测，人工神经网络模型往往是收敛性较差的，而且为训练其要消耗大量时间，不能跟一般的统计方式或者专家相提并论，随时都清楚地掌握输入输出之间的联系，但是因为因果转换的联系就好像“暗箱”让读者了解不到以上价格预测方式所产生的利弊问题，且网络繁杂，因此我们大致确定了当前现代方式的应用范围[38]。通过对现代价格预测方法的研究，可以看出，尤其是人工神经网络等技术，在对一种商品进行长期价格预测时，表现出了较好的性能。例如梁强等人从实验中验证了小波长期预测法在石油长期预测中的显著性，通过应用小波方法对石油的价格进行了预测[39]。

针对权重分类方式的相关组合预测模型，其中心原理是将相应的模型预测所得值跟权重的乘积当做最后的预测结论，即是给任何一个参加组合的模型分担其相应的权重[40]。当操作详细的组合时，第一个要对每个模型采用相同的时间序列进行预测结果，再以此计算加权系数，通过计算以及检测来估算相关误差，其中系数的大

小比例跟任意模型的预测精准度成正比例关系。最终，将这些预测模型的结果根据加权的方法生成最后的值。在此类组合预测模型中，这些模型的相似与否没有特殊要求，但是通常情况下其输入的数据是一样的[41]。也有某些特殊的情况，输入数据有可能不同，比如在获得加权系数的时候。例如那些对预测结论产生较大影响的相关因素，能通过分析使用不一样的加权对不一样的因素通过系数的计算，权重与模型之间的输出值呈一对一的关系，再将每个模型的对应的权重跟输出值进行合并最后得出预测结论。由于数据内会存在随机波动以及噪音等无法确定的因素，此类初衷是让不同的模型来解决数据预测以及数据预处理两块[42]。其中预测的这一部分首要承担数据的拟预测。在进行数据的预处理模块时，为减轻数据的冗余特性以及让模型过滤与预测结果不相关的或者相关性极小的情况，通常将非线性时间序列数据分解成较为稳定的子序列进行预测。预处理不仅能提升预测模型的效率，清洗脏数据。同时也可以减少时间复杂度、降低模型的计算负担[43]。

以上的文献可以看出，国内外有许多关于畜牧价格的预测方法，但大都是地域性畜牧动物，对于主要为放养的牲畜来说的价格预测应用并不多。现如今随着人们生活水平的不断提高，人们对食物的品质及口味上要求也在逐渐增长，因此本文选择羊肉品质较高的苏尼特羊进行价格预测研究。同时本文在上述文献的研究基础上，选择了基于核 PCA-动态贝叶斯网络模型及基于 WOE 分箱结合支持向量机回归模型进行价格预测实验。

1.3 主要工作和论文结构

1.3.1 主要工作

本研究内容整体包括三部分：一是对影响价格的气象因子进行相关文献的调查研究，然后在建立畜牧业大数据基础平台，再对影响苏尼特羊肉价格的气象影响因子进行核 PCA 降维，通过降低入参的数据量增加预测效率。二是构建了动态贝叶斯网络模型预测。三是选用 WOE 分箱结合支持向量机回归模型进行预测。四是对模型进行误差分析，得出价格预测拟合效果最好的模型。

1.3.2 论文结构

此论文划分如下五大部分：

第 1 章，绪论。当前章节说明了此次研究的课题的来源背景，以及研究意义所在，对当前和以往的国内外畜牧行业相关畜牧业产品价格预测的当前现状的研究和

调研，查阅相关畜牧业等价格预测的书籍，通过对畜牧产品价格的研究分析，对现阶段畜牧行业的发展情况作出分析概括。

第 2 章，畜牧业大数据基础平台的设计与实现，以维度建模的思想建立了大数据基础平台，并且开发了热部署接口平台，可以更方便快捷的管理接口平台，并对接口的安全做了校验。并且将采集到的数据清洗到大数据平台，再分层进行处理，最终加工到 ads 数据应用层。

第 3 章，核 PCA_动态贝叶斯网络的价格预测分析。本文首先采用核主成分分析方法，对影响羊肉价格的气象因子进行降维，减少入参数，从而提高了模型的预测效率；再构建动态贝叶斯网络模型，并与未进行核主成分分析降维的动态贝叶斯网络模型进行对比实验。最后用误差分析法分析对照前后的模型的效果。

第 4 章，结合 WOE 经验分箱的支持向量机回归价格预测分析。本章节首先采用 WOE 经验分箱得到的预测能力强的气象影响因子，并建立支持向量机回归模型对苏尼特羊肉价格进行预测，选用常用误差指标对模型进行评估分析。

第 5 章，总结与展望。此章节通过展示出本人在探究的过程中存在的不足之处与疑难问题，对其进行总结归纳，查找解决方案，积极处理问题，为下一步的研究讨论做好铺垫。

1.4 本章小结

这一章主要介绍了选题的来源和背景。并对国内外价格预测的研究现状以及农畜产品价格研究现状进行了调研，及现代畜牧业的发展情况调研。

2 畜牧业大数据基础平台设计与实现

2.1 需求分析

随着物联网设备和移动互联终端在畜牧行业中的使用,导致在畜牧业各个环节都产生了大量数据。如果牧场管理者能及时感知这些数据并做出相应的决策,可提高生产效率和经济效益。针对所述背景，设计畜牧业大数据基础平台，该平台需求概括如下：

(1) 在畜牧业生产过程中，会产生出各种各样的数据，并且数据量非常大。这些数据包含了结构化数据、半结构化数据、非结构化数据。由于传统的存储方法很难满足如数据量过大、数据类型多样等需求。因此急需新的数据中台满足海量畜牧业数据存储的需求。

(2) 畜牧业数据中的部分数据来自于物联网的设备采集，如果在短时间内有数据洪峰，会给服务器带来很大的压力。为了提升平台的可靠性，减轻服务器压力、提高数据处理效率，有必要引进数据缓存机制。

(3) 畜牧业的数据类型种类繁多并且分散在不同的服务器上，并由于数据类型数据来源不同、数据规模巨大，传统的数据管理方式并不适用于对畜牧业的海量数据进行挖掘和分析，所以必须对这些数据统一规则，并将这些数据进行整合，构成一个畜牧业数据中台。

基于以上情况分析，为了提高畜牧业数据处理能力，急需开发一个能弥补上述短板的畜牧业大数据基础平台。本课题通过使用 Kafka 解决不同数据源不同数据类型的数据采集，打破了不同数据源的数据孤岛问题，解决因数据生成速度快，消费速度慢和数据多样性带来的性能瓶颈，同时也满足数据在实时分析和离线分析的共享。通过数据仓库对畜牧业数据种类繁多的现状以数据仓库的方式进行整合，帮助牧民们可以更高效的挖掘数据价值，为草原畜牧业的发展赋能。

2.2 平台总体架构设计

畜牧业大数据平台以畜牧业的需求为出发点，它的平台设计一共有五个部分，如图 2.1 中所示，它的结构分别是：数据采集层、数据存储层、数据计算层、数据仓库层与数据应用层。

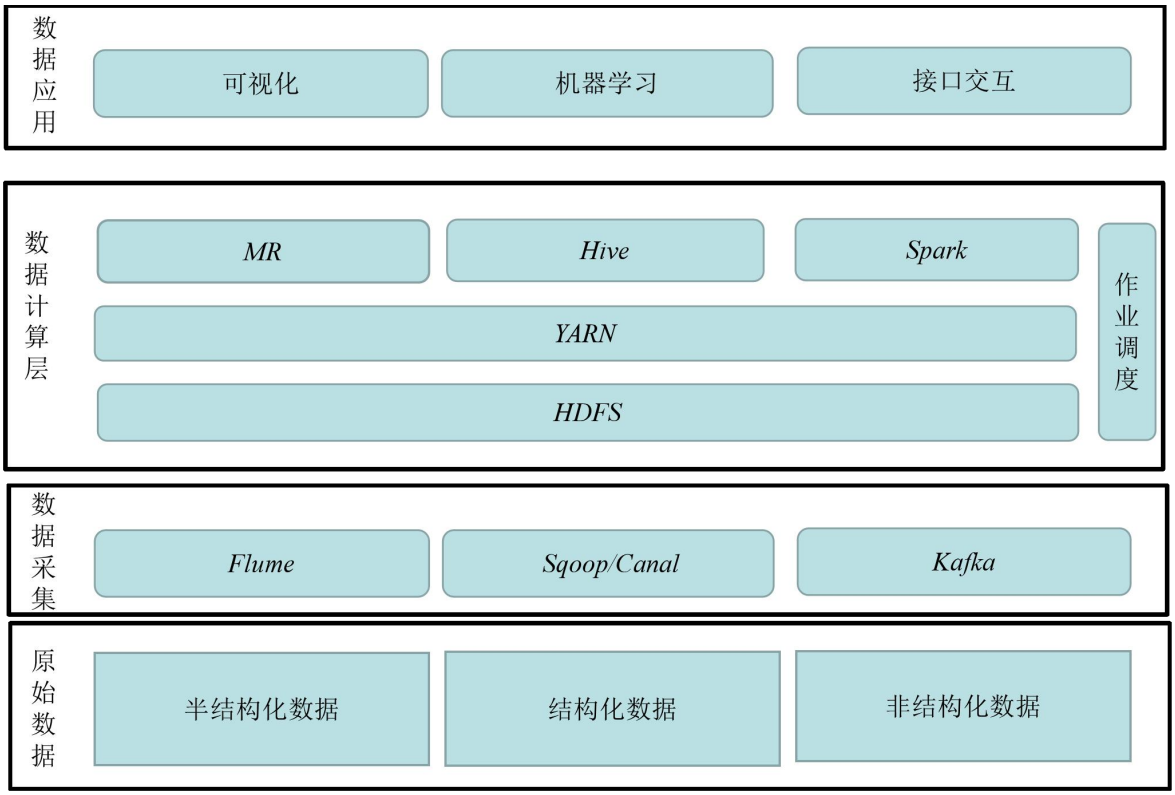

图 2.1　平台架构图

(1) 数据采集层

利用各种数据采集工具，实时地收集了草原畜牧业生产过程中所产生的数据，所收集到的数据分为半结构化数据和结构化数据两种。因为数据采集层对应不同的应用场景，所以在实现方式上会有差异，从而保证畜牧业大数据基础平台的安全性和可靠性。

(2) 数据存储层

将采集到的数据放在分布式文件存储系统(HDFS)中，由于 hive 的元数据是存储在 Derby 数据库，但是 Derby 数据库不支持多用户访问，所以将其对应的元数据存储在关系型数据库 Mysql 中。

(3) 数据计算层

计算层的计算引擎为 MapReduce 和 Spark，根据不同的业务情况，可以根据需要进行相应的选择。在一个计算任务开始之后，首先要以用户的配置为依据，向 YARN 提出任务运行资源的请求，在分配完资源之后，将 Spark 引擎作为一个例子，平台将计算任务所需的数据从文件系统中读取到 Spark 的内存中，在计算完成之后，再将计算结果输出到 HDFS 底层文件系统中[44]。

(4) 数据仓库层

数据仓库根据维度建模思想采用分层原则，由 ODS 原始数据层、DWD 数据明细层、DWS 数据服务层、ADS 主题层四部分组成。

(5) 数据应用层

数据应用层直接面向用户和数据需要方。可提供其接口满足不同用户对数据的交互需求。

2.3 平台模块设计实现

2.3.1 数据采集模块的实现

随着信息化技术在草原畜牧业中的应用日益广泛，物联网设备与服务器之间的数据交互也持续不断的进行，若采用人工或定时的方法对海量的数据进行同步，极有可能会导致数据写入时出现堵塞，进而影响系统的性能，严重时还会引起系统停机。因此，在这种情况下，可以利用采集工具 Flume 来采集畜牧业数据，并实时地将数据写入 HDFS, Flume 通过 JVM 进程以事件的形式将数据从源头传到目标端。这个过程包括三个主要的部分：Source、Channel、Sink。Source 是一个采集组件，它包含了多种数据源采集插件，比如 spooldir、 HTTP 等，它的作用是与数据源对接，从而获得数据。Channel 是一个传送信道组件，它的作用是把来自源的数据传送到 Sink[45]；Sink 将大量删除通道中的事件，并将它们写入存储系统。Flume 架构图如图 3.2 所示。

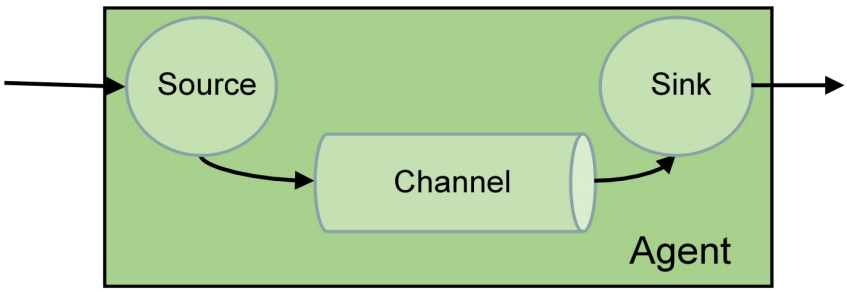

图 2.2　Flume 架构图

其中，在源与通道之间会建立一个事件处理器。Source 得到一个事件事件，它可以执行转换、清洁等操作，然后再通过通道选择器选择相应的通道。Flume 数据处理结构图如图 2.3 所示。

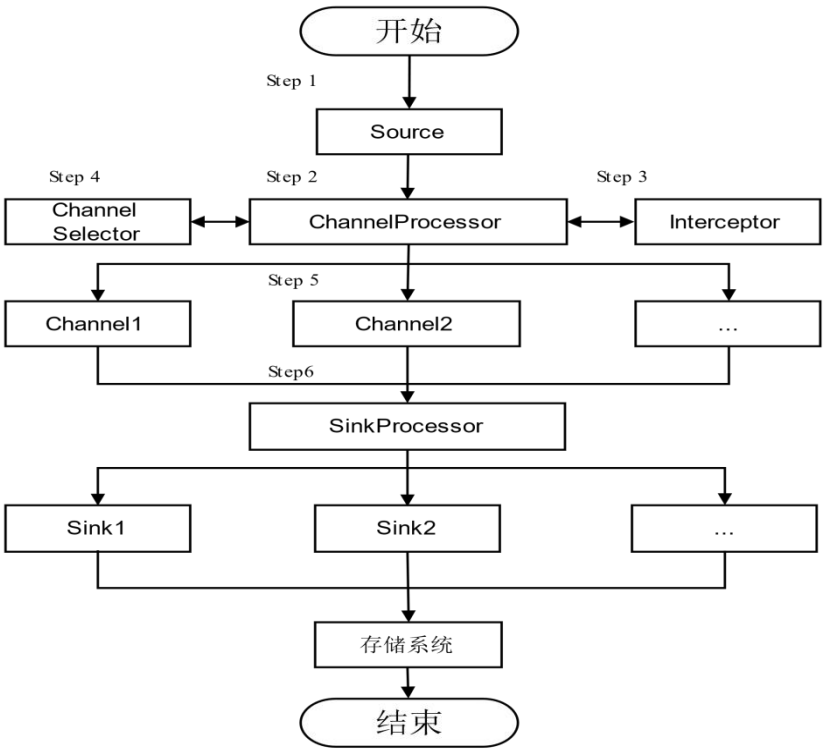

图 2.3　Flume 数据处理结构图

利用 Flume 技术，能够实现对畜牧业数据的实时收集，并且可以高效的解决需要定时同步海量数据所带来的例如数据倾斜等问题。由于草原畜牧数据有一部分来源于物联网设备采集，如果在短时间内有数据洪峰，会给服务器带来很大的压力。面对这种痛点，需要一个高性能的中间件来处理这样的突发事件。针对上述问题，本项目设计一种适用于此场景的采集方式，从而提升大数据平台的健壮性。

Kafka 有较高的吞吐量。是因为 Kafka 的生产者在写入数据的过程中，依次将数据追加到文件末尾。如图 2.4 和 2.5 所示，Kafka 通过追加写入消息，通过 offset 来表示读取第几条消息。

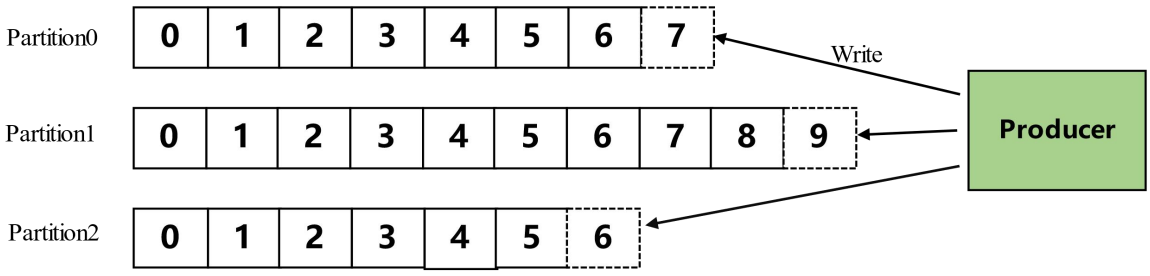

图 2.4　Kafka 顺序写

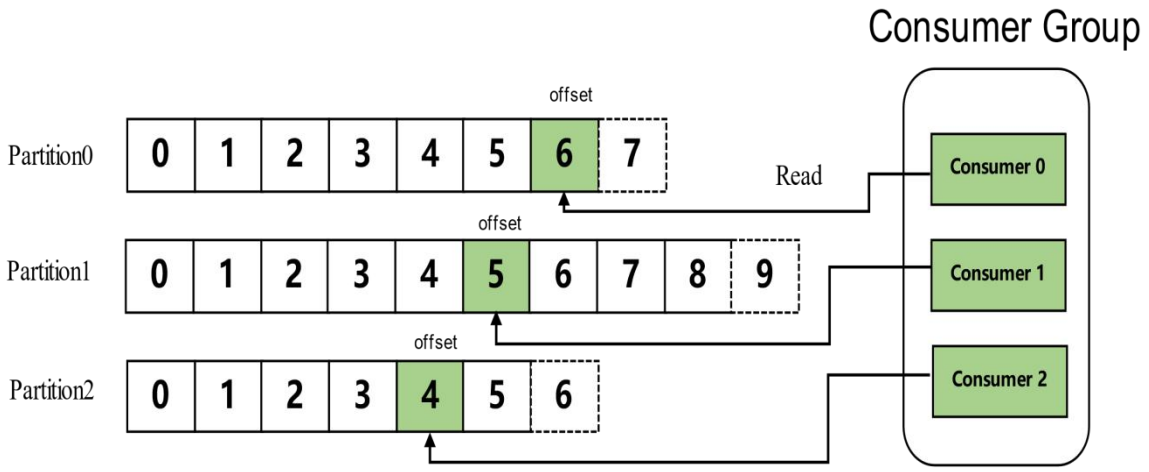

图 2.5　Kafka 顺序读

畜牧业大数据采集层通过 Flume 采集数据到 Kafka 中间件中，Kafka 能够控制和优化数据流在系统中的速度，从而解决在生产消息和消费消息过程中处理速度不一致的情况[46]。然后再通过 Flume 进行消费 Kafka 的数据写入存储层，这种模式非常实用。采集模型如图 2.6 所示。

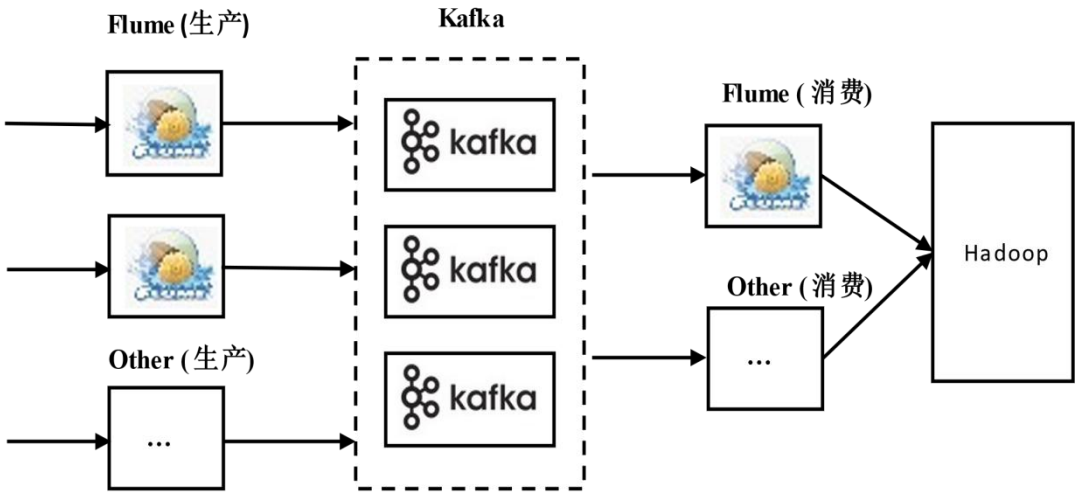

图 2.6　采集模型

为了完成畜牧业数据的收集，可利用 Kafka 数据库的高吞吐率、高时效性和高可扩展性。在畜牧业的应用场景中，物联网设备会实时地上传海量的数据，这些数据会被 Flume 推送到一个消息队列中，然后进行消费，从而完成数据的收集。这样，生产者与消费者就可以按照自己的需求进行生产与消费[47]。用户不需要担心是谁提供了这些数据，只要只需去指定的分区拉取数据；而且，生产者不必担心数据被谁拿走了，是否及时地拿走了，只要把这些数据放到一个队列中就可以了。该方法通过对数据传输速率的优化，有效地解决了数据洪峰问题。

Kafka 采用分布式集群部署在多台服务器上，以 Topic 作为基础单元对数据进行存储，在物理上把 Topic 分成一个或者多个 Partition，每个 Partition 通过创建副本分散到不同的 Brocker 来支持容错。数据采集过程中，数据的一致性很重要，Kafka 是一个高吞吐量的分布式消息系统，它的幂等特性是指在消息的生产和消费过程中，相同的消息能够被正确地保证只被处理一次，不会被重复处理。这是因为 Kafka 引入了幂等性生产者和消费者，以确保消息的唯一性[48]。在 Kafka 中，幂等性生产者的实现主要是通过消息的序列号来实现的。每个消息都会有一个唯一的序列号，当生产者发送一个消息时，会将序列号和消息内容一起发送给 Kafka 集群。Kafka 集群会检测消息的序列号，如果发现相同的序列号已经存在，则会认为这是一条重复消息，不会再次处理。幂等性消费者的实现主要是通过消息的偏移量来实现的。消费者会记录每个分区消费的最后一个消息的偏移量，当消费者重新启动时，会从记录的偏移量开始消费消息。如果有重复的消息出现，消费者会自动忽略，不会再次处理[49]。总之，Kafka 的幂等特性能够保证消息的唯一性，避免了重复处理和数据不一致的问题，提高了系统的可靠性和稳定性。Kafka 模型架构如图 2.7 所示。

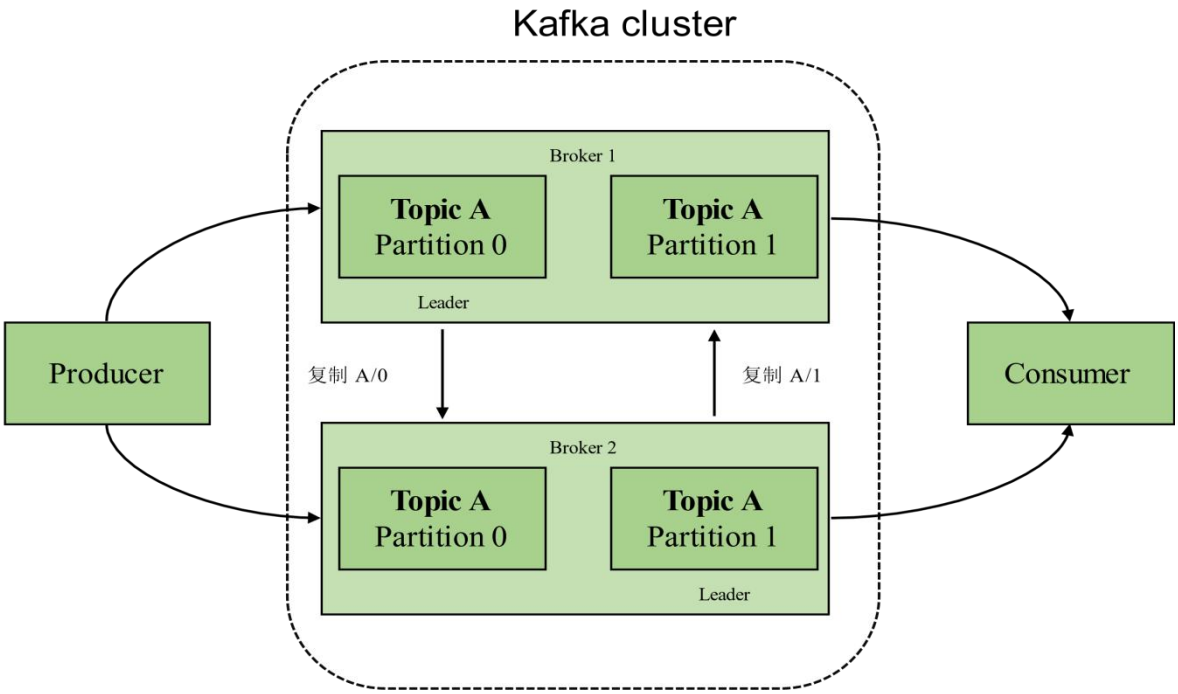

图 2.7　Kafka 模型架构

结合畜牧业大数据平台，对 Kafka 模型组件描述及功能实现如下：

（1） Producer

在 Kafka 中，信息的生产者将信息写到 Kafka 集群中，并在该平台上使用 Flume 将信息传输到 Kafka 集群。通过对源代码的二次开发，可以满足各种行业需要，并把收集来的数据吐到 Kafka。配置如下：

在 Flume 的工作目录下创建配置文件，搭建 Flume 到 Kafka 的配置伪代码如表 2.1 所示：

表 2.1　数据采集 flume_to_kafka 配置

flume_to_kafka 配置
定义组件 1. 配置 flume agent a. 设置 source：指定数据源 b. 设置 channel：指定数据缓存 c. 设置 sink：指定数据输出目标，即 Kafka 2. 配置 source a. 设置 type：指定数据源类型，如 exec、spooling directory 等 b. 设置相应的参数，如 exec 的 command、spooling directory 的 directory 等 3. 配置 channel a. 设置 type：指定数据缓存类型，如 memory、file 等 b. 设置相应的参数，如 file 的 path、memory 的 capacity 等 4. 配置 sink a. 设置 type：指定数据输出目标类型，如 Kafka b. 设置相应的参数，如 Kafka 的 brokerList、topic 等 5. 启动 flume agent

为确保 Kafka 在畜牧大数据平台上传输数据的可靠性，需要对其进行合理的设置。acks 指定了一个在生产者向数据写入成功前需要接收的分区数量，这个参数极大地影响了数据丢失的概率[50]。

acks=0: 如果将 acks 设置为 0，那么在发送完一条信息之后，生产者就不会再等待回复了，而是会立即判定这条信息是成功的。这样做能得到最大的吞吐率，但不能确保消息的可靠性，而且有可能导致消息的丢失。

acks=1: 当集群中的 leader 节点接收到一条消息时，服务器将向生产者发出一条消息，说明该消息已被成功发送。当一条消息未抵达 leader 节点时，生产者将收到一条错误的响应，为了避免数据丢失，生产者将再次发送该消息。

acks=-1: 每一个涉及到复制的节点都已经收到了一条消息，然后服务器将收到一个 ACK，以通知生产者已经收到了数据。即便其中一个服务器出现故障，由于保证多个服务器都能接收到消息，因此该集群仍能以此模式运行。

为了保证数据可靠性，本课题采取 acks=1。

（2）Topic

Kafka 生产者发布信息，而 Kafka 消费者则订阅信息，这些信息是在主题中持久保存的。在创建主题时，需要三个参数：topic（主题名称）、partitions（分区）、replication-factor（复制系数）[51]。分区和复制系数的设定对系统的性能有很大的影响，所以最好在一开始就把这些参数设定好。

选择分区数的基本原则：当 topic 的分区数量增加时，吞吐能力会得到提升，但是，越多的分区并不等于越高的吞吐量和处理能力，可以通过公式 2.1 得出：

$$n = \frac{T_t}{\max(T_p, T_c)} \quad \text{(式 2.1)}$$

其中：n 是分区数，T_t 表示期望吞吐量，T_p 表示 Producer 吞吐量，T_c 表示 Comsumer 吞吐量。

（3）Consumer

在 Kafka 中，信息的消费者对 Kafka 中的信息进行读取，并利用 Flume 将信息传输给 HDFS。也可以对 Sink Connector 做二次开发，以满足各种业务要求，最后把 Kafka 数据吐到其它数据库[52]。具体实现如下：

搭建 Kafka 到 Hdfs 的关键配置如下：

表 2.2 数据采集 kafka_to_hdfs 配置

kafka_to _hdfs 配置
定义 Flume Agent 名称；
配置 Kafka Source：
type：指定数据源类型为 KafkaSource。
zookeeperConnect：指定 Zookeeper 地址和端口。
topic：指定 Kafka Topic 名称。

batchSize：指定每次批量拉取的数据条数。

channels：指定数据源输出的 Channel。

配置 HDFS Sink：

type：指定 Sink 类型为 HDFS。

hdfs.path：指定 HDFS 路径，其中%Y-%m-%d 会被替换为当前日期。

hdfs.filePrefix：指定 HDFS 文件名前缀。

hdfs.fileSuffix：指定 HDFS 文件名后缀。

hdfs.rollInterval：指定 HDFS 文件滚动的时间间隔（单位：秒）。

hdfs.rollSize：指定 HDFS 文件滚动的文件大小（单位：字节），0 表示不限制文件大小。

hdfs.rollCount：指定 HDFS 文件滚动的文件数量，超过该数量会自动创建新文件。

hdfs.writeFormat：指定 HDFS 文件的写入格式。

channel：指定 Sink 从哪个 Channel 获取数据。

配置 Kafka Channel：

type：指定 Channel 类型为 Memory。

capacity：指定 Channel 的最大容量（单位：条）。

transactionCapacity：指定 Channel 的最大事务容量（单位：条）。

2.3.2 数据仓库模块的实现

ApacheHive 是一个以 Hadoop 为基础，能够对海量的结构性数据进行管理的数据仓库工具。Hive 定义了一种类似于 SQL 的语言，可用于读、写和管理存储大量数据。Hive 的实质是一个通过 HQL 转换到 MapReduce 的运算程序，它所管理的是一个分布式的文件管理系统中的数据。它省去了开发者编写 MapReduce 编程的时间，降低了学习的成本，而且操作简单，易于上手。在响应速度方面， Hive 与传统数据库相比更侧重于对大量数据的分析[53]。由于 Hive 的执行效率并不是很高，有些任务需要更高效率的任务则可交给 spark 计算引擎去计算。Hive 主要由用户接口、元数据(Metastore)、Hadoop、驱动器组成，Hive 执行流程图如图 2.8 所示：

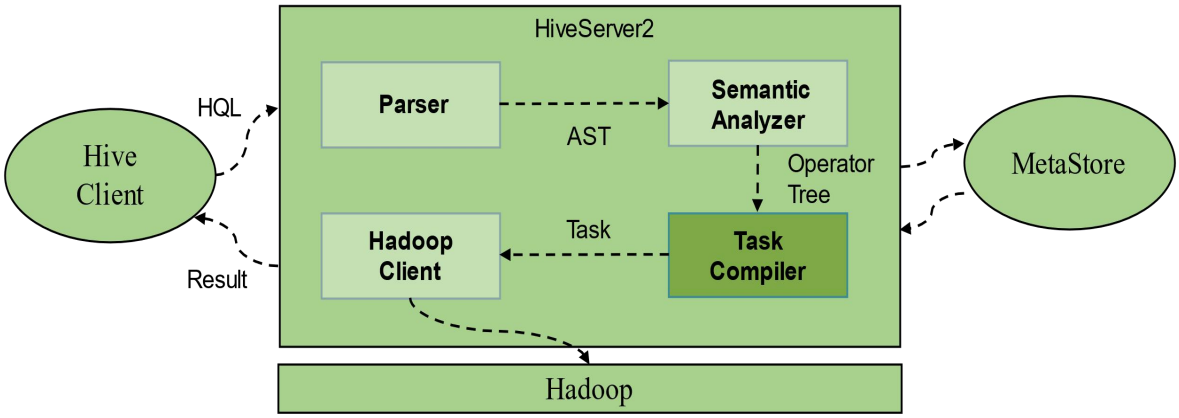

图 2.8　Hive 执行流程图

(1) 用户接口

hive 提供用户访问接口：JDBC/ODBC(JDBC 访问 Hive)、CLI(Command-line Interface)、WEBUI(通过浏览器访问 Hive)。

(2) 元数据(Metastore)

Hive 默认存储元数据是在 Derby 数据库中。通常推荐将元数据是存在 Mysql 关系型数据库中，元数据中包括 Hive 中的表名、字段、表的类型(管理表和外部表)、表的存储路径，库等。

(3) Hadoop

Hive 基于 Hadoop。处理的数据是存储在分布式文件管理系统(HDFS)。

(4) 驱动器

SQL 解析器（SQL Parser）将一条长长的、整行连在一起的 SQL 字符串转换为一个抽象文法树（AST)，并对其进行文法分析。编译器(Physical Plan)：将抽象语法树(Abstract Syntax Tree AST)编译成为一个计划且有逻辑性。优化器(Query Optimizer):优化逻辑实施计划。执行器(Execution):在逻辑上实施规划，变成可实施的实际规划。

构建精准高效的数据仓库是构建大数据平台的重要环节。随着畜牧业的信息化程度的提高，畜牧业的数据也更加复杂，这就对数据仓库的要求越来越高。数据仓库技术大部分应用在电商、物流、能源等方面，但是在草原畜牧业应用方面却极少有体现。而草原畜牧业具有场景复杂但数据对象统一的特点，因此，数据仓库的分层思想在畜牧业大数据平台上可以充分发挥出它的价值。通过整合数据，为畜牧业提供了高质量且有效的数据，对畜牧业的现代化信息化数字化的发展具有跨越性意义。数据仓库的分层设计，每个存储层都极其精简且健壮。这种设计有以下优点：

一是将数据处理的难度系数大大降低，使问题变得简单。二是使得各个数据表的存储位置具有明显的层次性，方便管理人员跟踪数据的流向。三是降低了对数据的重复利用，将中间层的数据导入到数据库中，从而降低了对数据的冗余。基于畜牧业的数据仓库分层由 ODS 原始数据层、DWD 数据明细层、DWS 数据服务层、ADS 主题层组成。数据仓库分层设计如图 2.9 所示。

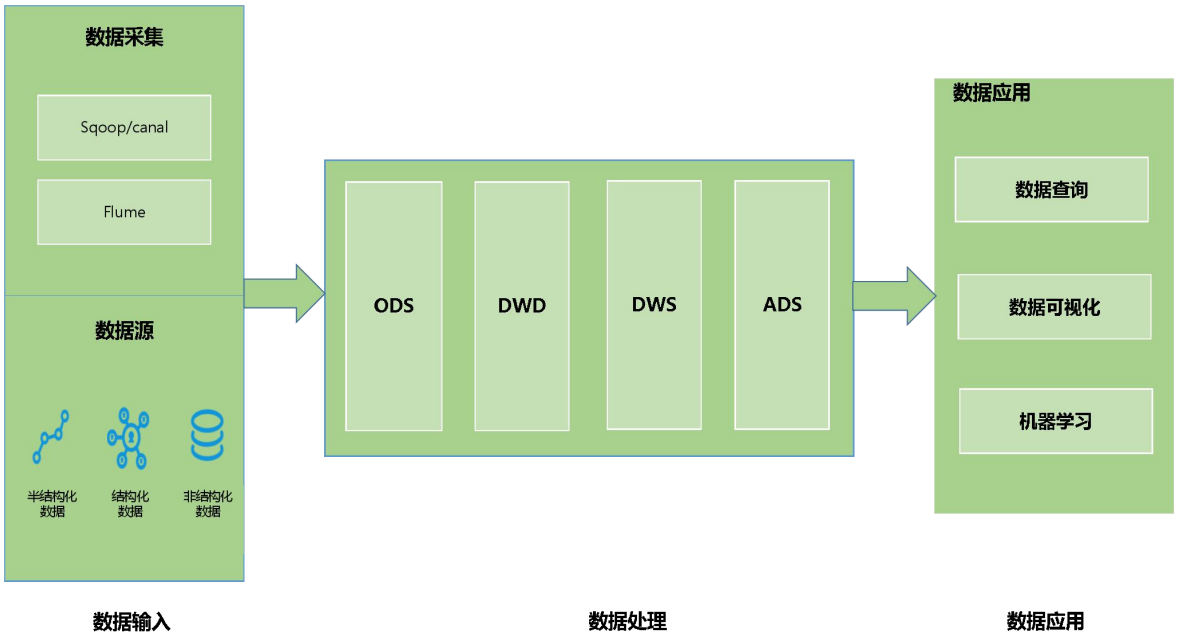

图 2.9　数据仓库分层设计

ODS (Operation Data Store)原始数据层：保持数据原貌不做任何更改，起到备份数据的作用。在建表时需要创建分区表，防止全表扫描，提高数据效率。在原始数据层需要考虑数据表的类型，如全量表、增量表和临时表。例如：类似于时间维度，地区维度，这类一般不会轻易发生变化的数据并且数据量较小的数据都会做全量表，类似于气象等回传数据记录当时状态的数据都会做增量表，偶尔业务方会用到的数据会存入临时表。

DWD (data warehouse detail) 层数据明细层：DWD 层涉及到数据建模，在大数据场景中，维度模型面向业务。这一层也进行脏数据的数据。逗号作为分隔用于获取的气象数据中，并且将其格式化，删除空数据，对于日期类型的数据的格式化处理。为了避免由于数据输入问题而对预测产生的影响，对数据中问题值和缺少值进

行处理解决目的是为了避免由于数据输入等问题导致影响预测的结果。通过观察相关实验数据，本次实验采用平均值的插补方法补充那些指标因子缺失值。针对不同数据表类型采取不同的存储策略。通过事实表和维度信息来进行分析，查询效率也更高。

因此本课题采取维度建模。

数据明细层中维度建模需要按照四种步骤：

(1) 选择业务过程

根据需求选取组织完成的一项业务操作型活动，如羊肉售卖业务。

(2) 声明粒度

在维度设计中，声明颗粒度是一个关键的步骤。粒度是数据仓库(Data Warehouse)中保存数据的细化程度。为了我们后期的分析我们选则最小粒度。

(3) 确认维度

对围绕业务过程事件的“谁、什么时候、什么、哪里、为什么和如何”等描述信息是由维度提供的，换言之，维度就是描述事实的角度。

(4) 确认事实

事实表是业务过程事件的衡量标准，具体为可以量化的值。

通过维度建模方法，可以确认草原畜牧业过程的总线矩阵子集如表 2.3 所示。

表 2.3　草原畜牧业过程的总线矩阵子集

	时间	所有者	牧民	地区	气象	牧区	资源
成羊体重	√	√		√	√		√
成羊放牧	√	√	√	√	√	√	√
成羊投产	√	√		√	√		√

通过草原畜牧业过程的总线矩阵了解到某个业务过程包含哪些维度。例如，成羊体重事实表。维度包括时间维、所有者维、地区维、气象维和其他维满足需求分析。

DWS (Data Warehouse Service)数据服务层：数据服务层采用分析主题作为建模驱动，根据应用和主题层面所需的指标，生成具有一致颗粒度的汇总表。例如：按照以天的粒度存储每日天气的变化聚合表。

ADS (Application Data Store) 主题层：ADS 层不包括建模设计，涉及表的是根据具体需求创建的。例如：后面进入到机器学习的数据所需要的所有样本数据及其特征都会加工到这里。

最后我们通过 Apache DolphinScheduler 进行全局脚本的一个调度管理任务，并且进行邮件通知服务，当任务调度中间件发生错误的时候会及时向我们发送邮件告警，当我们收到邮件告警根据邮件内容立刻处理出现的错误，以方便第二天正常的数据查询等服务。

2.3.3 可视化模块的实现

首先，将畜牧数据收集并上传至畜牧大数据平台中，然后对数据进行清洗、分层、建模，最后保留在畜牧大数据平台上。为了便于从事畜牧业的决策人员对此进行分析，将相关的数据用 Sqoop 再传送到 MySQL 中，MySQL 中的数据可以很好地将畜牧业生产的实际数据呈现出来。在苏尼特左旗，拥有比较完善的物联网设备（定位仪、图像采集装置、远程饮水控制装置等）的农户，实行信息化管理，采集养殖过程中产生的生产数据，进行挖掘和分析，为农户的决策提供强有力的支撑，实现“以数据赋能”的目标。

2.3.4 热部署接口服务

近年来各种新技术快速迭代发展。其中，接口的热部署技术逐渐被开发人员所重视。接口的热部署是指在不停止应用程序的情况下，更新应用程序的接口。接口的热部署技术具有以下几个好处及必要性。接口的热部署可以使得应用程序的更新更加快速、灵活。在传统的软件开发中，应用程序更新需要停止整个应用程序进行更新，而接口的热部署技术可以在不停止应用程序的情况下进行接口的更新，从而可以使得应用程序的更新更加快速、灵活。总之，接口的热部署技术在现代软件开发中具有非常重要的意义。它可以使得应用程序的更新更加快速、灵活，可以最大程度地减少应用程序的停机时间，也可以最大程度地减少应用程序的风险。因此，在现代软件开发中，接口的热部署技术是不可或缺的。为了保证方便接口的管理，数据接口安全，简化写接口流程。方便接口对接，减少代码开发量，为此实现接口认证，数据库管理接口平台。此服务可通过数据库配置实现简单接口的热生成，兼容手动添加接口，以及接口认证功能。并自动生成日志，每天生成，每个文件最大 30M，超过 30M 则生成新的日志文件，记录接口访问过程中的日志等信息。

本次对数据的 sql 配置表进行开发，简化开发代码量，以便更好的管理接口配置其伪代码如表 2.4 所示:

表 2.4　接口管理配置

接口管理配置

1.连接数据库

2.获取数据库 sql 集，定义 List

3.获取参数中接口名称 methodName

4.获取配置接口信息 ApiInformationEntityapiInfo

5.判断输入 id 是否为空，不为空则输入 methodName，否则为 id

6.获取配置接口名称 getMethodName

7.对比接口名称 apiName 和 methodName

8.获取配置参数 getParams

9.获取配置 getSql

10.添加 sql 集 sqlList.add(sql)

11.获取拼接 getSqlConditions

12.获取条件 conditions

13.获取判断结果 sqlConditionsMap

14.获取返回类型 returnType

15.包装成 json 根据返回类型返回参数

通过热部署接口的开发，可以实现接口的热生成，有效解决了更改底层逻辑给前端带来的影响。由于草原畜牧数据有一部分来源于物联网设备采集，如果数据接口安全得不到保障，那么可能存在数据安全问题，并且当后续接口开发任务突然增多的情况下，急需一个接口的管理平台，管理接口数据质量问题。根据以上问题设计热部署接口的开发，成为不可避免的工作，以提高大数据平台的鲁棒性。

（1）热接口配置：

数据库配置表，表名:[dws_api_information]。

表 2.5　数据库配置表说明

字段	释义	解释说明
ID	每个接口的唯一 ID	通过 uuid 生成每个接口的唯一 ID，后续通过 md5 加密生成私钥以进行接口的校验。
METHOD_NAME	接口的方法名	该名称为请求链接里的接口名

		称，会用于接口校验。
RETURN_TYPE	接口返回类型(body 中的结构)	返回类型中有[list][map][listOnlyOneContent]:[list]的升级，[listOnlyOneContent]，[mapOnlyOneContent]:[map]的升级，[mapOnlyOneContent]，[mashup]{list,map}:组合返回{}中的值不固定，可以为{list},{map},{list,map}等，若{}中的内容为 2 个甚至两个以上，则需要配合 SQL_CONDITIONS 列中 results 中 replace 的值为 add 来配合操作
PARAMS	入参定义	在此定义参数，格式为 map，其中 key 为参数名，value 为参数类型(目前支持 String, Integer)这里的参数名需要与 SQL 列中的占位符中的列名一致，还需要与 SQL_CONDITIONS 列中 conditions 中的列名一致
SQL	主要执行的 sql	若想自定义接口返回的结果集字段的格式，则可在 sql 中，select 中的字段后加 as "colume",其中双号必须加
SQL_CONDITION	sql 判断条件	增加 sql 的其他判断条件，动态的实现与前端的交互。
CREATE_DATE	创建时间	
CREATE_ID	创建者	
UPDATE_DATE	更新时间	

（2）接口请求以及加密 Key 的生成

在数据传输过程中，数据接口加密安全是非常重要的。数据接口加密可以有效地防止黑客攻击和数据泄露。一旦数据被窃取，用户的个人信息和财务信息将受到威胁。数据接口加密可以保护用户的数据，减少风险和损失。数据接口加密技术是通过将传输数据进行加密，使得黑客无法通过窃取数据来获取敏感信息。加密技术可以将数据通过 Md5 加密的形式转化成一种难以读取的形式，只有持有解密钥匙的用户才能解密。这种技术可以用于各种类型的数据传输，包括电子邮件、网站和移动应用程序。数据接口加密技术的作用不仅是保护用户的数据，还可以提高数据传输的效率和可靠性。加密技术可以防止数据被篡改或修改，保证数据的完整性和一致性。在今天的数字化世界中，数据接口加密安全已经成为了一项必要的技术。企业和个人都应该采用这种技术来保护自己的数据和隐私。以下为 key 加密过程：

//body 可视为接口参数

StringverifyString=body.get("appKey")+authKey+body.get("method")+body.get("param").toString()+timestamp;

Stringverify=DigestUtils.md5DigestAsHex(verifyString.getBytes())

以下接口参数为例，加密 key 的生成需要以下信息：

appKey: appKey 的信息，每个 appKey 对应一个唯一的接口。

authKey: 每个接口对应一个私钥，客户端与服务端各自保存一份，只是在计算加密 key 的时候使用，不互传。

methodName: 接口名称，如下例子中 method 的内容。

param: 接口参数信息，如下例子中 param 的内容。

timestamp: 时间戳，即接口请求时的时间戳，可通过 System.currentTimeMillis(); 获取到。

加密 key 的生成规则：

将 appkey,authKey,methodName,param,timestamp 转换为 String 类型，并按顺序进行相加，进行 MD5，生成的 uuid 则为加密 key。

加密 key 需要加在接口请求参数中。以下为接口请求参数示例：

{

"method":"methodname"，//接口方法名

"appKey":"36c94a4d9708422388b5827fe051f94e"，

"verify":"4f41260566d44681a4c55df9bef6d0f7"，

"timestamp":"1679452223486"，//时间戳

"param":{

"param1":"param1Val"，//请求参数

"param2":"param2Val"，

"param3":"param3Val"

}

}

2.4 平台资源规划配置

本课题通过 Apache 开源项目 Hadoop 构建畜牧业大数据基础平台，畜牧业大数据基础平台数据处理使用 MapReduce 和 Spark 计算引擎。海量数据的存储以及中间结果仍需要依赖于 HDFS。畜牧业大数据平台采用三台以 Centos 操作系统的服务器作为集群搭建环境，分别命名为 hadoop101，hadoop102，hadoop103。详细资源规划情况如表 2.6 所示。

表 2.6　集群部署规划

框架	hadoop101	hadoop102	hadoop103
HDFS	NameNode DataNode	DataNode	DataNode SecondaryNameNode
YARN	NodeManager	ResourceManager NodeManager	NodeManager
Zookeeper	Zookeeper Server	Zookeeper Server	Zookeeper Server
Hive	Hive		
Mysql	Mysql		
Sqoop	Sqoop		
Spark	Spark	Spark	Spark
Kafka	kafka	kafka	kafka
Flume	flume	flume	flume
Apache DolphinScheduler	dolphinScheduler		

2.5 本章小结

本章将畜牧区所产生的大量数据作为切入点，为了满足对这些大量数据的存储和计算处理的需要，构建了一个畜牧业大数据基础平台。以畜牧业场景对平台展开了合理的设计，选择了 Flume 与 Kafka 数据采集方式，来解决数据洪峰在平台上造成的影响。之后再对畜牧业数据进行分层和建模，以满足对不同业务过程进行分析的需要。最后开发了热部署接口平台，可以更方便快捷的管理接口平台，并对接口的安全做了校验，为生产部署提供参考和借鉴。

3 基于核 PCA_动态贝叶斯网络的价格预测

3.1 引言

价格预测的研究已被国内外学者高度重视，但这些研究大多数还停留在使用单变量模型对价格进行预测，并且使用的较为传统的数据标准化等处理方式，或者虽然考虑了诸多因素对价格的影响，但未考虑气象因子对价格波动和预测的影响。农畜产品作为自然环境下生长的产物，很容易受到自然因素的影响。羊肉价格会受到风速、降水、温度等因素影响。本项目充分考虑农畜产品价格的复杂的气象影响因素，提出优于传统 PCA 的核 PCA 降维的方法并结合动态贝叶斯网络预测价格，其预测效果拟合度预测精度都高于传统模型。

3.2 研究方法

1. 核 PCA 降维

核 PCA 降维(Kernel Principal Component Analysis) 又称为核主成分分析法，KernelPCA 是通过 PCA 来改进的，该方法将非线性可分数据转化为一种新的低维子空间用于线性分类。核主成分分析将气象因子转化为高维空间可以通过非线性映射来完成，使用主成分分析方法将气象因子从高维空间映射到其他的低维空间，并根据线性分类器来区分样本。能够以最小的信息损失对数据深层的挖掘[54]。由于核主成分分析的降维所具有的特性，使得核主成分分析的降维可被广泛应用于处理非线性问题。核主成分分析是对 PCA 算法进行非线性处理的一种改进。在此基础上，利用核主成分分析方法，对苏尼特羊肉的价格进行特征提取，并利用核主成分对原始的非线性原始气象因子进行高维线性化。即将其映射到一个高维的特征空间中，再用同一 PCA 处理。

假设输入空间中的 M 个样本 $x=(x_1,x_2,...x_M)(k=1,2,...,M)$，其中，$x_k \in R^N$ 是经数据预处理后的数据样本，引入映射 $\Phi: R^N \to F, x \to \Phi(x)$ 此时，样本点 $x_1,x_2,...x_M$ 在空间 F 变为 $\Phi(x_1),\Phi(x_2),...,\Phi(x_M)$，为了让映射后的数据样本在原点附近，假设

$$\sum_{k=1}^{M}\Phi(x_k)=0 \qquad \text{(式 3.1)}$$

相应的协方差矩阵为

$$H=\frac{1}{M}\sum_{j=1}^{M}\Phi(x_j)\Phi(x_j)^T \qquad \text{(式 3.2)}$$

那么，核主成分分析（KPCA）就是求解上述协方差矩阵 H 的特征向量λ与特征值v。根据矩阵的特征根与特征向量特性可知，$\lambda v = Hv$ 成立。因此，下式也同样成立：

$$\lambda[\Phi(x_k)\bullet v] = \Phi(x_k)\bullet Hv(k = 1,2,...,M) \quad \text{(式 3.3)}$$

上式中的特征向量v可以在$\{\Phi(x_1),\Phi(x_2),...,\Phi(x_M)\}$组成的空间内形成，那么：

$$v = \sum_{i=1}^{M}\alpha_i\Phi(x_i) \quad \text{(式 3.4)}$$

其中，α_i为常系数。

将公式 3.2，公式 3.3，公式 3.4 整理可得下式

$$\lambda\sum_{i=1}^{M}\alpha_i[\Phi(x_k)\bullet\Phi(x_i)] = \frac{1}{M}\sum_{i=1}^{M}\alpha_i[\Phi(x_k)\bullet\sum_{j=1}^{M}\Phi(x_j)]\cdot[\Phi(x_j)\bullet\Phi(x_i)] \quad \text{(式 3.5)}$$

为求上述矩阵的特征值与特征向量，需要求得 $\Phi(x_j)\cdot\Phi(x_j)(i = 1,...,M; j = 1,...,M)$ 即是样本x_i和样本x_j映射到高维空间再求内积；由于映射函数是将原样本映射到高维空间中，而对于维数可能很高，甚至可能是无穷维，因此，直接计算$\Phi(x_j)\cdot\Phi(x_j)$是非常困难的。如何能够避免这个问题，假设存在一个函数使得$K_{ij} = \Phi(x_j)\cdot\Phi(x_j)$成立，那么只要知道$K_{ij}$的函数形式，就能够避免直接计算高维甚至是无穷维的内积，存在这样的函数，使得$K_{ij} = \Phi(x_j)\cdot\Phi(x_j)$，这样的函数称为“核函数”。定义$M\times M$矩阵 $K: K_{ij} = \Phi(x_j)\cdot\Phi(x_j)$，上式可简化

$$M\lambda k\alpha = K^2\alpha \quad \text{(式 3.6)}$$

通过对上式的求解，能够得到符合要求的特征值λ和特征向量v。对于数据样本 x 在 F 空间中第 k 个特征向量v^k（其中，v^k 代表v 的第 k 个分量）方向上的投影为

$$\beta_k = v^k\cdot\Phi(x) = \sum_{i=1}^{M}\alpha_i[\Phi(x_i)\cdot\Phi(x)] \quad \text{(式 3.7)}$$

此时，β_k 是$\Phi(x)$的第 k 个主成分。最终，利用特征值$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$来计算贡献率 n。对累计贡献率设定一个阈值δ ，提取贡献率较高的前 m 个主成分，从而得到新的指标向量。

$$n=\sum_{i=1}^{m}\lambda_i \Big/ \sum_{i=1}^{M}\lambda_i \geq \delta, 0\leq\delta\leq 1 \qquad \text{(式 3.8)}$$

其中，λ_i为第i特征值，δ值为阈值，可根据问题需要取得，其δ值越小，筛选的特征越少，但也有可能产生较大的误差。

2. 动态贝叶斯网络原理

动态贝叶斯网络（Dynamic Bayesian Network,DBN）是一种以概率理论为基础和影响图为基础，将时变隐马尔可夫模型与贝叶斯网络模型融合形成的一种模型。其不仅保存了两种模型的优点，而且避免了缺点漏洞，可以随时间发生改变，也可以添加新的状态[55]。

动态贝叶斯网络（DBN）是贝叶斯网络（BN）在时间领域中的一种扩展。DBN是一个有向无环图，它是一个能反应系统随时间变化的模型。它能代表因果关系，顺序关系，或条件关系，并能由一般的常识性问题或专家知识构成[56]。贝叶斯网络（Bayesian Network）可以用条件概率表示事物间的因果关系，使贝叶斯网络在处理复杂问题方面具有很强的优势。尤其是，贝叶斯网络在建模上有很大的弹性，能够将连续和离散两个变量以不同的方法交叉运用。在实际应用中，不同的模型具有不同的优缺点，有些模型可能更适合处理某些类型的数据，而有些模型则更适合处理其他类型的数据。动态贝叶斯网络是任意过程的定向循环模型。他是由时间片组成的，每个片包括自己的变量。如图 3.1 及图 3.2 所示：

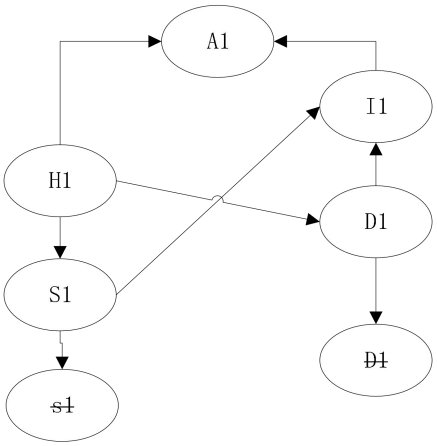

图 3.1　动态贝叶斯网络的初始状态

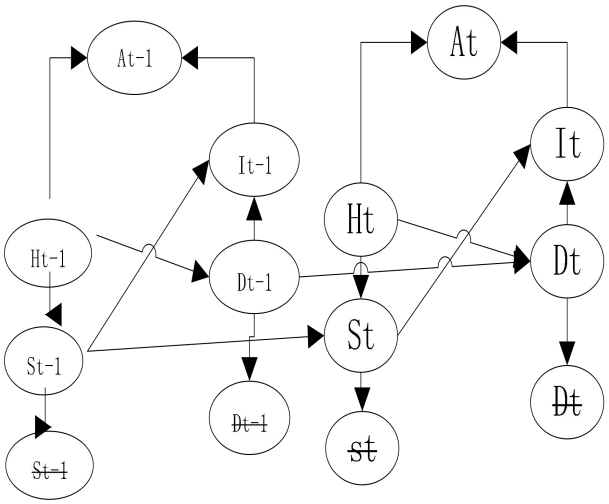

图 3.2　　动态贝叶斯网络的 2DBN

DBN 可以表示为(B1, B→)，其中 B1 可理解为先验网络，表示起始状态，而 B→则是转移网络。假设存在有一个 DBN 模型,B1 表示 x[1]上的联合概率，且 B→表示 x[1]与 x[2]之间的转移概率 P(x[t + 1]|x[t])，则 x[1],⋯,x[t]上的联合概率由以下表式 3.9：

$$P(x[1],x[2],...x[t]) = P_{B_0}(x[1])\prod_{t=1}^{t}(x[t+1]x[t]) \quad \text{(式 3.9)}$$

由此可知，动态贝叶斯网络可定义为(B1, B→)，其中 B1 表示贝叶斯网络。B→表示两个时间片的贝叶斯网络(2TBN)，它解释了转移模型 p(Zt|Zt-1)[57] 。如式 3.10：

$$P(Z_t \mid Z_{t-1}) = \prod_{i=1}^{N} P(Z_t^i \mid Pa(Z_t^i)) \quad \text{(式 3.10)}$$

其中 Z_t^i 是时间 t 的第 i 个节点，且为 Ut ,Xt, Yt 的分量。 $Pa(Z_t^i)$ 是 Z_t^i 的父项，其可以相同或者上一个时间片有关。2TBN 网络的首个时间片段中的节点无关联的参数。通过展现 2TBN 网络即可得到长度为 T 的序列的联合概率分布如式 3.11 所示：

$$P(Z_{it}) = \prod_{t=1}^{T}\prod_{i=1}^{N} P(Z_t^i \mid Pa(Z_t^i)) \quad \text{(式 3.11)}$$

3) 结合核 PCA_动态贝叶斯网络模型

通过将苏尼特羊肉价格做当前研究的对象，将从牧场气象数据中的十项影响苏尼特羊肉价格的相关气象因子进行分析研究，由于各指标之间存在一定的相关性，为避免数据冗余，本项目采用核主成分分析方法对其进行降维，以降低输入变量个

数，提高预测效率。利用核主成分分析方法对气象因子进行降维，从而有效地解决了大量的非线性映射问题，并利用动态贝叶斯网络对羊肉价格进行了预测。因此将相应的气象因子作为输入变量，将苏尼特羊肉价格作为输出变量，构建核 PCA 动态贝叶斯网络预测模型。

4）误差分析

以下为常用误差分析指标；

1.绝对误差

通常将预测的数据与真实数据的绝对值，称为绝对误差。 如式 3.12 所示:

$$e=|Y^{*}-Y| \tag{式 3.12}$$

式中 Y 代表真实苏尼特羊肉价格，Y *代表预测羊肉价格。

2.中误差

用最小二乘方法求出的实际误差，而中间误差值就是实际误差值与观测值之差的平方的平方。中误差是一种对被测群体中尺寸误差反应灵敏、能够全面反映被测对象的变化情况，因此在工程测量中得到了广泛的应用。如式 3.13 所示:

$$D=\lim_{\delta x\to 0}\frac{[\Delta]}{n} \tag{式 3.13}$$

其中△代表预测羊肉价格与实际价格误差，[]代表对预测羊肉价格与实际价格误差进行求和。△=Y-Y *（Y 代表真实苏尼特羊肉价格，Y *代表预测羊肉价格）则中误差如式 3.14 所示:

$$n=\pm\sqrt{\frac{[\Delta]}{n}} \tag{式 3.14}$$

其中 i 表示样本数量

3.极限误差

事实上，这些数值一般都是针对具体的测试条件而设定的，一般不会在实际测试中发生。因此，在偶然情况下，会产生两个错误的极限，这就是极限误差。边缘误差是区别误差与误差的一个界限，边缘误差与中间误差都是绝对误差。

5) 设计思路

其实现的基本流程如下图 3.3 所示，包括采集影响苏尼特羊肉价格因子及气象数据，建立大数据平台对采集到的数据进行预处理，核主成分降维，并且搭建动态贝叶斯网络预测模型、评估模型等环节。在构建预测模型过程中，最后在使用 MAE、MSE 等指标评估模型的预测效果。

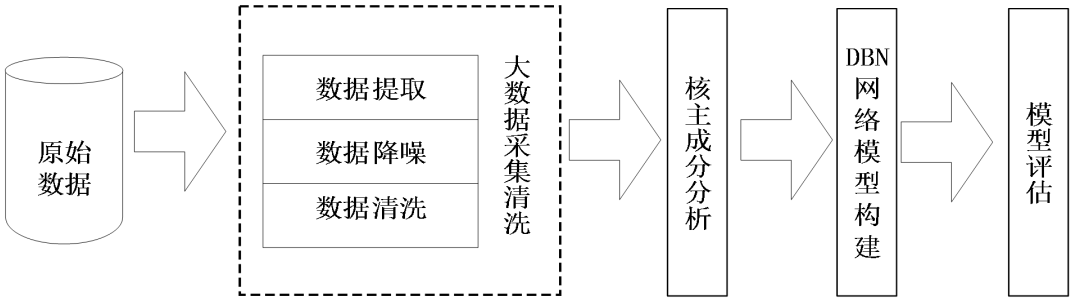

图 3.3　模型构建基本流程

3.3 实验材料及方法

3.3.1 实验材料

本次课题的试验区位置为内蒙古自治区锡林郭勒盟的苏尼特左旗某户牧民草场，夏季日照时数长，冬季短且寒冷漫长，该地主产业为畜牧业，苏尼特羊主要是在热季放养及在冷季喂精饲料。

3.3.2 数据来源

数据选取于内蒙古自治区锡林郭勒盟苏尼特左旗的智慧牧场的物联网设备观测到的气象数据及当地批发市场的苏尼特羊肉价格，其中以 2010 至 2020 年的数据作为测试集和训练集，以 2021 年的数据作验证集。由于苏尼特羊大多以放养为主。所以气象对于其体重的影响较大。

3.4 模型结果及分析

3.4.1 核 PCA 降维

苏尼特羊由于主要是放牧饲养，所以它的体重价格也会受到气象条件的影响。由于文中的影响因素太多，如果直接利用这些影响因素会增加模型的复杂度，甚至会导致得不到训练模型，模型得到的预测结果与实际值相差较大。因此，本文考虑要对影响因素进行降维，并且要能够最大程度的保留原来影响因子的信息。主成分分析能够提取原始信息中的所有变量，同时删去多余变量，并且消除变量之间的相

关性。但是，主成分分析对于线性数据处理具有很好的优势，并且很难提取非线性数据信息。为了提取影响因素中非线性信息，在主成分分析中通过引入非线性函数，使得将影响因子的数据通过非线性函数映射到更高维的空间，使得原始数据在高维空间中实现对非线性数据降维。同时运用核主成分分析提取影响羊肉价格的核影响因素，可以有效避免影响因子过多而导致模型的复杂度增加，以致得不到相应的训练模型。对于本文而言，影响因子共有 10 个，即原始空间是 10 维数据。为了更好的提取训练样本数据中非线性特征，将原数据集映射到高维空间中。但是，对于这个高维空间的维数是不确定的，可能是 11 维，但也可能会是更高维。为了解决维数灾难，采用一个核函数对源数据映射到更高维空间。因此，通过引入核函数后，原数据集可以在原来的 10 维空间中直接进行主成分分析。利用核 PCA 对影响因子进行处理，得到 10 个特征值及特征向量。为了能够提取核主成分、剔除冗余，取累计贡献率大于等于 80%，通过对气象影响因素降维，得到 5 个核影响因素，分别依次为：风速、降水量、温度、光照度、蒸发量。并且得到相应核主成分的特征值、贡献率以及特征值累计贡献率。累积贡献率结果展示，如图 3.4 所示：

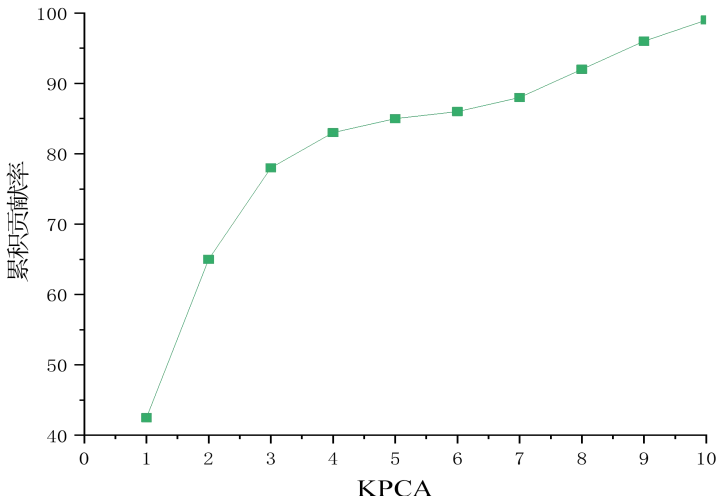

图 3.4　累计贡献率

由图 3.4 可知，1-5 个核主成分中的风速、降水量、温度、光照度、蒸发量累积贡献率在 85%以上。为了将所有的指标因子能通过数据更好的展示出来，所以选择这五个气象因子作为动态贝叶斯网络的输入参数进行预测研究。

3.4.2 核 PCA－DBN 预测结果

将苏尼特羊肉价格作为输出指标，对 2020 年及 2021 年所有月份的苏尼特羊肉价格进行预测。由于气象因子是一个不稳定的指标，在不同位置不同时间的观测值

不一定真正的反映当时的天气状态，但大体上可以反映当地的天气均值水平。所以在本次实验中为了减少由于数据维度过多产生的影响，先进行了核 PCA 降维，减少了变量数。由开始的十个气象因子减少至五个因子，最终该模型的输入参数为 5 个，输出参数为 1 个，将 2020 年及 2021 年所有月份的数据作为测试集，2010 至 2019 年所有月份的指标数据作为 2020 年的训练集，2010 至 2020 年所有月份的指标数据作为 2021 年的训练集。通过反复训练调参。最终结果显示效果最佳的状态是入参个数为 5 个，并且将该模型与传统的 PCA 降维后并结合动态贝叶斯网络模型及未经过降维的动态贝叶斯网络模型对比，还与传统的多项式回归模型及岭回归模型进行横向对比，展示效果模型预测效果如下图 3.5 所示。

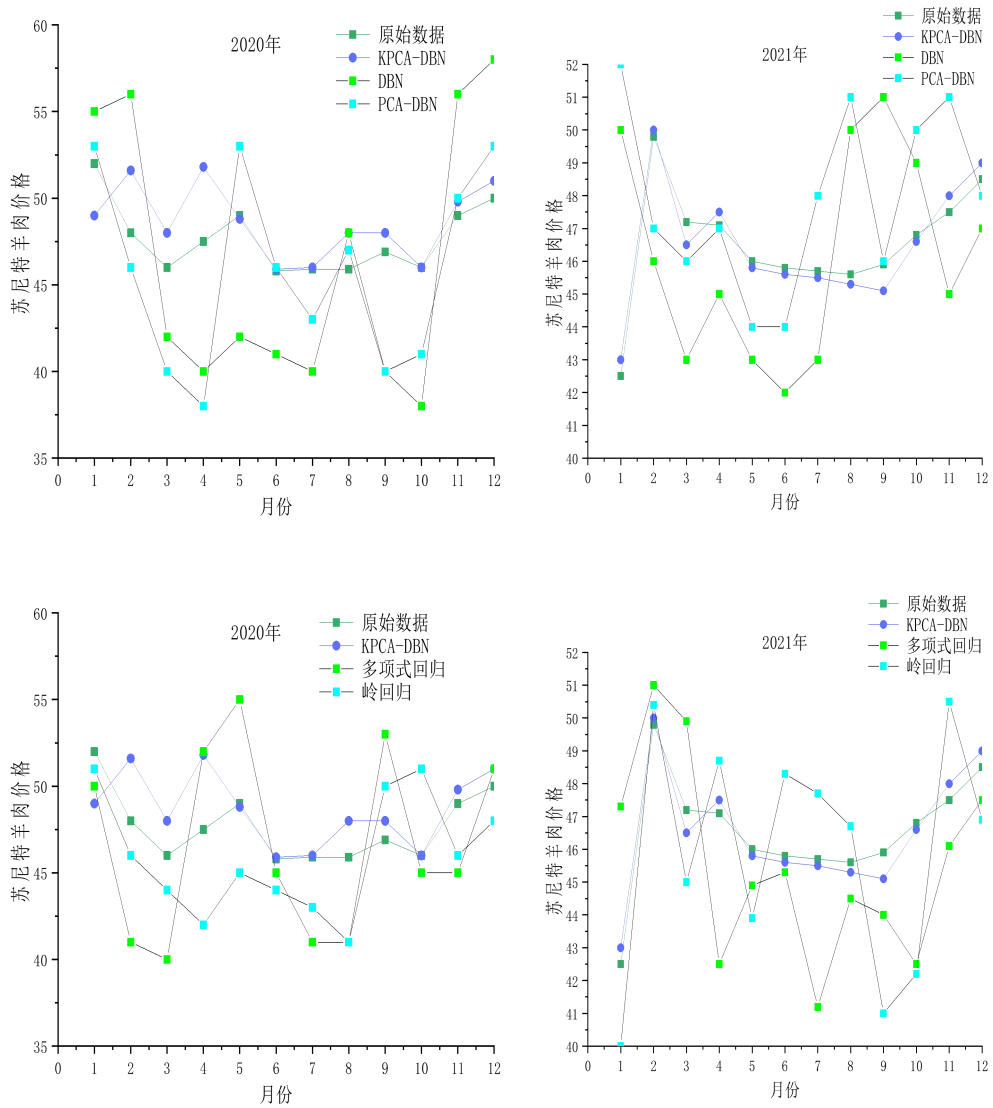

图 3.5　模型预测结果

由于动态贝叶斯网络在训练的过程中会出现略微的误差，并且输入参数选择的是影响价格的气象因子一部分数据，因此实际价格和预测得到的苏尼特羊肉价格之间会显示固定范围内的误差。通过图 3.5 也可以发现，结合 2020 年及 2021 年苏尼特羊肉价格的分析，KPCA 结合 DBN 的预测模型相较于 PCA 结合 DBN,及未进行降维的 DBN 的模型或者与传统的算法多项式回归、岭回归相比，无论是横向还是纵向对比，都可以更准确的显示出苏尼特羊肉价格在每个月内的浮动情况。

3.4.3 误差分析

在对所建模型进行性能效果评估中，最重要的是对其进行预测误差的分析。在现有的误差测量技术中，若采用不恰当的测量方法，会得到非常误导的结论。这样的结论可能会设计成一个过拟合的模型。过拟合（overfactor）是一种在已有数据集上能够很好地拟合，但是在新的数据集中却没有很好地预测效果[58]。目前常用的评估算法是通过增加数据量、增加样本数量等方式来实现的。怎样才能对模型的误差进行准确的度量，从而避免这类问题的发生。相比起预测模型，更精确的预测显得更为重要。因此，在测量误差时尤其要关注这点。避免由于数据过拟合或者欠拟合导致与实际的偏差。表 3.1 为相关误差分析结果展示。

表 3.1　误差分析表

月份	KPCA-DBN 相对误差（%）	DBN 相对误差（%）	PCA-DBN 相对误差（%）	多项式回归 相对误差（%）	岭回归 相对误差（%）
1	1.16	15.00	18.27	10.15	6.25
2	0.40	8.26	5.96	2.35	1.19
3	1.51	9.77	2.61	5.41	4.89
4	0.84	4.67	0.21	10.82	3.29
5	0.44	6.98	4.55	2.45	4.78
6	0.44	9.05	4.09	1.10	5.18
7	0.44	6.28	4.79	10.92	4.19
8	0.66	8.80	10.59	2.47	2.36
9	1.77	10.00	0.22	4.32	11.95
10	0.43	4.49	6.40	10.12	10.90

11	1.04	5.56	6.86	3.04	5.94
12	1.02	3.19	1.04	2.11	3.41

将得到的苏尼特羊肉价格的预测值与真实值进行比对验证，可发现前面的均相对误差为 0.85%，后者的均相对误差分别为 7.67%，5.47%，5.44%，5.36%，显然核 PCA-DBN 具有比较高的预测精准度。

3.5 本章小结

本章主要介绍了核 PCA 并结合动态贝叶斯网络的原理，研究对象为苏尼特羊。 通过对这些气象因子进行核 PCA 降维处理来提高动态贝叶斯网络的预测效率。 将核 PCA 降维后的气象因子作为入参，预测其羊肉价格。 同时，将其与未降维的气象因子的数据和一般的 PCA-DBN 模型及传统的多项式回归及岭回归进行比较，并将相对误差作为该模型的衡量指标。 结果表明:核 PCA 结合动态贝叶斯网络模型相比未降维的动态贝叶斯网络模型及传统模型具有较小的拟合误差和较高的预测精度。

4 基于 WOE 结合支持向量机回归模型的价格预测

4.1 引言

WOE 结合支持向量机回归模型相较于 KPCA 结合动态贝叶斯网络来说，后者需要先对数据进行核函数映射，然后才能进行分类和回归。这种方法对数据集的规模和维度都有一定的限制，同时也需要对数据分布做出一定的假设。综上所述，WOE 结合向量机回归模型是一种更优秀的预测建模方法，可以更好地适应大规模数据集，并且能够处理非线性问题，具有更高的预测准确率和鲁棒性。本项目充分考虑农畜产品价格的气象影响因素，为了提高计算效率，又提出了 WOE 法并结合支持向量机回归模型预测价格，其预测效果拟合度预测精度都高于传统模型。

4.2 研究方法

4.2.1 WOE 经验分箱

WOE(weight of evidence)直译为证据权重，是一种有监督的编码方式。它是本次研究中气象因子中重要的的特征转换方法，可以将非线性变量线性化处理，在提高解释性的前提下，还可以消除异常值对本次研究的影响，减少小概率事件对最终结果的权重[59]。WOE 的主要作用是描述预测变量气象因子与目标变量苏尼特羊肉价格之间的关系。WOE 的实质是表示当前分箱中负样本羊肉价格下降和正样本羊肉价格上涨各自占总体负样本和正样本比例的差异。其公式 4.1 如下所示

$$WOE_i = \log(\frac{Pi_1}{Pi_0}) = \log(\frac{\#B_i / \#B_T}{\#G_i / \#G_T}) \quad \text{(式 4.1)}$$

其中，WOE_i 为变量第 i 个分箱的 WOE 值；Pi_1 和 Pi_0 分别为变量第 i 个分箱中正负样本占总体正负样本的比例；$\#B_i$ 和 $\#G_i$ 分别为变量第 i 个分箱中正样本和负样本个数，$\#B_T$、$\#G_T$ 分别为总体正负样本数。

对公式 4.2 做一个简单的变换得到

$$WOE_i = \log(\frac{\#B_i / \#B_T}{\#G_i / \#G_T}) = \log(\frac{\#B_i / \#G_i}{\#B_T / \#G_T}) \quad \text{(式 4.2)}$$

观察变换后的公式，可以发现 WOE 值衡量的是当前分箱气象因子变化的比值，与总体的羊肉价格上涨跟下跌比值的差异。WOE 值的绝对值越大，意味着这种差异越大，该分组对正负样本的区分能力就越强；WOE 值越接近 0，意味着这种差

异越小，这个分组的区分能力就越差。当 WOE=0 时，即该分箱中气象因子的变化对于羊肉价格的上涨或下跌没有影响，代表这个分箱对样本的预测基本没有价值。

IV(信息价值)变量筛选，该系数的取值范围为[-1,1]。相关系数越接近 0，说明该变量的解释性越弱，越接近-1 或者 1 则解释性越强。IV 的公式定义（式 4.3）如下所示：

$$IV_i = (\frac{\#B_i}{\#B_T} - \frac{\#G_i}{\#G_T}) * \log(\frac{\#B_i / \#B_T}{\#G_i / \#G_T}) = (\frac{\#B_i}{\#B_T} - \frac{\#G_i}{\#G_T}) * WOE_i$$
$$IV = \sum_i^n IV_i \quad \text{(式 4.3)}$$

其中 IV_i 是变量第 i 个分箱的 IV 值，IV 是该变量整体的 IV 值。

4.2.2 支持向量机回归模型

1）核函数

在回归分析中，最主要的步骤就是调节参数，寻找最佳参数。 但在实际应用中，往往存在着两种情形：一种是训练样本不能被准确分类，另一种则是准确划分为两个超平面的情况，为此，本文提出了一种“核函数”。把样本映射到多维的特征空间，并在特征空间内实现样本的线性分离[60]。

如果 H 是与核函数相对应的再生内核希尔伯特空间，代表任意函数的单调递增区间，φ代表非负的损失函数，$\| h\|_H$ 表示在 H 空间中有关 h 的范数，一般如式 4.4 所示：

$$\min_{h \in H} F(h) = \delta(\| h \|_H) + \varphi(h(x_1), h(x_2), ..., h(x_m)) \quad \text{(式 4.4)}$$

目前常用的核函数有以下几种，线性核函数，如式 4.5 所示：

$$K(x_i, x_j) = x_i^T x_j \quad \text{(式 4.5)}$$

多项式核函数，对于任意整数 d，定义多项式核函数如式 4.6 所示:

$$K(x_i, x_j) = (x_i^T x_j + c)^d \quad \text{(式 4.6)}$$

$c \geq 0, d \geq 1$，且表示多项式次数。

Sigmoid 核函数，定义 Sigmoid 核函数如式 4.7 所示：

$$K(x_i, x_j) = \tanh(\beta x_i^T x_j + \theta)^d \quad \text{(式 4.7)}$$

其中 tanh 代表双曲正切函数，$\beta > 0, \theta > 0$。

rbf 核函数,定义 rbf 核函数如式 4.8 所示：

$$K(x_i, x_j) = \exp(-\| x_i^T - x_j \|^2 / 2\sigma^2)^d \quad \text{(式 4.8)}$$

其中σ 表示为核函数的带宽。

有许多常见的内核功能，而 RBF 内核功能是最实用的。对于较大的尺寸，较小的尺寸，大样本，小样本等问题都适用，本文选用 RBF 核函数，其中，RBF 核函数的宽度作为参数。

核函数的计算：假设原始样本内积为<x,z>，经过映射，样本内积为$<\varphi(x), \varphi(z)>$，则此时的核函数如式 4.9 所示：

$$K(x, z) = \varphi(x)^T \varphi(z) \quad \text{(式 4.9)}$$

将其公式展开，得到以如式 4.10 所示：

$$\begin{aligned} K(x, z) &= (x^T z)^2 \\ &= (\sum_{i=1}^{M} x_j x_i)(\sum_{j=1}^{M} x_j z_j) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i x_j)(z_i z_j) \end{aligned} \quad \text{(式 4.10)}$$

通常将在空间中的中心 x_0 与其任意点 x 直接的距离定义为欧式函数。

2）支持向量机回归

支持向量机（Support Vector Machine, SVM）是一种用于对数据进行分析处理的监督式机器学习模型。对于一组数据为线性可分的情况时，线性支持向量机可寻找超平面，来完成数据的处理与分析运算。因此，本文从气象因子中提取特征建立苏尼特羊肉预测模型是一项非线性问题。根据以上的论述，采用 SVR 算法，利用非线性 SVM 来建立一个对苏尼特羊的价格进行预测的模型是可行的[61]。在本文中，由气象因子和苏尼特羊肉价格组成的训练数据集为如式 4.11 所示：

$$data = \{(x_1, y_1), (x_2, y_2), ..., (x_i, y_i), ..., (x_u, y_u)\} \quad \text{(式 4.11)}$$

其中 x_1 表示气象因子，即从气象因子中提取多维特征向量，y 表示苏尼特羊肉的价格值，i =1,2, ,M 表示苏尼特羊肉价格的样本数量。则预测苏尼特羊肉价格的 SVR 算法公式可以表示如式 4.12：

$$f(x) = w^T X_n + b \tag{式 4.12}$$

其中 w 表示权向量，b 表示截距。

但是，针对样本数据，常规回归模型往往根据输出结果与实际苏尼特羊羊肉价格之差计算模型损失，仅在输出结果与实际结果一致时才将记 0。而支持向量机算法有一个偏离容忍范围，也就是一个阈值，如公式 4.13 中所表示。只有在与 y 之差的绝对值大于此阈值时，模型损失才会被计算。由于气象因素的采集很容易受环境及其自身的细微差异的影响，所以这些影响都会导致与 y 的不同。因此这一现象进一步证实了利用支持向量回归方法建立预测模型的正确性。本文 SVR 算法使用不敏感损失函数，公式如式 4.13：

$$l_{\in}(z)=\begin{cases} 0, & |z|\le\in \\ |z|-\in, & |z|>\in \end{cases} \tag{式 4.13}$$

当使用该损失函数后，得到公式 4.14：

$$\min_{w,b}\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{M}\lg(z) = \min_{w,b}\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{M}\lg(f(x_n,i)) - y_i \tag{式 4.14}$$

其中 z 值表示 $f(x)$ 与 y 之间的差值，即，C 表示惩罚因子。然后，引入松弛变量 $\in$ 和 ∂ 后，公式 4.13 可重写如式 4.15：

$$\min_{w,b}\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{M}(\in_i + \partial_i) = \min_{w,b}\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{M}\lg(f(x_n,i)) - y_i$$

$$A_{ij}=\begin{cases} f(x_n,i)) - y_i = (w^T x_n + b) - y_i \le \in + \partial_i \\ y_i - f(x_n,i)) \le \in + \partial_i \\ \in_i \ge 0, \partial_i \ge 0, i = 1,2,...,M \end{cases} \tag{式 4.15}$$

拉格朗日乘子 α 、 α 、 β 和 β 由拉格朗日乘子法得到的拉格朗日函数如式中 4.16：

$$\begin{aligned} &L(w,b,\in,\partial,\alpha,\alpha,\beta,\beta) \\ &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{M}(\in_i + \partial_i) - \sum_{i=1}^{M}\alpha_i \in_i + \sum_{i=1}^{M}\beta_i(f(x_{n,i}) - y_i - \in -\partial_i) \\ &+ \sum_{i=1}^{M}\beta_i(y_i - f(x_{n,i}) - \in -\partial_i) \end{aligned} \tag{式 4.16}$$

把公式(4.15)代入到公式(4.16)中，再将 L 对 w、b 、∈和∂求偏导，并把每个偏导数设为 0，计算得到式 4.17：

$$\begin{aligned} w &= \sum_{i=1}^{M}\beta_i x_{n,i} - \sum_{i=1}^{M}\alpha_i x_{n,i} = \sum_{i=1}^{M}(\beta_i - \alpha_i)x_{n,i} \\ \sum_{i=1}^{M}\beta_i &- \sum_{i=1}^{M}\alpha_i = \sum_{i=1}^{M}(\beta_i - \alpha_i) = 0 \\ c &= \beta_i - \alpha_i \end{aligned} \quad \text{(式 4.17)}$$

最后把核函数 $k(x_i, x_j)$ 带入公式(4.17)中，并把该式的值设为 0，然后得到预测苏尼特羊肉价格的 SVR 算法的如公式 4.18：

$$k(x_i, x_j) = \phi(x_i)^T \phi(x_j) \quad \text{(式 4.18）}$$

4.2.3 模型评价指标

模型的优劣无法用肉眼直接观测，必须用一些定性的指标对其进行评估。预测只能是对某一阶段或某一时刻的数值的估计。 在实际的实验中，实际数据与实际数据之间会存在一定的差距。 误差越大，模型的准确性越低。本次预测模型评价指标有以下几个：（其中 n 为本次研究中的样本总个数， y_i 为苏尼特羊肉价格最终真实售卖价格， y_j 为此次研究的预测价格）

(1) ME。这个数值是指在苏尼特羊的价格的实际值与模型预测值之间的误差之和在总体抽样中的平均，也就是预测的价格与观测值之差的平均值，该值越小代表模型效果越好，如式 4.19 所示：

$$ME = \frac{1}{M}\sum_{i=1}^{M}(y_i - y_j) \quad \text{(式 4.19)}$$

(2) MAE 。MAE 值指的苏尼特羊的价格的实际值和模型预测值之间误差的绝对值和占总样本数的平均值，该值越小表示该方法的实验效果越好，其值越大说明在预测过程中产生的误差越大；数值越低，表示误差越小，拟合的结果也就越好。其表达式如式 4.20 所示:

$$MAE = \frac{1}{M}\sum_{i=1}^{M}|y_i - y_j| \quad \text{(式 4.20)}$$

(3) 均方误差（MSE），该值指的苏尼特羊的价格的实际值和模型预测值之间误差的平方和占总样本数的平均值，其表达式如式 4.21 所示：

$$MSE = \frac{1}{M}\sum_{i=1}^{M}(y_i - y_j)^2 \quad \text{(式 4.21)}$$

(4) 均方根误差（RMSE）。该值指苏尼特羊的价格的实际值和模型预测值之间误差的平方和占总样本数的平均值的平方根，该值越小表示该方法的实验效果越好，其表达式如式 4.22 所示：

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i - y_j)^2} \quad \text{(式 4.22)}$$

(5) STD。该值统计模型预测值在一段时间内和其实际值的误差上下波动的幅度，其表达式如式 4.23 所示：

$$STD = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_j - \overline{y_i})^2} \quad \text{(式 4.23)}$$

(5) MAPE。平均绝对百分比误差（MAPE），当 MAPE <10 时，则呈现了较高的预测精度，其表达式如式 4.24 所示：

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_j}{y_i}\right| \times 100\% \quad \text{(式 4.24)}$$

(6) SSE。表示模型预测值和实际值对应点的误差的平方和，其表达式如式 4.25 所示：

$$SSE = \sum_{i=1}^{n} W_j (y_i - y_j)^2 \quad \text{(式 4.25)}$$

(7) R_2。R_2 用于回归模型中总方差与因变量方差的比例。R_2 很大，表示在因变量与自变量之间存在线性关系，如果回归模型拟合效果完美，则 SSE 等于零，则 R_2 为 1。R_2 小，表示因变量与自变量之间线性关系不明显。如果回归模型完全失败，SSE 等于 SST，没有方差可被回归解释，则 R_2 为零。其表达式如式 4.26 所示：

$$R_2 = 1 - \frac{\sum_i (y_i - y_j)^2}{\sum_i (\overline{y_i} - y_j)^2} \quad \text{(式 4.26)}$$

4.3 模型建立

4.3.1 实验数据

试验数据选取内蒙古自治区锡林郭勒盟苏尼特左旗的气象数据及苏尼特羊肉价格，其中以 2010 至 2020 年的数据作为测试集和训练集，以 2021 年的数据作验证集。由于苏尼特羊大多以放养为主，所以气象对于其体重的影响较大。

4.3.2 模型建立

本章节所采用的的预测模型是支持向量回归模型，并且对影响苏尼特羊肉价格的气象因子进行整理并做 WOE 经验分箱处理，根据 IV 值得到 8 个预测能力强的气象影响因子。

本次数据标准化处理采用无量纲化处理方法。如式 4.27 所示：

$$y_i = \frac{x_i - \bar{x}}{s} \tag{式 4.27}$$

这里的 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)^2}$ ，

IV 值的处理公式，如式 4.28 所示:

$$IV = \sum_{i=1}^{i}(\frac{y_i}{y} - \frac{n_i}{n})WOE_i = \sum_{i=1}^{i}(\frac{y_i}{y} - \frac{n_i}{n})In(\frac{y_i / y}{n_i / n}) \tag{式 4.28}$$

将标准化后的数据进行 37 比例划分，一部分为测试集一部分为训练集；

再采用支持向量机回归对苏尼特羊肉价格进行建模预测。

最后将预测的结果用 MAE、RMSE、MSE、MAPE、SMAPE 等方法进行模型评估。

4.3.3 结果分析

4.3.3.1 WOE 经验分箱及特征选择

我们对采集到的十个影响气象因子进行经验分箱，以气温为示例，介绍各个变量的分箱及正负样本数和所对应的 WOE 的值和 IV 值。计算结果如表 4.1 所示。

表 4.1　气温分档及其 WOE 值与 IV 值的计算

气温分档	正样本数量	负样本数量	WOE_i	IV_i
$(-28,-22]$	A_1	B_1	$In\frac{A_1 / A_T}{B_1 / B_T}$	$(\frac{A_1}{A_T} - \frac{B_1}{B_T})WOE_1$
$(-22,-16]$	A_2	B_2	$In\frac{A_2 / A_T}{B_2 / B_T}$	$(\frac{A_2}{A_T} - \frac{B_2}{B_T})WOE_1$
$(-16,-10]$	A_3	B_3	$In\frac{A_3 / A_T}{B_3 / B_T}$	$(\frac{A_3}{A_T} - \frac{B_3}{B_T})WOE_1$
$(-10,-4]$	A_4	B_4	$In\frac{A_4 / A_T}{B_4 / B_T}$	$(\frac{A_4}{A_T} - \frac{B_4}{B_T})WOE_1$

$(-4,2]$	A_5	B_5	$In\frac{A_5/A_T}{B_5/B_T}$	$(\frac{A_5}{A_T}-\frac{B_5}{B_T})WOE_1$
$(2,8]$	A_6	B_6	$In\frac{A_6/A_T}{B_6/B_T}$	$(\frac{A_6}{A_T}-\frac{B_6}{B_T})WOE_1$
$(8,14]$	A_7	B_7	$In\frac{A_7/A_T}{B_7/B_T}$	$(\frac{A_7}{A_T}-\frac{B_7}{B_T})WOE_1$
$(14,20]$	A_8	B_8	$In\frac{A_1/A_T}{B_1/B_T}$	$(\frac{A_8}{A_T}-\frac{B_8}{B_T})WOE_1$
$(20,26]$	A_9	B_9	$In\frac{A_8/A_T}{B_8/B_T}$	$(\frac{A_9}{A_T}-\frac{B_9}{B_T})WOE_1$
$(26,32]$	A_{10}	B_{10}	$In\frac{A_9/A_T}{B_9/B_T}$	$(\frac{A_{10}}{A_T}-\frac{B_{10}}{B_T})WOE_1$
汇总	A_T	B_T	$\sum WOE_i$	$\sum IV_i$

本论文采用的是经验分箱方法，因为每个变量都与现实情况密切相关，如果采用卡方分箱，等频分箱，等距分箱等方法，其分箱结果会与现实情况相差甚远。其气温分档及其 WOE 值结果如表 4.2 所示。

表 4.2　气温分档与其 WOE 值

气温分档	WOE 值
$(-28,-22]$	-1.6832
$(-22,-16]$	-1.2356
$(-16,-10]$	-0.4421
$(-10,-4]$	-0.1189
$(-4,2]$	0.0587
$(2,8]$	0.5834
$(8,14]$	0.6399
$(14,20]$	0.8153
$(20,26]$	0.6597
$(26,32]$	0.4176

由表 4.2 可以看出，在(-28,20]范围内，气温的分档位越大，其所对应的 WOE 值越大，而当气温档位在(14,20]之后，WOE 值随着气温的升高又开始减小，但其减

小的幅度并不大。说明在分档中是处于一个正常的趋势，并且是呈正相关趋势，其余气象因子的 WOE 如图 4.1 所示

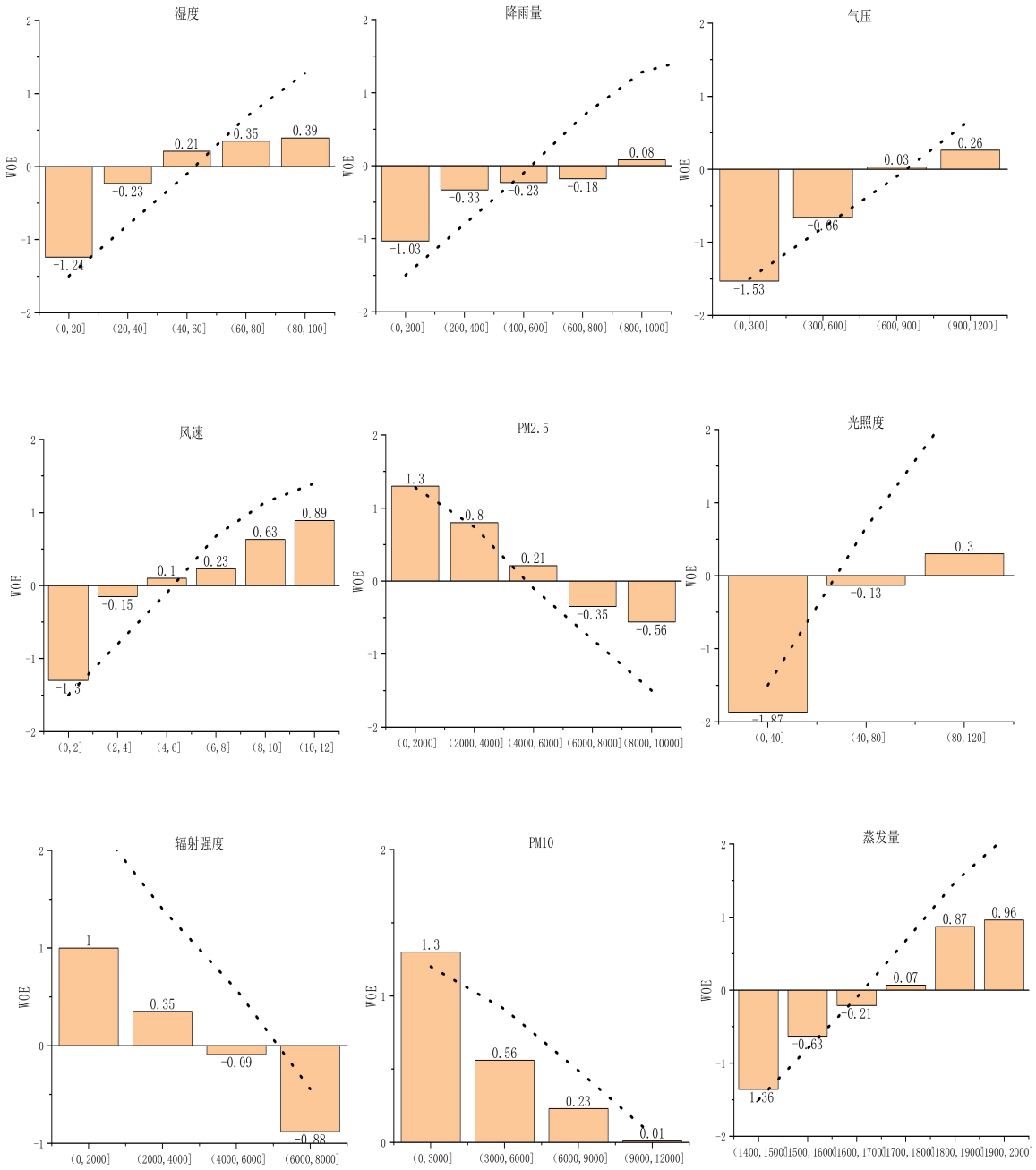

图 4.1　模型预测结果

在图 4.1 中，黑色的虚线表示该气象因子的 WOE 值，随着分类区间有着不同的趋势，而我们所选择的特征变量都是与正样本呈正相关的，例如，温度高，降水量高，为正样本的概率也是高的，其 WOE 值也应该是递增的且趋于稳定的。所以，我们将三个因素排除在外，即 PM10、 PM2.5、辐射强度。通过对各个特征变量的 WOE 值，计算相应的 IV，结果如图 4.2 所示。

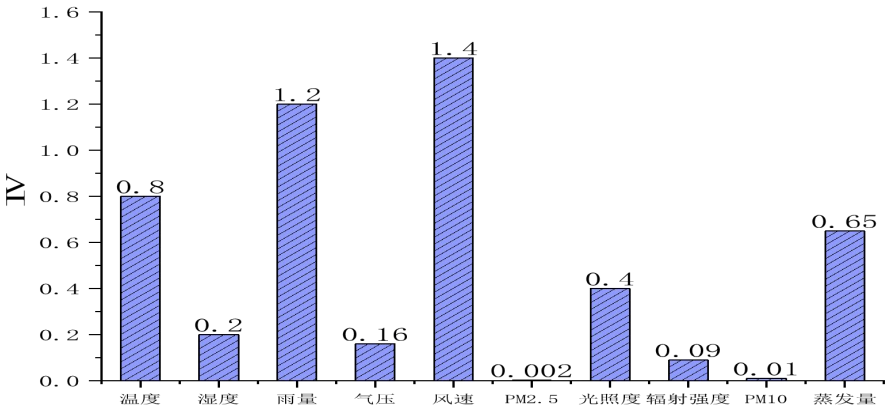

图 4.2　模型预测结果

我们先降气象因子的 IV 值进行排序筛选出预测能力强的变量，因为 IV 值越大也就意味着该气象因子越能体现所有样本的特征。IV 值在 0.02-0.1 之间为变量预测力弱的变量，在 0.1-0.3 之间为变量预测力中的变量，大于 0.3 的为预测力强的变量。由图 4.2 可以看出，PM10、PM2.5、辐射强度这三个气象因子与其他的气象因子相比，其 IV 值都小于 0.1，预测能力都较弱。所以选用预测能力强的变量温度、湿度、雨量、气压、风速、光照度、蒸发量。

4.3.3.2 支持向量机回归模型结果分析

本文选取 2010 至 2021 年的气象数据对苏尼特羊肉价格进行预测，将 2010 至 2019 年所有月份的指标数据作为 2020 年的训练集，2010 至 2020 年所有月份的指标数据作为 2021 年的训练集，分别将 WOE 结合支持向量机回归模型与传统的支持向量机模型及多项式回归、岭回归模型及原始数据作对比其预测结果如图 4.3 所示。

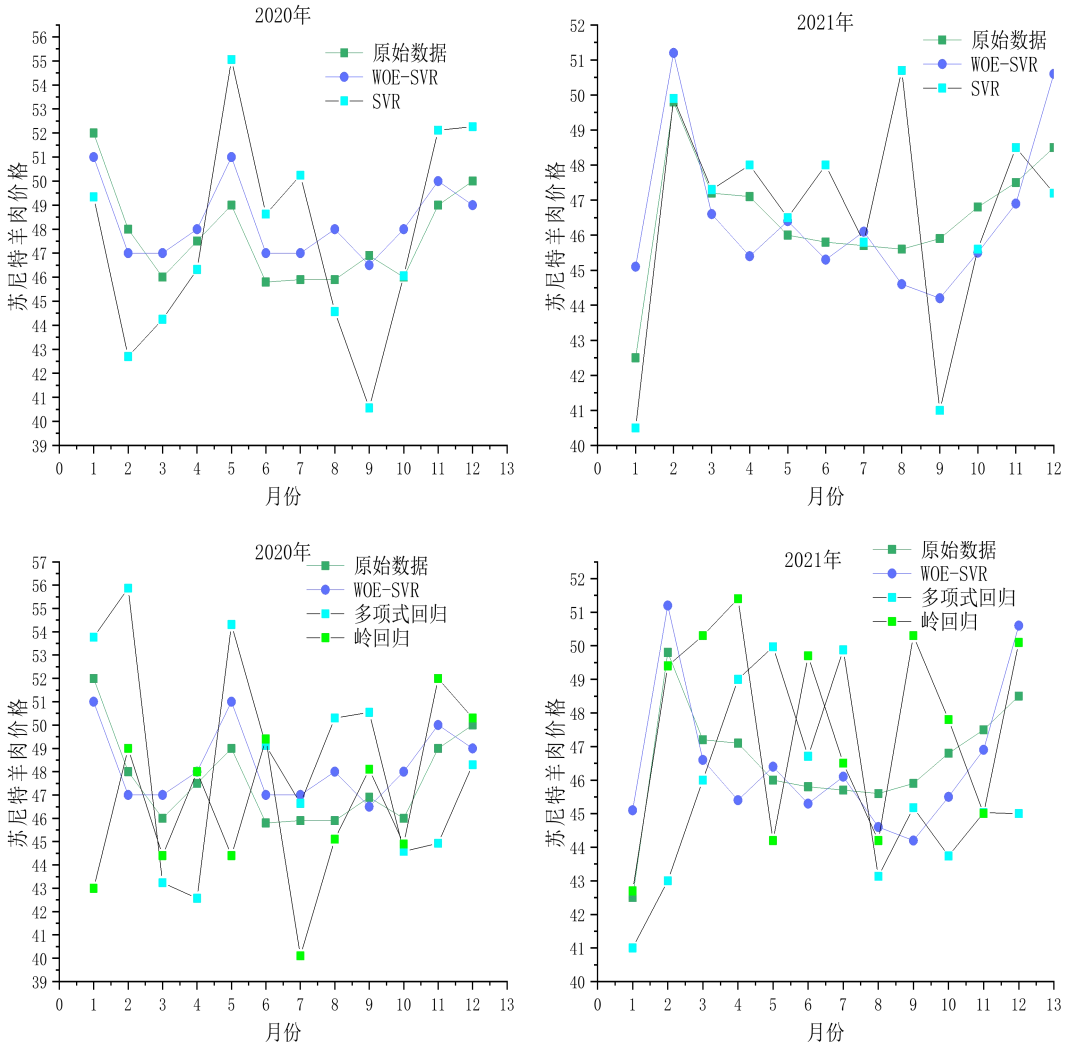

图 4.3　支持向量机回归预测结果图

通过对苏尼特羊肉价格进行建模预测，得出预测价格与真实价格的拟合曲线，从图中便可以容易看出 2020 年、2021 年中 WOE 结合支持向量机回归的模型相较于传统机器学习算法 SVR 模型、多项式回归、岭回归来比具有较好的拟合效果，其中误差相对很小，且拟合的趋势基本表现一致，所以该模型是可以预测实际价格的走势。

4.3.4 模型评估

将未分箱的 SVR 模型及传统的多项式回归和岭回归模型与 WOE 分箱的 SVR 模型数据进行对比，分别对这些模型使用平均绝对误差、均方根误差、均方误差、平均绝对百分比误差、对称平均绝对百分比误差进行模型评估，其结果如表 4.3 所示：

表 4.3　模型评估指标

模型评估	多项式回归	岭回归	SVR	WOE-SVR
MAE	21.25	19.11	18.13	16.35
RMSE	6.47	4.97	5.22	4.51
MSE	5.61	5.78	4.15	3.75
MAPE	14.77	16.25	16.75	14.27
SMAPE	21.53	16.83	18.25	16.22

指标值越小，代表模型误差越小，精度越高。由表 4.3 中的各项指标数据可以得出，经过 WOE 分箱后的模型各项误差指标都比未分箱的及传统多项式回归模型、岭回归模型指标更低，所以表明 WOE-SVR 拟合的效果更好，且均在可接受范围内，可以进行实际应用。

4.4 本章小结

本章采用 WOE 经验分箱法并结合支出向量机回归方法对苏尼特羊肉价格进行预测。得到的预测值与真实值之间存在一定的差距，但是误差都在可接受的范围之内，在整个后续的研究过程中，如果样本数量更多，或者气象因子等其他影响羊肉价格的因素更全面的话，拟合效果会比之前更好。在整个支持向量机回归建模的过程中，该模型通过了对比了与未分箱的支持向量机回归的模型及多项式回归、岭回归模型，显然分箱后的支持向量机回归模型具有更好的预测效果。

5 总结与展望

5.1 总结

畜牧放牧数据规模爆炸式增长，利用信息技术可以对牛羊的放牧售卖实现科学管理，为牧民提供辅助决策，为消费者带来合适的价格。本次课题通过选取研究对象为苏尼特羊进行讨论分析，并结合了影响苏尼特羊肉价格的主要气象因素，对苏尼特羊肉的价格采用核 PCA_动态贝叶斯网络模型及采用 WOE 结合支持向量机回归模型进行预测，下面所展示的为本次研究内容：

1）首先对大数据基础平台的搭建以及对价格预测文献的海量调查研究，首要对苏尼特羊肉价格所产生的气象因子进行收集并预处理，脏数据的清洗以及格式的处理。

2）在核 PCA_动态贝叶斯网络模型中，先采用核 PCA 对影响价格的气象因素进行数据降维处理，排除掉那些具备较差相关性的，得到主要的数据因子，再进入到动态贝叶斯网络模型。

3）在 WOE 结合支持向量机回归机模型，先采用 WOE 经验分箱法得到预测能力较强的因子，在进入支持向量机回归模型。

最后对模型预测结果进行误差分析评估模型的优劣。

5.2 展望

在此次的研究讨论中，由于研究的深度和时间并不是相当充分，后期仍有改进完善的空间。

1）通过结合产生影响的相关气象因子开展价格预测的讨论研究过程中，由于导致研究对象苏尼特羊价格的相关气象因素不止一个，同时也存在另外无法控制的因素影响价格因子结合，而在建模的过程中所选择的气象影响因子仅可以描述苏尼特羊肉价格的部分特性。因此预测模型的相关设计理念还具备更新达到更加完美的模型，更需要在此领域上进行下一步的精细研究和讨论。

2）本课题在分析苏尼特羊肉价格预测的过程中选取的是以往比较传统的预测方式进行讨论分析。即便预测的精准度上确实是有很大的提升，但是在之后的研究分析中，还是更希望可以使用到一些更加先进的，科学的算法。

参考文献

[1] 肖普辉. 现代畜牧业发展趋势与支撑体系建设[J].中国畜牧兽医文摘,2013,29(06):5-7.

[2] 魏秀娟. 数据时代畜牧业信息化建设刍议[J]. 中国畜牧杂志, 2014,50(10): 38-41.

[3] 朱亦斌,马亮.数据挖掘关联规则算法探讨[J].管理信息系统,2000(03):58-61+7.

[4] 杨昱梅,李继娜.基于 AHP 和 BP 神经网络的高校毕业生就业质量评价研究[J].中国教育学刊,2015(S1):148-149.

[5] 郭志荣,高峰,王其林.基于 RBF 网络和 MRAS 的鱼雷永磁同步电机无速度传感器控制方法[J].水下无人系统学报,2017,25(06):448-452.

[6] 孟猛,代昆豪,王杏,路靖.国内农产品价格形成机制研究综述[J].热带农业科学,2022,42(07):116-123.

[7] 靳占新,徐中一.大数据背景下物资价格预测方法[J].中国电力企业管理,2018(36):44-45.

[8] 王鹏,张利.大数据处理系统的研究进展与展望[J].高技术通讯,2015,25(Z1):793-801.

[9] 张喜红,王玉香.基于贝叶斯网络医学检验仪器故障诊断模型构建方法的研究[J].曲靖师范学院学报,2021,40(03):61-66.

[10] 张海妮.基于改进 RBF 网络的加工中心主轴热误差建模研究[J].自动化技术与应用,2019,38(01):60-64+74.

[11] 王国胤,刘群,夏英,胡军,马彬,纪良浩.大数据与智能化领域新工科创新人才培养模式探索[J].中国大学教学,2019(04):28-33.

[12] 闵昌兆.大数据时代人工智能在计算机网络技术中的应用分析[J].计算机产品与流通,2019(01):43.

[13] Drew, A, Linzer. Dynamic Bayesian forecasting of presidential elections in the states[J]. Journal of the American Statistical Association, 2013, 108(501): 124-134.

[14] Heath D,Swindle G. Introduction to Mathematical Finance[J]. Journal of the American Statistical Association, 2011, 57(466):563-563.

[15] Wang H , Bai C, Wei Q , et al. Inventory and pricing decisions when dealing with strategic consumers: A comprehensive analysis[J]. Computers & Operations Research, 2021, 136(5):105473.

[16] Avci E, Bunn D, Ketter W, et al. Agent-level determinants of price expectation formation in online double-sided auctions[J]. Decision Support Systems, 2019, 124.

[17] Towbin P, Weber S. Price Expectations and the U.S. Housing Boom[J]. Social Science Electronic Publishing, 2016, 15(182):1.

[18] 任永富.基于轻型注意力卷积网络的无人机遥感影像多目标检测[J].测绘技术装备,2022,24(04):89-94.

[19] Minsik, Lee, Hong, et al. Stock Price Prediction by Utilizing Category Neutral Terms : Text Mining Approach[J]. Journal of Intelligence and Information Systems, 2017, 23(2):123-138.

[20] 陈婷. NARX 动态神经网络的择时策略研究[D].上海师范大学,2019.

[21] Group E . Goldman Sachs cuts Brent price forecast[J]. Energy Argus petroleum coke, 2022(32):22.

[22] 傅如南,林丕源,严尚维,孙爱东.基于 ARIMA 的肉鸡价格预测建模与应用[J].中国畜牧杂志,2008(20):17-21.

[23] 马孝斌,王婷,董霞,王楚端.向量自回归法在生猪价格预测中的应用[J].中国畜牧杂志,2007(23):4-6.

[24] Fan W , Bifet A . Mining big data: Current status, and forecast to the future[J]. ACM SIGKDD Explorations Newsletter, 2014, 16(1):1-5.

[25] 罗超平, 翟琼, 李靖文. 基于时间序列数据的蔬菜价格波动特征及影响因子分析[J].西南大学学报(自然科学版),2013,35(04):26-31.

[26] Kawaguchi T, Matsunaga A, Watanabe A, et al. Prediction of changes in functional ability of inpatients with schizophrenia using logarithmic and linear regression modelling[J]. Hong Kong Journal of Occupational Therapy, 2018, 31(2): 76-85.

[27] Alessa A, Faezipour M. Preliminary flu outbreak prediction using twitter posts classification and linear regression with historical centers for disease control and prevention reports: Prediction framework study[J]. JMIR public health and surveillance, 2019, 5(2).

[28] Pacelli V, Bevilacqua V, Azzollini M . An Artificial Neural Network Model to Forecast Exchange Rates[J]. Journal of Intelligent Learning Systems & Applications, 2011, 3(2):57-69.

[29] Cui M, Zhu D, Guo L, et al. Usefulness of lumen area parameters determined by

intravascular ultrasound to predict functional significance of intermediate coronary artery stenosis[J]. Chinese medical journal, 2013, 126(09): 1606-1611.

[30] Gana R, Vasudevan S. Ridge regression estimated linear probability model predictions of O-glycosylation in proteins with structural and sequence data[J]. BMC Molecular and Cell Biology, 2019, 20(1): 1-31.

[31] Rajakumari K E , Kalyan M S , Bhaskar M V . Forward Forecast of Stock Price Using LSTM Machine Learning Algorithm[J]. International Journal of Computer Theory and Engineering, 2020, 12(3):74-79.

[32] 李文君.基于 GAM 模型的中国螺纹钢期货价格预测研究[J].中国证券期货,2021(03):16-23.

[33] 陈镜如.基于干预项修正 ARIMA 模型的煤炭价格预测研究[J].物流工程与管理,2021,43(09):132-135.

[34] 张春华,高铁梅,陈飞.经济时间序列频率转换方法的研究与应用[J].统计研究,2017,34(02):92-100.

[35] 金恒,过文俊.基于数据挖掘的异常财务数据识别方法研究[J].电子设计工程,2021,29(21):43-46+52.

[36] 李哲敏,李干琼.禽蛋市场价格短期预测[J].中国食物与营养,2010(06):36-40.

[37] 李琪阳,董雷.基于朴素贝叶斯的物联网设备指纹算法[J].电子设计工程,2021,29(21):155-158.

[38] 李咏豪,李伦波.朴素贝叶斯与 Softmax 回归在文本分类上的对比研究[J].电脑知识与技术,2021,17(28):131-132+137.

[39] 杜广宏,佘钰蔚,梁强,刘峰,张盟勃.致密薄储层综合预测技术在庆阳气田深层中的应用[J].石油地球物理勘探,2022,57(S2):141-146+231.

[40] 吴玮,郑子炜,卫栎,李琦,孙祎泽.基于组合模型的电价预测方法及应用[J].能源与节能,2022(12):10-14+24.

[41] 李冰箫,张世伟,郑舒宇,赵志帆.基于多元线性回归与 ARIMA 组合模型的水电功率预测研究[J].科学技术创新,2022(33):71-74.

[42] 胡凌云,杨威,王恒娜.一种新型高精度的区间数组合预测方法[J].统计与决策,2021,37(24):5-10.

[43] 崔云浩,朱军,孟浩.神经网络在农产品销量预测中的应用[J].现代农业科技,2021(17):262-266.

[44] 吴晓倩,权丽丽,陈诚,石磊.基于大数据决策树算法的学生成绩分析与预测模型仿真[J].电子设计工程,2020,28(24):138-141+146.

[45] 蒋亮,唐紫珺.基于 Flume、Kafka、HDFS 的多源数据采集系统[J].信息技术与信息化,2021(06):115-117.

[46] 张赛. 基于 Spark 的地震大数据并行处理系统的研究与实现[D].山东交通学院,2021.

[47] 高艳凯. 基于大数据平台的征信数据采集和数据处理研究[D].中北大学,2021.

[48] 冯冼. 基于 Kafka 的高并发消息机制优化研究[D].湘潭大学,2021.

[49] 杨燕艳,朱春燕.基于大数据技术的智慧医疗平台设计[J].信息记录材料,2022,23(06):173-176.

[50] 李晓辉.大数据技术架构下的高维数据挖掘算法分析[J].信息技术,2021(10):122-126.

[51] 石慧,陈培辉.基于大数据技术的房价数据采集及可视化分析应用[J].计算机时代,2021(08):71-75.

[52] 颜晓莲,章刚,邱晓红.Kafka 中改进型 Partition 过载优化算法[J].计算机技术与发展,2020,30(12):88-91.

[53] 徐新宇. 基于 Flink 的实时数据可视化系统的设计与实现[D].华中科技大学,2020.

[54] 吴鹏,马潮.基于小波包分解-KPCA-SVM 的压缩机气阀故障诊断技术研究[J].石油和化工设备,2021,24(11):53-56.

[55] 喻夏,周李涌.融合气象因子的苏尼特羊肉价格预测方法研究[J].智慧农业导刊,2022,2(13):37-40.

[56] 陈东宁,侯安农,姚成玉,侯鑫,邢然.一种新型动态贝叶斯网络分析方法[J].中国机械工程,2020,31(12):1394-1406+1414.

[57] 李博,张洪刚.基于高阶动态贝叶斯网络嵌入的路面异常检测算法[J].华南理工大学学报(自然科学版),2020,48(01):51-59.

[58] 林洪桦.测量误差分析及数据处理若干要点系列论文——移动平均式数据处理[J].自动化与信息工程,2020,41(05):1-6.

[59] 王娟娟. 基于数据挖掘技术的电信用户粘合度评分研究[D].山东师范大学,2022.

[60] Bahram, Choubin, Ehsan, et al. An Ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines[J]. The Science of the total environment,2018,651(Pt 2).

[61]郭淼.基于支持向量回归的大型客运站客流量预测应用研究[J].铁路计算机应用,2021,30(03):15-18.