山东大学
SHANDONG UNIVERSITY

《数据挖掘与机器学习》
课程论文

题　　目：基于人工智能方法的分类方法研究及应用

院　（部）：控制科学与工程学院

姓　　名：张辉、张良、明昊宇

指导教师：刘治平

完成日期：2023 年 2 月 3 日

目 录

摘 要

近年来，伴随着科学技术、互联网技术以及信息技术的迅猛发展，尤其是大数据信息时代的到来，海量的数据信息充斥着我们实际生活中的每一个领域。若希望大数据产生实质性的价值和意义，对大数据的处理过程是极其重要的，而数据分类问题作为大数据分析与数据挖掘中的关键内容，各种各样的分类技术和算法得到了高速发展，各类分类算法针对不同分类问题的性能是不一样的，本文通过 IRIS 数据集和 WINE 数据集，运用决策树、K 近邻分类器、感知机、逻辑回归、支持向量机、随机森林、朴素贝叶斯、多层感知机、TSK 模糊逻辑系统等算法进行两次分类实验，根据实验结果对比分析了各类算法的性能指标。

本文代码见 https://github.com/LittleLongFei/Classification。

关键词：数据挖掘；分类方法；机器学习；支持向量机

Research and Application of Classification Method Based on Artificial Intelligence Method

ABSTRACT

In recent years, with the rapid development of science and technology, Internet technology and information technology, especially the arrival of the big data information age, massive data information is flooding every field of our actual life. If big data is expected to generate substantial value and significance, the processing of big data is extremely important, and data classification is a key content in big data analysis and data mining. Various classification techniques and algorithms have been obtained. With the rapid development, the performance of various classification algorithms for different classification problems is different. This paper uses the IRIS data set and the WINE data set, using decision trees, K nearest neighbor classifiers, perceptrons, logistic regression, support vector machines, random forests, Naive Bayesian, multi-layer perceptron, TSK fuzzy logic system and other algorithms were tested twice, and the performance indicators of various algorithms were compared and analyzed according to the experimental results.

Code is available at this repository: https://github.com/LittleLongFei/Classification.

Key Words: data mining; classification method; machine learning; support vector machine

1 引 言

1.1 研究背景

数据挖掘技术就是从大量数据信息当中获取可用、有效信息的一个过程，从数据当中寻找、探索、开采知识的过程。同时，数据挖掘技术是现代互联网、计算机等信息技术高速发展下的产物，涉及信息化知识理论相对较多，包括数据库、统计学、电子学、人工智能等多个领域，数据挖掘技术是一项覆盖范围广阔、涉及内容烦琐复杂、融括领域较多的学科。关于数据挖掘技术的工作过程，具体如图 1.1 所示。

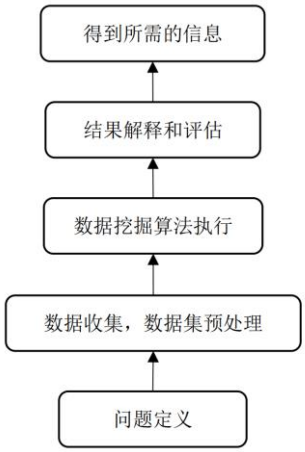

图 1.1 数据挖掘基本过程示意图

数据挖掘在主要任务方面，需要对其进行合理分类、科学预测、关联分析、类别汇集、时间顺序排列以及误差、缺陷分析等多项工作。其中，数据的合理分类是数据挖掘和分析过程中一个极为重要和关键的技术，始终是相关领域的讨论热点和热门研究主题，因为差异性的分类算法将导致出现各种不同的分类算法，例如决策树、K 近邻分类器支持向量机、朴素贝叶斯、多层感知机和逻辑回归等经典分类算法，以及 TSK 模糊逻辑系统这类具备机器学习能力的算法。同时各类分类算法的优劣又会对最终分类结果的可靠性、精准性以及大数据分析与数据挖掘的效率、质量造成直接性影响，所以在我们对规模系统庞大、数据信息量较高的数据进行分类时，需要合理选择分类算法，这对于相关任务的完成时至关重要的。

1.2 研究意义

分类问题是数据挖掘处理的一个重要组成部分，在机器学习领域，分类问题通常被认为属于监督式学习(supervised learning)，也就是说，分类问题的目标是根据已知样本的

某些特征，判断一个新的样本属于哪种已知的样本类。根据类别的数量还可以进一步将分类问题划分为二元分类(binary classification)和多元分类(multiclass classification)。

自从统计学诞生开始，就不断地出现新的分类算法及其各种改进方法，这些方法被广泛应用在社会的各行各业，尤其是医学和经济管理领域。

在医学领域，医生和学者们对心血管功能进行定量的判别与预测；探讨肺癌细胞核的有关体视学参数在肺癌诊断分型方面的意义；对因大肠癌而住院的病历按治愈和未愈分两组进行非条件多因素分类分析；通过分类探讨进展期胃癌淋巴结的转移规律。

在医学生物学领域中，Biometrics、Biometrical Journal 等学术刊物每年都刊登很多判别分析或逻辑回归分类的论文。在国内学术刊物中，这两种方法的应用也很多。医生和学者们借助于判别分析对心血管功能进行定量的判别与预测；利用判别分析探讨肺癌细胞核的有关体视学参数在肺癌诊断分型方面的意义；对因大肠癌而住院的病历按治愈和未愈分两组进行非条件多因素逻辑回归分析；通过逻辑回归探讨进展期胃癌淋巴结的转移规律。

在经济管理领域，对保险公司破产原因进行分析，量化保险公司倒闭前 5 年的公司金融问题信号，对金融风险概率显著性的评价；预测非寿险公司偿付能力，并检测显著影响非寿险公司偿付能力的因素；对分类预测失败的商务案例进行评价；利用多元判别分析和神经网络对上市公司财务困境进行预警分析；研究上市公司财务危机预警；基于个人消费信贷数据, 建立个人信用评分的判别模型；利用判别分析对商业银行监管和监控指标进行研究；建立分区域、分行业的逻辑回归财务预警模型等等。

1.3 研究现状

Han 等人将研究在不同类型数据库中挖掘知识的技术，包括关系数据库，事务数据库，面向对象数据库，空间数据库和活动数据库，以及全球信息系统[1]。Ritchie 等人引入多因素降维(MDR)作为降低多位点信息维数的方法，以改善与疾病风险相关的多态性组合的识别[2]。Hahn 等人描述了 MDR 方法和 MDR 软件包。Romero 等人调查了数据挖掘在传统教育系统中的应用，特别是基于 Web 的课程，众所周知的学习内容管理系统以及自适应和智能的基于 Web 的教育系统[3]。Lou 等人报告了一种广义 MDR(GMDR)方法，该方法允许调整离散和定量协变量，并且适用于各种基于人群的研究设计中的二分类和连续表型。这项工作是对数据挖掘在学习管理系统中的具体应用的调查，也是 Moodle 系统的案例研究教程[4]。

Kingsford C 等人对决策树方法进行了详细的介绍[5]。Khosravi 等人测试了四种基于决策树的机器学习模型，即逻辑模型树(LMT)、减少错误修剪树(REPT)、朴素贝叶斯树(NBT)和交替决策树(ADT)，用于伊朗北部哈拉兹流域的山洪敏感性映射[6]。Mistry 等人研究了一个大规模、与工业相关的混合整数非线性非凸优化问题，涉及梯度提升树和降低风险的惩罚函数[7]。Yang 等人提出了深度神经决策树(DNDT)-由神经网络实现的树模型[8]。Chu 等人提出了一种多层混合深度学习系统(MHS)，可以自动对城市公共区域个人处理的废物进行分类[9]。Heidari 等人使用最近提出的多层感知器(MLP)神经网络的蚱蜢优化算法(GOA)提出了一种新的混合随机训练算法[10]。Goyal 等人提出了一种特征选择技术，以使用基于多层感知器架构的非线性模型有效地估计工作量[11]。Ali 等人使用多层感知器神经网络(MLPNN)算法进行干旱预测[12]。Zhao 等人提出了一种视觉分析系统，旨在解释随机森林模型和预测[13]。Cheng 等人提出了一种鲁棒的随机森林方法来分析出行模式选择，以检查预测能力和模型可解释性[14]。Speiser 等人旨在能够在随机森林分类设置中评估不同的变量选择技术，以便根据专家和智能系统中的应用确定首选方法[15]。Chen 等人通过结合 Nagar-Bardini(NB)和 Nie-Tan(NT)非迭代算法求解输出区间 2 型模糊集的质心，讨论了区间 2 型模糊逻辑系统的模糊推理、类型约简和解模糊化块[16]。Castillo 等人使用连续卡尼克-孟德尔方法(CEKM)的近似，降低了处理区间类型 2 模糊系统以动态适应元启发式参数的计算成本[17]。Lee 等人提出了一种创新的混沌区间 2 型模糊神经振荡网络(CIT2-FNON)，用于全球金融预测[18]。Zhang 等人针对一类具有不确定扰动的非线性系统构建了基于命令滤波器的自适应模糊控制器[19]。Qu 等人根据城市交通流的时间序列，在模糊逻辑的理论框架下，提出了一种基于类型 2 模糊逻辑的预测方法[20]。

1.4 研究内容

本文采用了决策树(Decision Tree)、K 近邻分类器(K Neighbors)、感知机(Perceptron)、逻辑回归(Logistic Regression)、支持向量机(SVM)、随机森林(Random Forest)、朴素贝叶斯(Bayes)、多层感知机(MLP)和 TSK 模糊逻辑系统 9 类算法，针对 IRIS 数据集和 WINE 数据集，结合 Python 程序设计了两次分类实验，并从精确率、运行时间和 F-score 等指标对比了 9 种分类算法的性能。

1.5 本文结构安排

本文结构组织情况如下：在第二部分介绍了常见分类方法的理论基础与数学建模过

程，在第三部分将针对于本文所涉及到的分类方法进行对比实验，在第四部分，对上述实验结果进行总结分析。

2 分类方法的理论基础与数学建模

2.1 支持向量机

支持向量机（Support Vector Machine,SVM）是由 Vapnik 和 Cortes 在 20 世纪末提出的，可用于处理模式分类和非线性回归问题。在处理分类问题时，支持向量机将构造一个最优超平面，所谓最优超平面即是分类间隔最大的，这样可以更好地将两类样本分隔开。支持向量机算法在解决高维空间向量的问题，小样本的问题时有其独特的优势，能够从很大程度上解决维数较高给算法在实际应用中带来的难题。其主要的机制是利用通过训练和学习已经找到的超平面将样本空间进行二分类。当需要解决的问题是线性可分的时候，最优超平面在要求将两类问题正确地分类的基础上，还要求最优超平面最大化；而在解决这种线性不可分的问题时，只利用一个超平面根本不能使两类问题彻底的区分开的时候就需要添加松弛变量的使用，此时通过训练而得到的最优超平面称为广义最优分类超平面。

假设 D 为给定的数据集，数据集每个样本表示为二元组(X_i,y_i)(i=1,2,3,…,N)，X_i 表示第 i 个样本的属性集，y_i 表示的是与之相关联的类别号。令 $y_i \subseteq \{-1,1\}$，即每一个 y_i 只能取+1 或者-1。根据统计学当中的最小结构风险原理可知，只存在一个分类面能够使得属性为 n 维的间隔最大化该平面就被叫做为最大边缘超平面，也可以称之为最优超平面。

用公式来表示最优超平面可以表示为：

$$WX + b = 0 \tag{2-1}$$

其中，W 是权重向量，属性数为 n，内积为 WX，标量为 b。训练的样本属性为二维的，例如 X=(x1 , x2)，x1 与 x2 分别表示的是 X 的属性 A1 与 A2 的值。假如 w0 作为 b 的附加权重出现的话，那么可以把该分离超平面的形式改成公式：

$$w_0 + w_1 x_1 + w_2 x_2 = 0 \tag{2-2}$$

对权重做稍微的进行调整，目的是将定义好的超平面可以用如下公式表示：

$$H_1 : w_0 + w_1 x_1 + w_2 x_2 \geqslant 1 \quad \text{对于所有 yi=+1} \tag{2-3}$$

$$H_2 : w_0 + w_1 x_1 + w_2 x_2 \leqslant 1 \quad \text{对于所有 yi=-1} \tag{2-4}$$

总的来说，元组若是落在 H1 上或其上方，则属于+1 类，反之，若是落在 H2 上或其下方，则属于-1 类。H1 上面的任意一点到分离超平面的距离可以表示为 $\frac{1}{\|W\|}$，$\|W\|$=

$\sqrt{W \times W}$ 由分类超平面的定义可知，它也可以表示分离超平面到 H2 上面的任意一点的距离。由此可知，最大的边缘中间的距离等于 $\frac{2}{\|W\|}$，如下图 2.1 所示。

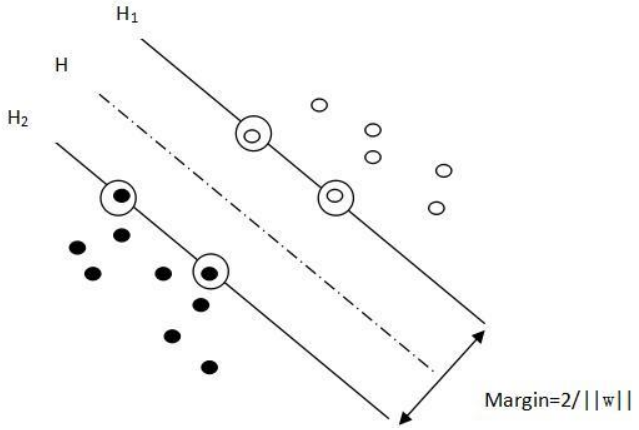

图 2.1 最优超平面

当需要被划分的支持向量为线性可分的情况下，分离面的需求是在将两类样本空间准确无误的区分开的前提下，还需要超平面的间隔最大化。两种样本分别用空心点和实心点去表示，分离面为中间的虚线，也被称为最优超平面。

2.2 决策树

决策树是一种类似二叉树或多叉树的树结构，是目前应用最为广泛的归纳推理算法之一。它是一种分类精度高且操作简单的逼近离散函数值的方法，能有效地处理有噪声或缺失的数据，因而称为一个非常实用的分类算法。

决策树是一种自上而下、分而治之的归纳过程，其本质过程就是一个贪心算法的过程，对于每一次对训练子集的分类都选择当前最大的属性作为分类属性。决策树的最大优点是：构造决策树之前并不需要了解分类背景相关的知识，只需要根据训练子集就可以构建决策树模型，如果当前训练子集不具有相同的决策，则继续按照该方法直至训练子集都属于同一类决策。决策树的每一个叶节点都代表一个类别，每一个非叶节点代表一个属性的测试，每一节点都是根据属性不同取值产生的分枝，每一从根节点达到叶节点的路径都代表唯一的规则。

决策树分类的步骤首先利用训练集建立决策树模型，然后利用已经建立好的模型对未知数据样本进行分类。方法是自根节点开始沿着枝干往下走，直至达到叶节点，此时叶节点就代表数据属于的类别。另外，在创建决策树时，由于噪声数据和孤立点的存在，在构建决策树的过程中可能会产生过度拟合，导致决策树中存在错误的节点，因此对于

构建的决策树需要做剪枝处理。

决策树剪枝分类预剪枝和后剪枝。预剪枝就是根据判断添加该节点是否会产生错误，若产生则不增加该节点。后剪枝是在已经产生的决策树中根据某种策略剪去错误的节点。

决策树的输入是一组带有规则的训练集，输出结果是一个类似树状的决策树模型。每一节点代表一个逻辑判断，每一条边是代表逻辑判断产生的结果，每一叶节点代表一个类别。构建决策树分为两步：首先根据训练集生成决策树模型，然后利用剪枝集对决策树中错误的节点进行修剪提高分类精度。

决策树分类算法很多，本文着重介绍 ID3 算法，该算法采用信息增益作为度量标准选择测试属性来划分样本信 息论表明：属性信息增益越大，系统越稳定，传递信息越充分，算法结合信息论原理，采用信息增益作为衡量标准，在每个非叶子节点选择信息增益最大的属性作为测试属性来划分样本，选取测试属性的每一个取值建立由该结点引出的分支，形成该节点的后续子节点，最终构造出决策树模型。如果我们缩短各分支节点与其叶子节点的 平均距离，则决策树模型的平均深度会变小，根据前述的分析可知：这可以有效提高决策树分类算法的速度和准确率。

具体实现过程如下：

输入：训练数据集中的候选属性集（此处用 attribute list 表示）

输出：决策树

（1）创建一个根结点

（2）若该结点中的所有样本均为同一类别 C，则：

（3）返回作为一个叶结点，以类别 C 标记；

（4）若 attribute list 为空，则：

（5）返回 N 作为一个叶结点，并标记为该结点所含样本中类别个数最多的类别；

（6）从 attribute list 选择一个信息增益最大的属性 test attribute，并将结点 N 标记为 test attribute

（7）对于 test attribute 中的每一个已知取值 ai，准备划分结点 N 所包含的样本集；

（8）根据 test attribute=ai 条件，从结点 N 产生相应的一个分支，以表示该测试条件；

（9）设 Si 为 test attribute=ai 条件所获得的样本集合；

（10）若 Si 为空，则将相应叶结点标记为该结点所含样本中类别个数最多的类别；

（11）否则将相应叶结点标志为 Generate decision tree 返回值。

2.3 K 近邻分类器

K 近邻分类器（最近邻分类器）具有简单有效的结构，是机器学习的经典非线性模型，K 近邻分类器以样本的特征向量为输入，特征向量构成特征空间，分类器输出为样本的类别分类时，以样本的个最近点作为局部空间，该样本的类别由局部空间内类别的数量进行表决。与其他模型不同，K 近邻分类器模型没有明确训练过程，而是使用数据集来划分特征空间，这样可以获得更高精度。作为 K 近邻分类器的唯一参数，邻域大小在分类性能中起着关键作用。如果邻域太小，模型将是复杂且过度拟合的，而如果太大，模型将太简单而不能获得令人满意的精度。此外，样本的类别分布不均也会影响 K 近邻分类器的表现。

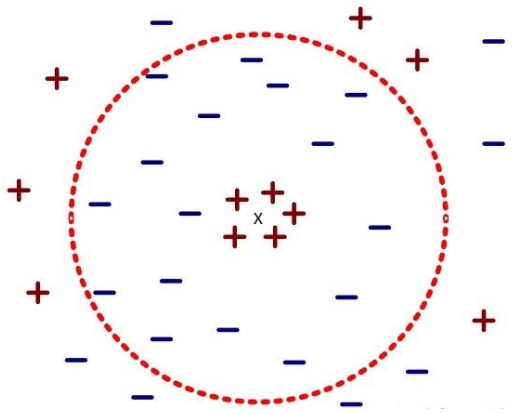

图 2.2 k 较大时的 k-最近邻分类

最近邻算法对每一个测试样例 $z=(x',y')$，计算它和所有训练样例 $(x,y)\in D$ 之间的距离（或相似度），已确定其最近邻列表 D_z，当得到最近邻列表后，测试样例就会根据最近邻中的多数类进行分类：

$$\text{多数表决：}\ y'=\arg\max\sum_{(x_i,y_i)\in D_z} I(v=y_i) \qquad (2\text{-}5)$$

其中，v 是类标号，y_i 是一个最近邻的类标号，$I(v=y_i)$ 是指示函数，如果其参数为真，则返回 1，否则，返回 0。在一些特殊场合，也可根据每个最近邻 x_i 距离的不同对其作用加权：$w_i=1/d(x',x)^2$。类标号可由下面公式确定：

$$\text{距离加权表决：}\ y'=\arg\max\sum_{(x_i,y_i)\in D_z} w_i\times I(v=y_i) \qquad (2\text{-}6)$$

算法具体实现过程如下：

（1）计算测试数据与各个训练数据之间的距离；

（2）按照距离的递增关系进行排序；

（3）选取距离最小的 K 个点；

（4）确定前 K 个点所在类别的出现频率；

（5）返回前 K 个点中出现频率最高的类别作为测试数据的预测分类。

2.4 朴素贝叶斯分类器

朴素贝叶斯分类器（Naive Bayes classifier）基于概率统计学的贝叶斯定理，是一种在先验概率与类条件概率已知的情况下，预测类成员关系可能性的模式分类算法。朴素贝叶斯分类算法成立的前提是：给定类标号，各属性相对于类别是独立的。根据这个假设条件可以得到它的结构，如下图 2.3 所示：

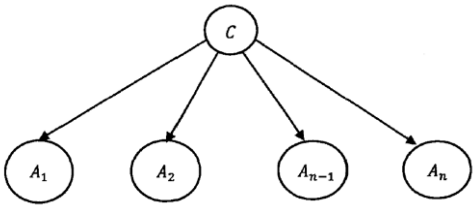

图 2.3 朴素贝叶斯算法结构图

其中 c 表示类标记的集合，即类别，A_i 为训练集中的第 i 个属性。

在分类问题中，设 X 表示属性集，Y 表示类变量，在两者关系不确定时，根据贝叶斯定理，可用先验概率 P(Y)、类条件概率 P(X|Y)和证据 P(X)表示后验概率 P(Y|X)：

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \tag{2-7}$$

当有了条件独立假设后，就不需要计算 X 每一个组合的类条件概率，只需对给定的 Y，计算每一个 Xi 的条件概率。在分类测量记录时，分类器对每个类 Y 计算后验概率：

$$P(Y \mid X) = \frac{P(Y)\prod_{i=1}^{d} P(X_i \mid Y)}{P(X)} \tag{2-8}$$

由于对所有的 Y，P(X)是固定的，因此只要找出对分子最大的类，而在求解类条件概率时，可采用两类方法，第一类是通过把每一个连续的属性离散化，然后用相应的离散区间替换连续属性值；第二类是可以假设连续变量服从某种概率分布，然后使用训练数据估计分布的参数。

2.5 逻辑回归分类器

逻辑回归分类器是基于线性回归的适用于二分类问题（经推广后也可用于多分类问

题）的分类器。在用于用于分类问题时，需要设定一个阈值来判断待预测类的所属类别。由于逻辑回归具有较好的可解释性和泛化性，因此常被视为机器学习和模式识别中最重要的分类模型之一。

逻辑回归分类器的基本思想就是将线性回归结果作用在某种非线性函数上（即逻辑斯谛分布函数），从而实现对结果的压缩和对分类的预测。通过将逻辑斯谛分布函数中的 z 用线性回归函数代入，即可得到逻辑回归公式：

$$P(y=1)=\frac{1}{1+e^{-\theta^T x}} \tag{2-9}$$

$$P(y=0)=\frac{e^{-\theta^T x}}{1+e^{-\theta^T x}} \tag{2-10}$$

可见，线性回归在逻辑斯谛分布函数的作用下，成功将回归结果压缩到(0，1)区间内，从而不仅解决了对异常值的敏感性，而且可以方便地在(0，1)区间选取某个值完成分类。

在得到逻辑回归公式后，由于是分段函数，且难以找到合适的损失函数，因此还无法直接求解。可以通过概率的思维，运用极大似然估计构造似然函数，并取对数似然函数，在根据最大似然函数得到逻辑斯谛回归的最优参数为：

$$\arg\max_{\theta} \sum_i y_i \times \ln h(\theta, x_i) + (1-y_i) l \times n(1-h(\theta, x_i)) \tag{2-11}$$

可根据逻辑斯谛回归分布函数的导数进行梯度下降迭代求解。

在逻辑回归中常用交叉熵损失函数，交叉熵损失函数和极大似然法得到的损失函数是相同的。可以定义如下的交叉熵损失函数：

$$L(x, y, \theta) = -y \ln h(\theta, x) - (1-y) \ln(1-h(\theta, x)) \tag{2-12}$$

2.6 随机森林

随机森林就是通过集成学习的 Bagging 思想将多棵树集成的一种算法：它的基本单元就是决策树。每棵树的按照如下规则生成：

（1）如果训练集大小为 N，对于每棵树而言，随机且有放回地从训练集中的抽取 N 个训练样本（就是 bootstrap sample 方法, 拔靴法采样）作为该树的训练集；从这里我们可以知道：每棵树的训练集都是不同的，而且里面包含重复的训练样本。

（2）如果存在 M 个特征，则在每个节点分裂的时候，从 M 中随机选择 m 个特征维度（m << M），使用这些 m 个特征维度中最佳特征(最大化信息增益)来分割节点。在森林生长期间，m 的值保持不变。

随即森林具有如下特点：

（1）能够处理具有高维特征的输入样本，而且不需要降维

（2）能够评估各个特征在分类问题上的重要性

（3）在生成过程中，能够获取到内部生成误差的一种无偏估计

（4）对于缺省值问题也能够获得很好得结果

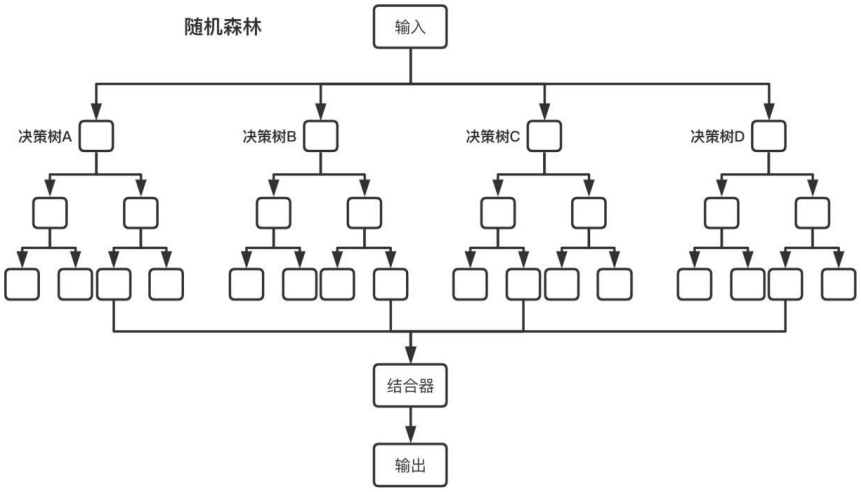

图 2.4 随机森林算法结构图

随机森林算法步骤：假设训练集 T 的大小为 N ，特征数目为 M，随机森林的大小为 K。遍历随机森林的大小 K 次：

（1）从训练集 T 中有放回抽样的方式，取样 N 次形成一个新子训练集 D；

（2）随机选择 m 个特征，其中 m < M；

（3）使用新的训练集 D 和 m 个特征，学习出一个完整的决策树得到随机森林。

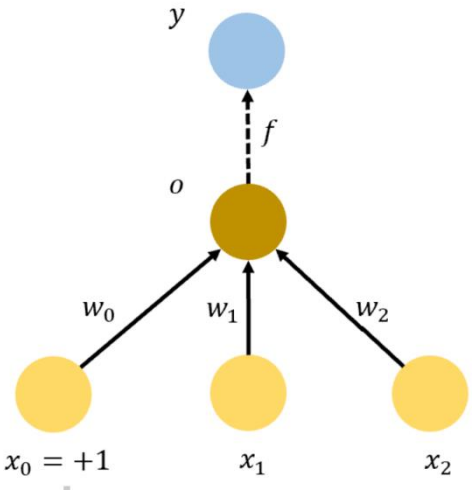

图 2.5 感知机结构图

2.7 感知机

感知机(perceptron)是二类分类的线性分类模型，其输入为实例的特征向量，输出为实例的类别，是神经网络的基础结构。感知机学习旨在求出将训练数据进行线性划分的分离超平面，为此，导入基于误分类的损失函数，利用梯度下降法对损失函数进行极小化，求得感知机模型。感知机学习算法具有简单，并且易于实现的特点，分为原始形式和对偶形式。感知机预测是用学习得到的感知机模型对新的输入实例进行分类。它是神经网络与支持向量机的基础。

感知机的结构由如下几部分构成：

（1）两个接受由外部输入的输入值的输入节点（黄色）：x_1、x_2

（2）一个由内部模型决定的阈值的输入节点（黄色）：x_0（这是为了使得感知机更加的灵活，而不仅仅依赖于外部输入值来确定输出值）

（3）三个输入节点到中间节点的连接权重（实线箭头）：ω_0-ω_2（这就是感知机将要学习的东西：不同的环境因子对预测结果的影响力分别有多大）

（4）一个由输入节点和连接权重，通过加权求和计算得到的中间值（深黄色）：o（这是先算出输入层带来的基础结果，为计算出最终的输出值做准备）

（5）一个用来对中间值做特殊变换，从而得到最终输出结果的激活函数：f（这一步可以用各种各样的函数，对初步计算得到的中间值做各种各样的变换，从而使得感知机更加灵活的输出，实现不同的效果、解决不同的问题。也可以不做任何变换，保留原始值）

2.8 多层感知机（神经网络）

多层感知器（Multilayer Perceptron,缩写 MLP）是一种前向结构的 人工神经网络 ，映射一组输入向量到一组输出向量。MLP 可以被看作是一个有向图，由多个的节点层所组成，每一层都全连接到下一层。除了输入节点，每个节点都是一个带有非线性激活函数的神经元（或称处理单元）。一种被称为 反向传播算法的监督学习 方法常被用来训练 MLP。多层感知器遵循人类神经系统原理，学习并进行数据预测。它首先学习，然后使用权重存储数据，并使用算法来调整权重并减少训练过程中的偏差，即实际值和预测值之间的误差。主要优势在于其快速解决复杂问题的能力。

多层感知机（MLP，Multilayer Perceptron）也叫人工神经网络（ANN，Artificial Neural Network），除了输入输出层，它中间可以有多个隐层，最简单的 MLP 只含一个隐层，即

三层的结构。

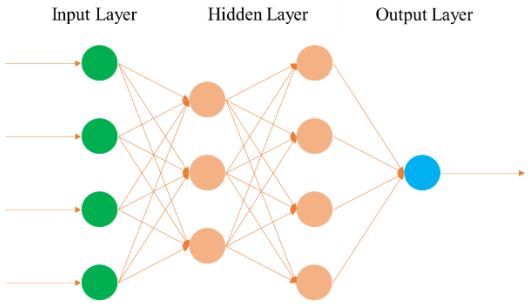

图 2.6 多层感知机结构

引入隐藏层的神经网络可以等价于仅含输出层的单层神经网络的问题，在于全连接层只是对数据做仿射变换（affine transformation），而多个仿射变换的叠加仍然是个仿射变换。解决问题的法之一是引线性变换，对隐藏变量使按元素运算的线性函数进变换，再作为下个全连接层的输。线性函数被称为激活函数。

（1）ReLU 函数：

ReLU（rectified linear unit）函数提供了一个很简单的非线性变换。给定元素x，该函数定义为：

$$\mathrm{ReLU}(x)=max(x,0) \tag{2-15}$$

（2）sigmoid 函数

sigmoid 函数可以将元素的值变换到 0 和 1 之间：

$$sigmoid(x)=\frac{1}{1+e^{-x}} \tag{2-16}$$

（3）tanh 函数

tanh（双曲正切）函数可以将元素的值变换到-1 和 1 之间：

$$tanh(x)=\frac{1-e^{-2x}}{1+e^{-2x}} \tag{2-17}$$

多层感知机就是含有少个隐藏层的由全连接层组成的神经络，且每个隐藏层的输出通过激活函数进变换。多层感知机的层数和各隐藏层中隐藏单元个数都是超参数。

2.9 TSK 模糊逻辑系统

Takagi-Sugeuo-Kang(TSK)模糊系统是由 Takagi、Sugeno 和 Kang 提出的，由于其良好的逼近性能而被广泛运用到系统辨识、模式识别、图像处理和数据挖掘等多个领域。模糊系统的主要问题是模糊规则的提取。常用的算法有模糊聚类法、神经网络和遗传算法。

该系统的优势在于 TSK 模糊逻辑系统是基于规则的，既可以利用误差等数据信息，也可利用专家的经验知识，这就为在预测系统中设计合适的修正子系统提供了灵活性；其次模糊逻辑无需建立对象的数学模型，而预测系统要建立起精确的数学模型是很困难的。最后 TSK 模糊逻辑系统属于非线性系统，一般生化过程预测对象都是非线性的，要对其进行误差反馈修正，调节器应该是非线性的，因此通过模糊系统的合理设计和调节，是可以得到理想的预测误差修正机制。

3 实 验

在本节将对上述内容提到的各类机器学习方法在分类问题上进行应用与验证，具体地，本节将涉及到一些经典的机器学习算法，决策树(Decision Tree)、K 近邻(K Neighbors)、感知机(Perceptron)、逻辑回归(Logistic Regression)、支持向量机(SVM)、随机森林(Random Forest)、朴素贝叶斯(Bayes)、多层感知机(MLP)，此外，TSK 模糊逻辑系统作为一种有效的机器学习方法一，也将作为对比方法参与本节的各个实验。

在接下来的内容中，首先，介绍了本节实验部分用到的性能指标；然后详细阐述了两个分类实验的过程以及结果；最后，对于本节内容进行了简要的总结。

3.1 性能指标

分类是监督学习中的一个核心问题。为了评价一个分类器的分类性能优劣，需要引入一些评估指标，常用的一些指标有混淆矩阵（Confuse Matrix）、准确率（Accuracy）、错误率（ErrorRate）、精准率（Precision）和召回率（Recall）、F1-score、ROC 曲线(Receiver Operating Characteristic Curve)、AUC（Area Under the Curve）、KS 曲线、Lift 值、P-R 曲线。接下来，以分类问题为例，介绍了分类问题中常用并且在本实验中用到的性能评估指标：精确率和 F1-score。

3.1.1 精确率(Precision)

精准率又称查准率，它是针对预测结果而言的，它的含义是在所有被预测为正的样本中实际为正的样本的概率，即在预测为正样本的结果中，有多大几率可以预测正确，其公式如下：

$$Precision=\frac{TP}{TP+FP} \tag{3.1}$$

此外，指标召回率也能反应相应的分类性能，单一依靠某个指标并不能较为全面地评价一个分类器的性能。一般情况下，精确率越高，召回率越低；反之，召回率越高，精确率越低。为了平衡精确率和召回率的影响，较为全面地评价一个分类器，引入了 F-score 这个综合指标。

3.1.2 F-Score

F-score 是精确率和召回率的调和均值，计算公式如下：

$$F_\beta=(1+\beta^2)\frac{Precision\times Recall}{\beta^2\times Precision+Recall},$$

$$=(1+\beta^2)\frac{(1+\beta^2)TP}{(1+\beta^2)TP+\beta^2FP+FN}. \tag{3.2}$$

其中，β的取值反映了精确率和召回率在性能评估中的相对重要性，通常情况下，β取值为 1，也就是 F1- score，表明精确率与召回率同样重要。

$$F_1=\frac{2\times TP}{2\times TP+FP+FN}, \tag{3.3}$$

此外，除了上述内容中介绍的精确率以及 F1-score，实验中还将训练时间（Training Time）作为一项评估指标，计算方式采用的是 SciKit-Learn 库中的函数。

表 3.1 IRIS 数据集部分样本

花萼长度	花萼宽度	花瓣长度	花瓣宽度	品种
5.1	3.5	1.4	0.2	Iris-setosa
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor
5.5	2.3	4.0	1.3	Iris-versicolor
6.5	2.8	4.6	1.5	Iris-versicolor
6.3	3.3	6.0	2.5	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica
7.1	3.0	5.9	2.1	Iris-virginica
6.3	2.9	5.6	1.8	Iris-virginica
6.5	3.0	5.8	2.2	Iris-virginica

3.2 IRIS 分类实验

3.2.1 IRIS 数据集

安德森鸢尾花卉数据集（Anderson’s Iris data set），也称鸢尾花卉数据集（Iris flower data set）或费雪鸢尾花卉数据集（Fisher’s Iris data set），是一类多重变量分析的数据集。最初是埃德加·安德森从加拿大加斯帕半岛上的鸢尾属花朵中提取的地理变异数据。它首次出现在著名的英国统计学家和生物学家 Ronald Fisher 1936 年的论文《The use of multiple measurements in taxonomic problems》中，被用来介绍线性判别式分析。其数据集包含了 150 个样本，都属于鸢尾属下的三个亚属，分别是山鸢尾、变色鸢尾和维吉尼亚鸢尾（Iris Setosa，Iris Versicolour，Iris Virginica）。四个特征被用作样本的定量分析，它们分别是花萼和花瓣的长度和宽度。基于这四个特征的集合，费雪发展了一个线性判别分析以确定其属种。该数据集测量了所有 150 个样本的 4 个特征，分别是：sepal length

（花萼长度）、sepal width（花萼宽度）、petal length（花瓣长度）、（花瓣宽度）。以上四个特征的单位都是厘米（cm）。

该数据集可以从 https://archive.ics.uci.edu/ml/datasets/iris 获取，表 3.1 展示了部分 IRIS 数据集的原始数据。

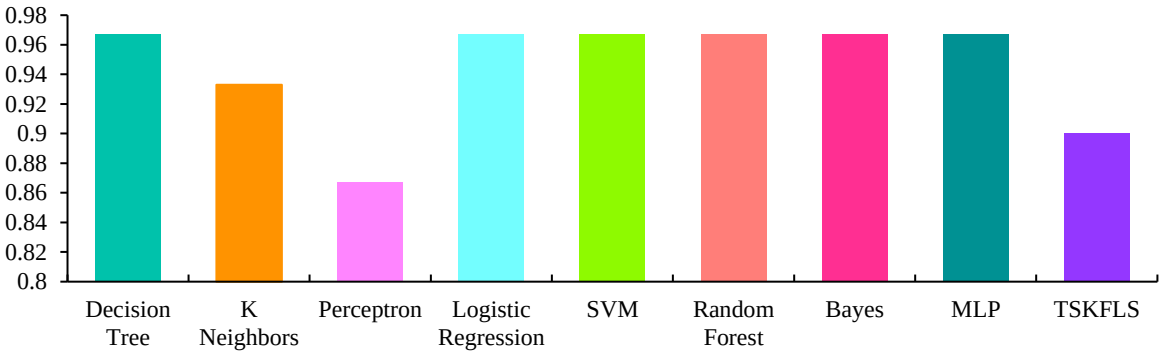

图 3.1 IRIS 分类实验各方法的性能评估指标——精确率

表 3.2 IRIS 分类实验各方法的性能评估指标

Method	Precision	F-score	Training Time(s)
Decision Tree	0.967	0.965	0.507
K Neighbors	0.933	0.930	0.805
Perceptron	0.867	0.848	0.003
Logistic Regression	0.967	0.965	0.008
SVM	**0.967**	**0.965**	**0.001**
Random Forest	0.967	0.965	0.032
Bayes	**0.967**	**0.965**	**0.001**
MLP	**0.967**	**0.965**	**0.001**
TSKFLS	0.900	0.895	0.571

3.2.2 IRIS 分类实验与讨论

该实验的对比方法有决策树、K 近邻、感知机、逻辑回归、支持向量机、随机森林、朴素贝叶斯、多层感知机、TSK 模糊逻辑系统。数据集共计有 150 条数据，实验中，将 80%的数据用于模型训练，并且将剩余的 20%的数据用于测试，训练数据和测试数据为随机选取，并且没有重合的数据，以下结果均为测试过程中获得的结果数据。

表 3.2 给出了本实验中的性能指标的结果，并将其可视化于图 3.1、图 3.2、图 3.3 中。根据实验中得出的结果数据，支持向量机、朴素贝叶斯以及多层感知机不管是在训练时间上还是精确率、F1 值都是这些方法中最优的方法，分类精确率达到 96.7%，在精确度

为三位小数点的情况下，训练时间仅为 0.001s。相比之前，决策树方法的精度虽然高但是却需要更多的训练时间。TSK 模糊逻辑系统虽然达到了很好的分类效果，然而与上述方法仍然存在较大的差距。

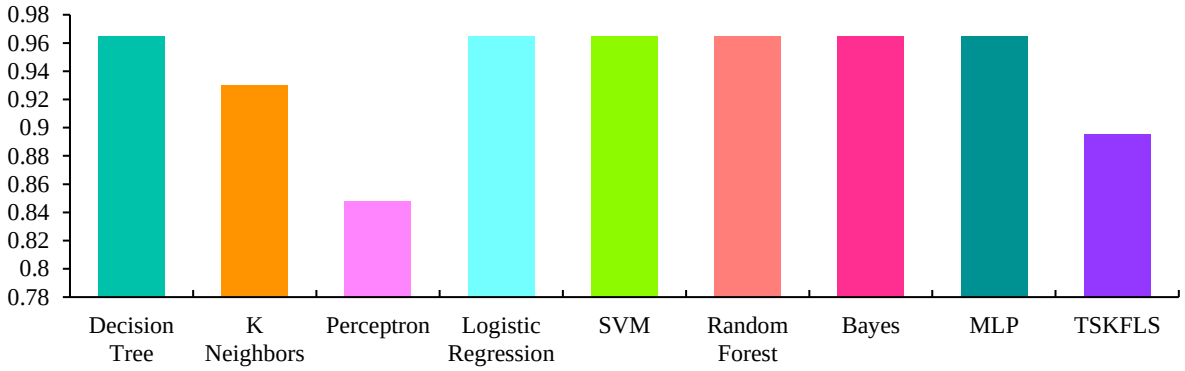

图 3.2 IRIS 分类实验各方法的性能评估指标——F1-Score

表 3.3 WINE 数据集的各个属性描述

属性	属性描述	属性类型
Class	类别	离散
Alcohol	酒精	连续
Malic acid	苹果酸	连续
Ash	灰	连续
lcalinity of ash	灰分的酸碱度	连续
Magnesium	镁	连续
Total phenols	总酚	连续
Flavanoids	黄酮类化合物	连续
Nonflavanoid phenols	非黄烷类酚类	连续
Proanthocyanins	原花色素	连续
Color intensity	颜色强度	连续
Hue	色调	连续
Proline	脯氨酸	连续
OD280/OD315 of diluted wines	稀释葡萄酒的 0280/ OD315	连续

3.3 WINE 分类实验

3.3.1 WINE 数据集

WINE 葡萄酒数据集是来自 UCI 上面的公开数据集，这些数据是对意大利同一地区种植的葡萄酒进行化学分析的结果，这些葡萄酒来自三个不同的品种。该分析确定了三种葡萄酒中每种葡萄酒中含有的 13 种成分的数量。从 UCI 数据库中得到的这个 wine 数

据记录的是在意大利某一地区同一区域上三种不同品种的葡萄酒的化学成分分析。数据里含有 178 个样本分别属于三个类别，这些类别已经给出。每个样本含有 13 个特征分量(化学成分)，分析确定了 13 种成分的数量，然后对其余葡萄酒进行分析发现该葡萄酒的分类。

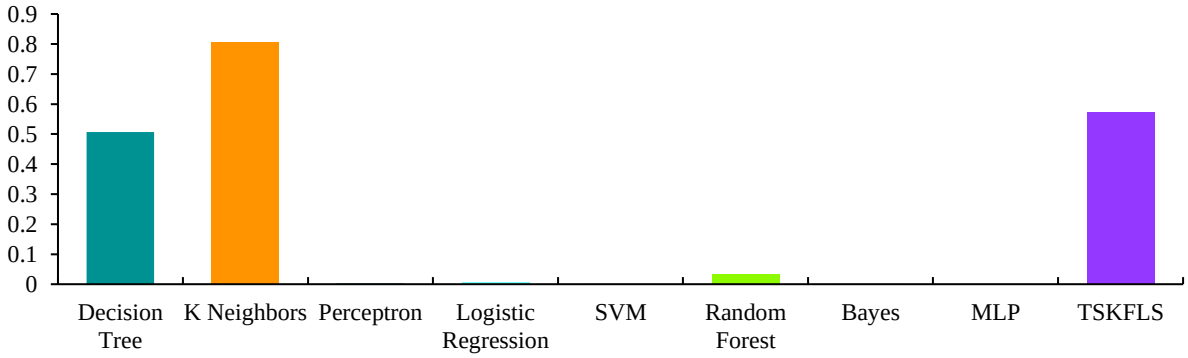

图 3.3 IRIS 分类实验各方法的性能评估指标——Training Time(s)

表 3.4 WINE 分类实验各方法的性能评估指标

Method	Precision	F-score	Training Time(s)
Decision Tree	0.944	0.951	0.563
K Neighbors	0.861	0.860	0.877
Perceptron	0.916	0.916	0.002
Logistic Regression	0.889	0.889	0.005
SVM	**0.999**	**0.999**	**0.001**
Random Forest	0.972	0.975	0.034
Bayes	0.964	0.970	0.001
MLP	0.933	0.965	0.001
TSKFLS	0.916	0.934	0.498

该数据集可以从 https://archive.ics.uci.edu/ml/datasets/wine 获取。

在 WINE 数据集中，这些数据包括了三种酒中 13 种不同成分的数量。文件中，每行代表一种酒的样本，共有 178 个样本；一共有 14 列，其中，第一个属性是类标识符，分别是 1/2/3 来表示，代表葡萄酒的三个分类。后面的 13 列为每个样本的对应属性的样本值。剩余的 13 个属性是，酒精、苹果酸、灰、灰分的碱度、镁、总酚、黄酮类化合物、非黄烷类酚类、原花色素、颜色强度、色调、稀释葡萄酒的 OD280/OD315、脯氨酸。其中第 1 类有 59 个样本，第 2 类有 71 个样本，第 3 类有 48 个样本。

具体属性描述如表 3.3：

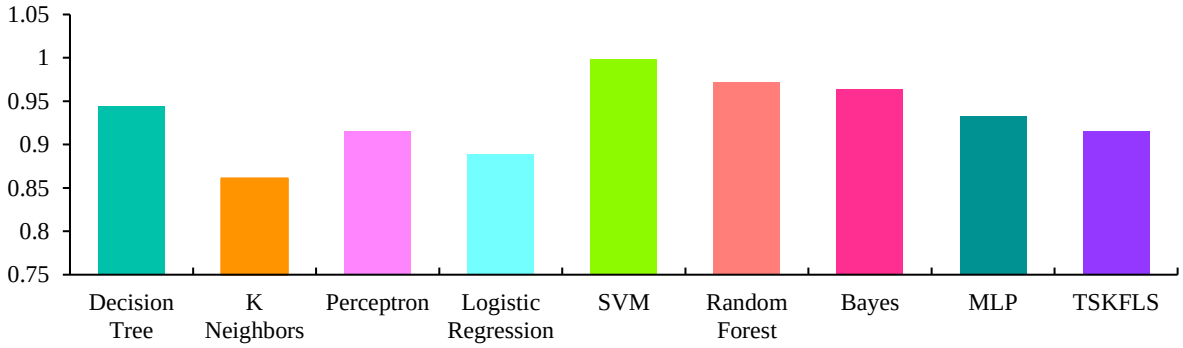

图 3.4 WINE 分类实验各方法的性能评估指标——精确率

3.3.2 实验与讨论

与实验一一致，该实验的对比方法仍然是决策树、K 近邻、感知机、逻辑回归、支持向量机、随机森林、朴素贝叶斯、多层感知机、TSK 模糊逻辑系统。实验中，将 70%的数据用于模型训练，并且将剩余的 30%的数据用于测试，训练数据和测试数据为随机选取，并且没有重合的数据，以下结果均为测试过程中获得的结果数据。

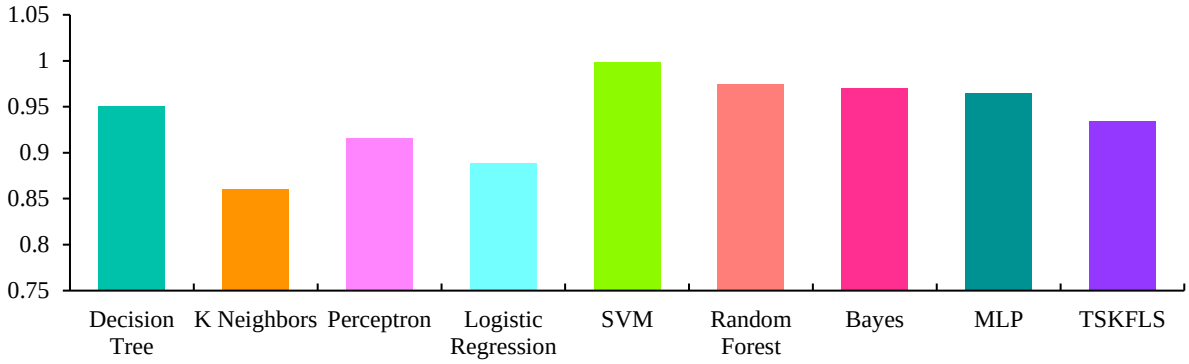

图 3.5 WINE 分类实验各方法的性能评估指标——F1-Score

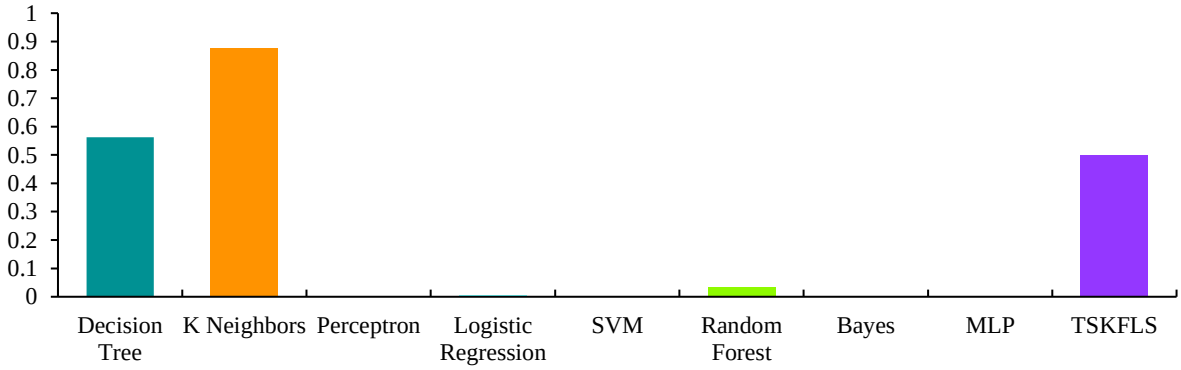

图 3.6 WINE 分类实验各方法的性能评估指标——Training Time(s)

表 3.4 给出了本实验中的性能指标的结果，并将其可视化于图 3.4、图 3.5、图 3.6 中。根据实验中得出的结果数据，支持向量机不管是在训练时间上还是精确率、F1 值都相较于其他方法有着最优秀的表现，分类精确率达到 99.9%，在精确度为三位小数点的情况下，训练时间仅为 0.001s。多层感知机和朴素贝叶斯虽然也可以节省大量的训练时间，但是在分类性能上稍逊色于支持向量机。

决策树方法和 TSK 模糊逻辑系统虽然达到了很好的分类效果，然而与上述方法仍然存在较大的差距，并且这两种方法需要的训练时间相对更长。

4 总 结

在本文中，首先回顾了人工智能领域一些常见的机器学习方法，并且简要的阐述了各种方法的应用情况；然后，在第二部分介绍了几种机器学习方法的数学描述以及建模过程;最后通过两个分类实验对于上述模型进行了实验验证,并且简要分析了实验结果。在实验分析中，支持向量机在两个实验中均表现出了最优秀的分类性能，并且相对于其他方法节省了更多的训练时间。反之，TSK 模糊逻辑系统虽然具备了机器学习的能力，然而在性能上仍然逊色于其他的方法。

在本文的研究中，有许多不足仍然值得改进，本文的重点在于研究不同的机器学习方法在分类问题上的性能差异，但是忽略了优化算法对于各类方法的影响，这些问题值得进一步讨论。

致 谢

在这里，我们向刘治平教授的治学态度以及学术风尚表示崇高的敬意，并且，感谢其在《数据挖掘与机器学习》这门课程中教授我们知识并指导我们完成本文的工作。

贡献与分工

在完成本文的过程中，张辉负责编写 9 种分类算法并进行 2 个分类实验，此外，还负责本文实验（第三部分）和总结（第四部分）两部分文章内容的撰写；张良和明昊宇负责文献资料的查阅以及本文第一部分、第二部分的撰写。

贡献度比重：张辉（34%）、张良（33%）、明昊宇（33%）。

参考文献

[1] Jiawei Han; "Data Mining Techniques", SIGMOD, 1996.

[2] M D Ritchie; L W Hahn; N Roodi; L R Bailey; W D Dupont; F F Parl; J H Moore;"Multifactor-dimensionality Reduction Reveals High-order Interactions Among Estrogen-metabolism Genes in Sporadic Breast Cancer", AMERICAN JOURNAL OF HUMAN GENETICS, 2001.

[3] Lance W. Hahn; Marylyn D. Ritchie; Jason H. Moore; "Multifactor Dimensionality Reduction Software for Detecting Gene-gene and Gene-environment Interactions", BIOINFORMATICS, 2003.

[4] Moty Ben-Dov; Ronen Feldman; "The Data Mining and Knowledge Discovery Handbook", 2005.

[5] Kingsford C, Salzberg S L. What are decision trees?[J]. Nature biotechnology, 2008, 26(9): 1011-1013.

[6] Khabat Khosravi; Binh Thai Pham; Kamran Chapi; Ataollah Shirzadi; Himan Shahabi; Inge Revhaug; Indra Prakash; Dieu Tien Bui; "A Comparative Assessment Of Decision Trees Algorithms For Flash Flood Susceptibility Modeling At Haraz Watershed, Northern Iran", THE SCIENCE OF THE TOTAL ENVIRONMENT, 2018.

[7] Miten Mistry; Dimitrios Letsios; Gerhard Krennrich; Robert M. Lee; Ruth Misener;"Mixed-Integer Convex Nonlinear Optimization With Gradient-Boosted Trees Embedded", ARXIV-MATH.OC, 2018.

[8] Yongxin Yang; Irene Garcia Morillo; Timothy M. Hospedales; "Deep Neural Decision Trees", ARXIV-CS.LG, 2018.

[9] Yinghao Chu; Chen Huang; Xiaodan Xie; Bohai Tan; Shyam Kamal; Xiaogang Xiong; "Multilayer Hybrid Deep-Learning Method For Waste Classification And Recycling", COMPUTATIONAL INTELLIGENCE AND NEUROSCIENCE, 2018.

[10] Ali Asghar Heidari; Hossam Faris; Ibrahim Aljarah; Seyedali Mirjalili; "An Efficient Hybrid Multilayer Perceptron Neural Network with Grasshopper Optimization", SOFT COMPUTING, 2019.

[11] Somya Goyal; Pradeep Kumar Bhatia; "Feature Selection Technique for Effective Software Effort Estimation Using Multi-Layer Perceptrons", PROCEEDINGS OF ICETIT 2019, 2019.

[12] Zulifqar Ali; Ijaz Hussain; Muhammad Faisal; Hafiza Mamona Nazir; Tajammal Hussain; Muhammad Yousaf Shad; Alaa Mohamd Shoukry; Showkat Hussain Gani; "Forecasting Drought Using Multilayer Perceptron Artificial Neural Network Model", ARXIV-PHYSICS.AO-PH, 2019.

[13] Xun Zhao; Yanhong Wu; Dik Lun Lee; Weiwei Cui; "IForest: Interpreting Random Forests Via Visual Analytics", IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, 2018.

[14] Long Cheng; Xuewu Chen; Jonas De Vos; Xinjun Lai; Frank Witlox; "Applying A Random Forest Method Approach to Model Travel Mode Choice Behavior", TRAVEL BEHAVIOUR AND SOCIETY, 2019.

[15] Jaime Lynn Speiser; Michael E Miller; Janet Tooze; Edward Ip; "A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling", EXPERT SYSTEMS WITH APPLICATIONS, 2019.

[16] Yang Chen; "Study on Centroid Type-Reduction of Interval Type-2 Fuzzy Logic Systems Based on Noniterative Algorithms", COMPLEX, 2019.

[17] Oscar Castillo; Patricia Melin; Emanuel Ontiveros; Cinthia Peraza; Patricia Ochoa; Fevrier Valdez; José Soria; "A High-speed Interval Type 2 Fuzzy System Approach for Dynamic Parameter Adaptation in Metaheuristics", ENG. APPL. ARTIF. INTELL., 2019.

[18] Raymond S. T. Lee; "Chaotic Interval Type-2 Fuzzy Neuro-oscillatory Network (CIT2-FNON) for Worldwide 129 Financial Products Prediction", INT. J. FUZZY SYST, 2019.

[19] Jing Zhang; Jianwei Xia; Wei Sun; Zhen Wang; Hao Shen; "Command Filter-based Finite-time Adaptive Fuzzy Control for Nonlinear Systems with Uncertain Disturbance", J. FRANKL. INST, 2019.

[20] Zhengyang Qu; Jun Li; "Short-term Traffic Flow Forecast on Basis of PCA- Interval Type-2 Fuzzy System", JOURNAL OF PHYSICS:CONFERENCE SERIES, 2022.