数据挖掘算法在葡萄酒信息数据分析系统中的研究

郝艳妮 吴素萍 田维丽
(宁夏大学信息工程学院 银川 750021)

摘 要 随着信息科技的快速发展,计算机中的经典算法在葡萄酒产业中得到了广泛的研究与应用。机器学习算法的特点是运用人工智能技术,在经过大量的样本集训练和学习后可以自动地找出运算所需要的参数和模型。针对数据挖掘中常用的机器学习算法进行相关的研究。以分类算法为例进行数据挖掘技术的研究。针对 SVM(支持向量机)泛化能力弱的缺点,给出了一种改进的 SVM-NSVM,即先对训练集进行精选,根据每个样本与最近邻类标的异同判断样本点的取舍,然后再用 SVM 训练得到分类器。针对 kNN(k-最近邻)训练数据集大的缺点,给出了一种改进的通过渐进的思想来寻找最近邻点。实验表明,与 SVM 相比,NSVM 在分类正确率、分类速度上有一定的优势。改进的 kNN 算法的复杂度明显降低。此外,设计了葡萄酒信息数据分析系统,利用数据挖掘方法对极大量的葡萄酒信息数据进行分析、对比与匹配,从而可挖掘葡萄酒的主要成分对比信息和营销潜在信息等;再对这些成分进行相应的分析,并与高质量葡萄酒中的成分进行相应的对比,最终得出葡萄酒的相关分析信息数据,其可帮助葡萄酒生产厂商对葡萄酒的成分含量、品质进行分析。

关键词 机器学习,数据挖掘,分类算法,葡萄酒,数据分析
中图法分类号 TP399 **文献标识码** A

Research on Data Mining Algorithm in Wine Information Data Analysis System

HAO Yan-ni WU Su-ping TIAN Wei-li
(School of Information and Engineering, Ningxia University, Yinchuan 750021, China)

Abstract With the rapid development of information technology, the classical algorithm in computer has been extensively studied and applied in wine industry. The characteristics of machine learning algorithm are to use the technology of artificial intelligence, and a large number of samples in the set of training and learning can atuomatically identify the model and parameters that operation needs. Related research was used in data mining machine learning algorithms in this paper. The research of data mining technology based on classification algorithm was taken as an example. And for the weak generalization ability of SVM(Support Vector Machine), we proposed an improved SVM-NSVM, in which the training set is selected precisely according to each sample, similarities and differences between the subject nearest class choice is decided, and then SVM is trained to get classifier. For big disadvantage of the training data set in kNN's(K-Nearest Neighbor), an improved progressive idea was given to find the nearest neighbor. Experiments show that, NSVM has more advantages than SVM in classification accuracy, speed classification. Complexity of the improved kNN algorithm is significantly reduced. In addition, the wine information and data analysis system was designed, and the data mining method was used to analyze, contrast and match the extremely large amount of wine information data so as to excavate the comparative information of the main components of wine and marketing potential information. Then these components was designed accordingly. With high-quality wine in the corresponding comparison of the ingredients, the final analysis analyze wine-related information and data can help wine producers analyze wine content and wine quality.

Keywords Machine learning, Data mining, Classification algorithm, Wine, Data analysis

随着人们生活质量的提高,葡萄酒成为人们喜爱的饮品,与此同时,宁夏已将葡萄酒作为优势特色产业,在贺兰山东麓计划建设百家葡萄酒庄。宁夏的葡萄酒文化正在快速盛行,人们对葡萄酒的质量提出了很高的要求,政府也越来越重视葡萄酒的信息数据。因此生产厂商对葡萄酒的成分数据进行分析显得弥足重要。

葡萄酒作为许多欧洲国家的一种重要商品,国外对葡萄酒的鉴别方法已经做了较多研究。葡萄酒的成分有 1000 多种[1],且它们之间有着复杂的关系。因此,若要采用科学的方法使存在于这些复杂关系的问题简单化,进而更加清楚地了解它们之间的关系,统计学方法无疑可以为葡萄酒的质量控制、预测、预报、区分提供一种有效的途径。葡萄酒的感官特性[2]取决于葡萄酒的质量等因素。感官评价,尤其是描述分析是确定和检验葡萄酒的感官质量与风格的有效手段[3-5],它可以促进葡萄酒工艺等的改善,指导人们如何鉴赏葡萄酒,是目前仪器分析所无法取代的。描述符[6]是描述分析的语言和工具,因此描述符的选择显得弥足重要。多元统计分析中聚类分析[11]等数学工具能把许多的描述语转化为综合性较强

本文受宁夏科技支撑计划项目(2015BY115),宁夏大学研究生创新项目(GIP201625)资助。
郝艳妮 女,主要研究方向为并行计算与高性能处理、计算机应用技术、农业信息化方向,E-mail:wgh2437547672@qq.com;吴素萍 女,教授,主要研究方向为计算机应用技术、农业信息化、高性能计算及图像处理,E-mail:wspg123@163.com(通信作者)。

的描述语，并且能反映原来多个描述语的信息，从而筛选出科学合理的描述符[7]。国外在这些方面的研究起步较早，而国内在这方面的研究则甚少。国内刘保东等[8]以 18 个不同含汁量的甜红葡萄酒样品作为分析样本，统计分析了与原汁含量相关的 10 项指标，利用计算机模式识别技术[10]，采用 5 种多元回归判别方法，对分析结果进行了回归识别，取得了令人满意的结果。

针对目前挖掘葡萄酒信息数据的准确率低、速度慢的特点，自主设计开发了一套葡萄酒信息数据分析系统。该系统可对极大量的葡萄酒信息数据进行分析、对比与匹配，从而可挖掘葡萄酒的主要成分对比信息和营销潜在信息等；再对这些成分进行相应的分析，并与高质量葡萄酒中的成分进行相应的对比，最终得出葡萄酒的相关分析信息数据，可帮助葡萄酒生产厂商更好地获取葡萄酒的成分含量、葡萄酒的品质、消费者的喜好等分析信息，从而更好地评估与提高葡萄酒的质量。

1 葡萄酒信息数据分析系统的设计

目前，大部分葡萄酒的数据收集与处理系统是基于桌面应用设计的，不但安装配置的步骤繁琐，而且得到的数据仅仅是从个人设备上收集到的，并不能共享他人的数据。通常要花费大量时间来等待数据收集，这样的葡萄酒数据收集和处理系统只能针对专业的数据分析人员，从而造成数据的浪费。

针对上述问题，设计一款葡萄酒信息数据分析系统，该系统采用无线网络技术将传感器采集到的葡萄酒的成分信息数据（包括：糖分[9]、水分）上传至服务器。用户只需使用 WEB 浏览器进行访问，不需要在本机上做任何配置，以便一些非专业人士使用，如生产厂商。实现数据的共享，可加快数据收集的速度，并提高数据分析的准确性。通过科学的算法进行数据的分析，以提高农作物的产量，减少在种植中的消耗，进而提高农民的收入。

该系统由 5 部分组成，包括两组传感器（包含 9 个传感器）、两个数据采集器（主数据采集器和扩展数据采集器）、GPRS 无线传输模块（内含 SIM 卡，用于数据传输）、远程 PC 端、电子鼻系统。该系统选用两组数据采集器，一组数据采集器通过有线连接 7 个水分传感器，另一组数据采集器通过有线连接 2 个传感器型电子鼻。一组作为主数据采集器，另一组作为扩展数据采集器。主采集器通过 RS232 串口直接获取扩展采集器获取的葡萄酒糖分传感器数据。主数据采集器与 GPRS 无线传输模块通过 RS232 串口线相连，远程 PC 端通过无线网络在主采集器中下载并实时监控葡萄酒的糖分数据。PC 端根据传感器采集到的数据进行数据分析，并与标准的葡萄酒成分信息进行对比与匹配，进而用户可根据最终的分析数据在酿酒时对葡萄酒的成分进行及时调整。该系统的采集监控结构设计如图 1 所示。

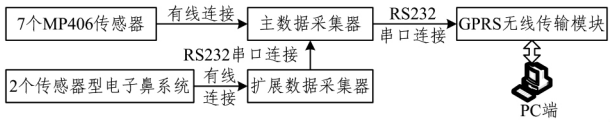

图 1 葡萄酒信息数据采集监控结构图

本实验通过水分传感器和传感器型电子鼻可采集到葡萄酒酿制过程中的水分含量、葡萄酒的糖分数据。其中，7 个水分传感器用来测量葡萄酒酿制过程中的水分含量，通过有线连接传输到主数据采集器，而 2 个传感器型电子鼻用来测量葡萄酒的糖分数据，通过有线连接传送到扩展数据采集器，进而通过 RS232 串口连接传送到主数据采集器，主数据采集器通过 RS232 串口连接传送到 GPRS 无线传输模块，无线传输模块再将数据传输到 PC 端。其中，PC 端葡萄酒信息数据分析系统的功能结构图如图 2 所示。

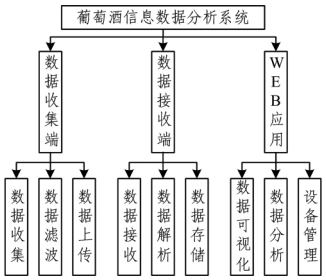

图 2 葡萄酒信息数据分析系统的功能结构图

该系统由 3 个模块构成，即数据收集端、数据接收和 Web 应用。其中数据收集端包括：数据收集、数据滤波和数据上传；数据接收端包括：数据接听、数据解析和存储数据；Web 应用包括：数据可视化、数据分析和设备管理。

2 分类算法的改进

2.1 SVM 算法的分析及改进

2.1.1 SVM 算法的简介

SVM 方法[13]通过非线性映射，将样本空间映射到一个高维的特征空间，从而使得原来的样本空间中的非线性可分的问题转化为在特征空间中的线性可分问题。支持向量机可简单描述为升维和线性化升维。一般的升维会使得计算的复杂性发生变化，SVM 方法巧妙地解决了该难题：应用核函数[15]的展开定理，即可不需要知道非线性映射的显示表达式，由于在高维特征空间中建立线性学习机，因此与线性模型相比，SVM 方法不仅几乎不增加计算的复杂性，而且在某种程度上避免了“维数灾难”。

支持向量机可用分类问题即寻找一个最优分类超平面，把此平面作为分类决策面。同时通过引进核函数很好地解决了在将低维空间向量映射到高维空间向量时带来的“维数灾难”。

2.1.2 SVM 算法的改进

SVM 算法虽然有较强的泛化能力，但当两类样本集混叠较严重时，SVM 的决策面可能由于过于复杂而降低了其泛化能力。

针对上述问题，本文给出一种改进的 SVM-NSVM：先对训练集进行修剪，根据每个样本与其最近邻类标的异同决定其取舍，然后再用 SVM 训练得到分类器。这种方法简单快捷，在分类正确率和分类的速度上均有较大的提高，并且也能适用于更大规模的训练样本集。

本文采取下面的策略对训练集进行修剪：首先找出每一个点的最近邻，然后对每一个点进行判断，如果该点与其最近邻属于同类，则保留此点；否则，删除该点。

采用马氏距离[16]计算两个向量之间的距离，对多元正态随机变量$(x_1,x_2,\cdots,x_n)$的密度函数为：

$$f(x_1,x_2,\cdots,x_n)=\frac{|\sum|^{-1/2}\mathrm{e}^{-1/2(x-u)'\Sigma^{-1}(x-u)}}{(2\pi)^{-n/2}} \tag{1}$$

其中，u 和 Σ 是样本总体的均值向量和协方差矩阵 $x=(x_1,x_2,\cdots,x_n)$。

对服从同一分布且协方差矩阵为 S 的两个随机向量 x，$y\in R^m$，其马氏距离为：

$$d(x,y)=\|x-y\|_M=\sqrt{(x-y)^{\mathrm{T}}S^{-1}(x-y)} \quad (2)$$

其中，$S=R^{m*n}$ 为随机变量的正定协方差矩阵。

下面是上述方法的实现算法。

给定一个训练集 $(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)$，$x_i\in R^n$，$y_i\in\{1,-1\}$，$i=1,2,\cdots,m$。将训练集表示为矩阵 $TR_{m\times(n+1)}=[XY]$，其中 $X=\begin{bmatrix}x_1\\ \cdots\\ x_m\end{bmatrix}$，$Y=\begin{bmatrix}y_1\\ \cdots\\ y_m\end{bmatrix}$。

修剪算法可表示如下：

(1)找出每一个向量的最近邻。

1)求出每个点与其他各点的距离，其到自身的距离可表示为∞；

2)找出最短距离和最近邻。

(2)判断每个向量的类标与其最近邻是否一致，如果一致则标记为 1，否则标记为－1。

(3)删除与最近邻类标不一致的向量。

经过以上步骤即可得到修剪后的训练集 TR。先利用最近邻对训练集进行修剪，再利用 SVM 训练得到分类器的方法称为 NSVM。

2.2　kNN 算法的介绍与改进

2.2.1　kNN 算法的简介

在模式识别中，kNN 算法[14]是将特征空间中最接近的训练样本进行分类的方法。

kNN 算法是一种比较适合于文本分类的分类算法。其核心思想是如果一个样本在特征空间中的 k 个最相邻样本的大多数属于某一个类别，则该样本也属于该类别，并具有类别中样本的特性。该方法在确定分类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

2.2.2　kNN 算法的改进

针对 kNN 算法[12]在分类阶段计算复杂度大、耗时长的缺点，本文给出一种改进的 kNN 算法。该算法的主要思想是：在训练阶段，首先计算训练集中两两向量之间的距离，然后对这些距离进行排序，针对每一个训练向量，设计一种独特的数据结构形式，将到该向量距离最短的前 K 个向量的集合和其他一些信息存储其中。在分类阶段，对于新到的待分类的测试向量，在训练集中随机性地找出一个训练向量，计算出该向量以及其数据结构中离该向量距离最短的 K 个向量到待分类向量的距离最短的向量。再以此向量为基点，计算出该向量以及其数据结构中离该向量距离最短的 K 个向量到测试向量的距离。如果已经计算过距离的向量，则不做重复计算。迭代执行，直到找到与测试向量的距离最短的前 K 个向量为止。再按照传统的 KNN 算法，计算出待分类的测试向量的所属类别。

3　数据挖掘算法在葡萄酒信息数据分析系统中的应用研究

优质的葡萄酒不仅来源于优质的葡萄原料，其酿造过程中的合理有效控制也扮演着举足轻重的分量，此外酿造过程的管控也是现今生产厂商很难把控的一个环节。基于该问题，我们自主开发了一个葡萄酒信息数据分析系统。该系统利用机器学习的数据挖掘算法对极大量的葡萄酒信息数据进行采集、分析、对比匹配，从而挖掘出葡萄酒的主要成分信息和营销潜在信息等，再对这些成分进行相应的分析，并与高质量葡萄酒中的成分进行相应的对比，最终得到葡萄酒的相关分析信息数据，以帮助葡萄酒生产厂商更好地获取葡萄酒的成分含量、葡萄酒的品质、消费者的喜好等分析信息，从而更好地评估与提高葡萄酒的质量。

3.1　实验准备

本文采用 UCI(University of CaliforniaIrvine)开源的数据集作为实验数据集，该实验主要分析了葡萄酒在酿制过程中糖分和水分含量的变化情况。现今一些葡萄酒生产厂商在生产葡萄酒的过程中由于所使用的原材料及生产车间受到气候、光照等相关因素的影响，因此不同的酒庄在酿造葡萄酒的过程中糖分和水分含量会出现一定的差异。本文实验采用的数据集中的葡萄酒的成分含量信息如表 1 所列，大部分酒庄在进行葡萄酒酿造过程中的标准成分数据信息如表 2 所列。

表 1　葡萄酒酿造过程中的标准成分信息数据

成分	含量/mg/L	比例/%
水分	75.5	68.6
糖分	4.0	3.6
其它	20.5	27.8

表 2　酒庄葡萄酒酿造过程中的成分信息数据

成分	含量/mg/L	比例/%
水分	75.3	68.5
糖分	4.2	3.8
其它	20.5	27.7

通过对本文使用的数据集中的数据进行整理，并计算出葡萄酒中糖分和水分的信息百分比含量，为系统的下一步分析做出了准备。

从表 2 可以观察到，葡萄酒的成分中水分和糖分数据占有很大的比重，其中水分的含量最高，根据糖分的含量可以判断该酒庄生产的葡萄酒是干型葡萄酒。

3.2　结果与分析

3.2.1　数据的分析

该系统主要通过数据挖掘算法中的分类算法来对该系统采集到的酿制葡萄酒过程中的信息数据进行分析，本文采用的分类算法是 SVM 算法和 kNN 算法。针对 SVM 泛化能力弱的缺点，提出了一种改进的 SVM-NSVM；针对 kNN 训练数据集大的缺点，给出一种改进的渐进的思想来寻找最近邻点。

(1)NSVM 算法与 SVM 算法的分析比较

改进的 SVM 算法首先通过对训练集进行修剪，根据每个样本与其最近邻类标的异同决定取舍，然后再用 SVM 训练得到分类器。实验表明，NSVM 比 SVM 在分类正确率、分类速度以及适用的样本规模上都表现出了一定的优越性。实验中采用的数据集为 UCI 开源的数据集，使用高斯核作为核函数，其中 $\sigma=0.5$，惩罚参数 $C=100$。该实验的数据集为 500 个 5 维向量的集合，每个向量附带一个类标。截取 5 维向量的前 K 个分量构成 K 维向量，取前 250 个样本作为训练集，后 250 个样本作为测试集。图 3 为 NSVM 与 SVM 的分类正确率的对比结果，可以看到 NSVM 的分类正确率有明显的提高。

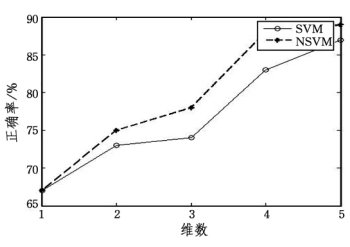

图 3　不同维数下 SVM 与 NSVM 的分类正确率比较(样本数为 500)

通过图 3 可以看到，改进的 SVM 算法较传统的 SVM 在分类正确率方面有更好的效果，因此改进的算法效果明显，具有很好的合理性。

(2)改进的 kNN 算法与传统 kNN 算法的分析比较

改进的 kNN 算法在对最近邻的选择过程中，放弃了传统算法中遍历所有样本的做法，而是通过渐进的思想来寻找最近邻点。通过实验可以看到该算法在保持与传统的 kNN 算法几乎一样的精度性能的前提下，可以明显降低算法的计算复杂度，并且可降低时间开销。实验中分别采用传统的 kNN 算法和改进的 kNN 算法对数据源进行了对比实验。实验所采用的数据来自传感器采集到的酿制葡萄酒过程中的文本信息数据，首先针对不同 k 值，对分类的效果进行了对比实验，实验结果如图 4 所示；此外在不同的 k 值下，对两种算法的时间消耗进行了对比实验，实验结果如表 3 所列。

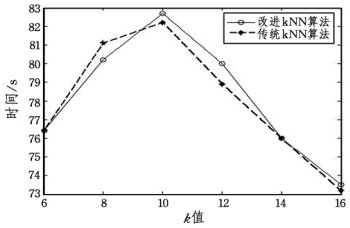

图 4 不同 k 值下的分类准确率对比

表 3 传统 kNN 算法和改进 kNN 算法的时间消耗比较

k	传统 kNN 算法消耗时间/s	改进 kNN 算法消耗时间/s
6	475	263
8	691	345
10	824	420
12	995	456
14	1187	563
16	1369	613

通过对比实验可以看出，不同的 k 值对分类的各项性能的影响明显不同。从图 4 可以看到，当 k 值为 10 时，准确率达到最高值。

通过该对比实验可以看出，在 k 的不同取值下，改进的 kNN 算法在时间消耗上有明显的优势，即该算法是可行的，可以很好地解决传统 kNN 算法计算时间消耗大的问题。

(3)改进算法与传统算法在该系统应用的分析比较

首先将传统分类算法中的 SVM 算法和 kNN 算法应用于自主开发的葡萄酒信息数据分析系统，并针对这两个算法各自的缺点进行改进；其次将改进的算法应用于该系统，改进的算法和传统算法的运行时间对比如图 5 所示。

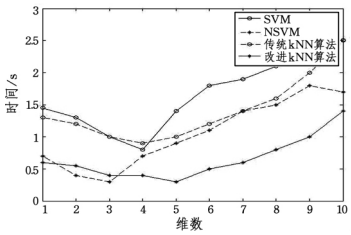

图 5 传统的分类算法与改进的分类算法在该系统的运行时间比较

从图 5 中可以看出，改进的算法较传统的算法在运行时间上明显降低，且随维数的增高，算法的运行时间也大致呈上升的趋势。

结束语 1)本文针对数据挖掘中常用的机器学习算法进行相关的研究，以机器学习中的分类算法为例进行数据挖掘技术的研究，并将该类算法运用到葡萄酒信息数据分析系统中，取得了很好的效果。

2)针对 SVM 泛化能力弱的缺点，给出一种改进的 SVM-NSVM，即先对训练集进行精选，根据每个样本与最近邻类标的异同判断取舍，然后再用 SVM 训练得到分类器。通过实验表明，改进方法在分类正确率、分类速度以及适用的样本规模上都表现出了一定的优越性。此外，针对 kNN 训练数据集大的缺点，给出一种改进的通过渐进的思想来寻找最近邻点。实验表明，改进的 kNN 算法较传统的 kNN 算法在精度性能几乎一样的前提下可明显降低算法的计算复杂度，也可降低时间开销。

3)文中给出一种葡萄酒信息数据分析系统，利用数据挖掘方法对极大量的葡萄酒信息数据进行采集、分析、对比与匹配，从而挖掘出葡萄酒的主要成分对比信息和营销潜在信息等，再对这些成分进行相应的分析，并与高质量葡萄酒中的成分进行相应的对比，最终得出葡萄酒的相关分析信息数据，其可帮助葡萄酒生产厂商更好地获取葡萄酒的成分含量、葡萄酒的品质、消费者的喜好等分析信息，从而更好地评估与提高葡萄酒的质量。

参 考 文 献

[1] ESCUDERO A，GOGORZA B，MELUSA M A，et al. Characterization of the aroma of a wine from maccabeo. Key role played by compounds with low odor activity values[J]. Journal of Agricultural and Food Chemistry，2004，52(11)：3516-3524.

[2] 李华. 葡萄酒的原产地域命名与感官评价[C]// 第二届国际葡萄与葡萄酒学术研讨会. 2001：29-34

[3] 李华. 中国葡萄酒原产地域产品命名系统[J]. 酿酒科技，2001，104(2)：63-68.

[4] NOBLE A C，ARNOLD R A，BUECHSENSTEIN J，et al. Modification of a standardized system of wine aroma terminology[J]. American Journal of Enology and Viticulture，1987，38(2)：143-146.

[5] DANZART M，SIEFFERMANN J M. Analyse sensorielle et mise en place dun laboratoir[J]. Revue Des Oenologues，2001，97(5)：31-35.

[6] 李华，刘勇强，梁新红，等. 运用多元统计分析确定葡萄酒感官特性的描述符[J]. 中国食品学报，2007，7(4)：114-119.

[7] 齐桂梅. 葡萄与葡萄酒中的酚元化合物[J]. 葡萄栽培与酿酒，1992，61(2)：41-42.

[8] FOW LES G W A. Acids in grapes and wines：Areview[J]. Journal of wine research，Grnande-Bretagne，1992，32(3)：25-411.

[9] 王运照，胡文忠，李婷婷，等. 冰葡萄酒酿造过程中糖分与乙醇变化的研究[J]. 食品工业科技，2015，10(17)：142-145.

[10] 李华，王庆伟，刘树文，等. 智能系统在葡萄酒产业中应用的研究进展[J]. 农业工程学报，2006，22(7)：193-199.

[11] 李艳芳. 基于多 Agent 系统的 Web 数据挖掘技术[J]. 计算机工程与设计，2007，28(6)：1267-1272.

[12] 王爱平，徐晓艳，李仿华，等. 基于改进 KNN 算法的中文文本分类方法[J]. 微型机与应用，2011，30(18)：8-13.

[13] 张巍，张功萱，王永利，等. 基于 CUDA 的 SVM 算法并行化研究[J]. 计算机科学，2013，40(4)：69-73.

[14] 熊亚军，廖晓农，李梓铭，等. KNN 数据挖掘算法在北京地区霾等级预报中的应用[J]. 气象，2015，41(1)：98-104.

[15] 丁世飞，齐丙娟，谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报，2011，40(1)：2-10.

[16] 黄飞，周军，卢晓东. 基于马氏距离的一维距离像识别算法仿真[J]. 计算机仿真，2010，27(3)：31-34.