Advances in Applied Mathematics 应用数学进展, 2019, 8(1), 41-49
Published Online January 2019 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2019.81005

Hans汉斯

Research on Security Selection by Naive Bayes Classifier Based on a New Feature Selection Method

Panpan Guo, Haijun Liu, Shuangshuang Li

School of Mathematics and Statistics, Zhengzhou University, Zhengzhou Henan
Email: 15690875640@126.com

Received: Dec. 18th, 2018; accepted: Jan. 2nd, 2019; published: Jan. 9th, 2019

Abstract

In this paper, a naive Bayes classifier for securities selection based on a new feature selection method is established. Firstly, in consideration of the trading data of 50 companies in Shenzhen Stock Exchange and 18 commonly used indicators, a new feature selection method, *i.e.* the combination of mutual information and principal component analysis, is adopted to select the value factors for classification. Secondly, a naive Bayes classifier is constructed with the data of the first 10 months, and the prediction accuracy of the classifier is tested with that of the last two months. The empirical analysis shows that the average accuracy of the classifier reaches 75%, which is of application value.

Keywords

Feature Selection, Mutual Information, Principal Component Analysis, Naive Bayes Classifier

基于一种新的特征选择方法的朴素贝叶斯分类器选择证券的研究

郭盼盼，刘海军，李双双

郑州大学，数学与统计学院，河南 郑州
Email: 15690875640@126.com

收稿日期：2018年12月18日；录用日期：2019年1月2日；发布日期：2019年1月9日

文章引用: 郭盼盼, 刘海军, 李双双. 基于一种新的特征选择方法的朴素贝叶斯分类器选择证券的研究[J]. 应用数学进展, 2019, 8(1): 41-49. DOI: 10.12677/aam.2019.81005

摘　要

本文提出了基于一种新的特征选择方法的朴素贝叶斯证券分类模型。首先，根据深交所50家公司2011年的交易数据和常用的18个指标，采取新的特征选择方法即互信息和主成分分析相结合选出用于分类的因子；其次，利用前10个月的数据建立朴素贝叶斯分类模型，用后两个月的数据检验模型的预测精度。实证分析表明模型的分类平均正确率达到75%，具有应用价值。

关键词

特征选择，互信息，主成分分析，朴素贝叶斯分类器

1. 引言

证券投资是金融研究领域的热门话题，如何选择证券是投资决策的关键。尽管投资者的盲目任意性和股票市场中的严重非线性给股票的预测与选择带来了很大的困难，事实表明，股票收益在一定程度上还是可以预测的[1]。有不少人尝试关于数据挖掘技术比如决策树[2] [3] [4]、分类器[5] [6] [7]及神经网络[8] [9] [10] [11]等选股的研究。

钱颖能和胡运发[5]使用 2002 年至 2004 年上海证券交易所的中报和年报的财务信息，利用朴素贝叶斯分类法对由超越市场指数而得到额外汇报的股票进行选择，结果表明朴素贝叶斯分类法在股票选择方面很有效；左辉和楼新远[6]使用证券分析师推荐的股票数据并从中选取 2007 年 1 月 8 日到 2007 年 10 月 29 日的数据，用“事件研究”方法分析其总体特征，寻找符合特征的股票以求得到超额回报，然后用朴素贝叶斯分类法选股。结果表明朴素贝叶斯分类法在股票的短线操作上有实用价值。骆桦和张喜梅[7]对沪深证券市场的能源股通过聚类分析选出对股票投资价值影响显著的财务指标构造样本特征集，再合理选取贝叶斯分类器的参数对股票分类。结果产生了 44.6%累计回报率，优于 32.4%的基准回报率。结果表明朴素贝叶斯分类法选股有较好的效果。如果利用不同的方法从较多的特征中筛选出有价值的特征，或许会得到更好的效果。

本文提出了基于一种新的特征选择方法的朴素贝叶斯证券分类模型，并且对深交所 50 家公司 2011 年的交易数据利用该模型分类，实证分析表明模型的平均正确率达到 75%，具有应用价值。

2. 预备知识

2.1. 特征选择

数据集中包含大量的特征，特征维度越高，计算越复杂，且其中包含的不相关特征和冗余特征会影响分类精度。特征选择可以定义为从原始 N 个特征中选出 M 个有价值特征的过程。特征选择方法可分为过滤式[12]、封装式[13]和混合式[14]方法。过滤式方法独立于分类算法评估选取的特征的质量，封装式方法需要用分类器来评估这种质量，混合式方法是前两种方法的结合。

2.2. 互信息

互信息是信息论里一种描述变量间相关性的信息度量。互信息的大小表示变量间包含共同信息的多少，变量耦合越强，互信息越大[15]。互信息对变量的分布类型没有要求，能够描述变量间的线性及非线性相关关系，故在变量选择中得到了广泛应用[16] [17]。

设两个离散随机变量 X 和 Y，$p(x,y)$ 是 X 和 Y 的联合概率分布函数，$p(x)$ 和 $p(y)$ 分别是 X 和 Y 的边缘概率分布函数，根据互信息理论[18]，随机变量 X 的熵 $H(X)$ 表示随机变量 X 的不确定度，可以定义为：

$$H(X)=-\sum_{x\in X}p(x)\log p(x) \tag{1}$$

条件熵 $H(X|Y)$ 表示在 Y 已知的条件下 X 的不确定度，可以定义为：

$$H(X|Y)=-\sum_{y\in Y}\sum_{x\in X}p(x,y)\log p(x|y) \tag{2}$$

互信息表示不确定度的减少量，可以用熵定义为：

$$I(X;Y)=H(X)-H(X|Y) \tag{3}$$

当 X 和 Y 相互独立即没有相关关系时，互信息为 0；当 X 和 Y 不相互独立即有相关关系时，互信息为正数，且相关性越强，互信息越大。

2.3. 主成分分析

2.3.1. 基本概念

主成分分析是将多指标化为少数几个综合指标的一种统计分析方法[19]。

定义 1 [19] 设某样本包含 p 个变量，分别用 $X_1,X_2,\cdots,X_p$ 表示，构成 p 维随机向量 $X=(X_1,X_2,\cdots,X_p)'$，其中均值为 μ，协方差矩阵为 Σ，称 $Y_i=a_i'X\quad(i=1,2,\cdots,p)$ 为 X 的第 i 主成分，如果：

1) $a_i'a_i=1\quad(i=1,2,\cdots,p)$；

2) 当 $i>1$ 时，$a_i'\Sigma a_j=0\quad(j=1,2,\cdots,i-1)$；

3) $Var(Y_i)=\max\limits_{a'a=1,a_i'\Sigma a_j=0(j=1,\cdots,i-1)}Var(a'X)$。

定义 2 [19] 设随机向量 $X=(X_1,X_2,\cdots,X_p)'$ 的协方差矩阵为 Σ，$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq 0$ 为 Σ 的特征值，$a_1,a_2,\cdots,a_p$ 为相应的单位正交特征向量，则 X 的第 i 个主成分为：

$$Y_i=a_i'X=a_{1i}X_1+a_{2i}X_2+\cdots+a_{pi}X_p\quad(i=1,2,\cdots,p) \tag{4}$$

定义 3 [19] 称 $\lambda_k\Big/\sum_{i=1}^{p}\lambda_i$ 为主成分 Y_k 的贡献率，称 $\sum_{k=1}^{m}\lambda_k\Big/\sum_{i=1}^{p}\lambda_i$ 为主成分 $Y_1,Y_2,\cdots,Y_m\ (m<p)$ 的累计贡献率。

2.3.2. 具体步骤

1) 用 Z-score 法对数据进行标准化变换

2) 求指标数据的相关矩阵

3) 求相关矩阵的特征根与特征向量

4) 计算主成分贡献率及累计贡献率，确定主成分(一般取累计贡献率为 85%~95%的特征值所对应的主成分。)

2.4. 朴素贝叶斯分类器

2.4.1. 基本概念

贝叶斯分类是一种可以预测给定样本属于某个特定类的概率的统计学分析方法。贝叶斯分类技术通过对已分类的样本子集进行训练，学习归纳出分类函数，利用训练得到的分类器实现对未分类数据的分类。其中朴素贝叶斯分类器是解决相应问题的最实际的方法之一。朴素贝叶斯分类基于一个简单的假设：给定目标值的属性值之间相互条件独立[20]。朴素贝叶斯分类器的原理是：给定待分类项，利用贝叶斯公式求解在此项出现的条件下各个类别出现的概率，哪个概率最大，就认为此待分类项属于哪个类别。

2.4.2. 朴素贝叶斯分类器

设研究对象的属性值为 $X=\left(x_1,x_2,\cdots,x_n\right)$，而目标值的属性值为 $Y=\left(y_1,y_2,\cdots,y_n\right)$，假设有 m 个类 $v_1,v_2,\cdots,v_m$。分类器考虑类的集合 m 并在其中寻找给定属性值 $X=\left(x_1,x_2,\cdots,x_n\right)$ 时可能性最大的类 $j\in m$，这种分类方法称为极大后验(*MAP*)分类，即：$v_{MAP}=\underset{v_j\in\{v_1,v_2,\cdots,v_m\}}{\arg\max}\ P\left(v_j\left|x_1,x_2,\cdots,x_n\right.\right)$，属性值已知的条件下极大后验分类就是 m 个类中概率最大的一类。利用贝叶斯公式将其整理为

$$v_{MAP}=\underset{v_j\in\{v_1,v_2,\cdots,v_m\}}{\arg\max}\frac{P\left(x_1,x_2,\cdots,x_n\left|v_j\right.\right)P\left(v_j\right)}{P\left(x_1,x_2,\cdots,x_n\right)} \tag{5}$$

$$=\underset{v_j\in\{v_1,v_2,\cdots,v_m\}}{\arg\max}\ P\left(x_1,x_2,\cdots,x_n\left|v_j\right.\right)P\left(v_j\right) \tag{6}$$

其中，$P\left(x_1,x_2,\cdots,x_n\left|v_j\right.\right)=\prod_i P\left(x_i\left|v_j\right.\right)$。在条件独立假设成立时，朴素贝叶斯分类等于极大后验分类，因而可得到朴素贝叶斯分类器的公式：

$$v_{NB}=\underset{v_j\in\{v_1,v_2,\cdots,v_m\}}{\arg\max}\ P\left(v_j\right)\prod_i P\left(x_i\left|v_j\right.\right) \tag{7}$$

如果类的先验概率 $P\left(v_j\right)$ 未知，则通常假设各类的先验概率相等，即：$P\left(v_1\right)=P\left(v_2\right)=\cdots=P\left(v_m\right)$。概率 $P\left(x_j\left|v_i\right.\right)$ 可以由训练样本来估计。这里用 m-估计 $P\left(x_i\left|v_j\right.\right)=\left(n_{ji}+mp\right)/\left(n_j+m\right)$ 来估计。其中，n_{ji} 是对应属性具有值 x_i 的类 v_j 的训练样本数，而 n_j 是类 v_j 的训练样本总数。p 所求概率的先验估计，m 为等效样本大小的常量。

3. 数据，指标与因子

3.1. 数据

本文所用数据来自于锐思数据库。选取深圳证券交易所 50 只 2011 年 1 月 4 日至 2011 年 12 月 31 日股票，对数据中进行简单的预处理，主要包括补全数据和复权。

3.2. 指标

所选指标有：Z_1：前收盘价、Z_2：收盘价、Z_3：开盘价、Z_4：最高价、Z_5：最低价、Z_6：成交额、Z_7：成交量、Z_8：中价、Z_9：5 日收盘价均值、Z_{10}：5 日成交额均值、Z_{11}：5 日成交量均值、Z_{12}：买卖指标 *AR*、Z_{13}：意愿指标 *BR*、Z_{14}：随机指标 *K*、Z_{15}：*D*、Z_{16}：*J*、Z_{17}：相对强弱指标 *RSI*、Z_{18}：日换手率。

3.2.1. 股票收益率

本文中的股票收益率是对数收益率，在 $\left[T,T+\Delta t\right]$ 内的计算公式为：

$$R_{i,T}=\ln\left(\frac{P_{i,T+\Delta t}+I_{i,T+\Delta t}}{P_{i,T}}\right) \tag{8}$$

其中，$R_{i,T}$ 是股票 i 在 T 时刻的收益率，$P_{i,T}$ 是股票 i 在 T 时刻的价格，$P_{i,T+\Delta t}$ 是股票 i 在 $T+\Delta t$ 时刻的价格，$I_{i,T+\Delta t}$ 是股票 i 在 $[T,T+\Delta t]$ 内的分红。

3.2.2. 日换手率

换手率也成周转率，指在一定时间内市场中股票转手买卖的频率。日换手率是指某一个交易日中某支股票当日的日成交量初一该股的流通股本，即换手率 = 某一段时期内的成交量/发行总股数 × 100%。

3.3. 因子的选取

计算每个原始指标与收益率之间的互信息。为了方便对股票数据的调用，本文按 1.xls--50.xls 的形式对存放数据信息的 Excel 表进行命名，借助 MATLAB 软件，通过编写程序一次性计算得到 50 家股票的这 18 个指标与收益率之间的互信息。其中前 5 家公司的结果如表 1 所示：

Table 1. Mutual information outcomes of the top five companies
表 1. 前 5 家公司的互信息

	表格 1	表格 2	表格 3	表格 4	表格 5
$I(Z_1;Z_{19})$	1.110359668	1.099889033	1.089842697	1.125920538	1.109784047
$I(Z_2;Z_{19})$	1.10881697	1.117758017	1.098010072	1.124252332	1.117001218
$I(Z_3;Z_{19})$	1.127126861	1.131420511	1.098508071	1.126748527	1.12507881
$I(Z_4;Z_{19})$	1.117911243	1.124494251	1.098405772	1.125080478	1.125917003
$I(Z_5;Z_{19})$	1.124473626	1.124494251	1.096169704	1.125080478	1.125917003
$I(Z_6;Z_{19})$	1.122911464	1.120409902	1.096169704	1.125920538	1.129343262
$I(Z_7;Z_{19})$	1.116737838	1.11559817	1.096650386	1.125920538	1.103221733
$I(Z_8;Z_{19})$	0.474802198	0.471361895	0.811832636	0.494832378	0.389729731
$I(Z_9;Z_{19})$	0.260569554	0.238567169	0.461843167	0.216786665	0.171769686
$I(Z_{10};Z_{19})$	0.414676474	0.388254771	0.73972281	0.44846479	0.325828778
$I(Z_{11};Z_{19})$	0.32151623	0.286835637	0.586627111	0.263599006	0.191785667
$I(Z_{12};Z_{19})$	0.091289193	0.026499282	0.17527769	0.076333218	0.065591362
$I(Z_{14};Z_{19})$	0.026951949	0.035696271	0.107708479	0.017773471	0.025841353
$I(Z_{15};Z_{19})$	0.035095403	0.038838678	0.110865222	0.021354449	0.030282553
$I(Z_{16};Z_{19})$	0.019787418	0.031444862	0.095732429	0.014057081	0.01621386
$I(Z_{17};Z_{19})$	0.072058185	0.054467933	0.221375307	0.080950792	0.089453431
$I(Z_{18};Z_{19})$	0.266973154	0.238567169	0.461843167	0.216786665	0.155561539

表中 $I\left(Z_j;Z_{19}\right)\left(j=1,2,\cdots,18\right)$ 表示第 j 个指标与收益率之间的互信息。从上述结果可以看出，5 个表格均显示指标 $Z_1,Z_2,Z_3,Z_4,Z_5,Z_6,Z_7$ 与收益率之间的互信息都大于 1.5，指标 $Z_8,Z_9,Z_{10},Z_{11},Z_{18}$ 与收益率之间的互信息都介于 0.1 和 1.0 之间，而有 4 个表格显示指标 $Z_{12},Z_{13},Z_{14},Z_{15},Z_{16},Z_{17}$ 与收益率之间的互信息都小于 0.1，只有表格 3 显示指标 $Z_{12},Z_{13},Z_{14},Z_{15},Z_{16},Z_{17}$ 与收益率之间的互信息介于 0.1 和 1.0 之间，此种情况占比不大，对指标的选择影响不大。故可以认为指标 $Z_1,Z_2,Z_3,Z_4,Z_5,Z_6,Z_7,Z_8,Z_9,Z_{10},Z_{11}$ 对收益率有显著影响，指标 $Z_{12},Z_{13},Z_{14},Z_{15},Z_{16},Z_{17}$ 对收益率无显著影响。此 5 个表格的结果可以反应整体情况，因此，选出 $Z_1,Z_2,Z_3,Z_4,Z_5,Z_6,Z_7,Z_8,Z_9,Z_{10},Z_{11},Z_{18}$ 作为主成分分析的指标。

对选出的指标进行主成分分析。以“东旭蓝天”即表格 40 为例，说明主成分分析过程。以预处理后的标准数据矩阵作为原始数据矩阵，计算其相关阵并绘制特征值图，如图 1 所示：

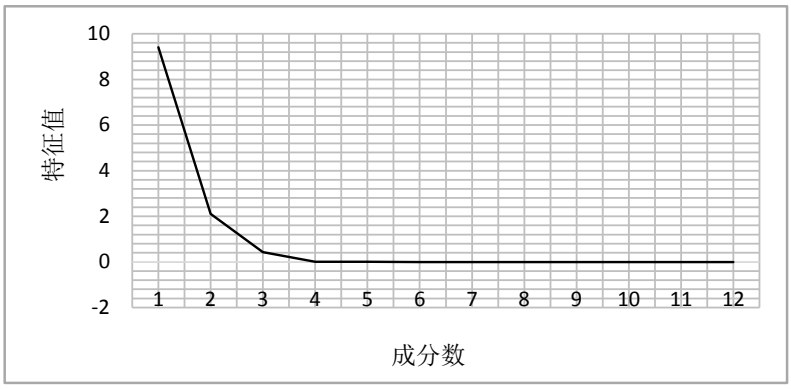

Figure 1. Eigevalues
图 1. 特征值

由图 1 可以看出，第二个主成分的变化趋势开始减慢，因此可以只考虑前两个主成分反映原有信息。相关阵的前两个特征值、对应的特征向量、贡献率及累计贡献率如表 2。

Table 2. Principal component result
表 2. 主成分结果

	主成分 Y_1	主成分 Y_2
Z_1：前收盘价	0.298643431	0.206278919
Z_2：开盘价	0.299813199	0.176391377
Z_3：最高价	0.300701394	0.170618405
Z_4：最低价	0.295734648	0.244725169
Z_5：收盘价	0.295985477	0.235804528
Z_6：中价	0.298082313	0.21570945
Z_7：5 日收盘价均值	0.294324182	0.257869676
Z_8：5 日成交金额均值	0.293524812	−0.098104744
Z_9：成交量	0.261981139	−0.493856614
Z_{10}：成交金额	0.286372458	−0.239864428
Z_{11}：5 日成交量均值	0.273257891	−0.336067968
Z_{18}：换手率	0.261699545	−0.494679098
特征值	9.411734966	2.111541604
贡献率	0.784311247	0.1759618
累计贡献率	0.784311247	0.96027304

从表 2 中可以看出，前两个主成分的累计贡献率已经达到了 96%，因此，只取前两个主成分，分别为：

$$Y_1 = 0.298643431Z_1 + \cdots + 0.273257891Z_{11} + 0.261699545Z_{18}$$
$$Y_2 = 0.206278919Z_1 - \cdots - 0.336067968Z_{11} - 0.494679098Z_{18}$$

4. 构建朴素贝叶斯分类器

将利用主成分分析所得的 2 个主成分 Y_1、Y_2 和收益率 R 训练朴素贝叶斯分类器规则。用均匀聚类法将主成分 Y_1、Y_2 离散化，分成 3 个类，将指标 R 离散化，分成 4 个类。因此，该分类器有 4 类，v_1 为低收益率类，v_4 为高收益率类，v_2, v_3 为普通收益率类；有 2 个样本，每个样本有 3 个属性值。以表格 40 为例，2011 年 11 月 9 日的主成分 Y_1、Y_2 和收益率 R 的值分别为−1.6887、1.3689、4.47，离散化后为 1、3、2。

将前 10 个月的数据作为训练集训练分类规则，后 2 个月的数据作为测试集检验分类规则的预测精度，通过 Matlab 软件编写程序一次性计算得到 50 家股票的分类结果。如表 3 所示。

Table 3. Classification Situation 1
表 3. 分类结果 1

表格	训练集正确率	测试集正确率	表格	训练集正确率	测试集正确率	表格	训练集正确率	测试集正确率
1	0.770491803	0.918918919	18	0.628415301	0.783783784	35	0.677595628	0.702702703
2	0.683060109	0.540540541	19	0.732240437	0.837837838	36	0.726775956	0.945945946
3	0.606557377	0.72972973	20	0.797814208	0.702702703	37	0.672131148	0.864864865
4	0.74863388	0.783783784	21	0.715846995	0.702702703	38	0.699453552	0.702702703
5	0.655737705	0.945945946	22	0.737704918	0.891891892	39	0.765027322	0.486486486
6	0.781420765	0.783783784	23	0.584699454	0.594594595	40	0.841530055	0.972972973
7	0.693989071	0.783783784	24	0.726775956	0.702702703	41	0.677595628	0.648648649
8	0.644808743	0.837837838	25	0.721311475	0.945945946	42	0.743169399	0.459459459
9	0.721311475	0.810810811	26	0.704918033	0.891891892	43	0.754098361	0.648648649
10	0.754098361	0.405405405	27	0.661202186	0.810810811	44	0.770491803	0.675675676
11	0.792349727	0.567567568	28	0.655737705	0.72972973	45	0.721311475	0.675675676
12	0.704918033	0.918918919	29	0.699453552	0.891891892	46	0.765027322	0.648648649
13	0.693989071	0.891891892	30	0.765027322	0.783783784	47	0.759562842	0.972972973
14	0.786885246	0.864864865	31	0.743169399	0.432432432	48	0.732240437	0.648648649
15	0.699453552	0.891891892	32	0.797814208	0.837837838	49	0.737704918	0.567567568
16	0.792349727	0.702702703	33	0.650273224	0.648648649	50	0.710382514	0.783783784
17	0.710382514	0.783783784	34	0.721311475	0.783783784	平均正确率	0.723180903	0.750355619

从上表结果统计得出：利用朴素贝叶斯分类器选股，50 家股票中，训练集正确率在 70%以上且测试集正确率在 40%以上的有 33 家，占比 66%，训练集正确率在 75%以上且测试集正确率在 40%以上的有 12 家，占比 24%。表明朴素贝叶斯分类器选股在一定程度上有很好的效果。

为了证明本文提出的方法更有效，用相同公司的数据，不利用互信息筛选因素只做主成分分析，分类结果如表 4 所示：

从上表结果统计得出：利用朴素贝叶斯分类器选股，50 家股票中，训练集正确率在 70%以上且测试集正确率在 40%以上的有 22 家，占比 44%，训练集正确率在 75%以上且测试集正确率在 40%以上的有 6 家，占比 12%。

从正确分类的比例和平均绝对误差两方面对比基于两种特征选择方法的朴素贝叶斯分类器的分类结果，如表 5 所示：

Table 4. Classification Situation 2
表 4. 分类结果 2

表格	训练集正确率	测试集正确率	表格	训练集正确率	测试集正确率	表格	训练集正确率	测试集正确率
1	0.721311475	0.945945946	18	0.639344262	0.702702703	35	0.693989071	0.837837838
2	0.721311475	0.513513514	19	0.666666667	0.621621622	36	0.710382514	0.891891892
3	0.677595628	0.648648649	20	0.655737705	0.675675676	37	0.639344262	0.486486486
4	0.677595628	0.432432432	21	0.672131148	0.513513514	38	0.595628415	0.864864865
5	0.655737705	0.486486486	22	0.74863388	0.72972973	39	0.628415301	0.594594595
6	0.737704918	0.756756757	23	0.666666667	0.72972973	40	0.803278689	0.756756757
7	0.683060109	0.702702703	24	0.699453552	0.567567568	41	0.732240437	0.675675676
8	0.584699454	0.378378378	25	0.612021858	0.891891892	42	0.775956284	0.918918919
9	0.666666667	0.918918919	26	0.601092896	0.864864865	43	0.704918033	0.432432432
10	0.775956284	0.432432432	27	0.704918033	0.864864865	44	0.737704918	0.945945946
11	0.74863388	0.864864865	28	0.62295082	0.540540541	45	0.726775956	0.675675676
12	0.595628415	0.675675676	29	0.710382514	0.621621622	46	0.775956284	0.648648649
13	0.595628415	0.432432432	30	0.672131148	0.648648649	47	0.666666667	0.540540541
14	0.704918033	0.72972973	31	0.743169399	0.540540541	48	0.704918033	0.702702703
15	0.628415301	0.405405405	32	0.693989071	0.621621622	49	0.710382514	0.702702703
16	0.781420765	0.72972973	33	0.546448087	0.675675676	50	0.655737705	0.432432432
17	0.737704918	0.486486486	34	0.661202186	0.351351351	平均正确率	0.685464481	0.656216216

Table 5. Classification Situation 3
表 5. 分类结果 3

	本文建立的朴素贝叶斯分类器		基于主成分分析的朴素贝叶斯分类器	
	训练集	测试集	训练集	测试集
正确分类的比例	72.30%	75.00%	68.50%	65.60%
平均绝对误差	0.041	0.118	0.047	0.126

从上表可以看出：本文建立的朴素贝叶斯分类器训练集和测试集正确分类的比例均高于基于主成分分析的朴素贝叶斯分类器正确分类的比例，其平均绝对误差均低于基于主成分分析的朴素贝叶斯分类器的平均绝对误差。从对比中得出本文提出的方法的分类结果优于不利用互信息筛选因素只做主成分分析的分类结果。

5. 结论

本文利用证券的交易数据并结合一种新的特征选择方法给出了一种朴素贝叶斯分类模型。实证分析表明：训练集正确率在 70%以上且预测精度在 40%以上的达到 66%，训练集正确率在 75%以上且预测精度在 40%以上的有 12 家，占比 24%。该分类器的平均正确率达到 75%，并且从正确分类样本属性值的比例和平均绝对误差两方面对比，本文提出的方法的分类结果均优于不利用互信息筛选因素只做主成分分析的分类结果。

参考文献

[1] Fama, E.F. and French, K.R. (1992) The Cross-Section of Expected Stock Returns. *The Journal of Finance*, **47**, 427-465. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

[2] 唐文慧. 基于数据挖掘技术的股价预测实证分析[D]: [硕士学位论文]. 成都: 西南财经大学, 2009.

[3] 雷炜, 叶东毅. 利用决策树技术对股票价格数据库进行数据挖掘[J]. 福建电脑, 2004(8): 52-53.

[4] 王领, 胡扬. 基于 C4.5 决策树的股票数据挖掘[J]. 计算机与现代化, 2015(10): 21-24.

[5] 钱颖能, 胡运发. 用朴素贝叶斯分类法选股[J]. 计算机应用与软件, 2007, 24(6): 90-92.

[6] 左辉, 楼新远. 基于贝叶斯分类的选股方法[J]. 电脑知识与技术, 2008, 2(10): 173-176.

[7] 骆桦, 张喜梅. 基于贝叶斯分类法的股票选择模型的研究[J]. 浙江理工大学学报(自然科学版), 2015, 33(3): 418-422.

[8] White, H. (1988) Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns. *IEEE International Conference on Neural Networks*, **2**, 451-458.

[9] Oliveira, F.A.D., Nobre, C.N. and Zárate, L.E. (2013) Applying Artificial Neural Networks to Prediction of Stock Price and Improvement of the Directional Prediction Index—Case Study of PETR4, Petrobras, Brazil. *Expert Systems with Applications*, **40**, 7596-7606. https://doi.org/10.1016/j.eswa.2013.06.071

[10] Zahedi, J. and Rounaghi, M.M. (2015) Application of Artificial Neural Network Models and Principal Component Analysis Method in Predicting Stock Prices on Tehran Stock Exchange. *Physica A Statistical Mechanics & Its Applications*, **438**, 178-187. https://doi.org/10.1016/j.physa.2015.06.033

[11] Qiu, M., Song, Y. and Akagi, F. (2016) Application of Artificial Neural Network for the Prediction of Stock Market Returns: The Case of the Japanese Stock Market. *Chaos*, *Solitons & Fractals*, **85**, 1-7. https://doi.org/10.1016/j.chaos.2016.01.004

[12] Almuallim, H. and Dietterich, T.G. (1991) Learning With Many Irrelevant Features. *Proceedings of the 9th National Conference on Artificial Intelligence*, Anaheim, 14-19 July 1991, AAAI Press, Volume 2.

[13] Domingos, P. and Pazzani, M. (1997) On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, **29**, 103-130. https://doi.org/10.1023/A:1007413511361

[14] Blum, A.L. and Langley, P. (1997) Selection of Relevant Features and Examples in Machine Learning. *Artificial Intelligence*, **97**, 245-271. https://doi.org/10.1016/S0004-3702(97)00063-5

[15] 唐勇波, 桂卫华, 彭涛, 等. 基于互信息变量选择的变压器油中溶解气体浓度预测[J]. 仪器仪表学报, 2013, 34(7): 1492-1498.

[16] 郭伟. 基于互信息的 RBF 神经网络结构优化设计[J]. 计算机科学, 2013, 40(6): 252-255.

[17] 韩敏, 刘晓欣. 基于互信息的分步式输入变量选择多元序列预测研究[J]. 自动化学报, 2012, 38(6): 999-1006.

[18] Cover, T.M. and Thomas, J.A. (1991) Elements of Information Theory. John Wiley & Sons, Inc, New York. https://doi.org/10.1002/0471200611

[19] 何晓群. 多元统计分析[M]. 第二版. 北京: 中国人民大学出版社, 2008.

[20] Tom M. Mitchell, 米切尔, 曾华军, 等. 机器学习[M]. 北京: 机械工业出版社, 2003.