应 用

基于主成分分析的朴素贝叶斯算法在垃圾短信用户识别中的应用

李琼阳, 田 萍

(许昌学院 数学与统计学院, 河南 许昌 461000)

摘 要: 在垃圾短信用户的识别问题中, 参与建模的用户行为消费数据存在极强的相关性, 直接使用朴素贝叶斯算法建模准确率极低. 为满足朴素贝叶斯算法要求建模属性条件独立的基本假定, 利用主成分分析对数据进行处理, 从而达到降维和属性独立的双重目的, 继而利用朴素贝叶斯算法进行建模. 结果表明, 基于主成分分析和朴素贝叶斯算法的组合模型效果显著. 可见在垃圾短信算法的识别中具有一定的实用价值和现实意义.

关键词: 属性独立; 主成分分析; 朴素贝叶斯; 组合模型

1 引言

朴素贝叶斯分类 (Naive Bayes Classification) 算法是目前公认的一种简单有效的分类算法, 它是一种基于概率的分类方法, 被广泛地应用于模式识别、自然语言处理、机器人导航、机器学习等领域. 朴素贝叶斯算法是基于特征项间条件独立的假设, 但实际应用中极少数问题能满足此假设. 为此许多学者致力于改进朴素贝叶斯算法, 以期提高算法的普适性和准确率. 改进之处主要体现在两个方面: 1) 在属性选择上预把控; 2) 衡量各属性对目标变量的影响程度, 对属性加权.

在属性选择上, Geenen P L 等人 [1] 提出一种基于互信息选择特征属性的方法, 并利用朴素贝叶斯算法在二分类问题上取得了极好的分类效果. 魏浩、丁要军 [2] 提出用属性关联度表示一个属性和类属性间的相关性, 反映这个属性对分类结果影响的程度; 用属性冗余度表示一个属性和其他属性之间相关性, 反映这个属性和其他属性间的依赖度. 王行甫、杜婷 [3] 提出利用 CFS 算法选择特征属性. 焦鹏等人 [4] 提出将属性先验分布的参数设置加入到属性选择的过程中, 并研究当先验分布服从 Dirichlet 分布及广义 Dirichlet 分布情况下的具体实践方案. 研究出一种加权朴素贝叶斯算法, 通过对不同的特征项提供不同的权值, 削弱特征项之间的相关性.

在属性加权上, 饶丽丽等人 [5] 提出在传统权重计算基础上, 考虑到特征项在类内和类间的分布情况, 另外还结合特征项间的相关度, 调整权重计算值, 加大最能代表所属类的特征项

收稿日期: 2018-01-08
资助项目: 许昌学院科研基金项目 (2019YB029)

的权重. Jiang L 等人 [6] 提出在训练集中深度计算特征加权频率, 估计朴素贝叶斯的条件概率. Wang S 等人 [7] 提出了两种自适应特征加权方法：1) 基于树的自适应特征加权；2) 基于信息增益率的特征加权 [7].

为了满足朴素贝叶斯算法的假定条件, 以期在垃圾短信用户识别的建模过程中取得良好的效果, 本文首先利用主成分分析方法对原始数据进行处理, 以达到降维和属性独立的双重目的, 继而利用朴素贝叶斯算法进行建模. 实证分析表明, 基于 PCA(主成分分析) 的朴素贝叶斯算法在垃圾短信用户识别中的准确率有显著提升.

2 主成分分析

主成分分析法是一种降维的统计方法, 设法将原来变量重新组合成一组新的相互无关的几个综合变量, 同时根据实际需要从中可以取出几个较少的综合变量尽可能多地反映原来变量的信息的统计方法叫做主成分分析或称主分量分析, 也是数学上处理降维的一种方法.

主成分分析算法的计算步骤：

1) 设表示训练样本的属性集, 有 m 个属性 $A_1, A_2 \cdots A_m$, 数据集中样本个数为 $n(n > m)$；数据样本 X_i 用一个 m 维特征向量来描述 m 个属性的值, 即：$x_{i1}, x_{i2} \cdots x_{im}$, 其中 $(x_{ij} \in A_j, 1 \le i \le n, 1 \le j \le m)$. 对样本进行如下标准化变换：

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad , \quad i = 1, 2, \cdots, n; j = 1, 2, \cdots, m \tag{1}$$

其中 $\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$, $s_j^2 = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}{n-1}$ 得标准化矩阵 z；

2) 对标准化矩阵 z 求相关系数矩阵 $R = [r_{ij}]_m$ 其中, $r_{ij} = \sum_{k=1}^{m} z_{ik} * z_{jk}, (1 \le k \le m)$；

3) 解样本相关矩阵 R 的特征方程 $|R - \lambda I_p| = 0$ 得 m 个特征根, 确定主成分

按 $\frac{\sum_{j=1}^{l} \lambda_j}{\sum_{j=1}^{m} \lambda_j} \ge 0.85$ 确定 l 值, 使信息的利用率达 85%以上, 对每个 $\lambda_j, 1 \le j \le m$, 解方程组 $Rb = \lambda_j b$ 得单位特征向量 b_j^0；

4) 将标准化后的指标变量转换为主成分 $U_{ij} = z_i' b_j^0, 1 \le j \le l, U_1$ 称为第一主成分, U_2 称为第二主成分,$\cdots$, U_l 称为第 l 主成分；

5) 对 l 个主成分进行综合评价；

对 l 个主成分进行加权求和, 即得最终评价值, 权数为每个主成分的方差贡献率.

3 朴素贝叶斯算法

朴素贝叶斯算法的分类原理是通过某对象的先验概率, 利用贝叶斯公式计算出其后验概率, 具有最大后验概率的类则为该对象所属的类. 朴素贝叶斯是在贝叶斯分类法的基础上提出的, 该算法满足一个简单的假定, 即在给定目标值时属性值之间相互条件独立.

朴素贝叶斯分类算法的工作过程如下：

1) 设表示训练样本的属性集, 有 m 个属性 $A_1, A_2 \cdots A_m$；C 表示类集合, 有 k 个类 $C_1, C_2, \cdots, C_k$；每个数据样本 X 用一个 m 维特征向量来描述 m 个属性的值, 即：$X = (x_1, x_2, \cdots, x_m)$, 其中 $x_i \in A_i (1 \le i \le m)$.

2) 对训练样本集进行统计, 计算得到每个特征属性在各类别的条件概率估计, 即:

$$P(x_1|C_1), P(x_2|C_1), \cdots, P(x_m|C_1)$$
$$P(x_1|C_2), P(x_2|C_2), \cdots, P(x_m|C_2)$$
$$\vdots$$
$$P(x_1|C_k), P(x_2|C_k), \cdots, P(x_m|C_k)$$

3) 对每个类别计算后验概率, 根据贝叶斯定理及朴素贝叶斯算法的假定可知:

$$P(C_i|X) = \frac{P(C_i) \cdot \prod_{j=1}^{m} P(x_j|C_i)}{P(X)} \tag{2}$$

4) 取最大后验概率项作为样本所属类别:

$$C(X) = \arg \max_{(1\le i\le k)} P(C_i) \cdot \prod_{j=1}^{m} P(x_j|C_i) \tag{3}$$

4 在垃圾短信用户识别中的实证分析

垃圾短信对人们的正常生活造成了极大的干扰, 工信部表示, 手机用户平均每周收到 10.7 条垃圾短信 [8]. 面对如此的境况, 准确高效的识别垃圾短信势在必行. 本文根据某运营商提供的用户消费行为数据, 利用第 2、3 部分介绍的主成分分析和朴素贝叶斯算法建立模型.

4.1 数据的收集和处理

某运营商提供了用户行为消费特征样本数据共 78258 条, 其中垃圾短信用户样本数据 11837 条, 用 1 标识; 正常用户 66421 条, 用 0 标识. 数据集有当月消费额、品牌、通话时长、发送短信条数、短信回复率、账户余额、是否为垃圾短信用户等共有 56 个属性. 下面将利用基于 PCA 的朴素贝叶斯算法来进行建模.

4.2 主成分分析进行数据处理

主成分分析可以达到降维和属性独立的双重目的. 本文利用主成分分析对运营商提供的用户消费行为数据进行建模, 最终将 56 个建模指标压缩为了 18 个综合指标, 各主成分的方差贡献率和累计方差贡献率如表 1 所示.

4.3 朴素贝叶斯算法建模

将原始建模数据利用主成分分析降维后, 变成了维数为 18 的数据集, 同时也保证了各维属性的相互独立性, 满足朴素贝叶斯算法在给定目标变量的前提下属性之间相互独立的基本假定, 现在可以利用朴素贝叶斯算法进行建模.

在数据集中按照 7:3 的比例进行分层随机抽样, 划分训练样本与检验样本. 基于 PCA 的朴素贝叶斯算法与传统朴素贝叶斯算法的建模结果对比如下表所示:

表中 (0,0) 表示为正确识别的正常短信用户样本数, (0,1) 表示将正常用户误判为垃圾短信用户的样本数, (1,0) 表示将垃圾短信用户误判为正常用户的样本数, (1,1) 表示正确识别的垃圾短信用户样本数.

$$\text{准确率} = \frac{(0,0)\text{ 样本数} + (1,1)\text{ 样本数}}{\text{总样本数}} \tag{4}$$

由上表可以看出基于 PCA 的朴素贝叶斯算法较传统的朴素贝叶斯算法在垃圾短信用户识别的准确率上有显著提升.

表 1　主成分方差贡献率和累计贡献率

主成分	方差贡献率	累计方差贡献率
Comp.1	29%	29%
Comp.2	9%	38%
Comp.3	6%	44%
Comp.4	6%	50%
Comp.5	5%	56%
Comp.6	4%	59%
Comp.7	3%	62%
Comp.8	3%	65%
Comp.9	2%	67%
Comp.10	2%	70%
Comp.11	2%	72%
Comp.12	2%	74%
Comp.13	2%	76%
Comp.14	2%	78%
Comp.15	2%	80%
Comp.16	2%	82%
Comp.17	1%	84%
Comp.18	1%	85%

表 2　模型改进前后效果对比

	基于 PCA 的朴素贝叶斯		朴素贝叶斯	
	训练集	测试集	训练集	测试集
(0,0)	43926	18841	29123	12596
(0,1)	2569	1085	17372	7330
(1,0)	2823	1228	637	237
(1,1)	5463	2323	7649	3314
准确率	90.2%	90.15%	67.12%	67.76%

5　结论

朴素贝叶斯算法是目前比较高效经济的分类算法之一, 也是常用的十大算法之一. 本文在朴素贝叶斯算法的基础上, 针对朴素贝叶斯算法的固有缺陷, 提出基于 PCA 的朴素贝叶斯算法, 该方法能有效处理相关属性, 使之尽可能满足朴素贝叶斯的理论假设, 同时又能最大限度的保留原始数据集的主要信息, 从而最大程度上提高模型的准确率.

在垃圾短信客户识别的实际应用过程中也发现, 基于 PCA 的朴素贝叶斯算法较传统的

朴素贝叶斯算法在准确率上有显著提升.

参考文献

[1] Geenen P L, Gaag L C V D, Loeffen W L A, et al. Constructing naive Bayesian classifiers for veterinary medicine: A case study in the clinical diagnosis of classical swine fever[J]. Research in Veterinary Science, 2011, 91(1): 64-70.
[2] 魏浩, 丁要军. 一种基于相关的属性选择改进算法 [J]. 计算机应用与软件, 2014, 31(08): 280-284.
[3] 王行甫, 杜婷. 基于属性选择的改进加权朴素贝叶斯分类算法 [J]. 计算机系统应用, 2015, 24(08): 149-154.
[4] 焦鹏, 王新政, 谢鹏远. 基于属性选择法的朴素贝叶斯分类器性能改进 [J]. 电讯技术, 2013(03): 329-334.
[5] 饶丽丽, 刘雄辉, 张东站. 基于特征相关的改进加权朴素贝叶斯分类算法 [J]. 厦门大学学报: 自然科学版, 2012, 51(4): 682-685.
[6] Jiang L, Li C, Wang S, et al. Deep feature weighting for naive Bayes and its application to text classification[J]. Engineering Applications of Artificial Intelligence, 2016, 52(C): 26-39.
[7] Wang S, Jiang L, Li C. A CFS-based feature weighting approach to naive bayes text classifiers[J]. Knowledge-Based Systems, 2016, 100(C): 137-144.
[8] 肖子玉, 吕姗. 信息安全与垃圾短信监控 [J]. 电信工程技术与标准化, 2010, 23(3): 60-64.

The Application of Naive Bayes Algorithm Based on Principal Component Analysis in Spam User Identification

LI Qiong-yang, TIAN Ping

(Xuchang University Mathematics and Statistics Faculty, Xuchang 461000, China)

Abstract: In the identification of spam SMS users, there is a strong correlation between user behavior and consumer data involved in modeling. Using naive Bayes algorithm directly in the modeling has a very low accuracy. In order to meet the basic assumptionin of attributes dependent that the naive Bayesian algorithm requires , this paper uses principal component analysis to process the data, so as to achieve the dual purpose of dimensionality reduction and attribute independent, and then use the Naive Bayesian algorithm to model. The results show that the combination model based on principal component analysis and Naive Bayes algorithm is effective. It can be seen that it has certain practical value and practical significance in the identification of spam messages.

Keywords: attribute independent; principal component analysis; naive Bayesian; combinatorial model