文章编号:1000－2367(2018)01－0077－07 DOI:10.16366/j.cnki.1000－2367.2018.01.012

化学成分-朴素贝叶斯分类算法的烟叶产地模式识别

吴圣超[1], 刘太昂[2], 葛炯[1], 沙云菲[1]

(1.上海烟草集团有限责任公司技术中心,上海 200082;2.上海帆阳信息科技有限公司,上海 200444)

摘 要:把总糖、还原糖、总氮、烟碱、总氯和总钾这 6 个成分含量作为影响烤烟烟叶产地的自变量,利用朴素贝叶斯分类算法(NBC)建立烤烟烟叶生产地的判别模型.结果表明,用朴素贝叶斯分类建立的烟叶产地识别模型建模、留一法、预报准确度分别为 91.24%、89.05%和 88.24%,而用支持向量机分类和 K 点最近邻分类建立的烟叶产地识别模型的准确率均低于朴素贝叶斯分类建立的模型.可见利用朴素贝叶斯分类算法对烟叶产地进行模式识别研究,可以很好地反映烟叶样本由于产地的不同带来的差异.因此可以将 NBC 算法引入到烟草行业的研究中.

关键词:烟草;朴素贝叶斯分类;模式识别

中图分类号:O69 **文献标志码**:A

烟叶在我国种植区域非常广泛,其中河南和云南是我国的种植烟叶大省.在烟叶生长过程中受内部遗传组成、外部环境条件和栽培措施等因素的影响,不同种植地区的烟叶品质差别很大,风格特征迥异[1].另外,在烟叶生产实践中,技术人员逐渐认识到不同地区烟叶的特色也不一样,如滇、闽、川等地区的烟叶属于清香型;河南、湖南等产地的烟叶属于浓香型;贵州、山东、东北等产地的烟叶属于中间香型等[2].近些年,烤烟烟叶的产地识别逐渐受到了广泛关注.

基于近红外分析技术(NIR),可以建立对烤烟产地的判别模型[3].马雁军等[4]采用 PPF(Projection of Basing on Principal Component and Fisher Criterion)投影方法分析样品不同成分与原产地的相似性.束茹欣等[5]利用主成分分析及 PPF 建立了对烟叶不同生产地分析模型.王毅等[6]利用 NIR 数据,研究了烟叶光谱与产地之间相似度值的计算方法.施丰成等[7]利用 NIR,通过 PLS-DA 算法成功判别了不同烟叶的不同产地.

利用烟叶化学含量的不同也可以建立生产所属省份的判别模型.张毅等[8]收集得到 2005～2009 年湖南、河南、福建和云南的 1 040 份样品,并检测其中 21 类常规物质.基于 Mining Tree,使用 C&RT analysis 建立了对不同烟叶归属地的识别模型.

卷烟生产中非常重视对烟叶中化学含量(总糖、还原糖、总氮、烟碱、总钾和总氯)的检测,产地不同化学含量也不尽相同.本研究尝试仅仅基于这 6 种常规化学成分作为影响因素,用来朴素贝叶斯分类、支持向量机分类、K 点最近邻分类建立烤烟烟叶产地的识别模型,探索烤烟烟叶产地识别的新方法.

1 材料与方法

1.1 样 品

收集了 274 个 2014 年生产的初烤烟叶样品作为建模集样品,其中 A 省 30 个,B 省 63 个,C 省 75 个,D 省 106 个.另外提供 68 个 2014 年生产的初烤烟叶样品作为预报集样品,其中 A 省 9 个,B 省 21 个,C 省 15 个,D 省 23 个.这总共 342 个样品的等级分别是 B2F,C3F 和 X2F.

收稿日期:2047-09-07;**修回日期**:2017-10-20.

基金项目:国家自然科学基金(21273145)

作者简介:吴圣超(1987－),男,江苏海门人,上海烟草集团有限责任公司助理工程师,主要从事卷烟产品配方研究,E-mail:378225220@qq.com.

通信作者:沙云菲(1980－),女,浙江鄞县人,上海烟草集团有限责任公司高级工程师,E-mail:shayf@sh.tobacco.com.cn.

烟末样品中的烟碱、总糖、还原糖、总氮、总钾和总氯这 6 类物质质量浓度采用连续流动法[9]测定,平行测定 3 次,取平均值.

1.2 方 法

1.2.1 朴素贝叶斯分类

朴素贝叶斯分类[10](Naive Bayesian Classifier,NBC)是基于贝叶斯定理(BC)的统计学分类方法,凭借着其优异的计算、较好的精确度及其厚实的理论体系而受到社会广泛认同.基本原理是预测一个未知类别的样本属于各个类别的可能性,选择其中可能性最大的一个类别作为该样本的最终类别[11].一般而言,NBC 中所有属性都能在不同程度上对分类产生影响,而不是仅有个别属性决定结果.

NBC 的原理是先得到某目标的先验概率,再用 BC 公式得到后验然率,也就是目标归于某一个分类下的可能性,最大然率所属类别即为该目标的分类[12-13].NBC 模型是从 BC 理论发展而来的,BC 理论中最核心的部分是 BC 公式.假设 M 维样本变量 $X=(X_1,X_2,\cdots,X_M)$,x 为 X 的一个样本,类标签为 $t(t=1,2,\cdots,T)$. 贝叶斯公式可以表示如下:

$$P(t\mid X=x)=\frac{P(t)\times P(X=x\mid t)}{P(X=x)}, \tag{1}$$

公式中,$P(X=x)$ 对于体系内存在的分类来说都相同,而 $P(X=x\mid t)$ 和 $P(t)$ 则是通过数据集训练得到.所以,对于每个样本 x 来说,不需要计算 $P(t\mid X=x)$ 的精确值,只需要求出使 $P(t)\times P(X=x\mid t)$ 值最大的那个类 t,就可以预测出该样本 x 所在的类.可是,$P(X=x\mid t)$ 的计算是不太具有可行性的.有鉴于此,学者们提出了 NBC 模型,它最重要的假设是:定义类别标签 t,假设样本属性互不相干.于是就有:

$$P(X=x\mid t)=\prod_{m=1}^{M}P(X_m=\frac{x_m}{t}). \tag{2}$$

现在贝叶斯公式变为:

$$P(t\mid X=x)=\frac{P(t)\times\prod_{m=1}^{M}P(X_m=\frac{x_m}{t})}{P(X=x)}. \tag{3}$$

1.2.2 支持向量机分类算法

支持向量机分类算法[14-15](Support vector classification,SVC)是 Vapnik 等在统计学习理论(Statistical learning theory,SLT)基础上提出的支持向量机算法(Support vector machine,SVM)的一种.SLT 体系以及 SVM 算法在处理少量样本的问题上的进展和成果十分瞩目,已经可以说是目前处理少量样本问题上的最佳选择了.

Vapnik 等在求解风险函数:

$$R[f]=\int_{x*y}(y-f(x))^2P(x,y)\mathrm{d}x\mathrm{d}y \tag{4}$$

为最小时,将结构风险函数 $R_h[f]$ 替代风险函数 $R[f]$,并证明了 $R_h[f]$ 可以用下列函数求极小而获得:

$$R_h[f]=\min_{s_h}\left\{R_{emp}[f]+\sqrt{\frac{h(\ln(\frac{2n}{h})+1)-\ln(\frac{\theta}{4})}{n}}\right\}, \tag{5}$$

其中 n 是建模集合的样本量,s_h 是 VC 维度的空间结构,h 是 VC 维度数目,R_{emp} 为经验风险函数.

对于分类问题,SVM 将分类样本映射到相对高维的坐标空间,并在高维空间内求出能对二类样本分类的最佳超平面方程,以此方程再来判别新未知样本的分类.即使用来预报小样本的集合,也可以做到较高的准确率.

1.2.3 K 点最近邻算法

KNN 法(K-nearest neighbor),也称为 K 点最近邻算法[16],最初的思路是先将事先分过类的已知样本和尚未分过类的样本分别计入高维坐标,考察未分过类的样本的 K 个附近的点.若近邻中某一类的点最多,则该未知样本就归到此类别.多维空间中,样本间距规定为欧几里得空间.两点 i 与 j 的间距 d_{ij} 可表示为:

$$d_{ij}=\left[\sum_{k=1}^{M}(x_{ik}-x_{jk})^2\right]^{\frac{1}{2}}. \tag{6}$$

有时为了方便计算，也可以采用绝对距离：

$$d_{ij}=\sum_{k=1}^{M}|x_{ik}-x_{jk}|. \tag{7}$$

KNN 法的优势在于对数据没有特殊要求，也不需要训练过程.但是劣势在于不会对已知点进行信息压缩，每次归类未知位置点需要把和所有的集合内的点的间距都运算一次，因此运算量十分巨大.另外，若该点附近某类点的量较大时，容易造成误判.

2　结果和讨论

2.1　烟叶化学指标

4 个省份烟叶常规成分数据的数据分析见表 1.

由表 1 可见：①D 省烟叶的总糖平均值为 31.86%，明显高于其他 3 个省份，说明 D 省的烟叶抽吸甜度可能更好，香气可能更为醇和饱满，烟叶的整体质量也是相对较高的；②A 省烟叶的还原糖和总钾平均值为 26.39%和 2.66%，明显高于其他 3 个省份，还原糖含量较高说明 A 省的成品烟的抽吸甜度也可能更好，香气可能更为醇和饱满，与总糖指标一样有助于烟叶评级的提高，另外总钾浓度高说明 A 省烟叶的燃烧性、吸湿性较好，烟叶的颜色和身份较好，有助于烟叶外观质量上的提升；③B 省烟叶的烟碱平均值为3.13%，明显高于其他 3 个省份，说明 B 省的烟叶整体生理满足感更强；④C 省烟叶的总氯浓度均值为 0.61%，显然高于其他 3 个省份，总氯含量较高说明 C 省烟叶吸湿性较大，可能会导致吸湿性过高进而降低其燃烧性，可能会对使用性能上造成一定的影响.

表 1　烟叶常规化学分析统计

化学成分	A 省	B 省	C 省	D 省
	平均值±标准差/%	平均值±标准差/%	平均值±标准差/%	平均值±标准差/%
总糖	30.17±5.28	26.25±5.80	23.80±4.43	31.86±6.69
还原糖	26.39±4.68	20.84±5.08	21.54±4.34	24.96±4.99
总氮	2.24±0.29	2.16±0.53	1.98±0.30	2.08±0.50
烟碱	2.62±0.91	3.13±1.60	2.47±0.73	2.32±0.87
总钾	2.66±0.53	1.77±0.91	1.31±0.44	1.58±0.48
总氯	0.22±0.10	0.28±0.13	0.61±0.22	0.34±0.22

2.2　主成分分析法(Principal component Analysis, PCA)投影图

本研究将烟末样品中的烟碱、总糖、还原糖、总氮、总钾和总氯这 6 种常规成分作为烤烟产地的 6 个自变量，构成了 1 个 6 维的高维空间，为了了解样本在高维空间的分布，本文用主成分方法[17-18]将烟叶样本由 6 维的高维空间投影到由第 1 主成分和第 2 主成分构成的二维平面，图 1 是烟叶样本的 PCA 投影图.由图 1 可以看出，4 省份烟叶样本在 6 种化学成分构成的高维空间中呈现混淆分布，不能将对烟叶产地进行区分.

2.3　NBC 模型整体准确性

图 1 显示出 A 省、B 省、C 省和 D 省烟叶样本在高维空间中呈现混淆分布，但可以尝试其他算法对烟叶产地进行判别.本文用 NBC 分类建立了烟叶产地的分类模型，模型的建模、留一法和预报准确率见表 2.由表 2 可以看出：以 6 种常规化学成分为因变量，用朴素贝叶斯分类建立烟叶产地的分类模型，其建模、留一法预测准确率分别为：91.24%、89.05%和 88.24%.

表 2　NBC 模型建立、留一法和预报准确率

建模准确率/%	留一法准确率/%	预报准确率/%
91.24	89.05	88.24

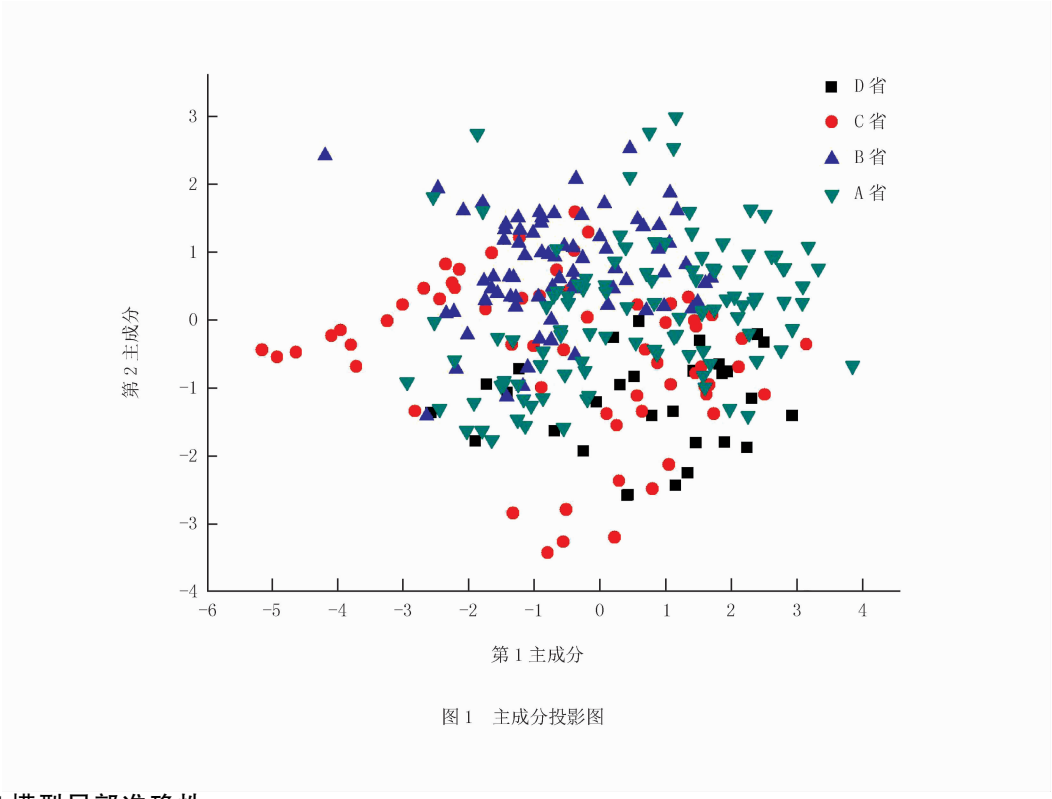

图 1 主成分投影图

2.4 NBC 模型局部准确性

表 2 呈现了 NBC 对 4 个产地整体分类判别的建模、留一法和预测准确率,没有显示出各个省份烟叶分类判别结果.表 3～5 的混淆矩阵[19]展示了更加详细的烟叶产地分类 NBC 模型的判别结果.由表 3 可以看出 A、B、C 和 D 省的 NBC 模型建模准确率分别是 96.67%、90.48%、95.89%和 88.68%.由表 4 可以发现 A、B、C 和 D 省的 NBC 模型的留一法准确率分别是:93.33%、87.30%、90.67%和 87.84%.由表 5 可以看出 A、B、C 和 D 省的 NBC 模型的预测准确率分别是:88.89%、90.48%、86.67%和 86.96%.

表 3 NBC 模型建模的混淆矩阵

类别	A 省	B 省	C 省	D 省
A 省	29	1	0	0
B 省	0	57	2	4
C 省	0	3	70	2
D 省	0	8	4	94

表 4 NBC 模型留一法的混淆矩阵

类别	A 省	B 省	C 省	D 省
A 省	28	2	0	0
B 省	0	55	2	6
C 省	0	4	68	3
D 省	0	8	5	93

2.5 SVM 模型的准确性

为了对比还采用 SVM 分类建立了 4 个烟叶产地的分类模型,模型整体的建模、留一法和预报的正确率见表 6.由表 6 可以看出,得到的烟叶产地的 SVM 分类判别模型的建模、留一法、预报的准确率分别为:90.87%、86.13%和 82.35%. 3 者的准确率均小于 NBC 模型.

表 5　NBC 模型预报的混淆矩阵

类别	A 省	B 省	C 省	D 省
A 省	8	1	0	0
B 省	1	19	0	1
C 省	0	1	13	1
D 省	0	3	0	20

表 7～9 的混淆矩阵展示了 4 个烟叶产地的 SVM 模型分类判别结果.由表 7 可以看出 A、B、C 和 D 省的 SVM 模型的建模准确率分别是 88.89%、87.74%、93.33%和 96.00%.由表 8 可以看出 A、B、C 和 D 省的 SVM 模型的留一法准确率分别是：82.54%、84.91%、83.33%和 92.00%.由表 9 可以看出 A、B、C 和 D 省的 SVM 模型的对未知样本预报准确率分别是：88.89%、77.27%、86.67%和 82.61%.

SVM 模型的 4 个烟叶产地分类的准确率大部分都小于 NBC 模型，特别是 SVM 模型对 B 省的预报准确率只有 77.27%，远小于 NBC 模型的 90.48%.

表 6　SVM 模型的建模、留一法和预报准确率

建模准确率/%	留一法准确率/%	预报准确率/%
90.87	86.13	82.35

表 7　SVM 模型建模的混淆矩阵

类别	A 省	B 省	C 省	D 省
A 省	56	5	0	2
B 省	10	93	0	3
C 省	0	1	28	1
D 省	2	1	0	72

表 8　SVM 模型留一法的混淆矩阵

类别	A 省	B 省	C 省	D 省
A 省	52	8	1	2
B 省	12	90	1	3
C 省	2	2	25	1
D 省	3	3	0	69

表 9　SVM 模型预报的混淆矩阵

类别	A 省	B 省	C 省	D 省
A 省	8	1	0	0
B 省	1	17	0	4
C 省	0	1	13	1
D 省	0	4	0	19

2.6　KNN 模型的准确性

此外，还使用了 KNN 分类建立了 4 个烟叶产地的分类模型，模型整体的建模、留一法和预报的准确率见表 10.由表 10 可以看出，得到的烟叶产地的 KNN 分类判别模型的建模、留一法、预报的准确率分别为：80.28%、80.28%和 79.71%.3 者的准确率远远小于 NBC 模型.

表 11～13 是 4 个烟叶产地的 KNN 分类模型.表 11 中可以看到 A、B、C、D 省的 KNN 模型的建模准确率分别为 73.01%、71.70%、83.33%、97.33%.由表 12 可以得到 A、B、C、D 省的 KNN 模型的留一法准确率

为73.01%、71.70%、83.33%、97.33%.由表 13 可以得到 A、B、C、D 省的 KNN 模型的对未知样本的预报准确率为 88.89%、68.18%、86.67%、82.61%.

KNN 模型的 4 个烟叶产地分类的准确率大部分都小于 NBC 模型,其中甚至有结果低于了 70%,如 KNN 算法对 B 省的预报准确率仅有 68.18%,可见 NBC 分类效果要远远好于 KNN 分类.

表 10 KNN 模型的建模、留一法和预报准确率

建模准确率/%	留一法准确率/%	预报准确率/%
80.28	80.28	79.71

表 11 KNN 模型建模的混淆矩阵

类别	A 省	B 省	C 省	D 省
A 省	46	11	2	4
B 省	15	76	5	10
C 省	0	2	25	3
D 省	1	0	1	73

表 12 KNN 模型留一法的混淆矩阵

类别	A 省	B 省	C 省	D 省
A 省	46	11	2	4
B 省	15	76	5	10
C 省	0	2	25	3
D 省	1	0	1	73

表 13 KNN 模型预报的混淆矩阵

类别	A 省	B 省	C 省	D 省
A 省	8	1	0	0
B 省	0	15	3	4
C 省	0	1	13	1
D 省	0	4	0	19

3 结 论

利用 A 省、B 省、C 省和 D 省 4 个省份烟叶化学成分的差异性建立了烟叶产地分类判别的 NBC 模型.结果表明,以烟碱、总糖、还原糖、总氮、总钾和总氯这 6 类成分质量浓度作为因变量建立起来的烟叶产地 NBC 模型可以达到很高的准确率,而利用 SVM 算法和 KNN 算法得到的结果则劣于 NBC 模型.由此可见,NBC 对烟叶产地预报模型可以很好地反映烟叶样本由于产地的不同带来常规化学成分的差异性.因此可以将 NBC 算法引入到烟草行业的研究中.

参 考 文 献

[1] 韩富根,卢红,闫克玉,等.烟草化学[M].北京:中国农业出版社,2010.

[2] 唐远驹.与烟叶特色相关的几个问题[J].中国烟草科学,2013,32(4):1-4.

[3] 陆婉珍,袁洪福,徐广通,等.现代近红外光谱分析技术[M].北京:中国石化出版社,2000.

[4] 马雁军,李雪莹,马莉,等.用近红外光谱和特征指标判别国产白肋烟产地及部位间相似性[J].中国烟草学报,2017,146:38-48.

[5] 束茹欣,蔡嘉月,杨征宇,等.应用近红外光谱投影模型法分析烟叶的产区与风格特征[J].光谱学与光谱分析,2014,34(10):2764-2768.

[6] 王毅,马翔,温亚东,等.应用近红外光谱分析云南主要烟叶生产基地之间的烟叶特性[J].光谱学与光谱分析,2013,33(1):78-80.

[7] 施丰成，李东亮，冯广林，等.基于近红外光谱的 PLS-DA 算法判别烤烟烟叶产地[J].烟草化学，2013，48(4)：56-59.
[8] 张毅，李强，王政，等.一种基于分类-回归决策树的烤烟产区识别模型[J].中国烟草学报，2014，20(6)：28-33.
[9] 孔浩辉，郭文，张心颖，等.连续流动法测定烟草中的蛋白质含量[J].烟草科技，2009，47(11)：38-41.
[10] 张亚萍，陈得宝，侯俊钦，等.朴素贝叶斯分类算法的改进及应用[J].计算机工程与应用，2011，47(15)：134-137.
[11] 刘红岩，陈海亮.Graph-NB：一种高效准确的多关系朴素贝叶斯分类算法[J].信息系统学报，2008，2(1)：1-11.
[12] 包小兵.基于朴素贝叶斯的 Web 文本分类及其应用[J].电脑知识与技术，2016，12(30)：220-222.
[13] 胡德敏，龚燕.基于谱聚类和扩展朴素贝叶斯的混合推荐算法[J].计算机应用研究，2016，33(12)：3709-3712.
[14] Vapnik.Statistical learning theory[M].New York：Wiley-Interscience，1998.
[15] 刘尚旺，段德全，崔艳萌，等.二次定位车牌分割及识别方法[J].河南师范大学学报(自然科学版)，2016，41(4)：151-155.
[16] 陈念贻，钦佩，陈瑞亮，等.模式识别方法在化学化工中的应用[M].北京：科学出版社，2000.
[17] Pearson K.On lines and planes of closest fit to systems of points in space[J].Philippine Magazine Series 6，1901，2(11)：559 - 572.
[18] Hotelling H.Analysis of a complex of statistical variables into principal components.[J].British Journal of Educational Psychology，1933，24(6)：417-520.
[19] 张静，宋锐，郁文贤，等.基于混淆矩阵和 Fisher 准则构造层次化分类器[J].软件化学，2005，16(9)：1560-1567.

Pattern recognition of the producing areas of flue-cured tobacco based on naive bayesian classifier algorithm base on the contents of chemical components

Wu Shengchao[1]，Liu Taiang[2]，Ge Jiong[1]，Sha Yunfei[1]

(1.Technology Center of Shanghai Tobacco Group Co.，LTD.，Shanghai 200082，China
2. Shanghai Fanyang Information Technology Co.，LTD.，Shanghai 200444，China)

Abstract：With the Naive Bayesian Classifier，a pattern recognition model of the producing areas of flue-cured tobacco was built. The model features were the contents of chemical components，including total sugar，reducing sugar，total nitrogen，nicotine，total chlorine and total potassium. The accuracy of the training set，LOOCV and the test set were 91.24%，89.05% and 88.24%，while the results of the SVM and the KNN could not get the same accuracy level as the NBC. The Naive Bayesian Classifier could be applied to pattern recognition of flue-cured tobacco samples of different origins.

Keywords：tobacco；naive Bayesian classifier；pattern recognition

[责任编校　赵晓华]