doi: 10.12068/j.issn.1005-3026.2018.01.004

基于OVO分解策略的智能卷烟感官评估方法

张忠良[1,2], 雒兴刚[1,2], 汤建国[3], 唐加福[1]
(1. 东北大学 信息科学与工程学院, 辽宁 沈阳 110819; 2. 杭州电子科技大学 管理学院, 浙江 杭州 310018;
3. 云南中烟工业有限责任公司 技术中心, 云南 昆明 650231)

摘　　要: 针对智能卷烟感官评估系统中涉及的多分类问题,采用"一对一"(one-versus-one, OVO)分解策略将复杂的多分类问题分解成多个易于处理的二分类子问题,然后针对这些子问题分别建立二值分类器,最后采用一定的聚合策略将二值分类器组合成多类分类器. 此外,分别采用基于动态分类器选择和基于距离相对竞争力加权法对OVO中的冗余二值分类器进行处理,从而降低其对OVO系统的消极影响. 为了验证所采用的方法在智能卷烟感官评估中的有效性,采用国内某烟草公司提供的数据集进行对比实验. 实验结果表明,在智能卷烟感官评估中基于OVO分解策略的多分类方法比传统方法具有更优的分类性能.

关 键 词: 多分类;一对一分解;聚合策略;卷烟感官质量;智能评估
中图分类号: TP 274　　**文献标志码**: A　　**文章编号**: 1005-3026(2018)01-0015-06

Intelligent Cigarette Sensory Evaluation Method Based on OVO Decomposition Strategy

ZHANG Zhong-liang[1,2], *LUO Xing-gang*[1,2], *TANG Jian-guo*[3], *TANG Jia-fu*[1]
(1. School of Information Science & Engineering, Northeastern University, Shenyang 110819, China; 2. School of Management, Hangzhou Dianzi University, Hangzhou 310018, China; 3. Technology Center, China Tobacco Yunnan Industrial Co., Ltd., Kunming 650231, China. Corresponding author: LUO Xing-gang, professor, E-mail: xgluo@mail.neu.edu.cn)

Abstract: Intelligent cigarette sensory evaluation system involves multi-class classification problems. The one-versus-one (OVO) decomposition strategy was employed to divide the multi-class classification problem into several easier-to-solve binary sub-problems. Then binary classifiers were established for these sub-problems. Finally, an aggregation strategy was adopted to combine the binary classifiers to be a multi-class classifier. In addition, dynamic classifier selection for OVO strategy (DCS-OVO) and distance-based relative competence weighting for OVO strategy (DRCW-OVO) were used to reduce the negative effect of the non-competent classifiers. In order to verify the effectiveness of the employed method in intelligent cigarette sensory evaluation, the experimental comparison by using the dataset from a Chinese tobacco company was carried out. The results indicate that the OVO decomposition strategy outperforms the classical methodology in intelligent cigarette sensory evaluation.

Key words: multi-class classification; one-versus-one (OVO) decomposition; aggregation strategy; cigarette sensory quality; intelligent evaluation

感官品质是评价卷烟质量的重要标准. 然而,由于烟草的物化指标与卷烟的感官质量之间存在着极其复杂的非线性关系,很难建立起两者之间有效的数学模型. 目前,人们主要通过人工评吸的方式对卷烟新产品开发和产品维护过程中涉及的卷烟感官质量进行评估. 然而,这种完全依赖评吸专家经验的方式容易受到其主观因素以及客观环境的影响,同时也对评吸专家的身体健康产生一

收稿日期: 2016-07-11
基金项目: 国家自然科学基金资助项目(71771070).
作者简介: 张忠良(1986-),男,浙江嘉兴人,东北大学博士研究生; 雒兴刚(1971-),男,新疆奇台人,东北大学教授,博士生导师; 唐加福(1965-),男,湖南东安人,东北大学教授,博士生导师.

定的负面影响. 此外,现代企业面对着越来越复杂的市场环境,需要卷烟企业在新产品开发和产品维护中,具有更高的效率和灵活性.

为了解决上述问题,研究人员开始采用数据挖掘技术对卷烟感官质量进行评价[1],试图从大量积累的烟草数据中获得烟叶物理化学指标和卷烟感官质量之间的映射关系,以辅助或代替评吸专家完成对卷烟感官质量的评价. 当前主要以神经网络[2]或支持向量机[3]来解决成品卷烟的智能化感官评估问题.

然而,卷烟感官质量具有多个等级,历史数据涉及多分类问题,而目前的研究主要采用分类算法直接处理多分类问题. 由于多分类问题的复杂性,采用这种直接的方式进行卷烟感官质量预测往往不能获得预期的效果. 另一方面,研究表明[4]采用一对一(one-versus-one, OVO)分解策略可以有效提高分类算法处理多分类问题的性能. 因此,本文采用 OVO 分解策略将智能卷烟感官评估中涉及的多分类问题分解成多个易于处理的二分类子问题,然后对每个子问题分别建立二值分类器,并采用聚合策略将二值分类器组合成多类分类器.

为了验证本文所采用方法在智能卷烟感官评估中的有效性,将其应用在烟草公司采集的数据集上,并比较了 6 种不同的聚合策略的效果. 此外,为了分析基分类器对本文方法的影响,分别采用决策树、神经网络和支持向量机作为基分类器进行了实验比较. 结果表明,不管采用何种基分类器,采用 OVO 分解策略均比经典的多类分类器的预测效果具有明显的优势.

1 数据和方法

1.1 数据

本文所采用的实验数据是国内某烟草公司在 2010 ~ 2012 期间所采集的日常检测数据,包括成品烟的常规化学成分以及相应的卷烟感官质量,数据样本个数为 684 个. 其中,常规化学成分包括总糖量等 13 项指标,将作为训练分类器时的输入属性;每个样本对应包括 5 项卷烟感官质量评估指标,分别为光泽、香气、杂气、刺激性、余味,作为分类器的决策属性. 针对每项评估指标,首先由多名卷烟感官评估专家根据感官质量评判标准(如表 1 所示)分别进行独立评估,然后通过计算所有专家给出的结果的平均值,从而得到相应指标的最终评估结果. 由于针对卷烟感官质量的评估指标有 5 项,因此本文共采用 5 组数据集.

表 1 感官质量判定标准

Table 1 Assessment criteria for sensory quality

评估指标	第一类	第二类	第三类
香气	细腻、丰满(32)	稍粗糙、充足(28)	较粗糙、淡薄(24)
光泽	油润(5)	较油润(4)	较暗淡(3)
刺激性	无刺激(20)	略有刺激(17)	较有刺激(15)
杂气	无杂气(12)	微有杂气(10)	略有杂气(8)
余味	舒适、纯净(25)	较舒适、较净(22)	尚舒适、尚净(20)

注:括号内为质量描述的最高分值.

1.2 方法

1.2.1 OVO 分解策略

OVO 分解策略针对多分类问题,将其分解成多个二分类子问题. 其基本思想如图 1 所示:首先将 m 类问题分解成 $m(m-1)/2$ 个二分类子问题,然后对每个子问题分别独立地建立二值分类器. 这样复杂的多分类问题就被分解成多个较易识别的二分类子问题.

为了识别未标识样本,OVO 方法一般需要两个步骤. 1)利用分类算法训练二值分类器用以区分成对类$\{C_i, C_j\}$;2)集成阶段. 置信度 $r_{ij} \in [0, 1]$表示二值分类器相对于类别 j 将未标识样本预测为类别 i 的概率,并且 $r_{ji} = 1 - r_{ij}$. 对每个成对类别分别进行相同的操作,就可以得到一个完整的分值矩阵:

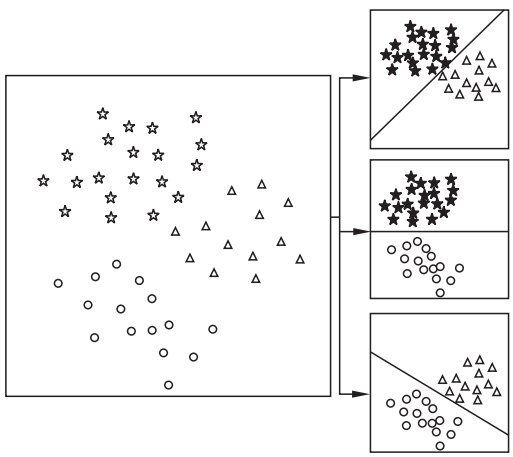

图 1 “一对一”分解策略示意图

Fig. 1 Illustration of “one versus one” decomposition

$$R=\begin{pmatrix} — & r_{12} & \cdots & r_{1m} \\ r_{21} & — & \cdots & r_{2m} \\ \vdots & & & \vdots \\ r_{m1} & r_{m2} & \cdots & — \end{pmatrix}. \tag{1}$$

最后,采用聚合策略可以得到未标识样本的输出结果.

1.2.2　聚合策略

由上文可知,OVO 分解策略的关键是将分值矩阵集成得到最终的输出结果. 本文采用以下几种方法.

1）投票法(voting strategy, VOTE)[4]. 该方法利用投票机制获得最终类别标识,获得最多投票的类别作为输出结果. 预测类别为

$$H=\arg\max_{i=1,\cdots,m}\sum_{1\leqslant j\neq i\leqslant m}s_{ij}. \tag{2}$$

其中,$s_{ij}=\begin{cases}1, & r_{ij}>r_{ji};\\ 0, & \text{其他}.\end{cases}$

2）加权投票法(weighted voting, WV)[5]. 在加权投票法中,每个二值分类器以置信度的形式给出结果,预测结果如式(3)所示:

$$H=\arg\max_{i=1,\cdots,m}\sum_{1\leqslant j\neq i\leqslant m}r_{ij}. \tag{3}$$

3）学习权值偏好法(learning valued preference for classification, LVPC)[6-7]. 该方法将绝对偏好、冲突程度和未知程度引入到最终类别的识别过程中,其决策规则:

$$H=\arg\max_{1\leqslant j\neq i\leqslant m}\sum_{1\leqslant j\neq i\leqslant m}p_{ij}+\frac{1}{2}c_{ij}+\frac{N_i}{N_i+N_j}I_{ij}. \tag{4}$$

其中:N_i 是类别 i 在训练集中的样本数;p_{ij} 和 p_{ji} 分别是对类别 i 和类别 j 的绝对偏好;c_{ij}是冲突程度;I_{ij}是未知程度. 相应的计算方法:

$$p_{ij}=r_{ij}-\min\{r_{ij},r_{ji}\}, \tag{5}$$

$$p_{ji}=r_{ji}-\min\{r_{ij},r_{ji}\}, \tag{6}$$

$$c_{ij}=\min\{r_{ij},r_{ji}\}, \tag{7}$$

$$I_{ij}=1-\max\{r_{ij},r_{ji}\}. \tag{8}$$

4）非支配准则(preference relations solved by non-dominance criterion, ND)[8-9]. 该方法的分值表引入了正规化模糊偏好关系. 最大程度不受支配的类别视作最终的输出结果,其判断规则:

$$H=\arg\max_{i=1,\cdots,m}\{1-\max_{1\leqslant j\neq i\leqslant m}r'_{ji}\}. \tag{9}$$

其中,r'_{ji}是相应的正规化分值表. r'_{ij} 和 R_{ij} 的计算方法:

$$r'_{ij}=\begin{cases}R_{ij}-R_{ji}, & \text{如果 } R_{ij}>R_{ji};\\ 0, & \text{其他}.\end{cases} \tag{10}$$

$$R_{ij}=\frac{r_{ij}}{r_{ij}+r_{ji}}. \tag{11}$$

1.2.3　冗余分类器处理技术

在传统的 OVO 方法中,将所有的二值分类器组合成多类分类器. 然而,冗余的二值分类器有可能导致错误的输出结果. 因此,在聚合过程中剔除冗余的二值分类器或者降低其负面影响可以有效地提高 OVO 的性能. 动态分类器选择的 OVO 策略和基于距离相对竞争力加权的 OVO 策略是两种处理 OVO 中冗余二值分类器的有效途径.

1）动态分类器选择的 OVO 策略(dynamic classifier selection for OVO strategy, DCS－OVO)[10],该方法考虑避免没有竞争力的子分类器对最终结果的负面影响. 首先,在训练集中确定待测样本的 k 个最近邻,然后观察邻域中包含的类别,最后将原始分值表中删去邻域中没有的类别的相应信息,从而获得新的分值表. 这里采用的最近邻数为类别数的 3 倍,如果邻域中只有一个类别,则最近邻数扩大到 6 倍. 最后针对新的分值表采用 WV 获得最终的输出结果.

2）基于距离相对竞争力加权的 OVO 策略(distance-based relative competence weighting for OVO strategy, DRCW－OVO)[11],该方法将距离引入到权值的计算中. 对于待测样本,训练集中与其越接近的样本所对应的类别将获得更大的权值. 对于训练集中的每一个类别,计算与待测样本最邻近的 k 个样本,并计算出该类别与待测样本的平均距离. 然后对于训练集中 m 个类别都进行如上操作,获得平均距离向量 $\boldsymbol{d}=(d_1,\cdots,d_m)$. 新的分值表就由 $r_{ij}\cdot w_{ij}$构成,采用 WV 获得最终的预测结果.

$$H=\arg\max_{i=1,\cdots,m}\sum_{1\leqslant j\neq i\leqslant m}r_{ij}\cdot w_{ij}. \tag{12}$$

其中,$w_{ij}=\dfrac{d_j^2}{d_i^2+d_j^2}$.

2　实验测试与分析

2.1　数据预处理

由前文所述可知,卷烟感官质量评估历史数据是多个评估结果的平均值,因此原始数据集中的数据不是离散值. 然而,针对卷烟感官质量的评估结果应该为离散值,因为卷烟感官质量评估专家基于表 1 以 0.5 为间隔对评估指标进行评价打分. 基于此,本文从卷烟感官质量评估的实际出发,采用等间隔的离散方法对决策属性进行离散化处理,离散结果如表 2 所示. 以光泽为例,打分评判标准为 3～5,因为实际数据集中没有小于

3.75 的数据,因此,将[3.75,4.25]内的样本标记为类别 1,(4.25,4.75]标记为类别 2,(4.75,5.00]标记为类别 3.

表 2 离散化结果

Table 2 Results of the discretization

评估指标	离散化描述
香气	[26.25, 26.75]∈类别 1, (26.75, 27.25]∈类别 2, (27.25, 27.75]∈类别 3, (27.75, 28.25]∈类别 4, (28.25, 28.75]∈类别 5, (28.75, 29.25]∈类别 6, (29.25, 29.75]∈类别 7, (29.75, 30.25]∈类别
光泽	[3.75, 4.25]∈类别 1, (4.25, 4.75]∈类别 2, (4.75, 5.00]∈类别 3
刺激性	[16.75, 17.25]∈类别 1, (17.25, 17.75]∈类别 2, (17.75, 18.25]∈类别 3
杂气	[9.25, 9.75]∈类别 1, (9.75, 10.25]∈类别 2, (10.25, 10.75]∈类别 3, (10.75, 11.25]∈类别 4
余味	[20.25, 20.75]∈类别 1, (20.75, 21.25]∈类别 2, (21.25, 21.75]∈类别 3, (21.75, 22.25]∈类别 4, (22.25, 22.75]∈类别 5

根据表 2 离散化之后的 5 组数据集的基本信息如表 3 所示.

表 3 数据集信息描述

Table 3 Summary description of datasets

感官指标	样本数	属性个数	类别数	数据分布
香气	684	13	8	2/56/34/60/9/134/339/50
光泽	684	13	3	42/118/524
刺激性	684	13	3	78/485/121
杂气	684	13	4	41/49/77/517
余味	684	13	5	24/78/64/407/111

此外,需要对每一个条件属性进行归一化处理,归一化方法:

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}. \tag{13}$$

其中:x_i 是选定属性的取值;$x_{\min}$ 和 $x_{\max}$ 分别表示该属性所有取值中的最小值和最大值.

本文采用“交叉验证法”(cross validation)进行实验评估. 首先,将数据集按照分层抽样的方式划分成大小相似的互斥子集;然后,每次采用其中的 4 个子集的并集作为训练分类器的训练数据集,剩下的那个子集作为测试数据集;最后,计算 5 次实验的平均分类准确率. 重复进行 5 次这样的实验,5 次的平均结果作为最终分类预测准确率.

2.2 基分类器及其参数设置

为了分析不同基分类器的影响,分别采用决策树、神经网络以及支持向量机作为基分类器进行实验. 所采用的基分类器的参数设置如表 4 所示.

2.3 实验结果

采用分类准确率作为各个方法的评价指标,即分类准确率 = (正确分类样本数/样本总数) × 100% . 表 5 ~ 表 7 是本文实验的结果,“Base”表示直接采用基分类器得到的结果,由于目前还没有有效的方法将支持向量机直接运用于多分类问题,所以表 7 中没有“Base”列. 另外,每个结果后面括号内的数值表示该方法在对应数据集中的效果的排列名次.

表 4 基分类器的参数设置

Table 4 Parameters setting for the base classifiers

算法	参数设置
决策树	特征选择标准 = Gini 剪枝 = 是 方法 = 分类 不纯节点中含有 10 个以上样本需分裂
神经网络	隐含层节点数 = 10 激活函数 = “logsig” 迭代次数 = 100
支持向量机	C = 1.0 容差参数 = 0.001 核函数类型 = 径向基函数 优化方法 = SMO

从表 5 ~ 表 7 中可以清楚地看出,在 3 种不同的基分类器的情况下,无论是针对每一个感官指标的预测效果,还是总体的平均结果,基于 OVO 分解策略的预测效果都明显优于传统的方法. 另外,不同的聚合策略所产生的效果也不尽相同:当决策树作为基分类器时,DRCW - OVO 的效果最好,其平均结果比经典的分类器(Base)提高了 5.32% ;当神经网络作为基分类器时,虽然 DRCW - OVO 取得了最好的平均结果,但是 ND 在香气和余味两个指标上优于 DRCW - OVO,另外,DRCW - OVO 的平均结果比经典分类器提高了 11.13% ;当 SVM 作为基分类器时,除香气指标外,DRCW - OVO 均获得了最好的效果. 综上

所述,基于 OVO 分解策略的多分类方法能够有效提高智能卷烟感官评估中对感官指标的预测效果.

表 5　以决策树为基分类器的卷烟感官质量分类准确率

Table 5　Classification accuracy of cigarette sensory quality with decision tree as the base classifier

数据集	Base	VOTE	WV	LVPC	ND	DCS – OVO	DRCW – OVO
光泽	86.50 (4)	86.44 (6)	86.49 (5)	86.64 (2)	86.61 (3)	86.35 (7)	88.13 (1)
香气	53.05 (7)	54.45 (6)	55.73 (3)	56.84 (2)	54.92 (4)	54.59 (5)	60.51 (1)
杂气	82.63 (4)	81.99 (7)	82.67 (3)	82.78 (2)	82.32 (6)	82.55 (5)	84.89 (1)
刺激	64.85 (2)	62.98 (7)	64.24 (4)	64.85 (2)	64.18 (5)	63.98 (6)	68.74 (1)
余味	63.27 (4)	62.36 (6)	63.30 (3)	63.89 (2)	62.31 (7)	62.51 (5)	66.67 (1)
平均结果	70.06 (4.2)	69.64 (6.4)	70.49 (3.6)	71.00 (2.0)	70.07 (5.0)	70.00 (5.6)	73.79 (1.0)

表 6　以神经网络为基分类器的卷烟感官质量分类准确率

Table 6　Classification accuracy of cigarette sensory quality with neural network as the base classifier

数据集	Base	VOTE	WV	LVPC	ND	DCS – OVO	DRCW – OVO
光泽	80.53 (7)	86.43 (5)	86.55 (4)	83.60 (6)	86.96 (2)	86.64 (3)	89.35 (1)
香气	47.37 (7)	59.16 (5)	60.61 (3)	54.43 (6)	61.81 (1)	60.88 (2)	59.45 (4)
杂气	77.87 (7)	83.34 (5)	84.25 (3)	78.13 (6)	84.45 (2)	83.64 (4)	85.33 (1)
刺激	71.64 (6)	72.81 (4)	72.90 (3)	70.91 (7)	73.01 (2)	72.02 (5)	74.12 (1)
余味	62.19 (7)	67.07 (5)	69.03 (3)	64.09 (6)	69.42 (1)	68.53 (4)	69.13 (2)
平均结果	67.92 (6.8)	73.76 (4.8)	74.67 (3.2)	70.23 (6.2)	75.13 (1.6)	74.34 (3.6)	75.48 (1.8)

表 7　以支持向量机为基分类器的卷烟感官质量分类准确率

Table 7　Classification accuracy of cigarette sensory quality with support vector machine as the base classifier

数据集	VOTE	WV	LVPC	ND	DCS – OVO	DRCW – OVO
光泽	88.19 (3)	88.22 (2)	87.26 (6)	88.16 (4)	88.02 (5)	89.56 (1)
香气	59.83 (5)	60.04 (3)	60.56 (1)	59.95 (4)	59.25 (6)	60.47 (2)
杂气	85.61 (4)	85.73 (2)	84.97 (6)	85.64 (3)	85.14 (5)	85.91 (1)
刺激	71.67 (5)	72.05 (3)	72.67 (2)	71.85 (4)	70.76 (6)	73.81 (1)
余味	70.41 (5)	70.44 (4)	70.82 (2)	70.50 (3)	70.20 (6)	71.44 (1)
平均结果	75.14 (4.4)	75.30 (2.8)	75.26 (3.4)	75.22 (3.6)	74.67 (5.6)	76.24 (1.2)

3　结　　论

本文采用了一种基于 OVO 分解策略的多分类方法对卷烟感官质量进行预测并进行了详细的实验比较. 为了验证本文方法的稳定性,分别以决策树、神经网络和支持向量机作为基分类器进行了实验. 分析了 4 种不同的聚合策略的性能,以及 2 种冗余分类器处理技术在智能卷烟感官质量预测中的性能. 实验结果表明,无论采用何种基分类器,与经典分类器应用于预测卷烟感官质量相比,基于 OVO 分解策略的多分类技术在智能卷烟感官质量预测中具有明显的优势. 并且,采用冗余分类器处理技术可以进一步提升系统的性能.

参考文献:

[1] 王强,李孟军,陈英武. 卷烟配方数据挖掘技术研究进展[J]. 中国烟草科学,2007,28(4):14 – 17.
(Wang Qiang, Li Meng-jun, Chen Ying-wu. Research progress in data mining technology on cigarette formulation [J]. *Chinese Tobacco Science*,2007,28(4):14 – 17.)

[2] 邵惠芳,许自成,李东亮,等. 基于 BP 神经网络建立烤烟感官质量的预测模型[J]. 中国烟草学报,2011,17(1):19 – 25.
(Shao Hui-fang, Xu Zi-cheng, Li Dong-liang, et al. The establishment of BP neural network based models for predicting tobacco leaf sensory quality [J]. *Acta Tabacaria Sinica*,2011,17(1):19 – 25.)

[3] 赵青松,李兴兵,唐小松. 基于支持向量机的烟叶感官品质评价[J]. 计算机工程与应用,2007,43(10):236 – 240.
(Zhao Qing-song, Li Xing-bing, Tang Xiao-song. Tabacum sensory evaluation based on the support vector machine [J]. *Computer Engineering and Applications*, 2007, 43 (10): 236 – 240.)

(下转第 25 页)