

多标号学习矢量量化的食用油掺伪检测

陈景波*

(常熟理工学院 电气与自动化工程学院, 江苏 常熟 215500)

(* 通信作者电子邮箱 ejbbjc@163.com)

摘要: 为了提高食用油掺伪检测效果, 基于食用油的高效液相色谱数据, 提出了一个新的多标号学习矢量量化算法 (ML-LVQ), 并应用于食用油的掺伪检测中。它每次调整两个原型使排序损失的上界最小, 并通过元标号分类器确定多标号的数目, 从而达到同时优化 ranking 准则函数和 bipartitions 准则函数的目的。在 9 类纯油以及它们的混合油样本的数据集上测试的结果表明, ML-LVQ 取得了比改进的 AdaBoost. RMH 算法更好的性能。

关键词: 多标号算法; 学习矢量量化算法; 元标号分类器; 高效液相色谱法; 食用油掺伪检测

中图分类号: TP181 **文献标志码:** A

Oil adulteration detection with multi-label learning vector quantization

CHEN Jingbo*

(School of Electrical and Automation Engineering, Changshu Institute of Technology, Changshu Jiangsu 215500, China)

Abstract: To improve the detection effect in oil adulteration, a new algorithm called ML-LVQ (Multi-Label Learning Vector Quantization) was proposed, which adapted Learning Vector Quantization (LVQ) to solve the multi-label learning problem on High Performance Liquid Chromatography (HPLC) data. It could minimize the upper bound of the ranking error, which would benefit the ranking measure. Moreover, the meta-labeler was used to identify the number of the labels for improving the bipartitions measure. The experimental results on nine classes of pure oil and their mixed oil samples show that the proposed algorithm is superior to the improved AdaBoost. RMH.

Key words: multi-label algorithm; Learning Vector Quantization (LVQ) algorithm; meta-labeler; High Performance Liquid Chromatography (HPLC) method; oil adulteration detection

0 引言

食用油的纯度问题由于涉及到食品安全和人类健康因而变得越来越重要。研究者发展出许多的分析技术来进行食用油的掺伪检测。典型的技术^[1]如理化检测方法、气相色谱方法、近红外光谱技术和高效液相色谱方法等。除了利用以上方法进行定性方法外, 还使用其他方法如主成分分析^[2]、线性判别分析^[3]、偏最小二乘方法^[4]等进行定量分析。

Huo 等^[5]使用高效液相色谱仪对纯油和混合油进行均衡等时采样。将掺伪的混合油看作是含有多个标号的样本, 而纯油仅含有一个标号。引入多标号 AdaBoost. RMH (Real AdaBoost with Multi-class Hamming Loss) 算法^[6-7]和改进的带元标号分类器的多标号 AdaBoost. RMH 对油样谱值特征数据进行分类。实验结果表明, 该算法不仅能对食用油是否为纯油进行区分, 还能有效地识别出混合油 (或纯油) 的组成成分, 并且改进的多标号算法能取得比经典的 AdaBoost. RMH 算法更好的性能。

为此, 本文基于学习矢量量化 (Learning Vector Quantization, LVQ)^[8-9]提出了一种新的多标号算法——多标号学习矢量量化 (Multi-Label Learning Vector Quantization, ML-LVQ) 算法, 并将该算法应用到食用油的掺伪检测中。相比于先前的多标号算法如 AdaBoost. RMH, ML-LVQ 算法充分地考虑到标号的排序问题, 结合元标号分类器, 能有效地解决多标号算法中确定阈值的问题。它通过调整原型的位置来近似地优化多标号的排序损失的上界, 从而减少多标号的排序

损失。元标号分类器^[10]的使用能有效地确定混合油成分的个数, 最小化标号的排序损失, 有助于帮助算法定性地确定混合油的各个成分的比重。

1 基于多标号分类的食用油检测

1.1 高效液相色谱仪采样

高效液相色谱仪^[11]的工作原理如下: 食用油待测物在不同的时间被注入色谱柱, 通过压力与固定相相互作用, 基于不同的相互作用, 检测器得到不同的物质发出的不同的峰信号, 通过分析对比可以得到这些食用油的种类。图 1 给出了 3 种不同种类的食用油的色谱图。从图 1 可以看出, 尽管 3 个谱图非常类似, 但仍然可以分辨出峰值的细微差别。

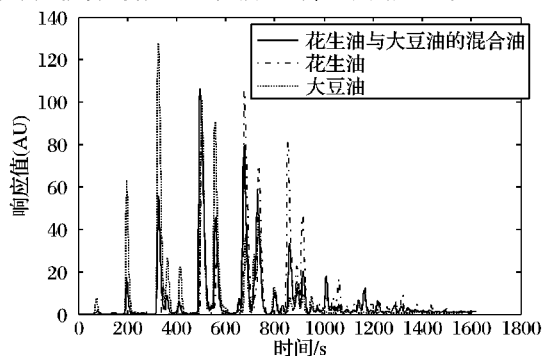


图1 三种食用油的色谱图

1.2 多标号 AdaBoost. RMH 算法

单标号分类主要关注样本只属于一个类别的情形, 而在

多标号分类^[12]中,样本可以属于多个类别,更极端的情况是,样本可以不属于任何类别或者属于所有的类别。给定标号集 Γ , 则任何样本都关联一个标号集 $S \subseteq \Gamma$ 和向量:

$$Y_l = \begin{cases} +1, & l \in S \\ -1, & \text{其他} \end{cases} \quad (1)$$

多标号 AdaBoost 算法能同时处理多个二元分类问题,它通过组合多个弱分类器(典型的如树桩分类器)来预测样本是否属于标号集 Γ 中的每一个,预测值的符号显示它是否属于特别的类别。给定样本 $\mathbf{x} \in \mathbf{R}^d$, AdaBoost 算法在不断变化的数据分布上产生一系列的弱假设 f_l , 最后得到最终的判别函数:

$$h_l(\mathbf{x}) = \sum_{i=0}^{T-1} \alpha_{i+1} f_{i+1}(\mathbf{x}); \quad l = 1, 2, \dots, L \quad (2)$$

1.3 评价准则

多标号的评价准则主要分为 bipartitions 准则和 ranking 准则。前者包括 micro-F1 和 macro-F1 准则,它在食用油检测中主要衡量了真实成分和预测成分之间的差别;而后者包括 avg-prec 和 one-error,主要用来计算标号的真实排序和预测排序之间的差别。它们的计算公式^[7]如下:

$$\text{mic-F1} = \frac{2TP}{2TP + FP + FN} \quad (3)$$

$$\text{mac-F1} = \frac{1}{L} \sum_{l=1}^L \frac{2TP_l}{2TP_l + FP_l + FN_l} \quad (4)$$

$$\text{one-error} = \frac{1}{N} \sum_{n=1}^N [h(\mathbf{x}_n) \notin t_n] \quad (5)$$

$$\text{avg-prec} = \sum_{n=1}^N \frac{1}{|t_n|} \sum_{l \in t_n} \frac{| \{ l' \in t_n \mid r(n, l') \leq r(n, l) \} |}{r(n, l)} \quad (6)$$

其中 TP_l , FP_l 和 FN_l 分别是正确正类、错误的正类和错误的负类的数目, TP , FP 和 FN 分别是它们的和, $r(n, l)$ 则表示第 n 个样本对于第 l 类的排序值。

2 多标号学习矢量量化算法

假设每一类 l 都对应一个原型向量的集合 $\{\mathbf{u}_{ls}, s = 1, 2, \dots, S\}$, 则给定样本 \mathbf{x}_n , 它关于类 l 的判别函数为:

$$g_l(\mathbf{x}_n) = \max_s \{ -\|\mathbf{x}_n - \mathbf{u}_{ls}\|^2 \} = -\|\mathbf{x}_n - \mathbf{u}_{ls}\|^2 \quad (7)$$

其中 $\|\cdot\|$ 表示欧氏距离。对于第 k 类的样本, 定义判别函数为 $g_k(\mathbf{x}_n) = -\|\mathbf{x}_n - \mathbf{u}_{ki}\|^2$, 那么关键串 $C = \{(k, l) \mid k \in t, l \notin t\}$ 上的排序损失为:

$$R = \frac{1}{N} \sum_n \sum_{k \in t_n, l \notin t_n} \frac{I[g_k(\mathbf{x}_n) \leq g_l(\mathbf{x}_n)]}{|t_n| L - |t_n|} \leq \frac{1}{N} \sum_n \sum_{k \in t_n} \frac{I[g_k(\mathbf{x}_n) \leq g_r(\mathbf{x}_n)]}{|t_n|} \leq \frac{1}{N} \sum_n \sum_{k \in t_n} \frac{2\sigma(\xi(g_r(\mathbf{x}_n) - g_k(\mathbf{x}_n)))}{|t_n|} \quad (8)$$

其中: $g_r(\mathbf{x}_n) = \max_{l \notin t_n} g_l(\mathbf{x}_n) = -\|\mathbf{x}_n - \mathbf{u}_{rj}\|$, $\sigma(\cdot)$ 是 sigmoid 函数 ($\xi > 0$)。

多标号学习矢量量化算法最小化正则化的排序损失的上界:

$$\tilde{R} = \sum_n \frac{1}{|t_n|} \sum_{k \in t_n} [\sigma(\xi d_k(\mathbf{x}_n)) + \alpha \|\mathbf{x}_n - \mathbf{u}_{ki}\|^2] \quad (9)$$

其中 $d(\mathbf{x}_n) = g_r(\mathbf{x}_n) - g_k(\mathbf{x}_n)$ 和 α 为正则化系数。令 $\varphi(\mathbf{x}_n) = \sigma(\xi d_k(\mathbf{x}_n))$, 则容易求得 $\varphi(\mathbf{x}_n)$ 的微分:

$$\begin{cases} \frac{\partial \varphi(\mathbf{x}_n)}{\partial g_{ki}} = -\xi \varphi(\mathbf{x}_n) (1 - \varphi(\mathbf{x}_n)) \\ \frac{\partial \varphi(\mathbf{x}_n)}{\partial g_{rj}} = -\frac{\partial \varphi(\mathbf{x}_n)}{\partial g_{ki}} \end{cases} \quad (10)$$

其中: $g_{ki} = -\|\mathbf{x}_n - \mathbf{u}_{ki}\|^2$, $g_{rj} = -\|\mathbf{x}_n - \mathbf{u}_{rj}\|^2$ 。给定样本 \mathbf{x}_n , 对于每个标号 $k \in t_n$ 仅仅需要更新两个原型向量:

$$\begin{cases} \mathbf{u}_{ki} = \mathbf{u}_{ki} - 2\eta(t) \left(\frac{\partial \varphi(\mathbf{x}_n)}{\partial g_{ki}} - \alpha \right) (\mathbf{x}_n - \mathbf{u}_{ki}) \\ \mathbf{u}_{rj} = \mathbf{u}_{rj} - 2\eta(t) \frac{\partial \varphi(\mathbf{x}_n)}{\partial g_{rj}} (\mathbf{x}_n - \mathbf{u}_{rj}) \end{cases} \quad (11)$$

其中 $\eta(t)$ 是第 t 次迭代的学习率。

为确定阈值, ML-LVQ 算法通过在新数据集 $\bar{D} = \{(\mathbf{x}_n, |t_n|)\}$ 上构造单标号分类器(即元标号分类器)来预测样本标号的数目。对于每个测试样本 \mathbf{x} , 它将输出该样本的标号数目 $k(\mathbf{x})$, 然后将判别值最高的 $k(\mathbf{x})$ 个类别作为标号集。算法如下:

算法1 多标号学习矢量量化算法。

初始化原型集合 \mathbf{m}_{is} ;

repeat

从训练集中随机选择一个样本;

for $k \in t_n$ do

从类别 k 的正类和负类中找到 \mathbf{x}_n 的最近邻原型向量 \mathbf{m}_{ki} 和 \mathbf{m}_{rj} ;

更新原型向量 \mathbf{m}_{ki} 和 \mathbf{m}_{rj} ;

$$\mathbf{u}_{ki} = \mathbf{u}_{ki} - 2\eta(t) \left(\frac{\partial \varphi(\mathbf{x}_n)}{\partial g_{ki}} - \alpha \right) (\mathbf{x}_n - \mathbf{u}_{ki})$$

$$\mathbf{u}_{rj} = \mathbf{u}_{rj} - 2\eta(t) \frac{\partial \varphi(\mathbf{x}_n)}{\partial g_{rj}} (\mathbf{x}_n - \mathbf{u}_{rj})$$

end for

until 算法收敛

输出: 将测试样本 \mathbf{x} 分给离它最近的原型所在的类别。

3 实验和结果

实验以中国不同地区的9种不同食用油以及它们的混合油为对象, 包括大豆油、棕榈油、芝麻油、玉米油、花生油、葵花油、米糠油、菜籽油和棉籽油, 每种食用油的样本数目见表1。数据集总共包含370个样本, 每个样本有1607维。每一维对应一个时刻的光谱响应值。

表1 实验样本

品种	样本数	品种	样本数	品种	样本数
大豆油	79	棕榈油	54	米糠油	24
花生油	88	芝麻油	117	纯油	246
葵花油	50	棉籽油	9	混合油	124
玉米油	15	菜籽油	58		

AdaBoost. LRMH 算法是一种改进的多标号 AdaBoost 算法, 它使用元标号模型通过预测测试样本的标号数目对经典的 AdaBoost. MH 算法的预测结果进行矫正, 将标号集中预测值最高的几个标号赋给测试样本。

本实验以 AdaBoost. LRMH 为参照, 比较了 ML-LVQ 算法与 AdaBoost. LRMH 算法在4个评价准则上性能。算法全部使用 Java 语言编写, 在 JDK1.6 环境下运行。为减少误差, 对数据集运行10次5-crossfold 取平均得到最终的性能值。算法使用 mic-F1 作为准则在1/3的训练集上验证超参数值。

两种算法均采用单标号 AdaBoost. RMH 作为元标号分类器, 其中弱分类器(树桩分类器)的数目取100。对于 ML-LVQ 而言, 它使用 K-means 聚类来初始化原型向量, 并通过将所有的属性归一化到 $[-1, +1]$ 上来保证随机梯度下降算法的稳定性, 缺省参数设置 $\eta(0) = 0.1 * cov$, $\eta(t) = \eta(0) * (1 - \frac{tN+n}{MN})$, $\alpha = 0$, $M = 40$, 其中 cov 是训练样本到最近邻的平均距离。AdaBoost. LRMH 算法选定超参数 T 来优化模型, 其

中 T 的取值范围为 $\{20, 40, 60, 80, 100\}$, 而 ML-LVQ 在集合 $\{1, 3, 5, 7, 9\}$ 上优化超参数 S 。

表 2 中测量运行时间: $S = 9, T = 100$ 。从中可知, ML-LVQ 算法在 4 个评价准则上都取得了比 AdaBoost. LRMH 更好的性能, 尤其在 $mac-F1$ 上 (从 90.97% 到 94.57%), 这表明 ML-LVQ 无论是在判别食用油 (纯油或混合油) 的组成成分方面还是在判别组成成分的比重方面都表现更优。

表 2 两种算法在 4 个评价准则上的比较

算法	评价准则/%				运行 时间/s
	one-error	avg-prec	mac-F1	mic-F1	
ML-LVQ	1.05	98.66	94.57	97.22	58.47
AdaBoost. LRMH	3.24	97.03	90.97	96.03	113.82

AdaBoost. RMH 算法的训练时间复杂度和测试时间复杂度分别为 $O(T(dN \log N + NL))$ 和 $O(TL)$, 而 ML-LVQ 的训练时间复杂度为 $O(NTLSd)$, 测试时间复杂度为 $O(LSd)$ 。表 2 的最后一栏显示了 ML-LVQ 取参数 $S = 9$ 和 AdaBoost. LRMH 取参数 $T = 100$ 时的运行时间比较。显然, ML-LVQ 在数据集上运行得更快一些, 近似地比后者快一倍。

4 结语

本文提出了一个多标号学习矢量量化算法 (ML-LVQ) 应用于食用油的掺伪检测, 它通过最小化排序损失的上界来优化多标号的排序 (ranking) 准则函数, 利用元标号分类器提高二分 (bipartition) 准则函数。前者能将混合油中的各成分按比重大小正确排序, 而后者能有效地区分混合油或纯油的组成成分。在 9 类食用油以及它们的混合所组成的样本上测试的结果显示, ML-LVQ 在 4 种准则函数上均取得了比改进的多标号 AdaBoost. RMH 算法更优的性能。

参考文献:

- [1] 宋玉峰, 王微山, 杨学军, 等. 食用油掺假检测方法研究进展[J]. 中国食物与营养, 2012, 18(3): 9-12.

(上接第 3110 页)

参考文献:

- [1] SUN J, FENG B, XU W B. Particle swarm optimization with particles having quantum behavior [C]// Proceedings of 2004 Congress on Evolution Computation. Piscataway: IEEE Press, 2004: 325-331.
- [2] 方伟, 孙俊, 谢振平, 等. 量子粒子群优化算法的收敛性分析及控制参数研究[J]. 物理学报, 2010, 59(6): 3686-3694.
- [3] 龙海侠, 须文波, 王小根, 等. 基于选择操作的量子粒子群算法[J]. 控制与决策, 2010, 25(10): 1499-1506.
- [4] SUN J, XU W B, FANG W. Quantum-behaved particle swarm optimization with a hybrid probability distribution [C]// Proceedings of the 9th Pacific Rim International Conference on Artificial Intelligence, LNCS 4099. Berlin: Springer-Verlag, 2006: 737-746.
- [5] LIU J, SUN J, XU W B. Improving quantum-behaved particle swarm optimization by simulated annealing [C]// Proceedings of the 2006 International Conference on Computational Intelligence and Bioinformatics, LNCS 4115. Berlin: Springer-Verlag, 2006: 130-136.
- [6] LIU J, SUN J, XU W B. Quantum-behaved particle swarm optimization with immune memory and vaccination [C]// Proceedings of the 2006 IEEE International Conference on Granular Computing. Piscataway: IEEE Press, 2006: 453-456.

- [2] CHRISTOPHER M. Pattern recognition and machine learning [M]. Berlin: Springer, 2006: 561-569.
- [3] HAI Z, WANG J. Detection of adulteration in camellia seed oil and sesame oil using an electronic nose [J]. European Journal of Lipid Science and Technology, 2006, 108(2): 116-124.
- [4] PATRICIA C M, JANSSEN H G, IRENE B M, et al. The use of multivariate modelling of near infrared spectra to predict the butter fat content of spreads [J]. Analytica Chimica Acta, 2007, 595(1/2): 176-181.
- [5] HUO Q G, JIN X B, ZHANG H M. Multi-label classification for oil authentication [C]// FSKD 2012: Proceedings of the 2012 International Conference on Fuzzy Systems and Knowledge Discovery. Piscataway: IEEE Press, 2012: 711-714.
- [6] SCHAPIRE R E, SINGER Y. Improved boosting algorithms using confidence-rated predictions [J]. Machine Learning, 1999, 37(3): 297-336.
- [7] SCHAPIRE R E, SINGER Y. Boostexter: a boosting-based system for text categorization [J]. Machine Learning, 2000, 39(2/3): 135-168.
- [8] JIN X B, HOU X W, LIU C L. Prototype learning with margin-based conditional log-likelihood loss [C] // ICPR 2008: Proceedings of the 2008 International Conference of Pattern Recognition. Piscataway: IEEE Press, 2008: 1-4.
- [9] JIN X B, LIU C L, HOU X W. Regularized margin-based conditional log-likelihood loss for prototype learning [J]. Pattern Recognition, 2010, 43(7): 2428-2438.
- [10] TANG L, RAJAN S, NARAYANAN V K. Large scale multi-label classification via metalabeler [C]// WWW'09: Proceedings of the 18th International Conference on World Wide Web. Piscataway: IEEE Press, 2009: 211-220.
- [11] 郭涛, 杜蕾蕾, 万辉, 等. 高效液相色谱法测胆固醇含量鉴别地沟油[J]. 食品科学, 2009, 30(22): 286-289.
- [12] TSOU MAKAS G, KATAKIS I, VLAHAVAS I. Mining multi-label data [M]. Berlin: Springer, 2010: 667-685.

- [7] COELHO L S. Novel Gaussian quantum-behaved particle swarm optimizer applied to electromagnetic design [J]. IET Science, Measurement and Technology, 2007, 11(2): 290-294.
- [8] 林星, 冯斌, 孙俊. 混沌量子粒子群优化算法[J]. 计算机工程与设计, 2008, 29(10): 2610-2612.
- [9] 许少华, 王皓, 王颖, 等. 一种改进的量子粒子群优化算法及其应用[J]. 计算机工程与应用, 2011, 47(20): 34-37.
- [10] XI M L, SUN J, XU W B. An improved quantum-behaved particle swarm optimization algorithm with weighted mean best position [J]. Applied Mathematics and Computation, 2008, 205(5): 751-759.
- [11] 许磊. 电力系统经济调度的放置研究[J]. 计算机仿真, 2012, 29(9): 324-327.
- [12] 白俊强, 尹戈玲, 孙智伟. 基于二阶振荡及自然选择的随机权重混合粒子群算法[J]. 控制与决策, 2012, 27(10): 1459-1464.
- [13] 刘耀年, 尹洪全, 张伟, 等. 基于改进粒子群算法的配电网状态估计[J]. 电测与仪表, 2012, 49(9): 24-27.
- [14] JIAN C L, ZHI H G, HUI Q W. Research on the application of the wavelet neural network model in peak load forecasting considering of the climate factors [C]// Proceedings of the 4th International Conference on Machine Learning and Cybernetics. Piscataway: IEEE Press, 2005: 538-543.