ОПРЕДЕЛЕНИЕ КОНТРАФАКТНЫХ АЛКОГОЛЬНЫХ НАПИТКОВ С ПОМОЩЬЮ КЛАСТЕРНОГО АНАЛИЗА В ПРОСТРАНСТВЕ ГЛАВНЫХ КОМПОНЕНТ ОПТИЧЕСКИХ СПЕКТРОВ ПРОПУСКАНИЯ

М. А. Ходасевич [1*], Г. В. Синицын [1], М. А. Гресько [2], В. М. Доля [2], М. В. Роговая [1], А. В. Казберук [1]

УДК 535.243:663.2

[1] Институт физики им. Б. И. Степанова НАН Беларуси, 220072, Минск, просп. Независимости, 68, Беларусь; e-mail: m.khodasevich@ifanbel.bas-net.by
[2] ООО "БИОСАН-Алкос", Москва, Россия

(Поступила 1 декабря 2016)

На примере исследования 153 промышленно выпускаемых сортов водочной продукции показана возможность решения задачи выявления контрафактных образцов с помощью введения унифицированной добавки с минимально допустимой для приборного обнаружения концентрацией и многопараметрического анализа измеренных в УФ и видимом диапазонах спектров пропускания. Путем применения иерархической кластеризации или метода С-средних в двумерном пространстве главных компонент достигнута 100 %-ная вероятность обнаружения контрафактной продукции.

Ключевые слова: *спектроскопия в ультрафиолетовой и видимой области спектра, метод главных компонент, кластерный анализ.*

Based on a study of 153 kinds of commercial vodka products, the possibility of identifying counterfeit samples is shown by introducing a unified additive with the concentration that is minimum acceptable for the instrumental detection and multivariate analysis of UV-Vis transmission spectra. The 100% probability of detection of counterfeit products is achieved through the use of hierarchical clustering analysis or C-means method in two-dimensional space of principal components.

Keywords: *UV-Vis-spectroscopy, principal component analysis, cluster analysis.*

Алкогольная продукция в зависимости от ее качества и потребленного количества может нанести вред здоровью и даже стоить жизни потребителям. Приблизительная оценка ежегодного количества смертей во всем мире по вине алкоголя составляет ~1.8 млн человек, или ~3 % [1]. В странах западной Европы этот показатель почти в два раза выше (~6 %) [2], а в странах центральной и восточной Европы — самый высокий в мире [3] (от 13 % в Польше до 25 % в Венгрии). Статистические данные по Беларуси и России сильно зависят от методики подсчета и составляют 3—19 % для Беларуси и 16—37 % для России [4]. Поэтому актуально понижение уровня смертности, связанной с потреблением алкоголя.

Одна из задач в рамках этой проблемы — выявление контрафактной алкогольной продукции на всех этапах ее производства и потребления: от завода-производителя до конечного потребителя. Применяемые методы борьбы с фальсификацией алкогольной продукции в основном касаются факторов, являющихся внешними по отношению к содержанию бутылок: акцизных марок, дозаторов, формы бутылки, наклеек, температурных меток, наличия лазерной маркировки уникального номера бутылки и др. Этими методами не защищается содержание бутылок — собственно сам алкогольный

IDENTIFICATION OF COUNTERFEIT ALCOHOLIC BEVERAGES USING CLUSTER ANALYSIS IN THE SPACE OF THE PRINCIPAL COMPONENTS OF THE OPTICAL TRANSMISSION SPECTRA

M. A. Khodasevich [1*], G. V. Sinitsyn [1], M. A. Gresko [2], V. M. Dolya [2], M. V. Rogovaya [1], A. V. Kazberuk [1] (*[1] B. I. Stepanov Institute of Physics, National Academy of Sciences of Belarus, 68 Nezavisimosti Prosp., Minsk, 220072, Belarus; e-mail: m.khodasevich@ifanbel.bas-net.by; [2] "BIOSAN-Alkos" Ltd., Moscow, Russia*)

напиток. Непосредственная защита напитка должна быть наиболее устойчива к фальсификации и учитывать потребительские интересы.

Состав производимых на территории Республики Беларусь и Российской Федерации водок характеризуется разнообразным ассортиментом вносимых добавок: сахар; натрий двууглекислый; кислоты уксусная, молочная, лимонная, соляная; мед; соль; калий марганцовокислый; крахмал картофельный; ароматные спирты и настои, получаемые из пищевого сырья; эфирные масла и ароматизаторы; пищевые добавки [4]. Такой ассортимент добавок наряду с возможными изменениями используемых спиртов и питьевой воды приводит к допустимому по технологическим инструкциям, рецептурам и нормативным документам разбросу показателей водок. Таким образом, выполнение требований нормативных документов [5] на физико-химические показатели водок (щелочность ≤2.0—3.0 $см^3$ в зависимости от используемого спирта, массовые концентрации уксусного альдегида ≤3—8 мг, сивушного масла ≤5—6 мг, сложных эфиров ≤5—13 мг, объемная доля метилового спирта ≤0.003—0.03 %) и легальный характер производства не приводят к абсолютной повторяемости их характеристик.

Для определения контрафактной водки с помощью измерений физико-химических параметров продукции, выпускаемой в пределах РБ и РФ в существующих условиях и при выполнении требований актуальных нормативных документов, требуется наличие периодически обновляемой базы данных по характеристикам каждого вида изготавливаемой алкогольной продукции, что практически неосуществимо. Возможный путь решения задачи защиты водки от подделки — унификация одной удовлетворяющей требованиям нормативных документов добавки в рецептурах всей легальной продукции. Выявление состава такой добавки с минимально допустимой для приборного обнаружения концентрацией затруднено вследствие широкого ассортимента ингредиентов водок.

Для выявления контрафактной алкогольной продукции на всех этапах ее производства и потребления необходимо применение недорогой и надежной технологии желательно при минимальных требованиях к пробоподготовке. Такой технологией может быть спектроскопия УФ, видимого и ближнего ИК диапазонов с последующей хемометрической обработкой. Спектроскопия стала стандартной технологией в фармацевтической [6], нефтехимической [7], пищевой [8—10] промышленности, биомедицине [11, 12] и т. д. из-за простоты, относительной дешевизны использования, экспрессного характера измерений и повышения качества получаемых результатов вследствие применения многопараметрических методов обработки информации. Эти методы позволяют обнаруживать и анализировать скрытые связи в огромных массивах спектральных данных. В настоящей работе применяется такой метод многопараметрического анализа, как метод главных компонент (МГК) [13] с последующим кластерным анализом. В последнее время МГК успешно используется для классификационных испытаний пищевых продуктов и алкогольных напитков с помощью спектроскопии комбинационного рассеяния [14], видимого [15], ближнего [16, 17] и среднего [18] ИК диапазонов, обработки изображения в сверхширокой спектральной области [19], колориметрии рассеянного света [20].

В настоящей работе спектры пропускания 153 образцов водок производства РБ и РФ измерены с помощью малогабаритного спектрометра OceanOptics 650 UV-Vis (650 активных пикселей, спектральное разрешение 2 нм, максимальное отношение сигнал/шум 300:1) в диапазоне 200—850 нм. Предварительная обработка спектров заключалась в сглаживании фильтром Савицкого—Голея полиномом третьей степени по девяти точкам. Водки с универсальной добавкой моделировали собой легально выпускаемые алкогольные напитки, без нее — контрафактные. Из-за большого количества образцов на рис. 1 представлены только усредненные спектры водок с универсальной добавкой и без нее, а также спектры, максимально отличающиеся от усредненных. Видно существенное перекрытие спектров водок с добавкой и без нее.

К матрице исходных спектральных данных размером 306 (количество образцов) на 651 (количество спектральных отсчетов) применен МГК — широко используемый метод анализа многопараметрических данных, предназначенный не только для исследовательского анализа больших массивов информации, но и для поиска выбросов, понижения размерности (ранга) данных, графического (маломерного) представления разделения данных на кластеры, проведения классификации и регрессии. МГК применен к центрированным матрицам зарегистрированных спектров пропускания. По зависимости суммарной дисперсии от количества главных компонент можно сделать вывод о достаточности двумерного пространства (79.4 % дисперсии данных в первой главной компоненте и 13.0 % во второй, суммарно в двух 92.4 %) для моделирования существенных различий в рассматриваемой выборке водок.

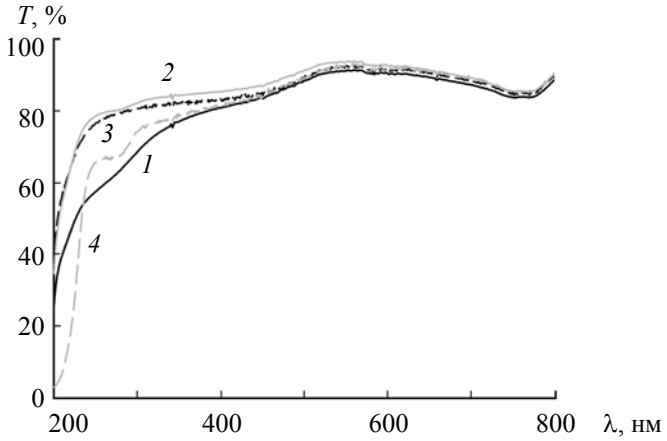

Рис. 1. Усредненные спектры 153 образцов водок с универсальной добавкой (*1*) и без нее (*2*) и спектры водок, максимально отличающиеся от усредненных (*3* и *4*)

После понижения размерности данных задача выявления контрафактной продукции может быть решена путем определения степени сходства исследуемых объектов с помощью кластерного анализа — статистической процедуры, упорядочивающей объекты в однородные группы по некоторым общим признакам [21]. В общем случае количество групп неизвестно и кластеризация относится к широкому классу задач обучения без учителя. В рассматриваемой задаче количество групп известно: две — легальная и контрафактная продукция. Поэтому результатом кластерного анализа должна быть классификация всех объектов на две непересекающиеся группы, чтобы каждая состояла из объектов, близких по используемой метрике.

Решение задачи классификации неоднозначно по нескольким причинам. Во-первых, не существует наилучшего универсального критерия качества классификации. Имеется ряд эвристических критериев, а также алгоритмов, не имеющих четко выраженного критерия, но осуществляющих достаточно качественную классификацию. При этом все они могут давать разные результаты. Во-вторых, результат классификации существенно зависит от метрики, выбор которой, как правило, субъективен.

Из методов классификации в настоящей работе использованы базирующийся на вероятностном подходе метод K-средних; иерархический анализ, предполагающий наличие вложенных групп различного порядка; метод разделения гауссовых многопараметрических смесей; основанный на применении нечеткой логики метод C-средних. Несмотря на значительные различия, перечисленные методы опираются на априорную гипотезу компактности расположения объектов: в пространстве главных компонент все близкие объекты должны относиться к одной группе, а все различные объекты находиться в разных группах.

При применении эвклидова расстояния и при одном из трех начальных выборов центроидов кластеров (случайный выбор сразу по всему набору данных, однородный выбор или использование только 10 % части выборки образцов) в вероятностном подходе K-средних общая сумма расстояний до центроидов неизменна. Среди рассмотренных мер расстояний (эвклидово, расстояние корреляции, сумма абсолютных разностей координат объектов в пространстве главных компонент и косинусная мера — разность единицы и косинуса угла между векторами из начала координат к точкам-образам образцов в пространстве пониженной размерности) лучшие результаты получаются при использовании косинусной меры — лишь 33 образца, моделирующих легальную продукцию, ошибочно определяются как контрафактные.

Для агломеративного иерархического анализа при использовании разных методов определения схожести объектов и различных мер расстояний (эвклидово, квадратичное эвклидово, Махаланобиса, городских кварталов, Минковского, косинусное, Чебышева) наименьшее количество ошибок наблюдается в случае метода Уорда (определение взвешенного квадрата расстояния между центроидами) с квадратичной эвклидовой мерой. Расстояние d_{ij} между образцами i и j в этом случае определяется как $d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)D^{-1}(\mathbf{x}_i - \mathbf{x}_j)'$, где D — матрица дисперсии выборки.

При классификации на две группы методом Уорда с применением квадратичной эвклидовой меры правильно определены все 153 образца, моделирующие контрафактные водки, и 121 образец, моделирующий легальную продукцию. Согласно [21], метод Уорда чаще других восстанавливает инту-

итивно наилучшую кластеризацию. Классификация исследуемой выборки выполнена также с помощью метода разделения гауссовых смесей. Правильно классифицированы 144 образца, моделирующие легальную водку, и 146 — контрафактную. Общая погрешность ~5 %. Использован также базирующийся на применении нечеткой логики метод классификации С-средних. Как и в случае иерархической кластеризации, вероятность обнаружения контрафактной продукции 100 %, легальной 79 %. Такой случай абсолютно достоверного обнаружения контрафактной продукции и лишь вероятностного легальной аналогичен скрининговым исследованиям в медицине, когда болезнь необходимо обнаружить обязательно, а ложный результат исследований здорового человека можно перепроверить иным методом. Следовательно, лучшие с точки зрения прикладных применений результаты классификации легальных и контрафактных водок в рассмотренном случае достигаются при применении иерархического метода или метода С-средних.

Таким образом, в процессе определения контрафактных и легальных водок, моделируемых с помощью введения унифицированной пищевой добавки с минимально допустимой для приборного обнаружения концентрацией, получена 100 %-ная точность выявления контрафактной продукции с использованием классификации образцов напитков в двумерном пространстве главных компонент методами иерархического анализа или С-средних. Применение описанного подхода к определению контрафактной водочной продукции позволит предпринять потенциально выполнимую попытку снижения уровня смертности, связанной с потреблением алкоголя.

[1] **G. Edwards.** Addiction, **92** (1997) 73—79
[2] **J. Rehm, B. Taylor, J. Patra.** Addiction, **101** (2006) 1086—1095
[3] **J. Rehm, U. Sulkowska, M. Manchuk.** Int. J. Epidemiol., **36** (2007) 458—467
[4] **Ю. Е. Разводовский.** Вопросы организации и информатизации здравоохранения, № 2 (2011) 15—20
[5] Межгосударственный стандарт ГОСТ 12712-2013 “Водки и водки особые. Общие технические условия”, Москва, Стандартинформ (2014)
[6] **M. Blanco, A.Peguero.** Trend. Anal. Chem., **29** (2010) 1127—1136
[7] **F. B. Gonzaga, C. Pasquini.** Anal. Chim. Acta, **670** (2010) 92—97
[8] **Lijuan Xie, Xingqian Ye, Donghong Liu, Yibin Ying.** Food Chem., **114** (2009) 1135—1140
[9] **D. Cozzolino, M. J. Kwiatkowski, R. G. Dambergs, W. U. Cynkar, L. J. Janik, G. Skouroumounis, M. Gishen.** Talanta, **74** (2008) 711—716
[10] **М. В. Роговая, Г. В. Синицын, М. А. Ходасевич.** Опт. и спектр., **117** (2014) 165—169
[11] **S. Kasemsumran, Y. Du, K. Maruo, Y. Ozaki.** Chemom. Intellig. Laboratory Systems, **82** (2006) 97—103
[12] **Z. Chuah, R. Paramesran, K. Thambiratnam, S. Poh.** Chemom. Intellig. Laboratory Systems, **104** (2010) 347—351
[13] **К. Эсбенсен.** Анализ многомерных данных, Черноголовка, изд-во ИПХН РАН (2005)
[14] **C. Frausto-Reyes, C. Medina-Gutierrez, R. Sato-Berru, L. R Sahagun.** Spectrochim. Acta, A, **61** (2005) 2657—2662
[15] **Li Xiaoli, He Yong, Fang Hui.** J. Food Eng., **81** (2007) 357—363
[16] **P. Heussen, H.-G. Janssen, I. Samwel, J. van Duynhoven.** Anal. Chim. Acta, **595** (2007) 176—181
[17] **O. Rodionova, L. Houmøller, A. Pomerantsev, P. Geladi, J. Burger, V. Dorofeyev, A. Arzamastsev.** Anal. Chim. Acta, **549** (2005) 151—158
[18] **D. Cozzolino, M. Holdstock, R. Dambergs, W. Cynkar, P. Smith.** Food Chem., **116** (2009) 761—765
[19] **A. Gowen, M. Taghizadeh, C. O’Donnell.** J. Food Eng., **93** (2009) 7—12
[20] **A. G. Mignani, L. Ciaccheri, H. Thienpont, H. Ottevaere, A. Cimato, C. Attilio.** Proc. Symposium on Photonics Technologies for 7th Framework Program (2006) 488—491
[21] **И. Д. Мандель.** Кластерный анализ, Москва, Финансы и статистика (1988)