КЛАССИФИКАЦИЯ АНАЛЬГЕТИЧЕСКИХ ЛЕКАРСТВЕННЫХ ПРЕПАРАТОВ В ПЕРВИЧНОЙ УПАКОВКЕ ПО СПЕКТРАМ ТЕРАГЕРЦОВОГО ДИАПАЗОНА С ИСПОЛЬЗОВАНИЕМ МНОГОПАРАМЕТРИЧЕСКИХ МЕТОДОВ

А. В. Ляхнович, Г. В. Синицын, М. А. Ходасевич *, Д. А. Борисевич

УДК 543.42:615.45

Институт физики НАН Беларуси,
Минск, Беларусь; e-mail: m.khodasevich@ifanbel.bas-net.by

(Поступила 22 июля 2021)

Выполнен сравнительный анализ эффективности нескольких многопараметрических методов распознавания классов образцов лекарственных средств в первичной упаковке по спектрам пропускания терагерцового диапазона. Показано преимущество метода опорных векторов, демонстрирующего ошибку классификации ≤5.7 %, для двух групп анальгетических препаратов с высокой степенью неоднородности образцов в группах.

Ключевые слова: *метод главных компонент, метод построения деревьев классификации и регрессии, линейный дискриминантный анализ, метод опорных векторов, терагерцовая спектроскопия.*

We compared the efficiency of some multivariate methods of recognition of drug classes in primary packages using THz transmission spectra. The advantages of the support vector machine are demonstrated with a classification error not exceededing 5.7% for two groups of analgetic drugs with a high level of heterogeneity of the samples in these groups.

Keywords: *principal component analysis, classification and regression tree, linear discriminant analysis, support vector machine, THz spectroscopy.*

Введение. В последние десятилетия активно разрабатываются методы и устройства бесконтактной диагностики на основе регистрации как изображений, так и спектров в терагерцовом (ТГц) диапазоне [1—6]. Спектроскопические методы анализа качества промышленных изделий, в том числе их состава, с использованием ТГц-излучения могут существенно расширить области применения сравнительно новых перспективных технологий и обеспечить безопасность продукции и, как следствие, повышение качества жизни населения. При этом используются отличительные свойства излучения данного диапазона. В частности, такая особенность, как способность проникать сквозь определенные виды оптически непрозрачных материалов, позволяет диагностировать целостность продукции, а также идентифицировать ее состав на соответствие требованиям технических условий непосредственно в упаковке [2]. При этом, например, для фармацевтических препаратов отсутствие необходимости вскрытия упаковки для проведения анализа позволяет избежать снижения качества лекарственных средств и прямых потерь, что в ряде случаев (ограниченных партий, дорогостоящих препаратов) может быть определяющим для обеспечения потребителя продукцией высокого качества.

При необходимости рассмотрения большого количества существенно различающихся образцов, принадлежащих ограниченному числу классов, эффективными могут оказаться многопараметрические методы анализа спектральных данных [5, 6]. Преимущество рассматриваемого подхода — исследование не специально подготовленных для лабораторных измерений образцов [7], а готовых форм промышленного производства без извлечения из упаковки.

CLASSIFICATION OF ANALGETIC DRUGS IN PRIMARY PACKAGES BY APPLYING MULTIVARIATE METHODS TO TERAHERTZ SPECTRA

A. V. Lyakhnovich, G. V. Sinitsyn, M. A. Khodasevich *, D. A. Borisevich (*B. I. Stepanov Institute of Physics of the National Academy of Sciences of Belarus, Minsk, Belarus; e-mail: m.khodasevich@ifanbel.bas-net.by*)

Измерения и образцы. Для измерения спектров пропускания образцов лекарственных средств использован импульсный спектрометр ТГц-диапазона реального времени T-Spec (TeraVil Ltd., Вильнюс, Литва), работающий совместно с источником фемтосекундных лазерных импульсов с λ = 1.03 мкм Flint (Light Conversion). Прибор использует принцип спектроскопии во временной области (time-domain) и способен регистрировать временной профиль в течение 100 мс. Последнее позволяет без существенных временных затрат проводить усреднение данных по 100 измерениям для каждого образца. Спектры пропускания образцов, вычисляемые через преобразование Фурье с помощью комплектного программного обеспечения для T-Spec, применялись для визуальной оценки корректности измерения. Для дальнейшего анализа использованы временные профили сигнала, пропорционального напряженности электрического поля ТГц-излучения. Особенности размещения образцов в ТГц-тракте спектрометра, которые связаны как со схемой фокусировки широкополосного излучения, так и с геометрией таблетированной формы образцов, аналогичны описанным в [6].

В качестве модельных объектов выбраны два класса распространенных лекарственных средств, принадлежащих группе анальгетических препаратов, — 4-ацетамидофенол (парацетамол) и ацетилсалициловая кислота (АСК). Дозировка препаратов 500 мг, при этом масса таблеток без упаковки 0.53—0.55 г для парацетамола и 0.57—0.61 г для АСК. Упакованы образцы лекарственных форм в ленточную композитную упаковку “конвалюта” на бумажной основе [8]. Таблетки различались по году изготовления (2010—2021 гг.), производителю (РУП “Белмедпрепараты”, ОАО Борисовский ЗМП), цвету маркировки и плотности заполнения маркировочными надписями апертуры образца. На упаковку некоторых образцов наклеены бумажные ценники предприятия торговли. Несколько таблеток, изначально упакованных производителем в пластиково-фольгированные блистеры, перемещены в освобожденные фрагменты “конвалют” непосредственно перед измерениями. Ряд образцов, извлеченных из бумажной упаковки, перемещен в упаковку препарата другого класса. Некоторые образцы хранились в лабораторных условиях вне упаковок на протяжении нескольких месяцев. Указанные действия осуществлялись как для расширения разнообразия обучающей базы многопараметрических методов, так и для проверки их эффективности в условиях фальсификации продукции и упаковки.

Измерения пропускания пустой упаковки демонстрируют снижение амплитуды сигнала временного профиля в ~2 раза в отсутствие выраженных спектральных особенностей. Соответственно, динамический диапазон измерений пропускания образцов препаратов по интенсивности сужался вчетверо по сравнению с результатами [6].

Результаты и их обсуждение. Исследованы 53 образца таблеток — 34 АСК и 19 парацетамола. Полученные с помощью быстрого преобразования Фурье спектры в диапазоне 0.2—1.1 ТГц представлены на рис. 1. Спектры пропускания образцов приведены к опорному сигналу, динамический диапазон ограничен шумом на уровне ~10^{-2}. Видно, что стандартная идентификация препаратов по различиям их ТГц-спектров в выбранной модельной паре проблематична из-за частичного перекрытия спектров. Существенно снизить погрешность идентификации лекарств с близкими спектральными свойствами способны многопараметрические методы обработки результатов измерений.

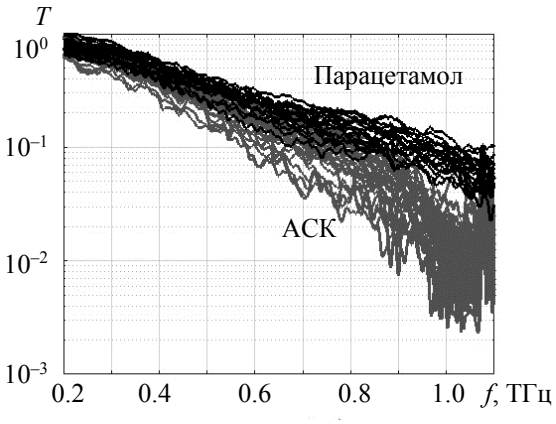

Рис. 1. Спектры пропускания исследуемых лекарственных веществ

При обработке спектров методом главных компонент (ГК) [9] исследована зависимость суммарной объясненной дисперсии от количества ГК. Определено, что дальнейший многопараметрический анализ следует проводить в пространстве четырех ГК, которое описывает 95.9 % дисперсии данных.

График счетов в первые две ГК, описывающие 89.1 % дисперсии спектров, представлен на рис. 2. Видно существенное перекрытие образцов двух рассматриваемых лекарственных препаратов. Образцы парацетамола 35 и 36 попали в область подпространства PC1-PC2, нахождение в которой характерно для таблеток АСК. Показательно, что эти образцы перед проведением измерений были вынуты из фабричных блистерных упаковок и помещены в конвалютные упаковки АСК.

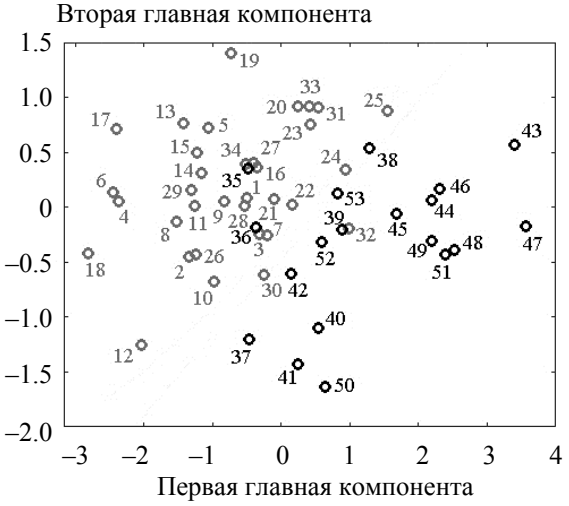

Рис. 2. График счетов спектров исследуемых таблеток в первые две главные компоненты

Для проведения классификации образцов по расположению в пространстве ГК спектров пропускания сравним три многопараметрических метода.

Первый метод — метод построения деревьев классификации и регрессии (CART) [10] — представляет собой структурированный в виде дерева набор бинарных правил классификации используемых данных, в нашем случае ГК. CART, как и другие методы многопараметрического анализа, может строиться с помощью кросс-валидации набора данных. Используем кросс-валидацию по 10 % всего количества образцов, реализуемую в пакете MatLab по умолчанию. Иерархическая структура правил приводит к решению о классификации, если оно может быть принято однозначно. Такое решение имеет аналогией лист дерева, положение которого однозначно определяется структурой разделения ветвей, начинающейся от ствола (весь массив данных) и заканчивающейся ветвью, на которой расположен рассматриваемый лист. Из существующих алгоритмов построения CART [11] для бинарной классификации наиболее часто используется построение всех возможных гиперплоскостей, делящих пространство входных переменных на две части, с последующим выбором разбиения, минимизирующего ошибку нахождения объекта одного класса в подпространстве другого. Следующий шаг алгоритма CART аналогично оперирует с подпространством, характеризующимся большей ошибкой классификации. Ограничивающими работу алгоритма условиями могут быть либо количество ошибочно классифицированных объектов, либо количество узлов принятия решений. Построенное дерево решений применяется для проведения кросс-валидации с целью избежать переобучения. На этом этапе обычно происходит уменьшение количества узлов принятия решений или уточнение правил переходов. Хорошая интерпретируемость модели CART позволяет игнорировать ее недостаток, который заключается в возможности получения локального и необязательно оптимального решения.

При ограничении количества узлов принятия решений шестью в пространстве четырех ГК построенному с помощью 10 % кросс-валидации дереву решений оказалось достаточно четырех узлов, не учитывающих PC3. При этом из 53 образцов неправильно классифицировано 8, т. е. вероятность ошибки 15.1 %.

Для разделения образцов на два класса в пространстве ГК также применим линейный дискриминантный анализ LDA [12]. Цель LDA — нахождение линейной функции, максимизирующей отношение дисперсий классов образцов и минимизирующей внутриклассовые дисперсии. Хорошие результаты применения LDA достигаются в условиях дискриминации классов, линейно разделенных в пространстве входных переменных, что неверно в рассматриваемом случае. Однако и в случае перекрывающихся классов при условии достаточного удаления их центроидов друг от друга и существенного различия дисперсии всей выборки образцов от дисперсий классов LDA позволяет получить малую вероятность ошибки классификации. Воспользуемся также обобщением LDA — квадратичным дискриминантным анализом QDA. QDA использует для разделения классов гиперповерхности второго порядка. Результаты применения LDA/QDA представлены на рис. 3. LDA ошибочно классифицирует пять образцов (вероятность ошибки 9.4 %), QDA — шесть (11.3 %).

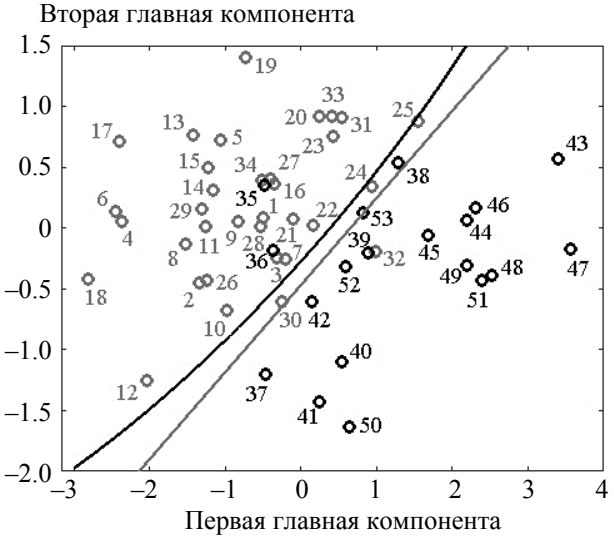

Рис. 3. Результаты применения LDA/QDA в пространстве PC1-PC2

Третий метод классификации — метод опорных векторов (SVM) [13] — представляет собой метод получения оптимальной границы двух классов в векторном пространстве независимо от вероятностных распределений векторов обучающей выборки. Цель SVM — поиск наиболее удаленной от обоих классов гиперплоскости, которая разделяет классы не только в обучающей выборке, но и при проведении проверки или кросс-валидации. В рассматриваемом случае проверка проводится путем кросс-валидации по 10 % образцов из всей исследуемой выборки.

Задача построения разделяющей гиперплоскости математически формулируется следующим образом. 53 вектора $\mathbf{x}_i$ в 4-мерном пространстве ГК, полученных из 802-мерных спектральных данных, характеризуются по принадлежности к классам переменной $y_i = \pm 1$. Разделяющая классы гиперплоскость определяется уравнением $\mathbf{wx} - b = 0$, где $\mathbf{w}$ — нормаль к ней. Построение оптимальной разделяющей гиперплоскости сводится к минимизации $\|\mathbf{w}\|$.

В рассматриваемом случае неразделимости данных метод SVM учитывает наличие ошибок в обучении. Набор дополнительных переменных $\xi_i \geq 0$ характеризует ошибку при классификации каждого объекта:

$$\begin{cases} \dfrac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{4}\xi_i \to \min_{w,b,\xi_i} \\ y_i(\mathbf{w}x_i - b) \geq 1 - \xi_i,\ 1 \leq i \leq 4 \\ \xi_i \geq 0,\ 1 \leq i \leq 4 \end{cases}$$

Коэффициент C позволяет при построении SVM модели регулировать соотношение между шириной гиперполосы, разделяющей классы, и суммарной ошибкой.

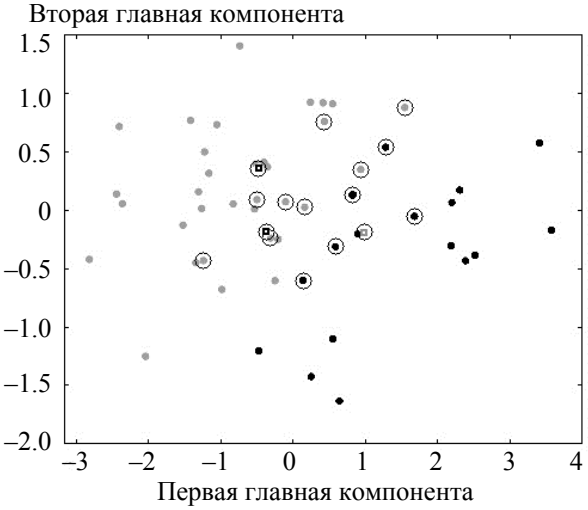

Рис. 4. Счета спектров исследуемых таблеток в PC1-PC2 и модель SVM с результатами классификации; ○ — опорные векторы, • — правильно классифицированные, □ — неправильно классифицированные

На рис. 4 представлен график счетов в PC1-PC2 с указанием положения опорных векторов и правильно и неправильно классифицированных образцов. Видно, что неправильно классифицированы два образца парацетамола и один АСК. Суммарная ошибка классификации 5.7 %.

Заключение. Продемонстрирована достаточно высокая эффективность многопараметрических методов анализа спектральных данных для классификации лекарственных средств без извлечения из первичной упаковки типа конвалюта по слабо различающимся спектрам пропускания терагерцового диапазона. Лучшее качество классификации продемонстрировано при использовании метода опорных векторов, который классифицирует образцы с ошибкой 5.7 % (три образца из 53) в отличие от линейного и квадратичного дискриминантного анализа (ошибки 9.4 и 11.3 % соответственно) и метода построения деревьев классификации (ошибка 15.1 %). Однако в каждом индивидуальном случае предполагаемого конкретного набора препаратов выбор оптимальной многопараметрической модели может отличаться от построенной модели для пары парацетамол—ацетилсалициловая кислота.

[1] **W. Suy, T. Chaoy, S. Yangy, Ch. Lin.** Seeing through a Black Box: Toward High-Quality Terahertz Tomographic Imaging via Multi-Scale Spatio-Spectral Image Fusion, arxiv.org/abs/2103.16932 (2021)
[2] **D. Molter, D. Hübsch, T. Sprenger, K. Hens, K. Nalpantidis, F. Platte, G. Torosyan, R. Beigang, J. Jonuscheit, G. von Freymann, F. Ellrich.** Appl. Sci., **11**, N 3 (2021) 950
[3] **L. Yang, T. Guo, X. Zhang, S. Cao, X. Ding.** Rev. Anal. Chem., **37**, N 3 (2018) 20170021
[4] **Y. Peng, C. Shi, Y. Zhu, M. Gu, S. Zhuang.** PhotoniX, **1**, N 12 (2020) 1—18
[5] **J. Deng, J. Ornik, K. Zhao, E. Ding, M. Koch, E. Castro-Camus.** Opt. Express, **28**, N 21 (2020) 30943—30951
[6] **Д. А. Борисевич, А. М. Гончаренко, А. В. Ляхнович, Г. В. Синицын, М. А. Ходасевич.** Журн. прикл. спектр., **88**, № 1 (2021) 144—149 [**D. A. Borisevich, A. M. Goncharenko, A. V. Lyakhnovich, G. V. Sinitsyn, M. A. Khodasevich.** J. Appl. Specrt., **88**, N 1 (2021) 132—136]
[7] **P. F. Taday.** Phil. Trans. R. Soc. Lond. A, **362** (2004) 351—364
[8] **Г. В. Аюпова, Г. М. Латыпова, О. И. Уразлина, А. А. Федотова.** Упаковка лекарственных средств: уч. пособие по фармацевтической технологии, Уфа, изд-во Башгосмедуниверситета (2009) 91
[9] **R. Bro, Age K. Smilde.** Anal. Method., **6** (2014) 2812—2831
[10] Medical Applications of Mass Spectrometry, Eds. K. Vékey, A. Telekes, A. Vertes, Elsevier (2008) 141—169
[11] **W. Loh.** Int. Stat. Rev., **82**, N 3 (2014) 329—348
[12] **L. A. Berrueta, R. M. Alonso-Salces, K. Héberger.** J. Chromatogr. A, **1158** (2007) 196—214
[13] **Y. Xu, S. Zomer, R. G. Brereton.** Crit. Rev. Anal. Chem., **36** (2006) 177—188