БЫСТРЫЕ МЕТОДЫ СНИЖЕНИЯ РАЗМЕРНОСТИ СПЕКТРАЛЬНЫХ ДАННЫХ ДЛЯ ИХ ОБРАЗНОЙ ВИЗУАЛИЗАЦИИ

В. А. Вагин [1*], А. Е. Краснов [2], Д. Н. Никольский [2]

УДК 535.34;536.631

[1] Научно-технологический центр уникального приборостроения Российской АН, 117342, Москва, ул. Бутлерова, 15, Россия; e-mail: vaguine@mail.ru
[2] Центр реализации государственной образовательной политики и информационных технологий, 125212, Москва, Россия; e-mail: krasnovmgutu@yandex.ru, nikolskydn@mail.ru

(Поступила 6 сентября 2018)

Разработаны быстрые методы снижения размерности спектральных данных, таких как данные ИК спектроскопии, хроматографии и т. д. В отличие от широко известных методов проецирования данных с размерностью N отсчетов в пространства меньшей размерности, имеющих вычислительную сложность порядка N×N, пропорциональную размерности ковариационных матриц данных, для снижения трудоемкости предлагается использовать новые методы, реализуемые в скользящем окне в n отсчетов. В результате быстрые методы имеют вычислительную сложность порядка n×N. Приводятся результаты компьютерных экспериментов по уменьшению размерности ИК спектров автомобильных бензинов. Задача понижения размерности ИК спектров актуальна как для их наглядной образной визуализации, так и для уменьшения мультиколлинеарности и снижения влияния шума при моделировании поведения или анализа параметров, зависящих от спектральных характеристик.

Ключевые слова: *спектральные данные, инфракрасный спектр, снижение размерности, визуализация.*

Fast methods for reducing the dimensionality of spectral data are developed. In contrast to the widely known methods of projecting data with a N dimension into spaces of a smaller dimension having a computational complexity of N×N order proportional to the dimension of the covariance data matrices, it is proposed to use new methods to reduce the complexity. They are feasible in a sliding window in n of counts. As a result, fast methods have the computational complexity of n×N order. The results of computer experiments of reducing the dimensionality of the IR spectra of automobile gasolines are represented. The problem of reducing the dimension of IR spectra is topical for their visualization as well as for decreasing the multicollinearity and the influence of noise while modeling the behavior or analysis of parameters depending on the spectral characteristics.

Keywords: *spectral data, infrared spectra, dimensional reduction, visualization.*

Введение. В [1, 2] показано, что при снижении размерности спектральные данные можно рассматривать как наборы положительно определенных отсчетов дискретных сигналов независимо от их физической природы. К таким наборам относятся, например, отсчеты спектров (поглощения, отражения, рассеяния [3, 4]), связанных с частотами или длинами волн соответствующих диапазонов; сигналов хроматографов, связанных с временем регистрации [5]; сигналов масс-спектрометров, связан-

---

FAST METHODS OF REDUCING THE DIMENSION OF SPECTRAL DATA FOR THEIR IMAGING VISUALIZATION

V. A. Vagin [1*], A. E. Krasnov [2], D. N. Nicol'skii [2] (*[1] Scientific and Technological Center of Unique Instrument-Making of the Russian Academy of Sciences, 15 Butlerova Str., Moscow, 117342, Russia; e-mail: vaguine@mail.ru; [2] Center of Realization of State Educational Policy and Informational Technologies, Moscow, 125212, Russia; e-mail: krasnovmgutu@yandex.ru, nikolskydn@mail.ru*)

ных с их массой, и т. д. Вне перечисленных областей под определение попадают также множество наборов физически разнородных показаний различных приборов и датчиков [3], параметров полей пакетов данных трафика компьютерных сетей [6]. Спектральные данные несут важную информацию о породивших их объектах и поэтому используются в задачах экспресс-анализа состояний этих объектов в различных областях медицины [7, 8], физики [9], химии [10—12], экологии [13], геологии [14], исследования космического пространства [15], пищевой промышленности [2, 3].

Важнейшая проблема экспресс-анализа состояний объектов по их спектральным данным формально сводится к решению задач кластеризации спектральных данных на стадии обучения (по совокупностям контрольных выборок), классификации данных или отнесению неизвестного спектра к тому или иному кластеру на стадии распознавания. Ввиду большой размерности используемых спектральных данных широко развиты методы ее снижения [2, 16]: метод главных компонент (PCA) путем ортогонального проектирования и его модификации (ICA, Fast ICA) на основе преобразований ковариационных матриц исходных данных; метод разбиения факторов на группы в факторном анализе; метод многомерного шкалирования; приближение матриц матрицами меньшего ранга путем их сингулярного разложения (SVD); метод опорных векторов (SVM); различные модификации дискриминантного анализа (линейный дискриминант Фишера (ЛДФ), линейный дискриминантный анализ (ЛДА) [17]); когнитивная визуализация [18]; нейросетевое проецирование (автоассоциативные сети (АС) [19], карты Кохонена [20], квазинейронное агрегирование [21]); алгоритмы оконных фурье-преобразований [22] и вейвлет-преобразований [23]; методы фазовых портретов динамических систем [24], фазовых портретов Гильберта и Френеля [25]. Подробный анализ этих методов приведен в [2] и показано, что для реализации большинства из них необходимы значительные вычислительные и временные ресурсы, так как сжатию данных с размерностью N отсчетов отвечает вычислительная сложность соответствующих алгоритмов порядка $N \times N$, пропорциональная размерностям ковариационных матриц исходных данных.

Цель настоящей работы — исследование применимости разработанных авторами быстрых методов снижения размерности спектральных данных, сохраняющих их геометрические и топологические особенности при минимальных затратах на обработку информации, к образной визуализации ИК спектров.

Экспериментальная часть. Проведены компьютерные эксперименты по снижению размерности ИК спектров автомобильных бензинов. В каждом эксперименте исследовались три метода снижения размерности спектральных данных, представленных дискретными функциями в N отсчетов на основе нейросетевого ортогонального проектирования с использованием преобразований Гильберта и Грама—Шмидта; хеширования фазовых портретов Гильберта; медианной статистики, вычисляемых в скользящем окне в n отсчетов.

ИК спектры бензинов получены на спектрометре АФ-3, разработанном в НТЦ УП РАН [26, 27]. Обработка информации проводилась с помощью исследовательского программного комплекса, включающего в себя программное обеспечение для сжатия спектральных данных [28, 29] и их распознавания на основе статистического метода Вальда [30], разработанного в Центре реализации государственной образовательной политики и информационных технологий (г. Москва). Для ускорения оценивания вариативности данных использован аналог критерия Линка [31] (размахов между максимальными и минимальными значениями).

ИК спектры, характерные для автомобильных бензинов с разными октановыми числами (АИ-80, АИ-92, АИ-95, АИ-98), получены на АЗС Москвы в ходе испытания подвижного аппаратно-программного комплекса экспресс-анализа качества бензинов. Несмотря на то что спектрометр АФ-3 позволяет работать в широком интервале пространственных частот (волновых чисел) среднего диапазона ИК спектра (450—4000 $см^{-1}$), использовались $N = 882$ отсчета спектральных данных в полосе 450—1299 $см^{-1}$ с шагом 1 $см^{-1}$ (разрешение спектрометра при погрешности измерения частот ± 0.1 $см^{-1}$). Выбор полосы объясняется чисто техническими причинами: так как сбор информации проводился не в лабораторных условиях, а в перевозимом на автомобиле спектрометре, в высокочастотной области спектра проявилось влияние аппаратных помех. Соответственно, результаты обработки в более узком (выбранном нами диапазоне) более наглядны (эффективны). ИК спектры двух марок бензинов приведены на рис. 1.

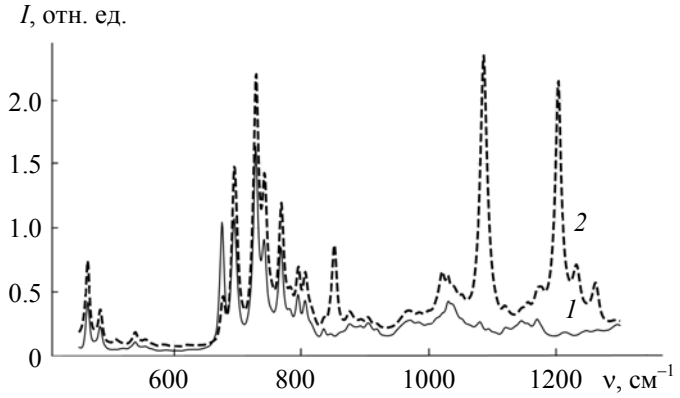

Рис. 1. ИК спектры автомобильных бензинов АИ-80 (*1*) и АИ-98 (*2*)

Результаты и их обсуждение. В вычислительном эксперименте использованы разработанные нами методы снижения размерности спектральных данных, отличающиеся простотой реализации и малой трудоемкостью по сравнению с PCA. Данные методы, опробованные в многочисленных вычислительных экспериментах, оказались эффективными при анализе многомерных информационных потоков положительно определенных данных сетевого трафика.

Первый метод снижения размерности ИК спектров связан с их нейросетевым проецированием на три ортогональные оси: первая образована с помощью выделения компонент спектральных данных, значения размахов которых превышают некоторый наперед заданный порог вариации; вторая — с помощью ортогонального преобразования Гильберта первой оси; третья — с помощью ортогональных преобразований Грамма—Шмидта первой и второй осей [28, 29]. Наперед заданный порог вариации является начальным приближением, необходимым для формирования первой оси (определяется чисто эмпирически, например, на уровне 0.5 максимального размаха спектральных данных). Как показано в [28], такой метод позволяет снижать размерность именно положительно определенных данных, при этом на первой оси (x_0) отображаются наиболее различающиеся (по введенному порогу) данные, а на второй (x_1) — наиболее близкие. В вычислениях использовано $n = 11$ отсчетов, что связано с эмпирическим выбором размера окна для дискретного цифрового фильтра Гильберта [28]. На рис. 2 приведен пример проецирования нескольких спектров каждой марки бензина в ортогональное 3*D*-пространство (x_0, x_1, x_2). В данном пространстве каждому из спектров соответствует одна точка. Выбранные пороги вариации 0.70 (рис. 2, *а*) и 0.19 (рис. 2, *б*) максимизируют критерий в виде отношения межгруппового расстояния кластеров к их внутригрупповому расстоянию (отношение дисперсии всех данных к сумме внутригрупповых дисперсий). Если по рис. 1 трудно судить о соотношениях между спектральными данными, то рис. 2 визуально позволяет делать это, так как при изменении порога вариации заметно изменяется эффект их кластеризации.

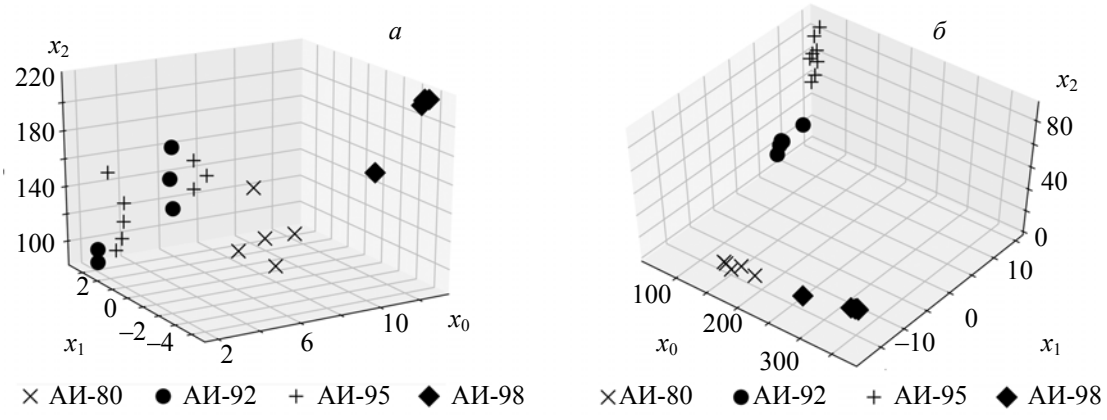

Рис. 2. 3*D*-Представления ИК спектров первым методом снижения их размерности, порог вариации 0.70 (*а*) и 0.19 (*б*)

Второй метод снижения размерности ИК спектров связан с формированием с помощью скользящего окна размером n отсчетов ($n << N$) преобразований Гильберта всех спектральных данных и их трансформаций в одномерные хеш-образы H. Каждому спектру $S(k)$ ставится в соответствие его Гильберт-образ $G(k)$, по которому формируется хеш-образ $H(k) = S(k) + G(k)$, где $k = 1, 2, \ldots, K = 882$ [32]. В вычислениях n = 11 отсчетов, что также связано с эмпирическим выбором размера окна для дискретного цифрового фильтра Гильберта. На рис. 3, *а* приведены 3*D*-образы одномерных нормированных вероятностных распределений $w(H)$ хеш-функций H, построенных для четырех марок бензинов. 3*D*-Образы построены на основе вычисления центральных моментов x_0, x_1, x_2 (средних значений, дисперсий и асимметрий) распределений $w(H)$ хеш-функций, образовавших неортогональные пространства. В 3*D*-представлении каждому из спектров соответствует одна точка. По сравнению с первым методом эффект кластеризации полученных 3*D*-образов исходных данных визуально проявляется хуже.

Третий метод снижения размерности ИК спектров связан с формированием с помощью скользящего окна (размером n = 3 отсчета) в каждом его текущем положении такой статистики, как медиана M [33]. На рис. 3, *б* приведены примеры 3*D*-образов одномерных нормированных вероятностных распределений $w(M)$ медианы, построенных для четырех марок бензинов. 3*D*-Образы, так же как и во втором методе, построены на основе вычисления центральных моментов x_0, x_1, x_2 (средних значений, дисперсий и асимметрий) распределений $w(M)$, сформировавших неортогональные пространства. В 3*D*-представлении каждому из спектров также соответствует одна точка.

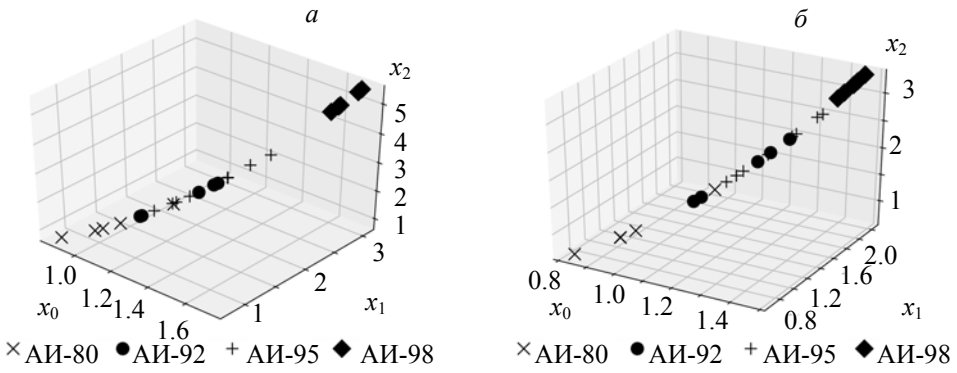

Рис. 3. 3D-Представления ИК спектров вторым (*а*) и третьим (*б*) методами снижения их размерности

Из рис. 2 и 3 видно, что эффект кластеризации первого метода (с управляемым порогом) по сравнению с другими более выражен, что подтверждается отношениями межгруппового расстояния кластеров к их внутригрупповому расстоянию (6.2/9.3, 2.8 и 1.9). В то же время два последних метода более простые. Приведенные на рис. 2 и 3 примеры демонстрируют лишь суть предлагаемых методов. Для применения численных оценок различных параметров исследуемых бензинов по сжатым спектральным данным необходимы их репрезентативные выборки [3].

Отметим, что вычислительная сложность алгоритмов приведенных методов снижения размерности спектральных данных составляет $\sim N \times n$. Это позволяет рассматривать методы как быстрые, что при увеличении размерности и числа спектральных данных может существенно сказаться на времени их анализа. В частности, эти методы могут быть полезны при спектральном анализе динамических процессов (мониторинге окружающей среды [13, 14], контроле химических реакций [10], технологических процессов [33]). Предлагаемые методы можно использовать и для любых однотипных наборов положительно определенных данных, например информационных потоков сетевых данных [34].

Заключение. Разработанные быстрые методы снижения размерности спектральных данных, в частности алгоритмическое и программное обеспечение, способствуют дальнейшему развитию программно-аппаратных комплексов для экспресс-анализа горюче-смазочных материалов по их ИК спектрам. Рассмотренные методы апробированы с использованием разработанного НТЦ УП РАН программно-аппаратного комплекса ПАК-Б, функционирующего на основе портативного переносного ИК-фурье-спектрометра [35]. Встраивание разработанных методов в ПАК-Б повысит как его бы-

стродействие, так и функциональные возможности за счет наглядного визуального представления спектральных данных. В дальнейшем авторы планируют провести дополнительные исследования применимости предложенных методов к кластеризации спектральных данных путем их сравнения с широко используемыми методами, например методом t-SNE [36].

Авторы выражают благодарность сотрудникам НТЦ УП РАН: г.н.с, д.т.н. А. А. Балашову и с.н.с. А. И. Хорохорину, а также с.н.с., нач. отдела контроля качества и физических методов АО ВНИИ НП, д.т.н. Е. И. Алаторцеву за помощь в работе.

[1] **A. E. Krasnov, S. A. Krasnikov, E. A. Chernov.** Materials Int. Sci.-Pract. Conf. “Innovative Information Technologies”, **2**, section 2, 21—25 April 2014, Prague, Moscow, HSE (2014) 664—670
[2] **Е. А. Чернов.** Метод сжатия и визуализации обобщенных спектральных данных объектов пищевой и химической промышленности, дис. ... канд. техн. наук, Москва, МГУТУ им. К. Г. Разумовского (2014) 6—8
[3] **А. Е. Краснов, С. А. Красников, А. В. Воробьева, Ю. Г. Кузнецова, Н. А. Краснова, Д. Ю. Анискин.** Основы спектральной компьютерной квалиметрии жидких сред, Москва, Юриспруденция (2006) 135—161
[4] **А. А. Балашов, В. А. Вагин, И. С. Голяк, А. Н. Морозов, А. И. Хорохорин.** Журн. прикл. спектр., **84**, № 4 (2017) 643—647 [**A. A. Balashov, V. A. Vaguine, Il. S. Golyak, A. N. Morozov, A. I. Khorokhorin.** J. Appl. Spectr., **84** (2017) 664—667]
[5] **А. К. Жерносек, И. Е. Талуть.** Аналитическая химия для будущих провизоров. Часть 1. Учебное пособие, Витебск, ВМГУ (2003) 282—309
[6] **Д. В. Бельков, Е. Н. Едемская.** Наук. праці ДонНТУ, Сер. Інформ., кіберн. обчисл. техн., **16**, № 204 (2012) 21—27
[7] **С. А. Лысенко, М. М. Кугейко.** Журн. прикл. спектр., **80**, № 3 (2013) 432—441 [**S. A. Lisenko, M. M. Kugeiko.** J. Appl. Spectr., **80** (2013) 419—428]
[8] **Г. Б. Толсторожев, М. В. Бельков, И. В. Скорняков, В. А. Бутра, В. И. Пехньо, А. Н. Козачкова, Н. И. Царик, И. П. Куценко, Н. И. Шарыкина.** Журн. прикл. спектр., **81**, № 3 (2014) 444—450 [**G. B. Tolstorozhev, M. V. Bel’kov, I. V. Skornyakov, V. A. Butra, V. I. Pekhnyo, A. N. Kozachkova, N. I. Tsarik, I. P. Kutsenko, N. I. Sharykina.** J. Appl. Spectr., **81** (2014) 463—469]
[9] **В. С. Бураков, А. В. Буцень, В. В. Кирис, Н. В. Тарасенко.** Журн. прикл. спектр., **80**, № 4 (2013) 604—609 [**V. S. Burakov, A. V. Butsen, V. V. Kiris, N. V. Tarasenko.** J. Appl. Spectr., **80** (2013) 589—594]
[10] **А. А. Балашов, В. А. Вагин, А. В. Висковатых, Г. А. Капралова, В. В. Крадецкий, А. И. Хорохорин, А. М. Чайкин.** Журн. прикл. спектр., **80**, № 1 (2013) 149—151 [**A. A. Balashov, V. A. Vagin, A. V. Viskovatykh, G. A. Kapralova, V. V. Kradeckiy, A. I. Khorokhorin, A. M. Chaikin.** J. Appl. Spectr., **80** (2013) 145—147]
[11] **М. Ю. Доломатов, Г. У. Ярмухаметова, М. М. Доломатова.** Журн. прикл. спектр., **84**, № 1 (2017) 132—137 [**M. Yu. Dolomatov, G. U. Yarmuhametova, M. M. Dolomatova.** J. Appl. Spectr., **84** (2017) 114—119]
[12] **М. Ю. Доломатов, Г. У. Ярмухаметова, М. М. Доломатова.** Журн. прикл. спектр., **85**, № 3 (2018) 443—447 [**M. Yu. Dolomatov, G. U. Yarmuhametova, M. M. Dolomatova.** J. Appl. Spectr., **85** (2018) 452—456]
[13] **А. Л. Куракин, В. А. Левченко, Л. И. Лобковский.** Журн. прикл. спектр., **80**, № 6 (2013) 913—919 [**A. L. Kurakin, V. A. Levchenko, L. I. Lobkovsky.** J. Appl. Spectr., **80** (2013) 905—911]
[14] **Г. Г. Райкунов, В. Л. Щербаков, С. И. Турченко, Н. А. Брусничкина.** Гиперспектральное дистанционное зондирование в геологическом картировании, Москва, Физматлит (2014) 26—89
[15] **Б. Е. Мошкин, А. В. Григорьев, В. А. Вагин, А. В. Шакун, Д. В. Пацаев, А. В. Жарков.** ПТЭ, № 6 (2017) 45—51
[16] **А. И. Орлов.** Науч. журн. КубГАУ, № 119(05) (2016) 1—16
[17] **В. Ю. Яньков, Е. А. Чернов.** Технологии XXI века в легкой промышленности (электронное научное издание), № 7, часть II, раздел 4, статья № 9 (2013) 1—6
[18] **В. В. Цаплин, В. Л. Горохов, В. В. Витковский.** Программные продукты и системы, № 3 (107) (2014) 22—25
[19] **С. Хайкин.** Нейронные сети: полный курс — Neural Networks: A Comprehensive Foundation, 2-е изд., Москва, Вильямс (2006) 509—572

[20] **В. Г. Манжула, Д. С. Федяшов.** Фундамент. исслед., № 4 (2011) 108—115
[21] **А. Е. Краснов, Ю. Л. Сагинов, Н. А. Феоктистова.** Приложение к журналу “Качество. Инновации. Образование”, **2**, № 5 (2015) 97—108
[22] **Ю. Е. Воскобойников.** Фильтрация сигналов и изображений: фурье и вейвлет алгоритмы (с примерами в Mathcad), Новосибирск, изд-во НГАСУ (2010) 86—89
[23] **P. S. Addison.** Physiol. Meas., **26** (2005) 155—199
[24] **Д. В. Аносов.** Дифференциальные уравнения: то решаем, то рисуем, Москва, изд-во МЦНМО (2008) 40—70
[25] **А. Е. Краснов, И. Н. Компанец.** Радиотехника, № 1 (2000) 55—60
[26] **А. А. Балашов, В. А. Вагин, С. А. Подлепа, М. А. Шилов.** Физические основы приборостроения, № 1 (2011) 122—131
[27] **В. А. Вагин.** ИК Фурье-спектрометры для научных исследований и применений, дис. ... д-ра техн. наук, Москва, НТЦ УП РАН (2009) 65—82
[28] **А. Е. Краснов, Д. Н. Никольский, А. А. Калачев.** Способ нейроподобного снижения размерности оптических спектров, патент РФ № 2 635 331, Бюл. № 31 (2017)
[29] **А. Е. Краснов, Д. Н. Никольский, А. А. Калачев.** Снижение размерности спектральных данных нейроподобным алгоритмом, св-во о гос. рег. программы для ЭВМ, РФ, № 2017612195 (2017)
[30] **K. G. Jayanta, D. Mohan, S. Tapas.** An Introduction to Bayesian Analysis. Theory and Methods, New York, USA, Springer Science+Business Media (2006) 23—26
[31] **В. П. Боровиков.** STATISTICA. Искусство анализа данных на компьютере, 2-е изд., Москва—Санкт-Петербург, ПИТЕР (2003) 409—504
[32] **A. E. Krasnov, E. N. Nadezhdin, V. S. Galayev, E. A. Zykova, D. N. Nikol’skii, D. S. Repin.** Int. J. Appl. Eng. Res., **13**, N 8 (2018) 5647—5654
[33] **А. Е. Краснов, В. М. Смирнов.** Способ управления созданием нанокристаллических структур на основе распознавания их оптических спектров, патент РФ № 2657101, бюл. № 16 (2018)
[34] **V. S. Galayev, A. E. Krasnov, D. N. Nikol’skii, D. S. Repin.** Int. J. Appl. Eng. Res., **12**, N 21 (2017) 10781—10790
[35] **А. А. Балашов, В. А. Вагин, Б. Е. Мошкин, В. И. Котлов, О. В. Хитров, А. И. Хорохорин.** ПТЭ, № 1 (2008) 179
[36] **Laurens van der Maaten.** J. Machine Learning Res., N 9 (2008) 2579—2605