МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФИЗИЧЕСКИЙ ФАКУЛЬТЕТ

Кафедра лазерной физики и спектроскопии

КОЛОДОЧКА Полина Сергеевна

КЛАССИФИКАЦИЯ ТИПОВ САХАРОВ С ПОМОЩЬЮ МНОГОПАРАМЕТРИЧЕСКОГО АНАЛИЗА УФ, ВИДИМЫХ И БЛИЖНИХ ИК СПЕКТРОВ ВОДНЫХ РАСТВОРОВ С ВЫБОРОМ СПЕКТРАЛЬНЫХ ПЕРЕМЕННЫХ

Реферат дипломной работы

Научные руководители:
доцент, доктор физико-математических наук,
главный научный сотрудник,
ГНУ «Институт физики имени Б.И. Степанова НАН Беларуси»
М.А. Ходасевич;
доцент, кандидат физико-математических наук, доцент,
кафедра лазерной физики и спектроскопии БГУ
Д.В. Горбач.

Минск, 2024

СОДЕРЖАНИЕ

ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ

CART	Classification and Regression Tree, дерево классификации и регрессии
HCA	Hierarchical Cluster Analysis, иерархический кластерный анализ
kNN	k Nearest Neighbors, метод k ближайших соседей
LDA	Linear Discriminant Analysis, линейный дискриминантный анализ
NIR	Near infrared, ближний инфракрасный диапазон спектра
PC	Principal Component, главная компонента
PCA	Principal Component Analysis, метод главных компонент
SVD	Singular Value Decomposition, разложение по сингулярным величинам
UV	Ultraviolet, ультрафиолетовый диапазон спектра
VIS	Visible, видимый диапазон спектра

РЕФЕРАТ ДИПЛОМНОЙ РАБОТЫ

Колодочка П. С.

Классификация типов сахаров с помощью многопараметрического анализа УФ, видимых и ближних ИК спектров водных растворов с выбором спектральных переменных.

Объем дипломной работы составляет 49 страниц, содержится 33 рисунка, 1 таблица, 14 использованных источников.

Ключевые слова: классификация сахара, спектральный анализ, метод главных компонент, кластерный анализ, CART, kNN, HCA, LDA, выбор переменных.

Объектом исследования дипломной работы являются 25 % водные растворы образцов свекловичного и тростникового сахара. Цель данной исследовательской работы заключается в том, чтобы с помощью UV-VIS-NIR спектроскопии разработать модель, основанную на применении метода главных компонент и методов кластерного анализа, для определения растительного источника рафинированных сахаров. Методы анализа спектров: метод главных компонент, методы кластерного анализа (метод построения классификационных деревьев, метод k ближайших соседей, иерархический кластерный анализ, линейный дискриминантный анализ) и метод выбора спектральных переменных.

К спектрам оптической плотности водных растворов сахаров применялись методы многопараметрического анализа. В результате работы были получены модели классификации тростникового и свекловичного сахаров с точностью 94-100 %. Использование метода выбора спектральных переменных позволило достичь 100 % точности классификации образцов свекловичного и тростникового сахара для всех методов классификации. Так же выбор спектральных переменных позволил ограничить количество используемых длин волн до 13.

В данном исследовании была получена достоверная классификация свекловичного и тростникового рафинированных сахаров. Это указывает на возможность замены стандартного дорогостоящего и трудоемкого метода релаксометрии ядерного магнитного резонанса на сравнительно дешевые и простые методы многопараметрического спектрального анализа для классификации растительного источника сахаров.

РЭФЕРАТ ДЫПЛОМНАЙ ПРАЦЫ

Калодачка П. С.

Класіфікацыя тыпаў цукроў з дапамогай шматпараметрычнага аналізу УФ, бачных і блізкіх ІЧ спектраў водных растворaў з выбарам спектральных зменных.

Аб'ём дыпломнай працы складае 49 старонак, змяшчаецца 33 малюнка, 1 табліца, 14 выкарыстаных крыніц.

Ключавыя словы: класіфікацыя цукру, спектральны аналіз, метад галоўных кампанент, кластарны аналіз, CART, kNN, HCA, LDA, выбар пераменных.

Аб'ектам даследавання дыпломнай працы з'яўляюцца 25 % водныя растворы ўзораў бураковага і трысняговага цукру. Мэта дадзенай даследчай работы заключаецца ў тым, каб з дапамогай UV-VIS-NIR спектраскапіі распрацаваць мадэль, заснаваную на ўжыванні метаду галоўных кампанент і метадаў кластэрнага аналізу, для вызначэння раслiннай крыніцы рафінаваных цукроў. Метады аналізу спектраў: метад галоўных кампанент, метады кластэрнага аналізу (метад пабудовы класіфікацыйных дрэў, метад k бліжэйшых суседзяў, іерархічны кластарны аналіз, лінейны дыскрымінантны аналіз) і метад выбару спектральных пераменных.

Да спектраў аптычнай шчыльнасці водных растворaў цукроў ўжываліся метады шматпараметрычнага аналізу. У выніку працы былі атрыманы мадэлі класіфікацыі трысняговага і бураковага цукроў з дакладнасцю 94-100 %. Выкарыстанне метаду выбару спектральных пераменных дазволіла дасягнуць 100 % дакладнасці класіфікацыі узораў бураковага і трысняговага цукру для ўсіх метадаў класіфікацыі. Гэтак жа выбар спектральных пераменных дазволіў абмежаваць колькасць даўжынь хваль величыней 13.

У дадзеным даследаванні была атрымана дакладная класіфікацыя бураковага і трысняговага рафінаваных цукроў. Гэта паказвае на магчымасць замены стандартнага дарагога і працаёмкага метаду рэлаксаметрии ядзернага магнітнага рэзанансу на параўнальна танныя і простыя метады шматпараметрычнага спектральнага аналізу для класіфікацыі раслiннай крыніцы цукроў.

ABSTRACT OF THE DIPLOMA THESIS

Kolodochka P. S.

Classification of sugar types using multiparameter analysis of UV, visible and near-IR spectra of aqueous solutions with a choice of spectral variables.

The volume of the thesis is 49 pages, contains 33 figures, 1 table, 14 sources used.

Keywords: sugar classification, spectral analysis, principal component analysis, cluster analysis, CART, kNN, HCA, LDA, selection of variables.

25% aqueous solutions of beet and cane sugar samples are the objects of research for the thesis. The purpose of this research work is to develop a model using UV-VIS-NIR spectroscopy to determine the plant source of refined sugars, based on the application of principal component analysis and cluster analysis methods. Spectrum analysis methods: principal component analysis, cluster analysis methods (classification trees, k nearest neighbors, hierarchical cluster analysis, linear discriminant analysis) and spectral variable selection method.

Spectral analysis methods were applied to the optical density spectra of aqueous sugar solutions. Models for the classification of cane and beet sugars were obtained with an accuracy of 94-100% as a result of the work. The method of selecting spectral variables made it possible to achieve 100% accuracy in the classification of beet and cane sugar samples for all classification methods. Also, the choice of spectral variables made it possible to limit it to 13 wavelengths.

A reliable classification of beet and cane refined sugars was obtained in this study. This indicates the possibility of replacing the standard expensive and labor-intensive nuclear magnetic resonance relaxometry method with relatively cheap and simple multivariate spectral analysis methods for classifying plant sources of sugars.

ВВЕДЕНИЕ

В современном мире методы анализа данных становятся все более сложными. Повышаются уровни автоматизации, количество регулярно анализируемых образцов растет, а объем данных на один образец увеличивается и может показаться ошеломляющим. Методы многопараметрического анализа помогают разобраться в этих данных, закономерности и особенности которых могут дать представление об изначальных образцах. Область исследования многопараметрических данных может стать основой для работы аналитиков и исследователей, чья задача – определение происхождения объектов (в том числе товаров и продуктов питания).

Рассматриваемая проблема заключается в классификации образцов. С помощью обучающей выборки можно построить модель для определения неизвестных объектов исследования.

В данной работе для исследования свекловичного и тростникового сахаров используются методы многопараметрического анализа и кластерного анализа. Метод главных компонент - один из основных многопараметрических методов в хемометрике. Он широко используется для исследовательского анализа данных, обнаружения выбросов, уменьшения ранга (размерности), графической кластеризации, классификации и регрессии.

Хемометрика - научная дисциплина, находящаяся на стыке химии и математики, ставящая целью получение данных с помощью математических методов обработки и исследования данных. Хемометрический подход к анализу данных в последнее время всё активнее применяется для решения различных задач [1, 2]. Основанная изначально для использования математических методов в химии, сейчас хемометрика широко используется и в смежных областях науки: физике, биологии, медицине, везде, где есть необходимость в анализе большого количества данных и поиске различного рода закономерностей.

Основное применение метода главных компонент (PCA, Principal Component Analysis) в хемометрике - уменьшение размерности. Если есть многомерные данные, применение метода главных компонент позволяет уменьшить размерность этих данных, чтобы наибольшие изменения в них были зафиксированы в пространстве меньшей размерности. PCA практически всегда является первым методом анализа при работе с многомерным набором данных.

Методы кластерного анализа предназначены для разбиения исходных данных на поддающиеся интерпретации группы. Таким образом, элементы, входящие в одну группу максимально «схожи», а элементы из разных групп

максимально «отличны» друг от друга. Одними из самых распространенных методов являются метод построения классификационных деревьев, метод k ближайших соседей и иерархический кластерный анализ.

В исследовательской работе используются два вида сахара: тростниковый и свекловичный. Визуально продукты не отличаются друг от друга, не имеют запаха и сладкие на вкус.

Очищенные сахара имеют белый цвет, однако, в зависимости от степени очистки, свекольный сахар может быть жёлтым. В желтоватом продукте меньшая доля сахарозы, но больший процент микроэлементов. Нерафинированный продукт несъедобен, обладает неприятным запахом и специфическим привкусом.

Тростниковый неочищенный сахар имеет коричневый оттенок и лёгкий карамельный привкус. К тому же, в нём есть небольшое количество мелассы (сиропообразная патока), в которой есть огромный комплекс полезных микроэлементов, витаминов и волокон растений.

Нерафинированные продукты отличить не составит труда, чего не скажешь об очищенных сахарах. Рафинированные тростниковый и свекловичный сахара содержат в себе более 99,9 % сахарозы, по этой причине отличить их очень сложно.

Стандартным в Европе методом определения вида сахара является релаксометрия ядерного магнитного резонанса, позволяющая определить тип сахарной примеси в меде [3], а также классифицировать и выявить фальсификацию тростникового, свекловичного и кокосового сахаров [4]. Однако ядерно-магнитный резонанс является дорогостоящим и трудоемким методом. Целесообразно найти более простой и дешевый метод определения свекловичного и тростникового сахаров или их фальсификации. Исследование по выявлению фальсификации черного чая [5] является примером успешной работы по выявлению подделок с помощью оптического спектрального анализа.

Цель данной исследовательской работы заключается в том, чтобы с помощью UV-VIS-NIR спектроскопии разработать модель, основанную на применении метода главных компонент и методов кластерного анализа, для определения растительного источника рафинированных сахаров.

Задачи:

- измерить спектры оптической плотности водных растворов образцов тростникового и свекловичного рафинированного сахара в UV-VIS-NIR областях
- применить метод главных компонент для анализа информации и уменьшения размерности данных
- построить модели классификации

- применить метод выбора спектральных переменных для улучшения точности моделей

ГЛАВА 1
МЕТОДЫ АНАЛИЗА И КЛАССИФИКАЦИИ МНОГОМЕРНЫХ ДАННЫХ

1.1 Многомерный анализ данных

Методы исследовательского анализа данных и кластерного анализа позволяют идентифицировать различные закономерности объектов, классифицировать их. Доступно множество численных методов, но все они имеют одну и ту же отправную точку. Данные выражаются в матричной форме, в которой каждый образец или объект описывается вектором измерений. Формируется пространство с таким количеством измерений, как количество переменных в векторе. Объект занимает одну точку в этом многомерном пространстве. Предполагается, что схожие точки (т. е. похожие выборки), будут иметь тенденцию группироваться в пространстве (Рисунок 1) [6].

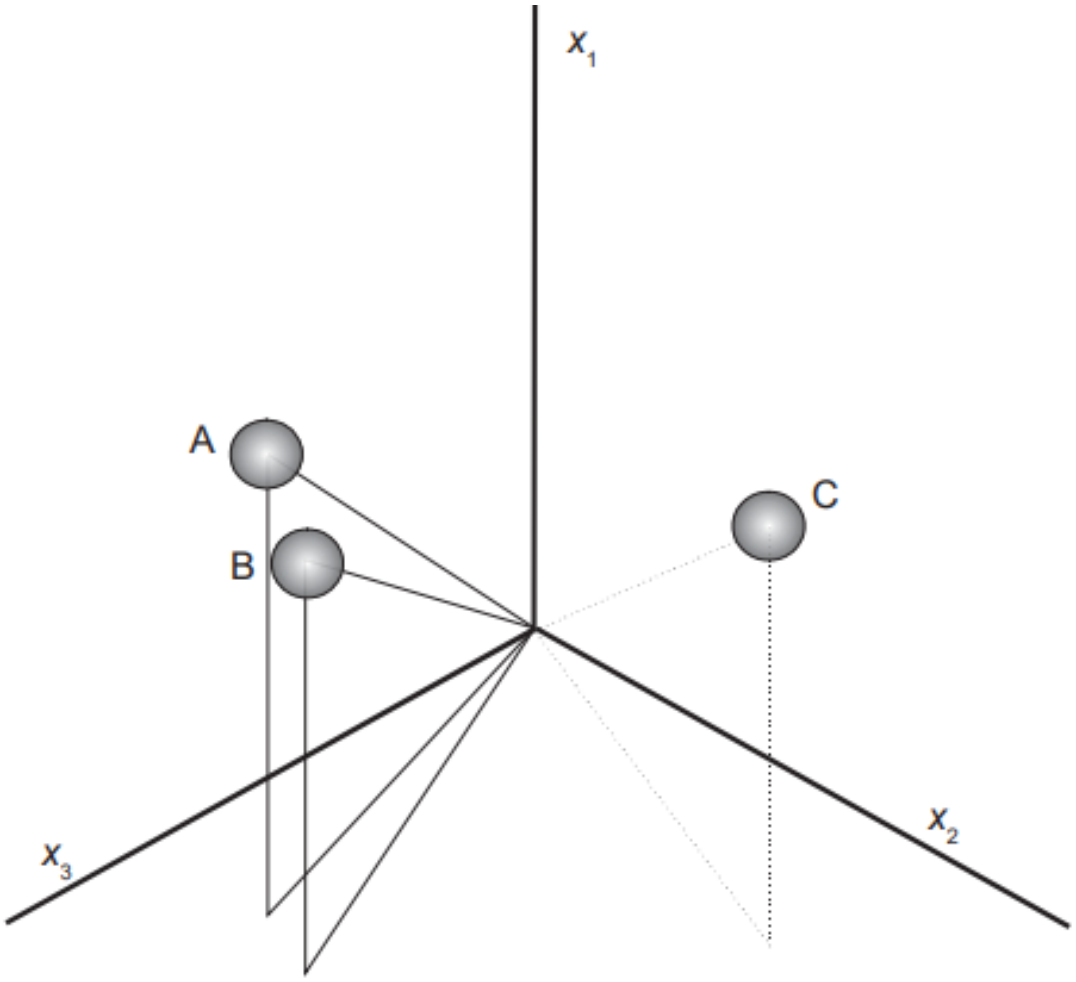

Рисунок 1 – Схематическое изображение точек многопараметрического пространства, представляющих исследуемые образцы

Три объекта A, B и C отображаются в трехмерном пространстве шаблонов, определенном переменными x1, x2 и x3. Объекты A и B расположены ближе друг к другу, являются более похожими, чем объект C.

Распространенным подходом анализа объектов будет графическая визуализация облака данных, ограниченного тремя переменными. Часто можно добавить четвертое измерение, но это предел для графических изображений. Для многомерных наборов данных альтернативным подходом является уменьшение размерности данных с минимальной потерей важных атрибутов, например дисперсии данных [7]. Методы многопараметрического анализа данных позволяют достичь этих целей.

1.2 Метод главных компонент

В исследовательской работе был использован метод главных компонент [6, 7, 8]. Он актуален для работы с большими массивами данных, когда основной задачей является анализ информации, выявление выбросов, уменьшение размерности и нахождение скрытых в спектрах зависимостей. Это универсальный метод, способный работать со сложными многопараметрическими данными. Вместо исходного множества переменных набор данных может быть выражен с помощью нескольких основных компонент. Вычисление главных компонент сводится к вычислению собственных векторов и собственных значений исходных данных [7].

- **Основные понятия**

Обрабатываемые данные должны представлять собой матрицу X, состоящую из i строк и j столбцов (рисунок 2). В данном случае образцами называются строки, пронумерованные индексом i=1…, I; а переменными являются столбцы j=1…, J [8].

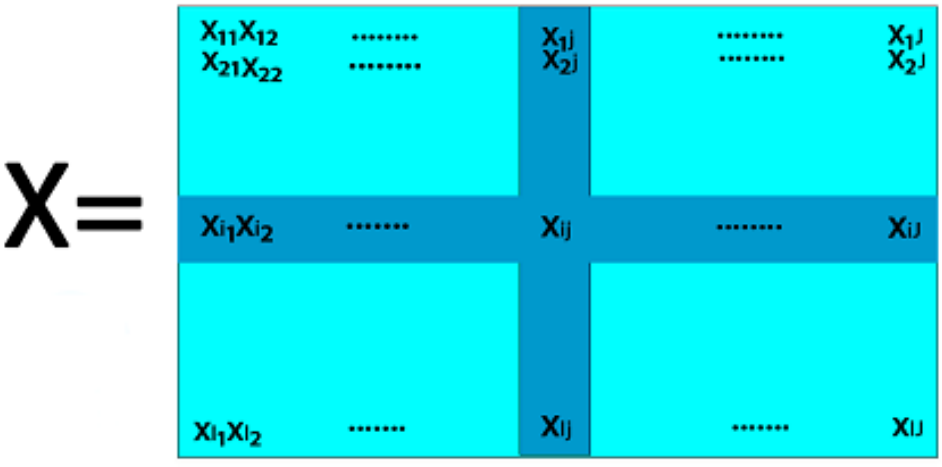

Рисунок 2 – Матрица X

Целью применения метода главных компонент является получение полезной информации из представленных данных. Определение «полезной информации» зависит от поставленной задачи. В исходных данных может содержаться необходимая информация, она даже может быть избыточной. Но существуют случаи, когда полезной информации может не быть вовсе.

Размерность исходных данных (количество образцов и переменных) влияет на результат поиска информации. Некорректно считать, что могут быть лишние данные. На практике это означает, что если получен спектр какого–то образца, то не нужно выбрасывать все точки, кроме нескольких характерных длин волн, а использовать их все, или, по крайней мере, значительную часть спектра.

Частью данных практически всегда является нежелательная составляющая. Данные, в которых нет полезной информации называются шумом. Что будет шумом, а что – искомой информацией, решается с учетом поставленной задачи. Погрешности в данных, вызванные шумом и избыточностью, приводят к появлению случайных связей между переменными.

- **Геометрическая интерпретация**

Можно описать суть метода с помощью геометрической интерпретации. Возьмем случай двух переменных x_1 и x_2.

Каждому образцу исходных данных соответствует точка на координатной плоскости. Центр множества точек, представляющих собой спектры, смещен относительно начала системы координат. В PCA отсутствует нулевая главная компонента, которая могла бы описать смещение, следовательно, предобработка данных должна обязательно включать в себя приравнивание нулю среднего значения множества спектров для каждой спектральной переменной – центрирование. Иногда также необходимо проводить нормирование – масштабирование с помощью деления на соответствующие стандартные отклонения. Нормирование редко используется в спектроскопии, так как все спектральные переменные имеют одинаковый характер и после масштабирования невозможно оценить влияние определенных спектральных интервалов на получаемые результаты.

После предобработки данных можно приступить к определению первой главной компоненты. На рисунке 3 изображено графическое представление двумерных данных. Покажем прямую, вдоль которой будет проходить максимальное изменение данных.

На рисунке 4 прямая подписана «PC1». Это первая главная компонента. Проецируем точки на линию. Полученные точки изображены красным.

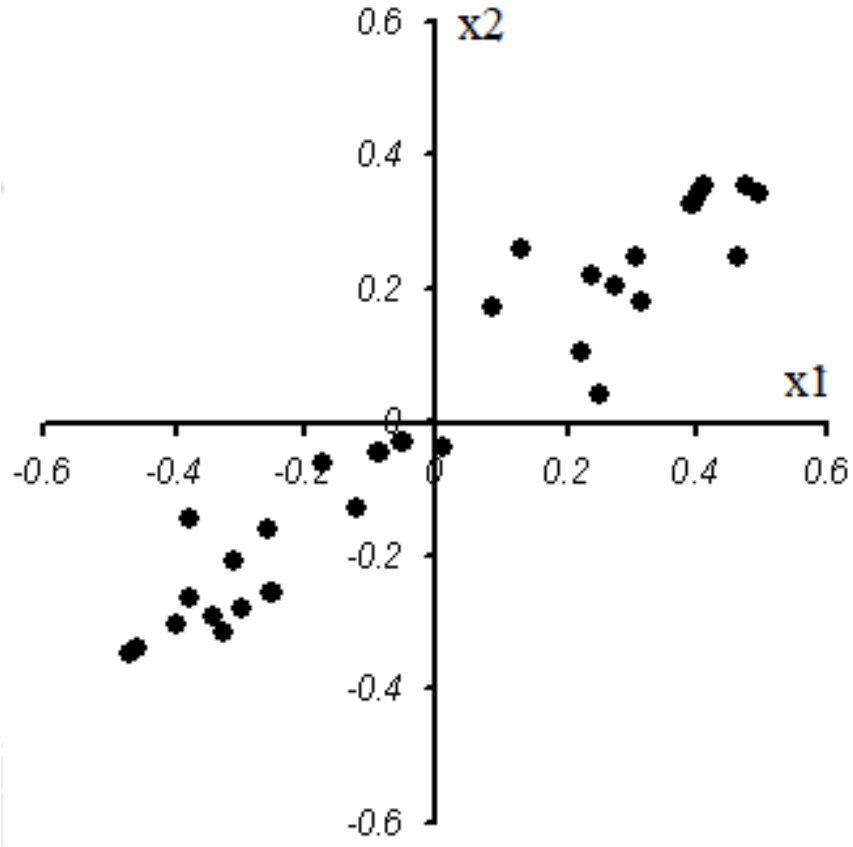

Рисунок 3 – Графическое представление двумерных данных

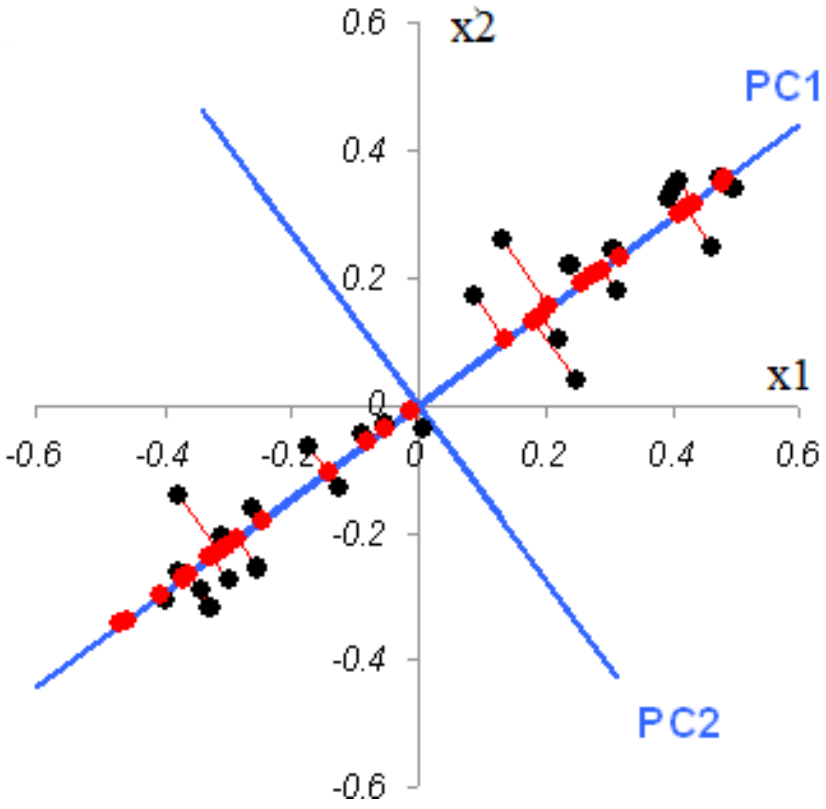

Рисунок 4 – Графическое представление эксперимента

Можно сделать предположение, что все точки должны были лежать на первой прямой. В таком случае отклонения от оси считаются ненужной информацией. Проверить шум ли это можно проделав ту же процедуру для остатков, то есть найти ось максимальных изменений для них [6].

PC1 является линейной комбинацией исходных переменных, фиксирующей максимальную дисперсию в наборе данных. Она определяет направление наибольшей изменчивости данных. Чем больше изменчивость, описанная в первой компоненте, тем больше информация. Последующие компоненты будут иметь меньшую изменчивость, чем первая главная компонента. PC1 определяет направление в многомерном пространстве спектральных переменных, которое минимизирует сумму квадратов расстояний между точками данных и линией.

Вторая главная компонента ортогональна первой, дисперсия данных вдоль неё максимальная из оставшихся возможных (рисунок 4, прямая «PC2»). Таким же способом находятся и последующие компоненты для пространств большей размерности. Они описывают оставшуюся вариацию, не коррелируя со всеми предыдущими компонентами [6].

- **Математическая интерпретация**

Метод главных компонент ищет в многомерном пространстве исходных спектральных переменных их линейные комбинации t_a, называемые главными компонентами:

$$\boldsymbol{t}_{\mathrm{a}} = \boldsymbol{p}_a^t \mathbf{x} = p_{a1}x_1 + \cdots + p_{aJ}x_J = \sum_{j=1}^{J} p_{aj}x_j \qquad (1)$$

Коэффициенты p_{aj} определяются таким образом, чтобы максимизировать дисперсию исходных данных с учетом ограничения:

$$p_{a1}^2 + \cdots + p_{a\mathrm{J}}^2 = 1 \qquad (2)$$

Декомпозиция матрицы X представлена на рисунке 5.

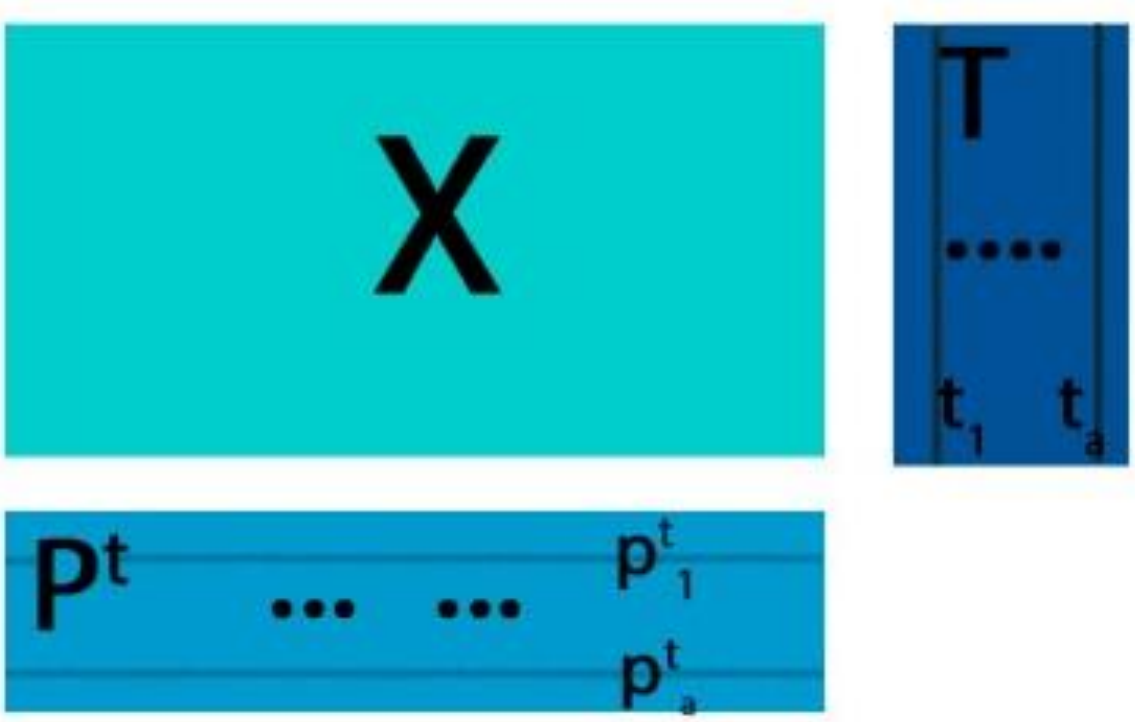

Рисунок 5 – Декомпозиция матрицы X

С помощью данного метода можно представить исходную матрицу данных $\boldsymbol{X}$ в виде произведения двух новых матриц $\boldsymbol{T}$ и $\boldsymbol{P}$ (рисунок 6) с остатками разложения, учитываемыми в матрице $\boldsymbol{E}$:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\mathrm{t}} + \boldsymbol{E} = \sum_{a=1}^{A} t_a p_a^{\mathrm{t}} + \boldsymbol{E} \qquad (3)$$

где A – количество главных компонент, необходимых для исследования, $\boldsymbol{T}$ – матрица счето́в размерностью I на A, $\boldsymbol{P}$ – матрица нагрузок размерностью J на A, $\boldsymbol{E}$ – матрица остатков размерностью как и $\boldsymbol{X}$ I на J. A значительно меньше I и J. Выбором количества PC и задается уменьшение размерности пространства представления многопараметрических данных [6].

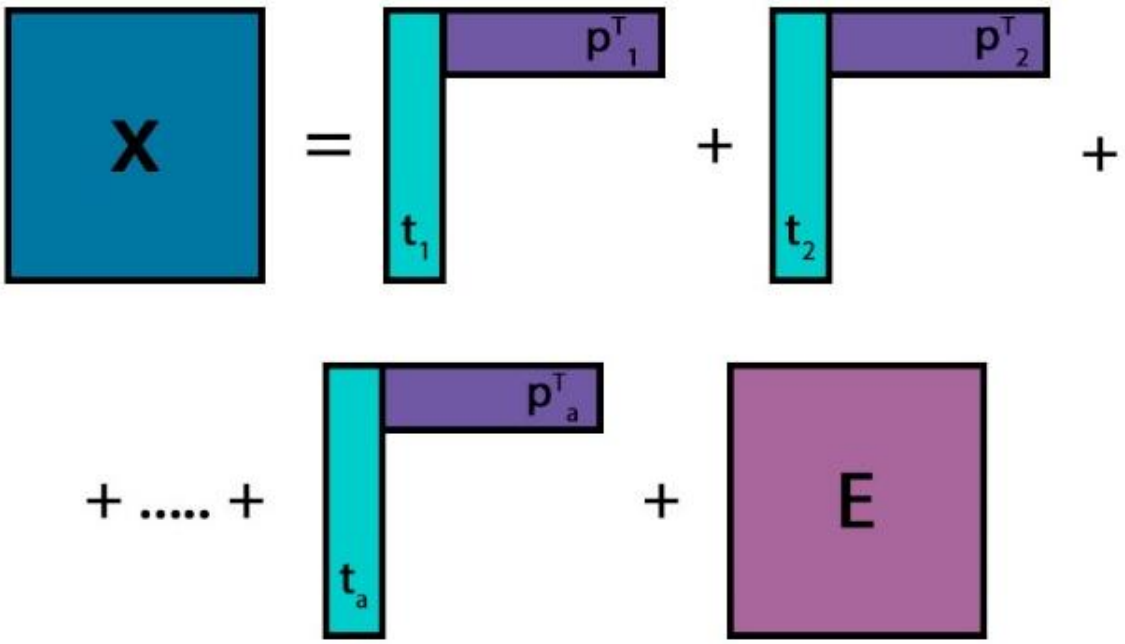

Рисунок 6 – Графическое представление разложения матрицы исходных данных методом главных компонент

Можно использовать любые алгоритмы, которые позволят вычислить собственные значения матрицы $\boldsymbol{X^tX}$. Наиболее популярным является разложением по сингулярным величинам (SVD – singular value decomposition):

$$\boldsymbol{X = USV^t} \tag{4}$$

где $\boldsymbol{U}$ и $\boldsymbol{V}$ – матрицы, образованные ортонормированными собственными векторами матриц $\boldsymbol{XX^t}$ и $\boldsymbol{X^tX}$, соответственно, а $\boldsymbol{S}$ – матрица, на диагонали которой находятся квадратные корни собственных значений матрицы $\boldsymbol{X^tX}$. $\boldsymbol{U}$, $\boldsymbol{S}$ и $\boldsymbol{V}$ легко преобразуются в матрицы счетов и нагрузок: $\boldsymbol{T}$=$\boldsymbol{US}$, $\boldsymbol{P}$=$\boldsymbol{V}$.

- **Счета и нагрузки**

Важную роль в исследованиях с помощью метода главных компонент играют графики счето́в и нагрузок. Матрица счето́в содержит в себе проекции исследуемых $\boldsymbol{J}$-мерных спектров в пространство выбранных главных компонент. При этом каждая строка представляет собой образ соответствующего спектра в новой системе координат PC, а столбец – проекции всех спектров на соответствующую PC. Столбцы в матрице счетов являются ортогональными.

Графики счетов несут в себе информацию, полезную для понимания того, как устроены данные. На графике счетов каждый образец изображается в координатах (t_i, t_j), просторанства PC1-PC2. Близость двух точек означает их схожесть, т.е. положительную корреляцию. Точки, расположенные под прямым углом, являются некоррелированными, а расположенные диаметрально противоположно имеют отрицательную корреляцию (рисунок 7).

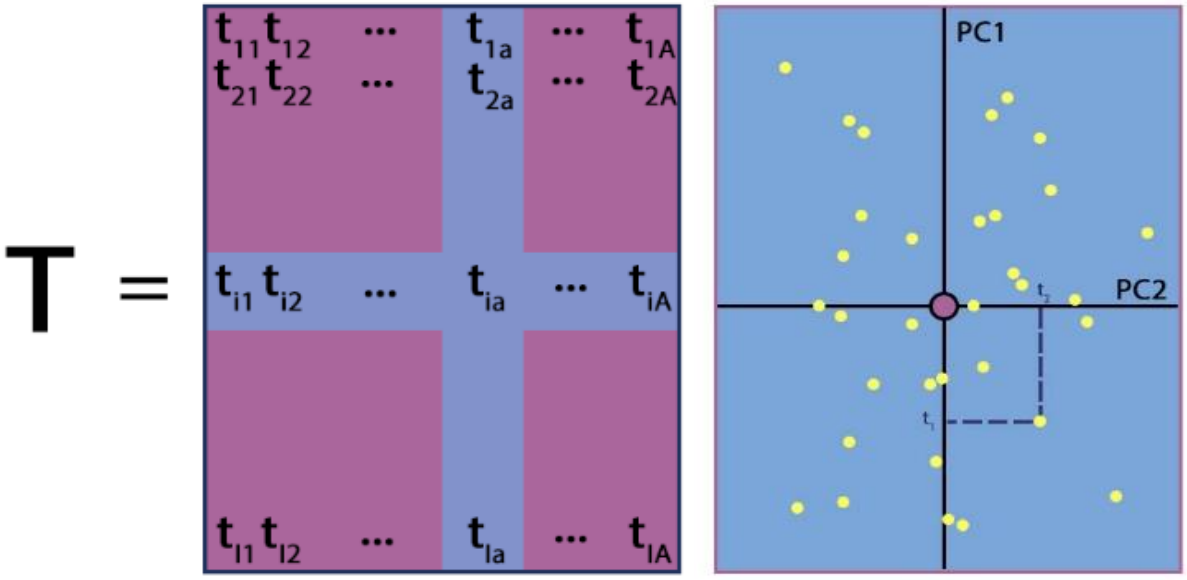

Рисунок 7 – Матрица счетов и график счетов

Матрица нагрузок является матрицей перехода из пространства спектральных переменных в пространство главных компонент. В строке матрицы нагрузок представлены коэффициенты, связывающие спектральные переменные и выбранную PC, а столбец представляет собой проекции спектральной переменной в пространство PC.

График нагрузок применяется для исследования роли переменных. На этом графике каждая переменная x_j отображается точкой в координатах (p_i, p_j) (рисунок 8). Анализируя его аналогично графику счетов, можно понять, какие переменные связаны, а какие независимы. Совместное исследование парных графиков счетов и нагрузок, также может дать много полезной информации о данных.

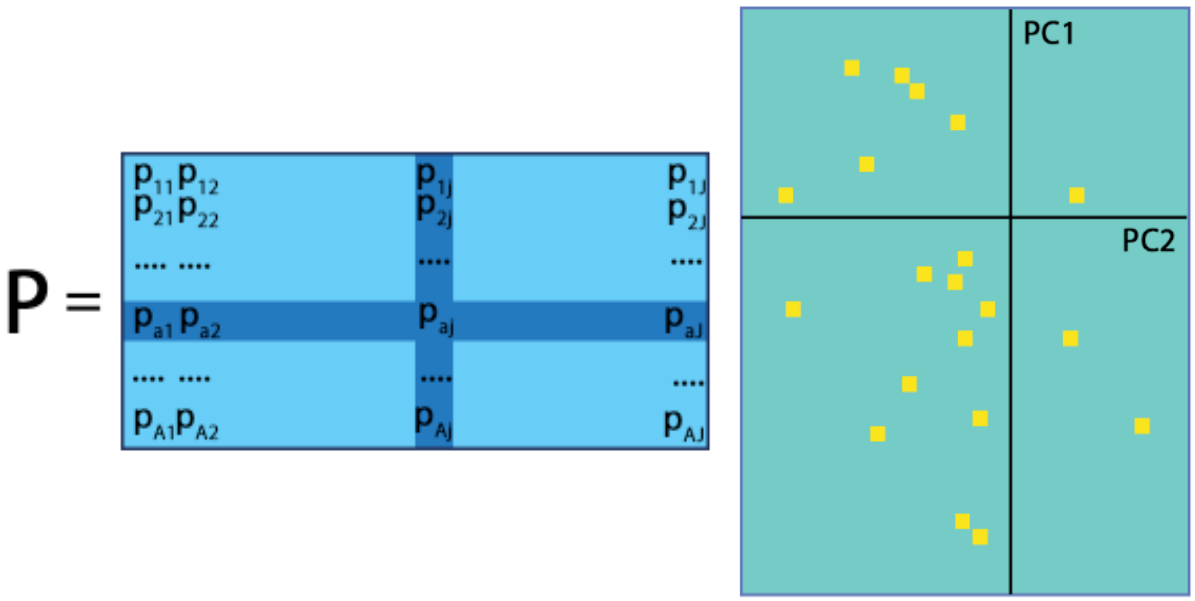

Рисунок 8 – Матрица нагрузок и график нагрузок

1.3 Методы классификации. Кластерный анализ

Целью кластерного анализа является разделение набора данных на группы, которые внутренне однородны и внешне различны. Классификация осуществляется на основе меры сходства или различия внутри групп и между ними. Выбор параметров играет важную роль, и разные варианты выбора могут привести к совершенно разным результатам [7].

- **Метод построения классификационных деревьев**

Дерево классификации и регрессии (CART, Classification and Regression Tree) [9] является частью популярных методов машинного обучения. Оно прогнозирует целевые переменные на основе заданных входных параметров. На рисунке 9 показана схематическая иллюстрация типичной модели дерева решений.

В литературе встречается множество алгоритмов построения дерева решений для задач классификации, однако CART является одним из наиболее часто упоминаемых алгоритмов дерева решений. Подобно другим алгоритмам дерева решений, алгоритм CART извлекает правила принятия решений из признаков и строит модель для прогнозирования целевых значений.

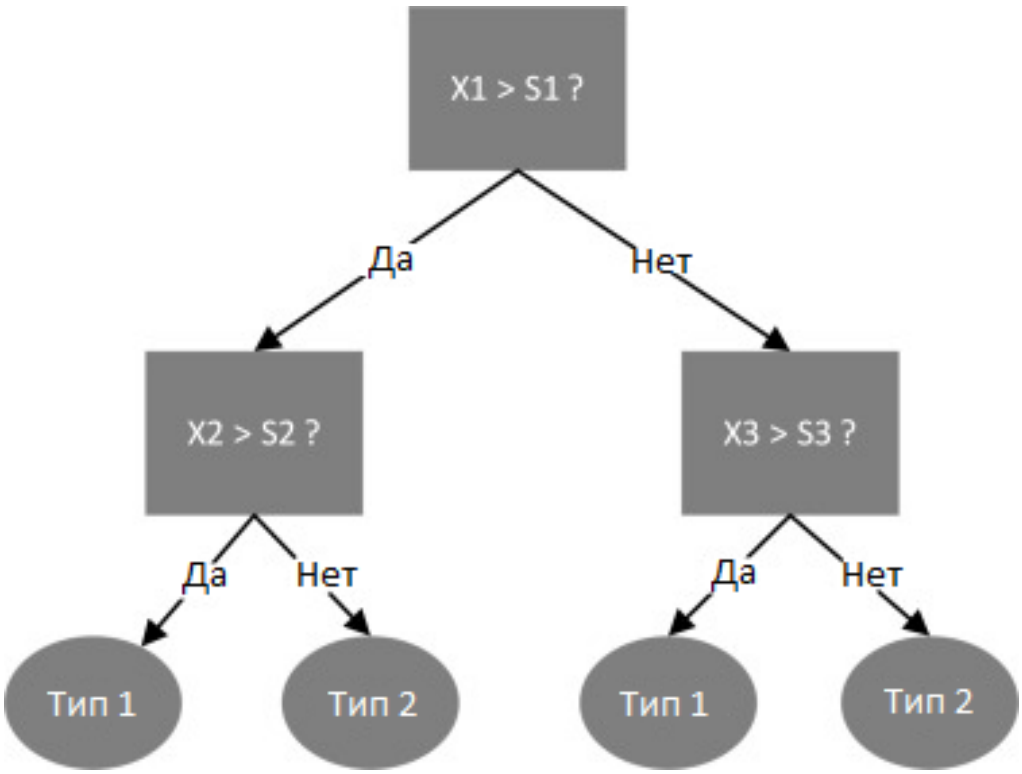

Рисунок 9 – Схематическая иллюстрация типичной модели дерева решений

В данной работе дерево классификации строится в пространстве главных компонент. Эта древовидная композиция представляет собой структурированный набор решений. Решения определяют правила классификации данных. Из этих правил создается иерархия вопросов. Последовательность ответов на вопросы определяет окончательное решение.

В данной работе используется наиболее часто встречающийся алгоритм построения CART. В пространстве главных компонент алгоритм разбивает все образцы на две группы различным образом. Далее выбирается вариант разбиения, при котором максимальное количество образцов одного класса попадают в одну группу. Это первый узел древа. Дальнейшее разбиение продолжается подобным образом до тех пока не будет достигнуто ограничивающее условие. Например, точность классификации или количество узлов принятия решений [9].

Однако, такое дерево классификации будет учитывать только особенности обучающей выборки. Поэтому необходима проверочная выборка или кросс-валидация. При этом обычно происходит уточнение условий или количества узлов принятия решений.

- **Метод k ближайших соседей**

Метод k ближайших соседей (kNN, k Nearest Neighbors) [10] — один из простейших методов классификации, который основан на оценивании сходства объектов, определяемого величиной расстояния между ними в пространстве главных компонент.

Для успешной работы правила kNN полагаются на знание большого количества правильно классифицированных ранее обучающих образцов. Основные принципы заключаются в том, что образцы, расположенные близко друг к другу в пространстве, скорее всего, будут принадлежать к одному и тому же классу [6].

Модель kNN работает, определяя расстояние между новым наблюдением и всеми наблюдениями, присутствующими в обучающих данных. После расчета расстояний алгоритм kNN выбирает k соседей, которые имеют самые короткие расстояния измеренного по некоторой метрике расстояния. Евклидово расстояние легко вычисляется и часто используется. Класс неизвестного объекта определяется по принадлежности к известным классам большинства из k ближайших объектов.

При классификации ближайшего соседа кластер определяется элементами граничного слоя, а объект классифицируется как принадлежащий к той группе, которая содержит его ближайшего соседа. Объект A будет отнесен к группе 1, а не к группе 2 (рисунок 10) [6].

Одним из основных преимуществ kNN является простота и интерпретируемость. Однако выбор подходящего значения для k имеет решающее значение, поскольку выбор слишком малых или слишком больших значений может привести к недостаточному обучению или переобучению соответственно.

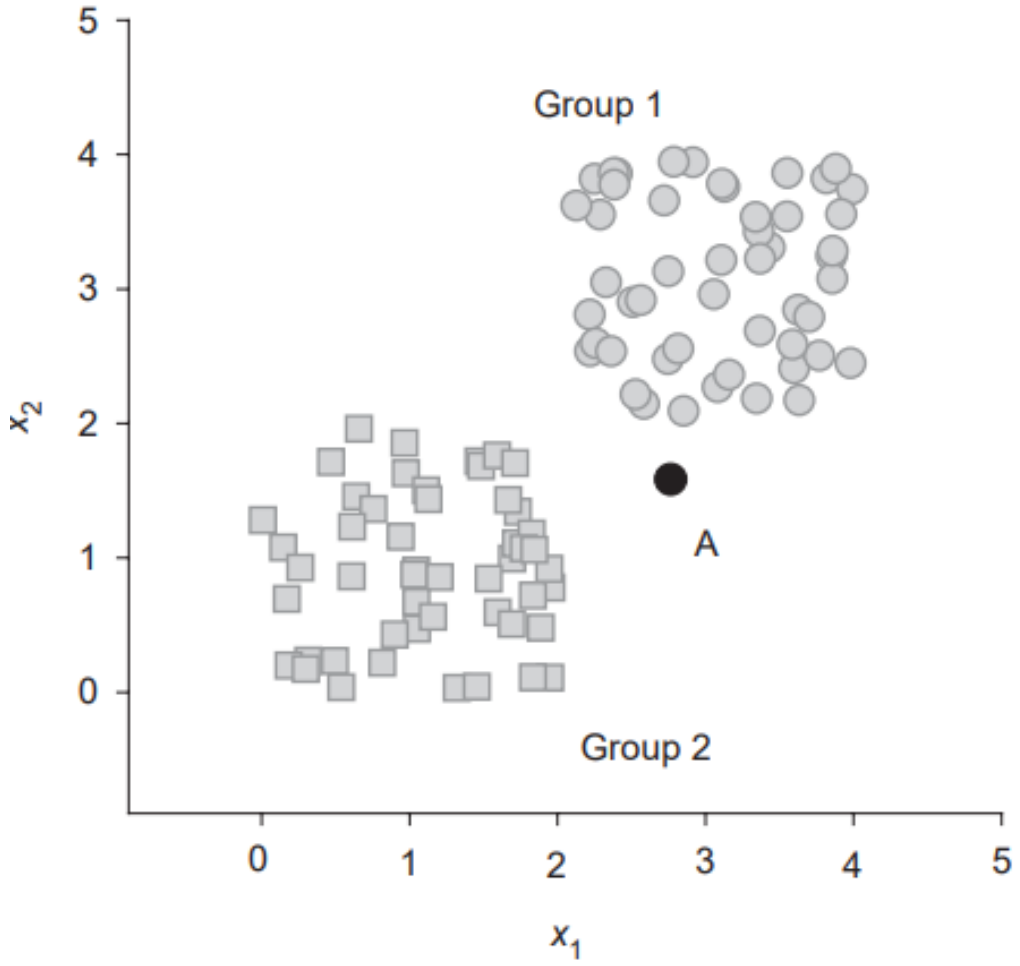

Рисунок 10 – Схематическая иллюстрация классификации методом kNN

- **Иерархический кластерный анализ**

Иерархический кластерный анализ (HCA, Hierarchical Cluster Analysis) [7] распределяет имеющиеся в массиве данных наблюдения на группы, не предлагая при этом алгоритм для предсказания класса, к которому будет отнесено то или иное наблюдение. Встречаются различные методы его реализации.

HCA так же реализует довольно распространённый для сравнения групп принцип: внутригрупповой разброс значений должен быть минимальным, а межгрупповой разброс – существенным. Однако, это условие не всегда соблюдается.

Выделяют два класса методов иерархической кластеризации:

- *Агломеративные методы*: новые кластеры создаются путем объединения более мелких кластеров;
- *Дивизивные* или *дивизионные методы*: новые кластеры создаются путем деления более крупных кластеров на более мелкие.

В данном случае используется агломеративный иерархический кластерный анализ. Используя исходную матрицу данных, сначала строится подходящая матрица мер сходства между объектами (например, евклидова матрица расстояний). Алгоритм вначале рассматривает каждый объект как отдельный кластер, а затем последовательно объединяет кластера, являющиеся наиболее близкими в соответствии с выбранной метрикой пространства. В данной работе объединение происходит до тех пор, пока не останется два кластера (рисунок 11).

Дендрограмма представляет собой графическое представление всего иерархического процесса и показывает, как данные объединяются в кластеры на различных этапах работы алгоритма [7].

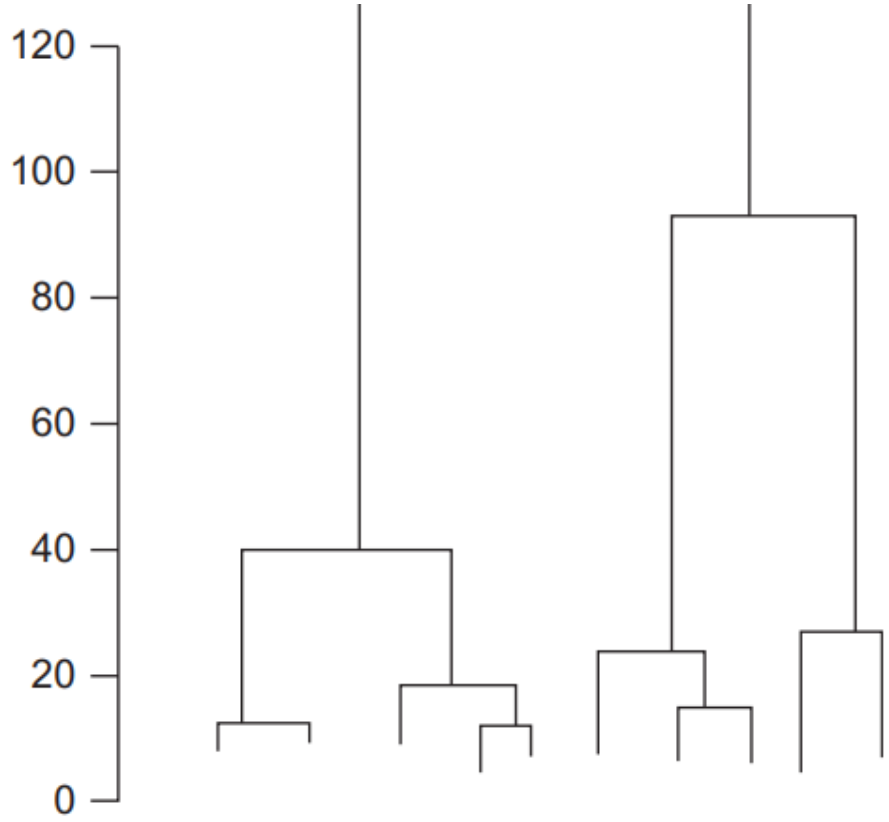

Рисунок 11 – Пример дендрограммы, иллюстрирующей иерархию кластеризации

- **Линейный дискриминантный анализ**

Линейный дискриминантный анализ (LDA, Linear Discriminant Analysis) [11] — это популярный алгоритм машинного обучения, который пытается смоделировать разницу между классами данных.

LDA стремится создать новые переменные, которые представляют собой линейные комбинации исходных переменных и которые лучше всего

указывают на различия между известными группами, в отличие от дисперсий переменных внутри групп [6].

Процесс выполнения LDA направлен на выведение и построение границы между известными классами объектов обучения с использованием статистических параметров. Эта граница получается с помощью дискриминантной функции, которая при применении к тестовому объекту дает значение или оценку, указывающую группу, к которой следует отнести новый объект.

Например, этот метод можно использовать для разделения данных по двум или более рекам на основе параметров качества воды или концентрации тяжелых металлов [12]. Исследователи могут использовать дискриминантный анализ, если группы уже определены в начале.

Математическая модель может быть создана как конечный результат дискриминантного анализа. Эту модель можно использовать для прогнозирования новых элементов группы. Более того, можно использовать модель для определения вклада различных переменных и изучения взаимосвязи между различными выбранными переменными и наблюдениями [13].

Данные NIR в сочетании с моделями LDA представляют собой чувствительную, эффективную, быструю, неразрушающую и точную методологию сбора информации об обработке, применяемой при сборе семян нута, точность которой достигает 94 % [11].

Линейный дискриминант или функция распределения задается уравнением линии, на которой расстояние между облаками данных максимально, а разброс внутри каждого облака минимизирован [7]. Это показано на рисунке 12. Все точки данных теперь можно спроецировать на эту линию.

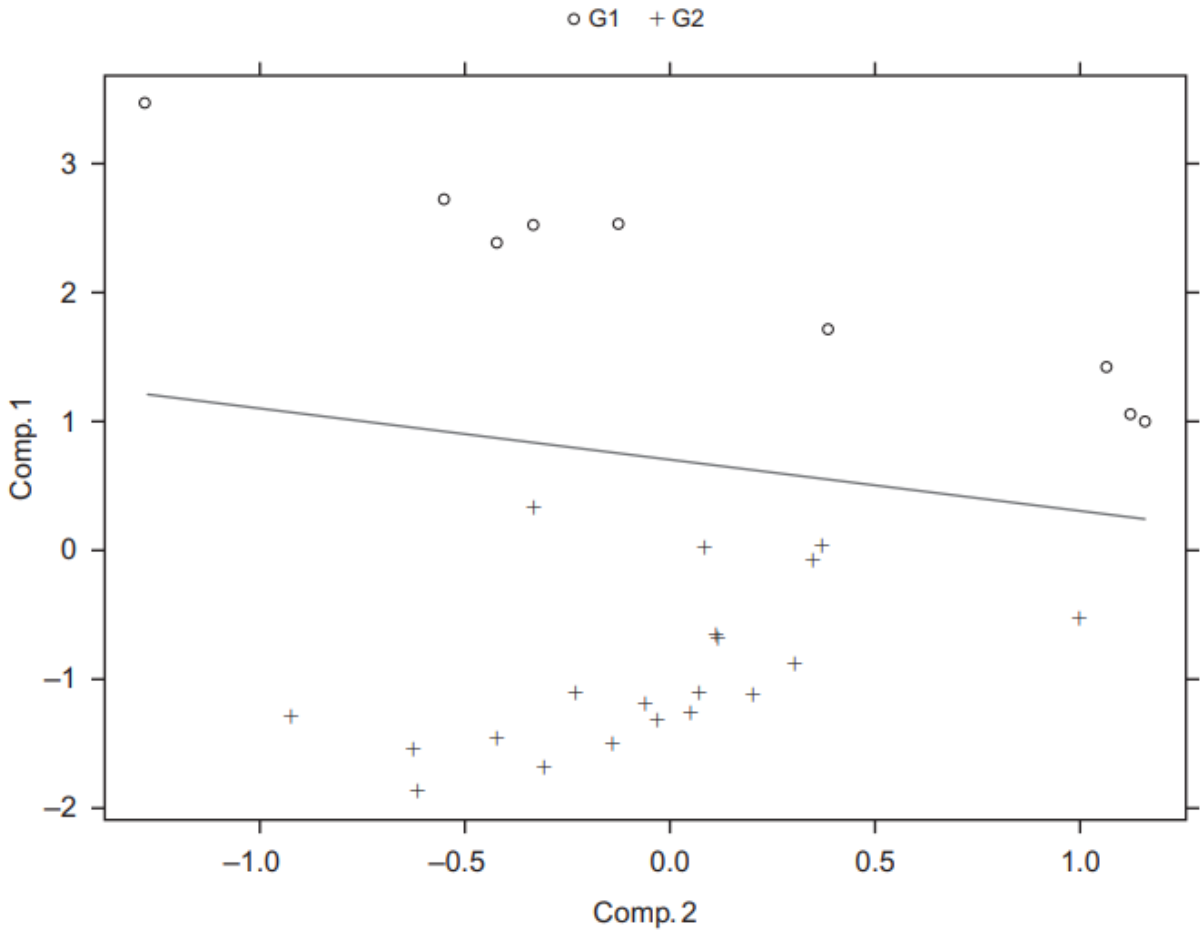

Рисунок 12 – Схематическая иллюстрация границы классификации, определенной в результате линейного дискриминантного анализа

1.4 Метрики расстояния

Наиболее часто используемой метрикой является евклидово расстояние – расстояние между двумя точками евклидова пространства, вычисляемое по теореме Пифагора [14].

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \quad (5)$$

где p обозначает размерность пространства, x_{ik} обозначает значение k-го признака для i-го наблюдения, x_{jk} обозначает значение k-го признака для j-го наблюдения.

Также будет использоваться корреляционное расстояние – это метод измерения корреляции между образцами.

$$d = 1 - \frac{(x_s - \bar{x}_s)(y_t - \bar{y}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\sqrt{(y_t - \bar{y}_t)(y_t - \bar{y}_t)'}} \quad (6)$$

где $\bar{x}_s = \frac{1}{n}\sum_j x_{sj}$ и $\bar{y}_t = \frac{1}{n}\sum_j y_{tj}$.

Диапазон коэффициента корреляции составляет [-1,1]. Чем больше абсолютное значение коэффициента корреляции, тем выше корреляция между X и Y. Когда X и Y линейно коррелируют, коэффициент корреляции равен 1 (положительная линейная корреляция) или -1 (отрицательная линейная корреляция).

ГЛАВА 2
ИЗМЕРЕНИЕ СПЕКТРОВ ОПТИЧЕСКОЙ ПЛОТНОСТИ

2.1 Описание спектрофотометра Shimadzu UV-3101PC

Для проведения спектральных измерений был использован спектрофотометр Shimadzu UV-3101PC (рисунок 13). Он обладает хорошими техническими характеристиками: спектральный диапазон спектрометра 190 – 3200 нм, паспортная погрешность 1%, разрешение 0,1 нм. Shimadzu UV-3101PC содержит два комплекта из трех решеток для охвата широкого диапазона длин волн от ультрафиолетового до ближнего инфракрасного.

Рисунок 13 – Спектрофотометр Shimadzu UV-3101PC

Особенности установки:

• Широкий динамический диапазон от -4,0 до 5,0 Abs

• Разрешение до 0,1 нм

• Двухлучевая схема

Таблица 1. Характеристики спектрофотометра Shimadzu UV-3101PC.

Диапазон длин волн	190 нм ~ 3200 нм
Ширина спектральной полосы (щели)	9 шагов в ультрафиолетовой/ видимой области; 0,1; 0,2; 0,5; 0,8; 1; 2; 3; 5; 7,5 нм; 12 шагов в ближней инфракрасной области; 0,4; 0,6; 0,8; 1,2; 2; 3; 4; 6; 8; 12; 20; 30 нм
Разрешение	0,1 нм
Точность длины волны	± 0,3 нм (при ширине щели 0,2 нм) в ультрафиолетовой / видимой области ± 1,6 нм (при ширине щели 1,2 нм) в ближней инфракрасной области. Автоматически возможна коррекция длины волны.
Повторяемость по длине волны	± 0,1 нм в ультрафиолетовой / видимой области ± 0,4 нм в ближней инфракрасной области
Переключение источников света	Источники света переключаются автоматически в сочетании со сканированием по длине волны. Длина волны, на которой переключаются источники света, выбирается в диапазоне от 282 до 393 нм с шагом 0,1 нм.
Фотометрическая система	Двухлучевая система измерения

2.2 Измерение спектров оптической плотности водных растворов тростникового и свекловичного сахара

Спектры оптической плотности 25 % водных растворов образцов сахара регистрировали на спектрофотометре Shimadzu UV-3101PC в диапазоне длин волн от 200 до 1380 нм с шагом 1 нм и шириной щели 1 нм. Для первой части работы использовались образцы сахара из 7 стран (Беларусь, Пакистан, Польша, Португалия, Россия, Румыния, Сербия). Из 102 исследованных образцов 55 являются растворами свекловичного сахара и 47 – тростникового. Для второй части работы, где применялся PCA с выбором спектральных переменных, были добавлены рафинированные тростниковые сахара из Кубы, Италии и России. Так же некоторые образцы (10 свекловичного сахара, 6 тростникового) были выделены для дополнительной проверки. Таким образом из 94 исследованных образцов 45 являются растворами свекловичного сахара и 49 – тростникового.

На рисунке 14 приведены спектры оптической плотности образцов тростникового и свекловичного сахаров.

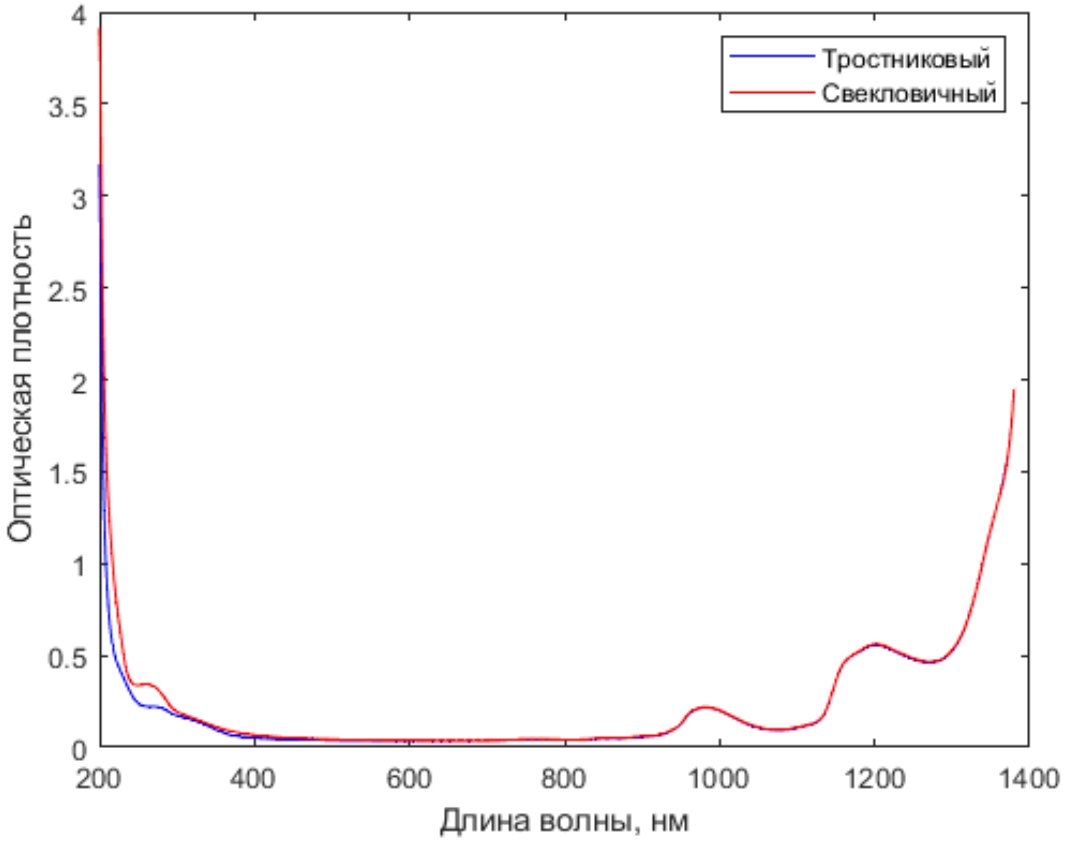

Рисунок 14 – Спектры оптической плотности тростникового и свекловичного сахаров

ГЛАВА 3
КЛАССИФИКАЦИЯ ТИПОВ САХАРОВ

3.1 Применения метода главных компонент к измеренным данным

Преобразование данных и построение графиков было проведено в программе MatLab.

Перед применением PCA спектры предварительно обрабатывали с помощью центрирования. Так как все спектральные переменные имеют одинаковый характер, то в нормировке нет необходимости. После предварительной подготовки можно начать использовать метод главных компонент.

При построении моделей классификации были достигнуты неудовлетворительные результаты при использовании всего (200 – 1380 нм) спектрального диапазона. Для достижения наилучшего результата был выбран спектральный диапазон 350-1000 нм (рисунок 15). В дальнейшем выяснится, что при использовании данного интервала, точность классификации возрастает.

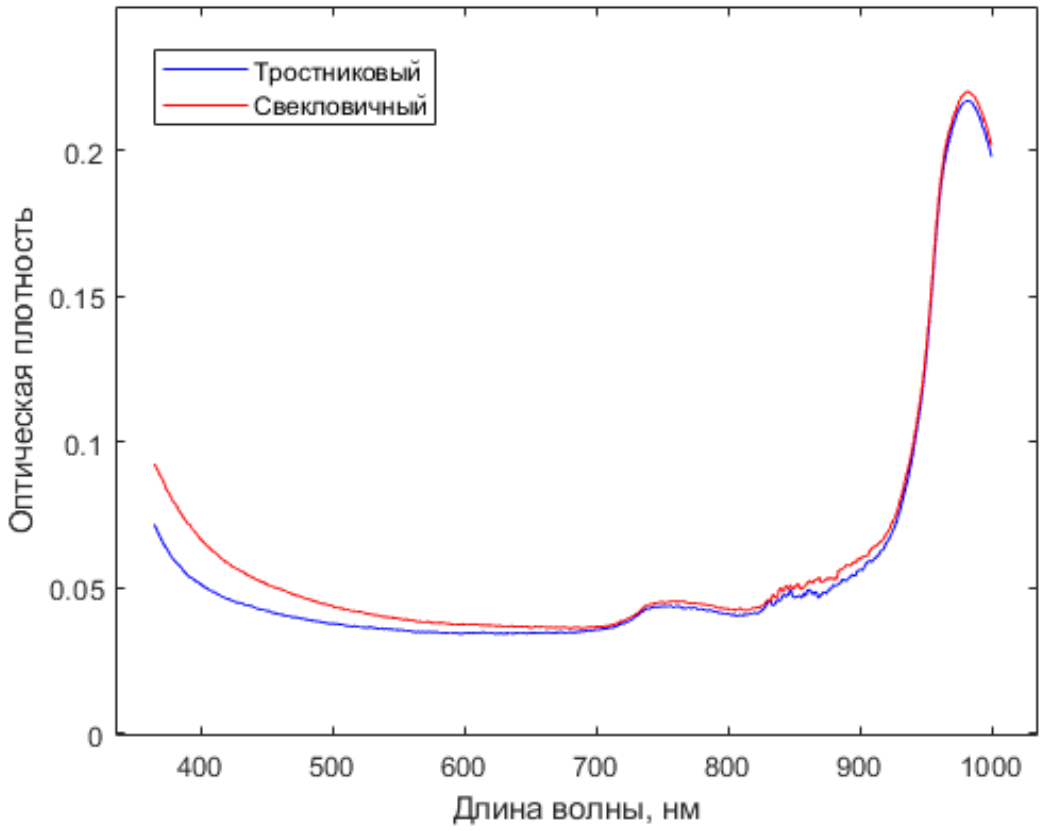

Рисунок 15 – Спектры оптической плотности тростникового и свекловичного сахаров в диапазоне 350-1000 нм

Перейдём к понижению размерности матриц исходных данных. Для дальнейшей работы ограничились использованием 10 главных компонент, которые содержат 99,98 % общей дисперсии данных (рисунок 16).

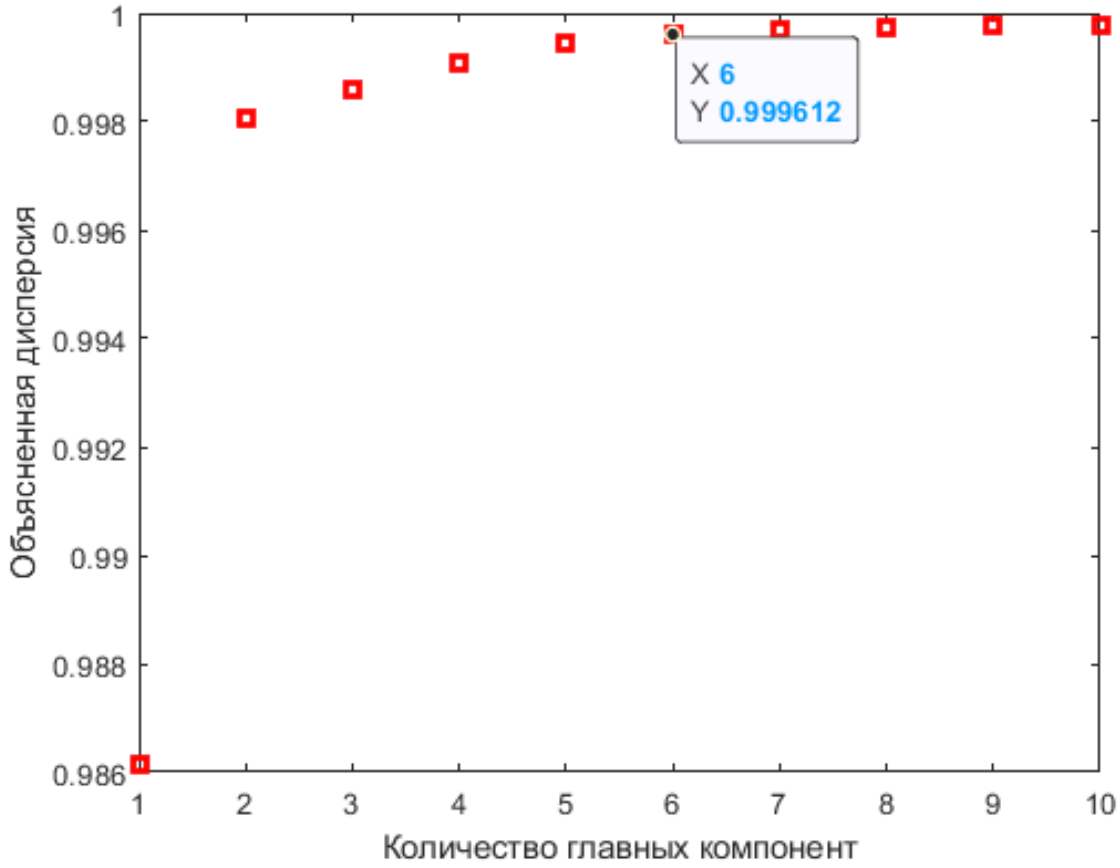

Рисунок 16 – График объясненной дисперсии для 10 главных компонент

Чем выше объясненная дисперсия, тем больше информации будет содержаться в компонентах до тех пор, пока они не начнут учитывать шум. На рисунке 4 показано, что первая главная компонента содержит более 98% информации. Суммарное количество информации в главных компонентах заметно увеличивается до шестой компоненты, но далее такого существенного увеличения не наблюдается.

3.2 Применение методов классификации

- **Метод построения классификационных деревьев**

Классификационное дерево [9] было построено на основе полученных главных компонент с применением 10-кратной кросс-валидации. Случайным образом выбирались девять десятых из набора данных, на их основе строилась модель. С помощью оставшейся части (одна десятая) проверялась полученная

модель. Кросс-валидация позволяет построить 10 моделей, из которых выбирается лучшая по точности.

Необходимо найти количество компонент, при использовании которых построенное дерево будет иметь минимальную ошибку (рисунок 17).

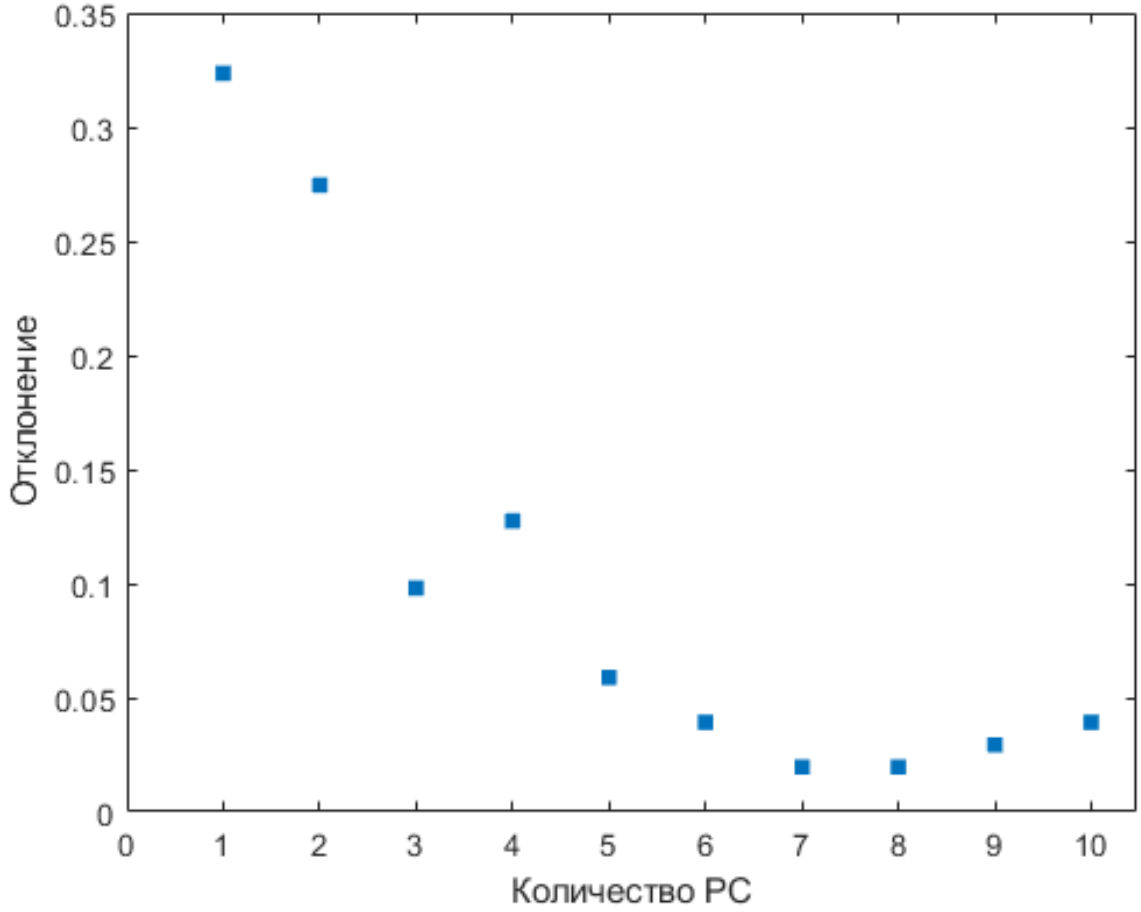

Рисунок 17 – График зависимости отклонения точности классификации модели CART при использовании различного количества главных компонент

В результате было установлено, что при выборе 7 главных компонент точность модели составила более 98 % [15, 16] (рисунок 18).

В данном случае алгоритм создал три узла выбора по трем компонентам. Получено четыре листа в которых содержится подмножество объектов, удовлетворяющих всем правилам ветви, которая заканчивается данными листами.

На рисунке 18 видно, что для построения модели использовались только 1, 3 и 6 главные компоненты. Можно сделать предположение, что они несут в себе наиболее полезную для классификации методом CART информацию.

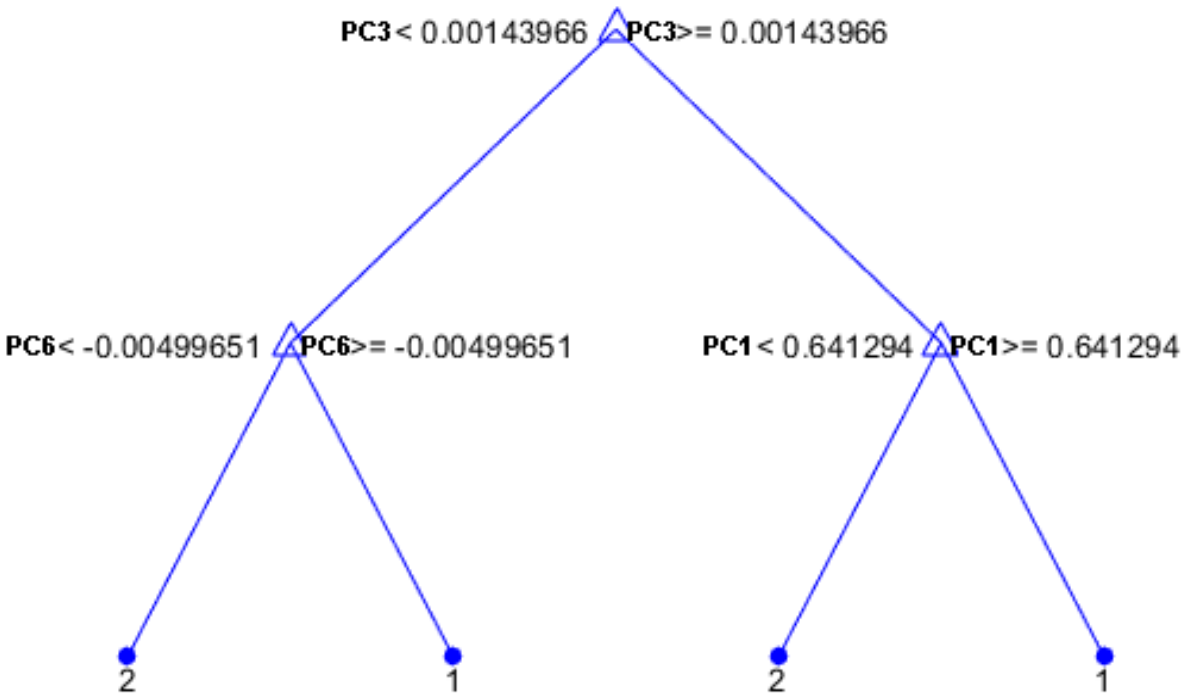

Рисунок 18 – Классификационное дерево для тростникового (1) и свекловичного (2) типов сахара

- **Метод k ближайших соседей**

Для построения модели с помощью kNN необходимо подобрать пространство, в котором будет наилучшая точность классификации. Простое включение всех компонент подряд для построения модели не позволяло достигнуть удовлетворительных результатов. Далее подбор компонент осуществлялся перебором всех возможных комбинаций десяти PC. Были построены модели для каждого варианта и выбрана наиболее точная.

На рисунке 19 изображены счета в пространстве третьей и шестой компоненты, которые так же использовались в методе CART.

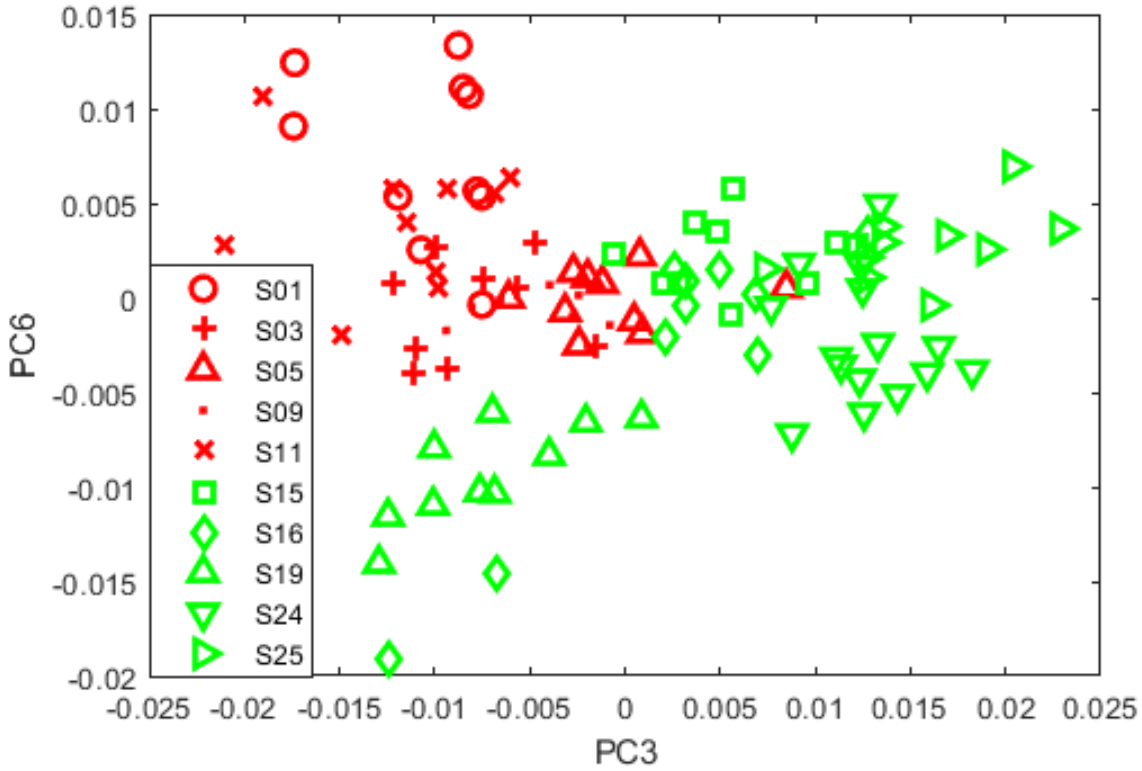

Рисунок 19 – График счетов в 3 и 6 главные компоненты (тростниковый сахар: S01-S11, свекловичный: S15-S25)

В пространстве 3 и 6 PC проведена дальнейшая кластеризация образцов сахара по их спектрам. Ошибка валидации при перепроверке с помощью обучающей выборки была минимальной – менее 3 %. Такой результат достигнут при учете 5 ближайших соседей (рисунок 20).

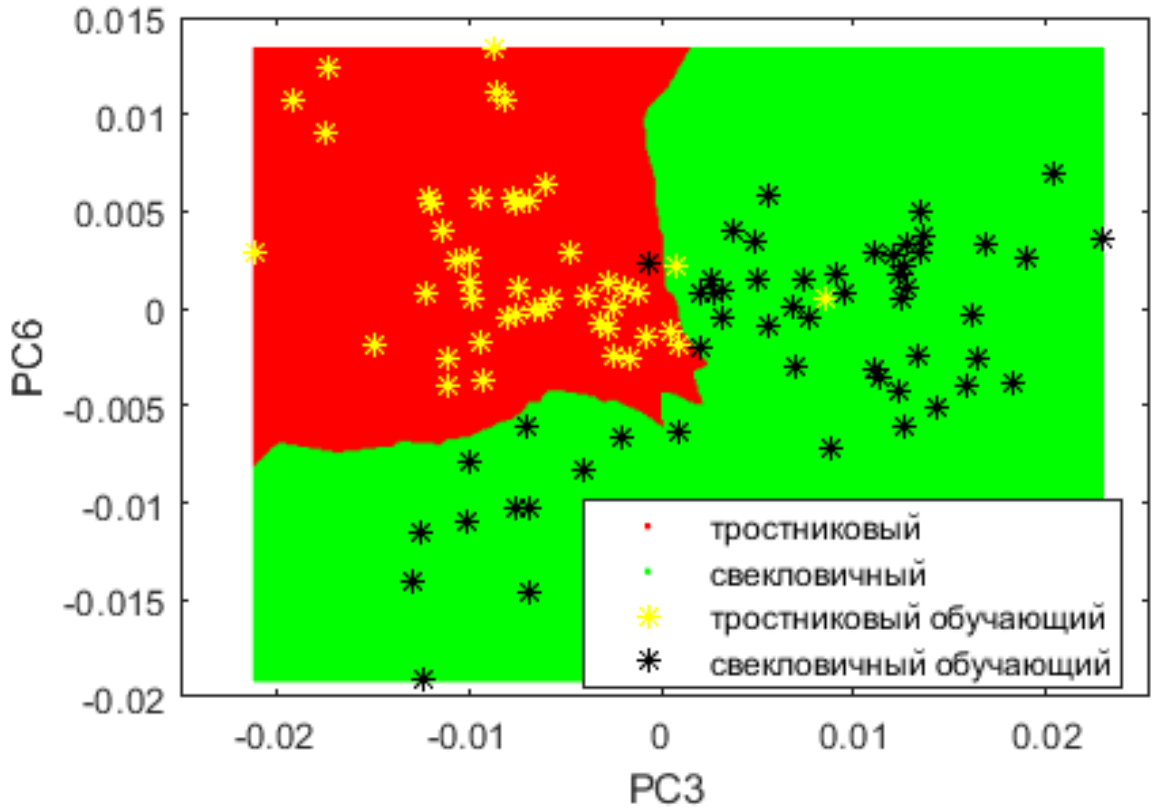

Рисунок 20 – График классификационной модели kNN в пространстве 3 и 6 главных компонент с наложением графика счетов

Далее модель была подвергнута дополнительной 10-кратной перекрестной проверке. Полученная точность классификации составляет около 94 %.

- **Линейный дискриминантный анализ**

Линейный дискриминантный анализ применялся к главным компонентам. На рисунке 21 можно наблюдать график зависимости отклонения модели от количества используемых главных компонент. Компоненты включались в модель в порядке нумерации. Таким образом 100 % точность была достигнута уже при использовании 1 и 2 PC.

Далее в пространстве 1 и 2 PC был построен график полученной классификационной модели с указанием границы между тростниковым и свекловичным сахаром (рисунок 22).

100 % точность модели после 10-кратной кросс-валидации достигнута при использовании 1 и 2 PC.

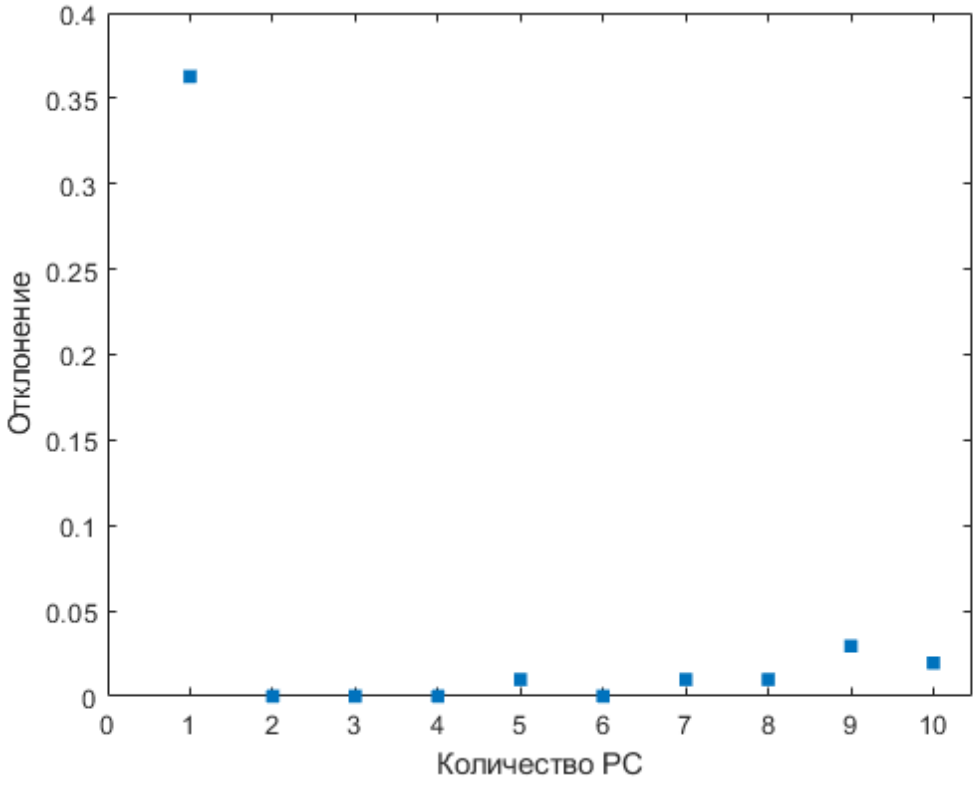

Рисунок 21 – График зависимости отклонения точности классификации модели LDA при использовании различного количества главных компонент

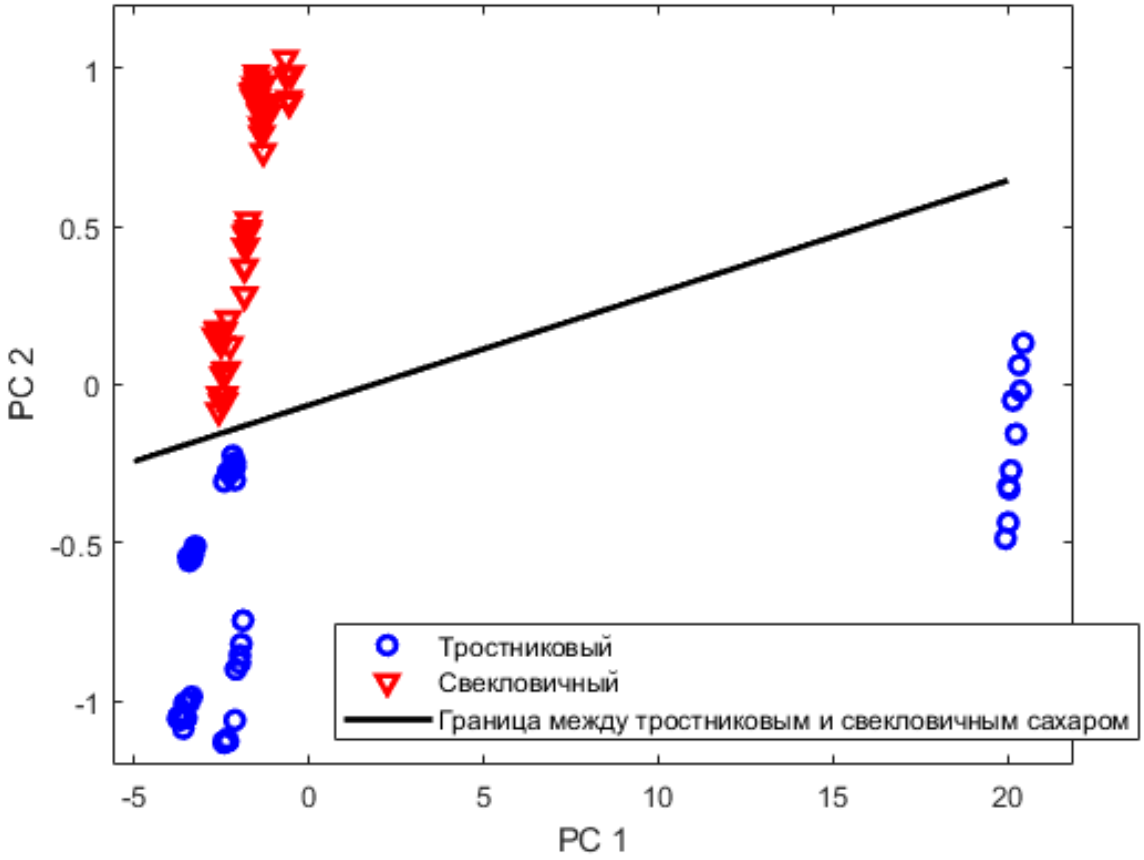

Рисунок 22 – График классификационной модели LDA в пространстве 1 и 2 главных компонент

- **Иерархический кластерный анализ**

Для построения модели с помощью метода HCA было выбрано двумерное пространство третьей и шестой главных компонент со стандартной евклидовой метрикой. Выбор компонент обусловлен тем, что в предыдущих методах (CART, kNN) в данном пространстве были достигнуты наилучшие результаты.

Выбор диапазона длин волн и главных компонент обусловлен предыдущими исследованиями, в которых такой метод обработки данных был оптимальным. В результате была достигнута точность 86 %, которая не является оптимальной.

Для повышения точности были исследованы другие комбинации 10 главных компонент. Построены модели для всех возможных комбинаций десяти PC. Так же были рассмотрены другие метрики, но это не помогло добиться лучших результатов. Из чего был сделан вывод, что необходимо сделать шаг назад и выбрать другой спектральный диапазон.

Для начала выбирался весь доступный диапазон длин волн (200 – 1381 нм) и постепенно уменьшался. Таким образом при выборе диапазона 200 – 320 нм (рисунок 23) точность классификации повысилась до 92 % для 1 и 2 компоненты (рисунок 24).

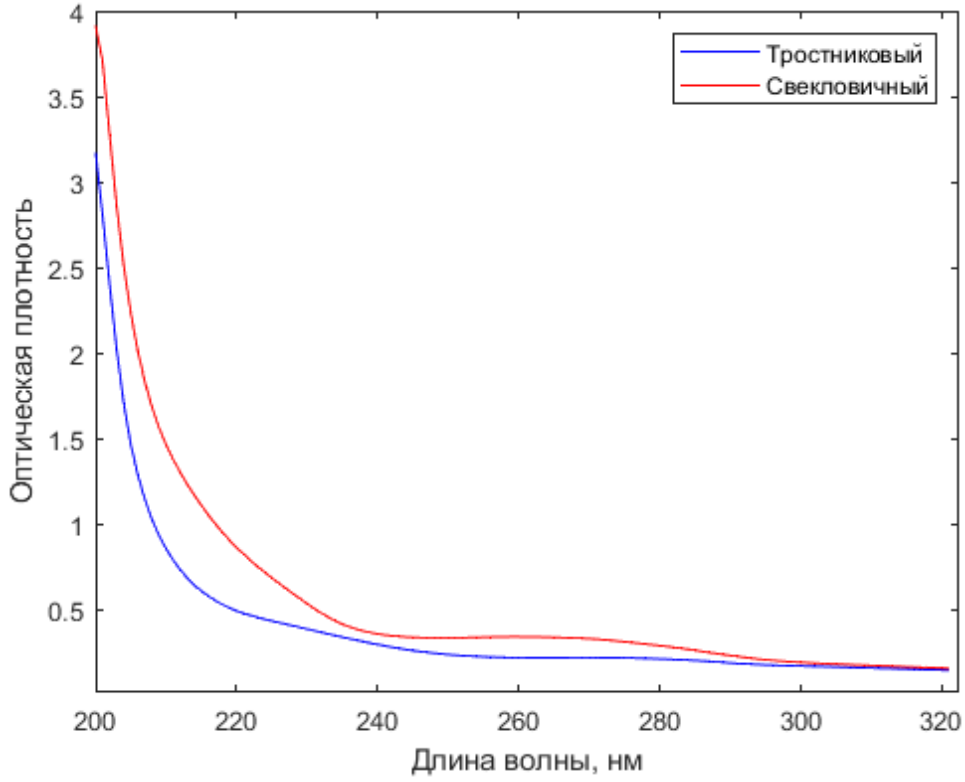

Рисунок 23 – Примеры некоторых спектров оптической плотности тростникового и свекловичного сахаров (200 – 320 нм)

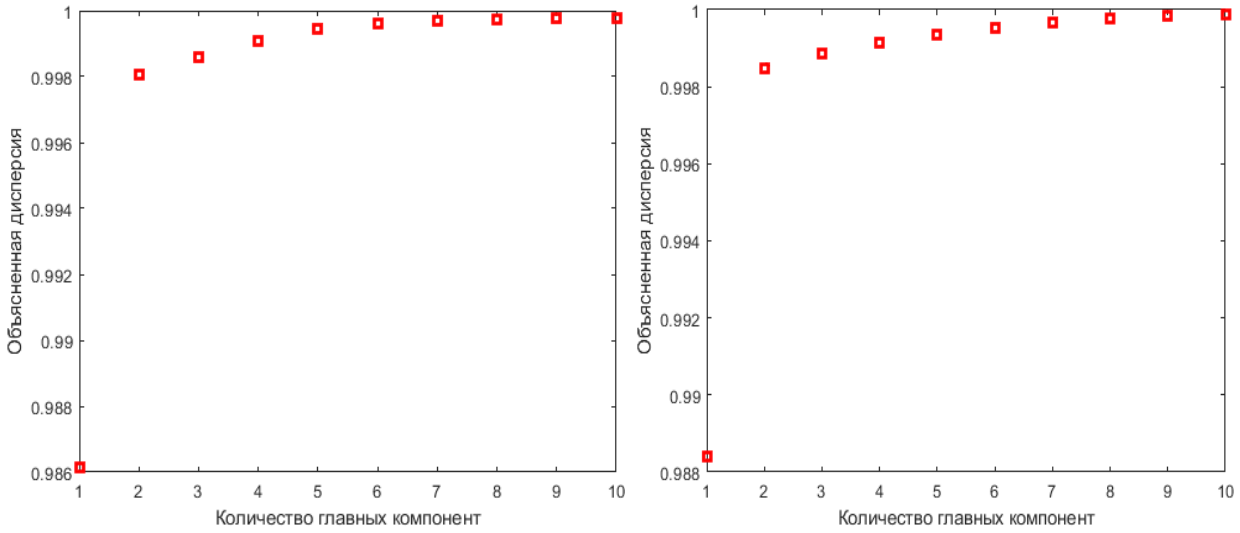

Рисунок 24 – График объясненной дисперсии для 10 главных компонент для 350-1000 нм (слева) и для 200-320 нм (справа)

На рисунках видно, что объясненная дисперсия для первых 10 компонент увеличилась. То есть компоненты стали описывать больше информации.

Далее была подобрана оптимальная метрика – корреляционное расстояние. С помощью перебора возможных комбинаций выбраны оптимальные компоненты – 2 и 8. В данном пространстве точность классификации составила 94 % (рисунок 25).

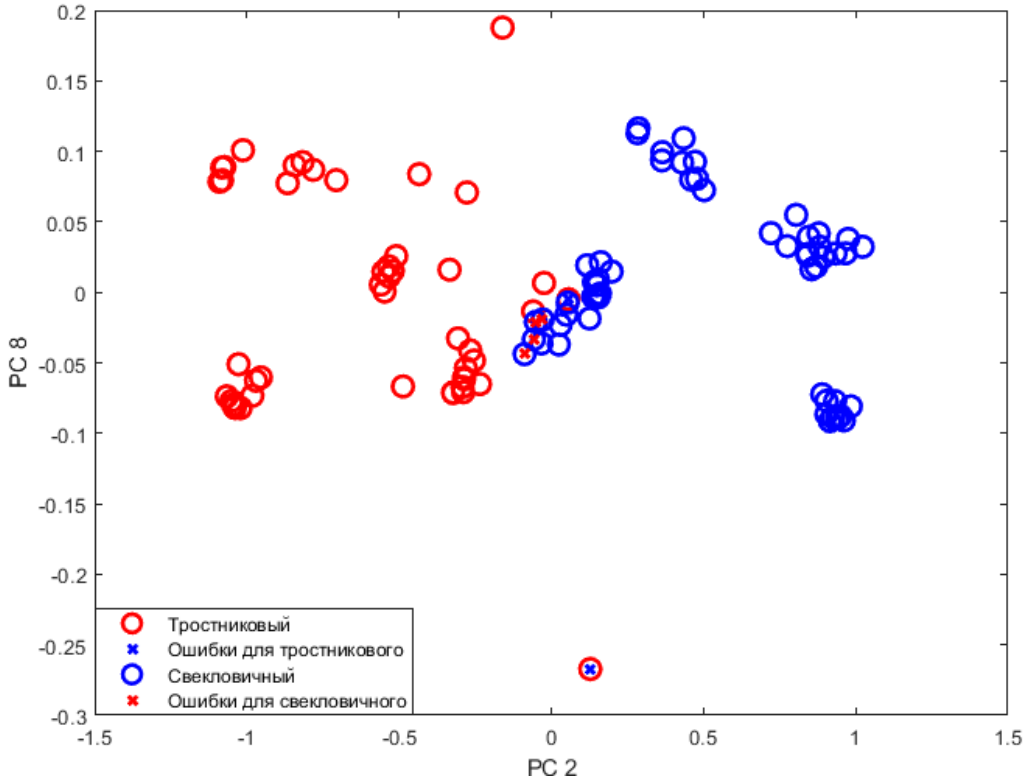

Рисунок 25 – График классификационной модели HCA в пространстве 2 и 8 главных компонент

На рисунке 25 с помощью крестиков указаны ошибочно классифицированные образцы. Видно, что они находятся достаточно близко к объектам другой группы. Один из образцов достаточно сильно отличается от остальных из-за чего тоже был неправильно классифицирован.

3.3 Выбор спектральных переменных

Для дальнейшей работы был использован весь доступный спектральный диапазон (200-1380 нм). Так же было изменено количество образцов. Матрица данных была снова центрирована. Далее вычислялся модуль средней по всем образцам величины разброса оптической плотности от усредненной величины для каждой спектральной переменной (рисунок 26).

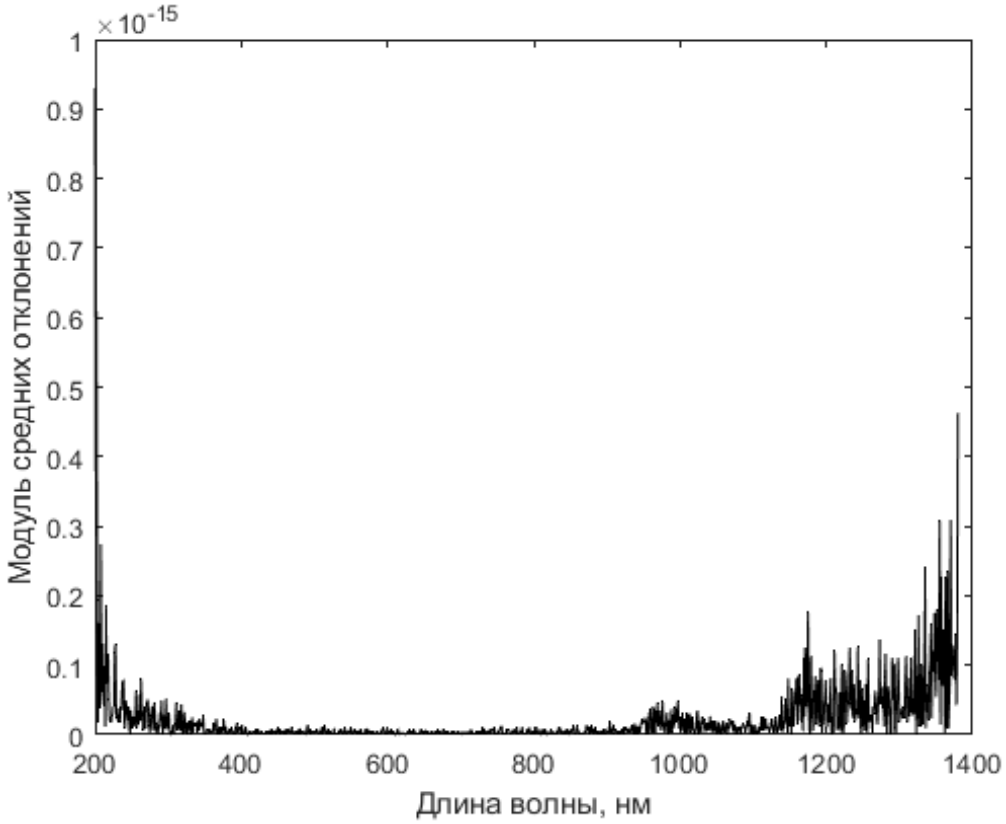

Рисунок 26 – График модуля средней по всем образцам величины разброса оптической плотности от усредненной величины для всех спектральных переменных

Для дальнейшей работы выбирались спектральные переменные в порядке убывания значений модуля средней по всем образцам величины разброса оптической плотности от усредненной величины. Таким образом

сначала использовались 11 спектральных переменных с наибольшими значениями (на 1 больше, чем количество главных компонент). К ним применялся PCA и выбирались две или три компоненты, в пространстве которых строилась модель с наименьшей ошибкой. Далее добавлялась еще одна спектральная переменная и процесс повторялся снова.

Таким образом было выбрано 13 спектральных переменных (рисунок 27) для дальнейшего построения моделей классификации. В дальнейшем выяснится, что данного количества достаточно для построения моделей со 100 % точностью классификации.

На рисунке 27 видно, что наибольшие величины разброса оптической плотности от усредненной величины находятся в диапазоне 200 – 215 нм (подробнее приведен на рисунке 28) и 1335 – 1380 нм (подробнее приведен на рисунке 29).

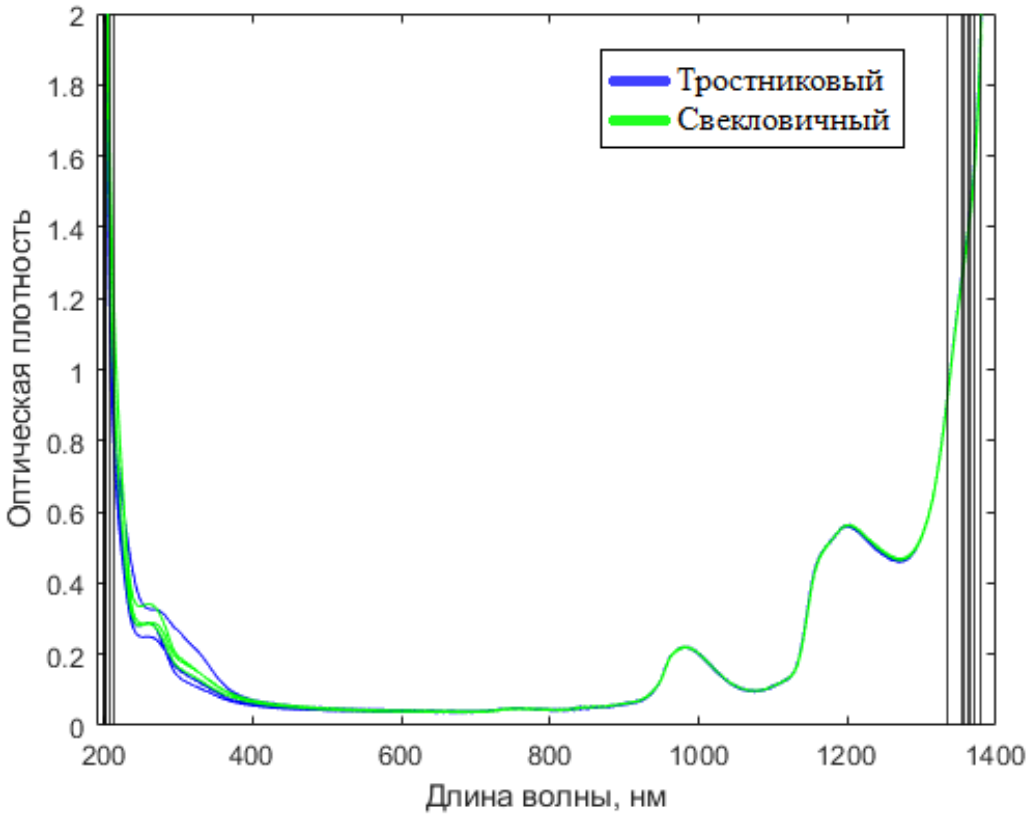

Рисунок 27 – Примеры некоторых спектров оптической плотности сахаров и выбранные спектральные переменные

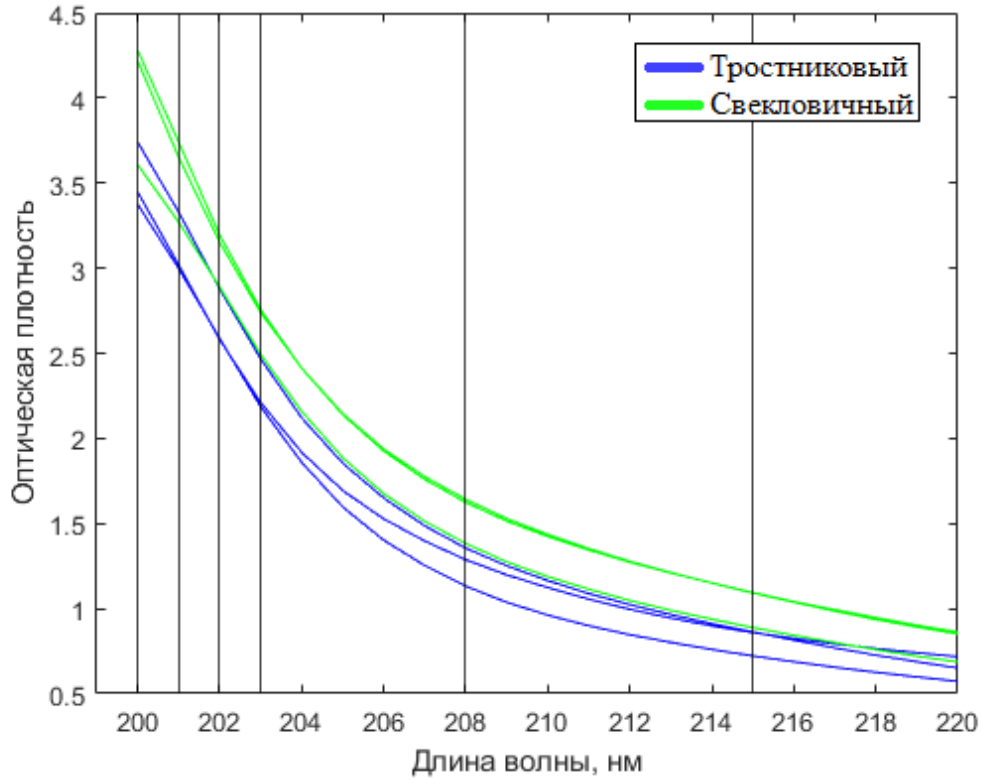

Рисунок 28 – Примеры некоторых спектров оптической плотности сахаров и выбранные спектральные переменные в диапазоне 200 - 220 нм

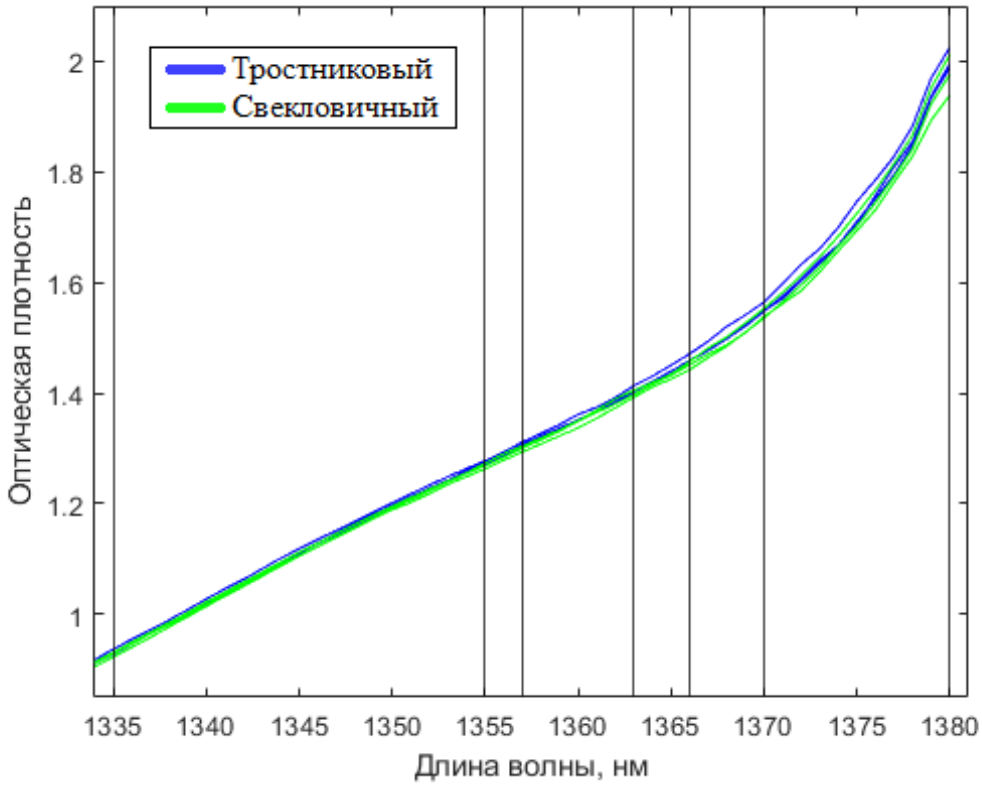

Рисунок 29 – Примеры некоторых спектров оптической плотности сахаров и выбранные спектральные переменные в диапазоне 1335 - 1380 нм

3.4 Результаты построения моделей с выбором спектральных переменных

Для метода построения классификационных деревьев была достигнута 100 % точность при построении модели в пространстве 1 и 4 компонент (рисунок 30). Для данной модели применялась 10-кратная перекрестная проверка.

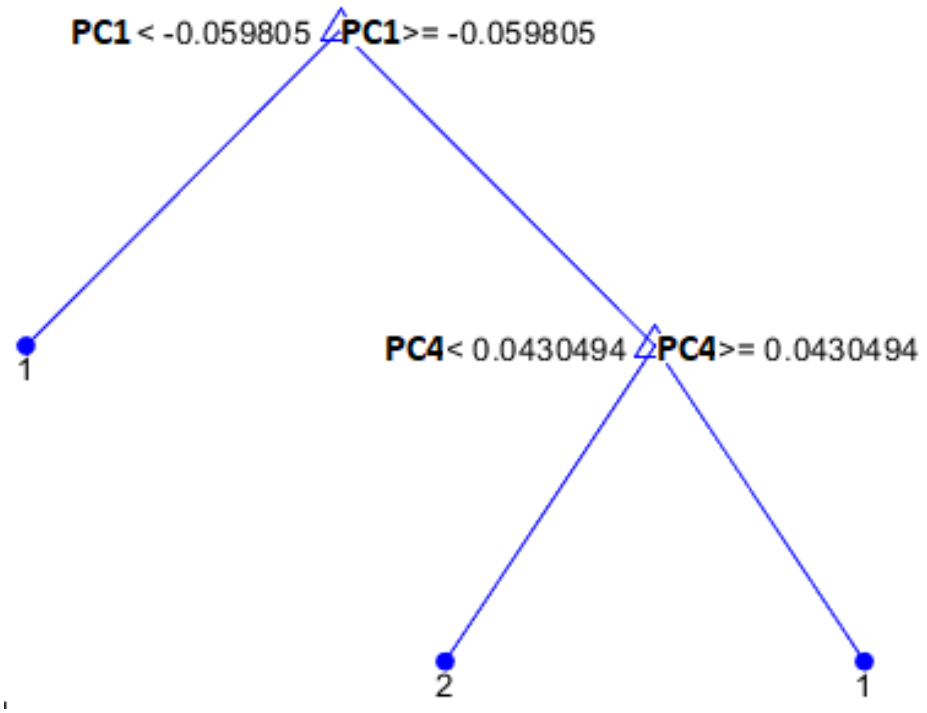

Рисунок 30 – Классификационное дерево для тростникового (1) и свекловичного (2) типов сахара с выбором спектральных переменных

В данном случае алгоритм создал два узла выбора и три листа по двум компонентам. Дерево классификации упростилось, в сравнении с моделью, построенной без выбора спектральных переменных.

Метод k ближайших соседей показал 100 % точность классификации при использовании 1 и 2 компонент и учете 4 ближайших соседей (рисунок 31). Использовалась стандартная метрика - евклидова. Данный результат так же был достигнут после кросс-валидации.

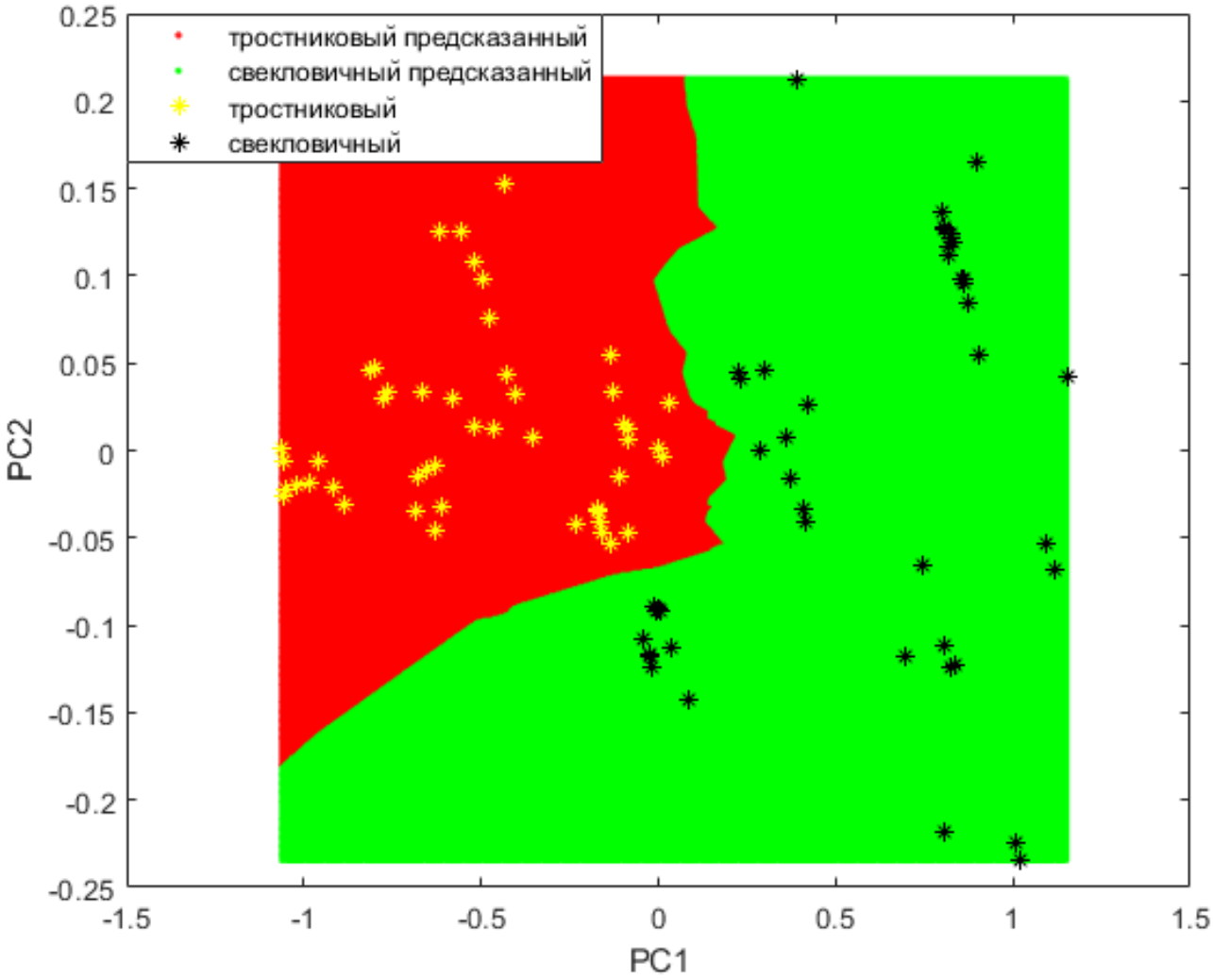

Рисунок 31 – График классификационной модели kNN с наложением графика счетов с выбором спектральных переменных

Была достигнута 100 % точность классификации для LDA (рисунок 32). Изменились главные компоненты для построения модели: до этого использовались 2 и 8, а после применения метода выбора спектральных переменных модель достигла 100 % точности при использовании уже 1 и 2 компонент. Можно сделать вывод, что компоненты стали содержать меньше шума, что позволило повысить точность классификации для других методов кластерного анализа.

100 % точность модели иерархического кластерного анализа была достигнута при использовании 1, 2 и 4 компонент при использовании корреляционной метрики (рисунок 33).

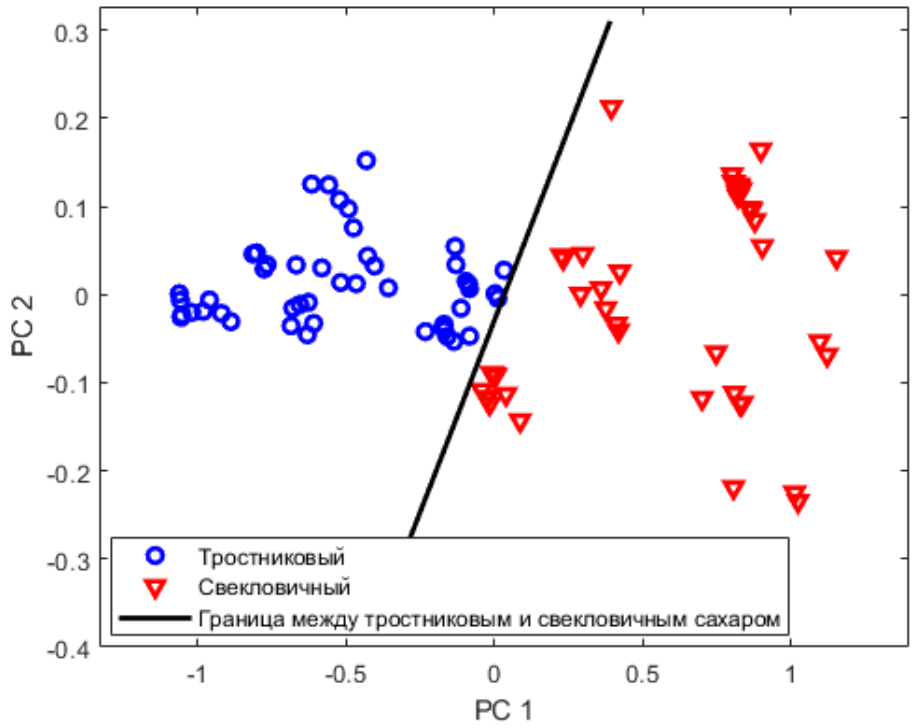

Рисунок 32 – График классификационной модели LDA в пространстве 1 и 2 главных компонент с выбором спектральных переменных

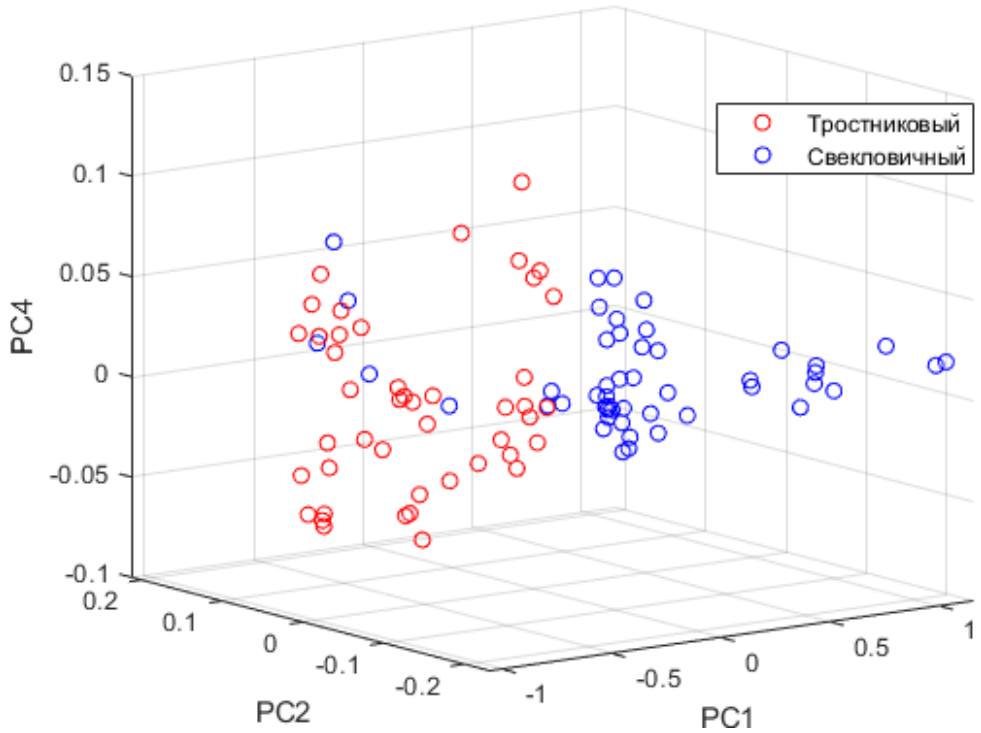

Рисунок 33 – График классификационной модели HCA в пространстве 1, 2 и 4 главных компонент с выбором спектральных переменных

При построении моделей классификации наилучшие результаты были достигнуты при использовании 1, 2 и 4 главных компонент практически во всех случаях.

ЗАКЛЮЧЕНИЕ

В работе продемонстрировано использование оптической спектроскопии и методов многопараметрического анализа для классификации свекловичного и тростникового сахара. Основой была комбинация UV-VIS-NIR спектроскопии поглощения, метода главных компонент и методов кластерного анализа. Получены модели классификации с точностью 94-100 %.

Использование метода выбора спектральных переменных по уменьшению модуля средней по всем образцам величины разброса оптической плотности для выбора маломерного пространства главных компонент и применения методов кластерного анализа в нем позволило достичь 100 % точности классификации образцов свекловичного и тростникового сахара для всех рассмотренных методов классификации. Так же выбор спектральных переменных позволил ограничить количество используемых длин волн до 13.

В данном исследовании была получена достоверная классификация свекловичного и тростникового рафинированных сахаров. Это указывает на возможность замены стандартного дорогостоящего и трудоемкого метода релаксометрии ядерного магнитного резонанса на сравнительно дешевые и простые методы многопараметрического спектрального анализа для классификации растительного источника сахаров.

Результаты работы представлены на 9 конференциях, 1 статья опубликована и 1 статья отправлена в печать.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Discrimination of geographical origin of extra virgin olive oils using terahertz spectroscopy combined with chemometrics / Wei L. [et al.] // Food Chemistry. – 2018 – Vol. 251 – P. 86-92.

2. Discrimination of Transgenic Rice containing the Cry1Ab Protein using Terahertz Spectroscopy and Chemometrics / Xu W. [et al.] // Scientific Reports. – 2015 – Vol. 5 – P. 11115.

3. Rachineni, K. Identifying type of sugar adulterants in honey: Combined application of NMR spectroscopy and supervised machine learning classification / K. Rachineni // Current Research in Food Science. – 2022 – Vol. 5 – P. 272-277.

4. Bachmann, R. Minor metabolites as chemical marker for the differentiation of cane, beet and coconut blossom sugar. From profiling towards identification of adulterations / R. Bachmann // Food Control. – 2022 – Vol.135 – P. 108832.

5. Wei, L. Rapid detection of carmine in black tea with spectrophotometry coupled predictive modelling / L. Wei, Y. Yang // Sun Food Chemistry. – 2020 – Vol. 329 – P. 127177.

6. Adams, M.J. Multivariate Classification Techniques / M.J. Adams // CHEMOMETRICS AND STATISTICS. – 2005 – P. 21-27.

7. Mishra, S. Multivariate Data Analysis / S. Mishra, A. Datta-Gupta // Applied Statistical Modeling and Data Analytics: A Practical Guide foe the Petroleum Geosciences. – 2018 – P. 97-118.

8. Bro, R. Principal component analysis / R. Bro, A. Smilde // Analytical Methods. – 2014 – Vol. 6 – P. 2812–2831.

9. Brown, S.D. Decision Tree Modeling / S.D. Brown, A.J. Myles // Comprehensive Chemometrics (Second Edition). – 2020 – P. 625-659.

10. Berrueta, L.A. Supervised pattern recognition in food analysis / L.A. Berrueta, R.M. Alonso-Salces, K. Héberger // Journal of Chromatography A. – 2007. – Vol. 1158 – P. 196–214.

11. FT-NIR and linear discriminant analysis to classify chickpea seeds produced with harvest aid chemicals / J.P.O. Ribeiro [et al.] // Food Chemistry. – 2021 – Vol. 342 – P. 128324.

12. Multivariate analysis of heavy metals concentrations in river estuary / A.F.M. Alkarkhi [et al.] // Environmental Monitoring and Assessment. – 2008 – Vol. 143 – P. 179–186.

13. Alkarkhi, A.F.M. Discriminant Analysis / A.F.M. Alkarkhi, W.A.A. Alqaraghuli // Applied Statistics for Environmental Science with R. – 2020 – P. 173-190.

14. Deza, M.M. Metrics on Normed Structures / M. M. Deza, E. Deza // Encyclopedia of Distances. – 2016 – P. 97-108.

15. Kolodochka, P.S. Classification of sugar types by UV-VID-NIR spectroscopy and multivariate analysis / P.S. Kolodochka, M.A. Khodasevich // THE 12th INTERNATIONAL CONFERENCE ON PHOTONICS AND APPLICATIONS. – 2022 – P. 244-247.

16. Ходасевич, М.А. Качественный и количественный многопараметрический спектральный анализ / М.А. Ходасевич, Д.А. Королько // VII конгресс физиков Беларуси. – 2023 – С. 105-106.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Колодочка П. С., Ходасевич М. А. Калибровка содержания тростникового сахара в смеси со свекловичным по спектрам оптической плотности водного раствора в УФ и видимой областях с помощью регрессии на главные компоненты // 79-я научная конференция студентов и аспирантов БГУ. – Минск 2022.

2. Kolodochka P. S., Khodasevich M. A. Classification of sugar types by UV-VIS-NIR spectroscopy and multivariate analysis// The 12th international conference on photonics and applications. – Con Dao, Ba Ria - Vung Tau, Vietnam 2022.

3. Колодочка П. С., Ходасевич М. А. Калибровка содержания тростникового сахара в смеси со свекловичным по спектрам оптической плотности водного раствора в УФ и видимой областях с помощью регрессии на главные компоненты // XIX Международная научная конференция «Молодёжь в науке – 2022». – Минск 2022.

4. Колодочка П. С., Ходасевич М. А. Классификация свекловичного и тростникового сахара с помощью применения метода ближайших соседей к главным компонентам спектров пропускания водных растворов // 87-ая научно-техническая конференция профессорско-преподавательского состава, научных сотрудников и аспирантов (с международным участием). – Минск 2023.

5. Ходасевич М. А., Королько Д. А. Качественный и количественный многопараметрический спектральный анализ // VII конгресс физиков Беларуси. – Минск 2023.

6. Колодочка П. С., Куликовская П. А., Ходасевич М. А. Применение методов кластерного анализа к главным компонентам спектров оптической плотности для классификации сахаров и пластмасс // VII Международная научно-техническая конференция "Прикладные проблемы оптики, информатики, радиофизики и физики конденсированного состояния". – Минск 2023.

7. Колодочка П. С., Ходасевич М. А. Классификация типов сахаров с помощью многопараметрического анализа UV-VIS-NIR спектров оптической плотности их водных растворов // 80-я научная конференция студентов и аспирантов БГУ. – Минск 2023.

8. Колодочка П. С., Ходасевич М. А. Классификация типов сахаров с помощью многопараметрического анализа UV-VIS-NIR спектров оптической плотности их водных растворов // XX Международная научная конференция "Молодежь в науке". – Минск 2023.

9. Колодочка П. С., Ходасевич М. А. Классификация географического происхождения лекарственного растительного сырья с помощью

многопараметрического анализа спектров оптической плотности его спиртовых растворов // 88-ая научно-техническая конференция с профессорско-преподавательского состава, научных сотрудников и аспирантов (с международным участием). – Минск 2024.

10. Колодочка П. С., Ходасевич М. А. Классификация географического происхождения лекарственных трав с помощью многопараметрического анализа спектров оптической плотности спиртовых настоек // Журнал Белорусского государственного университета. Физика. – отправлена в редакцию.