МИНОБРНАУКИ РОССИИ

Федеральное государственное автономное образовательное
учреждение высшего образования
«Южный федеральный университет»

Институт математики, механики
и компьютерных наук им. И. И. Воровича

Кафедра информатики и вычислительного эксперимента

Григорян Георгий Зоргевич

Использование стресс-функции для определения размерности пространства характеристик в задачах DATA MINING

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
по направлению подготовки
02.03.02 – Фундаментальная информатика и информационные технологии

Научный руководитель –
доц., к. ф.-м. н. Нестеренко Виктор Александрович

Допущено к защите:
заведующий кафедрой _____ Михалкович С.С.

Ростов-на-Дону — 2022

Задание

на выпускную квалификационную работу

студента 4 курса Григорян Георгия Зоргевича «ИСПОЛЬЗОВАНИЕ СТРЕСС ФУНКЦИИ ДЛЯ ОПРЕДЕЛЕНИЯ РАЗМЕРНОСТИ ПРОСТРАНСТВА ХАРАКТЕРИСТИК В ЗАДАЧАХ DATA MINING»

Цель работы: применение методов уменьшения размерности таких как PCA и Стресс-функция, и оценка качества редукции данных.

Для достижения указанной цели решить следующие задачи:

1. Оценить и выбрать методы уменьшения размерности.
2. Программно реализовать метод Главных Компонент.
3. Применить метод Главных Компонент.
4. Применить стресс-функцию.
5. Анализ полученных результатов.

Научный руководитель,
доц. Кафедры ИВЭ к.ф-м.н. Нестеренко В.А.

АНТИПЛАГИАТ
ОБНАРУЖЕНИЕ ЗАИМСТВОВАНИЙ

СПРАВКА

о результатах проверки текстового документа
на наличие заимствований

Южный Федеральный Университет

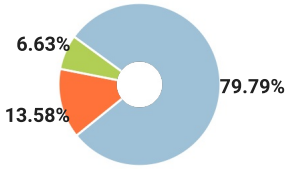
ПРОВЕРКА ВЫПОЛНЕНА В СИСТЕМЕ АНТИПЛАГИАТ.ВУЗ

Автор работы: Григорян Георгий Зоргевич
Самоцитирование рассчитано для: Григорян Георгий Зоргевич
Название работы: Использование стресс-функции для определения размерности пространства характеристик в задачах DATA MINING
Тип работы: Выпускная квалификационная работа
Подразделение: Институт математики, механики и компьютерных наук. Кафедра Информатики и вычислительного эксперимента.

РЕЗУЛЬТАТЫ

ЗАИМСТВОВАНИЯ	13.58%
ОРИГИНАЛЬНОСТЬ	79.79%
ЦИТИРОВАНИЯ	6.63%
САМОЦИТИРОВАНИЯ	0%

ДАТА ПОСЛЕДНЕЙ ПРОВЕРКИ: 17.06.2022



Модули поиска: ИПС Адилет; Библиография; Сводная коллекция ЭБС; Интернет Плюс; Сводная коллекция РГБ; Цитирование; Переводные заимствования (RuEn); Переводные заимствования по eLIBRARY.RU (EnRu); Переводные заимствования по Интернету (EnRu); Переводные заимствования издательства Wiley (RuEn); eLIBRARY.RU; СПС ГАРАНТ; Медицина; Диссертации НББ; Перефразирования по eLIBRARY.RU; Перефразирования по Интернету; Перефразирования по коллекции издательства Wiley; Патенты СССР, РФ, СНГ; СМИ России и СНГ; Модуль поиска "ЮФУ"; Шаблонные фразы; Кольцо вузов; Издательство Wiley; Переводные заимствования

Работу проверил: Нестеренко Виктор Александрович

ФИО проверяющего

Дата подписи:

Подпись проверяющего

Чтобы убедиться
в подлинности справки, используйте QR-код,
который содержит ссылку на отчет.

Ответ на вопрос, является ли обнаруженное заимствование корректным, система оставляет на усмотрение проверяющего. Предоставленная информация не подлежит использованию в коммерческих целях.

Оглавление

Введение

На протяжении многих лет объемы данных, предоставленных людям, стремительно увеличиваются. Из-за этого перед человечеством встает проблема извлечения полезной информации из них. Поэтому на помощь к людям приходит такая технология как Data Mining. Data Mining — это интеллектуальный анализ данных. Данная технология пришла на замену прикладной статистике, следовательно отсюда проистекает изобилие методов и алгоритмов. Сам же термин “Data Mining” часто переводится как добыча данных, извлечение информации. Одной из важных задач в Data Mining является уменьшение размерности.

Для чего же нужна редукция размерности пространства признаков? Во-первых, большое количество признаков требуют большего времени для вычислений. Во-вторых, большие вычисления более ресурсоемкие. В-третьих, в любых данных есть шум, который негативно влияет на обучение какой-либо модели. Кроме того, информацию, представленную в двумерном или трехмерном измерениях, можно легко визуализировать, чем при боле высоких измерениях. Есть множество методов, позволяющих сделать редукцию пространства признаков данных, но в своей работе я опишу и реализую Метод Главных Компонент либо же Анализ Основных Компонент (PCA, Principal Component Analysis), а также Стресс-Функцию.

Постановка задачи

Целью дипломной работы является применение методов уменьшения размерности таких как PCA и Стресс-функция, и оценка качества редукции данных.

В выпускной работе требуется выполнить следующие задачи

Основные этапы выполнения задач:

6. Оценка и выбор метода уменьшения размерности.
7. Программная реализация метода Главных Компонент.
8. Применение метода Главных Компонент.
9. Применение стресс функции.
10. Анализ полученных результатов.

1. Принцип работы алгоритма

Метод Главных Компонент — это способ выявления закономерностей в данных и выражения данных таким образом, чтобы подчеркнуть их сходства и различия. Поскольку закономерности может быть трудно найти в данных большой размерности, где роскошь графического представления недоступна, PCA является мощным инструментом для анализа данных. Суть алгоритма состоит в том, что уменьшение количества фич происходит за счет точности новых данных, так как главные компоненты являются линейной комбинацией признаков.

С геометрической точки зрения, главные компоненты представляют собой векторы данных, которые объясняют максимальное количество отклонений. Главные компоненты – новые оси, которые обеспечивают лучший угол для оценки данных, чтобы различия между наблюдениями были лучше видны. Поскольку существует столько главных компонент, сколько переменных в наборе, главные компоненты строятся таким образом, что первый из них учитывает наибольшую возможную дисперсию в наборе

данных.

Вычисление главных компонент может быть сведено к вычислению сингулярного разложения матрицы данных или к вычислению собственных векторов и собственных значений ковариационной матрицы исходных данных. В моем случае алгоритм реализован через нахождение собственных значений и векторов ковариационной матрицы.

2.1 Подготовка данных

Для демонстрации работы алгоритма был сгенерирован data set состоящий из вещественных чисел (Рис. 1). Размерность пространства данных равнялось 4 признакам, такая размерность взята для проверки редукции к разному количеству главных компонент.

x_1	x_2	x_3	x_4
1.0	2.73446908	11.01792737	98.47615116
2.0	4.35122722	12.38619263	20.46282367
3.0	7.21132988	11.48804931	67.17093415
4.0	11.24872601	24.10099224	77.66304739
5.0	9.58103444	30.21117481	142.25100662
6.0	12.09865079	32.83617975	127.60618803
7.0	13.78706794	42.4237342	199.82062948
8.0	13.85301221	51.14366228	110.97707692
9.0	15.29003911	47.1998583	218.21847896
10.0	18.0998018	60.19540684	269.12150528

Рис. 1. Таблица сгенерированных данных.

Для корректной работы PCA необходимо центрировать данные, так как метод очень чувствителен к дисперсиям, т.е. вычесть из каждого значения

столбца среднее арифметическое этого столбца (Рис. 2). Таким образом среднее арифметическое нормализованных данных будет равняться нулю.

x_1	x_2	x_3	x_4
-4,50000000	-8,09106677	-21,28239040	-34,70063301
-3,50000000	-6,47430863	-19,91412514	-112,71396050
-2,50000000	-3,61420597	-20,81226846	-66,00585002
-1,50000000	0,42319016	-8,19932553	-55,51373678
-0,50000000	-1,24450141	-2,08914296	9,07422245
0,50000000	1,27311494	0,53586198	-5,57059614
1,50000000	2,96153209	10,12341643	66,64384531
2,50000000	3,02747636	18,84334451	-22,19970725
3,50000000	4,46450326	14,89954053	85,04169479
4,50000000	7,27426595	27,89508907	135,94472111

Рис. 2. Таблица нормализованных данных.

2.2 Вычисление ковариационной матрицы

Для начала стоит разобраться что такое ковариация. Ковариация или корреляционный момент — мера зависимости одной случайной величины от другой. В нашем случае формула, по которой вычисляется ковариация значений двух измерений будет выглядеть так:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)}$$

Рис.3. Формула ковариации двух векторов.

где $\overline{X}$, $\overline{Y}$ среднее арифметическое значений измерений X, Y.

Ковариационная матрица для n-мерного пространства признаков будет выглядеть так:

$$C^{n*n} = \left(c_{i.j}, c_{i,j} = Cov\left(Dim_i, Dim_j\right)\right)$$

Рис.4 Ковариационная матрица

где Dim_x одно из возможных измерений.

$$\mathrm{C} = \begin{pmatrix} Cov(X,X) & Cov(X,Y) & Cov(X,Z) \\ Cov(X,Y) & Cov(Y,Y) & Cov(Y,Z) \\ Cov(X,Z) & Cov(Y,Z) & Cov(Z,Z) \end{pmatrix}$$

Рис.5 Пример ковариационной матрицы для данных с размерностью пространства.

В матрице коэффициентов ковариационной матрицы имеет значение знаки этих коэффициентов. Если знак – это:

- положительное число, то две переменные прямо пропорциональны, то есть второй увеличивается или уменьшается вместе с первым.
- отрицательное число, то переменные обратно пропорциональны, то есть второй увеличивается, когда первый уменьшается, и наоборот.

	x_1	x_2	x_3	x_4
x_1	9,16666666	14,34448508	52,34245948	193,94202309
x_2	14,34448508	23,84180375	81,40260965	300,65742902
x_3	52,34245947	81,40260965	312,47012667	1122,85505928
x_4	193,94202309	300,65742902	1122,85505928	5789,75302839

Рис. 5. Ковариационная матрица сгенерированных данных.

2.3 Извлечение собственных векторов и собственных чисел

Что бы определить Главные Компоненты необходимо извлечь собственные векторы (eigenvectors) и собственные числа (eigenvalues) из ковариационной матрицы. Главная Компонента – это новая переменная, комбинированная таким образом, что новые переменные не коррелированы между собой, и основная информация об исходных переменных помещается в первых компонентах. Для извлечения собственных чисел и собственных векторов был применен Метод Якоби – итерационный алгоритм применимый к вещественной симметричной матрице. Собственные векторы дают нам представление о направлении осей, где наблюдается наибольшая дисперсия т.е. большая часть информации, а собственные числа — значения, показывающие величину этой дисперсии у каждого собственного вектора. Таким образом отсортировав собственные векторы по их собственным числам, мы получаем Главные Компоненты в порядке их насыщенности информацией.

2.4 Формирование нового вектора фич

Для построения новой матрицы фич необходимо взять собственные векторы, обладающие наибольшей дисперсией и сформировать матрицу из этих векторов в столбцах.

$$FeatureVector = (ev_1, ev_2, \dots, ev_n)$$

где ev_k – собственный вектор.

2.5 Получение нового набора данных

Это последний шаг для метода главных компонент. После того, как был сформирован новый вектор фич, необходимо транспонировать его и умножить слева на транспонированный набор исходных данных.

$$FinalData = FeatureVector * MeanedData,$$
Рис.6. Проекция на ось главных компонент.

Где FeatureVector это транспонированная матрица собственных векторов со значениями в строках отсортированные сверху-вниз по порядку значимости, а MeanedData это транспонированные центрированные начальные данные. Цель этого перемножения переориентировать набор данных с исходной оси, на оси, представленные главными компонентами.

2.6 Восстановление данных.

Проекция данных дает огромные возможности для работы с ней, однако она не дает явного понимания какая информацию она содержит. Для понимания полной картины необходимо восстановить данные. Для этой процедуры все необходимое вычислено: Средние значения векторов признаков, данные спроецированный на оси главных компонент, собственные векторы.

$$RestoredData = FinalData * FeatureVectors + MeanVector,$$
Рис.7. Восстановление данных.

В данном случае восстановленные данные получаются умножением транспонированной матрицы собственных векторов на спроецированные данные справа. После умножения следует построчно прибавить вектор средних (MeanVector).

2. Применение Метода Главных Компонент

Для проверки реализации алгоритма было принято решение воспользоваться дата сетом Iris, который представляет себя данные о трех видах Ириса.

	sepal length	sepal width	petal length	petal width	target
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

Рис.8. Ирисы Фишера.

Первым делом было принято решение визуализировать данные.

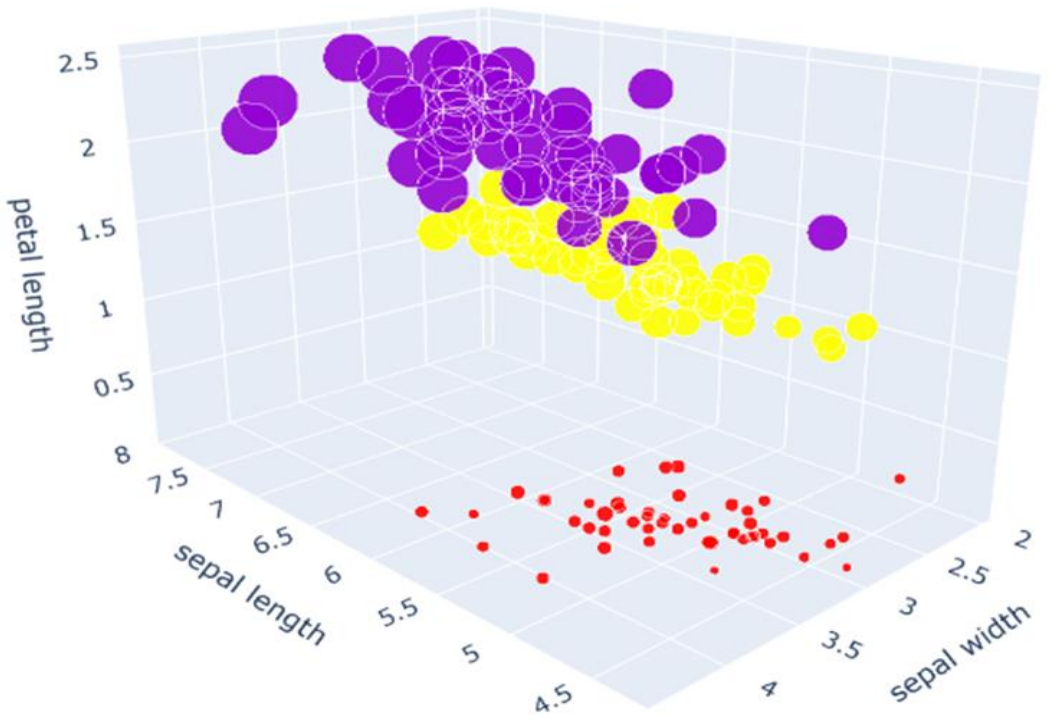


Рис.9. Представление Ирисов Фишера в размерности пространства 4.

На данном графике цветами представлены виды ирисов. Оси X, Y, Z заняты такими признаками как ‘petal length’, ‘sepal length, ‘sepal width’. Признак ‘petal width показывает размер окружности каждой элемента. Как видно из этого графика 4 признака избыточны для визуализации, в добавок они приносят сложность в восприятии графика. Попробуем уменьшить размерность. Нормализуем данные.

	sepal length	sepal width	petal length	petal width
0	-0.743333	0.446	-2.358667	-0.998667
1	-0.943333	-0.054	-2.358667	-0.998667
2	-1.143333	0.146	-2.458667	-0.998667
3	-1.243333	0.046	-2.258667	-0.998667
4	-0.843333	0.546	-2.358667	-0.998667
...	...	...	...	...
145	0.856667	-0.054	1.441333	1.101333
146	0.456667	-0.554	1.241333	0.701333
147	0.656667	-0.054	1.441333	0.801333
148	0.356667	0.346	1.641333	1.101333
149	0.056667	-0.054	1.341333	0.601333

Рис.10. Нормализованные данные.

Посчитаем ковариационную матрицу.

	sepal length	sepal width	petal length	petal width
sepal length	0.68569351	-0.03926846	1.27368233	0.5169038
sepal width	-0.03926846	0.18800403	-0.32171275	-0.11798121
petal length	1.27368233	-0.32171275	3.11317942	1.29638747
petal width	0.5169038	-0.11798121	1.29638747	0.58241432

Рис.9.Ковариационная матрица.

На ковариационной матрице видны зависимости между признаками, например, при росте параметра ‘sepal length’ уменьшается параметр ‘sepal width’ и наоборот.

Далее на основе ковариационной матрицы следует вычислить собственные векторы и собственные числа. Получаем следующий результат:

Собственные числа:

$$\begin{pmatrix} 0.02368303 \\ 0.07852391 \\ 0.24224357 \\ 4.22484077 \end{pmatrix}$$

Собственные векторы:

$$\begin{pmatrix} 0.31725455 \\ -0.32409435 \\ -0.47971899 \\ 0.75112056 \end{pmatrix} \begin{pmatrix} 0.58099728 \\ -0.59641809 \\ -0.07252408 \\ -0.54906091 \end{pmatrix} \begin{pmatrix} 0.65653988 \\ 0.72971237 \\ -0.1757674 \\ -0.07470647 \end{pmatrix} \begin{pmatrix} -0.36158968 \\ 0.08226889 \\ -0.85657211 \\ -0.35884393 \end{pmatrix}$$

Основываясь на значениях собственных чисел, мы видим, что 4 собственный вектор несет в себе наибольшую информацию о данных. Отсортируем их в порядке убывания значений собственных чисел. Попробуем уменьшить размерность пространства с 4 до меньших.

Для размерности пространства 3 график будет выглядеть следующим образом:

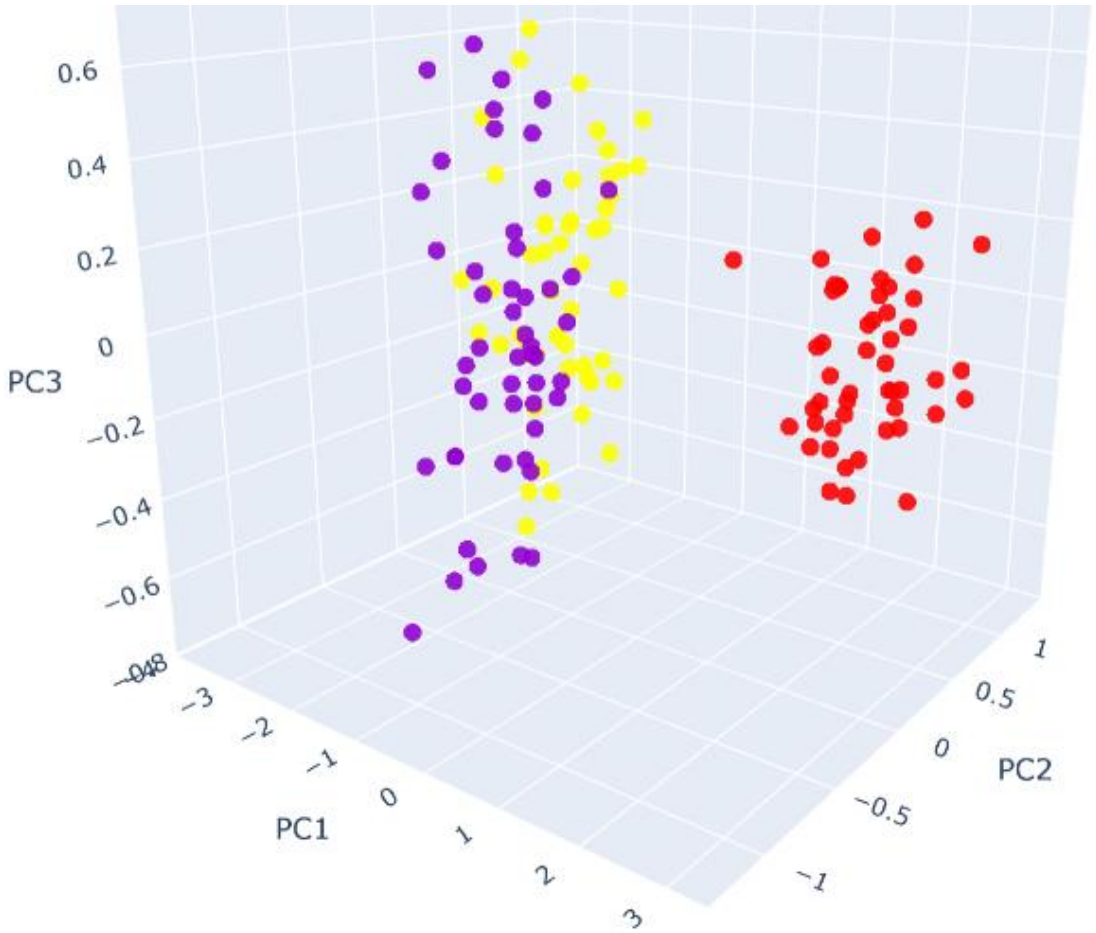


Рис.11. График размерности пространства 3.

Данные представлены в трехмерном пространстве, где осями являются 3 главные компоненты, а цвет отвечает за вид ирисов. Данные стало воспринимать проще, однако до конца не ясны границы между видами. Уменьшим размерность пространства до 2х.

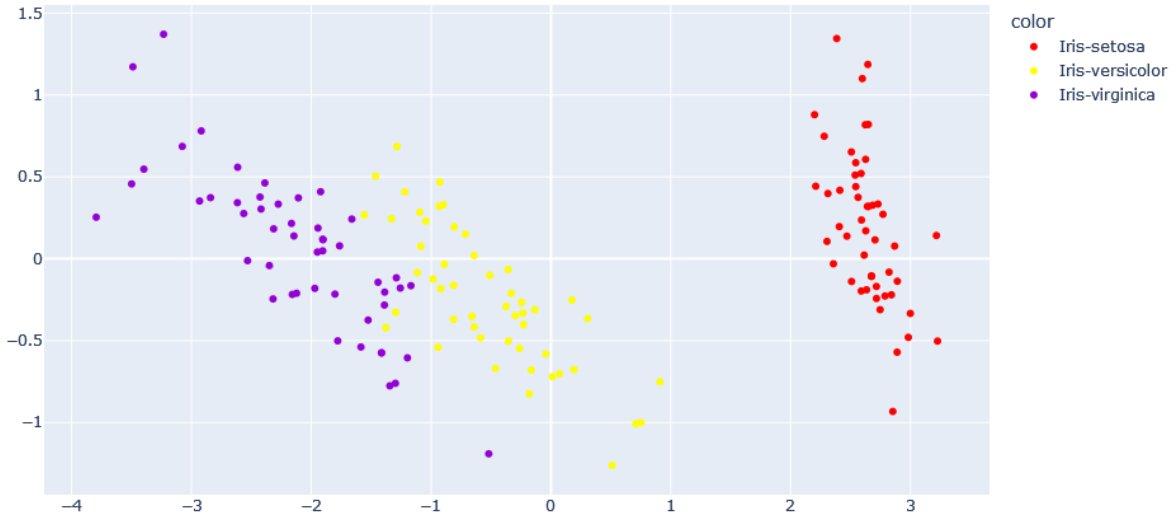


Рис.12. График размерности пространства 2.

Различия между видами ирисов стали более наглядны. Попробуем уменьшить размерность до 1.

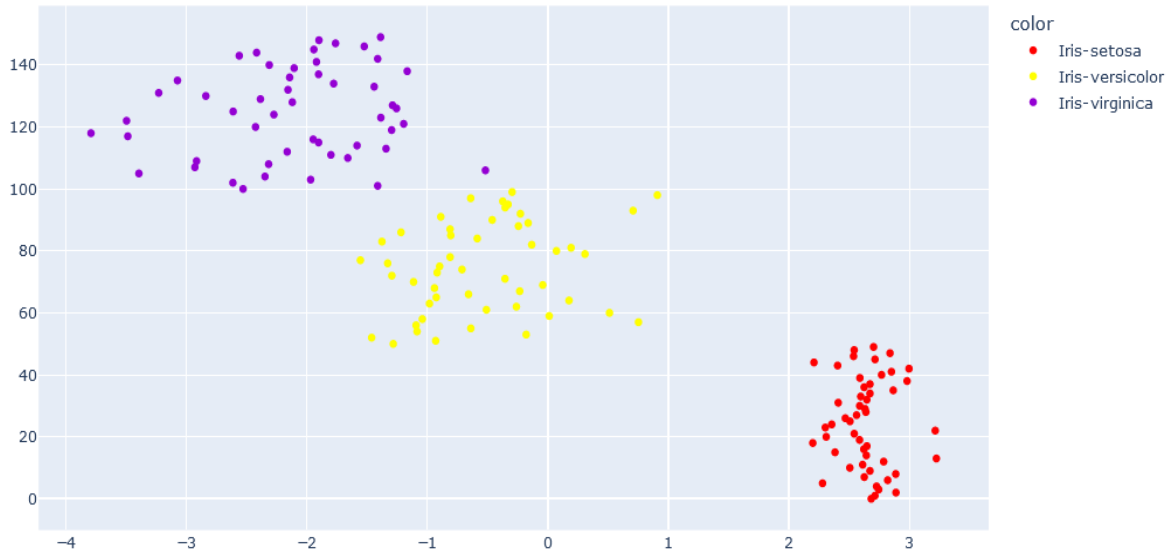


Рис.13. График размерности пространства 1.

На оси X- значение Главной компоненты, на оси Y – порядковый номер элемента. В отличии от рисунка 12 на рисунке 13 расстояния между точками намного меньше. И различия между тремя кластерами выражены лучшим образом.

3. Использование Стресс-функции. Оценка размерности пространства данных.

Во многих случаях PCA отличный и быстрый инструмент для редукции размерности пространства данных. Однако при нахождении вектора, при котором дисперсия максимальна может потеряться значительная часть данных. К примеру, возьмем два плоских диска одинокого размера расположенные друг на друге. PCA даст вектор максимальной дисперсии данных параллельный радиусу диска. Если спроецировать данные на вектор, и восстановить их, то потеряется большая часть из них, т.к. точки, находящиеся на 2ом диске, не были учтены. Для этого воспользуемся другим методом редукции пространства данных, таким как Стресс-функция. Принцип работы стресс-функции заключается в том, что точки размещаются так, чтобы попарные расстояния между ними в новом пространстве как можно меньше отличались от измеренных изначальных данных. Мера различий расстояний в исходном и новом пространстве называется функцией стресса.

Как работает стресс-функция:

1. Расчет матрицы приближения $D^2 = [d_{ij}^2]$, где d_{ij}^2 – евклидово расстояние между парами точек
2. Применение двойного центрирования: $B = -\frac{1}{2} C D^2 C$ с использованием центрирующей матрицы $C = I - \frac{1}{n} J_n$ где I — единичная матрица, а J_n матрица заполненная единицами.
3. Вычисление наибольшего собственного значения и соответствующего ему собственного векторов для матрицы B.

Попробуем уменьшить размерность пространства на основе стресс-функции. Для размерности пространства 3 получим следующий график со стандартным разбиением на виды ирисов:

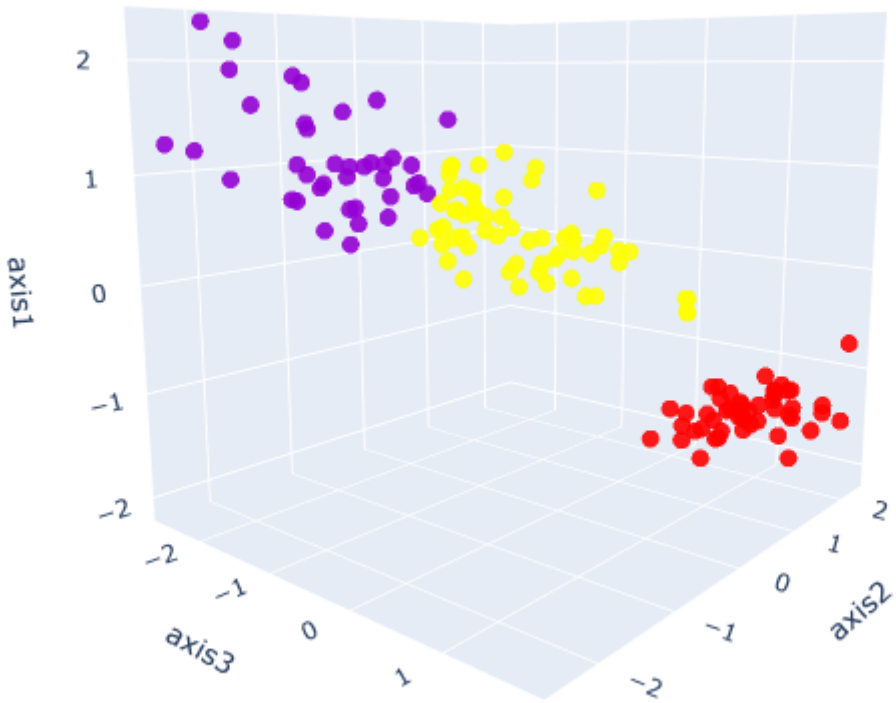


Рис.14. График размерности пространства 3 для стресс-функции.

График имеет значительное сходство с графиком четырехмерного пространства(рис.9). Снизим размерность до двух. Получим следующий график:

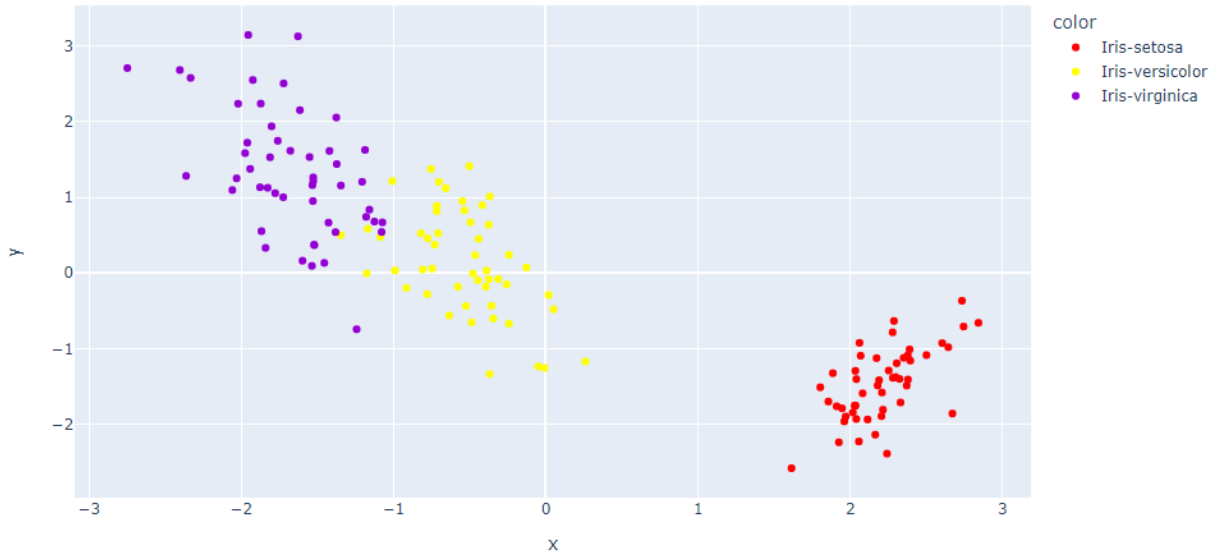


Рис.15. График размерности пространства 2 для стресс-функции.

Если рассмотреть график для размерности 3 с определенного ракурса, то будет видно сходство с рисунком 15.

Попробуем кластеризовать данные. Для этого используем один из методов кластеризации K-means. Изначально данные (MainData) были кластеризованы по видам ирисов. Применим K-means. Начальная кластеризация отличается от полученной в 16 пунктах. Назовем ее Origin. Кластеризуем данные, которые были получены использованием метода главных компонент. Сравним их со значением origin:

	origin	PC1	PC2	PC3	compare1	compare2	compare3
50	Iris-versicolor	Iris-versicolor	Iris-virginica	Iris-versicolor	NaN	False	NaN
52	Iris-virginica	Iris-versicolor	Iris-virginica	Iris-virginica	False	NaN	NaN
114	Iris-versicolor	Iris-virginica	Iris-versicolor	Iris-versicolor	False	NaN	NaN
146	Iris-versicolor	Iris-virginica	Iris-versicolor	Iris-versicolor	False	NaN	NaN

Рис.16. Точки, принадлежащие разным кластерам (PCA).

В данном случае PC1, PC2 и PC3 это названия кластеров, полученных на основе пространств, извлеченных с помощью PCA. Compare1, compare2, compare3 это значения полученные при сравнении кластеров origin с проекциями данных на оси Главных компонент, размерность которых соответственно равна 1,2,3. Как видно из таблицы между Origin и PC1 3 промаха, а между Origin и PC2 промах 1.

Так же кластеризируем данные, полученные с помощью уменьшения размерности при использовании стресс-функции.

	origin	Axis1	Axis2	Axis3	OA1	OA2	OA3
114	Iris-versicolor	Iris-virginica	Iris-virginica	Iris-versicolor	False	NaN	NaN
114	Iris-versicolor	Iris-virginica	Iris-virginica	Iris-versicolor	NaN	False	NaN

Рис.17 Точки, принадлежащие разным кластерам (Стресс-функция)

Axis1, Axis2, Axis3 аналогичные кластеры, как и PC1, PC2, PC3, только полученные на основе стресс-функции. OA1, OA2, OA3 это значения, полученные при сравнении с origin. Как видно из таблицы между Origin и Axis1, а так же Axis2 по одному промаху.

Оценим данные, полученные проецированием на оси главных компонент и при использовании стресс-функции. Для этого используем индекс Rand. Индекс оценивает, насколько много из тех пар элементов, которые находились в одном классе, и тех пар элементов, которые находились в разных классах, сохранили это состояние после кластеризации алгоритмом.

$$Rand = \frac{TP + TN}{TP + TN + FP + FN}$$

Элементы принадлежат одному кластеру и одному классу — TP

Элементы принадлежат одному кластеру, но разным классам — FP

Элементы принадлежат разным кластерам, но одному классу — FN

Элементы принадлежат разным кластерам и разным классам — TN

Отсюда получим различные значения для разных пар кластеризаций:

Для PCA:

Rand (Origin, PC1) = 0.9739597315436241

Rand (Origin, PC2) = 0.9911409395973154

Rand (Origin, PC3) = 1.0

Rand (MainData, PC1) = 0.8987919463087248

Rand (MainData, PC2) = 0.8737360178970918

Rand (MainData, PC3) = 0.8797315436241611

Для Стресс-функции:

Rand (Origin, Axis1) = 0.9911409395973154

Rand (Origin, Axis2) = 0.9911409395973154

Rand (Origin, Axis3) = 1.0

Rand (MainData, Axis1) = 0.8859060402684564

Rand (MainData, Axis2) = 0.8859060402684564

Rand (MainData, Axis3) = 0.8797315436241611

Построим графики для PC2 и Axis2, размерность пространства для них равна двум. На плоскости различия между начальным распределением ирисов и новыми кластерами видна наилучшим образом.

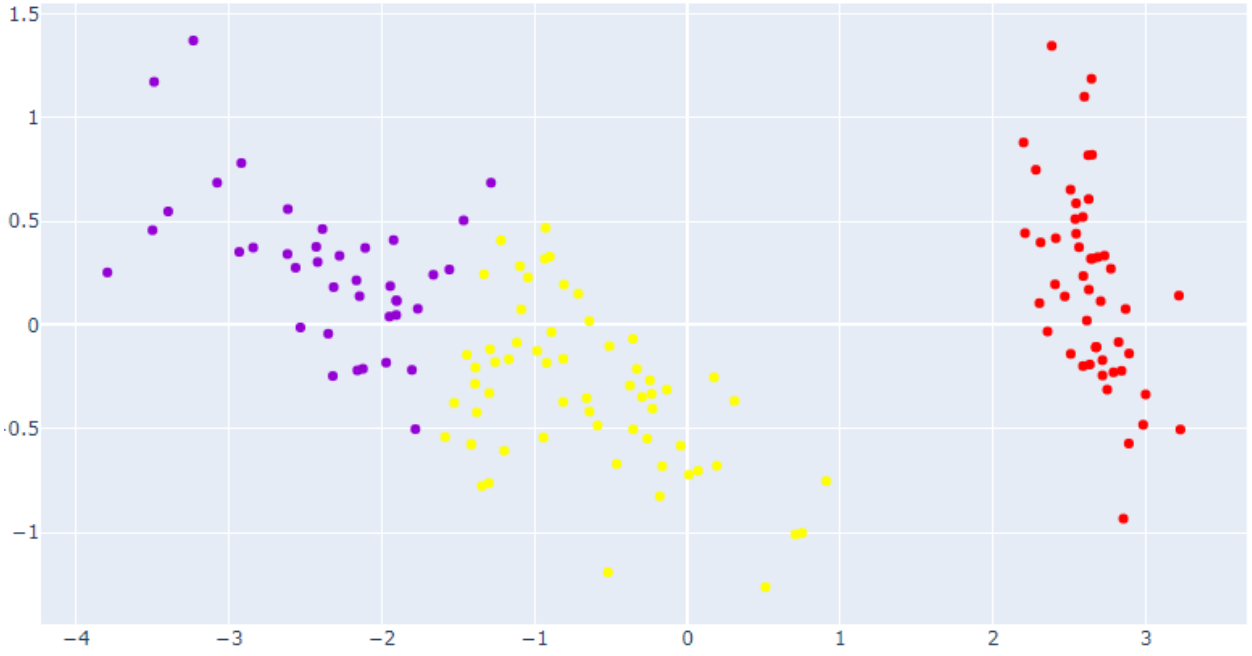


Рис.17 Кластеры PC2.

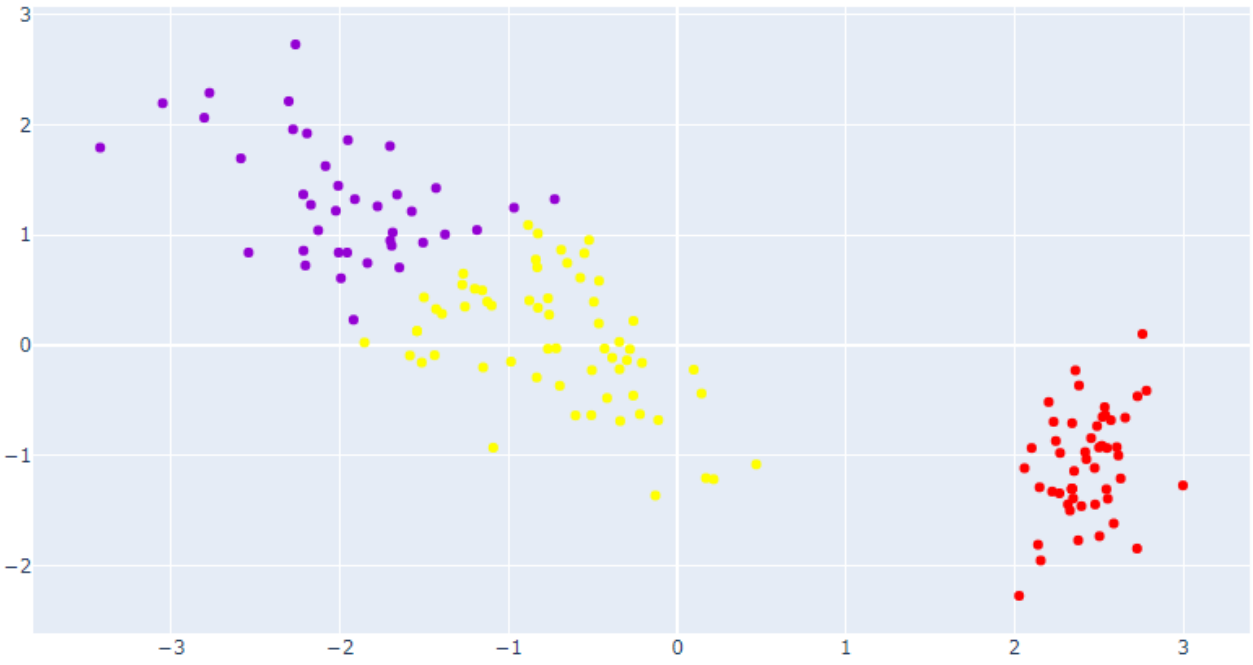


Рис.18 Кластеры Axis2.

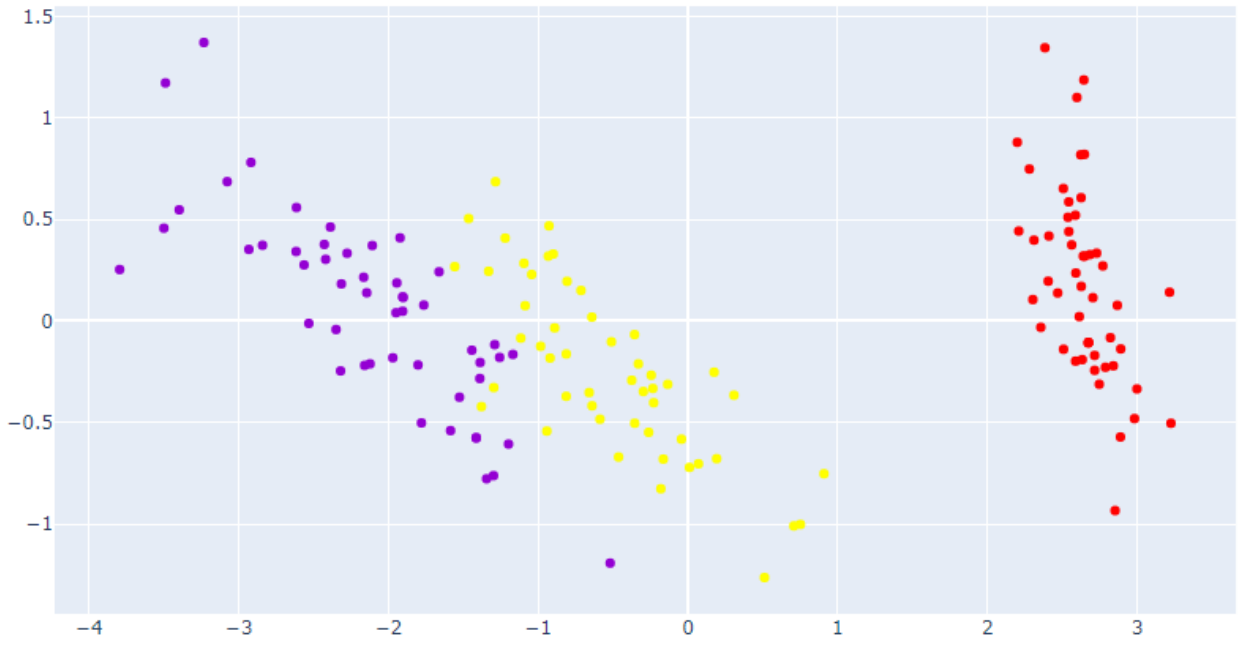


Рис.19 PC2 при начальном распределении ирисов.

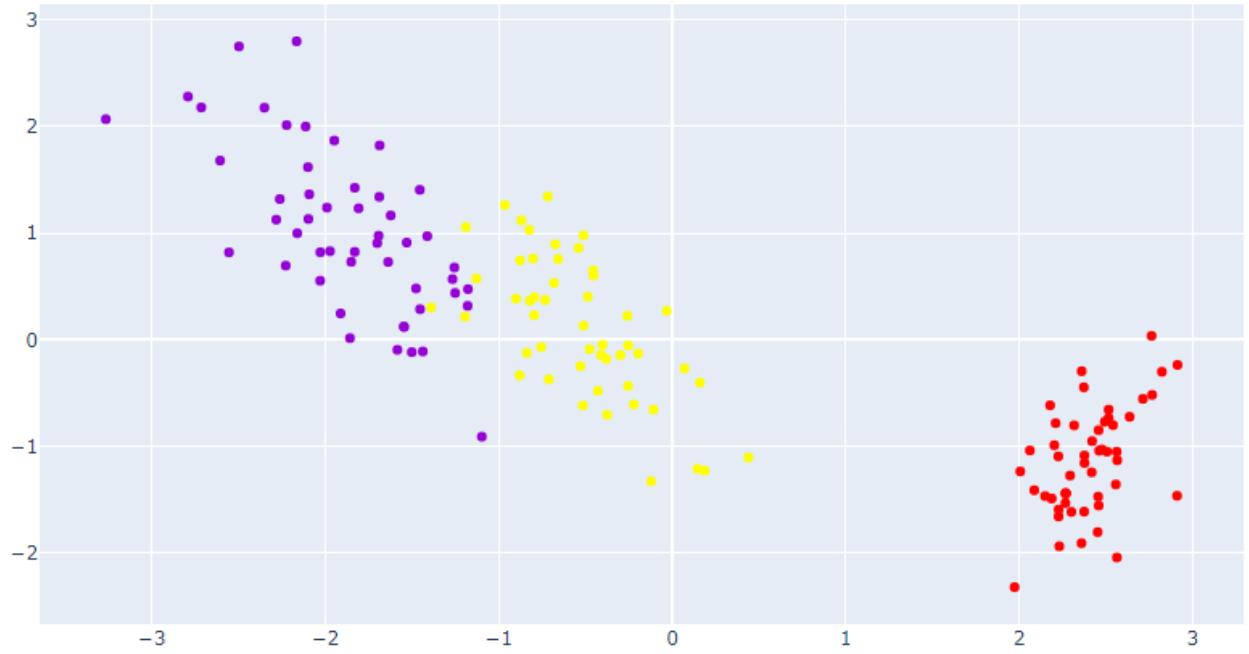


Рис.20 Axis2 при начальном распределении ирисов.

Как видно из различных графиков в отличие от PCA, метод стресс функции позволяет сохранить форму данных в приближенном к начальному виду.

Так же, оба метода были применены к такому датасету как параметры мужских и женских голосов, состоящий из 20 признаков.

	meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode	centroid	meanfun	minfun	maxfun
0	0.059781	0.064241	0.032027	0.015071	0.090193	0.075122	12.863462	274.402906	0.893369	0.491918	0.000000	0.059781	0.084279	0.015702	0.275862
1	0.066009	0.067310	0.040229	0.019414	0.092666	0.073252	22.423285	634.613855	0.892193	0.513724	0.000000	0.066009	0.107937	0.015826	0.250000
2	0.077316	0.083829	0.036718	0.008701	0.131908	0.123207	30.757155	1024.927705	0.846389	0.478905	0.000000	0.077316	0.098706	0.015656	0.271186
3	0.151228	0.072111	0.158011	0.096582	0.207955	0.111374	1.232831	4.177296	0.963322	0.727232	0.083878	0.151228	0.088965	0.017798	0.250000
4	0.135120	0.079146	0.124656	0.078720	0.206045	0.127325	1.101174	4.333713	0.971955	0.783568	0.104261	0.135120	0.106398	0.016931	0.266667
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3163	0.131884	0.084734	0.153707	0.049285	0.201144	0.151859	1.762129	6.630383	0.962934	0.763182	0.200836	0.131884	0.182790	0.083770	0.262295
3164	0.116221	0.089221	0.076758	0.042718	0.204911	0.162193	0.693730	2.503954	0.960716	0.709570	0.013683	0.116221	0.188980	0.034409	0.275862
3165	0.142056	0.095798	0.183731	0.033424	0.224360	0.190936	1.876502	6.604509	0.946854	0.654196	0.008006	0.142056	0.209918	0.039506	0.275862
3166	0.143659	0.090628	0.184976	0.043508	0.219943	0.176435	1.591065	5.388298	0.950436	0.675470	0.212202	0.143659	0.172375	0.034483	0.250000
3167	0.165509	0.092884	0.183044	0.070072	0.250827	0.180756	1.705029	5.769115	0.938829	0.601529	0.267702	0.165509	0.185607	0.062257	0.271186

Рис.21 Датасет характеристик мужских и женских голосов.

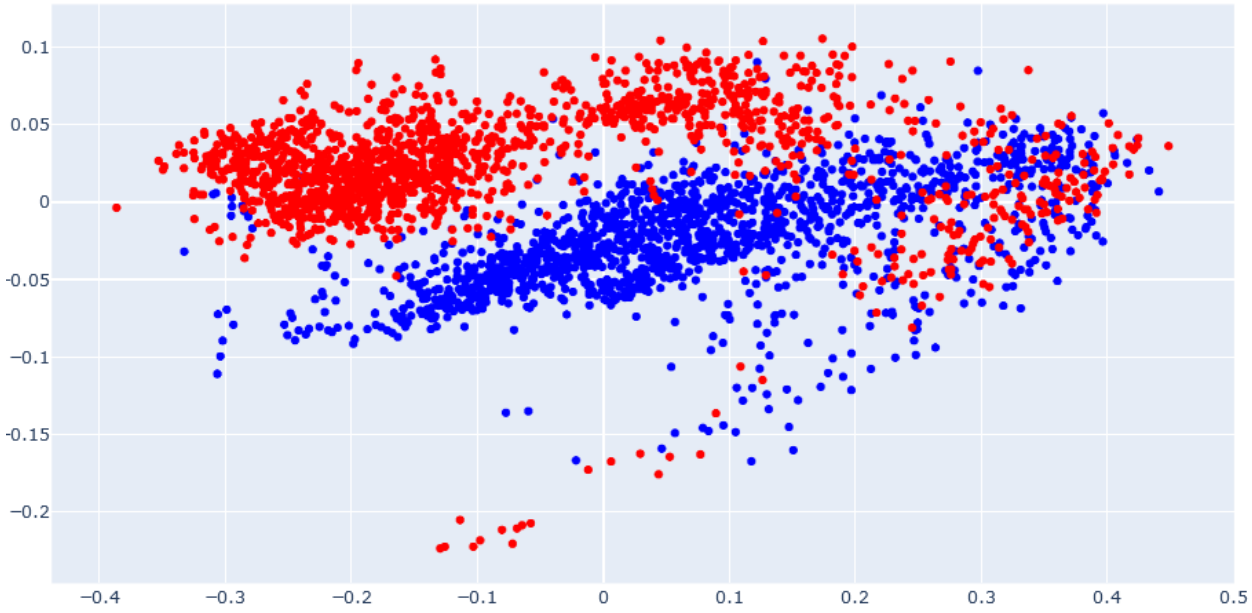


Рис.22 Применение PCA для голосов.

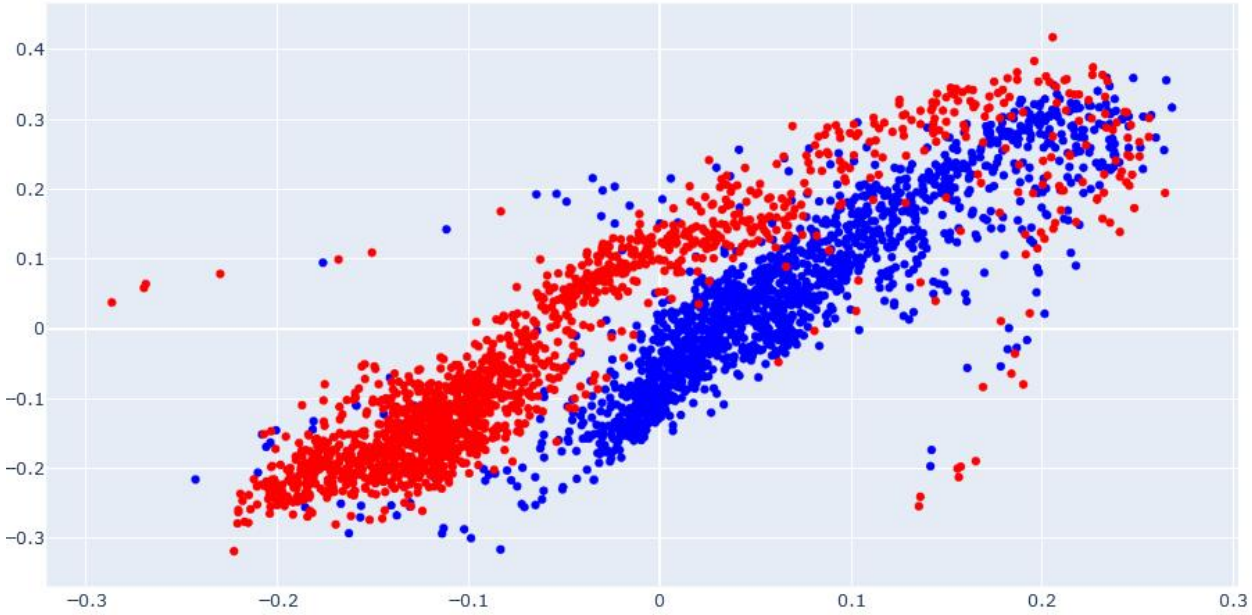


Рис.22 Применение Стресс-Функции для голосов.

Если основываться на результатах, полученных на ирисах Фишера, то Стресс-Функция позволяет сохранить приближенное положение точек в пространстве с начальным расположением. Однако в плане скорости вычислений PCA намного быстрее Стресс-Функции. Для датасета ирисов разница между ними не заметна, но для 3500 элементов скорость вычислений для PCA заняло долю секунды, а для стресс-функции в районе пары минут.

Заключение

В ходе работы был реализован и применен Метод Главных Компонент (PCA), а также была использована стресс-функция для уменьшения размерности пространства признаков. Было проведено сравнение двух методов. Так же был применен алгоритм кластеризации k-means для сравнения кластеров полученных на основе обработанных данных.

В работе были выполнены следующие задачи:

1. Оценка и выбор метода уменьшения размерности.
2. Программная реализация метода Главных Компонент.
3. Применение метода Главных Компонент.
4. Применение стресс функции.
5. Анализ полученных результатов.

Полученные знания позволяют в дальнейшем выбирать наилучший метод редукции данных, а также получить оптимальную размерность удобную для тех или иных задач.

Список литературы

1. Интуит, электронный курс по Data Mining, URL: https://intuit.ru/studies/courses6/6/info
2. Science Hunter, Две типичные задачи Data Mining, URL: https://www.sciencehunter.net/Blog/story/dve-tipichnyie-zadachi-data-mining
3. Анализ главных компонент, URL: https://www.helenkapatsa.ru/mietod-ghlavnykh-komponient/?
4. Ковариация, URL: https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%B2%D0%B0%D1%80%D0%B8%D0%B0%D1%86%D0%B8%D1%8F
5. Ковариационная матрица, URL: https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%B2%D0%B0%D1%80%D0%B8%D0%B0%D1%86%D0%B8%D0%BE%D0%BD%D0%BD%D0%B0%D1%8F_%D0%BC%D0%B0%D1%82%D1%80%D0%B8%D1%86%D0%B0
6. Лекция: Метод главных компонент, URL: http://math-info.hse.ru/f/2015-16/ling-mag-quant/lecture-pca.html
7. Jacobi eigenvalue algorithm, URL: https://en.wikipedia.org/wiki/Jacobi_eigenvalue_algorithm
8. Rand index, URL: https://en.wikipedia.org/wiki/Rand_index
9. Multidimensional scaling, URL: https://en.wikipedia.org/wiki/Multidimensional_scaling
10. Метод k-средних, URL: https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_k-%D1%81%D1%80%D0%B5%D0%B4%D0%BD%D0%B8%D1%85