ПРИМЕНЕНИЕ АЛГОРИТМА ISODATA В ЗАДАЧАХ КЛАССИФИКАЦИИ ОБЪЕКТОВ

Студент: Мельничук В.С

Группа: ФН12-51Б

Руководитель работы: Панкратов В.А

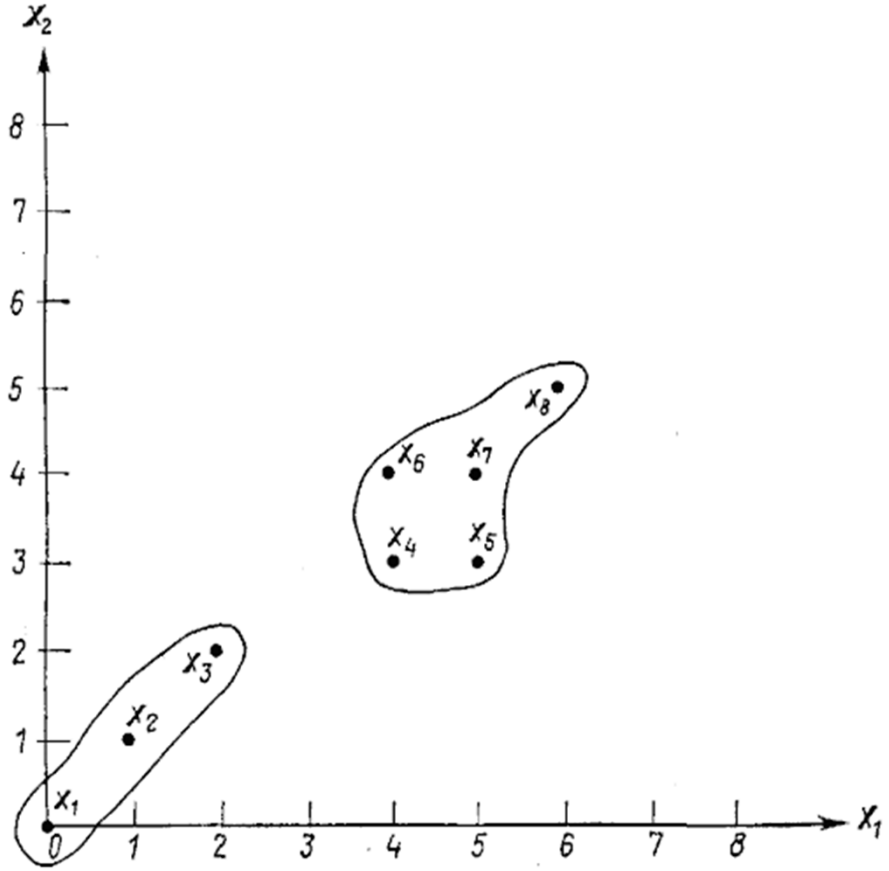
Цели и задачи курсовой работы

Цель:

- изучение алгоритма ISODATA и его применение на практике

Задачи:

- разработка математического обеспечения для реализации алгоритма ISODATA
- выполнение математического моделирования



Разбиение набора данных из 8 точек на 2 кластера с помощью алгоритма ISODATA (ИСОМАД)

Алгоритм ISODATA

Алгоритм ISODATA аналогичен процедуре, предусматривающей вычисление K внутригрупповых средних (K-means), поскольку и в этом алгоритме центрами кластеров служат выборочные средние определяемые итеративно.

В отличие от K-means ISODATA обладает набором вспомогательных эвристических процедур, которые вошли в алгоритм в результате осмысления эмпирического опыта его использования.

Алгоритм выполняется итеративно, каждую итерацию можно разбить на 14 шагов.

Одна итерация алгоритма ISODATA

Шаг 1. Определение глобальных гиперпараметров кластеризации:

K - необходимое (желаемое) число кластеров,

THETA_M - минимальное количество образов в одном кластере,

THETA_S - параметр, характеризующий среднеквадратичное отклонение,

THETA_C - параметр, характеризующий компактность,

I - максимальное количество итераций,

P - максимальное количество кластеров, которые можно объединить за один раз,

THETA_O - eps-изменение для продолжения алгоритма,

Шаг 1.1. Задание N начальных образов.

Одна итерация алгоритма ISODATA

Шаг 2. Заданные N образов распределяются по кластерам, соответствующим выбранным исходным центрам по правилу наименьших расстояний.

Шаг 3. Ликвидируются подмножества образов, в состав которых входит менее порогового числа элементов.

Шаг 4. Каждый центр кластера, локализуется и корректируется посредством приравнивания его выборочному среднему, найденному по соответствующему подмножеству.

Шаг 5. Вычисляется среднее расстояние между объектами, входящими в подмножество и соответствующим центром кластера по заданной формуле.

Шаг 6. Вычисляется обобщенное среднее расстояние между объектами, находящимися в отдельных кластерах, и соответствующими центрами кластеров.

Одна итерация алгоритма ISODATA

Шаг 7. (а) Если текущий цикл итерации последний, то переход к шагу 11.

(б) Иначе переход к шагу 8.

Шаг 8. Для каждого подмножества выборочных образов вычисляется вектор среднеквадратичного отклонения.

Шаг 9. В каждом векторе среднеквадратичного отклонения отыскивается максимальная компонента.

Шаг 10. Расщепление кластера на два новых кластера при нарушении условия компактности кластера.

Шаг 11. Вычисляются попарные расстояния между всеми парами

центров кластеров.

Шаг 12. Сортировка пар кластеров по попарному расстоянию.

Одна итерация алгоритма ISODATA

Шаг 13. Слияние кластеров при нарушении условия компактности кластеров.

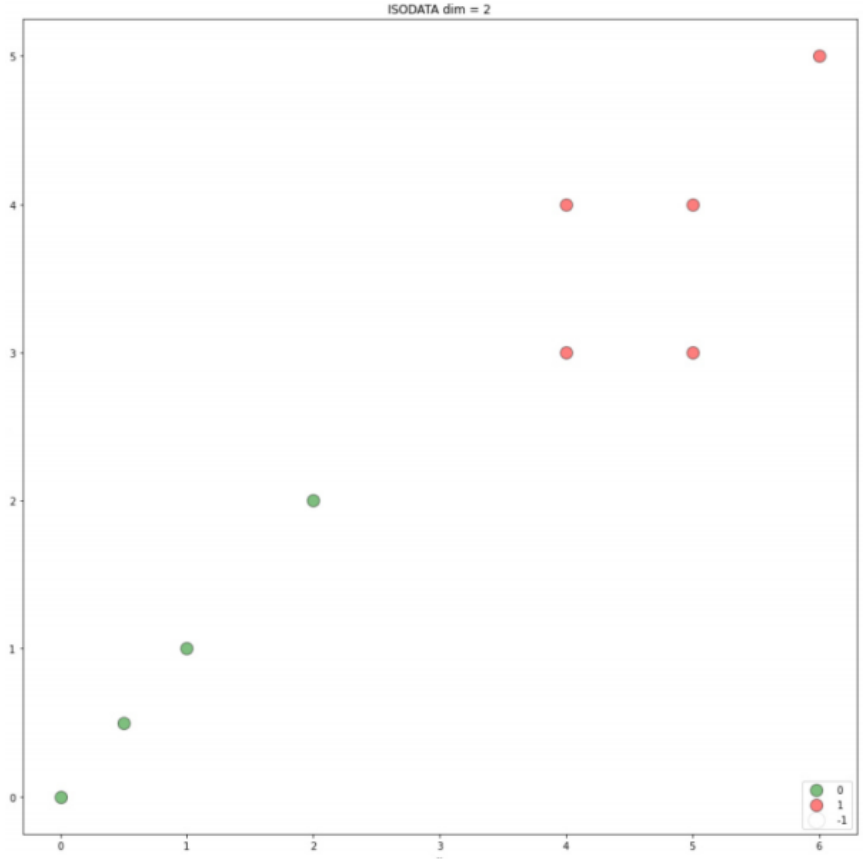
Шаг 14. Если текущий цикл итерации последний, то выполнение алгоритма прекращается. В противном случае следует возвратиться либо к шагу 1, если по предписанию пользователя меняется какой-либо из параметров, определяющих процесс кластеризации, либо к шагу 2, если в очередном цикле итерации параметры процесса должны остаться неизменными.

Завершением цикла итерации считается каждый переход к шагам 1 или 2.

Демонстрация работы ISODATA

Проведена реализация алгоритма ISODATA на языке Python 3.X. Выполнение кода и графическая иллюстрация происходила в облачной среде google.colab.

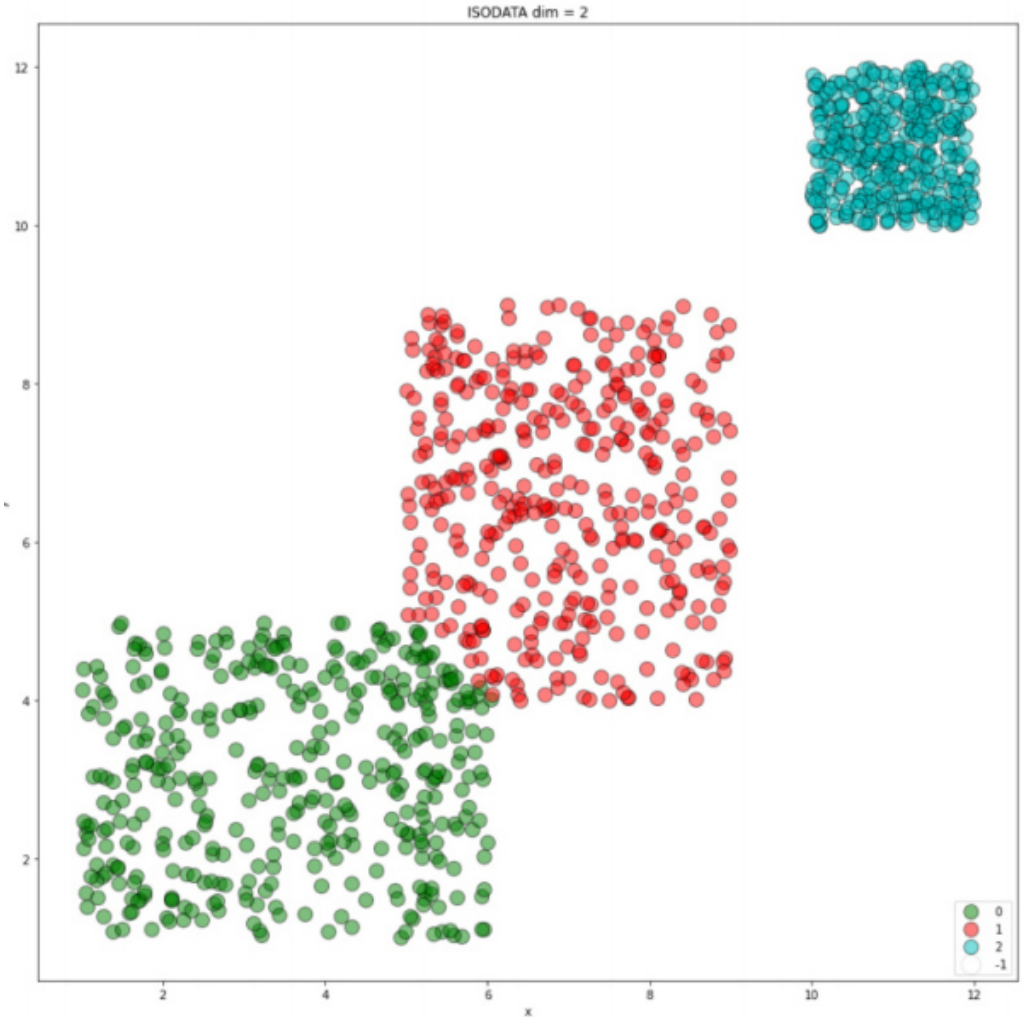
Алгоритм был реализован на основе его описания в книге «Принципы распознавания образов» Дж. Ту, Р. Гонсалес.



Разбиение набора данных из 9 точек на 2 кластера с помощью алгоритма ISODATA, реализованного на python 3.X

Демонстрация работы ISODATA

Образы представляют собой точки на плоскости, которые естественным образом разделяются на три квадратных кластера. Выборка состоит из 1200 точек, которые поровну разделены между тремя кластерами. Данные из одного кластера равномерно распределены в нем.

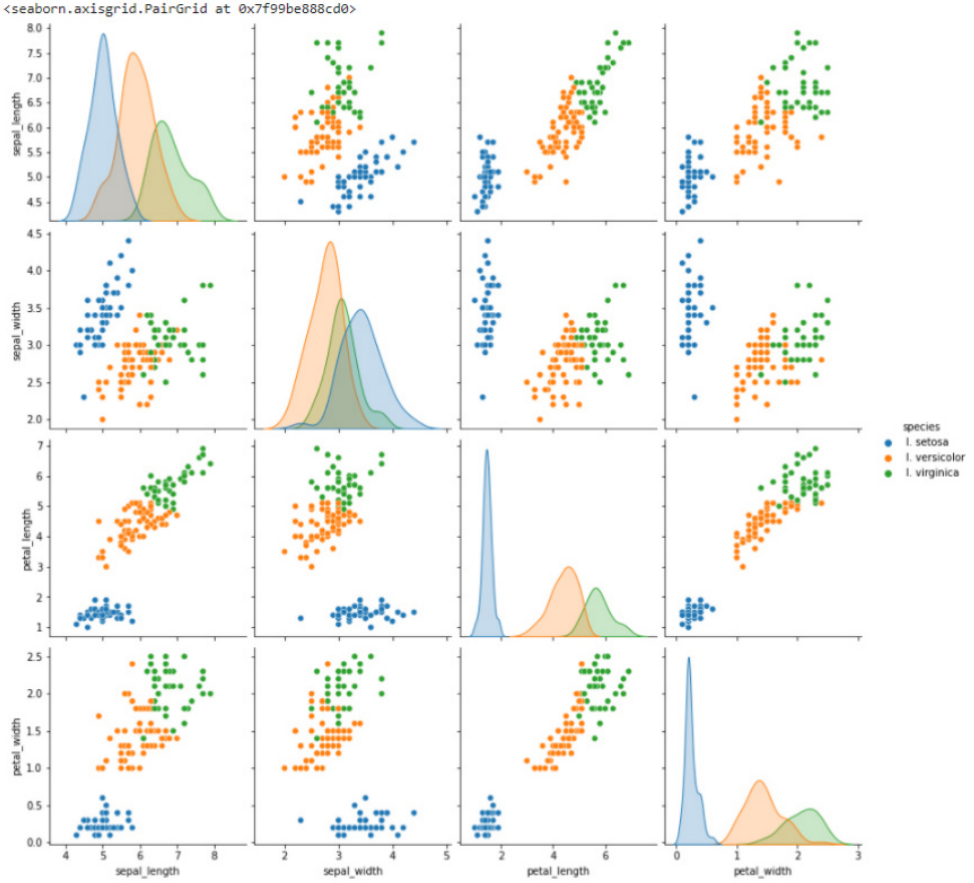


Демонстрация работы ISODATA

Датасет IRIS состоит из 150 образов. Образы представляют собой вектора из 4х элементов (4хмерное пространство) . Один из классов (setosa) линейно-разделим от двух остальных классов. Это хорошо видно на графиках, на которых по двум осям берутся 2 из 4х элементов 4хмерного пространства. На каждом из 12 попарных графиков класс setosa (обозначается на графике синим цветом) линейно-разделим и как следствие сильно обособлен относительно двух других(см. рис.).

Алгоритм ISODATA показал точность разбиения 89.33%.

Isodata




```
1 accuracy_v2(img_class, y)
```

```
Accuracy = 89.33
```

Демонстрация работы ISODATA

Эталонное распределение для датасета IRIS. Метки образов взяты из датасета IRIS, без использования алгоритма ISODATA.

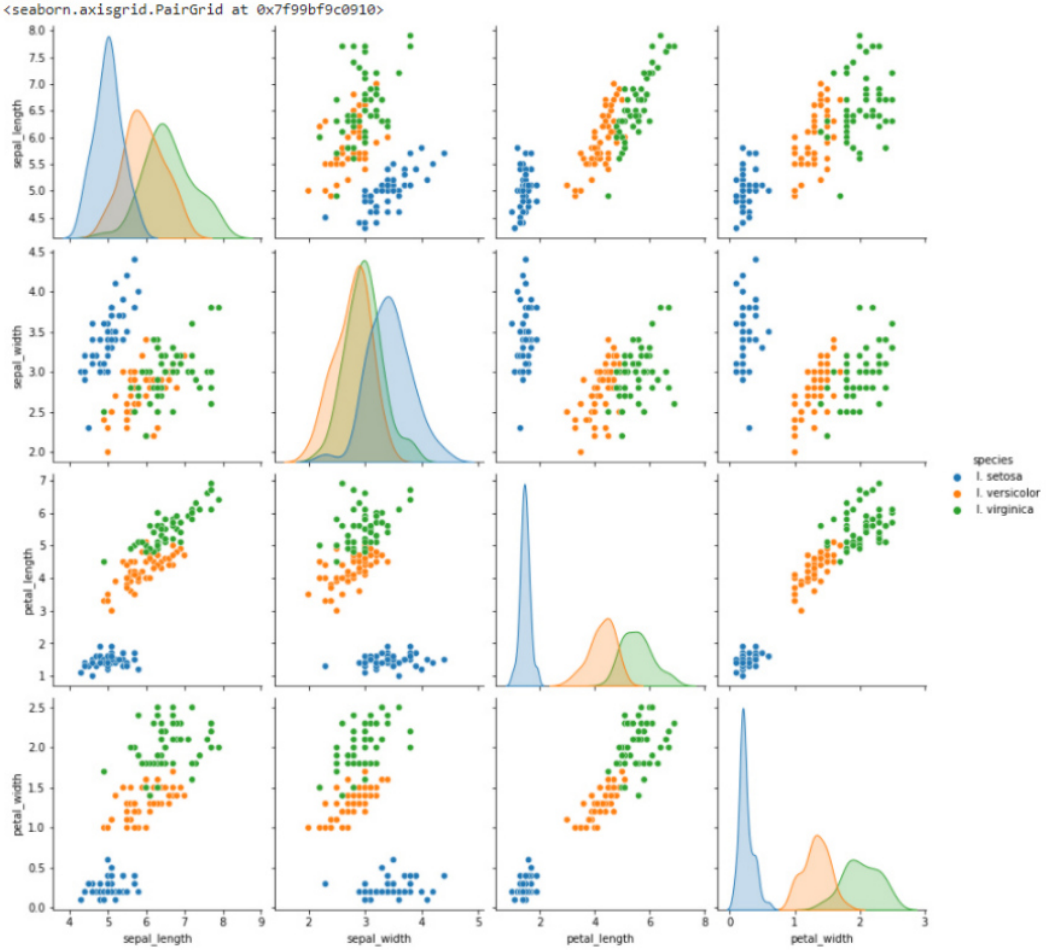
Для сравнения качества работы алгоритма ISODATA представлена НС (нейронная сеть с полносвязными слоями). НС показала точность разбиения на датасете IRIS - 93.33%. Сеть была обучена на половине датасета (75 образов).

```
Epoch 48/50
5/5 [==============================] - 0s 24ms/step - loss: 0.2971 - accuracy: 0.9254 - val_loss: 0.2245 - val_accuracy: 1.0000
Epoch 49/50
5/5 [==============================] - 0s 21ms/step - loss: 0.3078 - accuracy: 0.9403 - val_loss: 0.2175 - val_accuracy: 1.0000
Epoch 50/50
5/5 [==============================] - 0s 18ms/step - loss: 0.3210 - accuracy: 0.9254 - val_loss: 0.2154 - val_accuracy: 1.0000
<keras.callbacks.History at 0x7f994ab158d0>
```

```
[ ]  1 scores = model.evaluate(x_test,
     2                         y_test,
     3                         verbose=1
     4                         )
```

```
3/3 [==============================] - 0s 5ms/step - loss: 0.2781 - accuracy: 0.9333
```

Etalon
<seaborn.axisgrid.PairGrid at 0x7f99bf9c0910>



Демонстрация работы ISODATA

Датасет DIGITS состоит из 1796 образов. Образы представляют собой вектора из 64х элементов (64хмерное пространство). Датасет состоит из чернобелых картинок размерностью 8 на 8, значит в датасете расположены образы из 64хмерного пространства. Образы естественным образом разделяются на 10 кластеров (10 классов – числа от 0 до 9).

Алгоритм ISODATA показал точность разбиения 70.18%.

```
[ ]   1 accuracy_v2(img_class, y)

      Accuracy = 70.18
```

Для сравнения качества работы алгоритма ISODATA представлена НС (нейронная сеть с полносвязными слоями). НС показала точность разбиения на датасете DIGITS - 97.33%. Сеть была обучена на половине датасета (75 образов).

```
Epoch 48/50
26/26 [==============================] - 0s 12ms/step - loss: 0.0193 - accuracy: 0.9988 - val_loss: 0.1638 - val_accuracy: 0.9778
Epoch 49/50
26/26 [==============================] - 0s 15ms/step - loss: 0.0227 - accuracy: 0.9950 - val_loss: 0.1540 - val_accuracy: 0.9778
Epoch 50/50
26/26 [==============================] - 0s 14ms/step - loss: 0.0214 - accuracy: 0.9988 - val_loss: 0.1516 - val_accuracy: 0.9778
<keras.callbacks.History at 0x7f994a998750>
```

```
1 scores = model.evaluate(x_test,
2                         y_test,
3                         verbose=1
4                         )
```

```
29/29 [==============================] - 0s 5ms/step - loss: 0.0874 - accuracy: 0.9733
```

Выводы

В работе была проведена реализация алгоритма ISODATA и его сравнение с полносвязной НС. На двух датасетах IRIS и DIGITS была продемонстрирована лучшая точностью у НС.

Преимущество алгоритма ISODATA в исполнении без учителя, в то время как полносвязная НС обучалась с учителем.

Алгоритм был реализован на Python 3.x.

Спасибо за внимание