В связи с ограниченными сроками выполнения задания в качестве метода классификации выбран метод, демонстрирующий лучшие оценки качества в [https://github.com/ksopyla/svm\_mnist\_digit\_classification]. Оценки были получены с использованием перекрестной проверки (cross-validation) и решетчатого поиска (grid search). Таким образом, для решения задания используется нелинейный SVM с гауссовым ядром RBF (значения гиперпараметров gamma=0.05 и C=5).

Для многоклассовой классификации в scikit-learn реализованы две стратегии: «один против всех»(OVR) и «один против одного»(OVO). Для задачи классификации MNIST в случае выбора второй стратегии «один против одного» обучаются 45 двоичных классификаторов. Главное преимущество OVO при применении к многоклассовому SVM заключается в том, что 45 классификаторов последовательно обучаются на ограниченной подвыборке. Таким образом, сокращается время, необходимое для обучения на большом наборе данных, в сравнении с OVR.

Значение коэффициента корректных прогнозов (accuracy) при применении к тестовой выборке классификатора SVM, обученного на наборе изображений без предварительной обработки, представлено в таблице 1. Нормированная матрица неточностей представлена на рисунке 1. В результате применения обученного классификатора с заданными гиперпараметрами все объекты тестовой выборки отнесены к 9-ому классу.

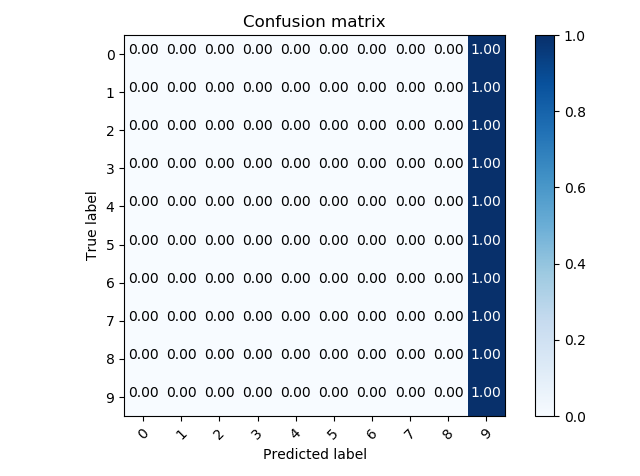


Рисунок 1 – Нормированная матрица неточностей для алгоритма SVM без предварительной обработки изображений

Нелинейные методы SVM чувствительны к масштабам признаков. Масштабирование можно выполнить следующими способами:

1) масштабирование признаков по минимаксу – приведение к диапазону (0, 1);

2) стандартизация признаков – приведение распределения признака к нулевому математическому ожиданию и единичной дисперсии;

3) стандартизация интенсивности отдельного изображения – приведение распределения интенсивности изображения к нулевому математическому ожиданию и единичной дисперсии.

Значения коэффициента корректных прогнозов, полученные в результате применения классификатора к тестовой выборке, для разных способов масштабирования данных также представлены в таблице 1. Нормированные матрицы неточностей представлены на рисунке 2.

Таблица 1 – Значения коэффициентов корректных прогнозов для разных способов предварительной обработки изображений

|  |  |  |
| --- | --- | --- |
|  | Коэффициент корректных прогнозов | Коэффициент корректных прогнозов с использованием предварительной геометрической нормализации изображения |
| Без предварительной обработки | 0,10 | - |
| Масштабирование по минимаксу | 0,98 | 0,99 |
| Стандартизация признаков | 0,93 | 0,95 |
| Стандартизация интенсивности | 0,89 | 0,92 |

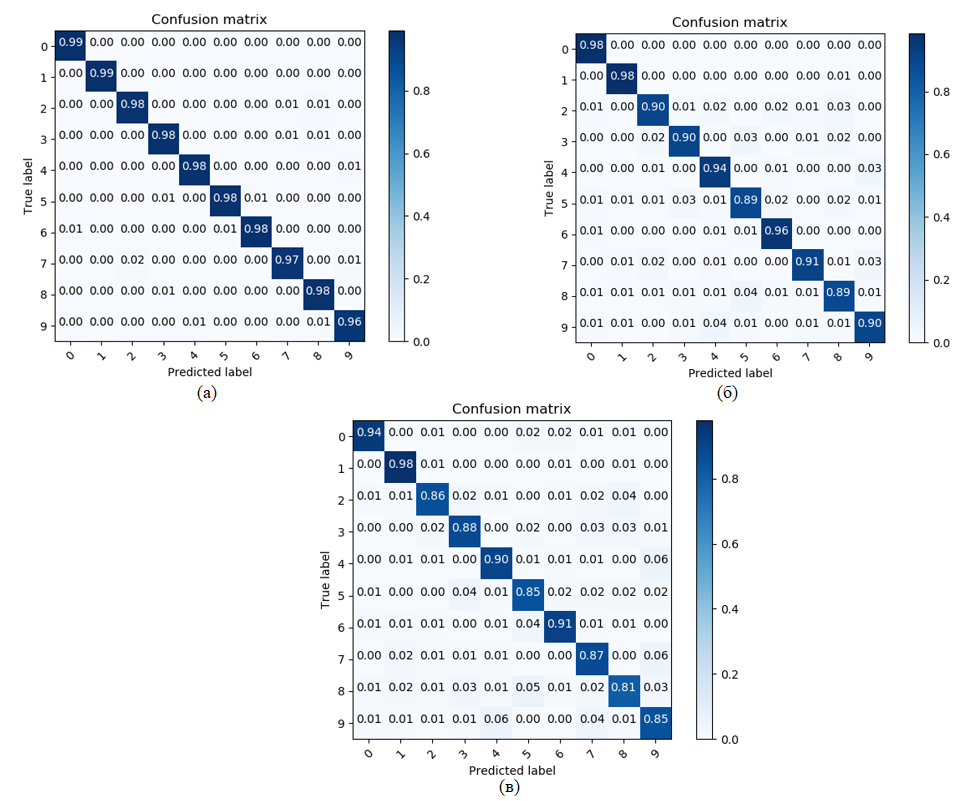


Рисунок 2 – Нормированные матрицы неточностей на тестовой выборке для разных способов масштабирования признаков (а – масштабирование по минимаксу, б – стандартизация признаков, в – стандартизация интенсивности отдельного изображения)

Лучший результат классификации по критерию наибольшего значения коэффициента корректных прогнозов достигнут в результате масштабирования признаков по минимаксу.

В результате анализа нормированной матрицы неточностей, соответствующей классификатору с наибольшим коэффициентом корректных прогнозов на тестовой выборке, можно сделать следующий вывод: наибольшая ошибка первого рода возникает при неверном отнесении объектов 7-го класса ко множеству объектов 2-го класса. Для оставшихся классификаторов наибольшая частота взаимного перепутывания достигается для объектов 4-го и 9-го классов, 3-его и 5-го.

Проведем анализ статистического распределения пикселей изображений отдельных классов. На рисунке 3 представлены математические ожидания пикселей по классам, на рисунке 4 представлены дисперсии пикселей по классам. Из визуального анализа характеристик, представленных на рисунках 3 и 4, следует вывод о высокой вариативности экземпляров внутри классов.

Применим метод геометрической нормализации изображения, устраняющий наклоны и перекосы символов. Для этого выбран алгоритм, основанный на моментах изображения.

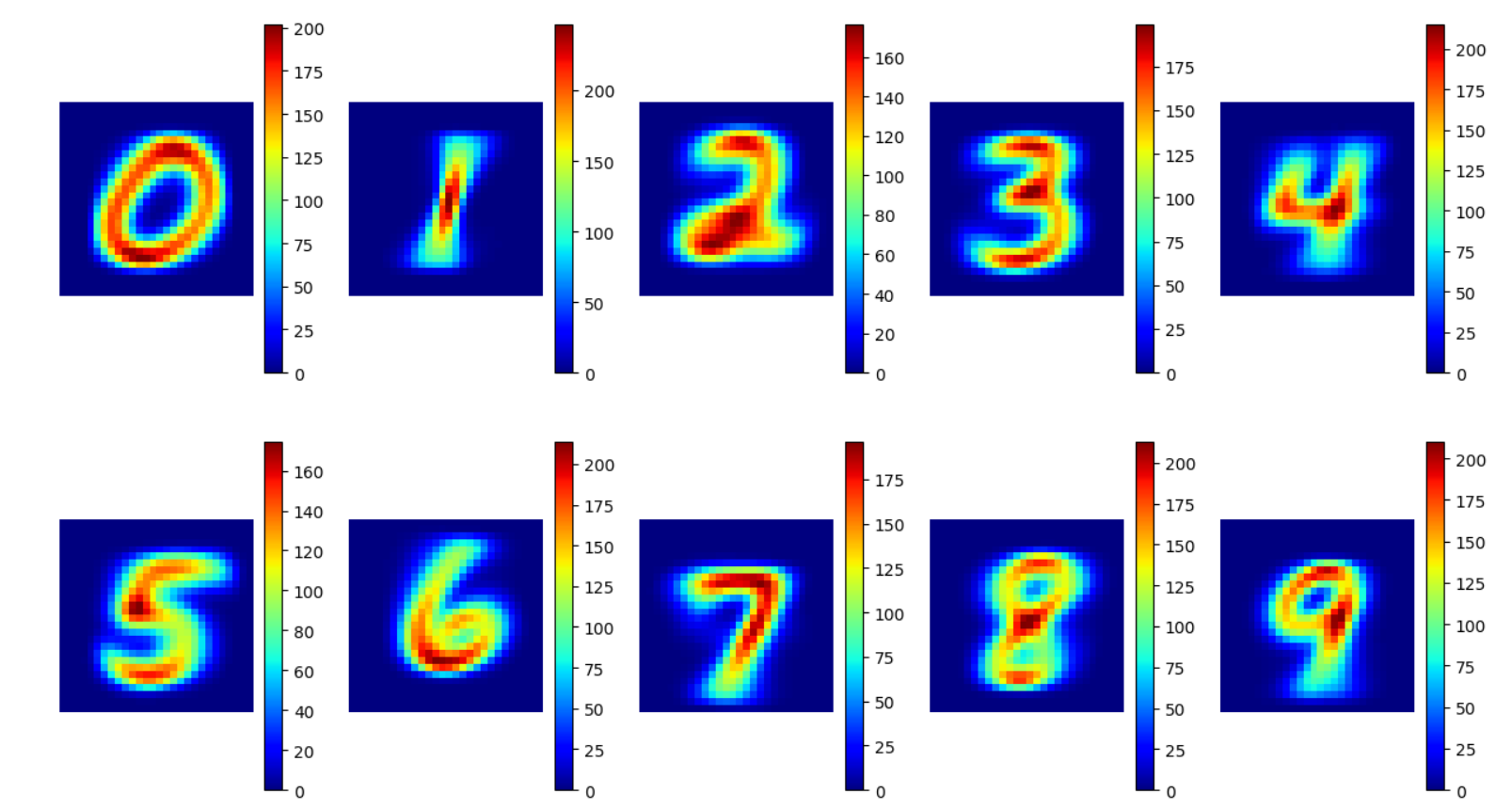


Рисунок 3 – Математическое ожидание пикселей классов

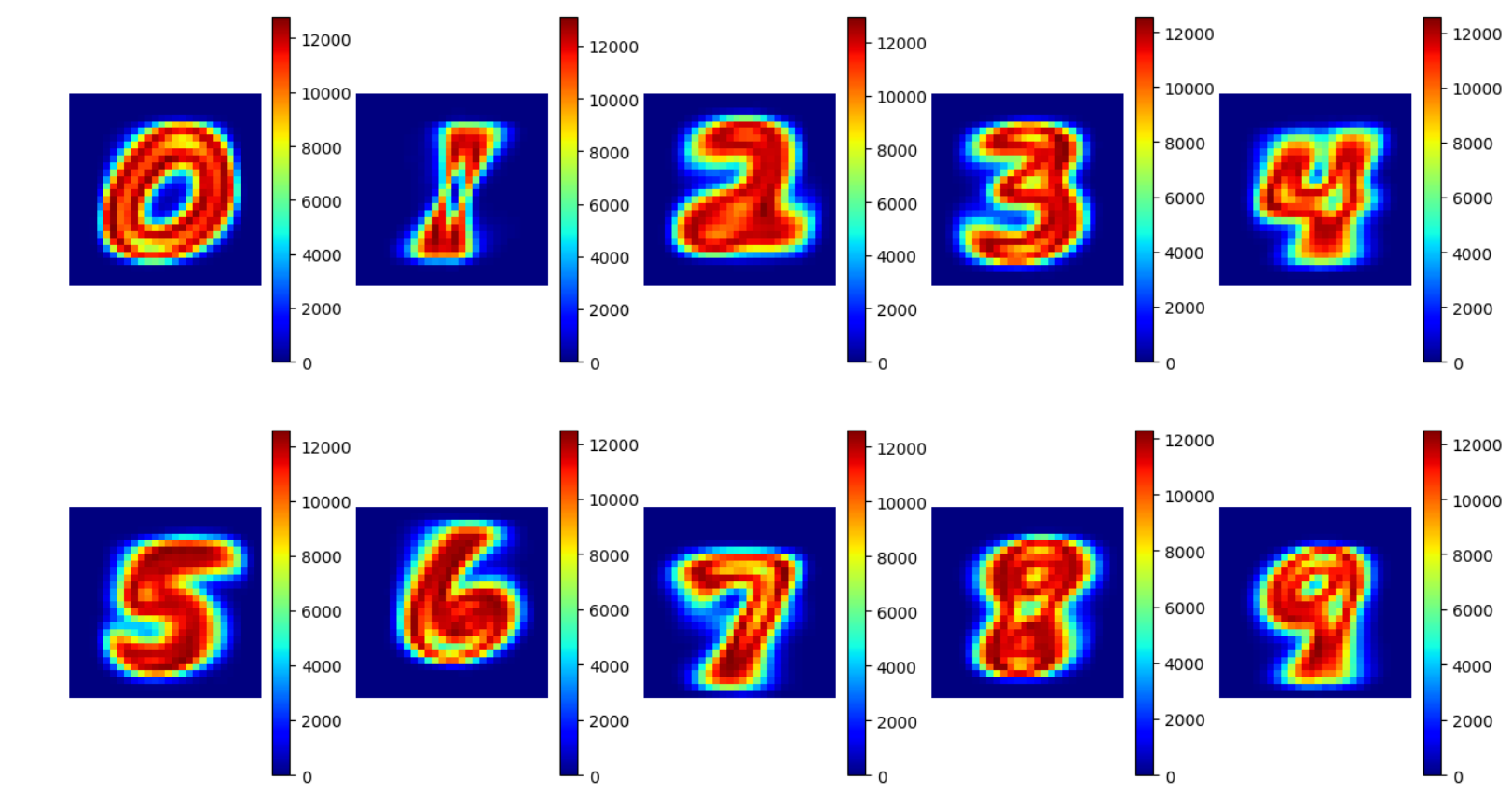


Рисунок 4 – Дисперсия пикселей классов

Результат работы алгоритма предварительной обработки для отдельных экземпляров изображений разных классов представлен на рисунке 5.

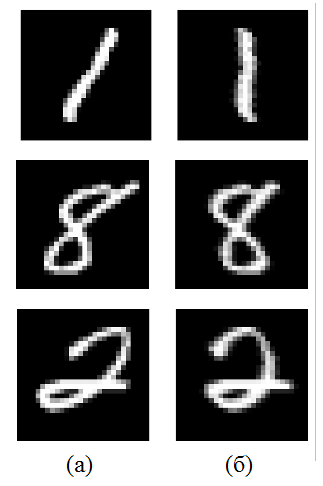


Рисунок 5 — Результат работы алгоритма предварительной обработки изображений (a – столбец с исходными изображениями, б – столбец с обработанными изображениями)

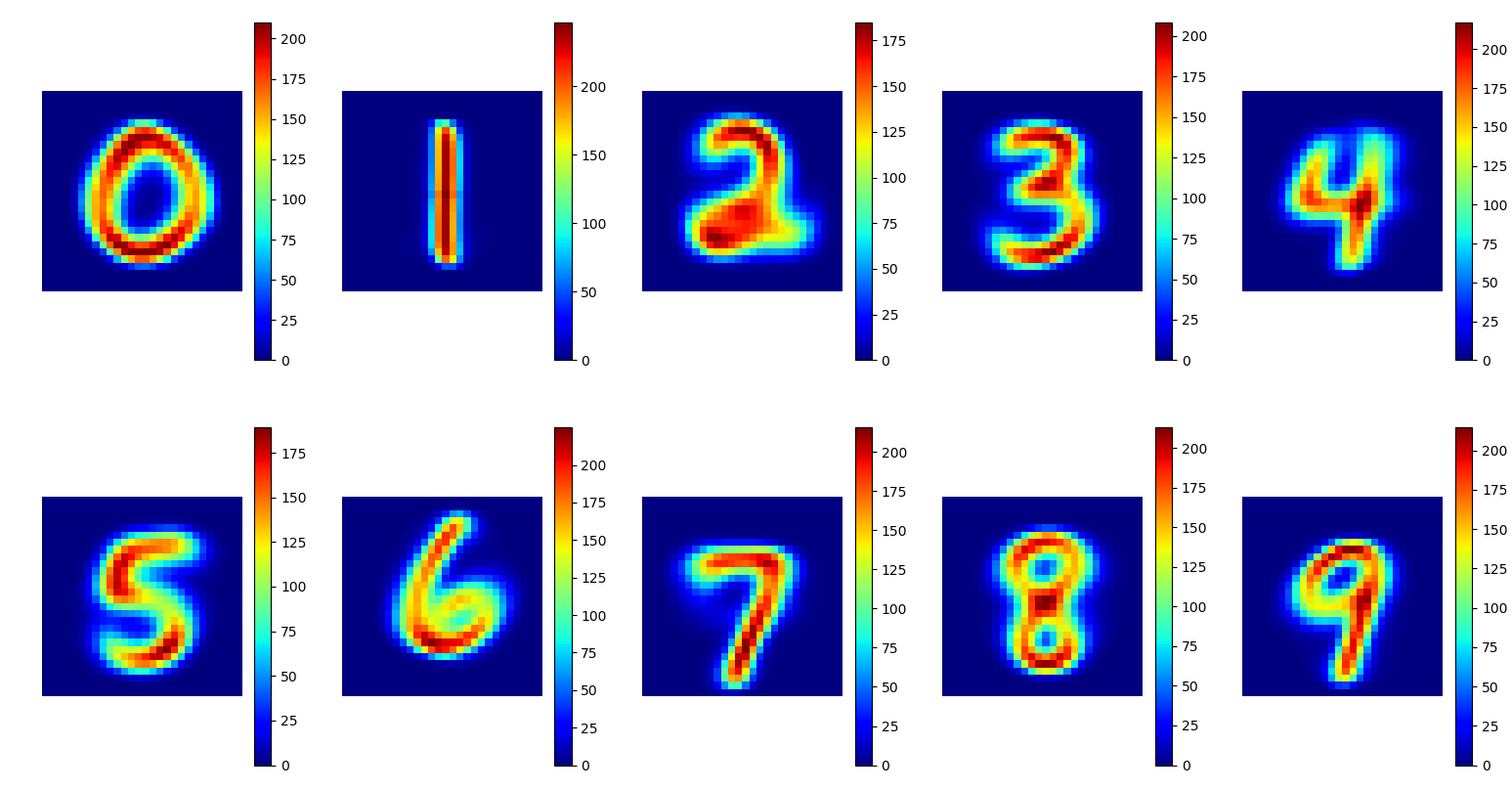


Рисунок 6 – Математическое ожидание пикселей классов после применения метода предварительной обработки

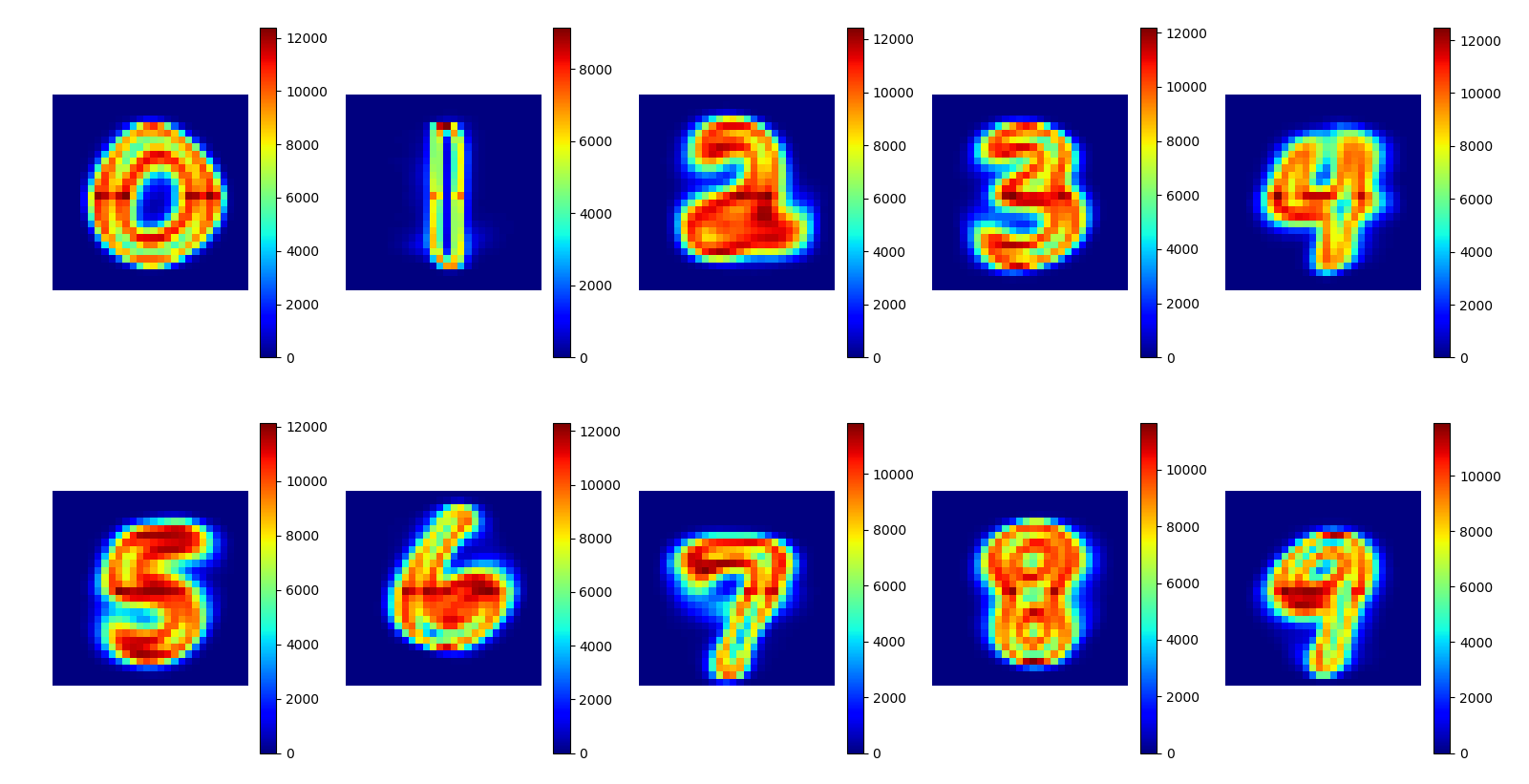


Рисунок 7 – Дисперсия пикселей классов после применения метода предварительной обработки

Из сравнительного анализа математических ожиданий и дисперсий пикселей изображений внутри классов до (рисунки 3, 4) и после (рисунки 6, 7) предварительной обработки следует вывод о снижении внутриклассовой дисперсии пикселей и выделении информативных признаков с помощью предложенного метода предварительной обработки изображений.

Итоговые оценки качества работы классификатора с применением выбранного алгоритма предварительной обработки изображений и последующими применением масштабирования признаков (для рассматриваемых способов масштабирования) представлены во втором столбце таблицы 1.

В результате применения метода предварительной обработки изображений (устранения наклонов цифр на изображениях) оценка качества работы классификатора увеличилась для всех способов масштабирования данных.

В скрипте main.py реализована предобработка с последующим масштабированием признаков по минимаксу.