УДК 519.87

Г. Д. Гефан [1], В. С. Попова [1], Н. С. Попова [1]

[1] Иркутский государственный университет путей сообщения, г. Иркутск, Российская Федерация

МЕТОД РАЗЛИЧЕНИЯ СХОДНЫХ ПОЧЕРКОВ С ПОМОЩЬЮ ЛИНЕЙНОГО КЛАССИФИКАТОРА

Аннотация. Метод простого линейного классификатора (ПЛК) предложено использовать для решения задачи различения сходных почерков. Почерк – уникальное свойство каждого человека, по которому возможна идентификация личности. Однако различить сходные почерки непросто, особенно при большом объеме данных. Результаты экспериментов, приведённые в работе, показали, что метод ПЛК позволяет достичь достаточно высокого уровня точности и надёжности при классификации графических образцов и может быть эффективным инструментом для различения сходных почерков.

Ключевые слова: задачи классификации, метод опорных векторов, линейный классификатор, различение сходных почерков, моделирование в Python.

G.D.Gefan [1], V.S. Popova [1], N.S.Popova [1]

[1] Irkutsk State Transport University, Irkutsk, Russian Federation

METHOD OF RECOGNIZING SIMILAR HANDWRITING USING A LINEAR CLASSIFIER

Abstract. The method of a simple linear classifier (SLC) is proposed to be used to solve the problem of recognizing similar handwritings. Handwriting is unique to each person, it can be used to identify a person. However, to recognize similar handwritings is not easy, especially with a large amount of data. The experimental results presented in this paper show that the SLC method can achieve a sufficiently high level of accuracy and reliability in the classification of graphical samples and can be an effective tool for recognizing similar handwritings.

Keywords: classification problems, Support Vector Machines, linear classifier, recognition of similar handwritings, modeling in Python.

Введение. На протяжении многих лет машинное обучение использовалось для решения множества задач в таких областях, как медицинская диагностика, техническая диагностика (компьютерное зрение, распознавание речи), экономика (кредитный скоринг, обнаружение мошенничества, биржевой анализ), офисная автоматизация (распознавание текста или рукописного ввода, обнаружение спама, категоризация документов) [1]. В общих чертах выделяют задачи классификации, кластеризации и регрессии.

Задачи классификации и их практическое значение. **Методы классификации**. Задачи классификации возникают в случае, когда необходимо присвоить объекту метку принадлежности к определённому классу на основе различных характеристик, называемых признаками [2, 3]. Классификация объектов находит широкое применение в системах безопасности, в управлении и контроле доступа, в системах по распознаванию знаков, человеческих лиц.

Существует немало методов классификации. К наиболее распространенным методам относятся: деревья решений, метод k-ближайших соседей, наивный байесовский классификатор, логистическая регрессия и метод опорных векторов. Последний метод рассмотрим чуть подробнее.

Основная идея метода опорных векторов (Support Vector Machine, SVM) состоит в построении линии (или гиперплоскости), оптимально разделяющей объекты выборки на два класса [4]. Необходимо разделить множества некоторой полосой так, чтобы, во-первых, эта полоса была как можно шире (для лучшего разделения двух классов), и, во-вторых, чтобы были минимизированы ошибки разделения [5-7]. Перечисленные оптимизационные требования (максимизации ширины полосы и минимизация ошибок) противоречат друг другу, и

критерий оптимизации должен быть сконструирован так, чтобы можно было регулировать их относительную важность. Векторы, оказывающиеся на границах разделительной полосы, называются опорными.

Описание простого линейного классификатора (ПЛК), его преимущества. Несмотря на свою популярность, SVM обладает рядом недостатков. Во-первых, он неустойчив по отношению к «шуму» (ошибочным точкам) в исходных данных. Если обучающая выборка содержит шумовые выбросы («объекты-нарушители»), то они будут существенным образом учтены при построении разделяющей гиперплоскости [8]. Во-вторых, есть регулирующий параметр алгоритма C, который надо подбирать [9, 10]. Более того, число переменных при решении задачи оптимизации равно числу обучающих векторов, что приводит к замедлению процесса обучения.

Алгоритм ПЛК следующий. Пусть имеется 2 класса тренинговых (обучающих) векторов $\mathbf{x_i}$. Присваиваем им метки z_i (одному классу +1, другому -1). Задаем условие оптимизации (1):

$$\sum_{i=1}^{n}(\mathbf{x}_i \cdot \mathbf{w} - b)z_i \to \max. \quad (1)$$

Здесь величина b – неизвестное расстояние от начала координат до границы (2):

$$b = \tfrac{1}{2}\left[\overline{\mathbf{x}\cdot\mathbf{w}}\Big|_{(1)} + \overline{\mathbf{x}\cdot\mathbf{w}}\Big|_{(2)}\right], \quad (2)$$

где слагаемые в скобках соответствуют усреднённым скалярным произведениям векторов одного и другого классов на неизвестный нормальный вектор разделительной гиперплоскости $\mathbf{w}$. Задаем ограничение вида $|\mathbf{w}| = 1$. Находим оптимальный нормальный вектор $\mathbf{w}^*$ и величину b^* как решение сформулированной задачи оптимизации.

Для классификации нового вектора $\mathbf{x}$ находим $\mathbf{x}\cdot\mathbf{w}^*$. Если величина $\mathbf{x}\cdot\mathbf{w}^* - b^* < 0$, то относим этот вектор к классу $z_i = -1$, иначе – к классу $z_i = 1$ [11].

Следовательно, преимущество данного метода состоит в том, что задача не зависит от какого-либо дополнительного параметра в отличие от метода опорных векторов, в котором необходимо задавать управляющий параметр C. Число переменных при решении задачи оптимизации равно размерности векторного пространства и не зависит от числа тренинговых векторов, что кардинально сокращает требуемые ресурсы.

Тестирование ПЛК на случайных векторах (двумерное нормальное распределение). Для тестирования ПЛК используем набор случайных двумерных векторов, координаты которых имеют нормальное распределение. Через a и σ обозначим математическое ожидание и среднеквадратическое отклонение (СКО) нормального распределения соответственно (таблица 1).

Таблица 1.

Данные для тестирования ПЛК

	1-ый класс		2-ой класс	
	X_1	X_2	X_1	X_2
a	3	5	5	7
σ	1	1	1	1

Решая задачу оптимизации, найдем параметры границы. Также посмотрим, сколько «ошибок» возникает, когда вектор одного класса опознаётся как вектор другого класса. Для этого возьмем 25 тренинговых векторов каждого класса. На графике (рис. 1) показаны данные двух классов, обозначенных красными и синими кружками, вместе с границей решения (линия

черного цвета), найденной методом ПЛК. Из рисунка видно, что в результате работы классификатора количество «ошибок» равно 4 (2 + 2).

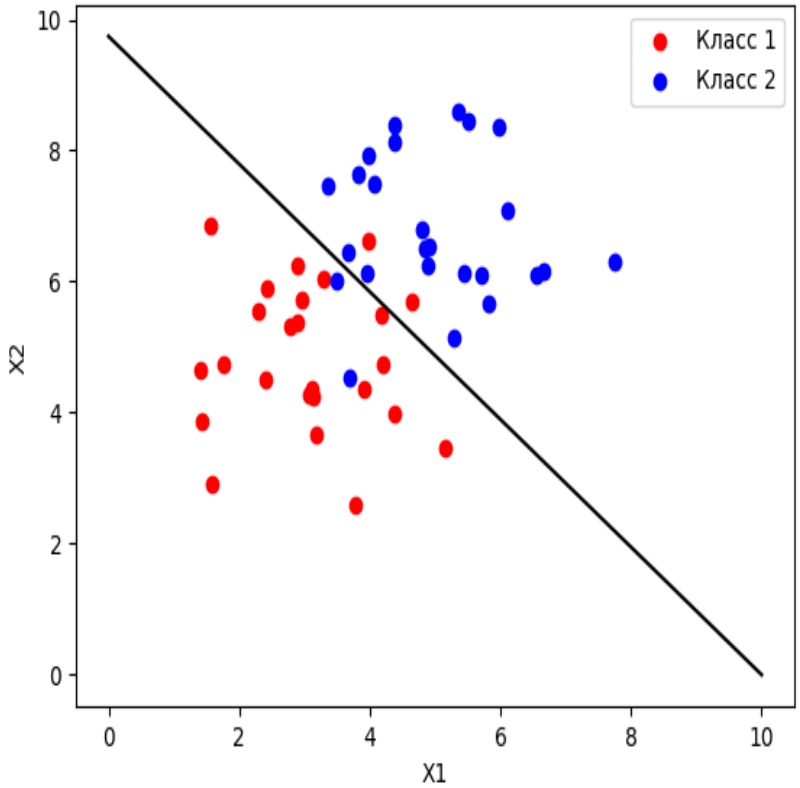

Рис. 1. Результат эксперимента со случайными векторами (двумерное нормальное распределение)

В таблицу 2 сведены результаты двадцати таких экспериментов, средние значения параметров границы по всей серии, а также теоретические параметры границы, которые в этом случае из соображений симметрии легко получить: $x_{(1)} + x_{(2)} = 10$ или $\mathbf{w}^* = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$, $b^* = \frac{10}{\sqrt{2}}$.

Таблица 2.

Результаты экспериментов

№	b^*	$\mathbf{w}^*$	
1	7,026	0,712	0,702
2	7,302	0,653	0,757
3	6,877	0,780	0,626
4	6,974	0,698	0,716
5	6,864	0,828	0,560
6	7,261	0,595	0,804
7	6,763	0,732	0,681
8	7,158	0,746	0,666
9	7,262	0,621	0,783
10	6,516	0,845	0,535
11	7,225	0,676	0,737
12	6,835	0,788	0,615
13	7,041	0,730	0,683

14	7,110	0,645	0,764
15	7,089	0,683	0,730
16	7,098	0,730	0,683
17	6,785	0,753	0,658
18	7,150	0,709	0,705
19	6,937	0,667	0,745
20	7,338	0,628	0,778
Среднее	7,031	0,711	0,696
Теоретическое	7,071	0,707	0,707

Таким образом, убеждаемся, что полученные оценки нормального вектора и расстояния до границы от начала координат не смещены, то есть колеблются вокруг теоретических значений.

Поэкспериментируем с другими классами векторов. Посмотрим, что будет происходить, если изменять только значение σ (например, в таблице 1 уменьшим её значение до 0,1). Результат приведен на рис. 2.

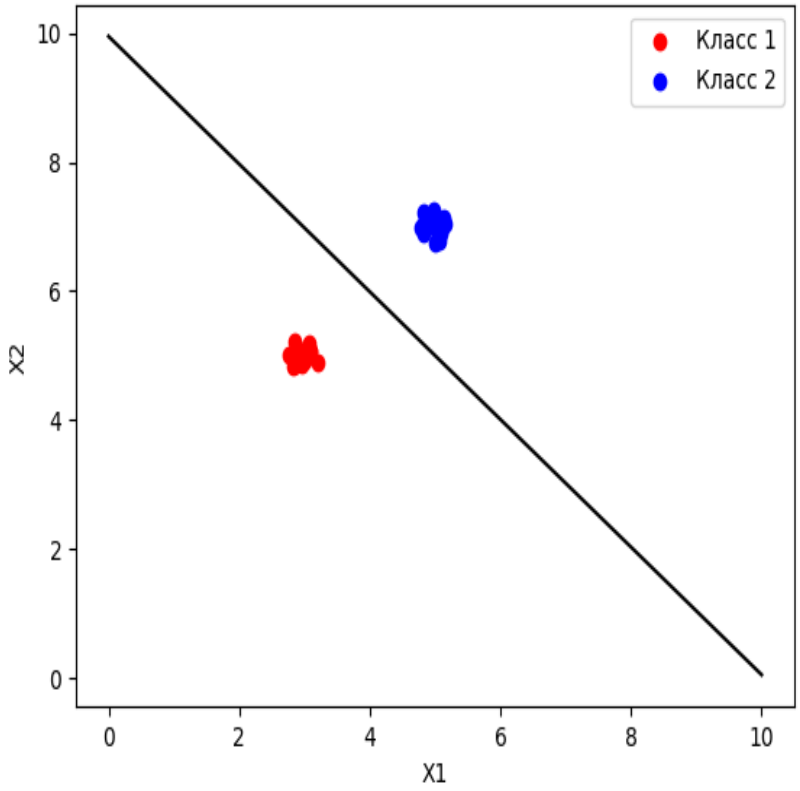

Рис. 2. Эксперимент с СКО равным 0,1

Очевидно, при одних и тех же значениях a с увеличением значения σ количество «ошибок» увеличивается, а при уменьшении значения σ – уменьшается. С изменением a центры классов могут располагаться достаточно близко (далеко) относительно друг друга, что может увеличить (уменьшить) количество «ошибок». Помимо этого возможно изменение теоретической границы.

В нашем случае модель реализовывалась на языке программирования Python, а генерация случайных векторов – с помощью его дополнительного модуля numpy.

Постановка задачи различения сходных почерков. Построение модели. Итак, решим одну из широко распространенных задач в системах по распознаванию рукописных текстов –

задачу различения сходных почерков – методом классификации данных, основанным на решении задачи линейного программирования. На данном этапе задача имеет следующую постановку: по имеющимся k росписям двух лиц построить модель, которая с наибольшей надёжностью определяла бы, какому из двух лиц принадлежит каждая роспись [12].

Для эксперимента были отобраны два лица («А» и «Б» – близнецы) со сходными почерками. Очевидно, что для одного и того же человека характерен некоторый разброс характеристик почерка от одного акта к другому. Так, было сфотографировано по 60 букв «з», написанных каждым лицом (то есть в нашем случае k равно 120). На рис. 3 приведены образцы почерков.

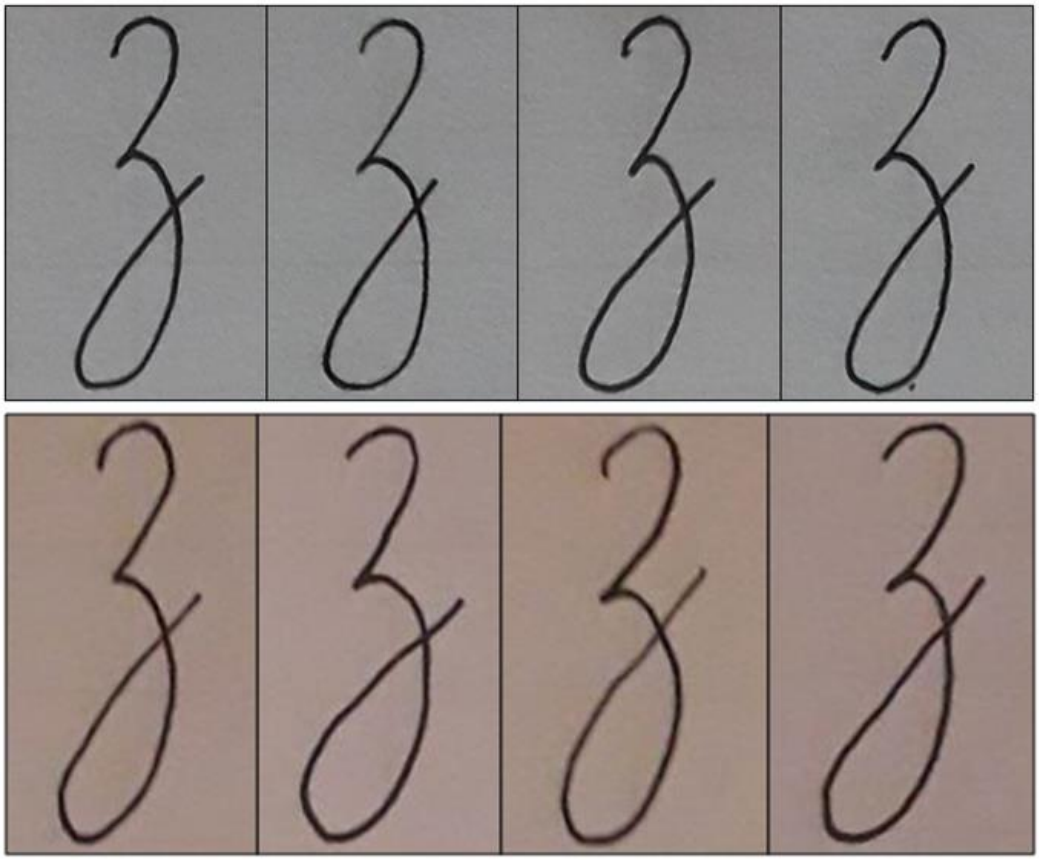

Рис. 3. Примеры почерков лиц «А» и «Б» соответственно

Все фотоснимки букв масштабировались и центрировались одинаково (рис. 4). Обработку получившихся изображений проводили с помощью языка программирования Python.

Сначала для получения векторов необходимо определить контур почерка, то есть кривую, соединяющую все непрерывные точки вдоль границы объекта. Затем – контрольные точки, на которых будет основываться способ параметризации векторов. В данной работе для обнаружения краев была применена специально предназначенная для этого функция cv.Canny().

Чтобы определить векторы, поставим задачу выделить, например, четыре «контрольные» точки (рис. 4), которые имеются на каждой росписи. На всех изображениях эти точки выбираются единообразно.

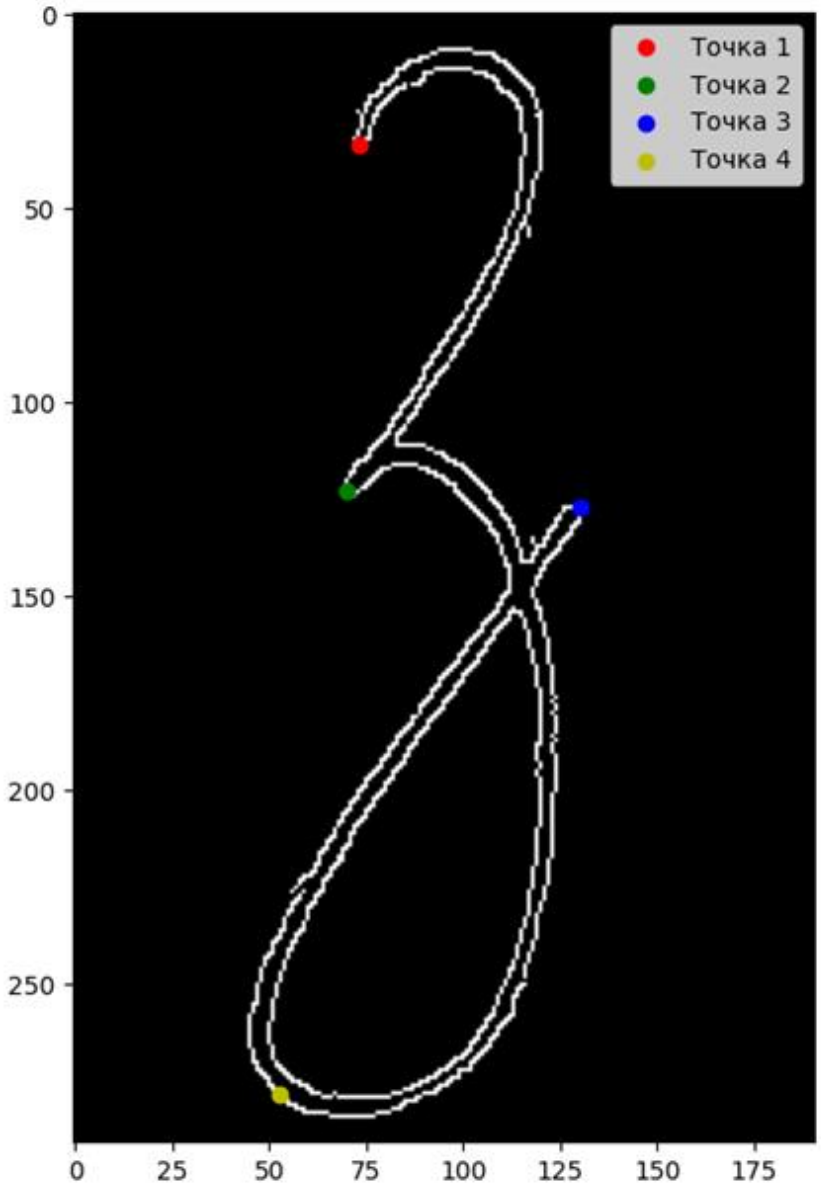

Рис. 4. Отображение «контрольных» точек

Далее рассчитываются расстояния между выбранными точками: 1-2, 2-3 и 3-4. Тем самым для каждой росписи получаем трёхмерный тренинговый вектор.

Эксперимент со 120 росписями дал следующие результаты для нормального вектора и расстояния до границы от начала координат (рис. 5): $\mathbf{w}^* = (0{,}379 \quad -0{,}246 \quad -0{,}892)$; $b^* = -131{,}764$.

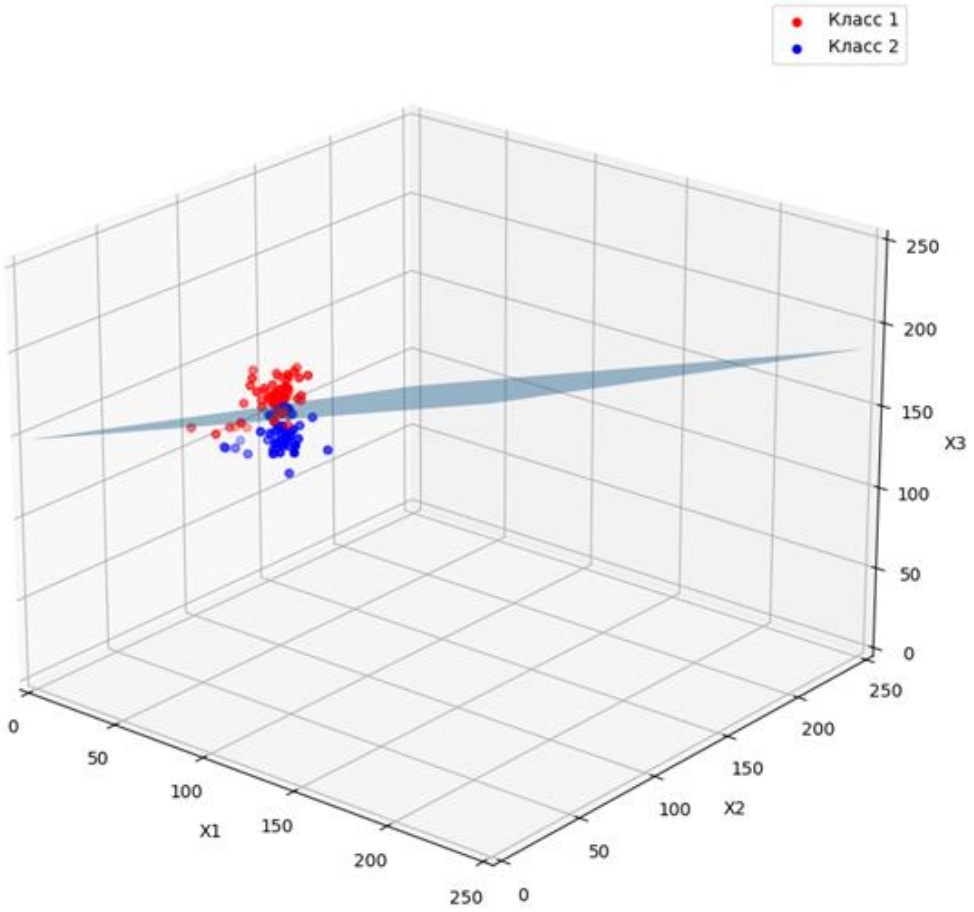

Рис. 5. Результат построения модели по 120 росписям двух лиц

Количество «ошибок» равно 20 (для 1-го класса их 9, для 2-го – 11).

Таким образом, доля правильных опознаний почерка составляет примерно 83,3%. Это указывает на то, что алгоритм классификатора прошел проверку успешно и может быть применен для последующих экспериментов.

Тестирование модели. Для более «честной» оценки пропустим через созданную модель уже новый, тестовый набор из 70 векторов от лица «А» и 30 векторов от лица «Б».

Результат работы ПЛК следующий: для 1-го класса количество ошибок равно 8, для 2-го – 8. Тем самым надёжность распознавания составила 84%.

Выводы. **Перспективы**. **Оценка эффективности метода для реальной ситуации**. В данной работе был создан и протестирован алгоритм ПЛК при решении задачи различения сходных почерков двух лиц (пары близнецов). Качество распознавания составляет около 84%. Модель реализована на языке программирования высокого уровня, Python.

Наша модель имеет узкую область применения, поскольку построена на данных образцах почерка пары близнецов (может применяться только в отношении них везде, где требуется аутентификация автора росписи: кем именно из близнецов была выполнена роспись).

В дальнейшем этот алгоритм может быть использован в качестве инструмента для проверки подлинности пользователя (речь идет о создании новой модели). Ведь различение сходных почерков является важной задачей в области криминалистики. Так, например, мошенники часто подделывают подписи в документах (договорах купли-продажи, дарения, завещаниях, накладных и других). Задача заключается в определении различий и сходств между почерками

разных людей с целью идентификации личности или выявления подделки. Допустим, первый класс остается неизменным (эталонным), а второй класс изменяемым (поддельным). В качестве образцов второго класса будут добавляться данные образцов почерка других лиц, которые пытаются подражать эталонной подписи.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Журавлев Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения. – М.: Фазис, 2005. – 159 с.

2. Кугаевских А.В., Муромцев Д.И., Кирсанова О.В. Классические методы машинного обучения. – СПб: Университет ИТМО, 2022. – 53с.

3. Marsland S. Machine Learning: An Algorithmic Perspective. CRC Press. – 2009. – 406 p.

4. Vapnik V.N. The Nature of Statistical Learning Theory. – Berlin : Springer – Verlag, 1995. – 334 p.

5. Brink H., Richards J., Fetherolf M. Real-World Machine Learning. Manning. – 2016. – 264 p.

6. Вьюгин В.В. Элементы математической теории машинного обучения: учебное пособие. – М.: МФТИ: ИППИ РАН, 2010. – 252 с.

7. Shalev-Shwartz S., Ben-David S. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press. – 2014. – 410 p.

8. Harrington P. Machine Learning in Action. Manning. – 2012. – 384 p.

9. Nefedov A. Support Vector Machines: A Simple Tutorial, 2016. – 35 p.

10. Bishop M. Christopher Pattern Recognition and Machine Learning. Springer. – 2006. – 738 p.

11. Гефан Г.Д., Иванов В.Б. Метод опорных векторов и альтернативный ему простой линейный классификатор // Информационные технологии и проблемы математического моделирования сложных систем. – Иркутск : ИрГУПС, 2012. – Вып. 10. – С. 84-94.

12. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов (статистические проблемы обучения). – М.: Наука, 1974. – 416 с.

REFERENCES

1. Zhuravlev Yu.I., Ryazanov V.V., Sen'ko O.V. *Raspoznavanie. Matematicheskie metody. Programmnaya sistema. Prakticheskie primeneniya* [Recognition. Mathematical methods. Program system. Practical applications]. Moscow, «Fazis» Publ., 2005, 159 p.

2. Kugaevskikh A.V., Muromtsev D.I., Kirsanova O.V. *Klassicheskie metody mashinnogo obucheniya* [Classical machine learning methods]. Saint Petersburg, 2022, 53 p.

3. Marsland S. Machine Learning: An Algorithmic Perspective. CRC Press. – 2009. – 406 p.

4. Vapnik V.N. The Nature of Statistical Learning Theory. – Berlin : Springer – Verlag, 1995. – 334 p.

5. Brink H., Richards J., Fetherolf M. Real-World Machine Learning. Manning. – 2016. – 264 p.

6. V'yugin V.V. *Elementy matematicheskoy teorii mashinnogo obucheniya: uchebnoe posobie* [Elements of the mathematical theory of machine learning: student textbook]. Moscow: MFTI: IPPI RAN, 2010, 252 p.

7. Shalev-Shwartz S., Ben-David S. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press. – 2014. – 410 p.

8. Harrington P. Machine Learning in Action. Manning. – 2012. – 384 p.

9. Nefedov A. Support Vector Machines: A Simple Tutorial, 2016. – 35 p.

10. Bishop M. Christopher Pattern Recognition and Machine Learning. Springer. – 2006. – 738 p.

11. Gefan G.D., Ivanov V.B. *Metod opornykh vektorov i al'ternativnyy yemu prostoy lineynyy klassifikator* [Support vector machine and its alternative simple linear classifier]. *Informacionnye tekhnologii i problemy matematicheskogo modelirovaniya slozhnyh system* [Information technologies

and problems of mathematical modeling of complex systems]. Irkutsk, IrGUPS, 2012, no. 10, pp. 84-94.

12. Vapnik V.N., Chervonenkis A.Ya. *Teoriya raspoznavaniya obrazov (statisticheskie problemy obucheniya)* [Theory of pattern recognition (statistical learning problems)]. Moscow, «Nauka» Publ., 1974, 416 p.

Информация об авторах

Григорий Давыдович Гефан – к. ф.-м. н., доцент кафедры «Математика», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: grigef@rambler.ru

Виктория Сергеевна Попова – студентка гр. БАС.5-22-1, Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: popovavika2017@yandex.ru

Надежда Сергеевна Попова – студентка гр. БАС.5-22-1, Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: Nm2nadia@yandex.ru

Authors

Grigory Davydovich Gefan – candidate of physical and mathematical sciences, associate professor of department of mathematics, Irkutsk State Transport University, Irkutsk, e-mail: grigef@rambler.ru

Victoria Sergeevna Popova – student, Irkutsk State Transport University, Irkutsk, e-mail: popovavika2017@yandex.ru

Nadezhda Sergeevna Popova – student, Irkutsk State Transport University, Irkutsk, e-mail: Nm2nadia@yandex.ru

Для цитирования

Гефан Г.Д., Попова В.С., Попова Н.С. Метод различения сходных почерков с помощью линейного классификатора // «Информационные технологии и математическое моделирование в управлении сложными системами»: электрон. науч. журн. – 2024. – №2. – С. 15-23 – Режим доступа: http://ismm-irgups.ru/toma/222-2024, свободный. – Загл. с экрана. – Яз. рус., англ. (дата обращения: 19.06.224)

For citations

Gefan G.D., Popova V.S., Popova N.S. Method of recognizing similar handwriting using a linear classifier // Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: ehlektronnyj nauchnyj zhurnal [Information technology and mathematical modeling in the management of complex systems: electronic scientific journal], 2024. No. 2. P. 15-23. [Accessed 19/06/24]