УДК 004.852
DOI: 10.15827/2311-6749.18.1.4

ПРИМЕНЕНИЕ МЕТОДОВ КЛАССИФИКАЦИИ ДЛЯ АНАЛИЗА ВИЗИТНЫХ КАРТОЧЕК В МОБИЛЬНОМ ТЕЛЕФОНЕ

С.А. Беляев, к.т.н., доцент, bserge@bk.ru; Т.В. Гордеева, студентка, tangord@mail.ru
(Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина) (СПбГЭТУ «ЛЭТИ»), ул. Профессора Попова, 5, г. Санкт-Петербург, 197376, Россия)

В статье рассмотрены подходы к решению задачи классификации, применяемые для анализа визитных карточек в мобильном телефоне. Выявлены достоинства и недостатки существующих приложений, специализирующихся на сканировании визиток. Предложена математическая модель выбранного подхода. Рассмотрен вариант архитектуры предлагаемого решения. Представлены результаты классификации на нескольких примерах. Сделаны выводы о работоспособности подхода и дальнейших улучшениях решения.

Ключевые слова: *Business Card Reader, OCR, многоклассовая классификация, наивный байесовский классификатор, метод опорных векторов, математическая модель.*

Визитная карточка – лицо делового человека. В современном мире она является инструментом, который отражает информацию о человеке или организации, такую как имя, фамилия, email, телефон, название организации и т.д. Зачастую количество визиток с каждым днем прибавляется и становится все сложнее хранить их и искать важные контакты.

Данную проблему помогает решить Business Card Reader (BCR), основанный на задаче распознавания визиток. Задача BCR – дать людям возможность извлечь информацию из визитных карточек, не прибегая к ручной установке адресов, телефонов, имен.

Процесс распознавания визиток проходит в несколько этапов:

– извлечение текста из изображений при помощи камеры мобильного телефона и оптического распознавания символов – Optical Character Recognition (OCR);

– разделение полученного текста на категории.

Существующие приложения

В настоящее время задача сканирования визиток реализована в платном приложении Business Card Reader Free от ABBYY [1], которое предоставляет бесплатно распознавание десяти визиток. BCR Free предлагает пользователям переносить данные визиток на 25 языках и имеет обширную словарную базу, что увеличивает точность распознавания. При этом в приложении нет возможности заново распознать уже имеющуюся визитку при ее редактировании.

Платные приложения Business Card Reader Pro [2] от SHAPE, ScanBizCards [3] от ScanBiz Mobile Solutions L.P., WorldCard Mobile [4] от Penpower Technology Ltd. доступны для устройств с iOS.

В основном многие существующие решения являются платными и поддерживаются не всеми платформами, а в бесплатных версиях не всегда включен весь необходимый функционал. В связи с этим целесообразна разработка нового приложения, компенсирующего данные недостатки.

Этапы распознавания визитных карточек

Для извлечения текста из визитной карточки можно воспользоваться любым OCR, например Tesseract [5]. При этом точность полученного результата снижается, если визитная карточка напечатана на сложном фоне, поэтому следует применять алгоритмы первоначальной обработки изображения для изоляции текста [6, 7].

Следующим этапом является классификация полученного текста. При этом невозможно перечислить все варианты в словаре, так как существует большое количество имен, фамилий, организаций и т.д. Поэтому необходимо выполнять автоматический разбор категорий.

Рассматриваемые подходы к решению задачи классификации используют обучение с учителем [8]: на обучающем наборе слов вычисляются статистические параметры модели, а затем данные параметры используются для предсказания класса неизвестных слов.

Текст визитной карточки может содержать большое количество категорий, поэтому задача сводится к многоклассовой классификации, которая имеет следующую формальную постановку:

X – пространство объектов;

Y – множество ответов;

$y: X \to Y$ – неизвестная целевая зависимость, значения которой известны только на объектах конечной обучающей выборки $X^l = (x_i, y_i)_{i=1}^l$.

Требуется построить алгоритм $a: X \to Y$, который способен классифицировать произвольный объект на всем пространстве X.

Стратегии многоклассовой классификации

Задачу можно свести к решению нескольких бинарных задач, так как большинство методов многоклассовой классификации либо основаны на бинарных классификаторах, либо сводятся к ним. Для этого рассмотрим стратегии «один-ко-многим» и «многие-ко-многим».

Идея стратегии «один-ко-многим» состоит в построении M классификаторов, которые отделяют каждый класс от остальных. Получим M задач бинарной классификации. Вычисляем оценки принадлежности каждому классу $p_i(i) \in R$.

Тогда решающее правило принимает вид

$$a(x) = \underset{m=1..M}{\operatorname{argmax}} p_m(x).$$

Данный алгоритм строит линейное число классификаторов, каждый из которых обучается на полной выборке.

Идея стратегии «многие-ко-многим» состоит в построении классификаторов для каждой пары классов. Получим $M(M-1)$ задач бинарной классификации. Вычисляем оценки принадлежности каждому классу $p_{mk}(x) \in \{0, 1\}$.

Тогда решающее правило принимает вид

$$a(x) = \underset{m=1..M}{\operatorname{argmax}} \sum_{k=1}^{M} p_{mk}(x).$$

Данный алгоритм строит квадратичное число классификаторов, при этом каждый из классификаторов обучается на небольшой подвыборке.

Методы классификации для анализа визитных карточек

В задаче классификации применяют такие алгоритмы, как наивный байесовский классификатор [9] и метод опорных векторов [10].

Рассмотрим их подробнее.

Наивный байесовский классификатор

В основе классификатора лежит теорема Баейса:

$$P(k|x) = \frac{P(x|k)P(x)}{P(k)}, \text{ где}$$

$P(k|x)$ – вероятность, что объект x принадлежит классу k;

$P(k|x)$ – вероятность встретить объект x среди всех объектов класса k;

$P(x)$ – безусловная вероятность встретить объект класса k в корпусе объектов;

$P(k)$ – безусловная вероятность документа объекта x в корпусе объектов.

Далее требуется рассчитать вероятность всех классов и выбрать класс с наибольшей вероятностью, то есть

$$k = \underset{k}{\operatorname{argmax}} \frac{P(x|k)P(x)}{P(k)}.$$

Предполагаем, что признаки объекта x зависят от класса k и не зависят друг от друга:

$$P(x|k) \approx P(x_1|k)P(x_2|k)\ldots P(x_n|k) = \prod_{i=1}^{n} P(x_i|k).$$

Теперь формула классификатора принимает вид

$$k = \underset{k}{\operatorname{argmax}} P(x) \prod_{i=1}^{n} P(x_i|k).$$

Метод опорных векторов

Рассмотрим задачу бинарной классификации.

Пусть объекты представлены вектором $x \in R_n$ Тогда классификатор принимает вид

$$k = \operatorname{sign}\left(\sum_{j=1}^{n} w_j x_j - w_0 \right), \text{ где}$$

x_i – признаки объекта x;

$w = (w_1, \ldots, w_n) \in R^n$, $w_0 \in R$ – параметры алгоритма;

$\langle w, x \rangle = w_0$ описывает гиперплоскость, разделяющую классы в пространстве.

Для выбора наиболее подходящего алгоритма для анализа визитных карточек в мобильном телефоне следует учитывать время обучения и качество классификации.

Сформулируем задачу в виде следующей математической модели: $M = (X, Y, X^l, G, A, K, W, F, P, H)$, где X – множество классифицируемых слов; $Y \in \{-1, 1\}$ – множество допустимых ответов; X^l – множество объектов обучающей выборки; $G = y^*: X \to Y$ – неизвестная целевая зависимость, значения которой известны только на объектах конечной обучающей выборки $X^l = (x_i, y_i)_{i=1}^l$; $A: X \to Y$ – алгоритм приближения неизвестной целевой зависимости на всем множестве X; $K = \{k\}$ – множество классов; W – множество признаков каждого класса; $F: X \to W$ – функция извлечения признаков W из классифицируемых объектов; P – множество вероятностей принадлежности классифицируемого слова каждому классу; H – множество классификаторов A.

Рассмотрим работу математической модели на примере. Пусть имеется множество классифицируемых слов X, состоящих из имен, фамилий и отчеств на русском языке, тогда множество K имеет вид K = {Фамилия, Имя, Отчество}. Пусть множество Y имеет вид $Y = \{0, 1, 2\}$, 0 – если $x \in X$ является именем, 1 – фамилией, 2 – отчеством.

В качестве множества признаков W извлечем триграммы с конца слов и добавим первый символ. Данный выбор обоснован тем, что большое количество фамилий на русском языке имеют схожие окончания (например, {О, В, А}, {К, О, В}, {Н, О, В}, {И, Н, А}, {Е, В, А}). Также имеются распространенные триграммы среди имен и отчеств.

Обучим классификаторы на конечной размеченной выборке X^l. Процесс обучения построим следующим образом: 70 % примеров из обучающей выборки используем для вычисления параметров модели, 30 % – для оценки качества классификатора. Также подберем наилучшие параметры алгоритмов (для метода опорных векторов) с помощью кросс-валидации.

На основании данных, взятых из [12, 13] (общее количество – 4 500 слов), проведены эксперименты. Их результаты показали, что наилучшие показатели классификации выявлены при использовании наивного байесовского алгоритма со стратегией «один-ко-многим» (табл. 1, 2).

Таблица 1

Результаты бинарной классификации

Категория	Время обучения, сек.	F-мера (micro), %
Наивный байесовский классификатор		
Фамилия-Имя	0,026	83,5
Фамилия-Отчество	0,021	99,45
Имя-Отчество	0,021	99,86
Метод опорных векторов		
Фамилия-Имя	0,125	74,34
Фамилия-Отчество	0,013	97,79
Имя-Отчество	0,04	98,4

Таблица 2

Показатели классификатора при использовании стратегии «один-ко-многим»

Алгоритм	Время обучения, сек.	F-мера (micro), %
Наивный байесовский классификатор	0,037	88,96
Метод опорных векторов	0,2612	80,38

Архитектура системы распознавания визитных карточек

С учетом предложенной математической модели и требований к ее решению разработана архитектура системы распознавания визитных карточек (рис. 1).

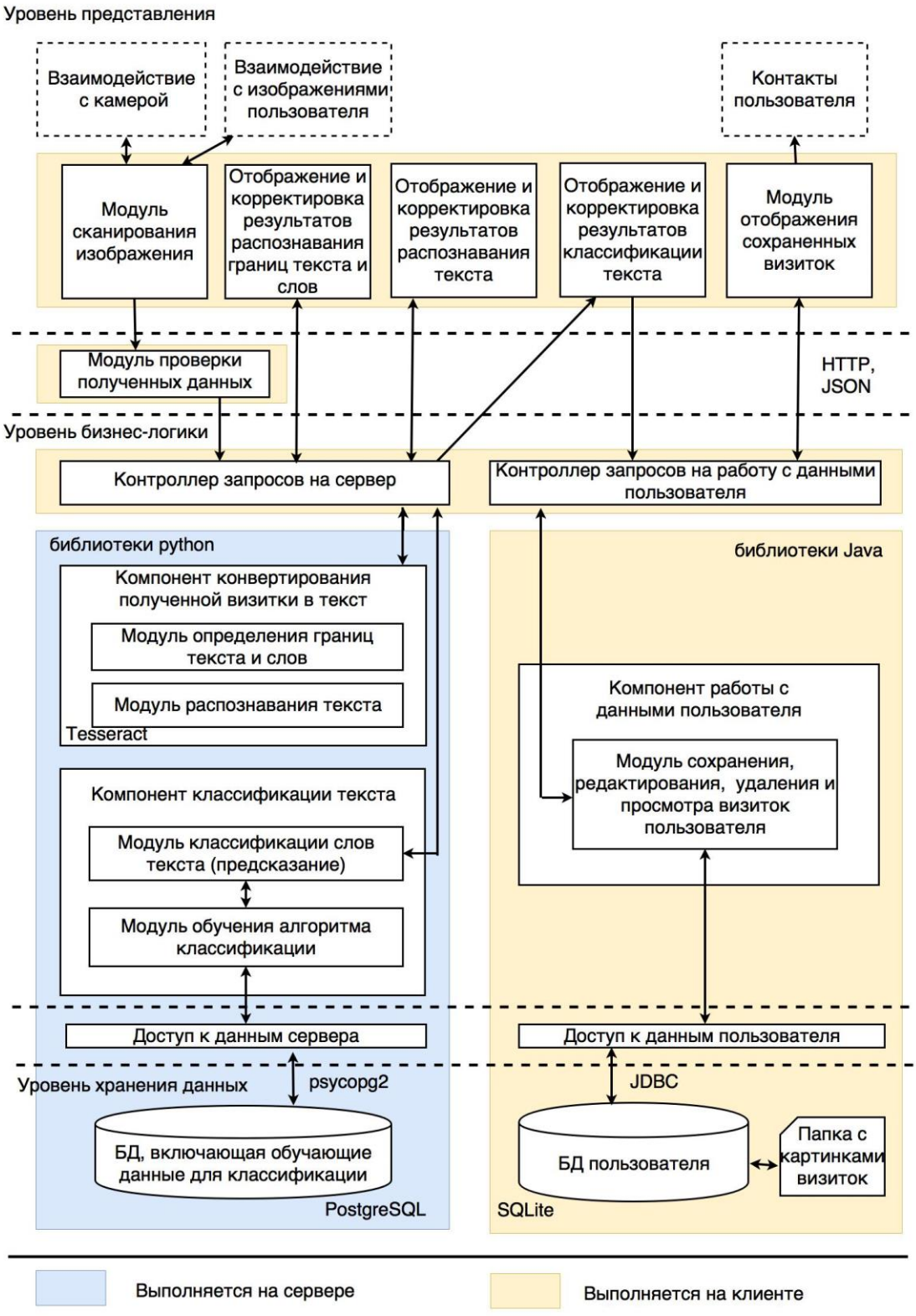

Рис. 1. Архитектура решения

В архитектуре выделены уровни представления, бизнес-логики и хранения данных.

На уровне представления выполняются считывание изображения, запрос на передачу данных серверу для распознавания и редактирование классифицированных данных, запрос на сохранение результатов

распознавания визитки. Также осуществляются просмотр и редактирование уже отсканированных и сохраненных визиток.

На уровне бизнес-логики выполняются операции по распознанию текста на визитках и по формированию ответа пользователю на запросы классификации, сохранения, изменения и удаления данных. Также на данном уровне формируются запросы в БД.

Уровень хранения данных содержит БД для классификации (имена, фамилии, должности и т.д.) и хранения данных пользователя (сохраненных визиток).

Для поиска телефонов и email предлагается использовать алгоритмы на основе регулярных выражений.

Схема работы компонента классификации представлена на рисунке 2.

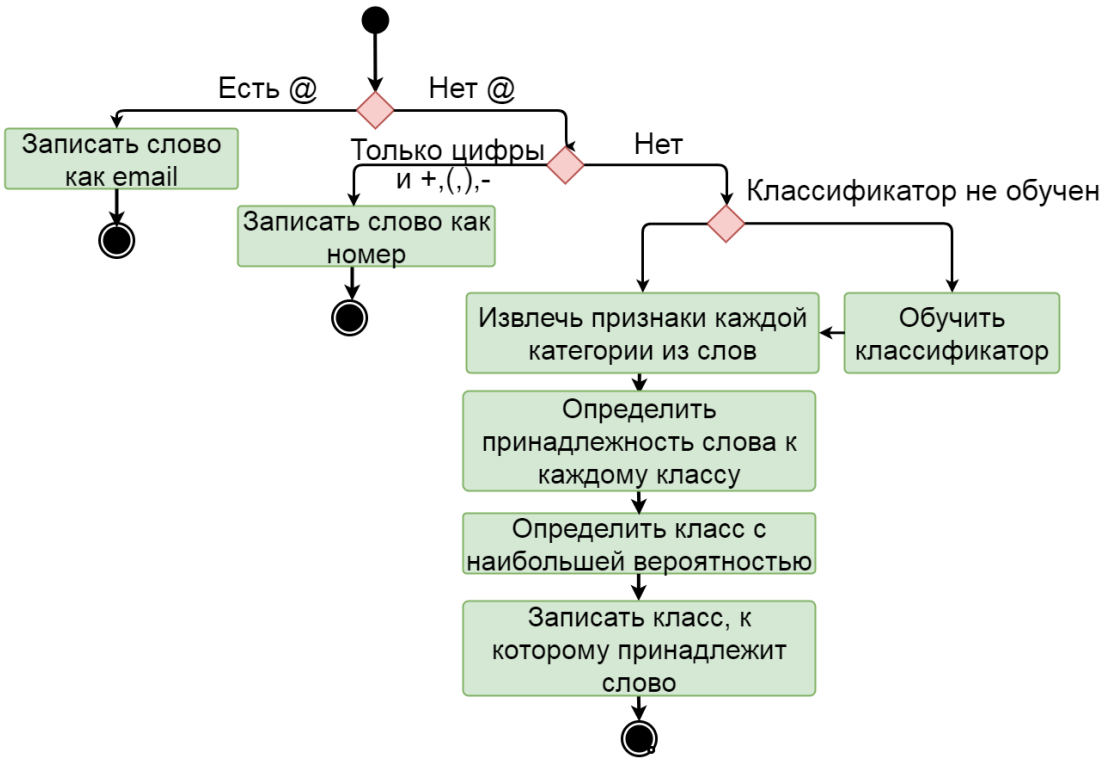

Рис. 2. Схема работы компонента классификации текста

Обучение классификатора проведем с использованием наивного байесовского алгоритма со стратегией «один-ко-многим» на размеченном корпусе, который состоит из имен, фамилий и отчеств в разном количественном соотношении.

Результаты экспериментов на тестовой выборке представлены в таблице 3.

Таблица 3

Результаты экспериментов

Количество слов	**Соотношение классов (имя/фамилия/отчество)**	**Время обучения, сек.**	**F-мера (micro), %**
4 500	1 500/1 500/1 500	0,033	89,6
7 500	3 000/3 000/1 500	0,059	86,1
15 000	7 500/6 000/1 500	0,156	84,8

По результатам экспериментов можно сделать вывод, что количество объектов обучающей выборки, принадлежащих определенному классу, влияет на точность классификации, так как при использовании стратегии «один-ко-многим» может возникнуть проблема с несбалансированными выборками. Поэтому следует оптимизировать количество объектов каждого класса при обучении классификатора.

Выводы

Предложенный подход при анализе визитных карточек позволяет проводить классификацию текста, содержащего фамилии, имена и отчества с точностью от 84 %. Данный показатель может быть увеличен за счет поиска дополнительных признаков классов. Также предстоит решить задачу выбора признаков для других классов, которые могут содержаться в тексте визитки. Например, организация, должность, адрес и другие.

Литература

1. Card Reader Free. ABBYY. URL: http://www.abbyybcr.com/en/ (дата обращения: 20.01.2018).
2. Card Reader Pro. SHAPE. URL: http://www.shape.ag/en/products/details.php?product=bcr (дата обращения: 20.01.2018).
3. ScanBizCards. ScanBiz Mobile Solutions. URL: https://itunes.apple.com/us/app/scanbizcards/id335047649?mt=8 (дата обращения: 20.01.2018).
4. WorldCard Mobile. Penpower Technology. URL: http://www.penpowerinc.com/product.asp?sn=392 (дата обращения: 20.01.2018).
5. Tesseract OCR. URL: https://opensource.google.com/projects/tesseract (дата обращения: 20.01.2018).
6. Nagabhushan P., Nirmala S. Text Extraction in Complex Color Document Images for Enhanced Readability. Scientific research, 2010, vol. 2, pp. 120–133; DOI: 10.4236/iim.2010.22015.
7. Mollah A.F., Basu S., Nasipuri M., Basu D.K. Text/graphics separation for business card images for mobile devices. Jour. of Computing, 2010, vol. 2, pp. 96–102. URL: http://www.academia.edu/3135051/Text_Graphics_Separation_and_Skew_Correction_of_Text_Regions_of_Business_Card_Images_for_Mobile_Devices (дата обращения: 20.01.2018).
8. Guido S., Müller A. Introduction to machine learning with Python. O'Reilly Media, 2016, 282 p.
9. Barber D. Bayesian reasoning and machine learning. Cambridge Univ. Press, 2012, 672 p.
10. Hastie T., Tibshirani R., Friedman J. The Elements of statistical learning. Springer, 2001, 552 p.
11. James G., Witten D., Hastie T., Tibshirani R. An introduction to statistical learning with applications in R. Springer, 2013, 426 p.
12. База данных имен и фамилий. URL: https://mydata.biz/ru/catalog/databases/names_db (дата обращения: 20.01.2018).
13. Программа подготовки отчетных документов для ПФР "Spu_orb". URL: http://www.pfrf.ru/branches/orenburg/info/~rabot/program (дата обращения: 20.01.2018).