Коррекция распознанного текста с использованием методов классификации

Д. Л. Шоломов

В статье представлен подход к классификации объектов с нефиксированным текстовым представлением. Данная задача возникает на этапе контекстной обработки результатов распознавания полей ввода в процессе распознавания структурированных текстовых документов. Данный классификатор был применен для распознавания наименований учебных заведений при вводе анкет школьников и студентов, а также и при обработке поля «Наименование товара» на отгрузочных разнарядках. В статье указаны качественные характеристики алгоритмов и особенности технической реализации.

В настоящее время в области информационных технологий широко используются системы распознавания структурированных документов. Обработка документа подразумевает его ввод с бумажного носителя, сортировку, маршрутизацию, распознавание, верификацию, а также экспорт во внешнюю информационную систему. Существенную долю среди обрабатываемых документов составляют стандартизированные формы, для которых состав, расположение и содержание полей ввода информации некоторым образом фиксированы. Для систем, специализированных на обработке стандартных форм априорное знание структуры документа, правил его заполнения, синтаксиса и семантики полей ввода является важной входной информацией, существенно расширяющей технологические возможности обработки документа и повышающей качество распознавания. Важной стадией процесса распознавания документа является стадия контекстной обработки результатов распознавания, на которой производится корректирование распознанных значений с учетом особенностей структуры поля. Например, в работах [5, 9, 10] авторами исследуются методы контекстной обработки, основанные на синтаксическом описании поля при помощи порождающих КС-грамматик Хомского в форме синтаксических диаграмм и БНФ, а также приводится классификация типов полей на формах и указываются методы применяющиеся для их коррекции.

При распознавании ряда документов возникает задача классифицировать значение поля, то есть отнести распознанное значение к одному или нескольким классам объектов.

Пример 1

Пусть имеется множество учебных заведений города Москвы. Будем называть реальные учебные заведения объектами. У каждого объекта есть множество вариантов написания, например, объект, соответствующий учебному заведению «Средняя Общеобразовательная Школа № 9 г. Москвы» может быть написан, как «Школа 9», «Московская Средняя Школа 9», «СОШ № 9 Волгоградского р-на Москвы» и вряд ли может быть написан, как «Школа № 11».

Пример 2

Поле «*Наименование товара*» на отгрузочных разнарядках, см. рис. 1, содержит текстовое описание товара. Товар может быть указан различ-

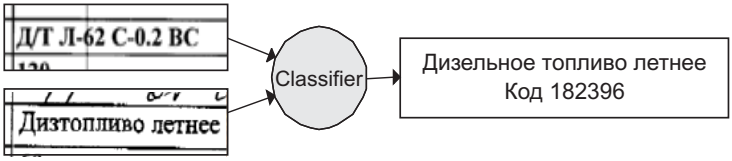

Рис. 1. Классификация поля «Наименование товара»

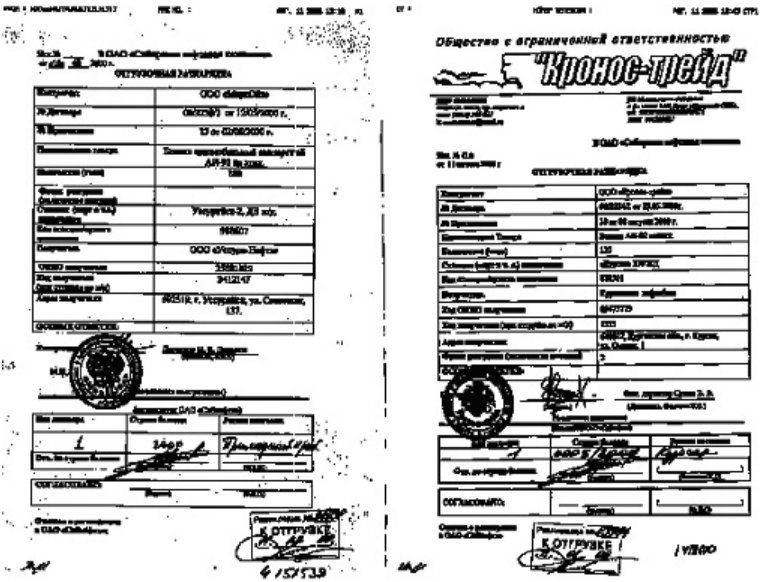

Рис. 2. Примеры отгрузочных разнарядок

ными способами. Например, дизельное топливо летнее может иметь такие текстовые представления как «Д/Т Л-62 С-0.2», «Дизтопливо летнее», «ДТЛ» и т. д. Требуется поставить в соответствие результату распознавания поля номер товара по определенной базе данных. Примеры отгрузочных разнарядок приведены на рис. 2.

1. Постановка задачи

В классификационной постановке задачи контекстной обработки результаты распознавания должны быть отнесены к одному или нескольким классам объектов. Введем формальное определение задачи классификации текстового фрагмента следующим образом.

Пусть задана пятерка $\langle \Omega, C, W, W_T, f_s \rangle$, где

- Ω — множество образцов;
- C — множество классов; $C = \{C_i\}$, если все классы известны заранее, либо $C = \{C_i, C^*\}$, где C^* — дополнение до $\{C_i\}$. C^* обозначает так называемый неизвестный класс, будем считать, что в случае если образец не принадлежит ни одному из классов C_i, он принадлежит C^*. Каждый класс C_i представлен описанием F_i;
- W — обучающая выборка образцов $W = \{\omega_i : \omega_i \in \Omega\}$;
- W_T — тестовая выборка образцов $W_T = \{\omega_i : \omega_i \in \Omega\}$;
- f_s — «экспертный» классификатор — функция $f_s : W \cup W_T \to P^{|C|}$, где

$$P^n = \left\{ \overline{p} \in R^n : \ \overline{p} = (p_1, \ldots, p_n), \quad p_i \in R, \quad \sum_{i=1}^{n} p_i = 1 \right\}.$$

  f_s обычно определяется человеком, который указывает — к каким классам и с какой вероятностью (либо оценкой) принадлежат образцы из тестовой и обучающей выборок W и W_T.

Требуется найти функцию классификации (классификатор) $f : \Omega \to P^{|C|}$, который наилучшим образом «приближает» f_s на тестовой выборке W_T. Оценить приближение можно различными способами, например, при помощи функции ошибки

$$e(f) = \sum_{\omega_i \in W_T} \mu(f(\omega_i), f_s(\omega_i)),$$

где $\mu(f_1, f_2)$ — функция метрики. В простейшем случае, когда классификация однозначная, т. е.

$$f(\omega) = (0, \ldots, 0, 1, 0, \ldots 0), \quad \text{а} \quad \mu(f(\omega_1), f(\omega_2)) = \begin{cases} 1, & f(\omega_1) = f(\omega_2), \\ 0, & f(\omega_1) \neq f(\omega_2). \end{cases}$$

$e(f)$ является числом ошибочных классификаций на тестовой выборке. Оптимальным классификатором является тот, у которого функция ошибки $e(f)$ минимальна.

Обычно в задачах классификации используется понятие признака. Признак (атрибут) — это некоторая характеристика образца, возможно не имеющая физической природы. Признак может принимать значения из некоторого множества A. В случае, если $A = \{0, 1\}$ — признак называется бинарным. Должна быть определена функция выделения признака $a : \Omega \to P^{|A|}$, которая по образцу выдает вероятностное распределение на множестве значений данного признака. В случае бинарного признака, т. е. принимающего значения 0 или 1, функция выделения выдает вероятность того, что признак присутствует в образце $a(\omega) = p$. В случае, если $p \in \{0, 1\}$ функция выделения признака a называется однозначной. Серьезными подзадачами классификации являются задачи определения множества признаков, а также выделения независимого подмножества.

При классификации текстового значения, образцом ω является результат распознавания текста. В некоторых случаях информация об альтернативах распознавания в целях оптимизации по скорости не используется, тогда образец ω — это строка текста. Описания F_i классов C_i — это, как правило, нормальные текстовые значения, либо уникальный код объекта, например, соответствующий коду класса C_i в некоторой базе данных.

В терминах определения задачи классификации в Примере 1 множество учебных заведений Москвы соответствует множеству классов $C = \{C_i\}$. Множество образцов $\Omega = \{\omega\}$ состоит из результата распознавания текстового поля с алфавитом, состоящим из букв русского языка, цифр и знаков препинания.

Пусть изначально для каждого учебного заведения указан набор вариантов его написания, путем разделения данного набора на две части получим обучающую и тестовую выборки. Варианты написания учебных заведений могут повторяться с учетом частоты их встречаемости. Это значит, что при помощи имеющегося набора вариантов задано множество W — обучающая выборка и функция «экспертного» классификатора f_s на множестве W. Также задана тестовая выборка W_T и функция f_s на W_T. Надо отметить, что в случае такого задания исходных данных функция f_s классифицирует текстовый образец однозначно. Образцы, которые невозможно экспертно классифицировать однозначно — удаляются из выборки. Задача состоит в указании алгоритма построения классификатора f, с наименьшей функцией ошибки $e(f)$, который по заданному образцу текста указывает к каким учебным заведениям и с какой вероятностью его можно отнести.

Подробнее информация по методам классификации дана в [2], на русском языке в более сжатой форме — в [13, 14], классификация при помощи нейронных сетей наиболее полно описана в [17].

Обзорной статьей по методам классификации текстовых документов является [8], по байесовым методам — [6]. Общая информация на русском языке приведена в [11]. Также работами по данной тематике являются [3, 7].

2. Признаки и функции выделения признаков

В качестве признаков $A = \{a_j\}$ выберем множество всех слов, встретившихся в образцах обучающей выборки. Слово — это подстрока в заданном алфавите, окруженная разделительными символами (пробел, знаки препинания и т. д.) Рассмотрим следующие варианты выбора функций $a_j(\omega)$ выделения признака a_j из образца ω.

а) Функция $a_j^{xT}(\omega)$ выделяет признак из текста, по наилучшим альтернативам символов на знакоместах результатов распознавания. Пусть строка **s** соответствует наилучшим альтернативам, тогда

$$a_j^T(\omega) = \begin{cases} 1, & \text{если} \quad a_j \text{ — подстрока } \mathbf{s}, \\ 0, & \text{иначе.} \end{cases}$$

В данном случае функция выделения признаков работает быстро, но даже при незначительной ошибке распознавания последствия могут носить весьма существенный характер, если ошибка распознавания произошла в ключевом признаке.

б) Функция $a_j^M(\omega)$ выделяет признак при помощи алгоритма MCHSR пристраивания текстового фрагмента к результатам распознавания, алгоритм описан в [16]. Если оценка, полученная от MCHSR превосходит некоторый порог t, то признак выделен, иначе — не выделен, т. е.

$$a_j^M(\omega) = \begin{cases} 1, & \text{MCHSR}(\omega, a_j) \geqslant t, \\ 0, & \text{иначе.} \end{cases}$$

При использовании MCHSR функция выделения признаков работает медленнее, но гораздо более устойчива к ошибкам распознавания.

в) Функция $a_j^P(\omega)$ так же, как и $a_j^M(\omega)$ выделяет признак при помощи алгоритма MCHSR, но принимает значение из диапазона $[0, 1]$.

$$a_j^P(\omega) = prob(\text{MCHSR}(\omega, a_j)).$$

Функция $prob$ преобразует оценку алгоритма MCHSR к вероятности того, что признак a_j содержится в образце ω. Функция $a_j^P(\omega)$

дает наиболее полную информацию о том содержится признак a_j в образце или нет, но при этом требуется привлекать специальные методы для преобразования оценок алгоритма MCHSR к вероятностям, см. например [1]. Это требует достаточно большого объема обучающего материала, при недостаточном его количестве можно использовать 2-х пороговую функцию $prob$, принимающую значения $\{0; 0{,}5; 1\}$.

3. Построение первичного классификатора

Пусть выбрано множество признаков $A = \{a_j\}$ и заданы функции выделения признаков из образца $a_j(\omega)$, которые выдают вероятность нахождения признака a_j в образце ω. Пусть обучающая выборка содержит образцы $W = \{\omega\}$, и функция f_s задана следующим образом.

$$f_s(\omega) = (0, \ldots, 0, 1_i, 0, \ldots, 0), \quad \text{если} \quad \omega \in C_i,$$

т. е. образцы классифицируются функцией f_s однозначно.

Введем следующие обозначения:

$$W_i \equiv \{\omega : \omega \in C_i\}$$

— множество образцов из обучающей выборки, которые f_s классифицирует, как принадлежащие классу C_i.

$$W_j^+ \equiv \{\omega \in W : a_j(\omega) \geqslant p^+\},$$

где p^+ — константа. Это множество образцов из обучающей выборки, в которых признак a_j содержится с вероятностью не меньше, чем p^+.

$$W_{ij}^+ \equiv \{\omega \in W_i : a_j(\omega) \geqslant p^+\}$$

— множество образцов из множества W_i, в которых признак a_j содержится с вероятностью не меньше, чем p^+.

$$p_{ij} \equiv \frac{|W_{ij}^+|}{|W_i|}$$

— вероятность того, что признак a_j принадлежит классу C_i (табл. 1), в случае если $a_j(\omega)$ выделяет признаки однозначно.

$$|W_{ij}^+|$$

— количество образцов из класса C_i, в которых присутствует признак a_j.

В случае, если $a_j(\omega)$ выделяет признаки не однозначно,

$$p_{ij} \equiv \frac{\sum\limits_{\omega \in W_i} a_j(\omega)}{|W_i|}.$$

Таблица 1

Вероятности принадлежности признаков a_j классам C_i

	a_1	a_j	a_k
C_1	p_{11}	...	p_{1k}
C_i	...	p_{ij}	...
C_N	p_{N1}	...	p_{Nk}
C^*	p_1^*	p_j^*	p_k^*
ω	$a_1(\omega)$	$a_j(\omega)$	$a_k(\omega)$

Обозначим

$$P_{ij}(\omega) = p_{ij}a_j(\omega) + \overline{p}_{ij}\overline{a}_j(\omega), \quad \text{где} \quad \overline{p} = 1 - p.$$

$P_{ij}(\omega)$ — вероятность того, что образец ω принадлежит классу C_i по признаку a_j.

$$P_i = \prod_{j=1}^{k} P_{ij}$$

— вероятность того, что образец ω принадлежит классу C_i.

Пусть $\overline{P} = (P_1, \ldots, P_N)$, тогда определим первичный классификатор следующим образом:

$$f_p(\omega) = \left(\frac{P_1}{|P|}, \ldots, \frac{P_N}{|P|}\right), \quad \text{где} \quad |P| = \sqrt{P_1^2 + \ldots + P_N^2}.$$

Построенный классификатор ESX, относится к так называемым байесовым классификаторам.

4. Сравнение функций выделения признаков

В п. 2 рассматривались варианты функций выделения признаков $a_j^T(\omega)$, $a_j^M(\omega)$, $a_j^P(\omega)$. В данном пункте приведем результаты сравнения качественных оценок первичного классификатора $f_p(\omega)$, использующего указанные функции выделения признаков.

Изначально список средних учебных заведений города Москвы состоял из 2131 наименований. Для тестирования первичного классификатора из множества учебных заведений было выделено подмножество,

Таблица 2

Результаты сравнения функций выделения признаков

720 образцов	$a_j^T(\omega)$	$a_j^M(\omega)$	$a_j^P(\omega)$
Классифицирован правильно, причем как наилучший вариант	422 (58,6 %)	549 (76,3 %)	587 (81,5 %)
Содержится среди предложенных вариантов	454 (63,1 %)	591 (82,1 %)	646 (89,7 %)
Классифицирован не правильно	266 (36,9 %)	129 (17,9 %)	74 (10,3 %)

состоящее из 500 наиболее репрезентативных, т. е. тех учебных заведений, к которым относилось наибольшее число образцов обучающей выборки. Обучение проводилось на множестве из 4272 образцов. Сравнение проводилось на тестовой выборке из 720 анкет с полем «*Наименование учебного заведения*» заполненным рукопечатно. Поле содержало название среднего учебного заведения города Москвы.

Результаты сравнения приведены в табл. 2.

Из таблицы видно, что использование алгоритма MCHSR дает существенное улучшение. Это связано с тем, что при рукопечатном заполнении качество распознавания на смешанном буквенно-цифровом алфавите относительно низкое и появление незначительных ошибок распознавания в текстовом фрагменте, содержащем существенный признак, приводит к ошибке классификации. Функцию $a_j^T(\omega)$ стоит использовать только в том случае, если в силу архитектуры программы к моменту классификации информация о распознавании в виде матрицы альтернатив распознавания утеряна, иначе предпочтительнее использование $a_j^P(\omega)$.

5. Задача с неизвестными классами

В случае, если изначально не все классы C_i известны, нужно учитывать возможность того, что классифицируемый образец не принадлежит ни одному из классов C_i, будем в этом случае говорить, что классифицируемый образец принадлежит некоторому неизвестному классу.

Обозначим как C^* — условный класс всех образцов встречающихся в данном поле. Будем использовать предположение о том, что образец не принадлежит ни одному из классов C_i, если вероятность

$$P_i = \prod_{j=1}^{k} P_{ij}$$

принадлежности данного образца любому из классов C_i рассчитанная на основе признаков a_j не существенно выше вероятности принадлежности образца классу C^*.

Введем следующие обозначения:

$$p_j^* = \frac{|W_j^+|}{|W|}$$

— вероятность того, что признак a_j принадлежит классу C^*, в случае если функции $a_j(\omega)$ выделяют признаки однозначно, p_j^* отражает вероятность встретить признак a_j в произвольном образце.

$$p_j^* = \frac{\sum\limits_{\omega \in W} a_j(\omega)}{|W|},$$

в случае неоднозначного выделения признаков.

$$P_j^*(\omega) = p_j^* a_j(\omega) + \overline{p}_j^* \overline{a}_j(\omega)$$

— вероятность того, что образец ω принадлежит классу C^* по признаку a_j.

$$P^*(\omega) = \prod_{j=1}^{k} P_j^*(\omega)$$

— вероятность того, что образец ω принадлежит классу C^*.

По нашему предположению, если вероятность принадлежности образца ω классам $\{C_i\}$ не существенно выше вероятности принадлежности классу C^*, то считаем, что образец не принадлежит ни одному из классов $\{C_i\}$. Исходя из этого, определим классификатор следующим образом:

$$f_p^*(\omega) = \left(\frac{P_1}{|P|}, \ldots, \frac{P_N}{|P|}, \frac{cP^*}{|P|} \right), \quad \text{где} \quad |P| = \sqrt{P_1^2 + \ldots + P_N^2 + (cP^*)^2}.$$

Константа c определяет — насколько выше должна быть вероятность принадлежности образца некоторому классу C_i, чем классу C^* для отнесения его к классу C_i.

6. Сглаживание

Алгоритм построения первичного классификатора $f_p(\omega)$, описанный в пункте 3 предполагает наличие довольно большой и репрезентативной обучающей выборки для того, чтобы правильно оценить вероятности P_{ij} и P_j^*. В случае, если обучающая выборка небольшая, вероятности $P_{ij}(\omega) = p_{ij} a_j(\omega) + \overline{p}_{ij} \overline{a}_j(\omega)$ часто оказываются нулевыми,

так как $|W_{ij}^{+}| = 0$, и как следствие этого $P_i = \prod_j P_{ij}$ тоже оказывается нулевой. В этих случаях предлагается использовать технику сглаживания для переопределения функции p_{ij} и тем самым компенсировать недостаток данных в обучающей выборке.

Рассмотрим основные случаи, в которых необходимо использовать технику сглаживания:

1. Признак a_j присутствует в образце, но в процессе обучения данный признак не встретился ни в одном образце $\omega \in W_i$ принадлежащему классу C_i, тогда $p_{ij} = 0$, $\overline{a}_j(\omega) = 0$, из чего следует, что $P_{ij} = 0$. В реальности, вероятность того, что признак может встретиться образце класса C_i может оказаться больше 0 из-за того, что выборка была не достаточно репрезентативна. Определим вероятность p_{ij} следующим образом:
$$p_{ij} = \frac{p_j^*}{\max\{a|W_i|, b\}},$$
где a и b — некоторые константы, p_j^* — вероятность появления признака в образце, $|W_i|$ — число образцов ω, отнесенных к классу C_i. Иными словами, вероятность того, что признак a_j встречается среди образцов класса C_i в константу раз меньше, чем средняя вероятность встретить признак a_j в произвольном образце. Константа пропорциональна количеству образцов в обучающей выборке, соответствующих C_i, и в то же время ограничена снизу константой b. Если выборка репрезентативна, то есть в ней имеется достаточно большое количество образцов, отнесенных к классу C_i, то $p_{ij} \to 0$ при $|W_i| \to \infty$.
2. Признак a_j не найден в образце ω, но в процессе обучения данный признак встречался во всех образцах $\omega \in W_i$ класса C_i. В данном случае $\overline{p}_{ij} = 0$, $a_j(\omega) = 0$, из чего следует, что $P_{ij} = 0$. В действительности вероятность того, что признак встречается в классе C_i может быть меньше 1 из-за недостаточного объема обучающих данных. В этом случае переопределим вероятности p_{ij} следующим образом:
$$p_{ij} = \frac{|W_{ij}^{+}|}{|W_i| + \delta},$$
где $|W_{ij}^{+}|$ — число образцов, соответствующих классу C_i, в которых найден признак a_j, $|W_i|$ — число образцов ω_{ij}, отнесенных к классу C_i, δ — константа. Другими словами, мы добавляем к множеству W_i некоторое количество, не обязательно целое, образцов, в которых признак a_j не встретился. При этом p_{ij} будет строго меньше 1. Если выборка репрезентативна, то есть в ней имеется достаточно большое количество образцов, отнесенных к классу C_i, в этом случае p_{ij} очень близка к 1, иначе p_{ij} может существенно отличаться от 1.

Таблица 3

Классификация с техникой сглаживания

935 образцов	$f_p(\omega)$	$f_s(\omega)$
Классифицирован правильно, причем как наилучший вариант	694 (74,2 %)	723 (77,3 %)
Содержится среди предложенных вариантов	744 (79,6 %)	779 (83,3 %)
Классифицирован не правильно	191 (20,4 %)	156 (16,7 %)

Обозначим классификатор, построенный аналогично первичному классификатору $f_p(\omega)$, использующий сглаживание, то есть с переопределенными функциями p_{ij} как $f_s(\omega)$, и $f_s^*(\omega)$ соответственно в случаях с использованием и без использования неопределенного класса.

Такое переопределение функций p_{ij} позволяет использовать классификатор для задач с недостаточно полной обучающей выборкой. В табл. 3 приведен результат использования техники сглаживания при классификации учебных заведений города Москвы. Сравнение проводилось на тестовой выборке из 935 анкет. Обучающая выборка состояла их 9068 образцов относящихся к 2131 учебным заведениям. Оба классификатора используют функции $a_j^P(\omega)$ выделения признаков.

Из таблицы видно, что использование сглаживания дает улучшение порядка 3–4 %, при этом количество ошибок уменьшается примерно на 20 %. Данный эксперимент проводился на множестве всех 2131 учебных заведений. Для каждого учебного заведения присутствовали, по крайней мере, 2 варианта наименования — полное и сокращенное, которые были взяты не из реальных анкет, а из специальной базы данных.

7. Проблема зависимости признаков

В пункте 2 в качестве признаков $A = \{a_j\}$ мы использовали множество всех слов, встретившихся в образцах обучающего множества. Такой выбор признаков, конечно, не лишен недостатков. Перечислим их:

1. При построении классификатора ESX были использованы функции

$$P_j(\omega) = \prod_{j=1}^{k} P_{ij}(\omega),$$

определяющие вероятность принадлежности образца ω классу C_i. Используя произведение, мы подразумеваем, что признаки a_j — независимые. Это, конечно же, не так. Часто кортеж слов — несколько

слов, следующих друг за другом, представляют одно понятие. В этом случае целесообразно рассматривать кортежи как один признак.

2. Подход не учитывает синонимы слов — различные слова могут представлять один и тот же признак. Для построения хорошего классификатора необходимо строить классы эквивалентности слов, соответствующих определенному признаку.

Учитывая указанные недостатки, была проведена следующая доработка классификатора в части выделения признаков.

Как уже было сказано, определение вероятностей P_i, как

$$P_j(\omega) = \prod_{j=1}^{k} P_{ij}(\omega) \tag{1}$$

не совсем корректно из-за того, что признаки a_j могут быть зависимы. В этом случае

$$P_i(\omega) = \sum_{(\sigma_1,\ldots,\sigma_k)\in 2^k} P\left(a_1^{\sigma_1} \in C_i, \ldots, a_k^{\sigma_k} \in C_i\right) P(a_1^{\sigma_1} \in \omega) \ldots P(a_k^{\sigma_k} \in \omega), \tag{2}$$

где $(\sigma_1, \ldots, \sigma_k)$ — наборы из 0 и 1, $a_j^1 \in C_i$ — означает, что $a_j \in C_i$, а $a_j^0 \in C_i$, что $a_j \notin C_i$.

В данной формуле подразумевается, что вероятность выделения одного признака из образца не зависит от вероятности выделения прочих признаков, т. е.

$$P\left(a_1^{\sigma_1} \in \omega, \ldots, a_k^{\sigma_k} \in \omega\right) = P(a_1^{\sigma_1} \in \omega) \ldots P(a_k^{\sigma_k} \in \omega).$$

По формуле условной вероятности

$$P\left(a_1^{\sigma_1} \in C_i, \ldots, a_k^{\sigma_k} \in C_i\right) = P\left(a_k^{\sigma_k} \in C_i | a_{k-1}^{\sigma_{k-1}} \in C_i, \ldots, a_1^{\sigma_1} \in C_i\right) \times$$
$$\times \ldots \cdot P\left(a_{k-1}^{\sigma_{k-1}} \in C_i | a_{k-2}^{\sigma_{k-2}} \in C_i, \ldots, a_1^{\sigma_1} \in C_i\right) \cdot \ldots \cdot P(a_1^{\sigma_1} \in C_i). \tag{3}$$

В случае независимости признаков a_j формула (3) принимает вид

$$P\left(a_1^{\sigma_1} \in C_i, \ldots, a_k^{\sigma_k} \in C_i\right) = P(a_k^{\sigma_k} \in C_i) \cdot \ldots \cdot P(a_1^{\sigma_1} \in C_i),$$

а вероятности $P_i(\omega)$, определенные при помощи формулы (2) совпадают с определением $P_i(\omega)$ при помощи (1).

Если признаки зависимы, то для правильного расчета вероятностей $P_i(\omega)$ требуется знать условные вероятности

$$P\left(a_j^{\sigma_j} \in C_i \mid a_{j-1}^{\sigma_{j-1}} \in C_i, \ldots, a_1^{\sigma_1} \in C_i\right),$$

что, как правило, практически не представляется возможным. Предлагается выделить только самые зависимые группы признаков, то есть

те признаки, которые практически всегда встречаются вместе. После выделения групп зависимых признаков в произведении (3) пропадут соответствующие члены. Остальные признаки, как и прежде, считаем независимыми. Если в задаче встречаются в основном слова либо независимые, либо сильно зависимые, то после этапа выделения слов, необходимо статистически определить словарные зависимости, после чего наиболее зависимые группы слов (словосочетания) объединить в отдельные признаки, при этом, удалив из множества признаков те признаки, которые соответствуют отдельным словам из словосочетаний.

Наряду с механизмом выделения словосочетаний, в алгоритм была добавлена возможность работы с классами эквивалентности признаков. Классы эквивалентности выделялись автоматически по следующей схеме.

1. *Отыскание эквивалентных словосочетаний.* Рассматривались все пары образцов (ω_1, ω_2) из обучающей выборки соответствующих одному классу C_i. В том случае если образцы различались в некотором небольшом подмножестве слов (например, 1–2 слова), эти подмножества объявлялись эквивалентными словосочетаниями.
2. *Определение частоты встречаемости* (*веса*) *словосочетаний*. Для каждого словосочетания определяем все образцы, которые его содержат, и тем самым получаем частоту встречаемости словосочетания.
3. *Определение вероятности замены словосочетаний эквивалентными.* Когда словосочетания выделены, определим — в каком количестве классов W_i наблюдалось данное словосочетание (обозначим N^t) и в каком количестве классов оно имело свой эквивалент (обозначим N^t_+). Тогда вероятность того, что данное словосочетание заменимо эквивалентом равна $P^t = N^t_+/N^t$.
4. *Селекция возможных замен словосочетаний*. Пусть возможен переход словосочетания s_1 в словосочетание s_2. Из полученного списка возможных замен отбираем только те, у которых вес словосочетания s_2 больше веса s_1 не менее чем в c_1 раз, т. е. $w(s_2)/w(s_1) > c_1$, а также вероятность замены данного словосочетания на эквивалентное $P^t > c_2$. Вес словосочетания $w(s)$ означает количество классов, в которых наблюдалось данное словосочетание. Селекция необходима для сокращения времени перебора.

Классы эквивалентности также искусственно были дополнены сокращениями слов и некоторыми другими трансформациями. При этом учитывалась возможность неоднозначного восстановления слова по сокращению, например, сокращение «об.» соответствует словам «общество» и «объединение».

В табл. 4 приведены результаты сравнения классификатора с учетом зависимости и эквивалентности признаков. Из таблицы видно, что учет

Таблица 4

Классификация с учетом зависимых и эквивалентных признаков

935 образцов	$f_s(\omega)$	$f_{se}(\omega)$
Классифицирован правильно, причем как наилучший вариант	723 (77,3 %)	758 (81,1 %)
Содержится среди предложенных вариантов	779 (83,3 %)	834 (89,2 %)
Классифицирован не правильно	156 (16,7 %)	101 (10,8 %)

зависимости признаков и использование классов эквивалентности дает существенное улучшение, количество ошибок сокращается примерно на треть.

8. Реализация и выводы

Классификатор был реализован в рамках системы массового ввода документов Cognitive Forms [12]. В 1999–2001 годах классификатор был успешно применен при вводе и обработке анкет школьников и студентов в Москве с целью регистрации категорий лиц, пользующихся льготами при проезде в метрополитене. При вводе анкет классифицировались средние и высшие учебные заведения города Москвы. Также классификатор использовался при обработке поля «*Наименование товара*» в проекте по вводу отгрузочных разнарядок для ОАО «Сибнефть». Всего при использовании классификатора ESX было введено более 2 000 000 документов.

Классификатор ESX показал высокое качество на тестовом наборе документов. Для поля «*Наименование учебного заведения*» оно составило около 90 %. Серьезным плюсом использования данного классификатора является наличие полностью автоматической процедуры обучения, также имеется возможность динамически дообучать классификатор в процессе работы. Использование техники сглаживания дает возможность обучать классификатор на недостаточно репрезентативной выборке. Также стоит отметить, что существенное улучшение качества работы дает процедура поиска зависимых признаков и построения классов эквивалентности словосочетаний. Кроме того, процедура селекции замен словосочетаний с учетом их веса позволяет использовать зашумленную обучающую выборку. В этом случае допустимы только замены неправильно распознанных фрагментов на распознанные правильно, так как вес последних больше.

Следует также отметить, что реализация классификатора в виде отдельных компонент допускает использование нестандартных функций выделения признаков из образца и функций построения набора признаков по обучающей выборке для решения специальных задач.

Литература

1. *Bouchaffra D.*, *Govindaraju V.*, *Srihari S.* A Methodology for Determining Probability of Correctness of Recognizer Scores // Proc. IEEE Conf. Computer Vision and Pattern Recognition, Santa Barbara, Calif., June 1998.
2. *Duda R. O.*, *Hart P. E.*, *Stork D. G.* Pattern Classification (2^{nd} ed.), John Wiley and Sons, 2001.
3. *McCallum A. K.*, *Nigam K.*, *Rennie J.*, *Seymore K.* Automating the Construction of Internet Portals with Machine Learning // J. Information Retrieval. 2000. V. 3. № 2. P. 127–163.
4. *Vassili V. Postnikov*, *Dmitry L. Sholomov.* Post-processing of OCR Results Using Automatically Constructed Partially Defined Syntax // Proc. of The International Conference on Machine Learning, Technologies and Applications. CSREA Press, June 2004, USA.
5. *Postnikov V. V.*, *Sholomov D. L.*, *Marchenko A. E.* FlexiDocs: The Template Driven Document Recognition Technology // Proceedings of the 6^{th} German—Russian Workshop on Pattern Recognition and Image Understanding (OGRW-6), 2003.
6. *Rennie J.* Improving multi-class text classification with naive bayes // Master's thesis, Massachusetts Institute of Technology. 2001.
7. *Salton G.*, *Allan J.* Selective Text Utilization and Text Traversal // Hypertext'93 Proceedings, November 14–18, 1993, Seattle, Washington, USA.
8. *Sebastiani F.* Machine learning in automated text categorisation: a survey // Pisa. IT. 1999.
9. *Sholomov D. L.* Syntactical Approach to Post-Processing of Fuzzy recognized Text // Proc. of The International Conference on Machine Learning, Technologies and Applications. CSREA Press, June 2003, USA. P. 115–121.
10. *Sholomov D. L.* Interpreting the Indistinctly Recognized Textual Constructions // Pattern Recognition and Image Analysis. 2003. V. 13. № 2. P. 353–355.
11. *Андреев А. М.*, *Березкин Д. В.*, *Морозов В. В.*, *Симаков К. В.* Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа // Интелтек изд-во, 2005.
12. *Арлазаров В. В.*, *Постников В. В.*, *Шоломов Д. Л.* Cognitive Forms — система массового ввода структурированных документов // Сб. «Управление информационными потоками», Москва, УРСС, 2002. С. 35–46.
13. *Мерков А. Б.* Основные методы, применяемые для распознавания рукописного текста. http://fornit2005.narod.ru/papers/methods.ps.
14. *Мерков А. Б.* О статистическом обучении // http://www.recognition.mccme.ru/pub/RecognitionLab.html/slt.pdf.

15. *Постников В. В.*, *Марченко А. Е.*, *Шоломов Д. Л.* Разбор структурированного документа в модели с нечеткой логикой // Сборник трудов ИСА РАН «Документооборот. Концепции и инструментарий.», Москва, УРСС, 2004. С. 71–82.
16. *Постников В. В.* Автоматическая идентификация и распознавание структурированных документов // Дис. ... к. т. н. М., 2001.
17. *Хайкин С.* Нейронные сети, полный курс. М.: Вильямс, 2005.
18. *Шоломов Д. Л.* Интерпретация нечетко распознанных текстовых конструкций // Сборник трудов 6-й Международной конференции «Распознавание образов и анализ изображений: новые информационные технологии». Великий Новгород, 2002.
19. *Шоломов Д. Л.* Синтаксический подход к пост-обработке нечетко распознанного текста // Сборник трудов ИСА РАН «Документооборот. Концепции и инструментарий». М.: УРСС, 2004. С. 193–207.
20. *Шоломов Д. Л.*, *Постников В. В.*, *Марченко А. А.*, *Усков А. В.* Пост-обработка результатов OCR распознавания использующая частично определенный синтаксис // Сборник трудов ИСА РАН «Интеллектуальные информационные технологии. Концепции и инструментарий». Т. 16. М.: КомКнига/URSS, 2005. С. 146–165.