НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2017 Том 17 № 1 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2017 Vol. 17 No 1 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 519.765

ИСПОЛЬЗОВАНИЕ ВЕРОЯТНОСТНОГО РАСПРЕДЕЛЕНИЯ НАД МНОЖЕСТВОМ КЛАССОВ В ЗАДАЧЕ КЛАССИФИКАЦИИ АРАБСКИХ ДИАЛЕКТОВ

О.В. Дурандин[a,b], Н.Р. Хилал[a,c], Д.Ю. Стребков[c], Н.Ю. Золотых[a]

[a] Нижегородский университет им. Н.И. Лобачевского, Нижний Новгород, 603950, Российская Федерация
[b] Национальный исследовательский университет «Высшая школа экономики», Нижний Новгород, 603155, Российская Федерация
[c] ООО «Диктум», Нижний Новгород, 603070, Российская Федерация
Адрес для переписки: oleg.durandin@gmail.com

Информация о статье
Поступила в редакцию 25.11.16, принята к печати 29.12.16
doi: 10.17586/2226-1494-2017-17-1-110-116
Язык статьи – русский

Ссылка для цитирования: Дурандин О.В., Хилал Н.Р., Стребков Д.Ю., Золотых Н.Ю. Использование вероятностного распределения над множеством классов в задаче классификации арабских диалектов // Научно-технический вестник информационных технологий, механики и оптики. 2017. Т. 17. № 1. С. 110–116. doi: 10.17586/2226-1494-2017-17-1-110-116

Аннотация

Предмет исследования. Предложен подход к решению задачи классификации, использующий информацию о распределении вероятностей на множестве классов в обучающей выборке. Алгоритм проиллюстрирован на одной из сложных задач автоматической обработки текстов на естественном языке – классификации арабских диалектов. **Метод.** Каждому объекту обучающей выборки сопоставляется распределение вероятностей над метками классов, вместо сопоставления единственной метки класса. Предлагаемый подход решает задачу с учетом распределения вероятностей над множеством классов для повышения качественных показателей работы классификатора. **Основные результаты.** Предложенный подход проиллюстрирован на примере задачи классификации арабских диалектов. Анализируемые данные, содержащие слова-метки, получены из социальной сети Twitter, относящиеся к шести арабским диалектам: саудовский, левантийский, алжирский, египетский, иракский, иорданский; использованы также сообщения на современном стандартном арабском языке (MSA). Показан рост качества классификации при учете вероятностного распределения над множеством классов в обучающей выборке. Показано, что даже относительно простой учет вероятностного распределения увеличивает точность предсказания с 44% до 67%. **Практическая значимость.** Предложенный подход и соответствующий алгоритм могут найти применение в случае, когда разметка данных экспертом требует значительных временных и финансовых ресурсов, но имеется возможность разработки эвристических правил. Реализация предложенного алгоритма позволит снизить затраты при подготовке данных без значительной потери точности классификации.

Ключевые слова
задача классификации, многоклассовая классификация, автоматическая аннотация, арабский диалект, классификация диалектов, меточный шум

PROBABILITY DISTRIBUTION OVER THE SET OF CLASSES IN ARABIC DIALECT CLASSIFICATION TASK

O.V. Durandin[a,b], N.R. Hilal[a,c], D.Y. Strebkov[c], N.Y. Zolotykh[a]

[a] Lobachevsky State University of Nizhni Novgorod (UNN), Nizhny Novgorod, 603950, Russian Federation
[b] Higher School of Economics National Research University, Nizhny Novgorod, 603155, Russian Federation
[c] "Dictum" Ltd., Nizhny Novgorod, 603070, Russian Federation
Corresponding author: oleg.durandin@gmail.com

Article info
Received 25.11.16, accepted 29.12.16
doi: 10.17586/2226-1494-2017-17-1-110-116
Article in Russian

For citation: Durandin O.V., Hilal N.R., Strebkov D.Y., Zolotykh N.Y. Probability distribution over the set of classes in Arabic dialect classification task. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2017, vol. 17, no. 1, pp. 110–116. doi: 10.17586/2226-1494-2017-17-1-110-116

Abstract

Subject of Research. We propose an approach for solving machine learning classification problem that uses the information about the probability distribution on the training data class label set. The algorithm is illustrated on a complex natural language processing task - classification of Arabic dialects. **Method.** Each object in the training set is associated with a

probability distribution over the class label set instead of a particular class label. The proposed approach solves the classification problem taking into account the probability distribution over the class label set to improve the quality of the built classifier. **Main Results.** The suggested approach is illustrated on the automatic Arabic dialects classification example. Mined from the Twitter social network, the analyzed data contain word-marks and belong to the following six Arabic dialects: Saudi, Levantine, Algerian, Egyptian, Iraq, Jordan, and to the modern standard Arabic (MSA). The paper results demonstrate an increase of the quality of the built classifier achieved by taking into account probability distributions over the set of classes. Experiments carried out show that even relatively naive accounting of the probability distributions improves the precision of the classifier from 44% to 67%. **Practical Relevance.** Our approach and corresponding algorithm could be effectively used in situations when a manual annotation process performed by experts is connected with significant financial and time resources, but it is possible to create a system of heuristic rules. The implementation of the proposed algorithm enables to decrease significantly the data preparation expenses without substantial losses in the precision of the classification.

Keywords
classification task, multiclass classification, automatic annotation, Arabic dialects, classification of dialects, label noise, class noise

Введение

Одной из проблем машинного обучения является решение задачи классификации. Имеется множество объектов, разделенных на классы. Кроме этого, задана обучающая выборка – множество размеченных (аннотированных) объектов, для каждого из которых известно, к какому классу он относится. Классовая принадлежность остальных объектов неизвестна. Требуется построить алгоритм, способный классифицировать (допуская, возможно, некоторые ошибки) произвольный объект из исходного множества. Как правило, алгоритмы для решения задачи классификации требуют большого объема размеченных данных. В большинстве случаев разметка выполняется экспертом и требует значительных временных и финансовых затрат. В то же время использование различных интернет-ресурсов может помочь получить значительные объемы данных со сравнительно малыми затратами. Зачастую возможно разработать ряд эвристических правил, которые смогут классифицировать собранные данные с определенной точностью. Разумеется, подобная классификация не сможет демонстрировать 100% точности, как это было бы в случае аннотирования экспертом.

Таким образом, можно говорить, что в ряде задач появляется меточный шум (label noise) [1, 2]: объекту обучающей выборки соответствует метка класса, отличающаяся от истинной метки класса, к которому в действительности принадлежит объект. В настоящей работе предпринята попытка рассмотреть этот шум и предоставить классификатору дополнительную информацию – распределение вероятностей над множеством классов для каждого объекта обучающей выборки. Эта вспомогательная информация может интерпретироваться как степень уверенности в метке класса, ассоциированной с объектом из обучающей выборки.

Итак, предположим, что каждому объекту i поставлен в соответствие вектор признаков x^i и вектор вероятностей его отнесения к соответствующему классу $(p_1^i, p_2^i, \ldots, p_K^i)$, где p_k^i – вероятность того, что объект i принадлежит классу k ($k = 1, 2, \ldots, K$).

Такая постановка задачи рассматривалась в [3], где для ее решения предлагается подход нечетких (fuzzy) множеств. В [4] к решению данной задачи применялся аппарат теории Демпстера–Шафера (Evidence theory). Та же задача была рассмотрена в [5] для решения проблемы оценки параметров в статистических моделях. Автор [5] предложил метод, базирующийся на максимизации обобщенного критерия правдоподобия (maximization of generalized likelihood criterion), который может быть интерпретирован как степень соответствия между статистической моделью и неопределенным наблюдением (uncertain observation). Этот метод был применен к данным, полученным в результате кластеризации конечных смешанных моделей (finite mixture models).

Здесь мы предлагаем иной подход, основанный на замене метки класса с соответствующей вероятностью (из соответствующего вероятностного распределения). В выполненных экспериментах предлагаемый алгоритм улучшает качество классификации.

Предложенный подход иллюстрируется на нетривиальной задаче – классификации арабских диалектов. Были использованы данные, собранные из социальной сети Twitter и автоматически аннотированные посредством эвристических правил. Эти правила не работают идеально: мы имеем дело с естественным языком и «живыми» данными социальных сетей.

Постановка задачи классификации с использованием вероятностного распределения на множестве классов

Пусть X – множество описаний объектов, Y – множество ответов. Кроме того, существует неизвестная целевая зависимость – отображение $f^*: X \to Y$, значения которой известны только на объектах обучающей выборки $\{(x^1, y^1), (x^2, y^2), \ldots, (x^N, y^N)\}$, где $x^i \in X$, $y^i \in Y$ ($i = 1, 2, \ldots, N$).

Задача обучения с учителем может быть сформулирована следующим образом: требуется восстановить зависимость f^*, т. е. построить решающую функцию $f: X \to Y$, которая аппроксимирует целевую

зависимость f^* не только на объектах обучающей выборки, но и на всем множестве объектов X [6]. Если $Y = \{1, 2, ..., K\}$, то говорят о задаче классификации с K классами. В этом случае множество X разбивается на классы $L_y = \{x \in X : f^*(x) = y\}$ ($y = 1, 2, \ldots, K$), и функция $f(x)$ возвращает номер класса y.

Пусть рассматривается задача классификации на K классов. Сделаем следующее предположение: обучающая выборка представлена как множество

$$T_{\text{train}} = \{(x^i, (p_1^i, p_2^i, \ldots, p_K^i)) : i = 1, 2, ..., N\}, \; p_k^i \geq 0, \; \sum_{k=1}^{K} p_k^i = 1 \; (i = 1, 2, \ldots, N),$$

где p_k^i – априорная вероятность того, что объект i принадлежит классу k. Проблема остается той же: построить решающую функцию (классификатор) $f: X \to Y$, которая дает аппроксимацию целевой функции f^*.

Задача автоматической классификации арабских диалектов

Словосочетание «Арабский язык» включает в себя различные варианты одного языка [7, 8]. Это может быть:

- классический арабский (язык Корана);
- MSA – современный стандартный арабский или официальный арабский;
- диалектный/живой арабский, который, в свою очередь, имеет несколько модификаций.

Диалекты арабского языка можно классифицировать по двум основным признакам – территориальному и социальному, при этом на социальное деление диалектов накладывается и географический фактор. Из этого следует, что в каждой арабской местности/городе/стране используется свой особый диалект арабского языка.

Если сравнивать диалекты с классическим арабским языком, то можно заметить, что диалекты Аравийского полуострова и более северных арабских стран более схожи с классическим арабским, чем диалекты стран северной Африки. При сравнении диалектов между собой можно обнаружить множество различий как с точки зрения фонетики, так и в морфологии и синтаксической структуре речи для разных диалектов [9, 10].

Если посмотреть, из чего состоит диалектная речь, то обнаруживается наличие в ней совокупности слов из классического арабского, слов, принадлежащих сразу к нескольким диалектам, а также наличие уникальных диалектных слов – слов-меток, дифференцирующих один конкретный диалект.

Используя слова-метки, можно вручную или автоматически классифицировать любой арабский текст. Например, если в арабском тексте не обнаруживается ни одного слова-метки, то, скорее всего, этот текст написан на классическом арабском языке. В случае обнаружения одного или нескольких слов-меток, принадлежащих к некоторому диалекту, скорее всего, и весь текст написан на этом диалекте.

В качестве слов-меток мы использовали самые частотные слова, т. е. те, вероятность присутствия которых в тексте очень высока. Например, большую часть слов-меток составляли местоимения, отрицательные и вопросительные частицы, а также диалектная лексика, не имеющая общих корней с синонимами из классического арабского языка (таблица).

Слово-метка	Эквивалент в классическом арабском	Грамматическое значение	Диалект
مش	ليس	отрицательная частица	левантийский, иорданский
شو	ماذا	вопросительная частица	левантийский
بدي	أريد	глагол	левантийский
عايزة	أريد	глагол	египетский
برضو	أيضا	наречие	египетский, левантийский
الحين	الآن	наречие	саудовский
نتوما	أنتم	местоимения	алжирский

Таблица. Примеры слов-меток

Очень часто встречаются тексты, в которых находятся несколько слов-меток, принадлежащим к двум или трем разным диалектам. Это затрудняет определение типа диалекта такого текста. Кроме того, еще сильнее усложняет задачу тот факт, что может встречаться текст, в котором отсутствуют какие-либо слова-метки, и в данном случае определить диалект возможно лишь по стилю, синтаксису и последовательности слов в предложении.

Достаточно хорошо справляются с трудностями задачи определения диалектов статистические подходы [10].

Описание анализируемых данных

Из социальной сети Twitter были собраны твиты, содержащие слова-метки. Эти слова-метки относятся к следующим шести арабским диалектам: саудовский, левантийский, алжирский, египетский, иракский, иорданский. Эти диалекты были выбраны как самые популярные диалекты в арабских странах. В дополнение к этим классам для экспериментов были добавлены сообщения седьмого класса – на современном стандартном арабском языке (MSA).

Твиты собирались в 2015–2016 г.г. Для части собранных твитов (4 156 837) диалекты были определены с помощью эвристических правил на основе анализа слов-меток, при этом были исключены твиты, содержащие слова-метки, относящиеся к разным диалектам. Для другой части (54 006) диалект был определен лингвистами-арабистами. Первый набор данных – автоматически размеченную выборку – будем обозначать $Tweets_{emp}$, второй набор – выборку, размеченную лингвистами, – $Tweets_{gold}$. Распределение твитов по диалектам для каждой из этих выборок представлено на рисунке.

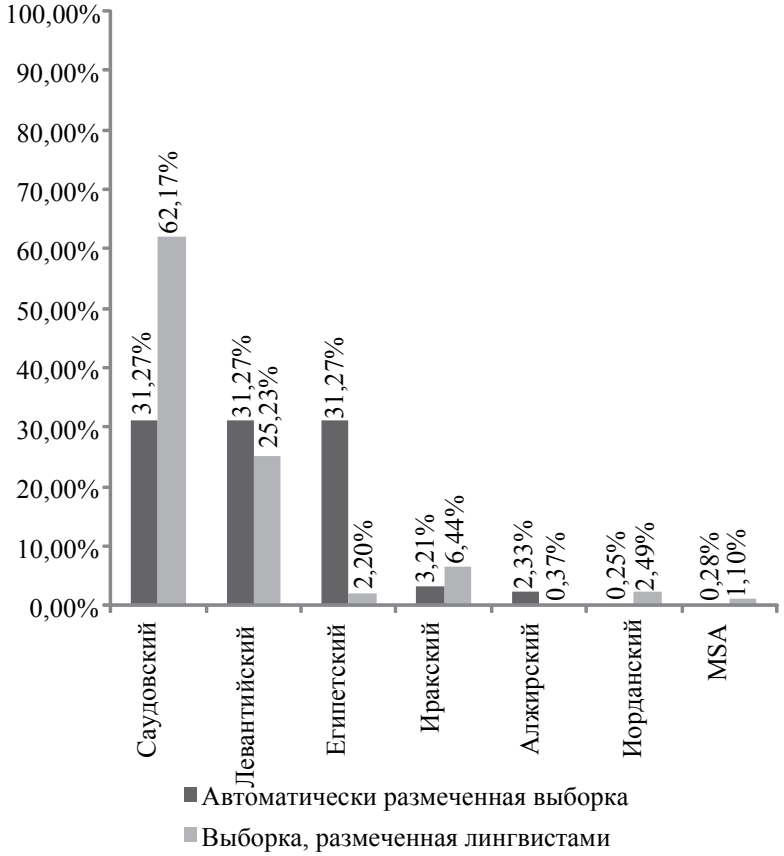

Рисунок. Распределение по классам в выборках

Заметим, что в автоматически размеченной выборке саудовский, египетский и левантийский диалекты распределены примерно поровну, в то время как оставшиеся классы содержат намного меньше твитов. Это обусловлено популярностью соответствующих диалектов в арабском сегменте социальных сетей. Несбалансированность между частями различных диалектов в выборке, размеченной лингвистами, обусловлена тем, что процесс разметки продолжается. Заметим, что в результате этого оценка качества классификации может быть смещенной.

Вычисление вероятностного распределения над множеством классов

Обозначим:

- $d_{k,\text{gold}}$ – событие, заключающееся в том, что твит принадлежит диалекту k;
- $d_{k,\text{emp}}$ – событие, заключающееся в том, что твит помечен эвристическим правилом как твит, принадлежащий диалекту k.

Тогда $P(d_{k,\text{gold}} \mid d_{k,\text{emp}})$ – условная вероятность того, что твит, помеченный эвристическим правилом как диалект k, действительно написан на этом диалекте. Следует отметить, что приведенная вероятность неизвестна, однако можно положить

$$P(d_{k,\text{emp}} \mid d_{k,\text{gold}}) = M_k / N_k,$$

где M_k – число твитов в $Tweets_{\text{gold}}$, помеченных как диалект k как лингвистами, так и с помощью эвристического правила; N_k – число твитов в $Tweets_{\text{gold}}$, помеченных как диалект k лингвистами. По формуле Байеса

$$P(d_{k,\text{gold}} \mid d_{k,\text{emp}}) = P(d_{k,\text{gold}})\, P(d_{k,\text{emp}} \mid d_{k,\text{gold}}) / P(d_{k,\text{emp}}),$$

где $P(d_{k,\text{gold}})$, $P(d_{k,\text{emp}})$ оцениваются по выборкам $Tweets_{\text{gold}}$ и $Tweets_{\text{emp}}$ соответственно.

Для каждого объекта i из $Tweets_{\text{emp}}$ находим соответствующие вероятности $\{(x^i, (p_1^i, p_2^i, \ldots, p_K^i)) : i = 1, 2, ..., N\}$, где K – общее число классов (в рассматриваемом случае 7). Для объекта i из $Tweets_{\text{gold}}$ полагаем $p_k^i = 1$, если твит i размечен лингвистами как твит, принадлежащий диалекту k, и $p_k^i = 0$ в противном случае.

Простой метод учета вероятностного распределения

Для рассматриваемой задачи стандартные алгоритмы классификации без учета вероятностного распределения на множестве классов показывают неудовлетворительные результаты.

В качестве признаков в задаче классификации используются символьные n-граммы (биграммы и триграммы) [11] и словарь слов-меток. Символьная n-грамма – это подстрока из n символов более длинной строки. Модели n-грамм базируются на систематической сборке и подсчете n-грамм с помощью «скользящего окна» длины n над корпусом документов [12]. В разных арабских диалектах используются разные механизмы словообразования (например, могут отличаться механизмы аффиксации), чем и обусловлена достаточная эффективность использования n-грамм в рассматриваемой задаче [9, 10].

В качестве обучающей выборки возьмем ⅓ данных (полученных в результате случайного изъятия) из $Tweets_{\text{gold}}$ и все данные из $Tweets_{\text{emp}}$, в качестве тестовой – оставшиеся ⅔ данных из $Tweets_{\text{gold}}$. В качестве метода построения классификатора будем использовать случайный лес (Random Forest) [13] – один из самых популярных алгоритмов, показывающий хорошие результаты на многих задачах.

В качестве показателей эффективности многоклассовой классификации были использованы взвешенные показатели точности P, полноты R и F_1-меры, определяемые по следующим формулам [14]:

$$P(I_l, \hat{I}_l) = \frac{1}{\sum_{l \in L} |I_l|} \sum_{l \in L} |I_l| \frac{|I_l \cap \hat{I}_l|}{|\hat{I}_l|}, \quad R(I_l, \hat{y}_l) = \frac{1}{\sum_{l \in L} |I_l|} \sum_{l \in L} |I_l| \frac{|I_l \cap \hat{I}_l|}{|I_l|},$$

$$F_1(I_l, \hat{I}_l) = 2 \frac{P(I_l, \hat{I}_l) \times R(I_l, \hat{I}_l)}{P(I_l, \hat{I}_l) + R(I_l, \hat{I}_l)}, \qquad F_1 = \frac{1}{\sum_{l \in L} |I_l|} \sum_{l \in L} |I_l| \times F_1(I_l, \hat{I}_l),$$

где $L = \{1, 2, \ldots, K\}$ – множество меток классов; y_l – множество объектов класса; $\hat{y}_l$ – множество объектов, отнесенных алгоритмом к классу l.

Алгоритм случайного леса без использования вероятностных распределений показал невысокое качество построенного классификатора: на тестовой выборке точность (precision) = 0,44, F_1-мера = 0,42. В целом такие результаты ожидаемы, поскольку стандартные алгоритмы классификации показывают плохие результаты на зашумленных данных [15].

Воспользуемся теперь следующим простым методом учета вероятностного распределения. Для каждого твита i из $Tweets_{\text{emp}}$ сгенерируем новую метку класса случайно в соответствии с вычисленным вероятностным распределением $(p_1^i, p_2^i, \ldots, p_K^i)$. Уже такой простой способ учета вероятностного распределения позволяет заметно повысить точность предсказания: алгоритм случайного леса показывает на тестовой выборке точность (precision) $P = 0{,}57$, F_1-меру $F_1 = 0{,}59$.

Алгоритм классификации на основе вероятностного распределения

Мы предлагаем следующий метод учета вероятностного распределения на множестве классов. Как и выше, предполагаем, что обучающая выборка T_{train} представляет собой набор объектов и вероятностные распределения над классами для каждого объекта обучающей выборки: $T_{\text{train}} = \{(x^i, (p_1^i, p_2^i, \ldots, p_K^i)) : i = 1, 2, ..., N\}$.

Метод строит ансамбль из M базовых классификаторов. Решающая функция представляет собой функцию голосования от всех построенных классификаторов. Псевдокод алгоритма приведен ниже.

Алгоритм 1. Генерация ансамбля классификаторов

1. Вход: $T_{\text{train}} = \{(x^i, (p_1^i, p_2^i, \ldots, p_K^i)) : i = 1, 2, ..., N\}$
2. M – количество базовых классификаторов
3. Выход: ансамбль обученных базовых классификаторов
4. *estimator_list* = {}
5. **for** $m = 1, 2, \ldots, M$:
6. $vector\_of\_classes_m = generate\_classes\_by\_prob\_distribution()$
7. $current\_estimator_m = learn\_base\_estimator(X_{\text{train}}, vector\_of\_classes_m)$

8. *estimator_list* = *estimators_list* ∪{*current_estimator*$_m$}
9. **return** *estimator_list*

При построении каждого базового классификатора используется обучающая выборка, в которой метка класса для объекта i генерируется случайно в соответствии с вычисленным вероятностным распределением $(p_1^i, p_2^i, \ldots, p_K^i)$. Псевдокод генерации меток классов *generate_classes_by_prob_distribution*() проиллюстрирован в алгоритме 2. Функция *learn_base_estimator*() реализует обучение базового классификатора (например, дерева решений).

Алгоритм 2. Реализация функции *generate_classes_by_prob_distribution*()

1. Вход: $T_{\text{train}} = \{(x^i, (p_1^i, p_2^i, \ldots, p_K^i)) : i = 1, 2, ..., N\}$
2. Выход: $C_{\text{train}} = (c_1, c_2, \ldots, c_N)$ – список меток классов
3. $C_{\text{train}} = \{\}$
4. **for** $i = 1, 2, \ldots, N$:
5. в соответствии с вероятностным распределением $(p_1^i, p_2^i, \ldots, p_K^i)$ случайно выбрать индекс класса c_class_i
6. $C_{\text{train}} = C_{\text{train}} \cup \{c\_class_i\}$
7. **return** C_{train}

В результате применения данного алгоритма качество классификации удалось значительно улучшить. В качестве базовых классификаторов использовались деревья решений. Было построено $M = 35$ базовых классификаторов. На тестовой выборке получены значения точности (precision) $P = 0{,}67$ и F_1-меры $F_1 = 0{,}70$.

Заключение

В работе предложен подход к решению задачи классификации, использующий информацию о распределении вероятностей на множестве классов в обучающей выборке. Алгоритм проиллюстрирован на одной из сложных задач автоматической обработки текстов на естественном языке – классификации арабских диалектов (всего $K = 7$ классов). Получено значительное увеличение точности классификации: точность предсказания поднялась с 44% до 67%. Представленный метод может использоваться в задачах, где есть возможность использовать эвристики для разметки обучающей выборки большого объема, что позволит значительно снизить затраты при подготовке данных без значительной потери в точности при решении задачи классификации.

Литература

1. Kearns M.J., Vazirani U.V. An Introduction to Computational Learning Theory. MIT Press, 1994. 221 p.
2. Flach P. Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press, 2012. 409 p.
3. Bezdek J.C., Keller K., Krisnapuram R., Pal N. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Springer, 1999. 776 p.
4. Denoeux T., Zouhal L.M. Handling possibilistic labels in pattern classification using evidential reasoning // Fuzzy Sets and Systems. 2001. V. 122. N 3. P. 409–424. doi: 10.1016/s0165-0114(00)00086-5
5. Denoeux T. Maximum likelihood estimation from uncertain data in the belief function framework // IEEE Transactions on Knowledge and Data Engineering. 2013. V. 25. N 1. P. 119–130. doi: 10.1109/tkde.2011.201
6. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 7th ed. Springer, 2013. 745 p.
7. Durandin O., Hilal N., Strebkov D. Automatic Arabic dialect identification // Computational Linguistics and Intellectual Technologies: Proc. Int. Conf. “Dialogue 2016”. Moscow, 2016.
8. Habash N.Y. Introduction to Arabic Natural Language Processing. Toronto: Morgan & Claypool, 2010. 186 p.
9. Heintz I. Arabic language modeling with stem-derived morphemes for automatic speech recognition. Ph.D. thesis. Ohio State University, 2010. 202 p.
10. Almeman K., Lee M. Toward developing a multi-dialect morphological analyser for Arabic // Proc. 4th Int. Conf. on Arabic Language Processing. Rabat, Morocco, 2012. P. 19–25.
11. Cavnar W.B., Trenkle J.M. N-gram-based text categorization // Proc. 3rd Annual Symposium on Document Analysis and Information Retrieval. 1994. P. 161–175.

References

1. Kearns M.J., Vazirani U.V. *An Introduction to Computational Learning Theory*. MIT Press, 1994, 221 p.
2. Flach P. *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 2012, 409 p.
3. Bezdek J.C., Keller K., Krisnapuram R., Pal N. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Springer, 1999, 776 p.
4. Denoeux T., Zouhal L.M. Handling possibilistic labels in pattern classification using evidential reasoning. *Fuzzy Sets and Systems*, 2001, vol. 122, no. 3, pp. 409–424. doi: 10.1016/s0165-0114(00)00086-5
5. Denoeux T. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Transactions on Knowledge and Data Engineering*, 2013, vol. 25, no. 1, pp. 119–130. doi: 10.1109/tkde.2011.201
6. Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 7th ed. Springer, 2013, 745 p.
7. Durandin O., Hilal N., Strebkov D. Automatic Arabic dialect identification. *Computational Linguistics and Intellectual Technologies: Proc. Int. Conf. “Dialogue 2016”*. Moscow, 2016.
8. Habash N.Y. *Introduction to Arabic Natural Language Processing*. Toronto, Morgan & Claypool, 2010, 186 p.
9. Heintz I. *Arabic language modeling with stem-derived morphemes for automatic speech recognition*. Ph.D. thesis. Ohio State University, 2010, 202 p.
10. Almeman K., Lee M. Toward developing a multi-dialect morphological analyser for Arabic. *Proc. 4th Int. Conf. on Arabic Language Processing*. Rabat, Morocco, 2012, pp. 19–25.
11. Cavnar W.B., Trenkle J.M. N-gram-based text categorization. *Proc. 3rd Annual Symposium on Document Analysis and*

12. Miao Y., Keselj V., Milios E. Document clustering using character N-grams: a comparative evaluation with term-based and word-based clustering // Proc. 14th ACM Int. Conf. on Information and Knowledge Management. 2005. P. 357–358.
13. Brieman L. Random forests // Machine Learning. 2001. V. 45. N 5. P. 5–32.
14. Zhang M.L., Zhou Z.H. A review on multi-label learning algorithms // IEEE Transations on Knowledge and Data Engineering. 2014. V. 26. N 8. P. 1819–1837. doi: 10.1109/tkde.2013.39
15. Segal M.R. Machine Learning Benchmarks and Random Forests Regression. Technical Report. Univ. California, San Francisco, 2004.

Авторы

Дурандин Олег Владимирович – аспирант, Нижегородский университет им. Н.И. Лобачевского, Нижний Новгород, 603950, Российская Федерация; старший преподаватель, Национальный исследовательский университет «Высшая школа экономики», Нижний Новгород, 603155, Российская Федерация, oleg.durandin@gmail.com

Хилал Надежда Риядовна – магистр, аспирант, Нижегородский университет им. Н.И. Лобачевского, Нижний Новгород, 603950, Российская Федерация; руководитель проекта – лингвист арабист, ООО «Диктум», Нижний Новгород, 603070, Российская Федерация, nadia.hilal@hotmail.com

Стребков Дмитрий Юрьевич – программист, ООО «Диктум», Нижний Новгород, 603070, Российская Федерация, dmitry.strebkov@gmail.com

Золотых Николай Юрьевич – доктор физико-математических наук, доцент, профессор, Нижегородский университет им. Н.И. Лобачевского, Нижний Новгород, 603950, Российская Федерация, Nikolai.Zolotykh@gmail.com

Information Retrieval, 1994, pp. 161–175.
12. Miao Y., Keselj V., Milios E. Document clustering using character N-grams: a comparative evaluation with term-based and word-based clustering. *Proc. 14th ACM Int. Conf. on Information and Knowledge Management*, 2005, pp. 357–358.
13. Brieman L. Random forests. *Machine Learning*, 2001, vol. 45, no. 5, pp. 5–32.
14. Zhang M.L., Zhou Z.H. A review on multi-label learning algorithms. *IEEE Transations on Knowledge and Data Engineering*, 2014, vol. 26, no. 8, pp. 1819–1837. doi: 10.1109/tkde.2013.39
15. Segal M.R. *Machine Learning Benchmarks and Random Forests Regression*. Technical Report. Univ. California, San Francisco, 2004.

Authors

Oleg V. Durandin – postgraduate, Lobachevsky State University of Nizhni Novgorod (UNN), Nizhny Novgorod, 603950, Russian Federation; senior lecturer, Higher School of Economics National Research University, Nizhny Novgorod, 603155, Russian Federation, oleg.durandin@gmail.com

Nadezhda R. Hilal – postgraduate, Lobachevsky State University of Nizhni Novgorod (UNN), Nizhny Novgorod, 603950, Russian Federation; Project manager – Linguist, “Dictum” Ltd., Nizhny Novgorod, 603070, Russian Federation, nadia.hilal@hotmail.com

Dmitrii Yu. Strebkov – software engineer, “Dictum” Ltd., Nizhny Novgorod, 603070, Russian Federation, dmitry.strebkov@gmail.com

Nikolai Yu. Zolotykh – D.Sc., Associate professor, Professor, Lobachevsky State University of Nizhni Novgorod (UNN), Nizhny Novgorod, 603950, Russian Federation, Nikolai.Zolotykh@gmail.com