Харченко К.А., Батенков К.А. Разработка системы автоматической классификации видов разрешённого использования земельных участков// Международный журнал информационных технологий и энергоэффективности. – 2026. – Т. 11 № 6 (68) Ч.1 с. 105–111

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/



УДК 004.852

РАЗРАБОТКА СИСТЕМЫ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ВИДОВ РАЗРЕШЁННОГО ИСПОЛЬЗОВАНИЯ ЗЕМЕЛЬНЫХ УЧАСТКОВ

[1] Харченко К.А., Батенков К.А.
ФГБОУ ВО «МИРЭА - РОССИЙСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ», Москва, Россия (119454, г. Москва, пр-т Вернадского, д. 78, стр. 4), e-mail: [1]kirill.kharchenko.02@yandex.ru

В данной работе разработан алгоритм автоматической классификации видов разрешённого использования земельных участков на основе методов обработки естественного языка и машинного обучения. Алгоритм включает предобработку текстовых данных, приведение их к векторным представлениям и применение гибридной иерархической модели для точного определения вида разрешённого использования. Приведено строгое математическое описание метода, включающее формализацию TF-IDF векторизации, логистической регрессии, косинусного сходства с прототипами классов и итоговой решающей функции. Результаты показали accuracy = 0,82 и Weighted F1 = 0,85.

Ключевые слова: ВРИ земельных участков, иерархическая классификация, гибридная математическая модель, TF-IDF, логистическая регрессия, косинусное сходство.

DEVELOPMENT OF AN AUTOMATIC CLASSIFICATION SYSTEM FOR PERMITTED USES OF LAND PLOTS

[1] Kharchenko K.A., Batenkov K.A.
MIREA - RUSSIAN TECHNOLOGICAL UNIVERSITY, Moscow, Russia (119454, Moscow, avenue Vernadsky, 78, b. 4), e-mail: [1]kirill.kharchenko.02@yandex.ru

This paper presents an algorithm for automatic classification of permitted uses of land plots based on NLP and machine learning. A rigorous mathematical framework is provided, including TF-IDF vectorization, logistic regression, cosine prototype similarity, and the hybrid decision function. Experimental results demonstrated accuracy = 0.82 and Weighted F1 = 0.85.

Keywords: Permitted land uses, hierarchical classification, hybrid model, TF-IDF, logistic regression, cosine similarity.

1. Формальная постановка задачи.

В работе решается задача автоматической классификации текстовых описаний видов разрешённого использования земельных участков. Пусть задано множество текстовых описаний:

$$X = \{x_1, x_2, \dots, x_n\}, \qquad (1)$$

где

X — множество текстовых описаний ВРИ;

n — число записей в датасете.

Требуется построить отображение:

$$f: X \to Y, \qquad (2)$$

где

Y — множество целевых классов ВРИ.

Задача рассматривается в иерархической постановке. Классификация выполняется в два этапа: сначала определяется укрупнённый (родительский) класс, затем уточняется подкласс внутри выбранной группы:

$$f(x) = f_2\big(x \mid f_1(x)\big), \qquad (3)$$

где

$f_1(x)$ — функция определения родительского класса;

$f_2(x \mid \hat{c})$ — функция уточнения подкласса при заданном родительском классе $\hat{c}$.

Родительский класс определяется как:

$$\hat{c}_p = argmax_\{c \in C_p\} P(c \mid x), \qquad (4)$$

где

C_p — множество родительских классов (укрупнённых групп ВРИ);

$P(c \mid x)$ — вероятность принадлежности текста x к классу c.

Подкласс определяется аналогично, но в пределах выбранного родительского класса:

$$\hat{c}_s = argmax_\{c \in C(\hat{c}_p)\} P(c \mid x, \hat{c}_p), \qquad (5)$$

где

$C(\hat{c}_p)$ — множество подклассов, принадлежащих родительскому классу $\hat{c}_p$.

2. Векторное представление текста (TF-IDF).

Для преобразования текстовых описаний в числовые векторы используется метод TF-IDF (Term Frequency — Inverse Document Frequency). Вес термина t в документе d определяется как:

$$w(t,d) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t), \qquad (6)$$

где частота термина с подлинейной нормализацией:

$$\mathrm{tf}(t,d) = 1 + \log_2 \mathrm{freq}(t,d), \text{ если } \mathrm{freq}(t,d) > 0, \qquad (7)$$

и обратная документная частота:

$$\mathrm{idf}(t) = \log\left(\frac{N}{\mathrm{df}(t)}\right), \qquad (8)$$

где

$freq(t, d)$ — число вхождений термина t в документ d;

N — общее число документов в коллекции;

$df(t)$ — число документов, содержащих термин t.

Каждый текст x_i представляется в виде вектора:

$$v_i = \big(w(t_1,x_i), w(t_2,x_i), \ldots, w(t_m,x_i)\big) \in \mathbb{R}^m, \qquad (9)$$

где m — размерность словаря (число уникальных термов, отобранных по критериям min_df и max_df). В работе используются словесные n-граммы порядка (1, 2), что позволяет учитывать как отдельные слова, так и устойчивые биграммы.

3. Классификатор на основе логистической регрессии.

В качестве базового классификатора на каждом уровне иерархии используется логистическая регрессия. Для задачи с K классами вероятность принадлежности текста x к классу c вычисляется с помощью функции softmax:

$$P(y = c \mid x) = \frac{\exp\left(w_c^T v + b_c\right)}{\sum_j \exp\left(w_j^T v + b_j\right)}, \qquad (10)$$

где

v — TF-IDF вектор текста x;

w^c — вектор весов для класса c;

b^c — смещение (bias) для класса c;

K — число классов на данном уровне иерархии.

Параметры модели оптимизируются минимизацией функции кросс-энтропийных потерь:

$$L = -\frac{1}{n}\sum_i \sum_c y_i^c \cdot \log P(y = c \mid x_i), \qquad (11)$$

где

y_i^c — индикатор принадлежности примера i к классу c (1 или 0);

n — число обучающих примеров.

4. Косинусное сходство с прототипом класса.

Для дополнительного уточнения классификации вычисляется сходство TF-IDF вектора текста с прототипом каждого класса. [1] Прототип класса c определяется как центроид TF-IDF векторов обучающих примеров данного класса:

$$p^c = \frac{1}{|D^c|}\sum_{x \in D^c} v(x), \qquad (12)$$

где

D^c — множество обучающих примеров класса c;

$v(x)$ — TF-IDF вектор текста x.

Сходство текста с прототипом определяется косинусной мерой:

$$S_{\text{proto}}(x, c) = \cos(v, p^c) = \frac{v \cdot p^c}{\|v\| \cdot \|p^c\|}, \qquad (13)$$

где

v — TF-IDF вектор классифицируемого текста;

p^c — прототип (центроид) класса c;

$\|\cdot\|$ — евклидова норма вектора.

Значение S_{proto} находится в диапазоне [0, 1] для неотрицательных TF-IDF векторов и характеризует семантическую близость текста к типичному представителю класса.[2]

5. Коэффициент пересечения ключевых слов.

Для каждого класса c формируется множество ключевых слов keywords(c), характерных для данной категории ВРИ. Коэффициент пересечения определяется как доля ключевых слов класса, присутствующих в тексте:

$$K_{\text{over}}(x, c) = \frac{|\text{keywords}(x) \cap \text{keywords}(c)|}{|\text{keywords}(c)|}, \qquad (14)$$

где

keywords(x) — множество слов текста x (после очистки и нормализации);

keywords(c) — заранее определённое множество ключевых слов класса c.

Данный признак позволяет учитывать предметно-ориентированную лексику, которая может быть недостаточно выражена в статистическом TF-IDF представлении.[3]

6. Предметно-ориентированный признак.

Предметно-ориентированный признак D(x) формируется на основе наличия кодов классификатора ВРИ в тексте. Если текст содержит числовой код вида «N.M» или «N.M.K», соответствующий записи в классификаторе, то:

$$D(x) = \begin{cases} 1, & \text{если } \mathrm{code}(x) \in \mathrm{Classifier} \\ 0, & \text{иначе} \end{cases}, \qquad (15)$$

где

code(x) — извлечённый из текста код классификатора;

Classifier — множество валидных кодов ВРИ.

При D(x) = 1 классификация может быть выполнена детерминированно по справочнику, минуя вероятностную модель.

7. Гибридная решающая функция.

Итоговое решение формируется гибридной функцией, комбинирующей вероятностные оценки классификаторов и дополнительные семантические признаки:

$$F(x) = \alpha \cdot P_p(c \mid x) \cdot P_s(c' \mid x, c) + \beta \cdot S_{\text{proto}}(x, c') + \gamma \cdot K_{\text{over}}(x, c') + \delta \cdot D(x), \qquad (16)$$

где

$P_p(c \mid x)$ — вероятность родительского класса c, определённая логистической регрессией (10);

$P_s(c' \mid x, c)$ — вероятность подкласса c' внутри родительского класса;

$S_{proto}(x, c')$ — косинусное сходство с прототипом подкласса (13);

$K_{over}(x, c')$ — коэффициент пересечения ключевых слов (14);

$D(x)$ — предметно-ориентированный признак (15);

$\alpha, \beta, \gamma, \delta$ — весовые коэффициенты, определяемые на валидационной выборке.

Итоговый класс определяется как:

$$\hat{c} = \mathrm{argmax}_{c'} F(x, c'), \qquad (17)$$

Если max F(x) < θ, где θ — порог уверенности, запись направляется на ручную проверку. Это обеспечивает контроль качества при промышленном использовании модели.

8. Метрики оценки качества.

Качество классификации оценивается по метрикам precision, recall и F1-score. Для класса c:

$$\mathrm{Precision}(c) = \frac{\mathrm{TP}^c}{\mathrm{TP}^c + \mathrm{FP}^c}, \qquad (18)$$

$$\mathrm{Recall}(c) = \frac{\mathrm{TP}^c}{\mathrm{TP}^c + \mathrm{FN}^c}, \qquad (19)$$

$$F1(c) = \frac{2 \cdot \mathrm{Precision}(c) \cdot \mathrm{Recall}(c)}{\mathrm{Precision}(c) + \mathrm{Recall}(c)}, \qquad (20)$$

где

TP^c — число верно предсказанных примеров класса c;

FP^c — число примеров, ошибочно отнесённых к классу c;

FN^c — число примеров класса c, пропущенных моделью.

Агрегированные метрики:

$$\text{Macro } F1 = (1/K) \Sigma^c F1(c), \qquad (21)$$

$$\text{Weighted } F1 = \Sigma^c (n^c / n) \cdot F1(c), \qquad (22)$$

где n^c — число примеров класса c, n — общее число примеров. Macro F1 придаёт одинаковый вес всем классам, Weighted F1 учитывает размер каждого класса.[4]

9. Результаты экспериментов.

Оценка качества проводилась на тестовой выборке из 1 188 записей. Полученные значения метрик по основным подклассам представлены в Таблице 1.

Таблица 1 - Метрики качества классификации по основным подклассам

Подкласс	**Precision**	**Recall**	**F1-score**	**Support**
2.1	0,92	0,89	0,91	66
2.3	0,92	0,85	0,88	26
2.6	0,64	0,60	0,62	45
3.1	0,99	0,91	0,95	185
3.2	0,82	0,82	0,82	34
3.3	0,62	0,69	0,66	36
3.4	0,94	0,85	0,89	72
3.5	0,94	0,85	0,89	132
3.6	0,78	0,80	0,79	45
3.8	0,77	0,55	0,64	42
3.9	0,76	0,90	0,83	29
4.1	0,74	0,73	0,74	67
4.4	0,95	0,86	0,90	251
4.6	0,82	0,80	0,81	50
4.9	0,98	0,75	0,85	65
5.1	0,77	0,74	0,76	23
6.9	0,64	0,90	0,75	20
Accuracy			0,82	1188
Macro avg	0,82	0,79	0,80	1188
Weighted avg	0,89	0,82	0,85	1188

Матрица ошибок классификации по основным подклассам представлена на Рисунке 1:

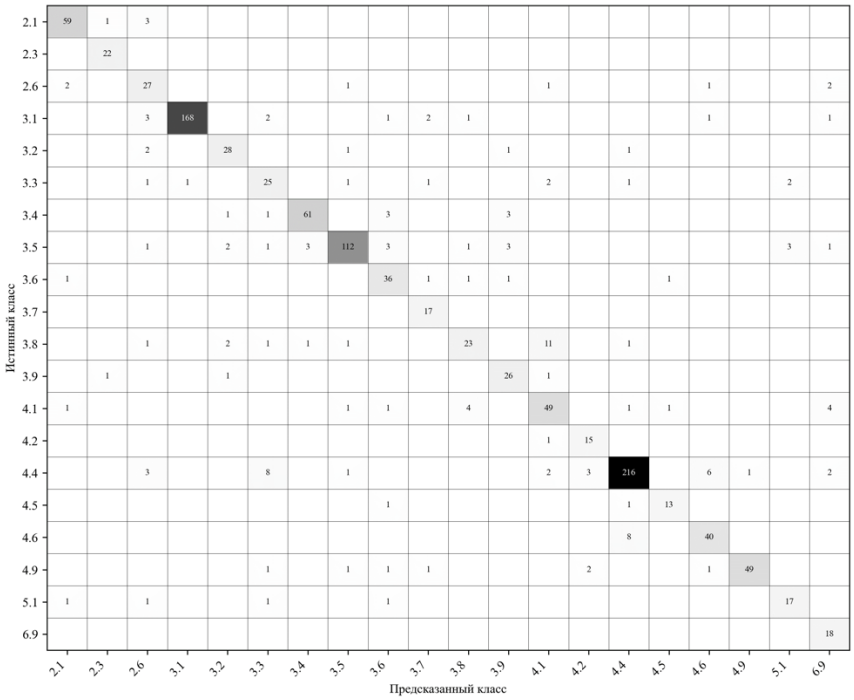


Рисунок 1 - Матрица ошибок классификации по основным подклассам

Анализируя полученные результаты, можно сказать, что разработанная модель обеспечивает достаточно устойчивое качество классификации текстовых описаний ВРИ. Средние значения accuracy = 0.82, Macro F1 = 0.80 и Weighted F1= 0.85, что говорит о приемлемой точности как в среднем по выборке, так и по отдельным подклассам. Самые высокие результаты показали подклассы 2.1, 3.1, 3.4, 3.5, 4.4 и 4.9, где значение F1-score находится в диапазоне от 0.85 до 0.95, что говорит об их достаточном объёме обучающих данных и более выраженными лексико-семантическими признаками классов. Более низкие значения получились у подклассов 2.6, 3.3, 3.8, 5.1 и 6.9, это связано с меньшим числом наблюдений и высокой семантической близостью к соседним классам. Анализ матрицы ошибок (Рисунок 1) показывает, что основная часть предсказаний сконцентрирована на главной диагонали, а возникающие ошибки носят объяснимый характер и возникают в основном между близкими по смыслу категориями.

Список литературы

1. Бурков А. Машинное обучение без лишних слов / А. Бурков. — СПб.: Питер, 2020. — 192 с.
2. Вьюгин В.В. Математические основы машинного обучения и прогнозирования / В.В. Вьюгин. — М.: Изд-во МЦНМО, 2021. — 399 с.
3. Марков С.С. Охота на электроовец. Большая книга искусственного интеллекта / С.С. Марков. — М.: ДМК Пресс, 2024. — 568 с.
4. Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval. — Cambridge University Press, 2008. — 482 p.
5. Bishop C.M. Pattern Recognition and Machine Learning. — Springer, 2006. — 738 p.

References

1. Burkov, A. Machine Learning Without the Whispers / A. Burkov. St. Petersburg: Piter, 2020. p.192
2. V. V. Vyugin. Mathematical Foundations of Machine Learning and Forecasting / V. V. Vyugin. Moscow: MCNO Publishing House, 2021. p.399
3. Markov S. S. Electric Sheep Hunting. The Big Book of Artificial Intelligence / S. S. Markov. Moscow: DMK Press, 2024. p.568
4. Manning, C. D., Raghavan, P., Schütze, H. Introduction to Information Retrieval. Cambridge University Press, 2008. p.482
5. Bishop C. M. Pattern Recognition and Machine Learning. - Springer, 2006. - p. 738