Математические структуры и моделирование
2025. № 3 (75). С. 122–126

УДК 004.85
DOI 10.24147/2222-8772.2025.3.122-126

SVM В ОПРЕДЕЛЕНИИ АВТОРСТВА ТЕКСТА: АКТУАЛЬНОСТЬ И ПЕРСПЕКТИВЫ ДАЛЬНЕЙШИХ ИССЛЕДОВАНИЙ

Р.А. Светлов[1]
студент, e-mail: roman.svetlov.dude@gmail.com
А.Н. Кабанов[2]
к.ф.-м.н., доцент, e-mail: kabanovan@omsu.ru

[1]Омский государственный университет им. Ф.М. Достоевского, Омск, Россия

[2]Нижневартовский государственный университет, Нижневартовск, Россия

Аннотация. В последние годы, с ростом популярности алгоритмов глубокого обучения, возникает вопрос об актуальности дальнейших исследований методов машинного обучения, включая SVM. Данная статья анализирует недостатки, преимущества и перспективы SVM, выявленные в других исследованиях, и делает вывод о том, что исследования SVM для классификации текста ещё не потеряли своей актуальности.

Ключевые слова: SVM, классификация, обзор.

Введение

Определение авторства текста с помощью метода опорных векторов (SVM, Support Vector Machine) остаётся актуальным, несмотря на рост популярности методов глубокого обучения. Метод демонстрирует высокую точность (до 98,8 %) в задачах классификации текстов, сравнимую с результатами нейросетей (например, 98 % у свёрточной нейронной сети (CNN, Convolutional Neural Network)), особенно на коротких текстах и при ограниченных вычислительных ресурсах. Однако прямые сравнения затрудняются тем, что условия тестирования часто не стандартизированы.

Перспективы SVM связаны с работой с редкими признаками (например, гапаксами), интеграцией гибридных моделей (SVM + нейросети) и разработкой универсальных многомодельных классификаторов. Вместе с тем метод имеет ограничения: нелинейный рост вычислительных затрат, сложности в выборе ядра и худшая эффективность на сгенерированных текстах. Цель статьи – обобщить текущие данные о применимости SVM, его преимуществах и недостатках, сделать выводы об актуальности данного метода.

В статье рассматриваются данные, полученные из нескольких систематических обзоров классификации текстов, а также некоторых работ, посвящённых попыткам улучшить результаты классификации.

Проблема сопоставления данных по разным наборам и параметрам классификации

В различных исследованиях используются различные наборы данных с различными условиями тестирования, что не даёт провести прямое и точное сравнение, что очевидным образом является проблемой для данной статьи.

Среди используемых наборов данных: 20Newsgroup, Amazon Review, Bike Review, Blogger, Chinese Microblogg, Counter, Gold, IMDb, PAN-12, Reuters, Spam-1000, Twitter, Webkb [1, с. 14], а также классическая литература и фанфикшен. О проблеме можно судить хотя бы по разбросу точности одного классификатора в различных наборах данных; для SVM, например, от 93 % на Reuters до 99 % на 20Newsgroup [1].

Актуальность

SVM демонстрирует высокую эффективность в задачах классификации текстов, сохраняя свою актуальность даже в условиях конкуренции с методами глубокого обучения. Согласно систематическому обзору [1], SVM является наиболее часто используемым алгоритмом в этой области (118 из 224 проанализированных работ). Как показано в таб. 1, его максимальная точность достигает 98,88 % на датасете 20Newsgroup – результат, который превосходит Naive Bayes (95,52 %).

Таблица 1. Лучшие показатели SVM в различных наборах данных

Датасет	**Accuracy (точность), %**	**Показатели других моделей, %**
20Newsgroup	98,88	NB (95,52)
Reuters	97,60	NB (97,89), kNN (96,64)
Spam-1000	95,20	NB (92,70), ANN (85,30)
Reuters	95,10	Traditional SVM (82,76)
Reuters	93	kNN (53), NB (24)

Это подтверждает, что SVM остаётся мощным инструментом даже при сравнении с современными нейросетями (например, CNN, показывающей 98 % точности на датасете Twitter, как показано в табл. 2).

Ключевые преимущества SVM:

- Одни из лучших результатов точности.
- Хорошо работает с нелинейным распределением данных.
- Устойчивость к переобучению. В отличие от нейросетей, SVM имеет теоретически обоснованную способность минимизировать риск переобучения за счёт максимизации разделяющего гиперплоского зазора между классами.
- Скорость обучения. SVM обучается быстрее, чем глубокие архитектуры, такие как CNN или длинная цепь элементов краткосрочной памяти (LSTM, Long Short-Term Memory), особенно на средних по размеру датасетах. Это делает

Таблица 2. Максимальные уровни точности, полученные из лучших наборов данных

Датасет	**Accuracy (точность), %**	**Алгоритм**
Enron8715	86	ANN
Bug Report	47,60	RNN
EHR	88,30	RF
Yelp Review	84,20	SVM
Twitter	98	CNN
Bike Review	79,25	RF
20Newsgroup	98,88	SVM
Amazon Review	91	LSA
IMDb	88,87	LR
Ohsumed	54,10	LDA
Movie Review	86,50	SVM
Spam-1000	95,20	SVM
Reuters	97,89	NB
Webkb	91,30	SVM

его предпочтительным для сценариев с ограниченными вычислительными ресурсами.

Недостатки в методах глубокого обучения:

- Зависимость от качества и объёма данных: нейросети требуют больших размеченных датасетов, тогда как SVM эффективен даже на малых выборках.
- Высокие вычислительные затраты: обучение глубоких моделей. Например, модель BERT, Bidirectional Encoder Representations from Transformers («двунаправленные презентации кодировщика для трансформеров») требует мощного GPU-ускорителя (графический процессор (GPU, Graphics Processing Unit)).

Перспективы

Несмотря на длительное изучение, SVM остаётся актуальным в задачах классификации текстов.

Одним из перспективных направлений является использование редких слов – гапаксов. [2].

В статье [3, с. 30] отмечается, что, по сравнению со статистическими методами, основанными на евклидовом расстоянии и косинусном сходстве, One-Class SVM + fastText может улучшить результаты до 15 %.

Статья [4] даёт информацию об использовании гибридных ансамблевых методов, состоящих из SVM и шести различных алгоритмов бустинга, причём SVM+CBC достиг точности 92 %.

В работе [5] рассматривается возможность повышения точности определения авторства текста за счёт интеграции данных о демографических характеристиках автора, таких как пол и возраст. Авторы предлагают архитектуру, в которой используется стек моделей: одна модель анализирует текст для классификации автора, а другая – предсказывает социально-демографические признаки, которые затем используются как дополнительные входные данные. Такой подход позволил повысить точность классификации в сравнении с более традиционными моделями.

В исследовании [6] представлено применение генетического алгоритма для отбора информативных признаков при решении задачи определения автора русскоязычного текста. В экспериментах обучение SVM на подмножестве из 400 наиболее информативных признаков позволило увеличить точность классификации на 10 % по сравнению с использованием всего исходного набора признаков. Генетический алгоритм помог сократить размерность пространства признаков, улучшить обобщающую способность модели и снизить вычислительную нагрузку. При этом глубокие нейронные сети показали более высокую точность (до 96 %), но требовали значительно больше времени на обучение.

Таким образом, перспективными направлениями улучшения SVM являются связанные с работой с редкими признаками и гибридными архитектурами. Однако большинство количественных оценок эффективности этих подходов требует дополнительной верификации.

Недостатки SVM

- Большие требования к ресурсам в процессе классификации. Нелинейный рост требований с ростом датасета.
- Сложности в подборе ядра под конкретную задачу [1, с. 9].
- SVM хуже нейросетей определяет сгенерированные тексты [3].

Заключение

В данной статье проведён обзор актуальности и перспектив SVM в конкуренции с методами глубокого обучения. Анализ показал, что, несмотря на хорошие результаты нейронных сетей, метод опорных векторов сохраняет не только актуальность, но и перспективы для дальнейших исследований. В то же время использование SVM имеет свои ограничения, часть которых можно решить гибридными моделями.

Литература

1. Palanivinayagam A., El-Bayeh C.Z., Damaševičius R. Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review // Algorithms. 2023. Vol. 16, No. 5. Art. 236.
2. Faltýnek D., Matlach V. Hapax remains: Regularity of low-frequency words in authorial texts // Digital Scholarship in the Humanities. 2022. Vol. 37, Iss. 3. P. 693–715.
3. Fedotova A., Romanov A., Kurtukova A., Shelupanov A. Digital Authorship Attribution in Russian-Language Fanfiction and Classical Literature // Algorithms. 2023. Vol. 16, No. 1. Art. 13.

4. Khan T.F., Sabir M., Malik M.H., Ghous H., Ijaz H.M., Nadeem A., Ejaz A. Comparative Analysis of Hybrid Ensemble Algorithms for Authorship Attribution in Urdu Text // Journal of Computers and Intelligent Systems. 2024. Vol. 3, No. 1. P. 81–91.
5. Deutsch C., Paraboni I. Authorship attribution using author profiling classifiers // Natural Language Engineering. 2023. Vol. 29. P. 110–137.
6. Куртукова А.В., Романов А.С., Федотова А.М., Шелупанов А.А. Применение методов машинного обучения и отбора признаков на основе генетического алгоритма в решении задачи определения автора русскоязычного текста для кибербезопасности // Доклады ТУСУР. 2022. № 1.

SUPPORT VECTOR MACHINE IN TEXT AUTHORSHIP ATTRIBUTION: RELEVANCE AND PROSPECTS FOR FURTHER RESEARCH

R.A. Svetlov[1]
Student, e-mail: roman.svetlov.dude@gmail.com
A.N. Kabanov[2]
Ph.D. (Phys.-Match.), Associate Professor, e-mail: kabanovan@omsu.ru

[1]Dostoevsky Omsk State University, Omsk, Russia
[2]Nizhnevartovsk State University, Omsk, Russia

Abstract. In recent years, with the growing popularity of deep learning algorithms, the question arises about the relevance of further research into machine learning methods, including SVM. This article analyzes the disadvantages, advantages, and prospects of SVM identified in other studies and concludes that SVM-based text classification research has not yet lost its relevance.

Keywords: SVM, classification, review.

Дата поступления в редакцию: 14.06.2025