УДК 004.932+811.512.1

ОЦЕНКА ЭФФЕКТИВНОСТИ ИСПОЛЬЗОВАНИЯ АЛГОРИТМА RANDOM FOREST ДЛЯ РАСПОЗНАВАНИЯ СИМВОЛОВ ХАКАССКОЙ ПИСЬМЕННОСТИ

А. А. Голубничий[1], А. Д. Яблонцева[2]

Хакасский государственный университет им. Н. Ф. Катанова, пр. Ленина, 92/1, 655017, г. Абакан, Россия,
[1]artem@golubnichij.ru, [2]91919919@mail.ru

В статье рассматривается применение алгоритма Random Forest для распознавания символов хакасской письменности. Представлены результаты оценки для распознавания символов из набора ExTL для ансамблей от 10 до 250 решающих деревьев.

Ключевые слова: распознавание символов, тюркские языки, расширенный кириллический алфавит, хакасская письменность, ExTL.

EVALUATION OF THE EFFICIENCY OF USING THE RANDOM FOREST ALGORITHM FOR RECOGNITION OF SYMBOLS OF THE KHAKAS SCRIPT

A. A. Golubnichiy[1], A. D. Yablontseva[2]

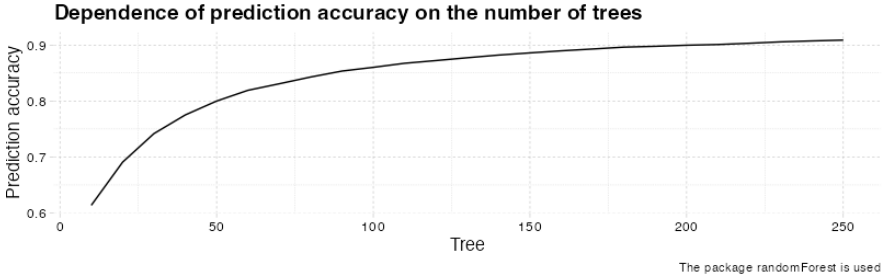
Katanov Khakass State University, ave. Lenin, 92/1, 655017, Abakan, Russia,
[1]artem@golubnichij.ru, [2]91919919@mail.ru

The article discusses the application of the Random Forest algorithm for recognizing the symbols of the Khakass script. The results of evaluation for character recognition from the ExTL set for ensembles from 10 to 250 decision trees are presented.

Keywords: character recognition, Turkic languages, extended Cyrillic alphabet, Khakass script, ExTL.

Современная письменность хакасского языка построена на основе кириллицы с расширением дополнительными символами: ғ, і, ң, ө, ү, ҷ. «Случайный лес» является одним из популярных алгоритмов решения задачи классификации и регрессии. В случае оценивания букв, в процессе распознавания символов, задача сводится к классификации конкретного изображения, посредством отнесения его к одной из групп (конкретному символу алфавита). Данный алгоритм широко применяется для решения задач с малым количеством категорий, например, распознавания рукописных символов арабских цифр в наборе данных MNIST [1]. Для распознавания символов в данном случае возможно использовать ансамбли из сравнительно небольшого количества деревьев, при увеличении количества групп для классификации задача в значительной степени усложняется. Исходным набором данных для проверки возможности применения алгоритма Random Forest послужил ранее разработанный набор данных ExTL, содержащий символы расширенного кириллического алфавита, используемого в современных тюркских письменностях [2]. Для построения модели использовался 101 символ, включающий основную кириллицу в нижнем и верхнем регистре, пунктуационные символы, а также 6 букв в двух регистрах для обозначения символов специфичных для хакасской письменности [3]. Однако вместо символа ҷ использовалось начертание данного символа с нижним выносным элементов справа – ҷ, так как в абсолютном большинстве источников печатного и электронного вида используется именно этот знак.

Для оценки возможности применения алгоритма Random Forest были произведены расчета точности распознавания символов с использованием от 10 до 250 решающих деревьев с шагом 10, с пятикратной повторностью. Усредненные оценки точности определения символов приведены на рисунке.



Зависимость точности определения символов хакасского языка от количества деревьев в алгоритме Random Forest

Как видно из рисунка, точность распознавания символов превышает 90% в случае увеличении числа решающих деревьев до 200. Дальнейшее увеличение деревьев в ансамбле не приводит к значимым улучшениям предсказания.

Библиографический список

1. Kussul E., Baidyk T. Improved method of handwritten digit recognition tested on MNIST database // Image and Vision Computing. 2004. Vol. 22, Iss. 12. P. 971–981. URL: https://doi.org/10.1016/j.imavis.2004.03.008 (дата обращения: 02.09.2022).
2. ExTL (Extended Turkic Languages). URL: https://extl.shinyapps.io/web_extl/ (дата обращения: 15.09.2022).

3. Голубничий А. А., Яблонцева А. Д. Формирование «белых списков» для создания универсальной системы распознавания тюркских языков, построенных на расширенных кириллических алфавитах // Научно-технический вестник Поволжья. 2022. № 4. С. 67–70.