АЛГОРИТМ ROTATION FOREST: СРАВНИТЕЛЬНЫЙ АНАЛИЗ С ИЗВЕСТНЫМИ АНСАМБЛЯМИ КЛАССИФИКАТОРОВ

К. Р. Панин

Согласно [1−2] использование комбинации классификаторов позволяет повысить точность классификации при решении практических задач. Среди всех имеющихся методов построения ансамбля классификаторов наиболее популярными являются «bagging» и «boosting» [3], которые основаны на манипуляциях с исходным обучающим множеством с целью построения нескольких классификаторов. Похожим на бэггинг методом является метод случайных подпространств (random subspace method, RSM [4]), основанный на создании вариативности при обучении с помощью выбора случайных подмножеств признаков. Широко известным примером использования бэггинга и RSM является случайный лес [5]. В настоящей работе рассмотрены популярные ансамбли классификаторов и предложен новый вариант ансамбля на основе комбинации деревьев решений, основанный на методе вращающихся деревьев [6] и представляющий собой его модифицированную версию. Метод заключается в осуществлении повторных случайных подвыборок из обучающего множества с последующим применением метода главных компонент и построения каждого базового классификатора на трансформированном признаковом пространстве, что позволяет снизить степень корреляции между ошибками отдельных классификаторов.

В данном исследовании выполнен сравнительный анализ известных методов построения ансамблей классификаторов бэггинг, AdaBoost, случайный лес и предложенного метода на основе вращающихся лесов. В качестве базового алгоритма использовался вариант деревьев решений. Наборы данных получены из архивов по машинному обучению UCI Machine Learning Repository [7] и KEEL-dataset repository [8]. Размер ансамблей варьировался от 10 до 100 с шагом 10. Эффективность классификации сравнивалась для ансамблей равных размеров. Для каждого набора данных и ансамбля эффективность классификации оценивалась с использованием 100 повторных разбиений объектов на обучающее и тестовое множества в пропорции 10:1, т.е. ансамбль строился на обучающем множестве, а эффективность классификации оценивалась на тестовом. Данный подход позволил оценить среднее значение и стандартное отклонение критерия эффективности классификации, результаты для каждого ансамбля представлены в таблице 2. В таблице 1 демонстрируется ранжированный список сравниваемых методов, согласно разнице

между числом раз, когда метод был лучше или хуже другого метода при проведении попарных сравнений.

Таблица 1

Ранжированный список сравниваемых методов

	Вращающиеся деревья	АдаБуст	Бэггинг	Случайный лес
Ранг	20	-12	-22	14
Количество побед	31	15	10	28
Количество поражений	11	27	32	14

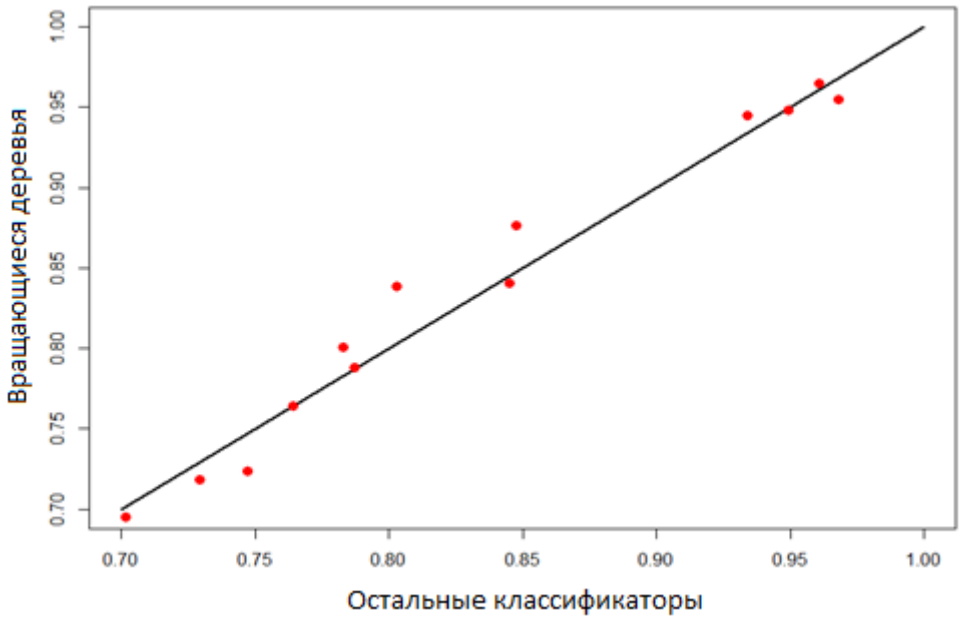

Рис. 1. Диаграмма демонстрирует точность предложенного метода к остальным методам по всем наборам данных

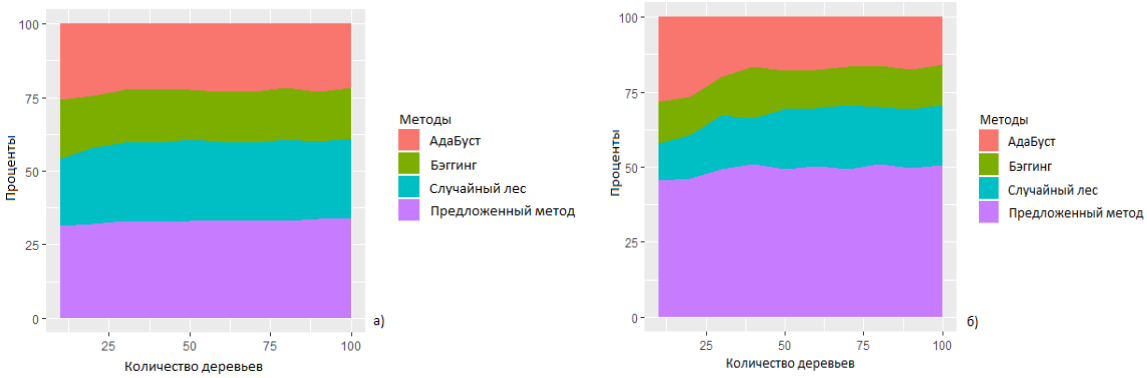

Рис. 2. График со слоями демонстрирует:
а – долю выигрышей для каждого метода на всех наборах данных в зависимости от количества деревьев, б – долю выигрышей для каждого метода на наборе данных balance

Таблица 2

Точность классификации и стандартное отклонение

Appendicitis	0.8486±0.0033	0.8427±0.0031	0.8472±0.0035	0.8572±0.0054•
•Balance	0.8895±0.0055•	0.8397±0.0050	0.8349±0.0029	0.8404±0.0061
BreastCancer	0.9665±0.0008•	0.9606±0.0017	0.9570±0.0015	0.9660±0.0020
Bupa	0.7061±0.0056	0.7099±0.0063	0.7202±0.0088	0.7263±0.0151•
Cleveland	0.5545±0.0035•	0.5217±0.0062	0.5351±0.0022	0.5390±0.0044
Ecoli	0.8033±0.0058	0.7849±0.0037	0.7627±0.0028	0.8056±0.0070•
•Heart	0.8469±0.0037•	0.8095±0.0084	0.8188±0.0061	0.8262±0.0107
•Ionosphere	0.954±0.0036 •	0.9377±0.0017	0.9212±0.0027	0.9331±0.0025
Iris	0.9477±0.0029	0.946±0.0025	0.9535±0.0027 •	0.9504±0.0016
Led7digit	0.7261±0.0035	0.7309±0.0019•	0.7195±0.0025	0.7216±0.006
Pima	0.7700±0.0027•	0.743±0.0035	0.7697±0.0026	0.7630±0.0061
Sonar	0.8355±0.0146•	0.829±0.0204	0.7800±0.0047	0.8177±0.0165
Vehicle	0.7319±0.0031	0.7699±0.0108•	0.7090±0.0032	0.7492±0.0017
Wine	0.9722±0.0066	0.9682±0.0007	0.9585±0.0035	0.9762±0.0043•

ЗАКЛЮЧЕНИЕ

В ходе исследования гипотеза о преимуществах предложенного ансамбля на основе вращающихся деревьев подтвердилась на большинстве наборов данных как видно из таблицы 2 и рисунка 1 на котором по оси Y представлена точность предложенного метода, а на оси X наилучшая точность классификации среди сравниваемых методов. Диагональной линия, соответствует одинаковым значениям точности. Большинство точек лежит выше диагональной линии означает преимущество предложенного метода. С помощью разработанного способа ранжирования ансамблей на основе попарного сравнения эффективности по совокупности всех анализируемых наборов данных получили, что метод вращающихся деревьев стоит на первом месте в списке ансамблей классификаторов, ранжированном по эффективности. На втором месте стоит случайный лес. При анализе эффективности ансамблей по совокупности наборов данных метод вращающихся деревьев выигрывает для

большинства размеров деревьев как видно из графика со слоями на рисунке 2.

Литература

1. Multiple Classifier Systems / J. Kittler & F. Roli (editors) // Proc. of 2nd International Workshop, MCS2001, (Cambridge, UK, 2-4 July 2001) / Lecture Notes in Computer Science. – Vol. 2096. – Springer-Verlag, Berlin.
2. *Vishwath P.* Fusion of multiple approximate nearest neighbor classifier for fast and efficient classification / P. Vishwath, M.N. Murty, C. Bhatnagar, // Information fusion. – 2004. – Vol. 5. – P. 239-250.
3. *Quinlan J.R.* Bagging, boosting and C4.5 / J.R. Quinlan // Proceedings of AAA/IAAI. – 1996. – Vol. 1. –P. 725-730.
4. *Skurichina M., Duin R. P. W.* Limited bagging, boosting and the random subspace method for linear classifiers // Pattern Analysis & Applications. 2002. Pp. 121–135.
5. *Breiman, Leo*. Random Forests // Machine Learning, 45(1), 5-32, 2001.
6. Kuncheva L.I. and C.J. Whitaker, Pattern recognition and classification, Wiley StatsRef-Statistics Reference Online, 2014.
7. *C.L. Blake and C.J. Merz*, "UCI Repository of Machine Learning Databases," 1998, http://www.ics.uci.edu/mlearn/MLRepository.html.
8. KEEL-dataset citation paper: J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing 17:2–3 (2011) 255–287.

СТАТИСИТЧЕСКАЯ ОБРАБОТКА РЕЗУЛЬТАТОВ ТЕСТИРОВАНИЯ ГРУПП ИСПЫТУЕМЫХ

О. В. Паркалова

Анализ результатов тестирования – это важная составляющая процесса контроля знаний. Статистическая обработка материалов по результатам тестирования решает две основные задачи: объективно оценивает результаты испытуемых и позволяет сделать вывод о качестве тестов. Поэтому важно иметь инструментарий, позволяющий оценить качество тестовых заданий и сделать вывод о целесообразности использования каждого отдельного задания.

Характеристики, вычисляемые по результатам выполнения тестов испытуемыми можно разделить на две категории: характеристики всего теста и характеристики отдельных вопросов.

Существует различный инструментарий, позволяющий вычислять эти характеристики. Это можно сделать в каком-либо математическом пакете, в табличном процессоре или с использованием специальных программных средств. Система управления обучением Moodle, например, имеет широкий набор средств по оценке результатов тестирования.