УДК 51-74

ТЕСТИРОВАНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ЗАДАЧ КЛАССИФИКАЦИИ И РЕГРЕССИИ ПОРЯДКОВОГО ПРИЗНАКА

Д.А. Редько

Научный руководитель: доцент, к.ф.-м.н., М.Е. Семенов

Национальный исследовательский Томский политехнический университет,

Россия, г. Томск, пр. Ленина, 30, 634050

E-mail: dar27@tpu.ru

MACHINE LEARNING ALGORITHMS TESTING FOR CLASSIFICATION AND REGRESSION PROBLEMS ON ORDINAL RESPONSE

D.A. Redko

Scientific Supervisor: Ass. Pr., PhD, M.E. Semenov

Tomsk Polytechnic University, Russia, Tomsk, Lenin str., 30, 634050

E-mail: dar27@tpu.ru

Abstract. *Optimal features selection for predictions play a key role in machine learning. In this paper, the main machine learning algorithms are analyzed on binary and ordinal classification and regression tasks. The Wine Quality DataSet has been used for numerical experiments.*

Введение. Машинное обучение является ведущей тенденцией развития промышленной сферы деятельности человека. Одной из ключевых проблем машинного обучения является выбор алгоритма обучения модели для решения задач классификации или регрессии в случае порядкового результирующего признака. Эта проблема связана с тем, что алгоритмы не учитывают возможность использования операции порядка и работают фактически как с бинарными (дихотомическими) данными.

Цель работы – изучение и практическое применение алгоритмов машинного обучения для решения задач классификации и регрессии для объектов, измеренных в порядковой шкале. Для проведения численных экспериментов использован набор данных Wine Quality Data Set [1].

Исходные данные. Исходный набор содержит 1599 записей, без пропусков и NaN ячеек и включает 11 вещественных признаков красного вина: fixed acidity (фиксированная кислотность), volatile acidity (летучая кислотность), citric acid (лимонная кислота), residual sugar (остаточный сахар), chlorides (хлориды), free sulfur dioxide (свободный диоксид серы), total sulfur dioxide (общий диоксид серы), density (плотность), pH (кислотность), sulphates (сульфаты), alcohol (содержание спирта, %), а также интегральный экспертный признак – quality (*качество*), приведенный в порядковой шкале: 0 – худшее, 10 – лучшее качество. Анализ данных показал, что в наборе признак *качество* принимает значения 3, 4, 5, 6, 7, 8 с частотами 10, 53, 681, 638, 199, 18 соответственно. При этом отношение 10/681=0,0146, что позволяет говорить о дисбалансе классов.

В работе проведены численные эксперименты с а) исходным набором, а также с б) бинаризированным набором: 0 – плохое, 1 – хорошее качество. Для отображения порядкового признака *качество* (quality) в бинарный мы использовали порог, равный 5. При этом доли элементов выборки с

метками 0/1 существенно не отличаются и равны 0,465/0,535 соответственно, что позволяет утверждать, что бинаризованные данные сбалансированы.

Предварительный анализ исходных данных с использованием матрицы корреляции (рис. 1) показал, что данные слабо коррелированы. Наибольшая зависимость для признака *качество* (quality) наблюдается с признаками *летучая кислотность* (volatile acidity), *сульфаты* (sulphates), *содержание спирта* (alcohol) и равна соответственно –0,39; 0,25; 0,48. После бинаризации данных корреляция незначительно уменьшилась и составила –0,32; 0,22; 0,43 соответственно. Хи-квадрат тест на независимость показал, что признаки free sulfur dioxide и total sulfur dioxide зависимы с результирующим признаком, а распределение значений *содержание спирта* (alcohol) отличается от нормального распределения (Shapiro-Wilk normality test, W = 0,92884, p-value < 2,2e–16).

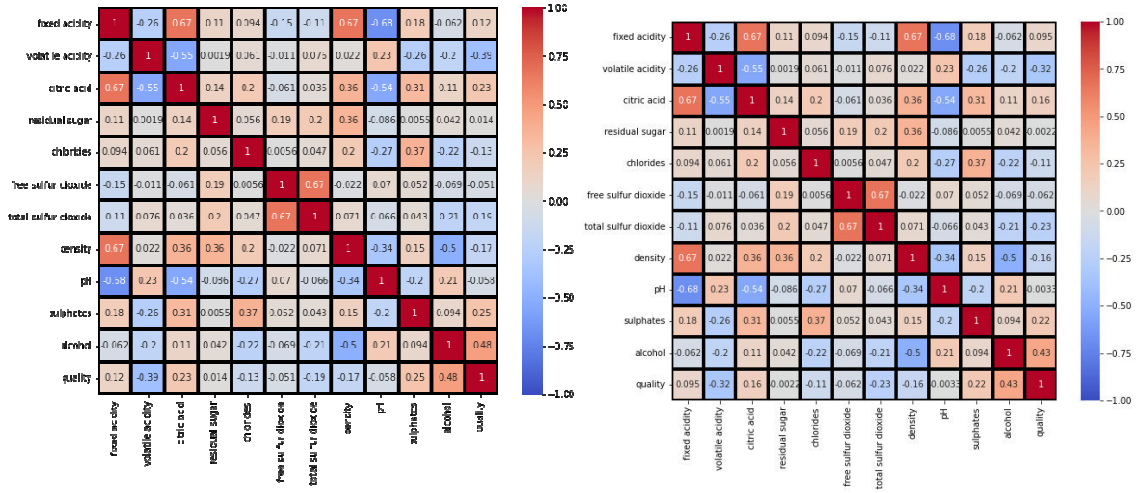

Рис. 1. Матрица корреляций для исходного набора данных (слева) и бинаризованного набора (справа)

Таблица 1

Доля правильно классифицированных значений признака "quality" для различных классификаторов

Классификатор	Бинарный признак		Порядковый признак		Кластеризация	
	Test	Train	Test	Train	Test	Train
ExtraTrees	0,7804	0,8339	0,6191	**0,9559**	0,9975	1,0000
Bagging	0,7767	0,9821	0,6097	0,7514	0,9975	1,0000
RandomForest	0,7729	0,8724	0,5984	0,7382	0,9925	1,0000
KNeighbors	0,7467	0,8095	0,5872	0,7091	0,9925	1,0000
DecisionTree	0,7410	0,8921	0,5816	0,7138	0,9875	1,0000
GradientBoosting	0,7410	0,7720	0,5590	0,8217	0,9625	0,9799
AdaBoost	0,7073	0,7485	0,5290	**0,5684**	0,9350	0,9599

Вычислительные эксперименты. Мы разделили исходный набор данных на обучающую (train) и тестовую (test) выборки в пропорции 70/30 и применили различные классификаторы. Результаты работы классификаторов приведены в Таблице 1. Бинарная классификация показывает удовлетворительные результаты: доля правильно классифицированных объектов на тестовой выборке не опускается ниже 0,7.

При этом качество классификации исходного набора данных на обучающей выборке колебался от 0,5684 (AdaBoost) до 0,9559 (ExtraTrees) и существенно падает на тестовой выборке, вплоть до 0,5290. Возможные причины – недостаточный объем данных для обучения и неспособность классификаторов учитывать наличие порядка в интегральном признаке. Для улучшения результатов классификации мы применили к исходным данным метод *k*-средних и выделили два кластера с 1179 и 420 элементами соответственно. Как видно из Таблицы 1 многие классификаторы показали абсолютное качество на тренировочной выборке, что свидетельствует о переобучении модели. Дальнейшие эксперименты были проведены с вариацией количества кластеров. Результаты экспериментов с выделением 3, 4, …, 15, 25, 30, 50 кластеров показали, что данная стратегия качественно не влияет на решение задачи классификации.

Развивая идею использования кластеров, мы проделали серию экспериментов с использованием случайного леса для решения регрессионной задачи для порядкового признака. Анализ показал, что три самых значимых признака – *содержание спирта* (alcohol), *летучая кислотность* (volatile acidity), *сульфаты* (sulphates). Для построения квантильной регрессии на основании леса [2, 3] мы использовали в качестве регрессора – *содержание спирта* (alcohol), который принимает вещественные значения из интервала [8,4; 14,9], M=10,42, SD=1,06 (рис. 2а). Анализ влияния выбранного регрессора на порядковую переменную *качество* показал, что качественное изменения распределения происходит в квантильных точках 0,1, 0,4 и 0,7 (рис. 2б).

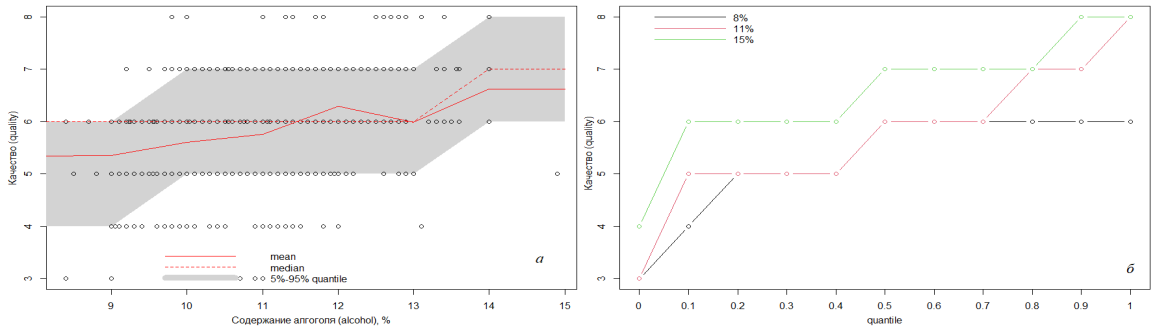

Рис. 2. Квантильная регрессия на основании леса из 500 деревьев с доверительным интервалом (а) и зависимость качества для различных уровней регрессора – содержание спирта: 8, 11, 15 % (б)

Заключение. Решение задач классификации и регрессии для результирующего признака, измеренного в порядковой шкале имеет свои особенности, которые нужно учитывать при выборе моделей машинного обучения. Для настройки параметров классификаторов и моделей регрессии требуется проведение дальнейших исследований.

СПИСОК ЛИТЕРАТУРЫ

1. Cortez P., Cerdeira A., Almeida F., Matos T. and Reis J. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems // Elsevier. – 2009. – V. 47(4). – P. 547-553.
2. Strobl C., Malley J., Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests // Psychol Methods. – 2009. – V. 14(4). – P. 323-48.
3. Meinshausen N., Ridgeway G. Quantile Regression Forests // J. Mach. Learn. Res. – 2006. – 7. – P. 984–987.