Сетевой метод автоматизации интеллектуального анализа данных в научных исследованиях

Е. Е. Котова, А. С. Писарев, И. А. Писарев
Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В. И. Ульянова (Ленина)»
eekotova@gmail.com, a_pisarev@mail.ru, pisarevivan@yandex.ru

***Аннотация*. В работе представлен метод автоматизации сценариев интеллектуального анализа данных для применения в научных исследованиях. В основе гибридная база данных информационных ресурсов (документов, изображений, видео и др.) и база метаданных, включающая средства обработки и интеллектуального анализа данных в веб-среде. Метаданные описывают более 50 методов машинного обучения, относящихся к задачам классификации информационных ресурсов на основе теоремы Байеса (Bayes), деревьев решений (Decision Trees), опорных векторов (Support Vector Machines - SVM), конволюционных нейронных сетей (Convolutional Neural Networks – CNN) и др. Приведены примеры алгоритмов разработки сценариев научных исследований.**

Ключевые слова: автоматизированная система научных исследований; методы машинного обучения; база метаданных; веб-среда; алгоритмы разработки сценариев

I. Введение

Разработка новых средств автоматизации анализа данных относится к направлению исследований в области автоматизации решения задач машинного обучения (Automated Machine Learning – AutoML), целью которого является анализ и извлечение новых знаний из больших объемов накапливаемой информации и данных в разных областях исследований.

Разработка методов AutoM совместно с методами предварительной обработки данных (текстовых, видео и изображений, сигналов) является объектом исследований в области обработки и анализа данных (Data Science) [1, 2].

Среди наиболее известных средств для автоматизированного анализа данных применяются программные продукты и платформы для анализа больших данных: WEKA [3], RapidMiner [4], KNIME [5], KEEL [6], SAP Predictive Analytics, Matlab (MathSoft), TenzorFlow [7] и др.

Метод автоматизированной обработки, классификации и регрессионного анализа данных реализоан в веб-среде ОнтоМАСТЕР-Ресурс [8]. Метод отличается построением сценариев и выбором методов машинного обучения, релевантных переменным (признакам и классам) данных, что позволяет повысить производительность научных исследований.

Сетевая архитектура ОнтоМАСТЕР-Ресурс позволяет проводить совместные научные исследования с использованием веб-интерфейсов и выполнения сценариев по анализу разнородных экспериментальных данных (текстов, видео и изображений, сигналов).

Результаты автоматизированного выполнения сценариев сохраняются в базе данных, что позволяет проводить всесторонний анализ и поиск наилучших моделей на основе многокритериального подхода.

II. Метод автоматизации интеллектуального анализа данных

Метод автоматизации обработки и интеллектуального анализа данных включает планирование и выполнение сценариев решения задач многоклассовой классификации и регрессионного анализа.

Рассматриваются прямая и обратная задачи.

Прямая задача предсказания выходной переменной (класса) Y от входных переменных (признаков) X представлена в виде отображения:

$$f : X \to Y ,$$

где f – модель анализа данных .

Различается постановка задачи в зависимости от типа данных. В случае, если области определения $D(X)$ и изменения $D(Y)$ имеют типы данных «numeric», решается прямая задача регрессионного анализа. Если $D(Y)$ имеет тип данных «symbolic», то решается задача классификации на основании признаков, содержащихся в данных.

Проблема поиска наилучших моделей для классификации и регрессии формулируется как обратная задача: найти отображение f^* (структуру модели и значения параметров настройки Z) при известных значениях независимых переменных (признаков) X и зависимой переменной (класса) Y, которое обеспечивает экстремум (минимум или максимум) функционала в виде

Работа частично выполнена при финансовой поддержке РНФ, проект № 17-71-20077.

аддитивной свертки целевых функций R_j при выполнении ограничений $R_i \in C$:

$$f^* = \arg\left(extr_{R_i \in C} \left\{ \sum_j \alpha_j R_j \left(f\left(X, Z\right), Y \right) \right\} \right).$$

В качестве целевых функций R_j используются количественные показатели производительности методов классификации и регрессии.

Возможны различные реализации оценок качества моделирования:

- по одному важнейшему критерию – при этом вес α_j этого критерия равен 1;
- – по одному критерию при проверке условия выполнения требований на ограничения для других критериев;
- многокритериальная оценка.

К общим показателям классификации и регрессии относятся:

- средняя абсолютная ошибка (mean absolute error – MAE);
- среднеквадратическая ошибка (root mean square error – RMSE);
- относительная абсолютная ошибка (relative_absolute_error – RAE);
- относительная среднеквадратическая ошибка (root relative squared error – RRSE).

Результаты классификации дополнительно характеризуется показателями:

- точности (accuracy) – отношением «правильно» (true positive) классифицированных случаев к общему их числу;
- коэффициентом каппа-статистики (kappa statistic).

Результаты регрессионного анализа дополнительно характеризуются коэффициентом корреляции (correlation coefficient – CC).

В разработанном методе осуществляется поиск моделей и выполнение методов, релевантных данным, автоматическое выполнение программ классификации и регрессионного анализа, сохранение результатов в базе данных, формирование отчетов с ранжированием моделей в соответствии с выбранными весами для частных критериев.

В публикациях по применению методов классификации среди наиболее часто используемых критериев отмечается точность метода, а при оценке методов регрессионного анализа – среднеквадратическая ошибка (RMSE) [9, 10].

В разработанном методе дополнительно предусмотрено вычисление модифицированного информационного критерия Акаике (AIC) что позволяет при выборе лучших моделей учитывать не только перечисленные выше критерии, но и сложность модели в зависимости от числа используемых входных переменных (признаков). В соответствии с принципом Оккама по критерию AICc более предпочтительными являются не только наиболее точные, но и наиболее простые модели.

В модифицированном информационном критерии Акаике при оценке качества моделей учитывается сложность (число параметров) и точность описания экспериментальных данных [11]:

$$AIC_c = 2k + n \ln \frac{RSS}{n} + \frac{2k\left(k+1\right)}{n-k-1},$$

где k – число параметров модели, n – размер выборки.

Критерий суммы квадратов остатков (Residual Sum of Squares – RSS):

$$RSS = \sum_{i=1}^{n} \left(y_i^{\text{model}} - y_i^{data} \right)^2,$$

$y_i^{\text{model}}, y_i^{data}$ – модельные и экспериментальные данные в моменты времени t_i .

$$RMSE = \sqrt{\frac{RSS}{n}}.$$

Поэтому, формула вычисления AIC с использованием полученной в результате классификации оценки RMSE имеет следующий вид для случаев $\text{RMSE} > 0$ и $n-k>1$ (при 100% точности AIC принимает условное большое отрицательное число):

$$AIC_c = 2k + n \ln \text{RMSE}^2 + \frac{2k\left(k+1\right)}{n-k-1}$$

III. Архитектура автоматизированной системы

Архитектура автоматизированной системы состоит из веб-интерфейсов исследователей, программных агентов, гибридной базы данных, базы знаний, онтологии методов и моделей [12].

Гибридная база данных содержит информационные ресурсы (документы, изображения, видео и др.). База знаний включает метаданные о средствах обработки и методах интеллектуального анализа данных. Метаданные описывают более 50 методов машинного обучения, относящихся к задачам классификации информационных ресурсов на основе теоремы Байеса (Bayes), деревьев решений (Decision Trees), опорных векторов (Support Vector Machines – SVM), конволюционных нейронных сетей (Convolutional Neural Networks – CNN) и др. Например, для классификации изображений в дополнение к методам, представленным в таблицах 1-3 были проведены исследования по использованию метода TenzorFlow [7].

Программная среда ОнтоМАСТЕР-Ресурс предоставляет возможность визуального редактирования сценариев, раскрытия блоков и назначения онтологических имен классов, реализующих заданные операции. На рис. 1 изображен пример сценария обработки и классификации изображений.

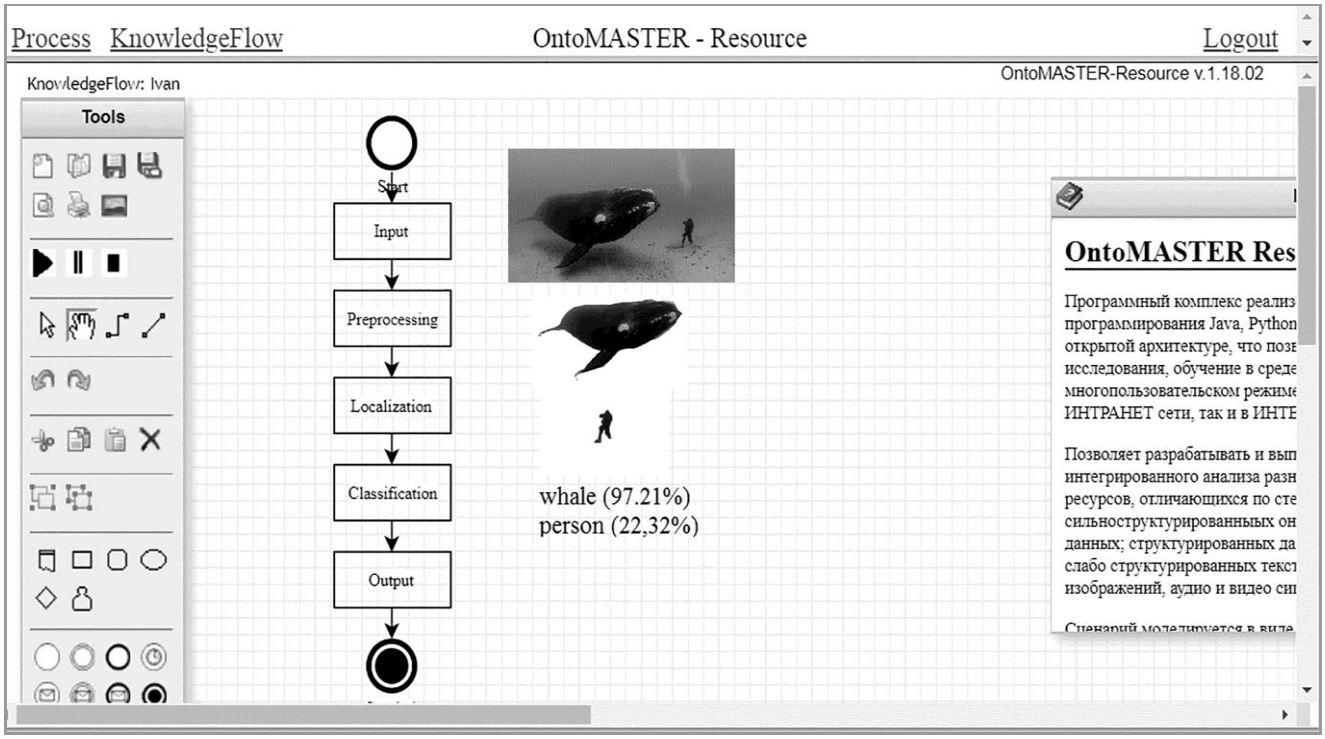

Рис. 1. Пример сценария обработки и классификации изображений

В табл. 1 перечислены методы, которые применены к численным (numeric) или символьным (symbolic) переменным (признакам).

ТАБЛИЦА I ПЕРЕЧЕНЬ МЕТОДОВ

№	Метод
1	AttributeSelectedClassifier
2	Bagging
3	BayesNet
4	CVParameterSelection
5	ClassificationViaRegression
6	DecisionStump
7	DecisionTable
8	FilteredClassifier
9	HoeffdingTree
10	IBk
11	InputMappedClassifier
12	J48
13	JRip
14	KStar
15	LMT
16	LWL
17	LibSVM
18	Logistic
19	LogitBoost
20	MultiClassClassifier
21	MultiClassClassifierUpdateable
22	MultiScheme
23	MultilayerPerceptron
24	NaiveBayes
25	NaiveBayesUpdateable
26	OneR
27	PART
28	REPTree
29	RandomCommittee
30	RandomForest
31	RandomSubSpace
32	RandomTree
33	SMO
34	SimpleLogistic
35	Stacking
36	Vote
37	WeightedInstancesHandlerWrapper
38	ZeroR

Методы, которые применены только к признакам в числовом виде, представлены в табл. 2.

ТАБЛИЦА II ПЕРЕЧЕНЬ МЕТОДОВ

№	Метод
1	AdaBoostM1
2	CostSensitiveClassifier
3	NaiveBayesMultinomial
4	NaiveBayesMultinomialUpdateable
5	SerializedClassifier

Методы, которые применены только к символьным переменным, представлены в табл. 3.

ТАБЛИЦА III ПЕРЕЧЕНЬ МЕТОДОВ

№	Метод
1	IterativeClassifierOptimizer
2	J48graft
3	NBTree
4	NaiveBayesMultinomialText
5	RandomizableFilteredClassifier
6	SGD
7	SGDText
8	SimpleCart
9	VotedPerceptron

IV. РЕЗУЛЬТАТЫ

Тестирование разработанного метода производилось на нескольких задачах классификации и регрессии.

Задача 1. Прогнозирование успешности обучения студентов. Известные оценки производительности экспертных систем (например, [10]) показывают, что возможно со степенью точности >70% прогнозировать эффективность учебной деятельности. На основе численных признаков *X*, полученных в результате диагностики индивидуальных когнитивных показателей студентов [13] и итоговых баллов академической успеваемости *Y* различными методами построены модели регрессионного анализа и классификации. Для классификации по результатам диагностики студенты были условно разделены на 3 группы: high (H), average (A) и special (S). Результаты экспериментов с обучающей выборкой (160 студентов) показали точность предсказания 100%, а с тестовой выборкой (60 студентов) 86%. Наилучшие результаты по точности, критерию AIC и критерию интерпретации получены по моделям деревьев решений (J48, LMT).

Задача 2. Классификация публикаций на примерах тематических областей знаний «Методы обработки изображений и сигналов в гидроакустике», «Инженерия знаний». Проведена предварительная обработка корпусов текстов: графематическая разметка, формирование частотного словаря, анализ частей речи и лемматизация, формирование векторной модели однословных и многословных терминов. Выбраны наиболее значимые по частоте признаки (более 5000). Тексты классифицированы с использованием нескольких моделей. Лучшими по критерию точности оказались деревья решений и байесовские модели. Приоритет отдан байесовским моделям, т. к. модели деревьев решений выделяли значительно меньшее число существенных признаков, что в последствии может привести к снижению точности классификации новых документов.

Задача 3. Классификация изображений на тестовом наборе «Image Segmentation Data Set» (UCI Repository of Machine Learning). Лучший результат 97.3 % показал метод Random Forest. На рис. 1 изображен результат классификации паттерна кита методом TenzorFlow с точностью 97.2%.

V. ЗАКЛЮЧЕНИЕ

Разработан метод автоматизированного анализа данных для решения задач классификации и регрессионного анализа. Из предлагаемых методов анализа данных выбирается лучший по комплексному критерию, что является особенностью данной системы. Данный подход позволяет сравнивать результаты, полученные разными моделями, учитывая не только точность, но и сложность модели (число параметров, признаков).

СПИСОК ЛИТЕРАТУРЫ

[1] Guyon I., Chaabane I., Escalante H. J., Escalera S., Jajetic D., Lloyd J. R., Statnikov A. A brief review of the ChaLearn AutoML challenge: any-time any-dataset learning without human intervention //Workshop on Automatic Machine Learning. 2016. Pp. 21-30.

[2] Olson R. S., Bartley N., Urbanowicz R. J., Moore J. H. Evaluation of a tree-based pipeline optimization tool for automating data science //Proceedings of the Genetic and Evolutionary Computation Conference 2016. ACM. 2016. pp. 485-492.

[3] Witten I. H., Frank E., Hall M. A., Pal C. J. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016. 629 p.

[4] Hanif M. H. M., Adewole K. S., Anuar N. B., Kamsin A. Performance Evaluation of Machine Learning Algorithms for Spam Profile Detection on Twitter Using WEKA and RapidMiner //Advanced Science Letters. – 2018. V. 24. No 2. Pp. 1043-1046.

[5] Fillbrunn A., Dietz C., Pfeuffer J., Rahn R., Landrum G. A., Berthold M. R. KNIME for reproducible cross-domain analysis of life science data //Journal of Biotechnology. 2017. V. 261. Pp. 149-156.

[6] Triguero I., González S., Moyano J. M., García S., Alcalá-Fdez J., Luengo J., Herrera F. KEEL 3.0: an open source software for multi-stage analysis in data mining //International Journal of Computational Intelligence Systems. 2017. V. 10. No 1. Pp. 1238-1249.

[7] Abadi M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M.TensorFlow: A System for Large-Scale Machine Learning //OSDI. 2016. V. 16. Pp. 265-283.

[8] Котова Е.Е., Писарев А.С., Писарев И.А. Программный комплекс анализа информационных ресурсов ОнтоМАСТЕР-Ресурс. Свидетельство о государственно регистрации программы для ЭВМ № 2018611107 от 24 января 2018 г.

[9] Strecht P., Cruz L., Soares C., Mendes-Moreira J., Abreu R. A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance //International Educational Data Mining Society. 2015. Pp.392-395

[10] Asif R., Merceron A., Ali S. A., Haider N. G. Analyzing undergraduate students' performance using educational data mining //Computers & Education. 2017. V. 113. Pp. 177-194.

[11] Burnham K. P., Anderson D. R. Multimodel inference: understanding AIC and BIC in model selection //Sociological methods & research. – 2004. V. 33. No. 2. Pp. 261-304.

[12] Pisarev I. A., Kotova E. E., Pisarev A. S., Stash N. V. Structure of knowledge base of methods for processing hydroacoustic signals //Young Researchers in Electrical and Electronic Engineering (EIConRus), 2018 IEEE Conference of Russian. IEEE, 2018. Pp. 1132-1135.

[13] Имаев Д.Х., Котова Е.Е. Оценка параметров динамических моделей обучаемых по результатам экспресс−диагностирования. // Известия СПбГЭТУ «ЛЭТИ». Современные технологии образования. 2015. № 1. С. 70–75.