Построение сценариев обработки многомерных пространственно-временных данных в условиях неопределенности

И. А. Писарев[1], Е. Е. Котова[2], А. С. Писарев[3]
Санкт-Петербургский государственный электротехнический
университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
[1]pisarevivan@yandex.ru, [2]eekotova@gmail.com, [3]a_pisarev@mail.ru

***Аннотация*. В статье представлен метод построения сценариев классификации объектов и траекторий их движения в реальном времени на основе обработки информации от большого числа пространственно-разнесенных источников информации (сенсоров) и анализа динамических массивов распределенных многомерных данных в условиях неопределенности. Метод реализован в адаптивной аналитической информационной системе. В приведенных примерах сценариев используются статистические, интервальные, нечеткие и стохастические методы описания неопределенности. При разработке сценариев адаптивной обработки и анализа данных применяются интегрированные онтологии областей знаний «гидроакустика», онтологии методов обработки сигналов, методов обработки изображений и методов анализа данных. Реализация сценариев автоматизированной обработки многомерных пространственно-временных данных осуществляется в сетевой программной среде OntoMASTER на основе стандартов семантического Web. Для поддержки проведения научных исследований реализуется динамический интерфейс построения сценариев и их тестирования в среде тренажера/имитатора. Среда обучения на основе предназначена для подготовки и переподготовки специалистов. Результаты работы могут быть применены как в научных исследованиях, так и в процессе обучения студентов.**

Ключевые слова: обработка информации; построение сценариев; условия неопределенности; вероятностные методы; классификация

I. Введение

В настоящее время актуальной проблемой является повышение производительности научных исследований на основе методов автоматизированного анализа данных.

Актуальность проблемы комплексной автоматизации всех процессов интеллектуального анализа данных подтверждается тем, что ее решению посвящен проект Discovery of Models, проводимый по инициативе управления перспективных исследовательских проектов (DARPA) [1].

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

Решение данной проблемы требует разработки эффективных инструментов для автоматизированной обработки и анализа больших объемов разнородных неструктурированных данных (видеоизображений, акустических сигналов) и ускорения процесса генерации новых знаний.

Сценарии поиска моделей и методов машинного обучения с оптимальной производительностью для конкретного набора данных включают выбор лучших моделей и оптимизацию метапараметров методов, например, количества деревьв в методе RandomForest [2, 3] или количества скрытых слоев в нейросетевой модели.

Проблема автоматизированного проектирования сценариев (workflow) процессов конвейерной (pipe) обработки и анализа данных исследуется при создании таких систем автоматизированного машинного обучения (Automated Machine Learning – AutoML), как RapidMiner [4, 5], WEKA [6], KNIME [7], KEEL [8], SAP Predictive Analytics, Matlab (MathSoft), TenzorFlow [9], Scikit-learn [10], Auto-WEKA [3, 11], Auto-Sklearn [12].

Важным направлением исследований в данной области является разработка сетевых сред для построения и автоматизированного выполнения сценариев комплексной автоматизированной обработки и интеллектуального анализа данных на основе стандартов семантического Web (Semantic Web) [13, 14].

В данной работе приведены результаты исследований по разработке метода построения сценариев автоматизированной обработки и интеллектуального анализа разнородных неструктурированных данных.

Метод реализован в сетевом программном комплексе OntoMASTER, разработанном авторами.

II. Метод

Метод автоматизации обработки и интеллектуального анализа данных включает построение и выполнение сценариев решения задач обработки данных и классификации. Сценарий представляется в виде многослойной сети составных и атомарных работ, которые могут выполняться человеком и/или интеллектуальными программными агентами.

Составные блоки (рис. 1) обозначают группу связанных работ и отображаются с символом «+» в левом верхнем угле интерфейса пользователя. Для детализации состава блока пользователю необходимо перейти в режим редактирования подсети. В свойствах атомарных блоков указываются уникальные обозначения из созданной онтологии процессов и параметры, необходимые для их выполнения.

Метод последовательного раскрытия неопределенности сценариев обработки и анализа данных реализован в сетевой программной среде OntoMASTER с динамическим интерфейсом на основе стандартов семантического Web [14]. Среда включает автоматизированную поддержку создания интегрированной онтологии областей знаний (например, области знаний «гидроакустика»), онтологии методов обработки и интеллектуального анализа сигналов, видеоизображений.

Реализация динамического интерфейса предназначена для организации среды поддержки научных исследований и среды обучения для подготовки и переподготовки специалистов на основе базы знаний (интегрированной онтологии) и среды тренажера/имитатора выполнения сценариев.

На рис.1 изображен пример сценария комплексной обработки и анализа видеоизображений, включающий этапы предобработки (сегментации, извлечения признаков объектов), локализации объектов в пространстве и времени, автоматической классификации объектов, трекинга траекторий движения, анализа, идентификации моделей движения и прогнозирования траекторий [14–16].

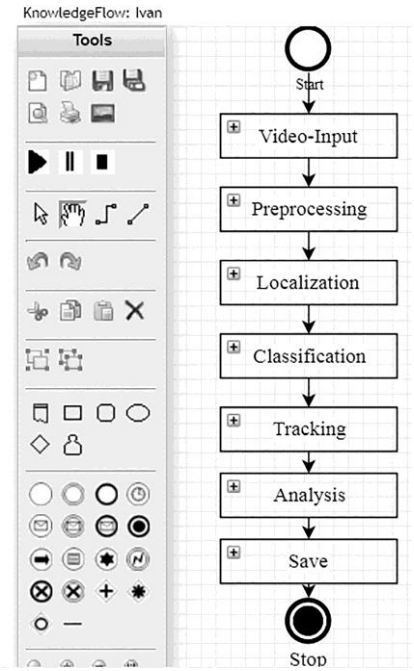

Рис. 1. Фрагмент обобщенного сценария обработки и анализа видеоизображений

Фрагмент сценария этапа построения модели классификации на основе обучающей выборки для заданного набора данных изображен на рис. 2. Список методов классификации, применимых к конкретному набору данных, формируется на основе запроса к онтологии, базе знаний и данных, в которых хранятся описания моделей и методов, их метапараметров, правила оценки применимости методов к структуре конкретных данных, сетевые адреса интеллектуальных агентов и сервисов.

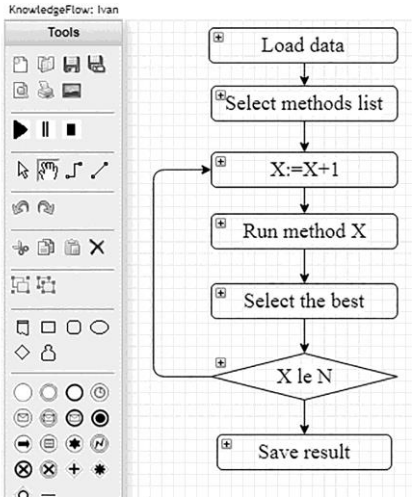

Рис. 2. Фрагмент детализации сценария построения и выбора модели классификации на основе обучающей выборки

Предусмотрены два варианта сценариев выполнения методов классификации. При автоматизированном выполнении сценария пользователю предоставляется возможность в Web интерфейсе изменять значения метапараметров методов, заданных по умолчанию. При выборе автоматического режима подключается блок поиска оптимальных значений метапараметров метода классификации.

Поиск оптимальных значений метапараметров осуществляется многократным решением обратных и прямых задач с применением подсистемы OntoMASTER, позволяющей на основе эвристических алгоритмов решать задачи глобальной оптимизации.

Задачи поиска оптимальных значений метапараметров методов анализа данных характеризуются наличием множества локальных экстремумов, нелинейностей, разнородных параметров (бинарных, целочисленных, вещественных, интервальных), неполнотой информации, большой размерностью. В подсистеме OntoMASTER применяются методы глобальной оптимизации с эвристическими стратегиями: эволюционные (Evolutionary) и генетические (Genetic) алгоритмы, алгоритмы роя (Particle swarm optimization) и колонии муравьев (Ant Colony optimization), алгоритм отжига (Simulated annealing), случайного поиска (Random optimization) и др.

Например, в число метапараметров метода RandomForest входят целочисленные и вещественные значения метапараметров: размер используемого фрагмента обучающего набора при формировании множества деревьев фрагмента в процентах от размера

обучающей выборки, число итераций, а в методе MultilayerPerceptron используются булевые (decay), целочисленные (batchSize) и вещественные метапараметры (learningRate).

Задача поиска значений метапараметров методов классификации формулируется в многокритериальной нечеткой постановке: найти отображение f^* (структуру модели и значения параметров настройки Z) по данным обучающей выборки при известных значениях независимых переменных (признаков) X и зависимой переменной (класса) Y, которое обеспечивает экстремум (минимум или максимум) функционала в виде аддитивной свертки целевых функций R_j при выполнении ограничений $R_i \in C$:

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

При поиске оптимальных значений метапараметров применяется интервальная шкала для задания ограничений на область поиска. Данный подход позволяет использовать в параллельном режиме несколько интеллектуальных агентов путем указания индивидуальных областей пространства поиска с последующим анализом и объединением полученных результатов.

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

Дополнительно при оценке качества метода классификации применяется модифицированный критерий Акаике, что позволяет сравнивать результаты, учитывая не только точность, но и сложность модели (число параметров, признаков).

Формула вычисления AIC_c с использованием полученной в результате классификации оценки RMSE имеет следующий вид для случаев $\mathrm{RMSE} > 0$ и $n - k > 1$ (при 100% точности AIC_c принимает условное большое отрицательное значение):

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

III. Результаты

Тестирование разработанного метода и программного обеспечения OntoMASTER проводилось на нескольких наборах данных. Результаты классификации сегментированных изображений «Image Segmentation Data Set» (UCI Machine Learning Repository) представлены в табл. 1 и 2. В таблицах представлены округленные значения показателей производительности (эффективности) решения задачи классификации различными методами. Общая выборка, включающая 1500 примеров, была разделена поровну на обучающий и тестовый наборы.

В соответствии со сценарием автоматического построения моделей классификации (рис. 2) были выбраны методы со списками метапараметров и их значениями по умолчанию. Возможность дополнительной ручной и автоматической настройки значений метапараметров реализована в Web интерфейсе системы OntoMASTER (рис. 3). Качество моделей проверялось по результатам классификации примеров из тестовой выборки.

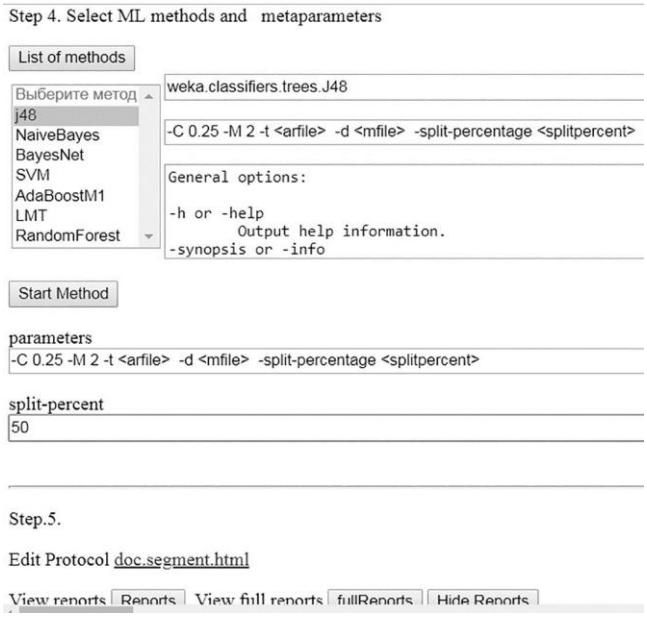

Рис. 3. Фрагмент интерфейса выбора методов и настройки значений метапараметров

В табл. I показан лучший результат классификации методом RandomForest как по точности – 97.33% правильно классифицированных примеров в тестовой выборке, так и по минимальному значению модифицированного критерия Акаике (AIC_c).

ТАБЛИЦА I РЕЗУЛЬТАТЫ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ

k	Name	N %	Aicc	Test %	Kappa	Mae	Rmse	Rae	Rrse
19	Random Forest	50	-3628.91	97.33	0.97	0.02	0.09	9.63	24.78
10	j48	50	-3106.93	94.4	0.93	0.02	0.12	7.73	35.55
18	LMT	50	-3299.44	94.13	0.93	0.02	0.11	8.74	30.94
19	Bayes Net	50	-2709.13	89.73	0.88	0.03	0.16	12.48	45.79

Затем был применен сценарий с дополнительной автоматической подстройкой метапараметров методов классификации. Наиболее точным оказался метод RandomForest – 97.47 %. На втором месте по точности оказался метод AdaBoostM1 – 96.4%.

Для сравнения полученных результатов была выполнена автоматическая классификация данных с помощью программного средства Auto-Weka [3], которая показала значение точности классификации 96.4%. Таким образом, метод и программная Web среда OntoMASTER

подтвердили свою эффективность при выполнении различных сценариев автоматической классификации.

ТАБЛИЦА II РЕЗУЛЬТАТЫ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ С ОПТИМИЗАЦИЕЙ ЗНАЧЕНИЙ МЕТАПАРАМЕТРОВ

k	Name	N %	Aicc	Test %	Kappa	Mae	Rmse	Rae	Rrse
19	Random Forest	50	-3622.01	97.47	0.97	0.02	0.09	9.63	24.78
19	AdaBoost M1	50	-3461.51	96.4	0.96	0.01	0.1	4.19	27.66

При тестировании программного обеспечения OntoMASTER использовался набор данных «Sonar» (UCI Machine Learning Repository) для классификации объектов по факторам, полученным в результате обработки измерений акустических сигналов [17]. В соответствии со сценарием автоматизированного построения моделей классификации (рис. 2) были выбраны методы со списками метапараметров и их значениями по умолчанию. Лучший результат по точности классификации (88.46%) на тестовой выборке был получен с помощью метода lazyIBk. В сценарии с дополнительной автоматической подстройкой метапараметров методов классификации наиболее точным оказался метод AdaBoostM1 – 89.42 %.

В сравнении с известными решениями задачи классификации данных «Sonar», не превышающими по точности 83% [18], с помощью разработанного авторами статьи метода и программных средств OntoMASTER [19] получены более точные результаты.

IV. Выводы

Разработан метод построения сценариев автоматизированного анализа данных для решения задач классификации. Многоуровневые сценарии объединяют операции обработки неструктурированных данных и машинного обучения, выполняемые человеком и интеллектуальными программными агентами, что позволяет последовательно раскрывать неопределенность и детализировать операции для достижения конечных целей исследований.

Из реализованных методов анализа данных выбирается лучший с автоматическим поиском оптимальных значений метапараметров. Подход, основанный на модифицированном критерии Акаике, позволяет сравнивать результаты, полученные на различных моделях, учитывая не только точность, но и сложность модели (число параметров, фактор-признаков).

Сравнение полученных результатов применения сценариев автоматической обработки и анализа данных с решениями, полученными с помощью других известных методов подтверждают эффективность разработанного метода, реализованного в сетевом программном комплексе OntoMASTER.

Результаты работы могут быть использованы как в научно-исследовательской деятельности для повышения эффективности научно-исследовательских, опытно-конструкторских работ, так и в учебном процессе.

Список литературы

[1] Brazdil P., Giraud-Carrier C. Metalearning and Algorithm Selection: progress, state of the art and introduction to the 2018 Special Issue. Machine learning. 2018. V.107. Pp. 1-14

[2] Breiman L. Random forests . Machine learning. 2001.V. 45. N. 1. Pp. 5-32.

[3] Kotthoff L., Thornton C., Hoos H., Hutter F., Leyton-Brown K. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. JMLR. 2017. 18(25). Pp. 1−5.

[4] Hofmann M., Klinkenberg R. RapidMiner: Data mining use cases and business analytics applications. Chapman and Hall/CRC Press. 2013. 518 p.

[5] Kietz J. U., Serban F., Bernstein A., Fischer S. Designing kdd-workflows via htn-planning for intelligent discovery assistance //5 th Planning to learn workshop WS28 at ECAI 2012. 2012. Pp.10-17.

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

[7] 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.

[8] 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.

[9] 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.

[10] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Vanderplas J. Scikit-learn: Machine learning in Python //Journal of machine learning research. 2011. V. 12. N. 10. Pp. 2825-2830.

[11] Thornton C., Hutter F., Hoos H. H., Leyton-Brown K. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms //Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. 2013. Pp. 847-855.

[12] Feurer M., Klein A., Eggensperger K., Springenberg J., Blum M., Hutter F. Efficient and robust automated machine learning. //In Advances in neural information processing systems. 2015. Pp. 2962-2970.

[13] Yao Y., Xiao Z., Wang B., Viswanath B., Zheng H., Zhao B. Y. Complexity vs. performance: empirical analysis of machine learning as a service //Proceedings of the 2017 Internet Measurement Conference. – ACM, 2017. Pp. 384-397.

[14] Pisarev I. A., Kotova E. E., Pisarev A. S., Stash N. V. Development of Scenarios for Automatic Processing and Data Mining in a Multi-Agent Environment //2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus). IEEE. 2019. Pp. 630-633.

[15] Pisarev A. S. Method for the parameters estimating of microparticles motion along the trajectories under uncertainty //2015 XVIII International Conference on Soft Computing and Measurements (SCM). IEEE. 2015. Pp. 211-213.

[16] Pisarev A. S., Rukolaine S. A., Samsonov A. M., Samsonova M. G. Numerical analysis of particle trajectories in living cells under uncertainty conditions //Biophysics. 2015. V. 60. N. 5. Pp. 810-817.

[17] Gorman R. P., Sejnowski T. J. Analysis of hidden units in a layered network trained to classify sonar targets //Neural networks. 1988. V. 1. N. 1. Pp. 75-89

[18] Bennasar M., Hicks Y., Setchi R. Feature selection using joint mutual information maximisation //Expert Systems with Applications. 2015. V. 42. N. 22. Pp. 8520-8532.

[19] Котова Е.Е., Писарев А.С., Писарев И.А. Программный комплекс разработки сценариев анализа данных ОнтоМАСТЕР-Сценарий. Свидетельство о государственной регистрации программы для ЭВМ № 2019613556 от 19 марта 2019 г.