DOI: https://doi.org/10.15407/rpra22.04.270

УДК 004.048; 524.6

PACS numbers: 07.05.Tp, 98.35.-a

А. А. ГОРБУНОВ [1], Е. А. ИСАЕВ [1,2], В. А. САМОДУРОВ [1,2]

[1] Национальный исследовательский университет “Высшая школа экономики”, ул. Мясницкая, 20, г. Москва, 101000, Россия

[2] Пущинская Радиоастрономическая обсерватория АКЦ ФИАН, г. Пущино, Московская обл., 142290, Россия

E-mail: agorbunov@hse.ru, eisaev@hse.ru, vsamodurov@hse.ru

ПРИМЕНЕНИЕ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ КЛАССИФИКАЦИИ БОЛЬШИХ ОБЪЕМОВ АСТРОНОМИЧЕСКИХ ДАННЫХ

Предмет и цель работы: *В процессе астрономических наблюдений собираются огромные объемы данных. БСА ФИАН (Большая сканирующая антенна Физического института Российской академии наук), используемая при исследовании импульсных явлений, ежедневно регистрирует 87.5 Гб данных (32 Тб в год). Целью представленной работы является разработка веб-сервиса для помощи экспертам в классификации новых астрономических наблюдений. Студия машинного обучения Azure Machine Learning Studio, поддерживающая алгоритм глубокой нейронной сети, используется в качестве инструмента для разработки веб-сервиса.*

Методы и методология: *Экспертами классифицированы 83096 индивидуальных наблюдений (на отрезке исследования июль 2012 – октябрь 2013). Свыше 89 % выборки соответствуют пульсарам, мерцающим источникам и быстрым радиотранзиентам, а остальные классы наблюдений относятся к аппаратурным сбоям, помехам, пролету спутника Земли, самолета. Всего выделено 15 классов наблюдений.*

Результаты: *Наличие подобной выборки, разделенной на классы, позволяет воспользоваться алгоритмами машинного обучения, с помощью которых станет возможной разработка автоматизированного сервиса для краткосрочного/долгосрочного мониторинга различных классов радиоисточников (в том числе радиотранзиентов различной природы), мониторинга ионосферы Земли, межпланетной и межзвездной плазмы, поиска и мониторинга различных классов радиоисточников. Под мониторингом в данном случае понимается автоматическая фильтрация и распознавание ранее неклассифицированных импульсных явлений. На текущий момент для автоматической фильтрации используются методы статистического анализа. В работе рассматривается альтернативный метод с использованием алгоритма машинного обучения – нейронной сети, которая обрабатывает поданные на вход первичные данные и, после обработки скрытым слоем, посредством выходного слоя определяет класс импульсного явления.*

Заключение: *Создание модели нейронной сети, обученной на выборке и выполняющей классификацию ранее неклассифицированных импульсных явлений, производится с помощью облачного сервиса Microsoft Azure Machine Learning Studio. Веб-сервис, созданный на основании модели, позволяет классифицировать как одиночные импульсные явления в режиме реального времени (запрос-ответ), так и выборку данных за определенный период (пакетная обработка).*

Ключевые слова: *большие данные, глубокие нейронные сети, классификация импульсных явлений*

1. Введение

В процессе астрономических наблюдений собираются огромные объемы данных. Радиотелескоп БСА ФИАН (Большая сканирующая антенна Физического института Российской академии наук), используемый при исследовании импульсных явлений, ежедневно регистрирует 87.5 Гб данных (32 Тб в год). Экспертами [1] классифицированы 83096 индивидуальных наблюдений (на отрезке исследования июль 2012 г. – октябрь 2013 г.). Свыше 89 % выборки соответствуют наблюдениям пульсаров, мерцающих источников и быстрых радиотранзиентов, а остальные классы наблюдений относятся к аппаратурным сбоям, помехам, пролету спутника Земли, самолета. Всего выделено 15 классов наблюдений.

Наличие подобной выборки, разделенной на классы, позволяет воспользоваться алгоритмами машинного обучения, с помощью которых станет возможной разработка автоматизированного сервиса для краткосрочного/долгосрочного мониторинга различных классов радиоисточников (в том числе радиотранзиентов различной природы), мониторинга ионосферы Земли, межпланетной и межзвездной плазмы, поиска и мониторинга различных классов радиоисточников. Под мониторингом в данном случае понимается автоматическая фильтрация и распоз-

нование ранее неклассифицированных импульсных явлений.

На текущий момент для автоматической фильтрации используются методы статистического анализа [1, 2].

Цель исследования – разработать веб-сервис, позволяющий автоматизировать процесс классификации больших объемов астрономических данных по импульсным явлениям.

Основные задачи, решаемые для достижения цели и определившие логику и структуру настоящей статьи:

– подготовка обучающей выборки на основе данных об импульсных явлениях;

– создание модели нейронной сети, обученной на выборке и выполняющей классификацию;

– разработка веб-сервиса с использованием инструментов Microsoft Azure Machine Learning Studio.

2. Методы и результаты исследований

Научное исследование, связанное с разработкой автоматизированного сервиса классификации импульсных явлений, проводилось в соответствии с двухшаговой моделью (рис. 1).

Базой для проведения исследования, описываемого в настоящей статье, является выборка данных о космических объектах, полученная в результате ручной классификации группой экспертов [1]. Процесс классификации, сбора данных, методы и результаты исследований подробно описываются в [1] и приведены в шаге 1 модели исследований (рис. 1).

Настоящая работа отражает результаты шага 2 (рис. 1). При исследовании были использованы алгоритм машинного обучения (глубокая нейронная сеть), метод отбора признаков для формиро-

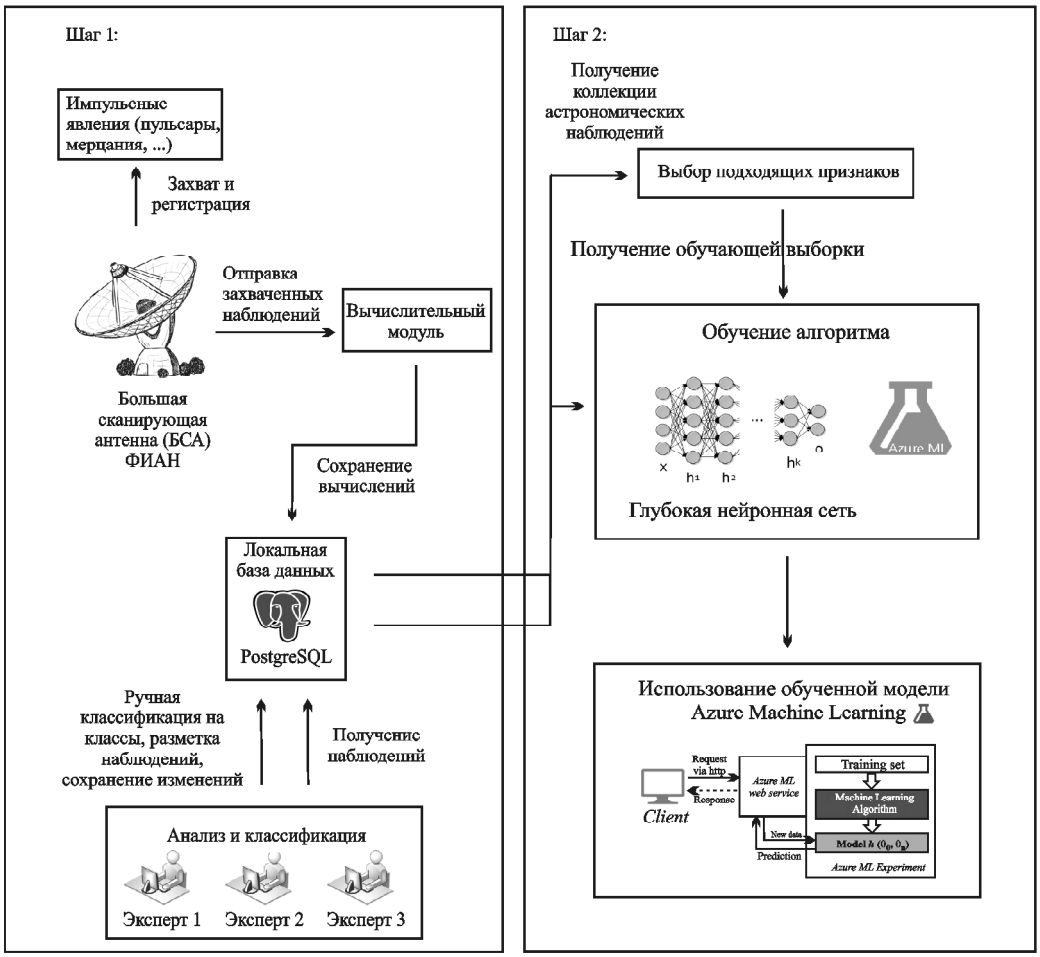

Рис. 1. Модель исследования

вания обучающей выборки, метод обучения глубокой нейронной сети с учителем [3] на основе сформированной обучающей выборки, метод перекрестной проверки на независимых данных. Для разработки и развертывания веб-сервиса автоматической классификации были использованы также возможности облачной инфраструктуры Microsoft Azure Machine Learning Studio.

Результатом этого исследования являются:

– обученная на выборке модель глубокой нейронной сети (рис. 2);

– веб-сервис, который использует обученную модель глубокой нейронной сети, помогает экспертам в классификации единичных импульсных явлений (в режиме реального времени), а также выборки, собранной за определенный период (пакетная обработка).

3. Обсуждение результатов

Для обучения глубокой нейронной сети выборка астрономических наблюдений (в период с июля 2012 г. по октябрь 2013 г.), полученная путем экспертной классификации, была поделена на две части: обучающая выборка (2013 г.) и данные для проверки точности модели после обучения (2012 г.). В качестве периода обучающей выборки был выбран 2013 г., так как именно в течение этого промежутка времени наблюдались все типы разнообразных импульсных явлений (пульсары, мерцания с разной продолжительностью и т. д.). Обучающая выборка за 2013 г. была разделена на обучающий (75 % данных) и тестовый (25 % данных) наборы. Оценка работы модели после обучения производилась с использованием тесто-

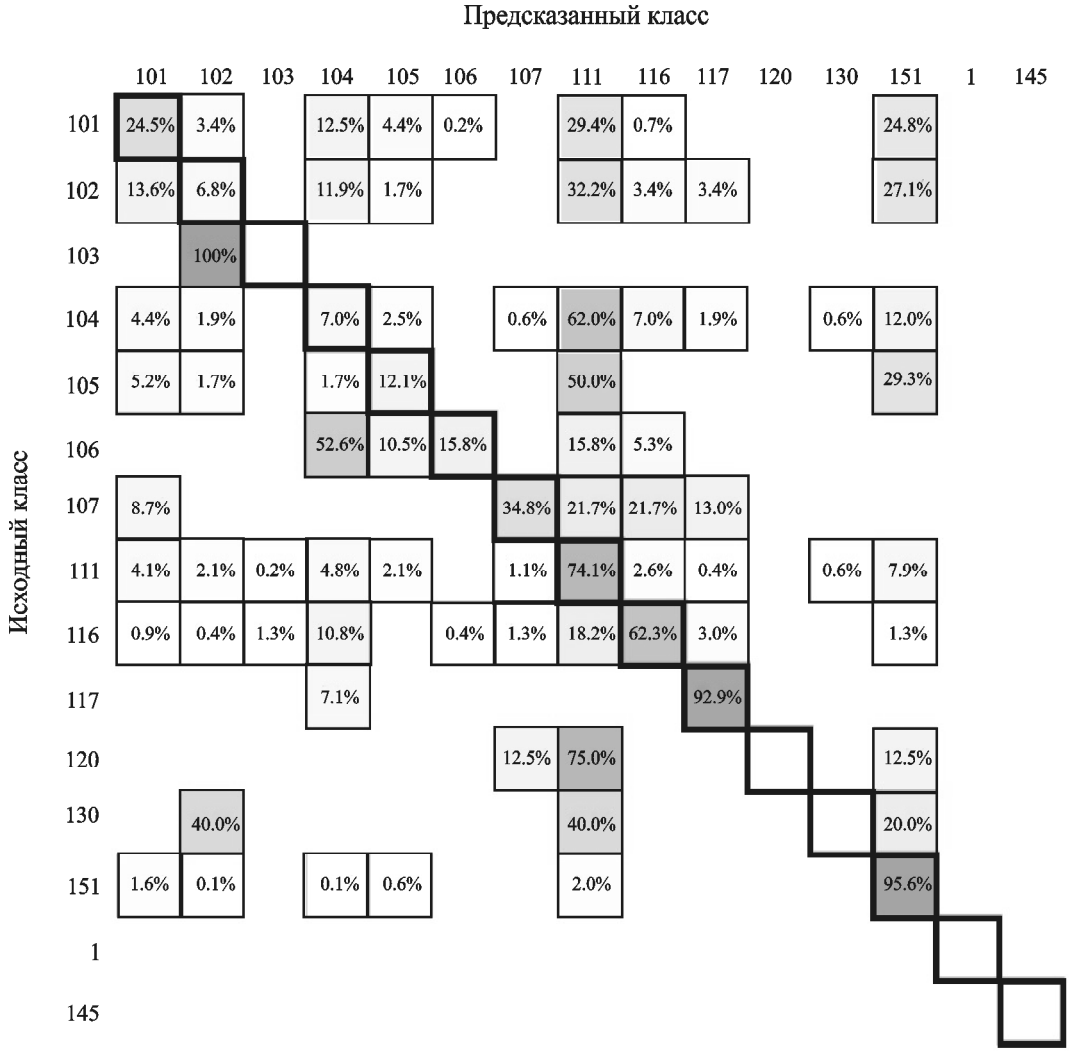

Рис. 2. Матрица неточностей (тестовый набор)

вого набора, результаты оценки приведены в матрице неточностей (confusion matrix) (рис. 2).

В строках матрицы – исходный класс импульсного явления, который подавался на вход модели глубокой нейронной сети; в столбцах – класс, предсказанный с помощью алгоритма машинного обучения.

На рис. 2 видно, что самые точные результаты классификации (95.6 %) представлены для объектов "151" – пульсаров, затем идут объекты "117" – мерцания долгопериодические, с периодом более 5 с (ионосфера), (92.9 %); "111" – мерцания короткопериодические, с периодом менее 1 с, (74.1 %); "116" – мерцания со средним периодом, 1÷5 с, (62.3 %).

При использовании модели для классификации новых импульсных явлений (рис. 3) также отмечается хороший процент распознавания объектов "151" – пульсаров (99.2 %), затем идут объекты "116" – мерцания со средним периодом, 1÷5 с, (77.7 %).

Результаты, представленные в матрице неточностей для тестового набора (рис. 2), получены после процедуры перекрестной оценки (cross-validation) работоспособности модели нейронной сети на независимых данных. Таким образом, исключается ситуация, при которой алгоритм показывает свою эффективность только на определенных данных.

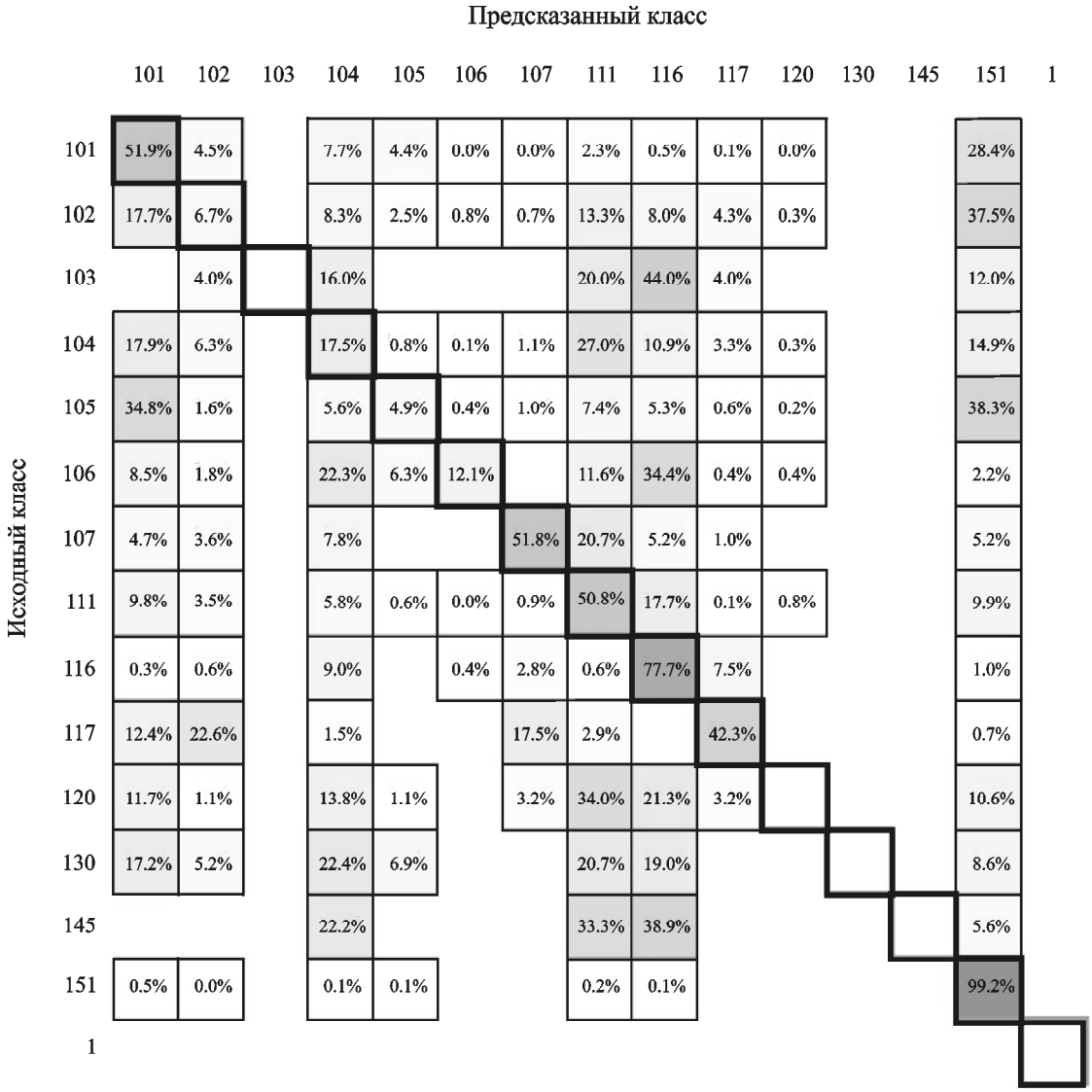

Рис. 3. Матрица неточностей (новые наблюдения)

Разработанная модель может быть использована для поиска пульсаров, мониторинга ионосферы Земли, поиска некоторых классов радиоисточников. Ожидается, что повышение точности классификации модели будет достигнуто после формализации процедуры принятия решения, которую выполняет каждый эксперт на основе своих накопленных знаний каждый раз, когда выполняет вручную классификацию того или иного импульсного явления. Таким образом, предполагается, что добавление новых признаков в обучающую выборку, знания из которой заложены в глубокую нейронную сеть, позволит улучшить способности модели к классификации.

4. Выводы

В работе отражены результаты исследования, связанного с разработкой автоматизированного веб-сервиса для классификации больших объемов наблюдений различных космических объектов на основе классификации потоковых данных об импульсных явлениях с помощью глубокой нейронной сети.

В соответствии с разработанной моделью исследования (рис. 1, шаг 2) выборка данных, полученная путем ручной классификации группой экспертов больших данных с помощью методов статистического анализа, была подготовлена для подачи на вход алгоритма машинного обучения, в данном случае глубокой нейронной сети. В мероприятия по подготовке обучающей выборки входил отбор признаков, тех колонок таблицы выборки, на основе которых возможно классифицировать поданное на вход импульсное явление. Для работы с алгоритмом глубокой нейронной сети использовалась облачная инфраструктура Microsoft Azure Machine Learning Studio. Обученная с помощью данного алгоритма модель показала свою эффективность на тестовой выборке для некоторых классов наблюдений (рис. 2) (для пульсаров – более 95 %) и при классификации новых ранее не классифицированных наблюдений (рис. 3) (для пульсаров – почти 100 %). Таким образом, погрешность работы алгоритма при классификации некоторых классов наблюдений составляет не более 5 % – результат значительно лучше, чем при автоматическом предварительном разборе импульсов (в среднем 15 % [1]), и сравнимый либо лучше, чем при ручном разборе данных. Ожидается, что повышение точности классификации модели будет достигнуто после формализации процедуры принятия решения, которую выполняет каждый эксперт на основе своих накопленных знаний каждый раз, когда производит вручную классификацию того или иного импульсного явления.

С использованием инструментов облачной инфраструктуры разработан веб-сервис, с помощью которого обученная модель глубокой нейронной сети может классифицировать импульсные явления в режимах потоковой (в реальном времени) и пакетной (выборка за определенный период) обработок. В качестве перспективной области использования разработанного веб-сервиса можно отметить программы поиска и краткосрочного/долгосрочного мониторинга различных классов радиоисточников (в том числе радиотранзиентов различной природы), мониторинга ионосферы Земли, межпланетной и межзвездной плазмы.

СПИСОК ЛИТЕРАТУРЫ

1. *Samodurov V. A., Dumsky D. V., Isaev E. A., Rodin A. E., Kazancev A. N., Fedorova V. A., and Belyatskij Yu. A.* The daily 110 MHz radio wave sky survey: statistical analysis of impulse phenomena from observation in 2012-2013 // Odessa Astronomical Publications. – 2016. – Vol. 29. – P. 167–170. DOI: 10.18524/1810-4215.2016.29.85206
2. *Taylor G. B., Ellingson S. W., Kassim N. E., Craig J., Dowell J., Wolfe C. N., Hartman J., Bernardi G., Clarke T., Cohen A., Dalal N. P., Erickson W. C., Hicks B., Greenhill L. J., Jacoby B., Lane W., Lazio J., Mitchell D., Navarro R., Ord S. M., Pihlström Y., Polisensky E., Ray P. S., Rickard L. J., Schinzel F. K., Schmitt1 H., Sigman E., Soriano M., Stewart K. P., Stovall K., Tremblay S., Wang D., Weiler K. W., White S., and Wood D. L.* First Light for the First Station of the Long Wavelength Array // J. Astron. Instrum. – 2012. – Vol. 1, No. 1. – id. 1250004. DOI: 10.1142/S2251171712500043
3. *Roman V. Ş. and Buiu C.* Automatic Analysis of Radio Meteor Events Using Neural Networks // Earth Moon Planets. – 2015. – Vol. 116, Is. 2. – P. 101–113. DOI: 10.1007/s11038-015-9473-y

REFERENCES

1. SAMODUROV, V. A., DUMSKY, D. V., ISAEV, E. A., RODIN, A. E., KAZANCEV, A. N., FEDOROVA, V. A. and BELYATSKIJ, YU. A., 2016. The daily 110 MHz radio wave sky survey: statistical analysis of impulse phenomena from observation in 2012-2013. *Odessa Astronomical Publications.* vol. 29, pp. 167–170. DOI: 10.18524/1810-4215.2016.29.85206
2. TAYLOR, G. B., ELLINGSON, S. W., KASSIM, N. E., CRAIG, J., DOWELL, J., WOLFE, C. N., HARTMAN, J., BERNARDI, G., CLARKE, T., COHEN, A., DALAL, N. P., ERICKSON, W. C., HICKS, B., GREENHILL, L. J., JA-

COBY, B., LANE, W., LAZIO, J., MITCHELL, D., NAVARRO, R., ORD, S. M., PIHLSTRÖM, Y., POLISENSKY, E., RAY, P. S., RICKARD, L. J., SCHINZEL, F. K., SCHMITT, H., SIGMAN, E., SORIANO, M., STEWART, K. P., STOVALL, K., TREMBLAY, S., WANG, D., WEILER, K. W., WHITE, S. and WOOD, D. L., 2012. First Light for the First Station of the Long Wavelength Array. *J. Astron. Instrum.* vol. 1, no. 1, id. 1250004. DOI: 10.1142/S2251171712500043

3. ROMAN, V. Ş. and BUIU, C., 2015. Automatic Analysis of Radio Meteor Events Using Neural Networks. *Earth Moon Planets.* vol. 116, is. 2, pp. 101–113. DOI: 10.1007/s11038-015-9473-y

A. A. Gorbunov [1], E. A. Isaev [1,2], and V. A. Samodurov [1,2]

[1] National Research University Higher School of Economics, 20, Myasnitskaya St., Moscow, 101000, Russia

[2] Pushchino Radio Astronomy Observatory of the P. N. Lebedev Physical Institute Astro Space Center, RAS, Pushchino, 142290, Russia

APPLICATION OF DEEP LEARNING NEURAL NETWORK FOR CLASSIFICATION OF BIG DATA OF ASTRONOMIC OBSERVATIONS

Purpose: In the process of astronomical observations vast amounts of data are collected. The BSA (Big Scanning Antenna) used in the study of impulse phenomena, daily logs 87.5 GB of data (32 TB per year). The aim of this work is to develop the web-service which assists the experts with classification of new astronomic observations. The Azure Machine Learning Studio which offers a Deep Neural Network algorithm is used as a tool for web-service developing.

Design/methodology/approach: Experts classified 83096 individual observations (on the segment of the study July 2012 – October 2013). Over 89 % of the sample correspond to pulsars, twinkling springs and rapid radiotransmitters, and all other classes of observations belong to hardware failures, interference, the flight of the Earth satellite and aircraft. There were allocated 15 classes of observations.

Findings: Such a sample, divided into classes allows using the machine learning algorithms. It has become possible to develop an automated service for short-term/long-term monitoring of various classes of radio sources (including radiotransmitted of different nature), monitoring the Earth's ionosphere, the interplanetary and the interstellar plasma, search and monitoring of different classes of radio sources. Monitoring in this case refers to the automatic filtering and detection of earlier unclassified impulse phenomena. Currently, for automatic filtering, statistical analysis methods are used. This paper considers an alternative method supposed to be using neural network machine learning algorithm which processes the input into raw data, and after processing by the hidden layer through the output layer determines the class of pulse phenomena.

Conclusions: Creating a neural network model, trained on a sample and classifying earlier unclassified impulse phenomena, is performed using the cloud service Microsoft Azure Machine Learning Studio. The Web service, created based on the aforesaid model, allows classifying both single impulse phenomena in real time (Request / Reply) and data sampling for a certain period (Batch processing).

Key words: big data, deep neural networks, impulse phenomena classification

А. А. Горбунов [1], Е. А. Ісаєв [1,2], В. О. Самодуров [1,2]

[1] Національний дослідницький університет "Вища школа економіки",
вул. М'ясницька, 20, м. Москва, 101000, Росія

[2] Пущинська Радіоастрономічна обсерваторія АКЦ ФІАН, м. Пущино, Московська обл., 142290, Росія

ЗАСТОСУВАННЯ ГЛИБОКИХ НЕЙРОННИХ МЕРЕЖ ДЛЯ КЛАСИФІКАЦІЇ ВЕЛИКИХ ОБСЯГІВ АСТРОНОМІЧНИХ ДАНИХ

Предмет і мета роботи: У процесі астрономічних спостережень накопичуються величезні обсяги даних. ВСА ФІАН (Велика скануюча антена Фізичного інституту Російської академії наук), яка використовується у дослідженні імпульсних явищ, щодня реєструє 87.5 Гб даних (32 Тб щороку). Метою роботи є розробка веб-сервісу для допомоги експертам у класифікації нових астрономічних спостережень. Студія машинного навчання Azure Machine Learning Studio, що підтримує алгоритм глибокої нейронної мережі, використовується як інструмент для розробки веб-сервісу.

Методи і методологія: Експертами класифіковано 83096 індивідуальних спостережень (на відрізку дослідження липень 2012 – жовтень 2013). Понад 89 % вибірки відповідають пульсарам, мерехтливим джерелам і швидким радіотранзієнтам, а решта класів спостережень відносяться до апаратурних збоїв, перешкод, прольоту супутника Землі, літака. Всього виділено 15 класів спостережень.

Результати: Наявність подібної вибірки, поділеної на класи, дозволяє скористатися алгоритмами машинного навчання, за допомогою яких уможливиться розробка автоматизованого сервісу для короткострокового довгострокового моніторингу різних класів радіоджерел (у тому числі радіотранзієнтів різної природи), моніторингу іоносфери Землі, міжпланетної та міжзоряної плазми, пошуку й моніторингу різних класів радіоджерел. Моніторингом у даному разі розуміємо автоматичну фільтрацію і розпізнавання раніше некласифікованих імпульсних явищ. Наразі для автоматичної фільтрації використовуються методи статистичного аналізу. В роботі розглядається альтернативний метод з використанням алгоритму машинного навчання – нейронної мережі, яка обробляє подані на вхід первинні дані і, після обробки прихованим шаром, за допомогою вихідного шару визначає клас імпульсного явища.

Висновок: Створення моделі нейронної мережі, що навчена на вибірці та виконує класифікацію раніше некласифікованих імпульсних явищ, виконується за допомогою хмарного сервісу Microsoft Azure Machine Learning Studio. Веб-сервіс, створений на основі моделі, дозволяє класифікувати як поодинокі імпульсні явища в режимі реального часу (запит-відповідь), так і вибірку даних за певний період (пакетна обробка).

Ключові слова: великі дані, глибокі нейронні мережі, класифікація імпульсних явищ

Статья поступила в редакцию 20.10.2017