2. Операционная система специального назначения Astra Linux Special Edition. [Электронный ресурс]. URL: http://www.cio-sibir.ru/files/Meet/2016/2016-10-07-Astra_Linux.pdf (дата обращения: 11.05.2020).

3. Галгали П., Гайтонде Р. Сравнение систем безопасности в AIX, Linux и Solaris // IBM developerWorks. 15.07.2007. [Электронный ресурс]. URL: https://www.ibm.com/developerworks/ru/library/au-compaixsolaris/index.html (дата обращения: 11.05.2020).

4. Исследование уровня безопасности операционной системы Linux. [Электронный ресурс]. URL: https://www.bestreferat.ru/referat-52957.html (дата обращения: 11.05.2020).

5. Ивашко Е. Система мандатного контроля доступа Smack // IBM developerWorks. 26.10.2010. [Электронный ресурс]. URL: https://www.ibm.com/developerworks/ru/library/l-apparmor-6/ (дата обращения: 11.05.2020).

6. ГОСТ Р 58256-2018. Управление потоками информации в информационной системе. Формат классификационных меток. Изд. офиц. М.: Стандартинформ, 2018. 8 с.

7. Девянин П. Модели безопасности компьютерных систем. Управление доступом и информационными потоками. 2-е изд., перераб. и доп. М.: Горячая линия–Телеком, 2013. 338 с.

УДК 004
doi:10.18720/SPBPU/2/id20-206

Сараджишвили Сергей Эрикович[1],
канд. техн. наук, доцент, доцент;
Морозов Юрий Алексеевич[2],
аспирант

ОСОБЕННОСТИ ОБУЧЕНИЯ НЕЙРОННЫХ СЕТЕЙ С ИСПОЛЬЗОВАНИЕМ LINKED OPEN DATA

[1, 2] Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Россия,
[1] SSaradg@yandex.ru, [2] stonefiz@gmail.com

Аннотация. В работе рассматриваются особенности обучения нейронных сетей с использованием открытых связанных данных. В рамках исследования проведен обзор публикаций, посвященных вопросам в этой области. В результате был описан подход обработки связанных данных для дальнейшего обучения и проведено тестовое обучение.

Ключевые слова: связанные открытые данные, машинное обучение, семантический веб, RDF.

Sergey E. Saradzhishvili[1],
Associate Professor;
Yuri A. Morozov[2],
Postgraduate

FEATURES OF LEARNING NEURAL NETWORKS USING LINKED OPEN DATA

[1, 2] Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russia,
[1] SSaradg@yandex.ru, [2] stonefiz@gmail.com

Abstract. The paper discusses the features of training neural networks using open linked data. The study reviewed publications on issues in this area. As a result, an approach to processing related data for further training was described and test training was conducted.

Keywords: linked open data, machine learning, semantic web, RDF.

Введение

Linked Open Data (LOD) или связанные открытые данные – это одна из самых мощных структур для хранения данных, а машинное обучение – одна из самых популярных парадигм для анализа данных. Несмотря на то, что за последние десять лет в обеих областях наблюдался рост популярности, их объединению уделяется относительно мало внимания.

Появление взаимосвязанных, физически распределенных и автономно поддерживаемых хранилищ LOD открывает возможности для прогнозирования и обнаружения знаний из таких данных.

Связанные данные являются результатом слияния более ранних идей и технологий, включая гипертекст, базы данных, онтологии, языки разметки и являются частью такой концепции как семантическая паутина (Semantic Web) [1]. Для того, чтобы быстро ознакомиться с основными источниками данных, достаточно посмотреть на известное облако связанных данных (Linked Data Cloud) [2]. На нем наибольшее количество данных, посвящены научным публикациям, затем следуют источники данных по биологии, открытые государственные данные и медиаинформация.

При классическом подходе машинного обучения предполагается, что каждый метод соответствует стандартному шаблону: входные данные представляют собой таблицу примеров, описываемых несколькими функциями с целевым значением для прогнозирования, а выходные данные представляют собой модель, предсказывающие целевое значение.

Однако классические подходы к машинному обучению ограничены в их применимости, поскольку собирать все данные в централизованном месте для анализа нежелательно и нецелесообразно из-за доступа, памяти, пропускной способности, вычислительных ограничений, безопасности и конфиденциальности. Одним из вариантов решения этих проблем является способ обучения моделей из хранилищ связанных данных.

Ряд различных методов машинного обучения могут применяться к связанным данным для различных целей. Основной причиной, по которой имеет смысл использовать эти данные является их большое количество, опубликованных в общем доступе, для работы с которыми могут быть использованы стандарты семантического веба.

При использовании связанных данных для машинного обучения, учитывая их нетипичную структуру возникает проблема, каким образом обратиться к ним для того, чтобы использовать в задачах обучения нейронных сетей и как организовать процесс предобработки для последующего использования. В этой статье, опираясь на существующие методы работы с RDF мы опишем свой подход для обучения нейронных сетей из хранилищ, связанных данных.

1. Обзор литературы

Многие существующие исследования и подходы (например, [3 – 5]) предлагают использовать целый набор различных технологий из стека Semantic Web – запросы SPARQL, онтологии, RDF и др. для взаимодействия с данными. Одним из главных недостатков является, что многие из них предполагают ручную разработку процедур выборки нужных данных, приводящую к формулированию разработчиком или исследователем запроса SPARQL для обработки структур RDF.

При реализации алгоритмов с таким подходом они будут громоздкими и требовать знаний языка запросов SPARQL, их код будет избыточным, а реализация будет выглядеть примерно так, как показано на рисунке ниже.

Однако эти подходы являются громоздкими и требуют обширных и избыточных знаний для работы исследователя.

Некоторые подходы предлагают наиболее простой путь для разработчика [6] напрямую взаимодействовать со слоем RDF, который является самым низким уровнем в стеке семантического веба. Избегая работы с запросами SPARQL и другими уровнями, знания о которых не являются приоритетными для специалиста в области машинного обучения.

```
rs = execute_query ("""
SELECT ? degree ( COUNT (? degree ) AS ? count )
WHERE {
{
    SELECT ( COUNT (?n) AS ? degree )
    WHERE
    {
        {
            ?n ? out_edge ? out
        }
    UNION
    {
        ? in ? in_edge ?n}
    }
    GROUP BY ?n
    }
}
GROUP BY ? degree """)
print_result ( rs )
#
# Example Result :
# ------------------
# | degree | count |
# ==================
# | 3 | 2 |
# | 4 | 3 |
# | 2 | 3 |
# ------------------
```

Рис. 1. Выборка данных для обучения через SPARQL

Основная идея RDF очень проста, а именно: операторы представляются в виде троек формы субъект–предикат–объект, причем каждая тройка выражает отношение (представленное ресурсом предиката) между ресурсами субъекта и объекта. Формально субъект выражается URI или пустым узлом, предикат – URI, а объект – URI или литералом, таким как число или строка.

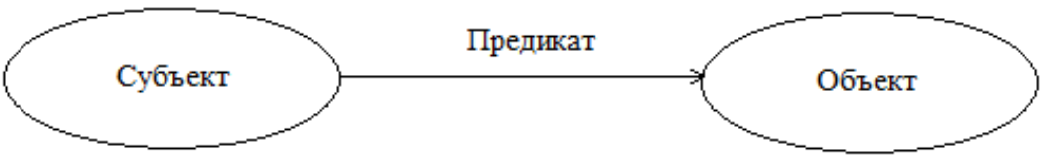

Рис. 2. Отношение в RDF

На рисунке 3 изображена схема отношений элементов в RDF.

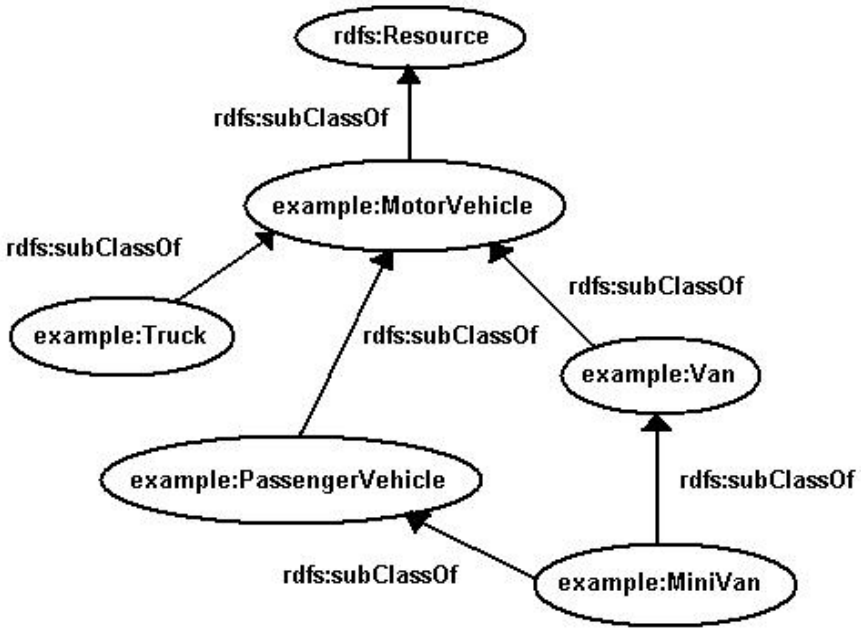

Рис. 3. Пример отношений в RDF

Структура в RDF подобна графу. Метод [7], а также метод [8] для генерации словосочетаний из RDF предлагают работать с данными RDF, как со множеством графов, где граф представлен набором взаимосвязанных троек. RDF представляется, как граф d из множества троек,

$$d=<t_1,t_2,...;t_n>, \quad (1)$$

где t_1 – первая тройка;

t_2 – вторая тройка;

t_n – последняя тройка.

2. Алгоритм предложенного метода

Мы будем рассматривать набор данных RDF как множественный граф с ресурсами, литералами и узлами на графе, как в формуле (1).

Каждая тройка t в множестве – содержит в себе субъект, предикат и объект. Граф $G = \{(s, p, o)|s \in S \wedge p \in P \wedge o \in O\}$, где s – субъект, p – предикат, o – объект, а S, P, O – множества субъектов, предикатов и объектов.

Для предобработки графа мы воспользуемся методом RDF2VEC, описанным в [9], который создает векторное представление для RDF. Для каждой тройки в полученном множестве троек мы получаем численное представление с помощью применения методологии RDF2VEC к каждому элементу и на выходе получаем матрицу X с числовыми значениями, в которую закодировано наше множество RDF.

На рисунке 4 представлен фрагмент кода на языке Python для применения RDF2VEC в процессе обработки данных.

```
def triples_to_vec (triples_list):
    X = []
    for triple in triplets_list:
        s,p,o = triple[:-1]
        es = get_RDF2vector(s)
        ep = get_RDF2vector(p)
        eo = get_RDF2vector(o)
        embd = np.concatenate((es, ep, eo)).flatten()
        X.append(embd)
    X = np.array(X)
    return X
```

Рис. 4. Фрагмент кода преобразования данных

Выводы и перспективы исследования

С использованием средств Python, Jupyter notebook и классификатора RandomForestClassifier в sklearn была произведена реализация обучения на тестовом наборе данных rdf, с предварительной обработкой через представление в множестве графов и преобразование в числовые вектора. Метрика точности accuracy_score составила 0.832, что говорит о том, что при таком подходе предсказание и обучение нейронных сетей будет верно, однако показатель не является идеальным, а также при графах с очень сильной глубиной возможно появление «шумов» при конвертации их в вектора. В дальнейшем планируется избавление от возможных «шумов» при преобразовании rdf для обучения.

Список литературы:

1. Berners-Lee T., Handler J., Lassila O. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities // Scientific American, May 2001.

2. Linked Open Data Cloud. [Electronic Source.] URL: https://lod-cloud.net/ (access date: 11.05.2020).

3. Bin S., Westphal P., Lehmann J., Ngonga A. Implementing scalable structured machine learning for big data in the SAKE project // 2017 IEEE International Conference on Big Data (Big Data 2017), December 11–14, 2017, Boston, MA, USA. Publisher: IEEE, 2018. P. 1400-1407. DOI: 10.1109/BigData.2017.8258073.

4. Venkata N., Kappara P., Ichise R., Vyas O. LiDDM: A Data Mining System for Linked Data. 2011. [Electronic Source] file:///C:/Users/Al/Downloads/LiDDM_A_Data_Mining_System_for_Linked_Data.pdf (access date: 11.05.2020).

5. Paulheim H., Fümkranz J. Unsupervised generation of data mining features from linked open data // Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics (WIMS'12). 2012. P. 1–12. DOI: 10.1145/2254129.2254168.

6. Bloem P., Vries G. K. D. Machine Learning on Linked Data, a Position Paper // Proceedings of the Linked Data for Knowledge Discovery ECML, 2014. DOI: 10.13140/2.1.2634.4963.

7. Lösch U., Bloehdorn S., Rettinger A. Graph kernels for RDF data // Simperl E., Cimiano P., Polleres A., Corcho Ó., Presutti V. (eds.). ESWC. Vol. 7295 of Lecture Notes in Computer Science., Springer, 2012. P. 134–148.

8. Sleimi A., Gardent C. Generating paraphrases from DBPedia using Deep Learning // Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web, 2016. P. 54–57. DOI: 10.18653/v1/W16-3511.

9. Ristoski P., Rosati J., Di Noia T., De Leone R., Paulheim H. RDF2Vec: RDF graph embeddings and their applications // Semantic Web. 2018. Vol. 10. P. 1–32. DOI: 10.3233/SW-180317.

УДК 519.8 : 004.65
doi:10.18720/SPBPU/2/id20-207

Моргунов Евгений Павлович[1],
канд. техн. наук, доцент,
доцент кафедры информатики и вычислительной техники;
Моргунова Ольга Николаевна [2],
канд. техн. наук, доцент,
доцент кафедры информатики и вычислительной техники;
Постойко Анастасия Юрьевна [3],
студент

РЕАЛИЗАЦИЯ МЕТОДА «АНАЛИЗ СРЕДЫ ФУНКЦИОНИРОВАНИЯ» В ВИДЕ РАСШИРЕНИЯ СУБД POSTGRESQL

[1, 2, 3] Сибирский государственный университет науки
и технологий имени академика М. Ф. Решетнева, Красноярск, Россия,
[1] emorgunov@mail.ru, [2] olgamorgunova@mail.ru,
[3] postoiko.anastasya@yandex.ru

Аннотация. Предложены усовершенствования технологии интеграции метода «Анализ Среды Функционирования» (Data Envelopment Analysis), предназначенного для оценки эффективности систем, в среду системы управления базами данных PostgreSQL, имеющей открытый исходный код. Показаны преимущества использования концепции репозитория и идеи многовариантных (мультиверсионных) вычислений.

Ключевые слова: эффективность систем, Анализ Среды Функционирования, АСФ, базы данных, репозиторий, PostgreSQL.