ISSN 2661-2666 (Online) International Scientific Journal Monte (ISJM) DOI: 10.33807/monte.20222486
Volume 6, (No).1 (2022): April

Using machine learning algorithms in Microsoft Azure ML to improve system search

Jonida Shehu[1*], Silvana Greca[2], Endri Xhina[3]

[1,2,3] University of Tirana, Faculty of Natural Sciences, Informatics Department
[*]jonida.shehu@fshn.edu.al

ABSTRACT

Machine learning algorithms have revolutionized predictive analysis, natural language processing, image classification, and information retrieval. Semantic AI is a new concept of using the power of machine learning and knowledge graphs for better search and system recommendations. This paper combines the knowledge graph information from an ontology with ML algorithms in a system where citizens can search for e-government services. The ontology describes linked entities of government institutions, legislations, public services, documentation, and their attributes. Service ranking rating and previous searches saved on the ontology provide users with better experience and search results. The experiment we model in Microsoft Azure ML uses data from the ontology and classifies public services with more accuracy and precision using machine learning algorithms by assigning weights to essential words when searching.

Keywords: ML algorithms; ontology; semantic search; e-government;

Introduction

Machine learning addresses how to build computers that improve automatically through experience. It is one of today's most rapidly growing technical fields, lying at the intersection of computer science and statistics and the core of artificial intelligence and data science (Kravets et al., 2015). Machine learning algorithms have been applied in different fields such as banking, e-government, medical, physics, and e-learning.

E-Government refers to the use of internet technology as a platform for exchanging information, providing services, and transacting with citizens, businesses, and other arms of government (Kamal, 2009). As administrative agencies serving the people, government agencies provide people with convenient and fast services. With the popularization and development of information technology, traditional inefficient government work methods can no longer adapt to the development of social information and cannot satisfy people's convenience and the need for fast administrative services;(Zhao, 2021). Every citizen can have a personalized experience with the use of ML applications.

Although there is a huge number of research in the literature related to ML applications, there is a lack of a comprehensive study focusing on the usage of this technology within governmental applications. (Charalampos et.al 2019).

In this paper, we aim to describe the use of machine learning algorithms to classify public electronic services more accurately based on citizen search. Our previous studies (Shehu & Xhina 2020) developed a web-based tool that searches our ontology OntoAL (Shehu & Xhina 2019) and translates life events into public services. This tool is dedicated to citizens and businesses and has a separate interface for employees to support them in modeling and altering public electronic services.

Literature Review

Many types of research have highlighted the improvement of systems as a combination between semantic web technologies and ML algorithms. In (Sveatshova & Zhou, 2020), the authors propose a system called SemML for ontology-enhanced ML pipeline development. It has several novel components. It relies on ontologies and ontology templates for task negotiation, data and ML feature annotation.

(Liao et al., 2021) designed an ontology for high-performance computing (HPC) to make training datasets and AI models FAIR. Their ontology provides controlled vocabularies, explicit semantics, and formal knowledge representations.

(Kim et al., 2017) investigates two approaches and finds a suitable solution that maximizes the advantages of both technologies. They suggest a novel integration idea to compensate each technology with semantic filtering. (d'Amato, 2020) presents envisioned research directions for further developing Machine Learning solutions for the Semantic Web. (Hsu & Lin, 2020) defined an integrated machine learning with semantic web framework into cloud computing and developed a linked data query platform (LDQP) to validate its feasibility.

Methods & Methodology

Many tools can process a large amount of data to classify, determine similarities or predict new data from current data. Some of the most popular tools are Weka (Frank et al. 2016), TensorFlow (Abadi et al. 2015), Google Cloud AutoML (Google 2022), and Microsoft Azure ML (Microsoft 2020). Some text classification libraries: Fast text (Facebook), ml5.js at collaboration with TensorFlow.

We have selected Microsoft Azzure ML to implement our experiment on machine learning algorithms based on data obtained from the ontology. Microsoft Azure ML is a cloud-based environment that can train, automate and create machine learning models (Microsoft 2022). The patterns that are created support programming languages like Python and R; however, this tool also has a simple interface for designing experiments without programming.

Through drag & drop functionalities, many modules can be included in the experiment. Regardless, we used some modules listed below:

- ✓ Ready-saved or locally uploaded data sets.
- ✓ Converters of data formats to CSV, TSV, dataset, ARFF, SVMLight.
- ✓ Data transformers, such as filters, intermediate data modifiers, data connectors, and data split used during the training process. For data manipulation, we can also mention grouping the values of several categories into one category and the elements of column selectors in the data set that allow the definition quickly of the columns, which we wish to include in the training.
- ✓ Filter based on feature selection, where columns specified have greater predictive power in the data set.
- ✓ Text analyzer for word processing.
- ✓ Machine learning module that would determine the initialization of the model by using neural networks for classification, training modules and model evaluation.

Data Analysis

Once the user provides feedback on the service, the user's search data are stored in the ontology at the distinguishing public service. These searches and the public service to which they belong will be exported from Protégé in CSV and uploaded to Microsoft Azure ML studio. To access the data, we drag & drop from the My Datasets submenu, the ontoAL.csv dataset, as shown in Figure 1.

In the following, on the data set, we will select only two columns that are "Search" and "Public Service". The data will be cleaned, split, and over them; we apply the "MultiClass Neural Network" machine knowledge module, which creates a multi-class classification model using a neural network algorithm. The experiment designer performs the whole procedure, even though the corresponding parameters are configured for each module used. For the Multiclass Neural Network module, we could specify some attributes:

- ✓ Create trainer mode->This option determines how the model will be trained. By selecting the value "Single Parameter," i.e., a single parameter, we are convinced of how the model will be configured. If we select the "Parameter Range" value, this would determine the uncertainty as to the best parameters. In this case, in the

Azure ML documentation (Microsoft 2022), for this value, it is suggested to specify a range of values and use the Tune Model Hyperparameters module to roam over all combinations and select the optimal configuration.

- ✓ Hidden layer specification->This option would specify whether to create the model relied on the basic neural network architecture by selecting the value "Fully-connected case," or we could use a personalized script in Net# language for the definition of hidden layers and their connections. For our experiment, we will select the "Fully-connected case" option.
- ✓ The number of hidden nodes-> This option allows us to customize the number of hidden nodes. The default value for a hidden layer would be 100 nodes.
- ✓ Learning rate-> This option determines the step size taken in each iteration before correction. The values for this option are from [0.1-1]. The definition of this feature is paramount because a small value can take a long time to train the data. Meanwhile, a considerable value can make the model train quickly but generates an unsustainable model (Microsoft 2022).
- ✓ Shuffle examples->This option is used to change the order of instances within the learning iterations.

Results

The preliminary results of the experiment are promising. With 1500 record searches, it is possible to predict the public service with an average accuracy of 98%; the average accuracy is 96% and the average sensitivity (recall) 95%.

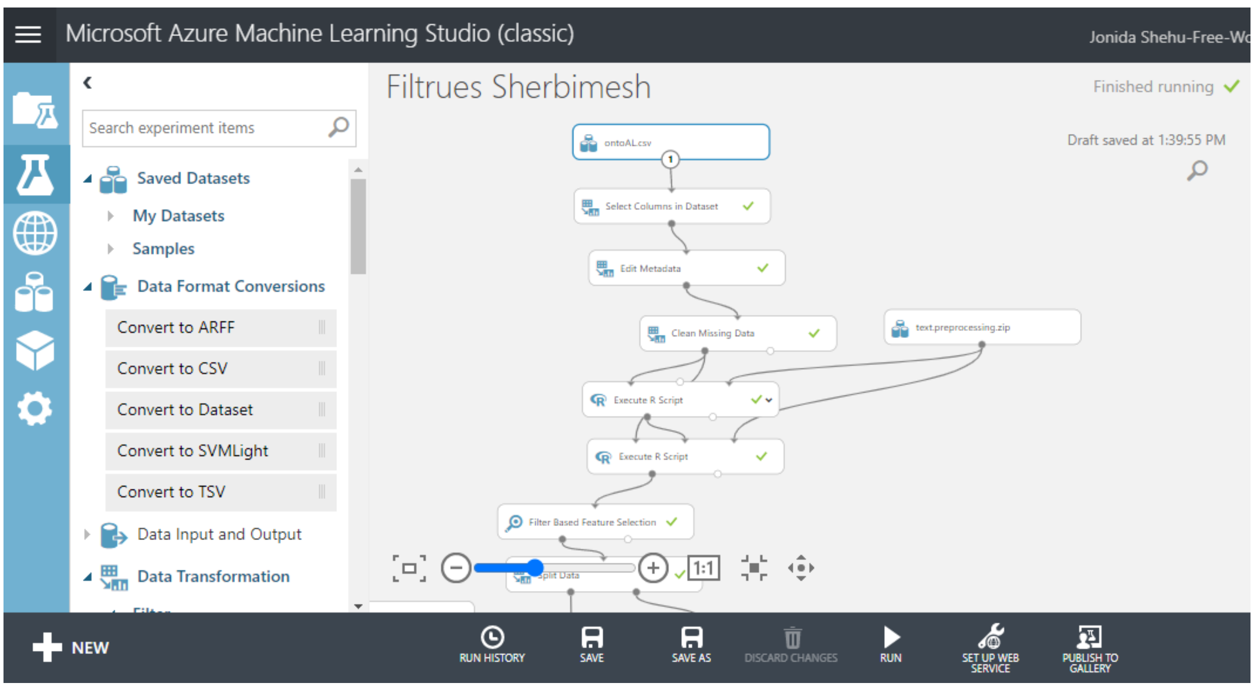

Figure 1. Experiment on Microsoft Azure Machine Learning Studio

Based on these outcomes, we can conclude that machine learning algorithms will significantly improve citizens' search on the ontoAL tool. The solution can be easily integrated because the Microsoft Azure ML Studio platform offers the possibility of accessing experiments as web services. Once created as an experiment service, it can use a programming language that supports an HTTP request and response. The official documentation of Microsoft provides guidelines and examples for implementation in C #, Python, R, and JavaScript.

For the JavaScript language, it is enough to use the object in JSON format to format input data and get after the results. To connect to the service, we need to install the request package. We can then use JavaScript code from the official Microsoft documentation.

Conclusions

Systems based on artificial intelligence increasingly require data closer to how people think and reason, which has led to the interaction with web semantics technologies. This paper emphasizes the importance of ML algorithms for improving search as a system that does not learn is incomplete. The experiment we created at ML Azure Learning Studio is based on the searches in the ontology of public services based on the legislature, naming, description, or even service category. Cleaning modules, data sharing, and neural networks have been applied to the data. The final results increase the accuracy by 10% compared to the previous results. We will automate the experiment by integrating it as a web service in the future.

References

Kravets, P., Burov, Y., Lytvyn, V., Vysotska, V., Ryshkovets, Y., Brodyak, O., Vyshemyrska, S. (2015). “Markovian Learning Methods in Decision-Making Systems”, Lecture Notes in Computational Intelligence and Decision Making, (423-437).

Alexopoulos, Ch., & Lachana, Z., Androutsopoulou, A., Diamantopoulou, V., Charalabidis, Y., Loutsaris, M. A. (2019). How Machine Learning is Changing e-Government. 354-363. 10.1145/3326365.3326412.

Zhao, W. (2021) "Construction and Application of E-Government Simulation Network Experimental Teaching Practice Environment under the Background of Machine Learning", Mobile Information Systems, vol. 2021, Article ID 1952428, 9 pages, 2021.

Kamal, M. M. (2009). “Mapping Factors Influencing EAI Adoption on the Adoption Lifecycle Phases in LGAs”, Handbook of Research on ICT- Enabled Transformational Government: A Global Perspective, Brunel University, UK. pp. 140-162, 2009.

Shehu, J. & Xhina, E (2019). “Modeling an ontology for public E-Government Services in Albania”, Proceedings of International Academic Conference on Research in Engineering and Technology, Barcelona, Spain, 5 - 7 September 2019.

Shehu J. & Xhina E. (2020). “A semantic search tool for E-government public services in Albania”, International Journal of Computer Science and Information Security (IJCSIS), Vol 3, ISSN : 1947-5500

Frank, E., Hall, M. A. & Witten, I. H. (2016): The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Abadi, M. Agarwal, A., Barham, P., Brevdo, E., Chen, Zh., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G. & Isard, M., TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. 2015.

Google: Vision & AI Machine Learning Products (2022), https://cloud.google.com/vision, aksesuar më 15 Janar 2022.

Microsoft: Azzure Machine Learning Studio Features https://azure.microsoft.com/en-us/services/machine-learning/#features, aksesuar më 15 Janar 2022

Microsoft: What is Azure Machine Learning? (2022) https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-azure-ml, aksesuar më 15 Janar 2022.

Microsoft: Multiclass Neural Network. (2020) https://docs.microsoft.com/enus/azure/machine-learning/studio-module-reference/multiclass-neural-network

Svetashova, Y. *et al.* (2020). Ontology-Enhanced Machine Learning: A Bosch Use Case of Welding Quality Monitoring. In: The Semantic Web – ISWC 2020. ISWC 2020. Lecture Notes in Computer Science(), vol 12507. Springer, Cham. https://doi.org/10.1007/978-3-030-62466-8_33

C. Liao et al., (2021). "HPC Ontology: Towards a Unified Ontology for Managing Training Datasets and AI Models for High-Performance Computing," 2021 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC), 2021, pp. 69-80, doi: 10.1109/MLHPC54614.2021.00012.

Kim, M., Kang, H., Kwon, S., Lee, Y., Kim, K. and Pyo, C., S. "Augmented ontology by handshaking with machine learning," 2017 19th International Conference on Advanced Communication Technology (ICACT), 2017, pp. 740-743, doi: 10.23919/ICACT.2017.7890191.

d'Amato, C. (2020). 'Machine Learning for the Semantic Web: Lessons Learnt and Next Research Directions'. 1 Jan: 195 – 203.

Hsu, I-Ch., & Lin. Y-H. (2020) Integrated machine learning with semantic web for open government data recommendation based on cloud computing, Journal of Software: Practices and Experience. Volume50, Issue12, Pages 2293-2312.