

MULTI-STEP METHODS FOR MACHINE LEARNING MODELS WITH WEB METRICS

Ivan Popchev, Daniela Orozova^{*,#}

*Received on August 2, 2023
Accepted on September 26, 2023*

Abstract

The purpose of the article is to outline methods as a multi-step process for developing an information flow with machine learning models that can be used for research, analysis, and training. Web metrics analysis of available web content can provide useful information about user behaviour. The results show that the formulated solutions can be successfully used for different tasks and can be adapted to new technologies and applications.

Key words: web metrics, machine learning, multistep methods

Introduction. Clearly machine learning is changing the way we interact with apps today. The machine learning algorithms presented in [1] here have been developed and applied in the field of agriculture. The ideas and preliminary results presented here are related to the work of the national research programme “Smart crop production”. The analysis of the data accumulated during the work can serve to track the movement of raw materials and products, reduce losses, increase trust, timely payments, etc.

[#]Corresponding author.

This work was supported by the Bulgarian Ministry of Education and Science under the National Research Programme “Smart crop production” approved by Decision of the Ministry Council No. 866/26.11.2020 and in the frames of Bulgarian National Recovery and Resilience Plan, Component “Innovative Bulgaria”, Project No. BG-RRP-2.004-0006-C02 “Development of research and innovation at Trakia University in service of health and sustainable well-being”.

DOI:10.7546/CRABS.2023.11.08

MITCHELL [2] provided the widely quoted and well-known definition of Machine Learning: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.”. Machine learning algorithms study data using various computational methods. Two main types of techniques are applied: *supervised* and *unsupervised learning*. The main task in the first type is to create a model from labelled data, which allows predictions to be made about future data. Algorithms of the second type take a data set containing only input values and find a structure or distribution in the data, with no indication of a known variable, the data is unlabelled, there are no training examples.

Another type is *reinforcement learning*: a computer program interacts with a dynamic environment in which it performs a certain goal. The program provides feedback in terms of rewards and punishments as it navigates its problem space.

The purpose of the article is to outline methods as a multi-step process for developing an information flow, using machine learning models of the accessible and popular Orange Data Mining system [3]. The methods can be used for research, analysis and training in the space of the agriculture as well as in the field of emerging technologies.

Web metrics. For a better understanding of the complex interactions of users in the web space, web analysis is applied. Web analytics is the collection, reporting, measurement and analysis of Internet data about website users. The web analytics process includes various web metrics defined in Google Analytics [4] as:

- page views – the number of web page hits by a human visitor;
- visitors – the number of different visitors to a given website;
- pages/visits – the number of pages a user visited during a given visit;
- time on site – duration (in seconds) of total time spent by all visitors on a website;
- stickiness – average time a visitor stays on a web page;
- frequency – average number of visits per visitor to a given site (loyalty indicator);
- recency – period of time (in number of days) since the visitor’s last visit to the site;
- length of visit – period of time (in seconds) that visitor spent on the website;
- depth of visit – the number of pages’ visitors have seen per visit, etc.

There are mainly two techniques used to analyze website traffic: server-side and client-side data collection. Information that can be obtained about customer behaviour includes, for example, number of pages viewed, language setting, browser screen resolution settings, etc. A number of companies provide web analytics software products such as Google, WebTrends, Nedstat and others. Google Analytics is the most widely distributed open source program [4]. To provide

a comprehensive view of online customers, data from various sources must be integrated and analyzed [5].

The fuzzy classification of the web indicator [6] is applied for the analysis of the attendance of the web sites accessed by users during the process. To estimate website traffic, a fuzzy system of rules is developed. The inductive fuzzy classification (IFC) method was applied [7]. The grouping of elements in a fuzzy set is done based on a defined membership function obtained by data induction. Percentile rank (PR) inductive fuzzy classification develops a fuzzy membership function using linguistic terms as: “low”, “medium”, “high” corresponding to the number of user visits:

- determines the belonging of the element x to the fuzzy class “high”:

$$\mu_{\text{high}}(x) := P(M < x)$$

- determines the element x to the fuzzy class “low”:

$$\mu_{\text{low}}(x) := 1 - \mu_{\text{high}}(x)$$

- determines the element x for “medium”:

$$\mu_{\text{medium}}(x) := 1 - \text{abs}(\mu_{\text{high}}(x) - 0.5) - \text{abs}(\mu_{\text{low}}(x) - 0.5).$$

An advantage of using methods based on fuzzy logic is the results are convenient for decision making. On the basis of data collected or generated as a result of interaction with web resources, patterns in user behaviour are sought.

Supervised learning – a multi-step method for machine learning models. The main techniques are: classification (class labels are discrete values) and regression (for continuous values) [8]. Tools for building classification and regression models are: Logistic Regression, Naïve Bayes Classifier, Support Vector Machines (SVM), Decision Trees; Artificial Neural Networks, k-nearest Neighbours, Linear Regression, Random Forests, Deep learning, etc. The main steps in creating the models, their evaluation and application are presented.

A multi-step method for building machine learning models:

Step 1. Data collection. Load the data using the “File” widget. They can be entered from Excel (.xlsx), a tab-delimited text file (.txt), a comma-separated data file (.csv), or a URL. The data can be previewed by selected columns or only a sample of them can be displayed. For example, with a “Scatter Plot” or “Distribution” widget.

Step 2. Selecting a model. Depending on the task, a specific regression or classification tool is chosen. With “File” dialogue box selects the target variable.

Step 3. Clear data. Problems in the performance of machine learning models can occur when some data is missing. This can negatively affect the accuracy of the results obtained. The radical approach is to delete the rows or columns that

have missing values. But if there is a high correlation between the column with missing values and the other columns, it is good to look for ways to replace the missing data. Using the “Impute” widget, different replacement methods can be used with: Distinct Value, Random Values, Model-Based Values.

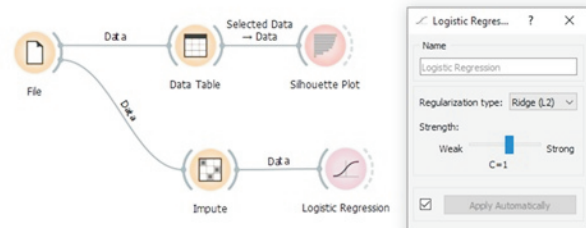


Fig. 1. Workflow with “logistic regression” widget

Step 4. Training the model. Selecting a widget to create and train the “Logistic Regression” model (Fig. 1). A different type of regularization can be set: L1 regularization (equivalent to the absolute value of the magnitude of the coefficients) or L2 regularization (equivalent to the square of the magnitude of the coefficients).

Step 5. Creating and training alternative models. Followed by application of alternative tools, on the data, as: SVM (Support Vector Machines), Neural Networks, Naïve Bayes, Decision Trees, etc.

Step 6. Evaluation of the performance of the models. Each of the models created is linked to the Test and Evaluation tool (Fig. 2). Both models and data must be linked to the test tool. After the models are evaluated using the “Test and score” tool, can be seen if their accuracy can be improved by tuning the parameters in the model.

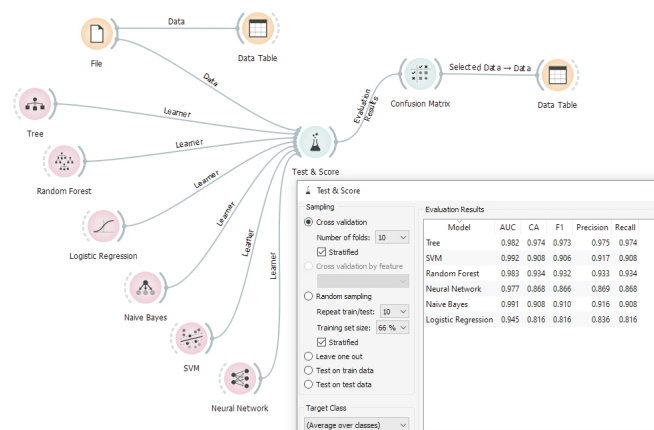


Fig. 2. “Test and score” workflow and results

Step 7. Visualization of the accuracy of the models. To visualize the results of evaluations, the “Test and score” widget can be linked with tools such as: “Confusion Matrix”, “ROC Analysis”, or others.

Step 8. Predicting new data. At this step, the model is ready for practical application. The model gains independence and makes its own conclusions.

An experiment. The study is related to the application of artificial intelligence to support Bulgarian crop production [9]. Following the eight steps of the algorithm, an approach for automated document type determination in the web space has been implemented [10]. Frequency of use is reported by analysing web metrics for the document. The determination of the significant words and the type of its content is done on the basis of subject-oriented ontologies in the field of smart crop production.

To determine the significant words for the document, the frequency of occurrence of the words is searched. In the experiment, we apply the software Word Cloud (wordclouds.com). A degree of proximity is determined for each word in the selected set with all words in the ontology dictionary. The degree of closeness can be determined by applying different algorithms, we use q-gram distance. In the experiment, real articles are determined for the subject area: smart crop production.

The National Genebank is a collection of plant genetic resources. It is linked to the European Electronic Catalogue of Plant Genetic Resources EURISCO (<http://eurisco.ipk-gatersleben.de>). The genebank is part of the national programme of the “smart crop production”. Ninety-four actual documents from the field and 82 documents from outside the subject area were used. About 160 words are extracted on average for each document, and for each word a similarity to the concepts of the subject ontology dictionary is calculated. One part of this data is used to train and create models, such as a target attribute – whether the resource is related to the domain or not. The remaining data is applied to test the models. A workflow is created by first loading the data with the keywords from the documents and their degree of proximity to the domain terms. Models are created applying the tools of the system (Fig. 3).

Similarly, the tools can be used for other analyses of collected data [11,12] such as: the length of time the visitor spends on the website; time elapsed since the visitor’s last visit to the site; length of visits; the depth of visiting the site, etc.

Unsupervised learning – methods for clustering. The main tasks solved are related to data clustering, dimensionality reduction, density estimation, etc. [13]. Cluster analysis is related to the distribution of a set of data into groups so that the data in one group are similar according to one or more criteria, and the data from different clusters are different [14]. Clustering techniques apply different assumptions about the structure of the data. HAN et al. [15] propose a categorization of methods into three categories: density-based methods, pattern-

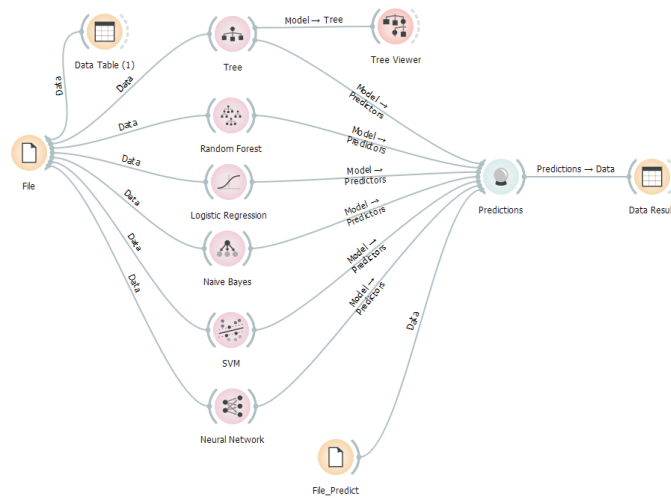


Fig. 3. Creating and evaluating the models

based methods, and network methods. Another categorization is based on the principle of induction is presented in ESTIVILL-CASTRO [16] In [17], the division of methods into two main groups is proposed: hierarchical methods and partitioning methods.

Algorithms based on partitions. These algorithms are defined as non-hierarchical. The number of clusters is determined in advance. Algorithms generate different partitions and then evaluate them according to certain criteria [17]. The k-means algorithm is one of the most popular algorithms of this type. It belongs to Exclusive Clustering group; each instance is placed in exactly one of the k clusters.

Method for building the workflow for clustering:

Step 1. Load Data. The data for analysis is selected via the “File” widget. An “Impute” tool can also be applied to clean data.

Step 2. Selecting the k-means or Louvain Clustering widget. This Orange system tools implements clustering algorithms (Fig. 4).

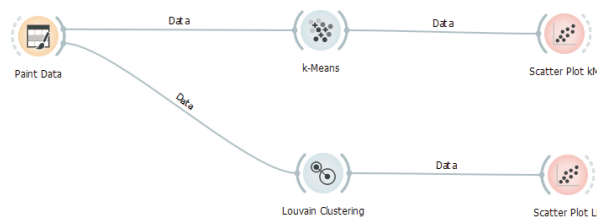


Fig. 4. Workflow with Clustering widgets and visualization of result

Step 3. Visualize the result. Clustering tools performs clustering on the data. A visualization widget such as a Scatter Plot should be connected to the workflow.

Hierarchical clustering. In Hierarchical Cluster algorithms, the number of clusters is not predetermined. The hierarchical clustering tool performs clustering of arbitrary object types based on the distance matrix between them and displays a corresponding dendrogram. It starts by assigning each element to a separate cluster. It then finds the most similar pair of clusters and merges them [18]. Next is a calculation of the similarities between the new cluster and each of the old clusters.

Method for constructing an information flow in hierarchical clustering:

Step 1. Data collection. The desktop and toolset must be loaded and the data for analysis has to be selected. An “Impute” tool can also be applied to clean the data.

Step 2. Selecting the Distance widget. Different metrics can be used to calculate distance when grouping the data: Euclidean, Manhattan, Minkowski metric.

Step 3. Selecting the Hierarchical Clustering widget. This widget implements an algorithm for hierarchical grouping. It performs the calculation of distances between objects. Shows the result of clustering by corresponding dendrogram.

Step 4. Preview the result. It has a large set of tools for visualization of the resulting set. For example, Data Table and Scatter Plot can be included in workflow.

An experiment. In the conducted experiment, we introduce 21 images in jpeg format of “crop production”. To apply hierarchical clustering, applying the Image Analytics tool [3], a source images file is loaded and visualized via the *Image viewer*. The experiment set a distance metric: cosine, for evaluating the proximity of the images with “Distances” widget. Applying the Hierarchical clustering widget, clusters are created from the input images and the result is a dendrogram, where a sunflower, rapeseed, linseed, corn appear in one cluster.

Scientific research related to the application of artificial intelligence, within the framework of the national programme for intelligent agriculture, aims to support activities of farmers and contribute to improvement of living conditions in rural areas.

Conclusion. As a result of the conducted research, multi-step methods have been formulated for the construction of: machine learning models for classification and regression; creating an information flow for clustering and hierarchical grouping. These methods represent practical approaches that can be used for research, analysis, and learning in various fields, using accessible and popular software platforms.

The authors plan to apply web metrics analytics to available web content to uncover patterns in user behaviour in the field of agricultural supply chain management. A number of risks can be generated due to inaccuracies in reporting

and handling web metrics. The authors' efforts will be directed to risk analysis [19] in the aspects of the evolution from risk management to enterprise global risk management.

REFERENCES

- [1] POPCHEV I., D. OROZOVA (2023) Algorithms for Machine Learning with Orange System, *Int. Journal of Online and Biomedical Engineering*, **19**(04), 109–123.
- [2] MITCHELL T. M. (1997) *Machine Learning*, McGraw-Hill Science/Engineering/Math, ISBN: 0070428077.
- [3] Orange Data Mining system, Retrieved from: <https://orange.biolab.si/training/introduction-to-data-mining/> [accessed: 19 July 2023].
- [4] CLIFTON B. (2012) *Advanced Web Metrics with Google Analytics*, 3rd Edition, New York, Wiley, USA.
- [5] OROZOVA D. (2019) Appropriate E-Test System Selection Model, *C. R. Acad. Bulg. Sci.*, **72**(6), 811–820.
- [6] MOBASHER B. (2007) Web Usage Mining. In: *Web Data Mining, Data-Centric Systems and Applications*, Springer Berlin Heidelberg, 449–483, DOI: 978-3-540-37882-2_12.
- [7] KAUFMANN M., A. MEIER (2009) An Inductive Fuzzy Classification Approach applied to Individual Marketing. In: *Proceedings of the 28th North American Fuzzy Information Processing Society Annual Conference (NAFIPS2009)*, Ohio, USA.
- [8] POPCHEV I., D. OROZOVA, S. STOYANOV (2019) IoT and Big Data Analytics in E-Learning, *Big Data, Knowledge and Control Systems Engineering (BdKCSE)*, 1–5.
- [9] POPCHEV I., L. DOUKOVSKA, I. RADEVA (2022) A framework of blockchain/IPFS-based platform for smart crop production, *International Conference Automatics and Informatics (ICAI)*, 265–270.
- [10] POPCHEV I., D. OROZOVA (2020) Text Mining in the Domain of Plant Genetic Resources, *10th International Conference on Intelligent Systems*, 596–600.
- [11] POPCHEV I., D. OROZOVA (2019) Towards Big Data Analytics in the E-learning Space, *Cybernetics and Information Technologies*, **19**(3), 16–25.
- [12] POPCHEV I., I. RADEVA, L. DOUKOVSKA (2023) Oracles Integration in Blockchain-Based Platform for Smart Crop Production Data Exchange, *Electronics*, **12**(2244), 1–20.
- [13] TRIPATHY B. K., A. SUNDARESWARAN, S. GHELA (2021) *Unsupervised Learning Approaches for Dimensionality Reduction and Data Visualization*, CRC Press.
- [14] KUBAT M. (2021) *An Introduction to Machine Learning*, Springer.
- [15] HAN J., M. KAMBER, J. PEI (2001) *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers.
- [16] ESTIVILL CASTRO V., J. A. YANG (2000) Fast and robust general purpose clustering algorithm, *Pacific Rim Int. Conference on Artificial Intelligence*, 208–218.
- [17] BAKSHI E. (2021) *Unsupervised Learning – Identify Machine Learning Tasks: Machine Learning Course*, Independently Published.
- [18] OROZOVA D., K. ATANASSOV (2019) Model of Big Data Map/Reduce Processing, *C. R. Acad. Bulg. Sci.*, **72**(11), 1537–1545.

- [¹⁹] POPCHEV I., I. RADEVA, I. NIKOLOVA (2021) Aspects of the evolution from risk management to enterprise global risk management, Engineering sciences, LVII, No.1, 16–30.

*Institute of Information
and Communication Technologies
Bulgarian Academy of Sciences
Akad. G. Bonchev St, Bl. 2
1113 Sofia, Bulgaria
e-mail: ivan.popchev@iict.bas.bg*

**Trakia University
Faculty of Economics
6015 Stara Zagora, Bulgaria
e-mail: daniela.orozova@trakia-uni.bg*