Algorithms for Machine Learning with Orange System

https://doi.org/10.3991/ijoe.v19i04.36897

Ivan Popchev[1], Daniela Orozova[2](✉)
[1]Bulgarian Academy of Sciences, Sofia, Bulgaria
[2]Trakia University, Stara Zagora, Bulgaria
daniela.orozova@trakia-uni.bg

Abstract—Emphasized is the need for new approaches and solutions for forming of increased information awareness, knowledge and competencies in the present and future generations to use the possibilities of emerging technologies for technological breakthroughs. The article presents basic machine learning tools of both types: supervised learning, which trains a model on known input and output data and predicts future results, and unsupervised learning, which finds hidden patterns or inherent structures in the input data. Algorithms for the processes of creating an information flow when applying the tools of the Orange system, which can be used for research, analysis and training, are formulated. Experiments related to smart crop production and analyses with different classification, regression and clustering algorithms. The results show that the formulated solutions can be successfully used for different tasks and can be adapted to new technologies and applications.

Keywords—emerging technologies, supervised and unsupervised learning, smart crop production, Orange system

1 Introduction

Today we witness to digitalization of all assets and economic agents in a uniform chain of the value and the integration in a general digital ecosystem. Unlimited possibilities are created for mutual up building, increasing and merging of technologies and societies, as well as for new technological breakthroughs in many areas. There is a need for new approaches and solutions for forming of increased information awareness, knowledge and competencies in the present and future generations to use the possibilities for technological breakthroughs. This raises new problems and therefor a necessity for the search of new solutions. The possibilities for carrying out and management of various processes become huge especially in a connected environment. This dynamical and heterogeneous environment must be able to adapt to the new characteristics and requirements which emerging technologies imposes.

It is important to mention that each of the emerging technologies is in a logically, scientifically proven and practically justified dependence on many different scientific areas. The complexity and mutual involvement of these emerging technologies is accompanied by a significant increase of the risk factors due to the all-embracing and in some aspects spontaneous digitalization which is a reason for anxiety to the formed disruption in the relation "human-digital environment" [3].

Potential risks in emerging technologies can be systematized in the following separate eight groups: Rd – privacy and data security; RL – change in labor market; Rp – mental distraction; RM – manipulation and echo camera; RF – fragmentation; RA – responsibility and accountability; RE – ecology, ecosystems and ethics; Rs – change in income/cost structure and ownership of assets.

Each risk has negative, often unknown, undefined in advanced impacts. This requires investigation and decision-making about the risk which can be in the following scheme of five phases:

phase 1: Identification of the risk;
phase 2: Quantitative/Qualitative evalutation of the risk and its characteristics;
phase 3: Choice of instrument and/or instruments for risk impact (standards, norms, rules, models, methods, algorithms);
phase 4: Risk management – direct impact on the environment or the object through the selected instruments;
phase 5: Monitoring, control and evaluation of the risk management which can be a sufficient reason for going back to previous phases.

The artificial intelligence and machine learning in particular is at the heart of the emerging technologies, because the connected to them scientific breakthroughs form directions whose functioning depends to greatest degree on the knowledge representation and imitation of the capabilities of the human's reasoning.

Various data mining software tools are available today. Among the most popular are: RapidMiner, RapidAnalytics, WEKA, PSPP, KNIME, Orange, Apache Mahout, jHepWork, Rattle, GhostMiner, XENO, SAS Enterprise Miner, Polyanalyst and IBM SPSS modeler. Each software has methods for analyzing and interpreting information from data sets. Orange software [1] is used in the conducted experiments. Orange is a software product for machine learning and data analysis through Python scripting and visual programming.

Orange Library is a hierarchically organized set of components. The main branches of the hierarchy of components are: data management and preprocessing for data input and output, classification, regression, association for association rules and frequent item sets mining, clustering, which includes k-means and hierarchical clustering approaches, evaluation with cross-validation and other sampling-based procedures, projections with implementations of principal component analysis, multi-dimensional scaling and self-organizing maps. The library is designed to simplify data analysis workflows as a combination of existing components.

Machine learning algorithms use computational methods to "learn" from the data. Two main types of techniques are applied: supervised learning, which trains a model on known input and output data so that it predicts future results, and unsupervised learning, which finds hidden patterns or inherent structures in the input data.

Machine learning tools are an essential part of the artificial intelligence of modern information systems. Through their skillful use, data analysis and work efficiency are accelerated. They are valuable to companies because they can increase efficiency, flag potential problems, find new revenue streams, identify areas of future growth.

The purpose of the article is to outline the basic algorithms in the form of a multi-step process for developing an information flow with Orange system's machine learning tools, that can be used for research, analysis, and training in the space of the emerging technologies.

2 Supervised learning with Orange system

The main task is to create a model from labeled data, which allows predictions to be made about future data. The main techniques are: classification when the class labels are discrete and regression when the score is a continuous value [6]. This type is called teacher-directed learning because the system is presented with sample inputs and corresponding outputs, and the goal is to learn a general rule that maps inputs to outputs.

Building a classification model is one of the most popular tasks in the field of machine learning when the class labels are discrete values. The goal is to assign objects to predefined classes. When the number of classes is 2, we have a binary classification, if it is 3 or more, the task is multiclass. In practice, classification models are useful in solving various business tasks. After creating the model, it is important to determine whether it performs optimally when new data is introduced. This requires multiple tests and a review of the various metrics that determine how well the model performs.

2.1 Orange system tools for classification and regression model building

The following is a review of the basic means of the Orange system for building a model for classification and regression by applying different algorithms: Logistic Regression, Naïve Bayes Classifier, Support Vector Machines (SVM), Decision Trees; Artificial Neural Networks [2]. This group also includes methods: k nearest Neighbors, Linear Regression, Random Forests, Deep learning, etc.

- *Logistic Regression*. It measures the relationship between a dependent variable and one or more independent variables by calculating probabilities using a logistic function. Predicts the probability of events occurring by fitting data to a function. In general, regressions can be used in applications such as: forecasting the revenue of a particular product; measure the success rate of marketing campaigns; calculation of credit scores, etc.
- *Naïve Bayes Classifier*. This method is based on the famous Bayes theorem for probabilities. It is based on the assumption that the effect of the attribute value of a given class is independent of the values of other attributes, an assumption called conditional independence of the class. The algorithm is used in applications from various fields.
- *Support Vector Machines*. A binary classifier algorithm in which the training process analyzes the input data and identifies patterns in the multidimensional space by categories. Given a set of points of two types in an N-dimensional space, this algorithm

generates an N-1 dimensional hyperplane to divide these points into two groups. For example, if we have several points of two types on a piece of paper that are linearly separated, the algorithm will find the straight line that separates the points of the two types and is as far away from them as possible. SVM algorithms are classified into two categories: linear (data is separated by a hyper-plane) and non-linear.

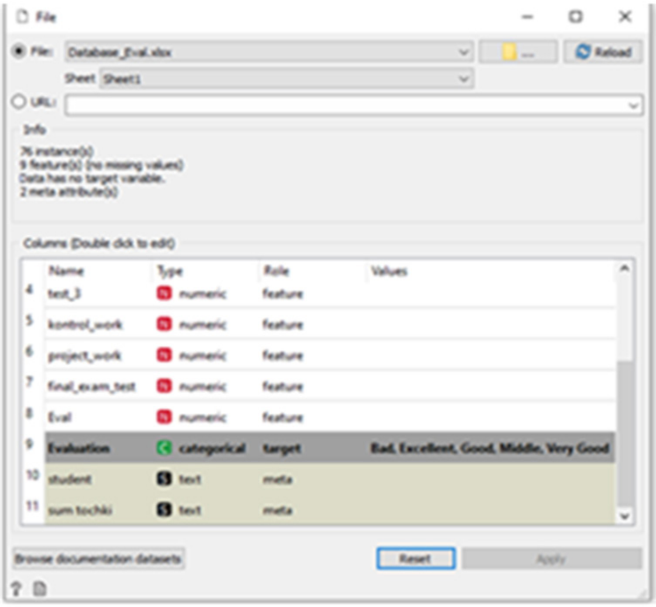
- *Decision Trees*. This type of algorithm represents the minimum number of questions (of yes/no type) that must be asked to calculate the probability of making a correct decision. The method allows to approach the problem in a structured and systematized way and to reach a logical conclusion.
- *Artificial Neural Networks*. These algorithms are inspired by the structure and function of bio-neural networks. They are most commonly used for regression and classification problems, but they are a vast field consisting of hundreds of algorithms and can be used to solve many different types of problems.

A number of tools for building classification and regression models have been built into the Orange system. The following are experiments illustrating the process of creating an information flow using the tools: Logistic Regression, Naïve Bayes Classifier, Support Vector Machines (SVM), Decision Trees and Artificial Neural Networks, etc. The following is a presentation of the basic steps that are taken to train a model using the Orange system.

2.2 A multi-step algorithm for building machine learning models using Orange's tools for classification and regression

Step 1. Data collection. At the beginning, a workflow is created, the desktop and the set of tools (widgets) are loaded. Data for analysis can be loaded via the "File" widget. They can be entered from Excel (.xlsx), a tabbed text file (.txt), a comma-separated data file (.csv), or a URL. For better understanding, the data can be visualized by some columns or extracts from them. For example, one can link the data file to the Scatter Plot tool and select the columns whose values will be plotted on the X and Y axes, colors used, shapes, sizes and other parameters. Another popular data visualization tool is "Distribution", which can show distribution in the dataset by a given attribute.

Step 2. Selecting a model and setting a target variable. Depending on the purpose and type of data, we choose a specific regression or classification tool and set the target variable. To do this, double-click on the "File" tool on the desktop, select the target variable and click Apply. An example of Figure 1.



● File: Database_Eval.xlsx

Sheet Sheet1

○ URL:

Info

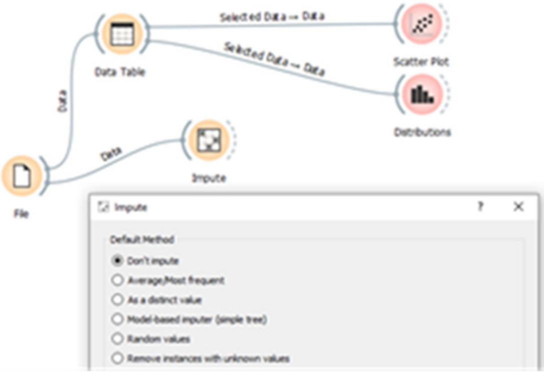
76 instance(s)
9 feature(s) (no missing values)
Data has no target variable.
2 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
4	test_3	N numeric	feature	
5	kontrol_work	N numeric	feature	
6	project_work	N numeric	feature	
7	final_exam_test	N numeric	feature	
8	Eval	N numeric	feature	
9	Evaluation	C categorical	target	Bad, Excellent, Good, Middle, Very Good
10	student	S text	meta	
11	sum tochki	S text	meta	

Fig. 1. Set a target variable using the "*File*" widget

Step 3. Clear data. When part of the data is missing, it causes problems in the performance of machine learning models and negatively affects the accuracy of the results obtained. In practice, various approaches are used by which the data are processed in a way that allows overcoming these problems. 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 rest of the features, it is good to leave it and look for ways to replace them. Using the Impute tool (Figure 2), different imputation methods can be represented. The default is the "Remove the rows with missing values" option. Other possible options are: Distinct Value, Random Values, Model-Based.

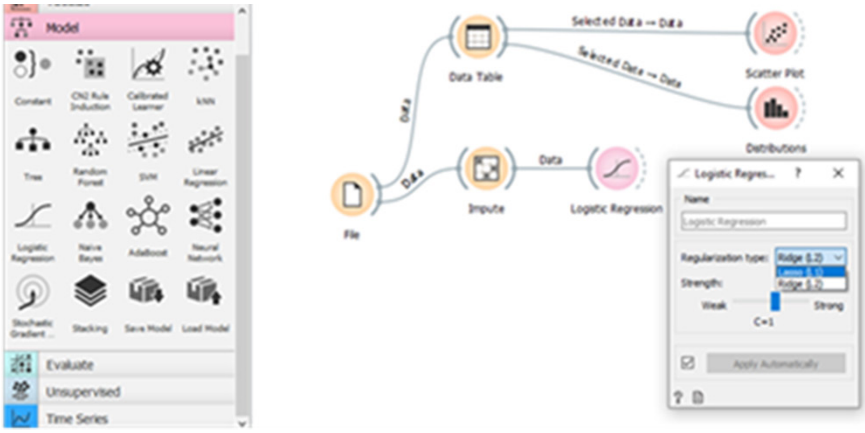


Default Method

● Don't impute
○ Average/Most frequent
○ As a distinct value
○ Model-based imputer (simple tree)
○ Random values
○ Remove instances with unknown values

Fig. 2. Application of "*file*" and "*impute*" widgets

Step 4. Training the model. Selection of a widget for creating and training the "Logistic Regression" model (Figure 3). Double click the widget and select the type of regularization you want to perform: performs L1 regularization (adds penalty equivalent to absolute value of the magnitude of coefficients) or performs L2 regularization (adds penalty equivalent to square of the magnitude of coefficients).



Name: Logistic Regression

Regularization type: Ridge (L2)

Lasso (L1)

Ridge (L2)

Strength:

C=1

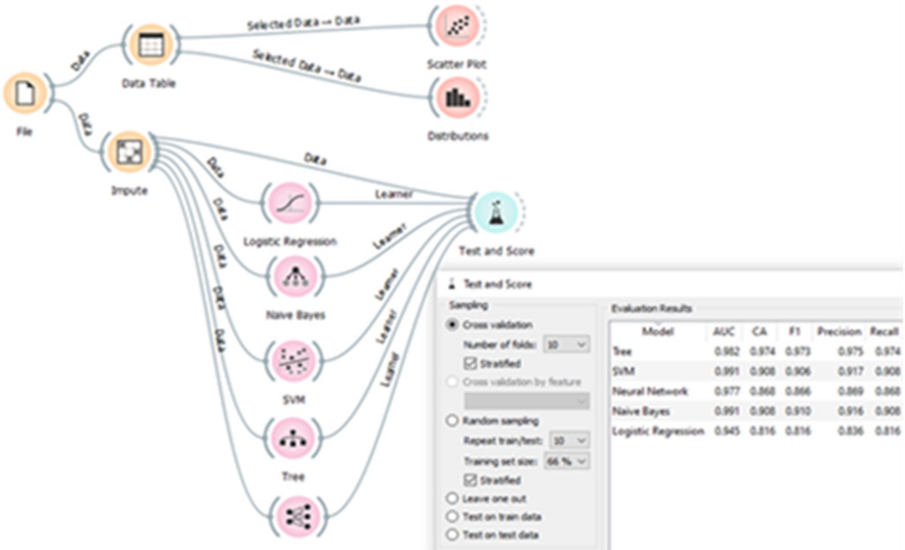
☑

Fig. 3. Creation and training of "*logistic regression*" model

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

Step 6. Evaluation of the performance of the models on the data. In the workflow, each of the created models is associated with a "Test and Score" widget.

Once the models are evaluated, it should be seen if their accuracy can be improved by tuning the parameters present in the model. Make sure you connect both the data and the model to the testing widget. The result of the operation of the tool "Test and Score" of the Orange system is a table with scores for: Accuracy, Precision, Recall and F1 Score for the created models is shown in Figure 4.



Test and Score

Sampling

● Cross validation

Number of folds: 10

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

○ Test on train data

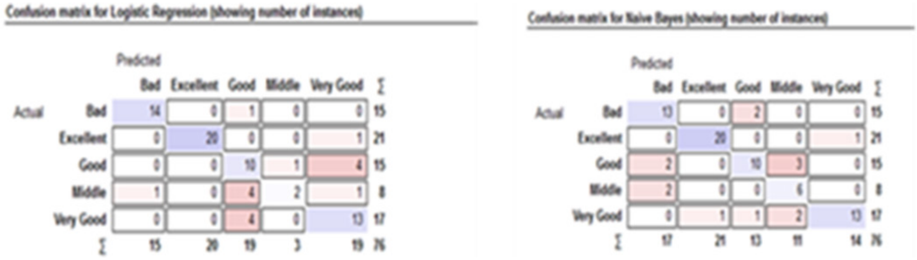
○ Test on test data

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Tree	0.982	0.974	0.973	0.975	0.974
SVM	0.991	0.908	0.906	0.917	0.908
Neural Network	0.977	0.868	0.866	0.869	0.868
Naive Bayes	0.991	0.908	0.910	0.916	0.908
Logistic Regression	0.945	0.816	0.816	0.836	0.816

Fig. 4. "*Test and score*" widget's workflow and results

Step 7. Visualization of the accuracy of the working models. To visualize the results better, drag and drop from the "Test and Score" widget to find "Confusion Matrix", ROC Analysis or another tool that can be used to test the different models and show the accuracy of their performance. For example, when selecting the "Confusion Matrix" tool and selecting the "Logistic Regression" or "Naïve Bayes" model, the corresponding confusion matrix is displayed, examples are shown in Figure 5.



Confusion matrix for Logistic Regression (showing number of instances)

		Predicted					
		Bad	Excellent	Good	Middle	Very Good	∑
Actual	Bad	14	0	1	0	0	15
	Excellent	0	20	0	0	1	21
	Good	0	0	10	1	4	15
	Middle	1	0	4	2	1	8
	Very Good	0	0	4	0	13	17
	∑	15	20	19	3	19	76

Confusion matrix for Naive Bayes (showing number of instances)

		Predicted					
		Bad	Excellent	Good	Middle	Very Good	∑
Actual	Bad	13	0	2	0	0	15
	Excellent	0	20	0	0	1	21
	Good	2	0	10	3	0	15
	Middle	2	0	0	6	0	8
	Very Good	0	1	1	2	13	17
	∑	17	21	13	11	14	76

Fig. 5. Sample results from "*Confusion matrix*" tool

Step 8. Anticipating new data. Through the trained algorithms, when setting a new combination of values for the selected components, the output variable is expected to be determined. At this step, we consider the model ready for practical application. The model gains independence and makes its own conclusions based on data sets and training.

In Figure 6 presents the prediction workflow using the Prediction widget of the Orange system. The new data is fed through a Test.xlsx file that has the same structure as the original data table, but the output variable column is not set. From the Evaluate menu, the Predictions widget is selected, which performs the predictions on the data from the Test.xlsx file and determines the predicted value of the column. The resulting file has an extension (.ows) and can be opened with any version of the Orange system.

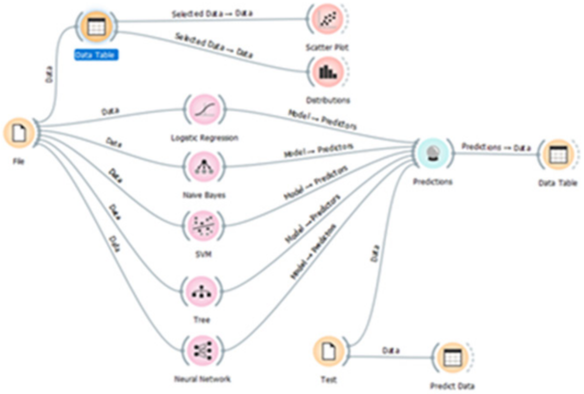


Fig. 6. Workflow with prediction widget

2.3 An experiment through the Orange system

In the example, following the eight steps of the algorithm, an approach for automated document analysis is implemented. Determination of the significant words in the document and the type of its content is performed, based on subject-oriented ontologies in the field of smart crop production. The research is aimed at the application of artificial intelligence to support Bulgarian crop production [4]. To find the significant words for the document we look for the frequency of occurrence of the words, in this case using the Word Cloud software (wordclouds.com). By pasting a text, document or URL, the software generates a list of all words and images in the text and the number of occurrences. Filters can be imposed on the resulting word list and its size. Each word from the resulting set is then searched in the dictionary with the concepts of the subject ontology associated with the domain. A degree of proximity is determined for each word in the selected set with all words in the ontology dictionary and the smallest value obtained is taken. The degree of closeness can be determined by applying different algorithms. In our technology solution we use q-gram distance. The q-gram distance between two strings x and y is defined in the following way:

$$D_q(x,y) = \sum_{v \in \Sigma^q} | G(x)[v] - G(y)[v] | \tag{1}$$

where $G(x)[v]$ denotes the total number of the occurrences of v in x. The proposed approach calculates the degree of closeness of the concepts and relations, retrieved from the ontology, to the keywords, retrieved from the user's document.

In this way, expressions that are "central" to the document are searched for, using functions for the frequency of occurrence of the words or phrases; the number of words that appear in the title of the document or in the headings of sections; position of the sentence or phrase in the document and in the section. In the experiment, real articles are determined for the subject area – smart crop production from genebank (data on plant genetic resources).

The National Collection of Plant Genetic Resources includes cultivated species and their wild relatives. It is linked to the European Electronic Catalog of Plant Genetic Resources EURISCO (http://eurisco.ipk-gatersleben.de). The introduction of the European database requires an annual report on the species registered in the current year in the database. The Genebank is a component of the national program of the "smart crop production".

In the conducted experiment, we used 82 off-topic documents along with 94 actual subject-area articles. On average, about 160 central words are extracted for each document, and for each of them a similarity to the concepts of the subject ontology dictionary is calculated based on a q-gram function. Thus, by analyzing the documents, a new data set is created with extracted central words and their degree of proximity to the concepts of the field. 2/3 of this data is fed to training classification algorithms, with the target attribute being whether the document is related to the domain or not.

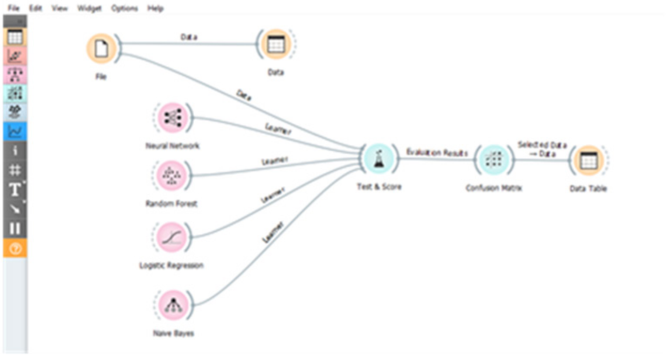


Fig. 7. Creating and evaluating the models

To solve the classification problem, we apply the techniques of the Orange Data Mining System application. We create a workflow process by loading the data's keywords, the degree of proximity and the document belonging to the area (0/1). We consistently create models applying the system's tools: Neural Network, Random Forest, Logistic Regression and Naïve Bayes.

After training, testing is performed with the remaining 1/3 of the data and a score is determined for the accuracy and precision of each of the algorithms using the Test&Score tool. The work process is shown in Figure 7. Our goal is to create a classifier that assigns a new document to the classes {domain unrelated = 0, domain related = 1}. The output of the Test&Score tool is a table of scores, regarding efficiency, precision, recall (sensitivity) and F1 score for the created models.

2.4 Reinforcement learning

Reinforcement Learning [12] also belongs to Supervised Learning machine learning algorithms. The characteristic of this type of algorithms is that it imitates a psychological model, in which the system is given "rewarding" and "punishing" signals in order to maximize the probability of receiving a "reward" and minimize "punishments" [14]. This training is typically applied in the absence of a predefined "correct" training data set. Such an approach is different from the supervised learning approach, where the goal is to reduce the deviation according to pre-set correct data (input/output).

Reinforcement learning occurs as a sequence of trial actions that gradually lead to reinforcement of good actions and avoidance of inappropriate ones. A cumulative reward feature is triggered. The result of the training is an optimal strategy for action in any situation. A strategy is optimal if it manages to maximize the sum of all rewards received during its execution. The agent makes decisions in discrete steps, and each subsequent decision depends on the current state of the environment.

3 Unsupervised learning with Orange system

Algorithms of this type take a dataset containing only input values and find a structure or distribution in the data, with no indication of a known variable or reward function, the data is unlabeled, there are no training examples [5]. Main types of solved tasks are: Dimensionality reduction, Density estimation, Clustering.

Cluster analysis is the distribution of a set of observations into subsets (clusters), so that the observations in the same cluster are similar according to one or several predefined criteria, and the observations from different clusters are different [7]. Different clustering techniques work with different assumptions about the structure of the data, which are often defined by some similarity metric and evaluated, for example, by internal compactness, or closeness between members of the same cluster, and difference between clusters.

Many clustering methods have been developed, each using a different induction principle. In [8] propose the division of clustering methods into two main groups: hierarchical and partitioning methods. Han and Kamber [9] propose a categorization of the methods into three additional main categories: density-based methods, pattern-based clustering, and network methods. An alternative categorization based on the induction principle of different clustering methods is presented in Estivill-Castro [10].

3.1 Algorithms based on clustering

These algorithms generate different partitions and then evaluate them according to some criteria. They are called non-hierarchical because each instance is placed in exactly one of k mutually exclusive clusters. The desired number of clusters is required to be entered in advance.

One of the most commonly used clustering algorithms is the k-means clustering algorithm. This type of algorithm belongs to Exclusive Clustering because the data is clustered so that if a certain data belongs to a given cluster, it cannot be included in another cluster. Another type of Overlapping Clustering, uses fuzzy sets to group data so that each point can belong to two or more clusters with different degrees of membership.

The algorithm performs three steps: determines the central coordinate (centroid), determines the distance from each object to the center, groups the objects based on the smallest distance. Terminates when centroids stop moving or some threshold is reached (e.g. number of iterations). Following is a sample workflow for applying k-means tool to a data set and visualizing the result using "Scatter Plot" tool (Figure 8).

Algorithm for building the information stream:

Step 1. Creation of Data. The desktop and toolset must be loaded and the data for analysis is selected via the "File" widget. An "Impute" tool can also be applied to clean the data.

Step 2. Selecting the k-means widget. This Orange system tool implements the k-means clustering algorithm.

Step 3. Preview the result. The k-means tool itself does not visualize a result, for this purpose a visualization tool, for example a Scatter Plot, must be connected to the data stream.

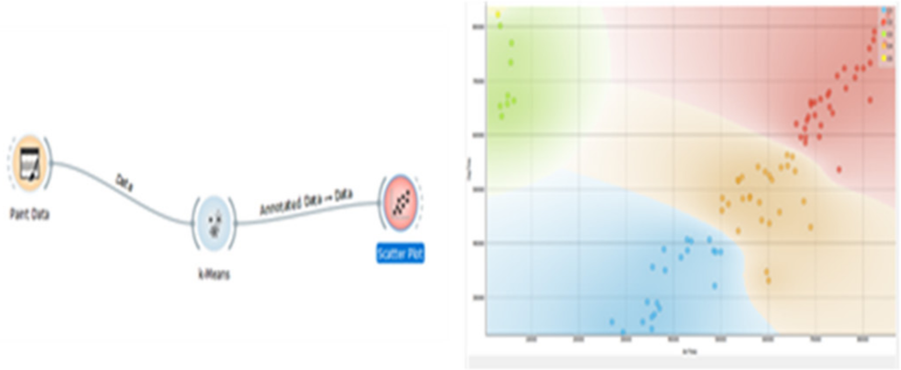


Fig. 8. Workflow with *k-means* widget and visualization of result via scatter plot

3.2 Hierarchical clustering

When the number of clusters is not predetermined, hierarchical clustering procedures (Hierarchical Cluster) are used. These algorithms start by declaring each point as its own cluster and then merge two most similar clusters until a stopping criterion is satisfied. The wide variety of procedures arises from the metric used between different objects. Examples of distance calculation metrics: Euclidean distance; Sum of Absolute Differences (SAD or L1 Norm); Sum of Squared Differences (SSD or L2 Norm); Mean Squared Error (MSE) etc.

Hierarchical clusterization in Orange computes a hierarchical clustering of arbitrary object types from the matrix of distances between them and displays the corresponding dendrogram. It starts by assigning each element to a cluster, so with N elements there are N clusters, each containing only one element. Next, the closest (most similar) pair of clusters is found and merged into one cluster. Next is a calculation of distances (similarities) between the new cluster and each of the old clusters. The steps are repeated until all elements are grouped into one cluster of size N.

An algorithm for constructing an information flow in hierarchical clustering involves applying the following steps (Figure 9):

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. To calculate the distance between elements. Different geometric metrics can be used to group the data to calculate distance. A Euclidean metric is a measure of the distance between points plotted on the Euclidean plane. Manhattan metric is a measure where distance is calculated as the sum of the absolute value of the differences between two points plotted on the Cartesian coordinate system. The Minkowski distance metric is a generalization of the distance metrics from the Euclidean metric and the Manhattan metric.

Step 3. Selecting the Hierarchical Clustering widget. This tool of the Orange system implements an algorithm for the hierarchical grouping of arbitrary types of objects by the calculated distances and displays the corresponding dendrogram. A dendrogram is a graph-tree in which each node represents one step of the clustering process. It also carries additional information about the distance between the two clusters.

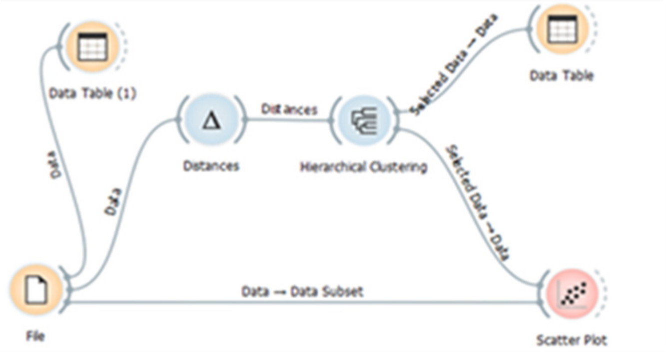
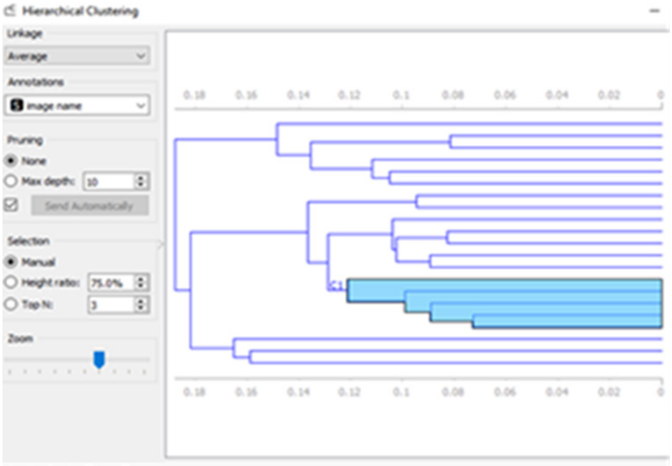


Fig. 9. Hierarchical clustering workflow

Step 4. Preview the result. Through different instruments, the result can be visualized in different ways, for example through the Data Table and Scatter Plot.

To apply hierarchical clustering applying the Image Analytics widget, a source image file is loaded [1]. In the conducted experiment, we introduce 21 images in jpeg format of domestic animals. Images can be visualized via the Image viewer widget.



Hierarchical Clustering

Linkage

Average

Annotations

image name

Pruning

● None

○ Max depth: 10

☑ Send Automatically

Selection

● Manual

○ Height ratio: 75.0%

○ Top N: 3

Zoom

Fig. 10. Dendrogram – result of the work of the hierarchical clustering widget

Next is setting the distance metric: cosine, for evaluating the proximity of the images using the Distances widget. Applying the Hierarchical clustering widget, clusters are created from the input images and the result is a dendrogram shown in Figure 10, where a camel, cow, bull, horse and donkey appear in one cluster C1.

Scientific research related to the application of artificial intelligence and block chain technologies, conducted within the framework of the national programs on intelligent agriculture and intelligent animal husbandry, aims to support the development of agriculture as a high-tech, sustainable, high-productivity and attractive sphere of the economy that will contribute to improving the living conditions of farmers and rural areas in general [11, 13].

4 Conclusions

Emerging technologies are characterized by a radical novelty, fast grow and impact on the other technologies. In other words, such technologies emerge and evolve in time and have the potential to cause significant impact on the connected to them processes of knowledge production. Despite their indeterminacy, can be proposed algorithms in the form of a multi-step process for developing an information flow with Orange system's machine learning tools, that can be used for research, analysis, and traing.

Orange is a software product that is constantly being developed and supplemented. Mathematical modeling is implemented more easily with the new tools in the different versions of the product. For example, Multi-Criteria Decision Making is used when decisions need to be made that involve several different criteria. Taylor polynomials are used to quickly calculate the value of the function in any region of interest. The Marcov chains are a tool for predicting value estimates in the future based on the current values of the data in the data set.

Exactly which algorithm to choose when solving a given problem depends on many factors [14, 15]. For example, from the size and type of data, also from its quality and quantity. It depends on what the result would serve as well as the time available to reach a result. The field under consideration is highly dynamic, with the development of both notions of natural intelligence and many related sciences.

Machine learning has spawned a new set of concepts and technologies, new algorithms, analysis tools, and more. The growing complexity and interdependence of technologies is accompanied by a significant increase in risk factors, combined with not always adequate means of ensuring information security.

As a result of the research related to the purpose of the work, a multi-step algorithms was formulated, using Orange's tools, for constructing: machine learning models for classification and regression; an information flow on clustering; an information flow in hierarchical clustering. These algorithms represent practical approaches, that can be used for research, analysis, and training in various fields.

It should be held experiments with quantitative data on the technological and economic negative impact of digital risks and applications of various algorithms for multicriteria decision making, with an emphasis on fuzzy logic, neural networks and other intelligent techniques.

Future efforts will be focused on researching new technologies and risks, building practical models for risk analysis and tools to construct conventional harmony "artificial intelligence – natural intelligence".

5 Acknowledgment

This work was supported by the Bulgarian Ministry of Education and Science under the National Research Program "Smart crop production" approved by Decision of the Ministry Council №866/26.11.2020 г. and by the Scientific Research Fund "BG PLANTNET establishment of national information network genebank – plant genetic resources", project KP-06-N36.

6 References

[1] Orange DataMining system [Online]: https://orange.biolab.si/training/introduction-to-data-mining/

[2] Popchev I., & Orozova D., (2019), Towards Big Data Analytics in the E-learning Space, Cybernetics and Information Technologies, Vol. 19, No. 3, pp. 16–25. https://doi.org/10.2478/cait-2019-0023

[3] Popchev I., & Orozova D., (2020), Towards a Multistep Method for Assessment in E-Learning of Emerging Technologies, Cybernetics and Information Technologies, Vol. 20, No 3, pp. 116–129. https://doi.org/10.2478/cait-2020-0032

[4] Popchev I., & Orozova D., (2020), Text Mining in the Domain of Plant Genetic Resources, Proceedings of 2020 IEEE 10th International Conference on Intelligent Systems, pp. 596–600, ISBN 978-172815456-5. https://doi.org/10.1109/IS48319.2020.9200174

[5] Tripathy B. K., Sundareswaran A., & Ghela S., (2021), Unsupervised Learning Approaches for Dimensionality Reduction and Data Visualization, CRC Press, ISBN-101032041013. https://doi.org/10.1201/9781003190554

[6] López César Pérez, (2022), Machine Learning. Supervised Learning Techniques: Regression. Examples with SAS and MATLAB, ASIN:B09QH7S84F.

[7] Kubat Miroslav, (2021), An Introduction to Machine Learning, Third Edition. Springer, ISBN-103030819345. https://doi.org/10.1007/978-3-030-81935-4

[8] Bakshi Elisabeth, (2021), Unsupervised Learning – Identify Machine Learning Tasks: Machine Learning Course, Independently Published, ISBN 9798729137855.

[9] Han J., & Kamber M., (2001), Data Mining: Concepts and Techniques. Morgan Kaufmann Publish.

[10] Estivill Castro, V., & Yang, J. A., (2000), Fast and Robust General Purpose Clustering Algorithm. Pacific Rim International Conference on Artificial Intelligence, pp. 208–218. https://doi.org/10.1007/3-540-44533-1_24

[11] Vera Toktarova, (2022), Model of Adaptive System for Mathematical Training of Students within eLearning Environment, International Journal of Emerging Technologies in Learning (iJET), Vol. 17, No. 20, pp. 99–117. https://doi.org/10.3991/ijet.v17i20.32923

[12] Millington, I., (2006), Artificial Intelligence for Games. Reinforcement Learning, Elsevier Inc. Ch.7.6, pp. 612–628.

[13] Doukovska L., (2021), Artificial Intelligence to Support Bulgarian Crop Production, Engineering Sciences, LVIII, pp. 30–48. https://doi.org/10.7546/EngSci.LVIII.21.04.03

[14] Charitopoulos A., Rangoussi, M., Koulouriotis. D., (2022), Blending E-Learning with Hands-on Laboratory Instruction in Engineering Education an Experimental Study on Early Prediction of Student Performance and Behavior, International Journal of Emerging Technologies in Learning (iJET), Vol. 17, No. 20, pp. 213–230. https://doi.org/10.3991/ijet.v17i20.33141

[15] Wei Na, (2020), A Data Mining Method for Students' Behavior Understanding, International Journal of Emerging Technologies in Learning (iJET), Vol. 15, No. 06, pp. 4–17. https://doi.org/10.3991/ijet.v15i06.13175

7 Authors

Ivan Popchev, Bulgarian Academy of Sciences, Sofia, Bulgaria.
Daniela Orozova, Trakia university, Stara Zagora, Bulgaria.

Article submitted 2022-11-17. Resubmitted 2023-01-10. Final acceptance 2023-01-13. Final version published as submitted by the authors.