Classification of chest X-Ray images using Orange Data Mining Tool

Lilyana S. Koleva, Georgi A. Filipov

In this article, the capabilities of the Orange Data Mining software product to classify chest X-ray images are shown. Viewed images of patients are divided into three categories: healthy, with Covid-19 and with viral pneumonia. Several classification methodologies are reviewed and compared, such as: neural networks, nearest neighbor method, decision tree, etc.

Keywords – data mining, image classification, neural networks, orange.

Класификация на рентгенови изображения на гръдния кош с помощта на Orange Data Mining Tool (Лиляна С. Колева, Георги А. Филипов). *Статията разглежда възможностите на софтуерният продукт Orange Data Mining да класифицира изображения рентгенография на гръден кош. Разглежданите изображения на пациенти са разделени в три категории: здрави, с Covid-19 и с вирусна пневмония. Разглеждат се и се сравнява няколко методологии за класифициране, като: невронни мрежи, метода на най-близките съседи, дърво на решенията и др.*

Introduction

In general, data mining is the analysis of data from different perspectives and summarizing it into useful information that can be used to increase sales and reduce costs. Data mining software is one of the tools for analyzing data. It allows the information to be analyzed, categorized and to summarize the interrelationships discovered. Data mining is the process of discovering relationships or patterns among dozens of fields in large relational databases [1].

The goal of this paper is to classify and group chest X-ray images according to their content using the Orange Data Mining software, extracting information from them and converting them into a form suitable for machine learning. To demonstrate how by feeding the images to a neural network, logistic regression and other classification methods, can hierarchically group or classify a large batch of pictures and extract meaningful information or set of imagines according to their content.

Data mining

Data mining is a process of analyzing the stored databases in the direction of extracting new useful information by revealing the deep and hidden relationships between seemingly unknown and unrelated quantities. Its important feature is that it provides the ability to process multidimensional arrays and extract multidimensional dependencies, while automatically revealing exceptional situations - data and cases not included in general regularities. Data mining analysis automatically makes hypotheses to reveal dependencies between different components and parameters.

At the core of modern Data mining technologies is the concept of patterns or models reflecting the fragmented multifaceted relationships between data. These templates represent a set of regularities, selection of data according to given properties, which are appropriately presented in forms easily accessible to users. To create these templates, methods are applied that do not limit the basic assumption in the model structure and the type of distributed values of the analyzed indicator [2].

Data mining software analyzes relationships and recurring patterns in warehoused transaction data based on open-ended user queries. There are several types of analytical software - statistical, machine learning and neural network. In principle, one of four types of interrelationships is sought:

- *Classes* - Stored data is used to localize information into predefined groups. For example, a flower shop might use consumer order data mining to determine when customers visit and what they typically order. This information can be

used to increase store traffic by offering appropriate bouquets and ordering more of the flowers in demand, thus reducing losses from less desirable flowers for the season;

- *Cluster* - Data is grouped by logical relationships or user preferences. For example, data may be extracted to identify market segments or similar user behavior;
- *Associations* - Data can be mined to identify associations;
- *Consistent Patterns* - Data is mined to predict patterns of behavior and trends. For example, a sporting goods retailer might predict the likelihood of purchasing a backpack based on the purchase of a sleeping bag and hiking boots [3].

The data mining process can be divided into the following four main stages:

Stage 1: Data collection: Relevant data for applying analysis are identified and collected.;

Stage 2: Preparing the data: This stage involves a set of steps to prepare the data for extraction. Begins with data exploration, profiling, and preprocessing, followed by correcting errors and other quality issues in obtained data;

Stage 3: Data Extraction: This is done by applying one or more data processing algorithms;

Stage 4: Data analysis and interpretation: The results obtained from data mining are used to create analytical models that can help in decision making.

Orange Data Mining Tool

Orange is an excellent software package for machine learning and data mining. It best supports data visualization and is component-based software. It is written in the Python computer language.

Because it is component-based software, Orange's components are called "widgets". These modules range from data visualization and preprocessing to algorithm evaluation and predictive modeling.

Widgets offer basic functions such as:

- It displays a data table and allows you to select functions;
- Data reading;
- Predictors for training and comparing learning algorithms;
- Visualization of data elements, etc.

Furthermore, Orange brings a more interactive and enjoyable atmosphere to the mundane analysis tools. It is quite interesting to use.

The data coming into Orange can be quickly formatted according to the desired model and can be easily moved, if necessary, through simple drag-and-drop actions. It allows users to make better decisions in a short amount of time through quick data comparison and analysis [5].

Methodology

Artificial neural networks

Artificial neural networks (ANN) are models based on biological neural networks. An artificial neural network is a system for parallel information processing that possesses the ability to store and utilize experimental knowledge. It models the activity of its biological equivalent “the brain” in the following two aspects:

- Information is accumulated in the ANN through a process of learning;
- The strength of connections between individual nodes (neurons) are modeled by weights on the corresponding links, which are used to store information.

In general, ANNs consist of simple information processing units called neurons or nodes. Neurons are interconnected, and the weights of the connections between them determine the strength of the respective links. The input information for each neuron is the weighted sum of signals from the other neurons. This information is accumulated within the neuron, and its output signal is determined through the use of an activation or transfer function [6].

The most commonly encountered architectures of ANNs consist of several distinct (sequential) layers of elements (neurons), where the elements in the lowest layer serve as input devices for the network. They perceive signals from the external environment, while the elements in the topmost layer act as output for the network, producing the result of the network's operation, which is essentially based on the input signals and the weights of the connections in the system. Often, in these ANNs, the connections are unidirectional and link the elements of one layer to the elements of the layer immediately above it. Depending on the number of layers in the network, we speak of two-layer neural networks (consisting of only an input and an output layer, lacking a so-called hidden layer) and multilayer neural networks (having at least one hidden layer).

Random Forest model

The Random Forest (RF) model is a machine learning algorithm used for tasks such as classification, regression, and more. It is based on the idea of an ensemble of decision trees.

In the RF model, a large number of decision trees are created, where each tree is trained on different

subsets of the data (randomly selected subsets of the training set) and with randomly selected features from the feature set. During classification, each individual tree model provides its own classification response, and during regression, it predicts its own value. The results from all the trees are then combined, using either the average value (for regression) or voting (for classification) from the individual trees, to obtain the final prediction.

The RF model is known for its ability to handle overfitting and provide reliable predictions even with a small number of trees. It is widely used in practice due to its effectiveness, easy implementation, and capability to handle large datasets [7].

k-Nearest Neighbors

The k-Nearest Neighbors (k-NN) model is a machine learning algorithm used for classification and regression tasks. It is based on the principle that similar examples have similar classifications or values for the target variable.

In classification, the k-NN model works as follows: For a given new example, the algorithm searches for the k closest training examples from the training set, where k is a predefined parameter. Then, using voting, the model determines the classification of the new example by considering the most frequently occurring class among its neighbors.

In regression, the k-Nearest Neighbors model uses the average value of the target variable from the k nearest neighbors as the predicted value for the new example [7].

Image Classification with Orange Data Mining Tool

A schema has been built in Orange for classifying chest radiographs categorized into three categories (Fig. 1): 50 images of patients who have had a coronavirus infection, 50 images of patients who have had viral pneumonia, and 50 images of normal radiographs from healthy patients [8]. For testing, we will provide 27 images from each category, totaling 81 shuffled and uncategorized images of lung radiographs.

a) *b)* *c)*

Fig.1. Chest X-ray images: a) healthy patients, b) coronavirus infection, c) viral pneumonia [8].

Before starting the image processing, the images need to be embedded and transformed into a suitable format for machine learning. This is done by passing them through a so-called "embedder," which represents a pre-trained ANN embedded within Orange. For the purposes of this example, the Google Inception v3 neural network is used. It is a deep learning model trained for image recognition tasks. The activations from the penultimate layer of the model, which represent the images as vectors, are used for embedding [9].

To embed the images into the neural network and obtain additional features for them, it is necessary to connect the loaded images from the Import Images tool (Fig. 2) to the Image Embedding tool (Fig. 3).

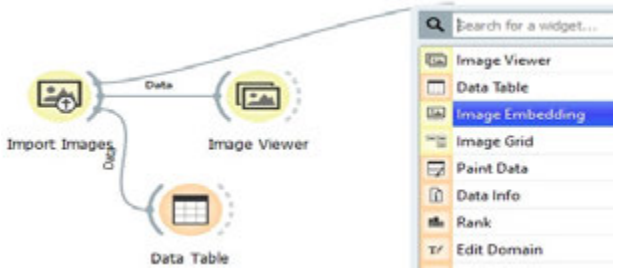

Fig.2. Import Images Tool.

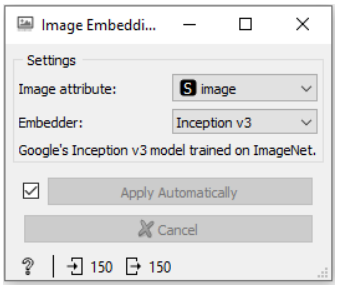

Fig.3. Image Embedding tool.

The Image Embedding tool takes the loaded images as input and applies the pre-trained neural network (Inception v3 in this case) to extract meaningful features from the images. These features are then used as inputs for further analysis or classification tasks.

After the images are embedded and classified, they are used to train various models that can later be used to classify new images into the correct categories. Four types of models have been trained:

- Artificial neural network with 350 hidden neurons structure (Fig. 4 – "1");
- Logistic regression model (Fig. 4 – "2");
- Random Forest model (Fig. 4 – "3");
- k-Nearest Neighbors (Fig. 4 – "4").

Fig. 4 shows a scheme where the four training models are connected to the Test and Score tool – "5",

which performs the training and evaluation of the trained models. After that, the Confusion Matrix tool – "6" is connected, allowing you to see where each model made mistakes in classifying the images.

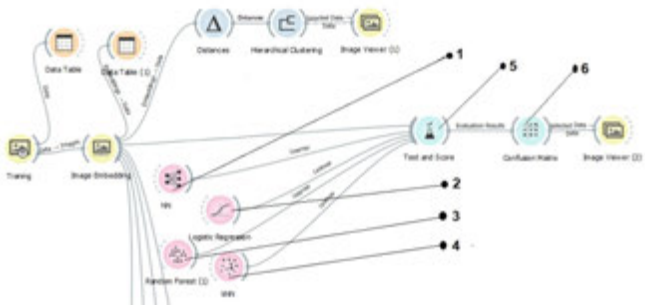

Fig.4. Scheme of the four training models, labeled as follows: 1 - Three-Layer Neural Network with 350 neurons in the hidden layer; 2 - Logistic Regression Model; 3 - Random Forest Model; 4 - k-Nearest Neighbors Model; 5 - Test and Score Tool; 6 - Confusion Matrix Tool.

Table 1 shows the result of training the selected models, where it can be seen that the Logistic regression model achieves the best classification accuracy with a CA (Classification Accuracy) value of 0.798 or 79.8% correctly classified images.

Fig. 5 displays the confusion matrix of the Logistic regression model. The diagonal of the matrix represents the correctly classified images for each category.

Table 1

Training results

Model	ACU	CA	F1	Precision	Recall
kNN	0.903	0.780	0.780	0.799	0.780
Rand. Forest	0.865	0.713	0.712	0.711	0.713
N350	0.921	0.793	0.790	0.790	0.193
Logistic Reg.	0.925	0.798	0.798	0.798	0.798

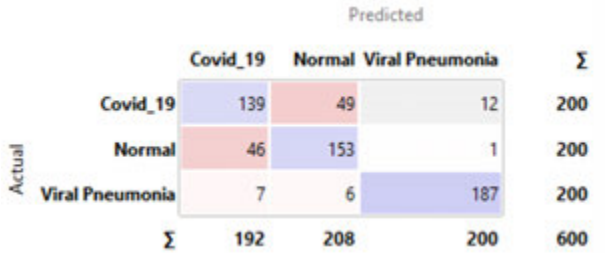

Fig.5. Confusion matrix of the Logistic regression model.

During the training, the Orange algorithm utilizes 4 iterations of all provided images. This means that each model classifies each individual image 4 times, resulting in the total number of images per category increasing from 50 to 200.

From Fig. 5, it can be observed that the most accurate logistic regression model correctly categorizes 153 patients as healthy. However, it misclassifies 46 healthy patients as having a coronavirus infection and one patient as having viral pneumonia. The model correctly categorizes 139 patients with a coronavirus infection, but incorrectly categorizes 49 as healthy and 12 as having viral pneumonia. In the case of viral pneumonia, the model appears to have the highest percentage of correctly classified images: 187 from 200 patients with viral pneumonia. However, according to the model, 6 patients are healthy and 7 have a coronavirus infection. The assumption that healthy individuals are sick is not dangerous and will be corrected when a healthy person is referred for treatment. In this case, it is more concerning that X-rays indicating coronavirus and normal pneumonia are labeled as healthy.

The trained models have been tested with dataset of 81 new uncategorized chest X-ray images. The testing of the trained models with the new image dataset is conducted in several steps. First, the models are duplicated from the trained models to preserve the setting obtained during training. The new images need to be loaded into Orange and "embedded". Once the images are embedded, they are passed through the trained models, which are connected to the "Predictions" widget. The predictions are then displayed in form of table, showing how each model has categorized the images and which images were misclassified.

Fig. 6 presents the entire constructed scheme, including the connection of the selected models to the "Predictions" widget. This allows for an overview of how the models have classified the images into categories and which images were misclassified.

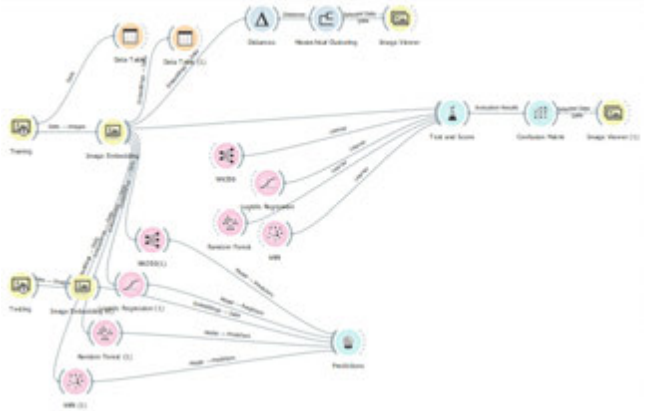

Fig.6. Scheme and connection of the trained models to the Predictions tool.

Fig. 6 contains also widget named "hierarchical clustering". The function of the widget is to group unclassified data/images into clusters. In this particular case, due to the great similarity of the chest

X-rays, the automatic grouping using the widget does not give good enough results. Therefore, to evaluate the image classification models, the images are pre-divided into 3 categories: Covid-19, Normal and Viral Pneumonia.

	NN350	Logistic Regression (1)	kNN	Random Forest (1)	image name
1	Covid...	Covid_19	Covid...	Covid_19	COVID-352
2	Covid...	Covid_19	Normal	Covid_19	COVID-353
3	Covid...	Covid_19	Covid...	Covid_19	COVID-354
4	Covid...	Covid_19	Covid...	Covid_19	COVID-355
5	Covid...	Covid_19	Covid...	Covid_19	COVID-356
6	Covid...	Covid_19	Normal	Covid_19	COVID-357
7	Covid...	Covid_19	Normal	Covid_19	COVID-358
8	Covid...	Covid_19	Covid...	Covid_19	COVID-359
9	Covid...	Covid_19	Covid...	Covid_19	COVID-360
10	Covid...	Covid_19	Normal	Normal	COVID-361
11	Covid...	Covid_19	Covid...	Normal	COVID-362
12	Covid...	Covid_19	Covid...	Normal	COVID-363
13	Normal	Covid_19	Normal	Covid_19	COVID-364
14	Covid...	Covid_19	Normal	Normal	COVID-365
15	Covid...	Covid_19	Normal	Covid_19	COVID-366
16	Covid...	Covid_19	Normal	Covid_19	COVID-367
17	Covid...	Covid_19	Covid...	Covid_19	COVID-368
18	Covid...	Covid_19	Normal	Covid_19	COVID-369
19	Covid...	Covid_19	Covid...	Covid_19	COVID-370
20	Covid...	Covid_19	Covid...	Covid_19	COVID-371
21	Normal	Normal	Normal	Covid_19	COVID-372
22	Covid...	Covid_19	Covid...	Covid_19	COVID-373
23	Covid...	Covid_19	Normal	Covid_19	COVID-374
24	Covid...	Covid_19	Covid...	Covid_19	COVID-375
25	Covid...	Covid_19	Covid...	Covid_19	COVID-376
26	Covid...	Covid_19	Normal	Covid_19	COVID-377
27	Covid...	Covid_19	Covid...	Normal	COVID-378

Fig.7. A sample of the classification results of the new images.

Fig. 7 displays a sample of the results provided by the "Predictions" application. The white columns contain a list of classified images using the four trained models, while the gray column represents the actual input images, which have been pre-named. From the figure, it can be seen that the k-nearest neighbors (kNN) method has misclassified the highest number of patients with a coronavirus infection as healthy, whereas logistic regression has made only one error.

Conclusion

The article explores the capabilities of the software product Orange Data Mining to classify images using four different methods: Artificial Neural Networks, Logistic Regression, k-Nearest Neighbors, and Random Forest. It has been shown that Logistic regression provides the best results in terms of accurately classifying chest X-ray images of healthy individuals, those with viral pneumonia, and those with a coronavirus infection, followed by Artificial neural networks.

However, categorizing the images proved to be challenging for the trained models, as they failed to classify the images in the "normal" category, mistakenly resembling them to viral pneumonia.

From the conducted experiments, it can be seen that the use of Logistic regression and Artificial neural networks can be helpful in interpreting chest X-ray images. However, due to the high similarity between the images, the results should be reviewed by a specialist before making a final decision about an individual's health.

REFERENCES

[1] Maman, O. What is Data Mining and Data Extraction – A Full Overview: https://netnut.io/what-is-data-mining-and-data-extraction-full-overview/

[2] Kolev, G., V. Lozeva, E. Koleva. Digital Twin Software Overview with Text Mining Techniques. 2022 International Conference Automatics and Informatics (ICAI), Varna, Bulgaria, 2022, pp. 153-158.

[3] How Data Mining Works: A Guide: https://www.tableau.com/learn/articles/what-is-data-mining

[4] Craig Stedman, Data Mining. 2021, https://www.techtarget.com/searchbusinessanalytics/definition/data-mining

[5] Top 15 Best Free Data Mining Tools, https://myservername.com/top-15-best-free-data-mining-tools#1_Xplenty.

[6] Mladenov, V., S. Yordanova. Faculty of Automation. Fuzzy control and neural networks (in Bulgarian), Technical University - Sofia, 2006.

[7] Hoarau, A., A. Martin, J.-Ch. Dubois, Y. Le Gall. Evidential Random Forests, Expert Systems with Applications, Vol. 230, 2023, 120652.

[8] Kaggle, https://www.kaggle.com/

[9] Image Embedding: https://orangedatamining.com/widget-catalog/image-analytics/imageembedding/

Assist. Prof. Dr. Eng. Lilyana S. Koleva – *She is an Assistant professor at the University of Chemical Technology and Metallurgy, Sofia. Her fields of interest are: modelling, parameter optimization, automation, electron beam technologies, data analysis.*

e-mail: sura@uctm.edu

Eng. Georgi A. Filipov - *He has completed a Master's degree in "Information Technology" at UCTM - Sofia, in 2023.*

He has an interest in programming, embedded systems, microcontrollers, the Arduino platform, precision mechanics, mechatronics, and machinery.

e-mail: mr.georgifilipov@gmail.com.

Received on: 17.06.2023