Decision tree approach for IRIS database classification

Marcelina Lachowicz
Institute of Mathematics
Silesian University of Technology
Kaszubska 23, 44-100 Gliwice, Poland
Email: marcelina.lac@o2.pl

***Abstract*—Data classification is one of important topics in information technology. There are many methods used in this process. This article presents classification of iris flowers by the use of decision tree. In the system was implemented a procedure to used an open data set for classification of applied types of flowers by numerical features describing them. Results show that proposed model is very simple but also efficient in numerical classification.**

I. Introduction

In information processing very important role is for data analysis and decision processes. From information hidden in data we can get knowledge about various things. In general we can use various methods to process the data. Artificial intelligence gives us many interesting approaches to data science.

In general input data is organized in smaller groups called classes in which final classification is done. By the use of this kind of decision making processes we can estimate many things. In [1] was implemented a method to estimate energetic efficiency. In [2], [3] was done a prediction on wind farming, while in [4], [5] cancer classification from medical images was done. Among methods of data science very often decision trees are used. Some of first approaches to use decision trees as classifiers were presented in [6], where toxic hazards estimation was done using decision trees. In [7] decision trees were used to help on remote sensing to classify land images. Recently there are many optimized decision tree structures developed for specific examples of input data.

In [8] was presented how to join decision tree with bee algorithm on the way for faster data classification. In [9] decision trees were joined with Bayesian methods to efficiently classify smoking cessation. There are many approaches where decision trees give very good results. An interesting survey over various decision tree approaches and their implementations was presented in [10].

In this article i show how to implement a simple python procedure based on decision tree classifier.

Implemented idea was used to recognize iris flowers from open data set. Results show that proposed implementation works well returning very good results.

II. Ronald Aylmer Fisher

Ronald Aylmer Fisher was born on February 17, 1890 in London, and died on July 29, 1962 in Adelaide. He was a British geneticist and statistician. The Briton graduated from Gonville and Caius College at the University of Cambridge and worked as a professor of eugenics at the London School of Economics in 1933-1943 and professor of genetics at the University of Cambridge (1943-1957). Anders Hald described him as ”a genius who almost created the foundations of contemporary statistics”, and Richard Dawkins as ”the greatest heir of Darwin” and a member of the Royal Society in London (Royal Society). The geneticist created, among others, maximum likelihood, analysis of variance (ANOVA) and linear discriminant analysis. He also dealt with methods of hypothesis verification using statistical methods (in anthropology, genetics, ecology) and was one of the creators of modern mathematical statistics. He is also known for the development of experimental results at the Rothamsted Institute of Agricultural Research (1919 - 1933) and as the author of the *Statistical Methods for Research Workers* (1925), *Statistical Methods and Scientific Inference* (1956).

III. Proposed classifier

The principle of proposed decision model is based on decision tree.

1) Decision Tree: A decision tree is a (graphical) method of supporting the decision-making process. It is used in decision theory and in machine learning. In the second application, it helps in acquiring knowledge based on examples.

Algorithm - it works recursively for each node. We have to decide whether the node will be:

1) leaf - we end this recursive call,
2) a branch node according to the values that the given attribute takes and for each child node we create a recursive call to the algorithm with the list of attributes reduced by the attribute just selected.

Building a tree - The tree consists of nodes: decisions and states of nature and branches. Rectangles are decisions, and states of nature are circles. We start with the root. At the very beginning we have the first given sepal length (figure

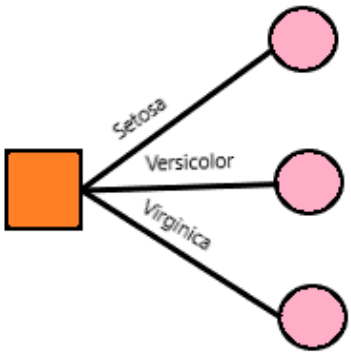

Fig. 1. First step in proposed decision tree reasoning.

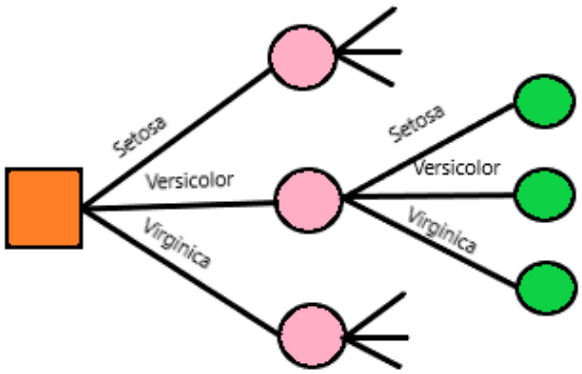

Fig. 2. Second step in proposed decision tree reasoning.

1), then we analyze the second variable - sepal width (figure 2). In this way, we continue the construction of the whole tree.

A. Coding

```
from sklearn import tree
import numpy as np
iris = open("iris.txt","r")
list = []
for line in iris:
data = line.split(",")
list.append(data)
list2 = [[float(column) for column in row]
for row in list]
print(list2)
iris_res = open("iris-result.txt","r")
res = []
for line in iris_res:
data = line.split(",")
res.append(data)
res2 = [[int(column) for column in row]
for row in res]
print(res2)
pr = tree.DecisionTreeClassifier()
pr = pr.fit(list2, res2)\\
```

The code was written in Python. As you can see, we have introduced the data of IRIS databases to the program. Then we entered the function to learn our network through a library that Python has:

```
sklearn.tree. DecisionTreeClassifier
```

IV. The IRIS database

The IRIS database contains a set of iris flower measurement and was first made available by Ronald Fisher in 1936. This is one of the most well-known collections, in addition, as we'll see in a moment is also very simple. The set of irises consists of 4 measurements of flower petals and a leaf: width and length. There are three types of flowers:

- **Versicolor** - This flower is found in North America and develops up to a height of 80 centimeters. The leaves of this plant have a width of more than one centimeter, and the roots form large and thick clumps. A well-developed plant has 6 blue petals and blooms from May to July, while large seeds can be observed appearing in autumn.

- **Setosa** - the flower is found in Canada, Russia, north- east Asia, China, Korea, southern Japan and Alaska. The plant has half-green leaves, high branched stems and purple-blue flowers similar to lavender (there are also pink and white flowers). The roots are shallow, large and rapidly spreading.

- **Virginica** - this flower is native to North America. The leaves are 1 to 3 centimeters long and sometimes longer than the flower stalk. The plant has 2 to 4 erect or arching, bright green. The roots are spread underground. The seeds are light brown and differently shaped, and are born in three-part fruit capsules. The petals vary in color from dark purple to pinkish- white. These plants bloom from April to May and have from 2 to 6 flowers.

V. Our example

Our database has 150 data (50 for each type of flower), of which 120 we used to learn the artificial neural network, and 30 (10 for each species) to test the artificial neural network. The question then arises: Do the data groups we have received correspond to the three species of iris? To see this, let's look at the error matrix.

A. Confusion Matrix

Confusion Matrix is the basic tool used to assess the quality of the classification. In our table, we consider three classes of abstraction as a result of which we get a 3 x 3 matrix. The table of errors arises from the intersection of the predicted class and the class actually observed. Our matrix is presented in figure 3. Now i want to analyze measure of results to show how proposed classification works.

1) Analysis of the Confusion Matrix: **Terminology and derivations from a confusion matrix:**

1) For Versicolor
   - *sensitivity, recall, hit rate or true positive rate* (**TPR**)
     $TPR = \frac{TP}{P} = \frac{TP}{TP+FN} = \frac{10}{10+0} = \frac{10}{10} = 1$
     $TPR = 1 - FNR$
   - *specificity, selectivity or true negative rate* (**TNR**)
     $TNR = \frac{TN}{N} = \frac{TN}{TN+FP} = \frac{20}{20+0} = \frac{20}{20} = 1$
     $TNR = 1 - FPR$
   - *precision or positive predictive value* (**PPV**)
     $PPV = \frac{TP}{TP+FP} = \frac{10}{10+0} = \frac{10}{10} = 1$
     $PPV = 1 - FDR$
   - *negative predictive value* (**NPV**)
     $NPV = \frac{TN}{TN+FN} = \frac{20}{20+0} = \frac{20}{20} = 1$
     $NPV = 1 - FOR$
   - textitmiss rate or false negative rate (**FNR**)
     $FNR = \frac{FN}{P} = \frac{FN}{FN+TP} = \frac{0}{0+10} = \frac{0}{10} = 0$
     $FNR = 1 - TPR$
   - *fall-out or false positive rate* (**FPR**)
     $FPR = \frac{FP}{N} = \frac{FP}{FP+TN} = \frac{0}{0+20} = \frac{0}{20} = 0$
     $FPR = 1 - TNR$
   - *false discovery rate* (**FDR**)
     $FDR = \frac{FP}{FP+TP} = \frac{0}{0+10} = \frac{0}{10} = 0$
     $FDR = 1 - PPV$
   - *false omission rate* (**FOR**)
     $FOR = \frac{FN}{FN+TN} = \frac{0}{0+20} = \frac{0}{20} = 0$
     $FOR = 1 - NPV$
   - *accuracy* (**ACC**)
     $ACC = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{10+20}{10+20+0+0} = \frac{30}{30} = 1$
   - F_1 *score* - harmonic mean of precision and sensitivity
     $F_1 = 2 \cdot \frac{PPV \cdot TPR}{PPV+TPR} = \frac{2TP}{2TP+FP+FN} = \frac{20}{20+0+0} = \frac{20}{20} = 1$
   - *Matthews correlation coefficient* (**MCC**)
     $MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}} = \frac{300}{10 \cdot 10 \cdot 20 \cdot 20} = \frac{300}{200} = \frac{3}{2}$
   - *informedness or Bookmaker Informedness* (**BM**)
     $BM = TPR + TNR - 1 = 1 + 1 - 1 = 1$
   - *Markedness* (**MK**)
     $MK = PPV + NPV - 1 = 1 + 1 - 1 = 1$
2) For Setosa
   - *sensitivity, recall, hit rate or true positive rate* (**TPR**)
     $TPR = \frac{TP}{P} = \frac{TP}{TP+FN} = \frac{10}{10+0} = \frac{10}{10} = 1$
     $TPR = 1 - FNR$
   - *specificity, selectivity or true negative rate* (**TNR**)
     $TNR = \frac{TN}{N} = \frac{TN}{TN+FP} = \frac{20}{20+0} = \frac{20}{20} = 1$
     $TNR = 1 - FPR$

- *precision or positive predictive value* (**PPV**)
  $PPV = \frac{TP}{TP+FP} = \frac{10}{10+0} = \frac{10}{10} = 1$
  $PPV = 1 - FDR$
- *negative predictive value* (**NPV**)
  $NPV = \frac{TN}{TN+FN} = \frac{20}{20+0} = \frac{20}{20} = 1$
  $NPV = 1 - FOR$
- textitmiss rate or false negative rate (**FNR**)
  $FNR = \frac{FN}{P} = \frac{FN}{FN+TP} = \frac{0}{0+10} = \frac{0}{10} = 0$
  $FNR = 1 - TPR$
- *fall-out or false positive rate* (**FPR**)
  $FPR = \frac{FP}{N} = \frac{FP}{FP+TN} = \frac{0}{0+20} = \frac{0}{20} = 0$
  $FPR = 1 - TNR$
- *false discovery rate* (**FDR**)
  $FDR = \frac{FP}{FP+TP} = \frac{0}{0+10} = \frac{0}{10} = 0$
  $FDR = 1 - PPV$
- *false omission rate* (**FOR**)
  $FOR = \frac{FN}{FN+TN} = \frac{0}{0+20} = \frac{0}{20} = 0$
  $FOR = 1 - NPV$
- *accuracy* (**ACC**)
  $ACC = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{10+20}{10+20+0+0} = \frac{30}{30} = 1$
- F_1 *score* - harmonic mean of precision and sensitivity
  $F_1 = 2 \cdot \frac{PPV \cdot TPR}{PPV+TPR} = \frac{2TP}{2TP+FP+FN} = \frac{20}{20+0+0} = \frac{20}{20} = 1$
- *Matthews correlation coefficient* (**MCC**)
  $MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}} = \frac{300}{10 \cdot 10 \cdot 20 \cdot 20} = \frac{300}{200} = \frac{3}{2}$
- *informedness or Bookmaker Informedness* (**BM**)
  $BM = TPR + TNR - 1 = 1 + 1 - 1 = 1$
- *Markedness* (**MK**)
  $MK = PPV + NPV - 1 = 1 + 1 - 1 = 1$

3) For Virginica

- *sensitivity, recall, hit rate or true positive rate* (**TPR**)
  $TPR = \frac{TP}{P} = \frac{TP}{TP+FN} = \frac{10}{10+0} = \frac{10}{10} = 1$
  $TPR = 1 - FNR$
- *specificity, selectivity or true negative rate* (**TNR**)
  $TNR = \frac{TN}{N} = \frac{TN}{TN+FP} = \frac{20}{20+0} = \frac{20}{20} = 1$
  $TNR = 1 - FPR$
- *precision or positive predictive value* (**PPV**)
  $PPV = \frac{TP}{TP+FP} = \frac{10}{10+0} = \frac{10}{10} = 1$
  $PPV = 1 - FDR$
- *negative predictive value* (**NPV**)
  $NPV = \frac{TN}{TN+FN} = \frac{20}{20+0} = \frac{20}{20} = 1$
  $NPV = 1 - FOR$
- textitmiss rate or false negative rate (**FNR**)
  $FNR = \frac{FN}{P} = \frac{FN}{FN+TP} = \frac{0}{0+10} = \frac{0}{10} = 0$
  $FNR = 1 - TPR$
- *fall-out or false positive rate* (**FPR**)
  $FPR = \frac{FP}{N} = \frac{FP}{FP+TN} = \frac{0}{0+20} = \frac{0}{20} = 0$
  $FPR = 1 - TNR$
- *false discovery rate* (**FDR**)
  $FDR = \frac{FP}{FP+TP} = \frac{0}{0+10} = \frac{0}{10} = 0$
  $FDR = 1 - PPV$

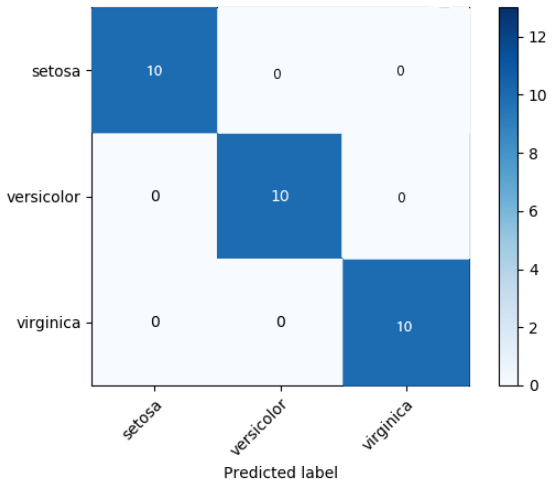

Fig. 3. Confusion matrix for classification results by the use of implemented decision tree method

- *false omission rate* (**FOR**)
  $FOR = \frac{FN}{FN+TN} = \frac{0}{0+20} = \frac{0}{20} = 0$
  $FOR = 1 - NPV$
- *accuracy* (**ACC**)
  $ACC = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{10+20}{10+20+0+0} = \frac{30}{30} = 1$
- F_1 *score* - harmonic mean of precision and sensitivity
  $F_1 = 2 \cdot \frac{PPV \cdot TPR}{PPV+TPR} = \frac{2TP}{2TP+FP+FN} = \frac{20}{20+0+0} = \frac{20}{20} = 1$
- *Matthews correlation coefficient* (**MCC**)
  $MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}} = \frac{300}{10 \cdot 10 \cdot 20 \cdot 20} = \frac{300}{200} = \frac{3}{2}$
- *informedness or Bookmaker Informedness* (**BM**)
  $BM = TPR + TNR - 1 = 1 + 1 - 1 = 1$
- *Markedness* (**MK**)
  $MK = PPV + NPV - 1 = 1 + 1 - 1 = 1$

Legend:

- **P** - condition positive - the number of real positive cases in the data,
- **N** - condition negative - the number of real negative cases in the data,
- **TP** - true positive,
- **TN** - true negative,
- **FP** - false positive,
- **FN** - false negative

We present our error matrix in the form of a table in figure 3, in which the poems correspond to the species of iris, to which the data point belonged, while the columns tell the genre to which it was qualified. The elements inside the table specify the number of data points corresponding to the species of iris specified in the row header assigned to the data group specified in the column header.

Let us note the perfect compatibility between data groups and species of iris. Each point has been correctly classified.

VI. Conclusions

Proposed reasoning was easy to implement. Results show decisions were very good and all inputs were classified correctly. The code of the reasoning in python used a library for artificial intelligence where the method was coded.

In future research i want to develop another method for data classification based on other probabilistic methods where decision between classes will be related to statistical measures.

References

[1] G. Capizzi, G. L. Sciuto, G. Cammarata, and M. Cammarata, "Thermal transients simulations of a building by a dynamic model based on thermal-electrical analogy: Evaluation and implementation issue," *Applied energy*, vol. 199, pp. 323–334, 2017.

[2] S. Brusca, G. Capizzi, G. Lo Sciuto, and G. Susi, "A new design methodology to predict wind farm energy production by means of a spiking neural network–based system," *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, vol. 32, no. 4, p. e2267, 2019.

[3] G. Capizzi, G. L. Sciuto, C. Napoli, and E. Tramontana, "Advanced and adaptive dispatch for smart grids by means of predictive models," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6684–6691, 2017.

[4] M. Woźniak, D. Połap, G. Capizzi, G. L. Sciuto, L. Kośmider, and K. Frankiewicz, "Small lung nodules detection based on local variance analysis and probabilistic neural network," *Computer methods and programs in biomedicine*, vol. 161, pp. 173–180, 2018.

[5] M. Wozniak, D. Polap, L. Kosmider, C. Napoli, and E. Tramontana, "A novel approach toward x-ray images classifier," in *2015 IEEE Symposium Series on Computational Intelligence*. IEEE, 2015, pp. 1635–1641.

[6] G. Cramer, R. Ford, and R. Hall, "Estimation of toxic hazarda decision tree approach," *Food and cosmetics toxicology*, vol. 16, no. 3, pp. 255–276, 1976.

[7] M. A. Friedl and C. E. Brodley, "Decision tree classification of land cover from remotely sensed data," *Remote sensing of environment*, vol. 61, no. 3, pp. 399–409, 1997.

[8] H. Rao, X. Shi, A. K. Rodrigue, J. Feng, Y. Xia, M. Elhoseny, X. Yuan, and L. Gu, "Feature selection based on artificial bee colony and gradient boosting decision tree," *Applied Soft Computing*, vol. 74, pp. 634–642, 2019.

[9] A. Tahmassebi, A. H. Gandomi, M. H. Schulte, A. E. Goudriaan, S. Y. Foo, and A. Meyer-Baese, "Optimized naive-bayes and decision tree approaches for fmri smoking cessation classification," *Complexity*, vol. 2018, 2018.

[10] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.