Revista Publicando, 4 No 13. (1). 2017, 563-573. ISSN 1390-9304

Comparison of Methods for classifying data for division into groups when learning English

Alfiya Rafailovna Baranova[1], Karina Irikovna Kalimullina[2]

1. **Kazan Federal University, baranova.alfiyarafailovna@mail.ru**
2. **Kazan Federal University**

ABSTRACT

The article deals with the possibility of applying methods of data analysis, in particular data classification, with the aim of dividing the students into groups when learning English. Purpose of the research is to compare methods for classifying data such as a method of k-nearest neighbors, and decision tree. It is necessary to define what the optimality criterion is and to compare models based on these methods. The result of the research detects two models, allowing the input data to divide students into groups to study the English language, and an analysis of the feasibility of their application.

Keywords: language, teaching, student, data analysis, machine learning, classification problem, k-nearest neighbors method, decision trees.

Articulo recibido: 05-11-2017
Aprobación definitiva: 04-12-2017

Revista Publicando, 4 No 13. (1). 2017, 563-573. ISSN 1390-9304

1. INTRODUCTION

The importance of learning foreign languages increases every day and the learning process has many difficulties and unsolved problems. However, the process of sorting students into groups is also can be difficult. As a rule, teachers use tests for distribution students to groups. Only tests themselves can not reliably and uniquely determine the language level of a student. In order to prevent additional difficulties in the system of testing of students' knowledge and at the same time to increase the accuracy of these studies, it is possible to use different data analysis methods. Groups which the subjects will be divided into are known in advance, and correspond to different levels of knowledge of a foreign language. This condition allows the possibility of setting the problem of data classification for division into groups to study a foreign language, in particular the English language. It is necessary to understand the term "classification" to apply the methods of solving this problem in practice. Classification is a systematic distribution of the studied objects, phenomena, processes by species, types, for any essential features for ease of study; grouping of initial concepts and arrange them in a certain order, reflecting the degree of similarity. Thus, by classification we mean the assignment of objects (dependent variables) to one of the previously known classes(Lecture 5: Data analysis problems. Classification and clustering. – URL: http://www.intuit.ru/studies/courses/6/6/lecture/166 (the date of access: 10.09.2017).). Classification refers to a supervised learning strategy, also called managed learning (Lecture 5: Data analysis problems. Classification and clustering. – URL: http://www.intuit.ru/studies/courses/6/6/lecture/166 (the date of access: 10.09.2017).). Classification can be binary, where the dependent variable can take only two values (for example, yes or no, 0 or 1), and non-binary when considering the set of classes for the dependent variable. It should be noted that we will have to deal with non-binary classification. In addition, the classification can be one-dimensional (one feature) and multidimensional (two or more features). Obviously, based on the methods of multidimensional classification, you can get much more accurate results, and their application is preferable in a situation close to real (4). An entity determines which of the predefined classes the object belongs to. It's called a classifier. To conduct

Articulo recibido: 05-11-2017
Aprobación definitiva: 04-12-2017

Revista Publicando, 4 No 13. (1). 2017, 563-573. ISSN 1390-9304

classification using mathematical methods, it is necessary to have a formal description of an object that can be operated using a mathematical classification apparatus. Each object (database record) carries information about some property of the object. The entire dataset is divided into two parts: training and test sets. The training set includes the data used to train the model. It contains both the input and output values of the examples. Output values are designed to train the model. The test set also contains input and output values. However, the output values are used to verify the correct operation of the model.

There are many classification methods, but we need to understand which ones suit the best to divide students into groups for learning English. Thus, we will have to compare several basic methods of solving the classification problem and choose the optimal one from them.

2. METHODS

[1] The comparison should begin with the analysis of the k-nearest neighbors method. This method is one of the simplest methods for solving the classification problem. However, according to the authors, the results obtained by this method can be quite reliable. It is assumed that there is already some number of objects with an exact classification (i.e., the class to which the object belongs to is known in advance). Levels of knowledge of the English language were chosen as classes, and tested different skills (e.g. listening, reading) were adopted as attributes. The k-nearest neighbors method is based on the rule that an object is considered to belong to the class to which most of its nearest neighbors belong to (**Cover T. M., Hart P. E. Nearest neighbor pattern classification / T. M. Cover, P.E. Hart, IEEE Transactions on Information Theory 13, 1967, pp. 21–27.**). “Neighbor” is an object that is close to the sample in a particular sense. Here it is advisable to determine how the degree of proximity of objects will be measured. For this a certain metric (i.e., a function of the distance) has to be introduced. Next, it is necessary to select the parameter k. In practice, it is most often assumed that k = (√N) (K-Nearest Neighbors. – URL: http://www.machinelearning.ru/wiki/index.php?title=KNN (the date of access: 10.10.2017).).

Articulo recibido: 05-11-2017
Aprobación definitiva: 04-12-2017

Revista Publicando, 4 No 13. (1). 2017, 563-573. ISSN 1390-9304

[2] Also for comparison we consider the method of decision tree for classification problem. This method consists in carrying out the process of dividing the data into groups until homogeneous (or almost homogeneous) sets are obtained (**Beginner's Guide to Decision Trees for Supervised Machine Learning. – URL: https://www.quantstart.com/articles/Beginners-Guide-to-Decision-Trees-for-Supervised-Machine-Learning (the date of access: 20.09.2017)**). A set of rules that gives such a partition will allow then to make a prediction (i.e., to determine the most likely class number) for the new data. Due to the comparative review of methods it was found out that the method of decision tree is one of the most effective. The method implements the principle of "recursive division", also called "Divide and conquer" strategy (Maze Generation: Recursive Division. – URL: http://weblog.jamisbuck.org/2011/1/12/maze-generation-recursive-division-algorithm (the date of access: 19.10.2017).). In the decision nodes starting from the root, one should select the characteristic which value is used to divide all data into 2 classes. The process continues until the stopping criterion is fulfilled. This is possible in the following situations (**Lecture 9: Methods of classification and prediction. Decision trees. – URL: http://www.intuit.ru/studies/courses/6/6/lecture/174 (the date of access: 3.10.2017)**):

1. All (or almost all) data of this decision node belongs to the same class;
2. There are no attributes on which it is possible to build a new partition;
3. The tree exceeded a predetermined "limit of growth" (if it was pre-installed).

These models can be implemented using the mathematical package R.

3. RESULTS

For building models we used 5 classes: Beginner+Elementary, Pre-Intermediate, Intermediate, Upper-Intermediate, Advanced+Proficient. The minimum and maximum levels of English proficiency were included in the nearest to them in the group. It is assumed that the number of subjects being taught that belong to these groups is relatively small.

The number of skills that could be selected as attributes is very large, so we chose only those that did not require a significant increase in the test task. In the result it was obvious that the test of speaking cannot be included in the test, and the test of writing

Articulo recibido: 05-11-2017
Aprobación definitiva: 04-12-2017

Revista Publicando, 4 No 13. (1). 2017, 563-573. ISSN 1390-9304

skills can increase the size of the data to be processed. These models were used in the following attributes: reading, listening, knowledge of grammar, vocabulary. Each of these skills can be verified using existing tests.

We used 10 questions to review each of the first two skills and 25 to test the knowledge for each other. The results obtained from the tests were converted as a percentage of the number of correctly performed tasks and divided by 100 to represent them in numbers from 0 to 1. Table 1 shows a fragment of the training sample.

Table 1. A fragment of the training sample.

Reading	Listening	Grammar	Vocabulary	Class
0,5	0,4	0,96	0,64	3
0,1	0,1	0,28	0,2	1
0,7	0,7	0,8	0,88	3
0,7	0,8	0,92	0,8	3
0,7	1	0,56	0,56	2
0,1	0,3	0,32	0,48	2
0,6	0,6	0,88	0,8	3
0,4	0,5	0,76	0,56	3
0,9	0,6	1	0,96	4
0,3	0,3	0,36	0,32	1
0,1	0,2	0,48	0,44	2
0,5	0,7	0,96	0,6	3
0,4	0,5	0,84	0,48	3
0,2	0,1	0,2	0,4	2
0,7	0,9	0,96	0,88	5
0,4	0,1	0,6	0,76	3
0,5	0,6	0,92	0,68	3
0,4	0,1	0,6	0,8	3
0,5	0,7	0,96	0,48	3
0,6	0,6	0,84	0,56	3

Articulo recibido: 05-11-2017
Aprobación definitiva: 04-12-2017

Revista Publicando, 4 No 13. (1). 2017, 563-573. ISSN 1390-9304

0,1	0,2	0,32	0,24	1
0,7	0,7	0,96	0,72	3
0,2	0,3	0,28	0,48	2
0,2	0,1	0,36	0,36	1
0,8	0,7	0,8	0,68	3

At the next stage we check the work of our model on the test sample. Parts of the results of the model for test samples are shown in Table 2.

Table 2. Results of the model for test samples.

Reading	Listening	Grammar	Vocabulary	Class
0,4	0,5	0,76	0,8	3
0,7	0,6	0,88	0,72	3
0,8	0,9	0,72	0,8	4
0,7	0,7	0,84	0,72	3
0,4	0,6	0,6	0,72	3
0,4	0,3	0,88	0,8	3
0,5	0,1	0,92	0,6	3
0,5	0,3	0,92	0,72	3
0,1	0,3	0,28	0,24	1
0,1	0,2	0,44	0,24	2
0,2	0,1	0,32	0,44	2
0,7	0,9	0,72	0,72	3
0,1	0,1	0,4	0,24	2
1	1	0,96	0,92	5
0,1	0,2	0,24	0,2	1
0,5	0,2	0,8	0,56	2
0,7	1	0,8	0,8	3
0,8	0,6	0,84	0,84	4

Articulo recibido: 05-11-2017
Aprobación definitiva: 04-12-2017

0,7	0,7	0,72	0,96	3
0,3	0,1	0,44	0,24	2
0,6	0,6	0,84	0,76	3
0,3	0,3	0,2	0,44	2
0,2	0,3	0,32	0,36	1
0,7	0,6	0,88	0,76	3
0,1	0,1	0,4	0,2	2

[3] We can estimate the accuracy of our forecast by building a cross-validation table (**CrossTable. Cross Tabulation With Tests For Factor Independence – URL: https://www.rdocumentation.org/packages/gmodels/versions/2.16.2/topics/CrossTablehttps://www.rdocumentation.org/packages/gmodels/versions/2.16.2/topics/CrossTable (the date of access: 18.09.2017)**).

Table 3. Part of the results of the decision tree.

test_data_labels		**data_test_pred**				
	1	2	3	4	5	Row Total
1	18	0	1	0	0	19
2	0	20	2	0	1	23
3	1	2	22	0	0	25
4	2	0	0	11	1	14
5	0	1	0	1	18	20
Column Total	21	230	25	12	20	100

Reading	Listening	Grammar	Vocabulary	Class
0,5	0,4	0,96	0,72	3
0,4	0,5	0,76	0,8	3
0,7	0,6	0,88	0,72	1
0,8	0,9	0,72	0,8	4

Articulo recibido: 05-11-2017
Aprobación definitiva: 04-12-2017

Revista Publicando, 4 No 13. (1). 2017, 563-573. ISSN 1390-9304

0,7	0,7	0,84	0,72	3
0,4	0,6	0,6	0,72	2
0,4	0,3	0,88	0,8	2
0,5	0,1	0,92	0,6	3
0,5	0,3	0,92	0,72	3
0,1	0,3	0,28	0,24	4
0,1	0,2	0,44	0,24	1
0,2	0,1	0,32	0,44	2
0,7	0,9	0,72	0,72	3
0,1	0,1	0,4	0,24	3
1	1	0,96	0,92	2
0,1	0,2	0,24	0,2	5
0,5	0,2	0,8	0,56	3
0,7	1	0,8	0,8	3
0,8	0,6	0,84	0,84	3
0,7	0,7	0,72	0,96	3
0,3	0,1	0,44	0,24	3
0,6	0,6	0,84	0,76	1
0,3	0,3	0,2	0,44	3
0,2	0,3	0,32	0,36	2
0,7	0,6	0,88	0,76	1

4. DISCUSSION

[4] We built a model that uses sample data with the predefined data for prediction. We can compare the tables with the results of the model for the test sample and the real data to estimate the accuracy of the model. However, the R package provides the ability not to check the results manually, but to set the accuracy of the forecast using the built-in functions. Therefore, we used the cross-validation method. We used the function CrossTable and built a table to estimate the forecast (**Starkweather J. Cross Validation techniques in R: A brief overview of some methods, packages, and**

Articulo recibido: 05-11-2017
Aprobación definitiva: 04-12-2017

Revista Publicando, 4 No 13. (1). 2017, 563-573. ISSN 1390-9304

functions for assessing prediction models. / J. Starkweather – URL: http://it.unt.edu/sites/default/files/crossvalidation1_jds_may2011.pdf (date of access: 6.10.2017)). In the lower right corner of the table we can see how many objects were used to test the model. If we present the cross-validation table as a matrix, the number of correct forecasts can be seen on the main diagonal. Thus, the values of the other columns of each row are incorrectly distributed objects. In order to know the efficiency of the model, it is sufficient to count the number of erroneous predictions. In our model, the total number of test objects is 100, and the number of errors is 12. We get a prediction accuracy of 88%. We can't exactly say that this is the optimal forecast, so we checked the dependence of factors on each other. We received that the factors are not correlated, and therefore this model can be used.

Consider the results of a model built using decision trees. The forecasts obtained in the model are also compared with real data. We also composed a cross-validation table for the analysis of the method. It showed that the method allowed 5 errors. Thus, the accuracy of the forecast is 95%.

However, we now can compare the results of decision tree method with the results of the k-nearest neighbors method. We can conclude that the efficiency of the model constructed by the decision tree method is higher than that of the k-nearest neighbors method. As we see from the table above, the decision tree method showed a better result than the k-nearest neighbors method did.

5. CONCLUSIONS

Thus, we built two models to study the problem of dividing the students into groups when learning English. Both of these models show acceptable results and can be used in practice. However, from the point of view of the accuracy of forecasts, the decision tree model showed better results. But we also understand that high precision is achieved at the expense of complicating the model, in particular the learning process. Therefore, from the point of view of simplicity, the model k-nearest neighbors is optimal.

It's also worth noting that although these models are also based on knowledge assessment through testing, the separation process is more structured, but is not more labor-intensive.

Articulo recibido: 05-11-2017
Aprobación definitiva: 04-12-2017

In the future, it is possible to add more methods for comparison and to reveal the optimality of the methods by more criteria.

6. ACKNOWLEDGEMENTS

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University..

7. REFERENCES

Beginner's Guide to Decision Trees for Supervised Machine Learning. – URL: https://www.quantstart.com/articles/Beginners-Guide-to-Decision-Trees-for-Supervised-Machine-Learning (the date of access: 20.09.2017).

Cover T. M., Hart P. E. Nearest neighbor pattern classification / T. M. Cover, P.E. Hart, IEEE Transactions on Information Theory 13, 1967, pp. 21–27.

CrossTable. Cross Tabulation With Tests For Factor Independence – URL: https://www.rdocumentation.org/packages/gmodels/versions/2.16.2/topics/CrossTablehttps://www.rdocumentation.org/packages/gmodels/versions/2.16.2/topics/CrossTable (the date of access: 18.09.2017).

Kashina O. A. Data analysis in the environment R / O. A. Kashina – URL: https://edu.kpfu.ru/course/view.php?id=833 (the date of access: 5.10.2017).

K-Nearest Neighbors. – URL: http://www.machinelearning.ru/wiki/index.php?title=KNN (the date of access: 10.10.2017).

Lecture 5: Data analysis problems. Classification and clustering. – URL: http://www.intuit.ru/studies/courses/6/6/lecture/166 (the date of access: 10.09.2017).

Lecture 9: Methods of classification and prediction. Decision trees. – URL: http://www.intuit.ru/studies/courses/6/6/lecture/174 (the date of access: 3.10.2017).

Articulo recibido: 05-11-2017
Aprobación definitiva: 04-12-2017

Maze Generation: Recursive Division. – URL: http://weblog.jamisbuck.org/2011/1/12/maze-generation-recursive-division-algorithm (the date of access: 19.10.2017).

Predictive modeling, supervised machine learning, and pattern classification. – URL: https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html (the date of access: 11.10.2017).

Starkweather J. Cross Validation techniques in R: A brief overview of some methods, packages, and functions for assessing prediction models. / J. Starkweather – URL: http://it.unt.edu/sites/default/files/crossvalidation1_jds_may2011.pdf (the date of access: 6.10.2017).

Articulo recibido: 05-11-2017
Aprobación definitiva: 04-12-2017