JOURNAL OF BALTIC SCIENCE EDUCATION

ISSN 1648-3898 /Print/
ISSN 2538-7138 /Online/

CLASSIFICATION OF THE EDUCATIONAL TEXTS STYLES WITH THE METHODS OF ARTIFICIAL INTELLIGENCE

Tereza Horáková,
Milan Houška,
Ludmila Dömeová

Abstract. *Modern educational methods emphasize the necessity to transfer knowledge instead of data or information within the educational process. Thus it is important to the educational texts supporting the educational process contain knowledge in a particular textual representation. But it is not trivial to decide whether the particular piece of text contain knowledge or not. The solution is to measure the similarity between the particular text structure and a typical structure of a knowledge-designed text. This research aims at analysing the classification ability of three commonly-used classification techniques: artificial neural networks (ANNs), classification and regression trees (CARTs) and decision trees (bigMLs) to separate texts or text fragments into two groups. The texts in the first group contain mainly data and information (common texts), the texts in the other group contain knowledge in one of the particular knowledge representations (knowledge texts). The sample of 120 text fragments was used for the analysis. The results show that the ANN techniques are significantly more able to make the right classification of the text than the CART or bigML ones, and evidence good classification abilities. Thus the ANN approach could broaden the set of methods used for evaluation of difficulty of educational texts or textbooks.*

Keywords: *artificial intelligence, classification and regression trees, educational texts, knowledge representation, knowledge unit, production rules, stylometric analysis.*

Tereza Horáková, Milan Houška, Ludmila Dömeová
Czech University of Life Sciences Prague, Czech Republic

Introduction

Educational texts are an integral part of the educational process and play an important role for the transfer of knowledge. Texts, graphs, figures, pictures or other objects integrated in the textbooks are able to naturally represent more kinds of knowledge. On the other hand, to reach the current qualitative level of the transfer of knowledge, it is not enough to design any educational text or textbook. There are the standard requirements put on an educational text of high quality, in particular:

- correctness of the content;
- free of the grammar mistakes;
- easy to read for the students;
- cover common didactic elements such as educational objectives, keywords, summary, questions for self-evaluation of the level of knowledge reached, references, etc.;
- and others.

Beside the standards, other additional criteria should be respected, such as the difficulty of the text subject to the target group of the readers (students) or the literary style of the presentation of the text. There is a standard set of measures used for the measuring the difficulty of the educational texts or textbook. The most commonly-used measures are connected with the complex difficulty rate of the text and with its two components: syntactic difficulty rate and semantic difficulty rate (Arya, Hiebert & Pearson, 2010). Auxiliary measures for these indicators (e.g. coefficient of density of scientific and factual information per noun, average number of adverbs per sentence, average number of adverbs per complex of sentences and many others) are described in Hrabí (2012) or Hűbelová (2010). Nevertheless, as the measuring the difficulty of the educational text is not a key topic of this research, the measures could not be described in detail here.

In education, the classification of literary styles of the educational texts could be researched from more points of view. Cortina-Borja & Chappas (2006) quantified the literary style of various forms of media, including the new ones (broadsheet and tabloid newspapers, technical periodicals and

television news scripts). It allowed them to investigate the richness of vocabulary exhibited in these texts under the proposition that the writing style usually varies depending on the targeted readership or audience. Graham et al. (2012) state that in literature, there is an established set of techniques that have been successfully leveraged in the statistical analysis of literary style. The most common purpose of the analysis is to answer questions of authenticity and attribution. In their work, Graham et al. (2012) suggest that the progress made and statistical techniques developed in understanding visual processing as it relates to natural scenes can serve as a useful model and inspiration for visual stylometric analysis.

This is a challenge for the stylometric analysis to identify the literary style, if an author uses an expert knowledge in its raw form (as stored in an expert system) as the source of knowledge for an educational text. In this case a specific kind of an education text is designed. See Fig. 1 to compare the designs of an educational text based on production rules as a kind of the representation of procedural knowledge (knowledge text) with a piece of the same text of no specific design (common text). From the viewpoint of the particular domain (mathematical modelling), the pieces of text (A) and (B) contain the same "message" to the reader; the difference is in design only.

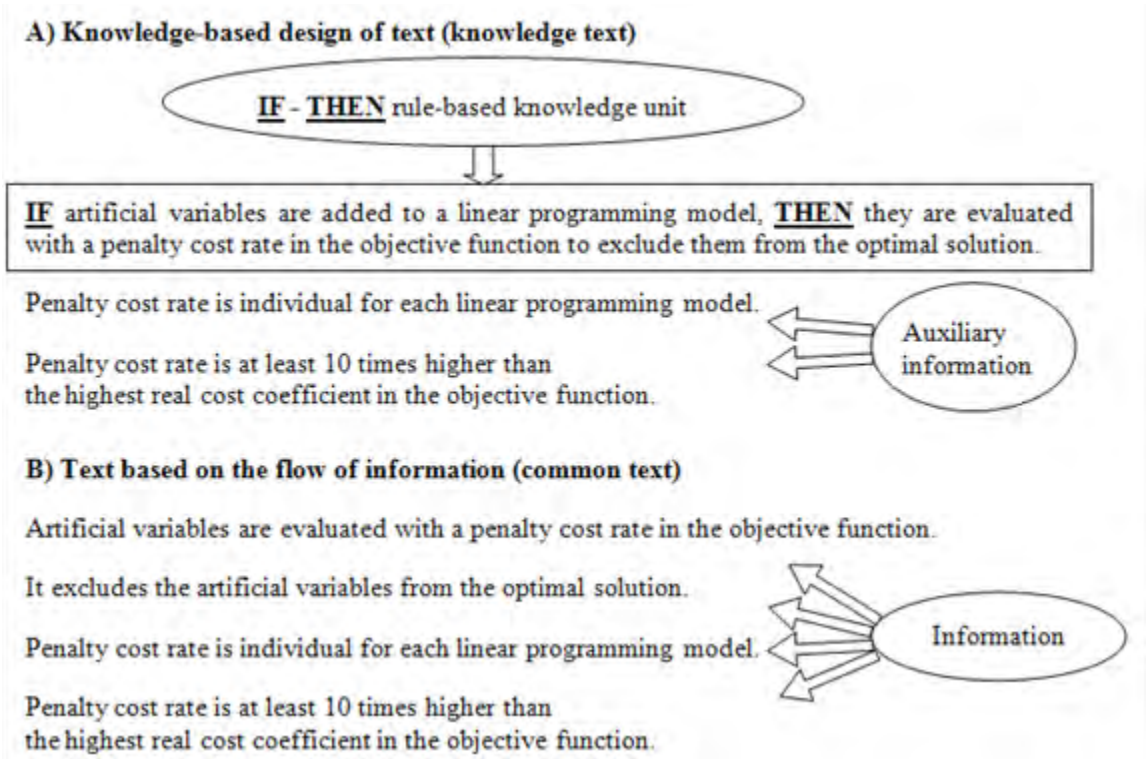

Figure 1: Educational text of a knowledge-based design and of a common style.

Source: own work, redesigned based on Dömeová et al. (2008).

A piece of the knowledge text (A) is always based on one particular production rule or knowledge unit (see Materials and Methods for further definitions), which represent the knowledge transferred to the reader (student, user, etc.). The knowledge unit could be accompanied with a subordinated auxiliary text containing additional information. The auxiliary text also improves the readability of the text to the reader, because it is extremely difficult to read a text made from the knowledge units only. On the contrary, the text with no intentional orientation to transfer knowledge (B) is made as a sequence of individual information.

In thinking about the utilization of knowledge texts in education, the next step should be to measure the efficiency of the knowledge transfer through such designed texts and compare it with the efficiency reached with the common texts. It is not simple to write a complete educational text on a particular topic or even a complete textbook using the knowledge-based style; only significant increase of the efficiency of the transfer of knowledge makes the knowledge style of the texts worth using.

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

The experiment on measuring the efficiency of knowledge using the knowledge text showed how thin the line between a knowledge text and a common text is. Moreover, the line is not crisp. The majority of real texts will be classified as "rather knowledge text" or "rather common text" instead of "surely knowledge text" or "surely common text". That is why it is necessary to have a classification algorithm of a good performance available to distinguish among the text styles.

However, a new problem reveals within the task on the classification of large amount of educational texts has been solved: it is necessary to decide, which of the methods is the most suitable for the classification of the educational texts. When processing large volumes of text, techniques of artificial intelligence are used to classify objects. Classes can be distinguished according to their topic but also to other characteristics of text; for example, common e-mails and spam messages are distinguished (Mohammad & Zitar, 2011; Puniškis, Laurutis & Dirmeikis, 2006). Tools for text classification include decision trees, linear discriminant analysis, regression analysis, artificial neural networks (Puniškis et al., 2006) and definitely a range of others, including various mutual combinations. An important new direction of research is represented by the pattern recognition which deals with studies and proposals of systems capable of recognizing typical structures in texts and data (Mohammad & Zitar, 2011).

This research is aimed at to determine the ability of three widely-used classification techniques: artificial neural networks (ANNs), classification and regression trees (CARTs) and decision trees (bigMLs) to distinguish the educational texts of a common style and educational text of a specific design using the formal representations of knowledge such as production rules or knowledge units (knowledge texts). For this purpose the same data (fragments of knowledge texts and common texts) that were used by Horáková & Houška (2014b) are analysed and, using the methods of artificial intelligence (artificial neural networks and classification regression trees), the aim is to find models capable of classifying the knowledge texts based on characteristic concepts and to distinguish them from common educational texts as well as to estimate new results for arbitrary unknown texts, i.e. whether they are of knowledge character (knowledge texts), or of information character (common text), based on experience (training set of data).

Methodology of Research

Classification Methods and Strategies

In general, AI is defined as the intelligence of an artificial entity (computer) used to solve complex problems (Ludger, 2009). An important mission of AI is to teach the machine how to complete tasks that normally require human intelligence (Kumar & Thakur, 2012). According to Mařík et al. (2004) AI is a wider topic than machine learning. Three basic methods of AI are related to knowledge processing: knowledge representation, knowledge obtaining, and derivation including retrieval (Kumar & Thakur, 2012). According to Shukla & Vilay (2013) AI systems can be taught new concepts and they can draw conclusions. New discoveries in the field of artificial intelligence can be used to improve existing methods of teaching and the introduction of a transdisciplinary approach (Flogie & Aberšek, 2015). Sklenák, Berka, Rauch, Stross & Svátek (2001) states that from the point of view of AI methods, distributed artificial intelligence and machine learning is in the foreground of concern of experts; this is also proven by numerous current publications (e.g. Tu et al., 2012, Navarro Silvera et al., 2014, Tayyebi & Pijanowski, 2014). For example, machine learning and its methods are used in the field of data mining techniques where it is necessary to find certain dependencies, patterns and trends based on characteristics of data stored in databases, e.g. in Nauman & Thompson (2014), Moradi et al. (2013), Tayyebi & Pijanowski (2014). Model induction, i.e. training the classifier using patterns, is typical for machine learning (Hüllermeier, 2008).

CARTs and ANNs are included in the software Statistica as data mining tools. Text-mining and ancillary data mining techniques are very often used to analyse text documents (Lin, Hsieh & Chuang, 2009). There are even some pedagogical experiments, where text-mining techniques are used as an analytical tool (Horáková & Rydval, 2015; Poole, 2016). Text mining techniques are also used for example for mining sentiments of teaching evaluation (Leong, Lee & Mak, 2012), to characterize patterns in teachers' narratives and value-added patterns (Cavicchiolo, Alivernini & Manganelli, 2015), or for representations of study and students' academic motivation (Alivernini, Cavicchiolo, Palmerie & Girelli, 2015). Text and data mining techniques can be used as a tool for style identity, which is an important factor for textbook analysis. Prediction of the text style based on the occurrence of words (terms or concepts) is commonly used in stylometry (Kestemont et al., 2016). Even some authors have created an automatic software that with some degree of accuracy classifies whether the text or textbook was created by a specific author style (Tang & Cao, 2015).

Multi-layer Perceptrons networks (MLPs) are highly nonlinear tools that are usually trained using iterative techniques. The most recommended techniques for training neural networks are the BFGS (Broyden-Fletcher-Goldfarb-Shanno) and Scaled Conjugate Gradient algorithms. These methods perform significantly better than the more traditional algorithms such as Gradient Descent but they are, generally speaking, more memory intensive and computationally demanding. Nonetheless, these techniques may require a smaller number of iterations to train a neural network given their fast convergence rate and more intelligent search criterion (Bishop, 2005). The methods used to train Radial Basis Function (RBF) networks is fundamentally different from those employed for MLPs. This mainly is due to the nature of the RBF networks with their hidden neurons (basis functions) forming a Gaussian mixture model that estimates the probability density of the input data (Bishop, 2005). For RBF with linear activation functions, the training process involves two stages. In the first stage, the location and radial spread of the basis functions using the input data (no targets are considered at this stage) were fixed. In the second stage, the weights connecting the radial functions to the output neurons were fixed. Thus, it is exact and does not require an iterative process.

The set of neuron activation functions for the hidden and output neurons available in *STATISTICA Automatic Neural Networks* is consist of the Identity function, Logistic sigmoid function, Softmax function and Exponential function and other functions. According to Bishop (2005), these five mentioned functions are more suitable for classification problem.

Bishop (2005) states that the error function is used to evaluate the performance of a neural network during training. The sum-of-squares error function is primarily used for regression analysis but it can also be used in classification tasks. Nonetheless, a true neural network classifier must have an error function other than sum-of-squares (Sum of Sq.), namely Cross Entropy Error Function (Entropy) (Jiřina, 2003).

Data, Information and Knowledge

For the purposes of this research the differences among data, information and knowledge are understood from the technical point of view, as defined in computer sciences, knowledge engineering or systems engineering.

Classical computer science uses the term data to denominate numbers, texts, sounds, pictures or other sense perceptions represented in a form understandable for a computer (Sklenák et al., 2001). Data is raw. It simply exists and has no significance beyond its existence (in and of itself). It can exist in any form, without relation to usability. It does not have meaning of itself. In computer parlance, a spreadsheet generally starts out by holding data (Ackoff, 1989).

Information is represented by data and their meaning. It depends on aggregation of data into context. According to Choo (2001), the observer makes sense of noticed data through a process of cognitive structuring which assigns meaning and significance to perceived facts and messages. What meanings are constructed depends on the schemas and mental models of the actor. Information is data that has been given meaning by way of relational connection (Ackoff, 1989). This„meaning" can be useful, but does not have to be. In computer parlance, a relational database makes information from the data stored within it. Information is a flow of messages. The patterns and relationship in the data is pointed out and discussed. The data is made informative and must be put into a context and linked like data.

Due to the absence of a widely-accepted definition of knowledge, an object-oriented approach to knowledge as characterized in Knowledge Engineering is used. Knowledge Engineering perceives knowledge as an object that can be identified, stored and eventually passed on for further use. Knowledge can be represented in several ways. Usually, production rules, decision tables, semantic networks and frames are used for this purpose (Mařík et al., 2004). Thanks to the methods and representations of knowledge in knowledge engineering, artificial intelligence (AI) has changed the view of information and it puts knowledge as a form of generalization and abstraction above it (Sklenák et al., 2001). Knowledge discovery, which is defined according to Dubois & Prade (1996) as a non-trivial process of discovering valid, new, understandable and possibly useful knowledge in texts, can be considered a new independent scientific branch (Hüllermeier, 2008).

In accordance with this approach, the model of a knowledge unit (Dömeová et al., 2008) as a possible representation of knowledge has been created as well. Knowledge unit is the basic component of so-called knowledge texts (Houška & Rauchová, 2013) designated for education. Knowledge texts have already been tested in the education process as a tool for transferring the knowledge as compared to common educational texts (Rauchová & Houška, 2013a, Horáková & Houška, 2014a, Horáková et. al., 2014)), their efficiency of creation has been measured (Rauchová

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

& Houška, 2013b), and the quantitative profile of knowledge text and its difference from text that is common from the linguistic point of view has been determined statistically, too (Rauchová et al. 2014). The research follows the results of Horáková & Houška (2014b) where, among others, the hypothesis of the difference of the number of characteristic concepts of knowledge texts as compared to common texts was statistically proven; knowledge text has shown significantly higher rate of these concepts (terms) which, as stated by Horáková & Houška (2014b), is determined by the structure of the knowledge unit (see the following chapter for further details).

Knowledge Unit and Knowledge Text

In this research, "knowledge text" is understood as a specific form of text that expresses knowledge explicitly. Based on the research of Dömeová et al. (2008), it was determined that production rules and their extended versions, i.e. knowledge units, constitute a representation of knowledge in text of a very good suitability. Dömeová et al. (2008) express the knowledge unit formally as:

$$KU=\{X,Y,Z,Q\}$$

Where *X* represents a problem situation, *Y* represents an elementary problem that will be solved within the problem situation *X*, *Z* represents the goal of solving the elementary problem, *Q* represents the solution of the elementary problem *Y*.

Then, knowledge unit can be of the following form (Dömeová et al., 2008): "When it is necessary to solve problem *Y* within the problem situation *X* with the goal *Z*, then apply the solution *Q*."

Furthermore, knowledge texts and common texts as terms are distinguished in this research. Houška & Rauchová (2013) proposed the methodology for creating knowledge text from common text, where knowledge text can also contain additional information on top of knowledge units and standard production rules. The examples of the educational texts on the industrial waste processing (see Enviregion, 2014, in Czech, translation by the authors) in knowledge-based design and in the common style follow.

> *"The waste arisen from industry production differs in comparison with the one arisen from households in more properties. It differs in the composition influenced with the kind of the production. It can often contain elements, which are of the hazardous character for people as well as for the nature (toxic, explosive, flammable, etc.). That is the reason for special manipulation for such waste. Individual productions generate waste of different properties and thus there is no unique procedure for processing it. Waste from the chemical productions is often really dangerous and has to be modified before processing. Metallurgy also produces a large amount of dangerous waste. Food productions generate waste that could be transformed into a fertilizer and used in agriculture. Building industry can often recycle the waste in order to be re-used for the production of building materials or for building the houses."* (Enviregion, 2014)

Its knowledge form (the original text modified by the authors) can be presented as follows:

> *"If we consider the waste arisen from industry production and describe its properties, then it differs from the households one in more characteristics influenced with the source of the waste. If it contains elements denoted as hazardous for people or nature (toxic, explosive, flammable, etc.), then we should manipulate with the waste carefully. When we consider the industrial waste and describe its processing, we should bear in mind that each production generates a different kind of the waste, and thus there is no unique way of processing the waste. If dangerous waste is processed, the manipulation procedure should be described in detail in order to prevent the consequences to the environment, e.g. using the modification of the waste from chemical production aimed at the reduction of the content of the toxic metals, such as cadmium, nickel, lead, etc. When we deal with the waste processing and aim at exploiting the maximum value obtained from the waste, then we can e.g. transform the food production waste into fertilizers, building production waste into building material, etc."* (Enviregion, 2014).

The complete research sample (all pairs of normal and knowledge texts in Czech) is available at: *http://pef.czu.cz/~houska/Agris_2014/Sample.pdf*.

It was discovered that not all texts can be rewritten in the knowledge form. Knowledge texts have been tested in the education process as a possible tool for transferring the knowledge as compared to common educational texts (Rauchová & Houška 2013a, Horáková & Houška 2014a), their efficiency of creation has been measured (Rau-

chová & Houška 2013b), and the quantitative profile of knowledge text and its difference from text that is common from the linguistic point of view (syntactic and semantic analysis of texts according to Hůbelová (2010)) has been determined statistically, too in Rauchová et al. (2014b).

The research follows the results of Horáková & Houška (2014b) where, among others, the hypothesis of the difference of the number of characteristic concepts of knowledge texts as compared to common texts was statistically proven.

Authors have included terms that cover all parts of a production rule and knowledge unit, i.e. subordinating and coordinating (consequential and causal) conjunctives as well as particular terms (in Czech language - *když, pokud, kdyby, protože, poněvadž, jelikož, jestliže, -li, přestože, ačkoli, třebaže, i když, ač, že, aby, jakmile, až, než, nežli, zatímco, proto, a proto, a tak, tudíž, a tudíž, tedy, a tedy, neboť, vždyť, totiž, neb, je-li potřeba, zapotřebí, v rámci, hodlat, mít v úmyslu, za účelem, z důvodu, jednalo by se, má-li se, s cílem, je třeba, má-li být*) among the observed concepts. Knowledge text contains significantly higher number of these concepts (terms) from the statistics point view, as compared to common text (Horáková & Houška 2014b).

Data Sample and Procedure of the Research

The procedure of the presented research has consisted of 5 following steps (Figure 2).

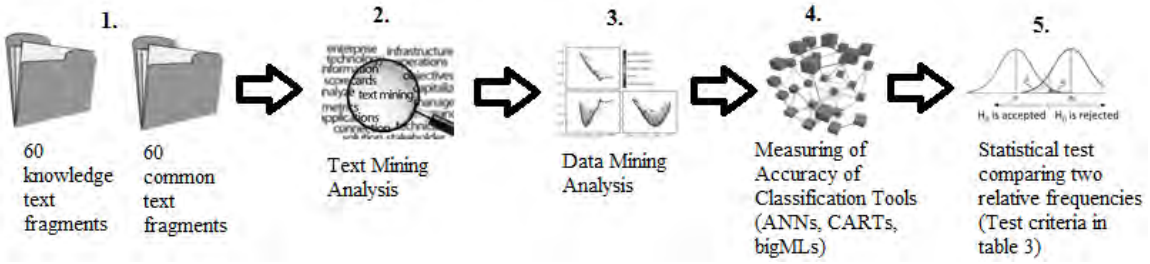

Figure 2: Scheme of the procedure of the presented research.

The first step was focused on preparation in total 120 text fragments written in the Czech language (available at *http://pef.czu.cz/~houska/Agris_2014/Sample.pdf*), where half of the fragments were common texts and the other half were knowledge texts. The average extent of text fragments was 205 words, the extent of the smallest fragment was 161 words and the extent of the longest one was 311 words.

Then in the second step text fragments were analysed by text mining tools in software STATISTICA 12 (StatSoft) program using its STATISTICA Text Miner module. Term document matrix (Srivastava & Sahami, 2009) was created for selected word concepts (according to Horáková & Houška, 2014b). The rows of the matrix represent particular cases (text fragments) and the columns represent the absolute frequencies of occurrence of observed selected concepts (terms), see above in a chapter *Knowledge Unit and Knowledge Text*. The term document matrix of dimensions [*n***m*] consisted of *n* = 120 rows and *m* = 25 columns. The whole term-document matrix is available at http://pef.czu.cz/~houska/TT_2015/Freq_Matrix_1.xls.

As the third step the models for the classification of text fragments were created in the STATISTICA Data Miner module; in particular, these were ANNs and CARTs. For the sakes of comparison, classification decision trees were also created using the free version of the Internet "machine-learning" product named BigML (available at *https://bigml.com*). Jiřina (2003) states that working with artificial neural networks in the Statistica program requires to divide the data into three sets: training one, testing one, and validating one. Bishop (2005) says that typically, the ratio of this division is 50:25:25 or 70:15:15, however it is recommended to choose the ratio that gives more balanced results. That is why for this research, the ratio 70:15:15 was selected because this ratio gives more balanced results in comparison with the ratio 50:25:25. In the resulting software reports, the performance is observed for each of these sets; the models without considerable fluctuations in performance for particular sets were selected.

The fourth step has consisted of the prediction, i.e. classification of hitherto unknown cases, 30 text fragments (1/4 of the original range of training and testing data) were selected: 15 text fragments are of the knowledge type and 15 of the common type. The complete term-document matrix of the sample for prediction is available at:

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

http://pef.czu.cz/~houska/TT_2015/Freq_Matrix_2.xls. The accuracy of prediction of each selected classifiers has been observed (Figure 3).

The last fifth step of the procedure was a statistical analysis, i.e. the definition of statistical hypotheses and statistical testing by selected parametric test (see next two following sub-chapters). The aim of this step is to compare which artificial algorithm provides statistically higher accuracy of classification performance.

Statistical Hypotheses

In accordance with design of the research the following research hypotheses have been defined: H_0: "*The selected artificial neural networks provide the same relative frequencies of correct classification of knowledge texts or common educational texts as the selected classification trees and regression trees.*"

Twenty hypotheses for ANN vs. CART have been tested. *H_0(ij): "ANN(i) achieves the same relative frequency of correct classification into the class of knowledge texts or standard educational texts as CART(j).", i {1,2,3,4,5}, j {1,2,3,4}.*

Where *i* denotes the sequence number of the ANN, *j* denotes the sequence number of the CART.

Twenty hypotheses for ANN vs. bigML have been tested. *H0(im): "ANN(i) achieves the same relative frequency of correct classification into the class of knowledge texts or standard educational texts as bigML(m).", i {1,2,3,4,5}, m {1,2,3,4}.*

Where *i* denotes the sequence number of the ANN, *m* denotes the sequence number of the bigML.

Data Analysis

Parametric two-sample test of the hypothesis of two matching relative frequencies is used to test the formulated research hypothesis *H0*, i.e. whether there is no difference in the error rate of selected classification tools of artificial neural networks and of classification and regression trees. Test criterion *U* is calculated using the following formula:

$$U = \frac{p_1 - p_2}{\sqrt{\left(\frac{\overline{p}(1-\overline{p})}{n_1} + \frac{\overline{p}(1-\overline{p})}{n_2}\right)}}.$$

Where p_1 is the relative frequency of the observed attribute in the first set, p_2 is the relative frequency of the observed attribute in the second set, n_1 is the range of the first set, n_2 is the range of the second set; the parameter is calculated using the following formula:

$$\overline{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}.$$

The calculated test criterion *U* is compared to the tabulated value of standardised normal distribution (*Uα*) for the selected significance level (α = 0.05), $U_{0.05} = -1.645$. Parametric test based on normal distribution was selected according to results of Shapiro-Wilk W test (Lindsey, 2009) (data set comes from normal distribution), In case of $|U|>|U\alpha|$), the zero hypothesis is rejected ($H_0: \pi_1 = \pi_2$, where π_1 is the relative frequency of the observed attribute in the first set and π_2 is the relative frequency of the observed attribute in the second set) in favor of the alternative hypothesis ($H_A: \pi_1 \neq \pi_2$) (Lindsey, 2009).

Results of Research

Five artificial neural networks showing balanced performance for each data set (training, testing and validating) were selected for classification. 4 of the networks are of the MLP (Multilayer Perceptron Network) type, 1 network is of the RBF (Radial Basis Function) type. Detailed characteristics of particular networks, i.e. including the type of the error function, type of activation of a hidden layer, output activation function, number of hidden layers (indicated as the middle number in the name of the artificial neural network), of training algorithms and

of the characteristics of performance are shown in Table 1. Table 2 shows a report for the networks characterized above in Table 1, it shows in detail the accuracy of their ability to assign text to two selected classes - knowledge texts and common texts (note that these are also marked as "normal texts" in the report).

Table 1. Characteristics of artificial neural networks.

Number of artificial neural network	Name of artificial neural network	Performance of training	Performance of testing	Performance of validating	Training algorithm	Error function	Activation of a hidden layer	Output activation function
ANN 1	MLP 49-19-2	100.00	94.44	100.00	BFGS	Entropy	Logistic	Softmax
ANN 2	RBF 49-22-2	82.14	72.22	94.44	RBF	Entropy	Gaussian	Softmax
ANN 3	MLP 49-13-2	97.62	83.33	100.00	BFGS	Sum of sq.	Identity	Exponential
ANN 4	MLP 49-10-2	100.00	88.89	94.44	BFGS	Sum of sq.	Logistic	Exponential
ANN 5	MLP 49-16-2	97.62	77.78	94.44	BFGS	Sum of sq.	Identity	Exponential

Source: own work, processed in the STATISTICA 12 (StatSoft) program

Table 2. Ability of classification of particular neural networks (report).

Artificial Neural Network	Classification	Text type - Normal	Text Type - Knowledge
MLP 49-19-2	Total	60	60
	Correct	60	59
	Incorrect	0	1
RBF 49-22-2	Total	60	60
	Correct	48	51
	Incorrect	12	9
MLP 49-13-2	Total	60	60
	Correct	59	56
	Incorrect	1	4
MLP 49-10-2	Total	60	60
	Correct	59	58
	Incorrect	1	2
MLP 49-16-2	Total	60	60
	Correct	59	54
	Incorrect	1	6

Source: own work, processed in the STATISTICA 12 (StatSoft) program

According to Razi & Athappilly (2005), four classification binary trees of the CART type were selected for comparison as examples of the most used classification tools. The trees differ in the number of nodes (terminal and non-terminal ones). Each decision tree has different importance (which is not the same as the statistical significance) of selected terms that serve as the basis for dividing the nodes, i.e. classification. The graphical illustration of the importance of particular terms for particular trees (CART 1–4) is available at: *http://pef.czu.cz/~houska/TT_2015/Grafic_Summary_CARTs.doc.;* the vertical axis shows the importance (scaled from 0 to 1 where 1 corresponds to the most important division variable), the horizontal axis shows the division variables, i.e. terms (concepts) that are used for the classification into two classes (knowledge text, common text). The illustration of importance of the selected discrimination variables (selected Czech terms) for dependent variable *text type* for each CART, i.e. CART 1, CART 2, CART 3, and CART 4 is available as well at: *http://pef.czu.cz/~houska/TT_2015/Grafic_Summary_CARTs.doc.* Also at this mentioned websites is showed reports that contain the wide structures of particular classification trees. For nodes that are not being divided any more (terminal nodes), not all items are filled in the respective row. However,

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

the structure of a classification tree is also obvious in the traditional graphical representation of trees. Some illustration (mentioned above) shows the traditional graphical representation for the selected decision tree (CART 3).

Furthermore, 4 models (bigML1, bigML2, bigML3, bigML4) of decision trees were created for the sakes of comparison in the free version of the bigML machine-learning system. As the reports are quite extensive, detailed structure of particular models can be found at: *http://pef.czu.cz/~houska/TT_2015/Models_bigML.xls*. There is also available the summary importance of particular division variables for all 4 models in total. It is obvious that the variable "tak" has the highest influence. The graphical summary of importance for all 4 models and structure of four models with the focus on the fourth model and its outside branch on the right side is available at: *http://pef.czu.cz/~houska/TT_2015/Grafic_Summary_CARTs.doc*.

Particular artificial intelligence classifiers (5 artificial neural networks, 4 CARTs and 4 bigML decision trees) were presented with hitherto unknown text fragments (with the range of n = 30); their brief characteristics can be found in chapter Material and Methods, section Data. Prediction was carried out based on observed parameters (term-document matrix) and the correctness of assigning the text (classification) to the corresponding class (common text, knowledge text). Particular outputs of prediction for all classifiers can be found at: *http://pef.czu.cz/~houska/TT_2015/Prediction.xls*. Relative frequencies of correct classification of text fragments into groups are obvious in the outputs, see Figure 3.

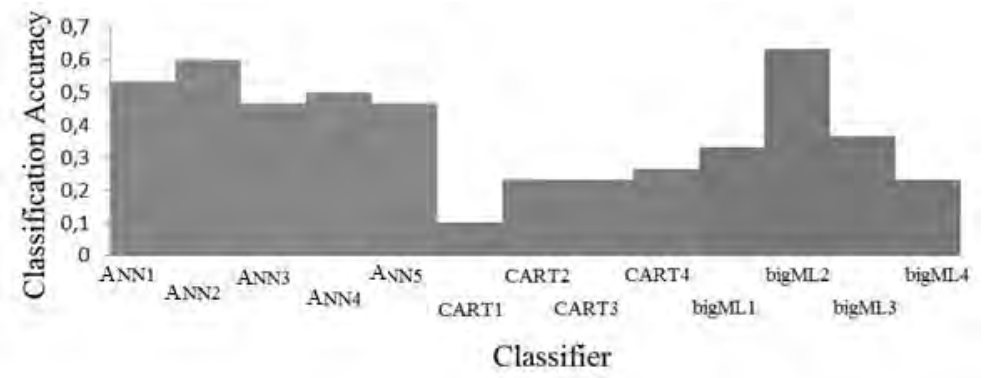

Figure 3: Graphical representation of the accuracy of prediction of particular artificial intelligence tools.

Three artificial neural networks classified correctly one half or more of unknown text fragments, i.e. ANN 1 (16 correct, 14 incorrect), ANN 2 (18 correct, 12 incorrect), ANN 4 (15 correct, 15 incorrect); 2 artificial neural networks succeeded for fewer than one half of the text fragments, i.e. ANN 3 (14 correct, 16 incorrect), ANN 5 (14 correct, 16 incorrect). All classification and regression trees that were used classified fewer than one half of unknown text fragments, i.e. CART 1 (3 correct, 27 incorrect), CART 2 (7 correct, 23 incorrect), CART 3 (7 correct, 23 incorrect), CART 4 (8 correct, 22 incorrect). 1 model of bigML decision trees (i.e. bigML2, 19 correct, 11 incorrect) classified correctly more than one half of unknown documents; the other 3 succeeded for fewer than one half of unknown documents (i.e. bigML1: 10 correct, 20 incorrect; bigML3: 11 correct, 19 incorrect; bigML4: 7 correct, 23 incorrect). Test criteria for 40 null hypotheses are shown in Table 3.

Table 3. Test criteria for 40 sub-hypotheses.

U	CART 1	CART 2	CART 3	CART 4	bigML1	bigML2	bigML3	bigML4
ANN 1	3.607865	2.389755	2.389755	2.108185	1.56315	–0.78558	1.297498	2.389755
ANN 2	4.05999	2.880476	2.880476	2.605251	2.070197	–0.26553	1.808397	2.880476
ANN 3	3.151444	1.894662	1.894662	1.607399	1.054093	–1.2975	0.785584	1.894662
ANN 4	3.380617	2.143199	2.143199	1.858698	1.309307	–1.0421	1.0421	2.143199
ANN 5	3.151444	1.894662	1.894662	1.607399	1.054093	–1.2975	0.785584	1.894662

Source: own work, processed in the MS Office Excel 2013 program

Test criteria marked **bold** ($|U|<|Ua|$) correspond to the combinations of classifiers that result in matching frequencies for the correctness of classification; test criteria marked black ($|U|>|Ua|$) correspond to the combinations of classifiers that result in different frequencies for the correctness of classification; e.g. the test criterion *U* when comparing ANN 1 to CART 1 is 3.607865 which is higher than the tabulated value for the selected significance level alpha (0.05) thus we reject the zero sub-hypothesis in favour of its alternative, i.e. ANN 1 does not achieve the same relative frequency of correct classification into the class of knowledge texts or standard educational texts as CART 1.

Discussion

The results show that except one combination (CART 4 vs. ANN 5), artificial neural networks (ANN 1, ANN 2, ANN 3, ANN 4, and ANN 5) provide more accurate prediction in comparison to the classification and regression trees (CARTs: CART 1, CART 2, CART 3, and CART 4). This is in accordance with the statements by Jiřina (2003). Comparison of artificial neural networks, CARTs and classical regression was also carried out e.g. by Razi & Athappilly (2005). According to them, the highest MAPE (mean absolute percentage error) error rate corresponds to the use of regression, then CARTs, and the lowest rate corresponds to artificial neural networks; however, when mutual tests of correct classifications using artificial neural networks and CARTS were carried out, the authors did not confirm a statistically significant difference which also became evident in the results included above when comparing some artificial artificial neural networks and CARTS from the bigML system. Lin et al. (2010) examined 5 modeling techniques: Artificial Neural Networks (ANNs), Radial Basis Function (RBF), Support Vector Regression (SVR) and Multivariate Adaptive Regression Splines (MARS). Out of these, SVR was evaluated as the best one; on the other hand, it requires more input data than other methods. Significant factors that can influence the accuracy of classification of artificial neural networks include the size of samples of the training set, testing set and validating set as well as the size of samples for prediction. For example, Rajan et al. (2009) or Mohammad & Zitar (2011) managed to find ANNs with better prediction abilities quite thanks to samples of a wider range. Tayyebi & Pijanowski (2014) found that ANNs are more accurate but their interpretability is worse. Satisfactory results can be obtained by using genetic algorithms combined with systems based on immunologic reactions of living organisms (Mohammad & Zitar, 20011). Puniškis et al. (2006) recommends a combination of multiple methods as well. Other factors include types of used algorithms or specifics of the solved task.

The weak point of a classification task for dividing knowledge texts and common texts is that even in spite of designating the text as the knowledge one, it can contain, according to the methodology of Houška & Rauchová (2013), not only production rules and knowledge units but also additional, particularizing pieces of information without knowledge potential; this can result in certain violation for classifiers within the scope of sought-after relationships and connections among particular variables.

On the other hand, the classification of the educational texts according to their style (knowledge text vs. common text) can contribute to the development of Content Knowledge (CK) of textbooks and educational texts in connection with a model of TRACK (Technological Pedagogical Content Knowledge) developed by Mishra & Koehler (2006). Moreover, classification of the texts can improve the methods for measuring the TRACK model (Archambault & Barnett, 2010).

Therefore, the future goal will be to expand the set of samples, to find new artificial neural networks capable of correct classification, and to extend the work with artificial intelligence methods by including fuzzy logic as e.g. Huang et al. (2014) or Chen et al. (2010) did, in order to be able to determine the rate of affiliation of the text being "more of the knowledge type", or "more of the common type". This could lead to the advantages not only within the scope of the textbook analyses and textbook creation for schools and universities but also for the preparation of the text materials within professional or vocational education and training.

Three groups of beneficiaries can profit from the results of the analysis: the readers of the text (students, users, etc.), the authors of the educational texts, and analysts/evaluators of the textbooks. As found out in the previous research, the readers are unable to perceive the differences among the text styles in general. The response from the readers was equivalent independently on the methods of the measurement; neither objective measurement using the fMRI approach Horáková & Houška (2016), nor subjective feedback through a questionnaire survey (Rauchová & Houška, 2013a). On the other hand, intentional structure of the knowledge text allows to motivate the reader to pay the attention to the most important statements in the text, as shown by Reichelt et al. (2014). The authors can apply this approach to be sure that they are really codifying knowledge instead of pure information or even data only (Dömeová et al., 2008). Moreover, when a production rules based knowledge base from an expert system is

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

taken as a source for developing an educational text, the author receives the feedback, whether he/she incorporates the knowledge from the expert system successfully into his/her text. The above-analyzed algorithms could also serve to the text analysts/evaluators of the textbook as another analytical approach (beside text difficulty and readability analyses (Hrabí, 2012; Hűbelová, 2010)) providing additional measures.

Finally, Franzolin & Bizzo (2015) define laxity as a "tendency to alter scientific knowledge, when presenting it in textbook form, to make it accessible to learners. The rigorism is the opposite of laxity, i.e., the tendency to approximate knowledge taught to scientific knowledge, seeking to transmit it correctly, with a commitment to scientific principles". Thus, if we are able to identify whether the text is according to knowledge or common text style, we will be able in looking at the knowledge taught in schools to find a balance between the "laxity" and "rigorism" according to the level of education.

Conclusions

The ability of three widely-used classification techniques: artificial neural networks (ANNs), classification and regression trees (CARTs) and decision trees (bigMLs) to distinguish the knowledge-based style and the common style of the educational texts was measured and discussed in this paper. Different methods of artificial intelligence can provide satisfactory results based on the particular purpose of use, on the possibilities of obtaining the input parameters, or on the volume of training databases as well as the target ones.

120 text fragments from textbooks have been tested; 60 text fragments were written in a knowledge style, 60 of them were written in a common style. Most of the ANNs provided higher success rate of classification in comparison with the CARTs; on the other hand, most of the bigML decision trees achieved the same success rate as the ANNs.

These finding are useful for knowledge content analysis in many didactical knowledge models, in particular for textbook creation and textbook analysis and for decision about content of text which could be in/excluded into/from textbook.

The further research will focus on the applications of other artificial intelligence algorithms or combinations of multiple AI methods. The use of deeply-elaborated algorithms serving for the classification of images or for spam detection is also worth considering. Replacing a bivalent classification (1 … common text, 0 … knowledge text, or vice versa) with a discrete or a continuous scale could also extend the set of methods used for the evaluation of text difficulty.

As it has been already mentioned, measures and indicators of semantic or syntactic difficulty of educational texts or textbooks are well known. The approach presented in this article can raise a measure on knowledge content of the text, e.g. a number of pieces of knowledge per paragraph, or other. Extending the set of training texts would also contribute to the improvement of accuracy which can generally contribute to textbook analysis methodology. Moreover, if we are able distinguish texts according to their style (with explicit or implicit expressed knowledge) we can better find the balance between laxity and rigorism of knowledge taught in schools.

Acknowledgements

The research is supported by the grant project of the Internal Grant Agency of the FEM CULS Prague "Determining the neuropsychological characteristics of learning for different kinds of educational texts using neurotechnologies", No. 20151047.

References

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis, 16* (1), 3-9.

Alivernini, F., Cavicchiolo, E., Palmerio, L., & Girelli, L. (2015). Representations of study and students' academic motivation. *Procedia Social and Behavioral Sciences*, *205*, 302-305. doi: 10.1016/j.sbspro.2015.09.086.

Archambault, L. M., & Barnett, J. H. (2010). Revisiting technological pedagogical content knowledge: Exploring the TPACK framework. *Computers & Education*, *55* (4), 1656-1662. doi: 10.1016/j.compedu.2010.07.009.

Arya, D., Hiebert, E., & Pearson, P. (2011). The effects of syntactic and lexical complexity on the comprehension of elementary science texts. *International Electronic Journal of Elementary Education*, *4* (1), 107-125.

Bishop, C. M. (2005). *Neural networks for pattern recognition*. Oxford: Oxford university press.

Cavicchiolo, E., Alivernini, F., & Manganelli, S. (2015). A mixed method study on teachers' diaries: Teachers' narratives and value-added patterns. *Procedia Social and Behavioral Sciences*, *205*, 485-492. doi: 10.1016/j.sbspro.2015.09.048.

Chen, F., Chen, Y., & Kuo, J. (2010). Applying moving back-propagation neural network and moving fuzzy neuron network to predict the requirement of critical spare parts. *Expert Systems with Applications*, *37*, 4358-4367.

Choo, Ch. W. (2001). The knowing organization as learning organization. *Education + Training, 43* (4/5), 197-205. doi: 10.1108/EUM0000000005482.

Cortina-Borja, M., & Chappas, C. (2206). A stylometric analysis of newspapers, periodicals and news scripts. *Journal of Quantitative Linguistics*, *13* (2-3), 285-312. doi: 10.1080/09296170600892538.

Dömeová, L., Houška, M., & Beránková Houšková, M. (2008). *Systems approach to knowledge modelling*. Hradec Králové: Graphical Studio Olga Čermáková.

Dubois, D., & Prade, H. (1996). What are fuzzy rules and how to use them, *84*, 169-185. In Hüllermeier, E. 2008. Fuzzy sets in machine learning and data mining. *Applied Soft Computing Journal*, *11*, 1493–1505. doi: 10.1016/j.asoc.2008.01.004.

Enviregion (2014). *Textbook on environmental education*. Retrieved: 15/5/2014, from http://ucebnice3.enviregion.cz/

Flogie, A., & Aberšek, B. (2015). Transdisciplinary approach of science, technology, engineering and mathematics education. *Journal of Baltic Science Education*, *14* (6), 779-790.

Franzolin, F., & Bizzo, N. (2015). Types of deviation in genetics knowledge presented in textbooks relative to the reference literature. *Procedia - Social and Behavioral Sciences*, *167*, 223-228, doi: 10.1016/j.sbspro.2014.12.666.

Graham, D. J., Hughes, J. M., Leder, H., & Rockmore, D. N. (2012). Statistics, vision, and the analysis of artistic style. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2012, *4* (2), 115-123. doi: 10.1002/wics.197.

Horáková, T., & Houška, M. (2014a). On improving the experiment methodology in pedagogical research. *International Education Studies*, *9*, 84-98.

Horáková, T., & Houška, M. (2014b). Quantitative differences among normal and knowledge texts on agriculture waste processing. *AGRIS on-line Papers in Economics and Informatics*, *4*, 59-68.

Horáková, T., Houška, M., Luhanová, K., & Černíková, K. (2014). Comparative analysis of quantitative indicators of normal and knowledge texts. *Proceedings of the 10th International Scientific Conference on Distance Learning in Applied Informatics*. Štúrovo, Slovakia. Prague: Wolters Kluwer, 621-631.

Horáková, T., & Houška, M. (2016). Differences among knowledge and normal educational texts: a fMRI Study. In *11th International Scientific Conference on Distance Learning in Applied Informatics,* Štúrovo, Slovakia: Wolters Kluwer, 523-532.

Horáková, T., & Rydval, J. (2015). Using text mining for the improvement of didactic tools in language acquisition. *Proceedings of the 12th International Conference Efficiency and Responsibility in Education (ERIE 2015)*, Prague, 163-173.

Houška, M., & Rauchová, T. (2013). Methodology of creating the knowledge text. In: *Proceedings of the 10th International Conference on Efficiency and Responsibility in Education*. Prague: Czech University of Life Sciences Prague, Faculty of Economics and Management, Department of Systems Engineering, 2013, 197-203, [CD ROM].

Huang, W., Oh, S., & Pedrycz, W. (2014). Design of hybrid radial basis function neural networks (HRBFNNs) realized with the aid of hybridization of fuzzy clustering method (FCM) and polynomial neural networks (PNNs). *Neural Networks*, *60*, 166-181.

Hrabí, L. (2012). Natural science textbooks for the fourth grade and their text difficulty. *Envigogika*, *7* (2), 1-7. doi:10.14712/18023061.322.

Hůbelová, D. (2010). Analyses textbooks regional geography for primary school. *Geographical Information, 14* (1), 55-63.

Hüllermeier, E. (2008). Fuzzy sets in machine learning and data mining. *Applied Soft Computing Journal*, *11*, 1493–1505. doi: 10.1016/j.asoc.2008.01.004.

Jiřina, M. (2003). *How to work with ANN at STATISTICA - artificial neural network*. Prague: StatSoft.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., & Daelemans, W. (2016). Authenticating the writings of Julius Caesar. *Expert Systems with Applications*, *63*, 86-96. doi: 10.1016/j.eswa.2016.06.029.

Kumar, K., & Thakur, G. S. M. (2012). Advanced applications of neural networks and artificial intelligence: A review. *International Journal of Information Technology and Computer Science*, *4* (6), 57–68. doi: 10.5815/ijitcs.2012.06.08.

Leong, C. K., Lee, Y. H., & Mak, W. K. (2012). Mining sentiments in SMS texts for teaching evaluation. *Expert Systems with applications*, *39* (3), 2584-2589. doi: 10.1016/j.eswa.2011.08.113.

Lin, F. R., Hsieh, L. S., & Chuang, F. T. (2009). Discovering genres of online discussion threads via text mining. *Computers & Education*, *52* (2), 481-495. doi: 10.1016/j.compedu.2008.10.005.

Lin, C. W., Hong, T. P., & Lu, W. H. (2010). Linguistic data mining with fuzzy FP-trees. *Expert Systems with Applications*. doi: 10.1016/j.eswa.2009.12.052.

Lindsey, J. K. (2009). *Introduction to applied statistics: A modelling approach*. New York: OXFORD University Press.

Ludger, G. F. (2009). *Artificial intelligence- structures and strategies for complex problem solving*, 5th edition, Pearson. In Kumar & Thakur, 2012.

Mařík, V., Štěpánková, O., Lažanský, J. et al. (2004). *Artificial Intelligence I – IV.* Prague: Academia Praha.

Mishra, P., & Koehler, M.J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108* (6), 1017-1054. doi: 10.1111/j.1467-9620.2006.00684.x.

Mohammad, A. H., & Zitar, R. A. (2011). Application of genetic optimized artificial immune system and neural networks in spam detection. *Applied Soft Computing Journal*, *11* (4), 3827–3845. doi: 10.1016/j.asoc.2011.02.021.

Moradi, G., Mohadesi, M., & Moradi, M. R. (2013). Prediction of wax disappearance temperature using artificial neural networks. *Journal of Petroleum Science and Engineering*, *108*, 74-81.

Nauman, T. W., & Thompson J. A. (2014). Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. *Geoderma*, *213*, 385-399.

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Navarro Silvera, S. A., Mayne, S. T., Gammon, M. D., Vaughan, T. L., Chow , W., Dubin, J. A., Dubrow, R., Stanford, J. L., West, A. B., Rotterdam, H., Blot, W. J., & Risch, H. A. (2014). Diet and lifestyle factors and risk of subtypes of esophageal and gastric cancers: classification tree analysis. *Annals of Epidemiology*, *24*, 50-57.

Puniškis, D., Laurutis, R., & Dirmeikis, R. (2006). An artificial neural nets for spam e-mail recognition. *Electronics and Electrical Engineering*, *5* (69), 73-76.

Rajan, K., Ramalingam, V., Ganesan, M., Palanivel, S., & Palaniappan, B. (2009). Automatic classification of Tamil documents using vector space model and artificial neural network. *Expert Systems with Applications*, *36*, 10914-10918.

Rauchová, T., & Houška, M. (2013a). Efficiency of knowledge transfer through knowledge texts: statistical analysis. *Journal on Efficiency and Responsibility in Education and Science*, *6*, 46-60.

Rauchová, T., & Houška, M. (2013b). A calculation scheme for measuring the efficiency of knowledge texts for vocational education. *Procedia - Social and Behavioral Sciences*, *106*, 10-19.

Razi, M. A., & Athappilly, K. (2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications*, *29*, 65-74.

Reichelt, M., Kammerer, F., Niegemann, H.M., & Zander, S. (2014). Talk to me personally: Personalization of language style in computer-based learning. *Computers in Human Behavior, 35*, 199-210. doi: 10.1016/j.chb.2014.03.005.

Srivastava, A. N., & Sahami, M. (2009). *Text mining: Classification, clustering, and applications*. New York: CRC Press Taylor & Francis.

Shukla, S. S., & Vilay, J. (2013). Applicability of artificial intelligence in different fields of life. *International Journal of Scientific Engineering and Research (IJSER)*, *1* (1-3). Retrieved 15/12/2016, from http://www.ijser.in/archives/v1i1/MDExMzA5MTU=.pdf.

Sklenák, V., Berka, P., Rauch, J., Stross, P., & Svátek, V. (2001). *Data, information, knowledge and internet.* Prague: C. H. Beck, Prague.

Tang, X., & Cao, J. (2015). Automatic genre classification via n-grams of part-of-speech tags. *Procedia-Social and Behavioral Sciences*, *198*, 474-478. doi: 10.1016/j.sbspro.2015.07.468.

Tayyebi, A., & Pijanowski, B. C. (2014). Modeling multiple land use changes using ANN, CART and MARS: Comparing tradeoffs in goodness of fit and explanatory power of data mining tools. *International Journal of Applied Earth Observation and Geoinformation*, *28*, 102-116.

Tu, Y., Chang T., Chen, C., & Chang, Y. (2012). Estimation of monthly wind power outputs of WECS with limited record period using artificial neural networks. *Energy Conversion and Management*, *59*, 114-121.

Received: *January 15, 2017* Accepted: *May 22, 2017*

Tereza Horáková	PhD Student, Research Assistant at Department of Systems Engineering, Faculty of Economics and Management, Czech University of Life Sciences Prague, Kamýcká 129, Prague, 165 00, Czech Republic. E-mail: horakovat@pef.czu.cz Website: http://home.czu.cz/rauchova/
Milan Houška	PhD, Associate Professor at Department of Systems Engineering, Faculty of Economics and Management, Czech University of Life Sciences Prague, Kamýcká 129, Prague, 165 00, Czech Republic. E-mail: houska@pef.czu.cz Website: http://home.czu.cz/houska/
Ludmila Dömeová	PhD, Associate Professor at Department of Systems Engineering, Faculty of Economics and Management, Czech University of Life Sciences Prague, Kamýcká 129, Prague, 165 00, Czech Republic. E-mail: domeova@pef.czu.cz