

Classification of Special Web Reviewers Based on Various Regression Methods

Kristína Machová, Marian Mach, Miroslava Hrešková

Department of Cybernetics and Artificial Intelligence, Technical University of Košice, Letná 9, 04200 Košice, Slovakia, kristina.machova@tuke.sk, marian.mach@tuke.sk, miroslava.hreskova@student.tuke.sk

Abstract: The paper is from the field of antisocial behaviour recognition in online discussions. It focuses on the extraction of knowledge about special web reviewers as authorities or trolls, and trying to distinguish one from the other based on their opinions and credibility. This paper presents a statistical application of supervised learning for creating a prediction model applied to find authorities (or trolls) among authors of online comments. The model can be used for differentiating between authoritative and non-authoritative reviewers. Standard methods such as linear and logistic regression and genetic programming were applied. The objective of designed approach is to model dependency of the reviewer variable on independent predictors representing special reviewing. Values of those independent predictors (variables) are extracted from the data about structure as well as text content of online discussions. The model can offer important information for social web users, who search for truthful and reliable information in the process of creating their own opinions.

Keywords: reviewers classification; authority; troll; linear regression; logistic regression; genetic programming

1 Introduction

Currently, social media are being studied very intensively. Particularly, the extraction of summarized knowledge from online discourse content has been a popular topic of research in recent years. The volume of discourse data, which is extracted every day from social networks is too large and time consuming to be processed by a human. Discourse data are cumulated on different online platforms, such as Facebook, Google+, Twitter, Disqus, Quora, etc. Within posts on social networks, useful information is mixed with misleading information, such as fake news or trolls' reviews. Thus, it is important to search for authoritative sources and reviewers. An authoritativeness evaluation becomes an increasingly important aspect of research on online discussion media. This evaluation can generate a list of reviewers, who are ordered according to their authoritativeness. Usually, other users follow the most credible reviewers. Our work can bring a useful insight, for example, when we have to rely on the opinions given by reviewers, who are reviewing papers for conferences or journals.

Within this publication, we are focusing on the extraction of knowledge about specific reviewers of online communities. Our aim is to distinguish between classes of reviewers – authorities or trolls – automatically, in order to know whether we can trust their opinions. This approach is a part of research on the recognition of antisocial behaviour in online communities. When classifying a reviewer into the authority or non-authority (troll) class, we have to address a classification problem. On the other hand, we can consider this problem as a regression problem, when we want to estimate the exact value of authoritativeness or trolling of some reviewer from a given interval. We have used the regression analysis, particularly linear, nonlinear, polynomial or symbolic regression as a measure of authoritativeness estimation, and logistic regression for web reviewer classification into the class of authority or non-authority.

2 Related Works

Our aim is to search for authorities and trolls of an online discussion. Other similar problems are authorship attribution, authorship verification of web reviewers and author profiling. Authorship attribution and identification is a problem of labelling an unknown document with a correct author from a given list of potential authors. The authorship attribution is based on measuring the similarity between the authors' interests (including writing styles) and the given text [1]. According to [2], character n-grams are considered among the best predictive features for the authorship attribution task. The paper presents an experiment with different machine-learning methods. The model for the authorship attribution can be successfully used for authorship verification as well. The objective of authorship verification [3] is to determine if a specific author has written a given text. The aim of the author profiling method is to learn all dimensions of the author profile. The paper [4] describes a method of learning an author profile with the focus on two dimensions: age and gender. The aim of this learning method is to identify differences in writing between a man and a woman within given dimensions.

Similar task was solved in work [5], which presents a methodology for detecting fake profiles on Twitter social network, and consequently, to associate them to real profiles in the same social network. This approach is based on an analysis of posts content generated by both fake and real profiles. Machine learning methods, such as Support Vector Machines, Random forests, K-NN and Naïve Bayes classifiers, were used. The results of accuracy were rather worse. The highest achieved accuracy was 0.68.

The work [6] addresses problems related to detecting communities of authority and estimating the influence of such communities in dynamic social networks. The work detects communities sharing common interests – called "meta-communities", using topic modelling and the between-ness centrality. The authors

empirically demonstrate the suitability of presented approach for the community-of-authority detection.

The approach [7] focuses on spreading of rumours in online discussions on Twitter. It is used to find an opinion leader using centrality measure metric on social network analysis. Aside from defining the centrality measure, it defines the special weighted edge. Twitter creates different kinds of relationship that can be turned into an edge, but not all the relationships have the same impact on spreading rumours. Thus, the study experiment considered edge weighting and centrality weighting. The study found that the edge with the ability to spread to a wider audience (quote, retweet, and reply) tends to have a bigger impact on finding an opinion leader. The study also finds that a low in-degree weight, high between-ness weight and low or no PageRank weight could give a 100% agreement upon other evaluation algorithms for finding the opinion leader.

Bougoussa et al. [8] proposes a parameter-less mixture model-based approach. They represent each user with a feature vector composed of information related to their social behavior and activity in an online community. Then they propose a statistical framework, based on the multivariate beta mixtures, in order to model the estimated set of feature vectors. Therefore, the probability density function is estimated and the beta component which corresponds to the most authoritative users is identified.

The work [9] uses an assumption that the influencer has greater effect on the online social network than the average member. According to this definition, the influencer can be considered as an authority. In this work, an approach for influencer detection is designed, using semantic analysis to filter out irrelevant interactions, and achieving a simplified graph representation allowing the detection of true influencers. The approach is an interesting aggregation of both, semantic and social web.

According to paper [10], digital revolution has drastically changed people's lives. They try to distinguish the differences between two other user types - digital immigrants and natives. Digital natives are characterised by their highly automatic and quick response in a hyper-textual environment. Digital immigrants are characterised by their main focus on textual elements and a greater proneness to reflection. The main goal of the present research is to investigate the effect of affective priming on prosocial orientation in natives and immigrants by using a mobile application.

The majority of existing methods have certain limitations. There is usually a lack of automatic mechanism to formally differentiate between authoritative and non-authoritative users. So we tried to find a formula for estimation of authoritativeness and trolling of an online reviewer.

3 Authoritative Posting

The measure of “authority” can be represented by an ability of people to influence others, positively or negatively, and to lead them to achieve a certain goal. The concept of “authority” comes from the Latin word “augere”. It denotes a person with opinions, attitudes or decisions respected by other group members. The authority can be derived from the relations between people (web users), but also from positions and hierarchies [11]. There are many kinds of authorities. Often, we divide them into *formal authorities* (measure of a person’s influence based on their formal position, regardless of personality traits) and an *informal or natural authorities* (result of personal capability, adequate self-confidence, spontaneous influence on others, no usage of pressure or force). Our work focuses on detection of informal – natural authority.

An important question to be answered is how the authority of social media can be defined. Reviewers insert their posts with their opinions, ideas and attitudes to an online discussion forum and thus create a “discourse content”. The discourse content represents data for modelling an authority of online discussion, which is a special kind of authority. It is related not only to the content but also to the structure of an online discussion. We tried to find attributes (variables, predictors) representing the authoritative posting. The values of these attributes should be extracted from the structure of an online discussion as well as short texts of the posts. To achieve our objective, we need to define:

- attributes (variables, predictors) – values of which can be extracted from an online discussion,
- dependency of the variable “Authority” on the independent variables – attributes selected in the first step. This dependency should be represented by an estimation or discrimination function.

We will use capital letter A in “Authority” for a case, when Authority is not a general concept but an exact variable.

3.1 Attributes of Authority from the Structure

Each online discussion can be represented by a tree (Figure 1). The following information can be obtained from the discussion tree:

- *Number of discussion posts* by a given reviewer
- *Number of reactions* to posts(s) by a selected reviewer
- *Position in the discussion tree* expresses the mean level of posts by a reviewer.
- *Word number* represents the ratio of the number of words in the reviewer’s comments and the total number of words in a discussion.

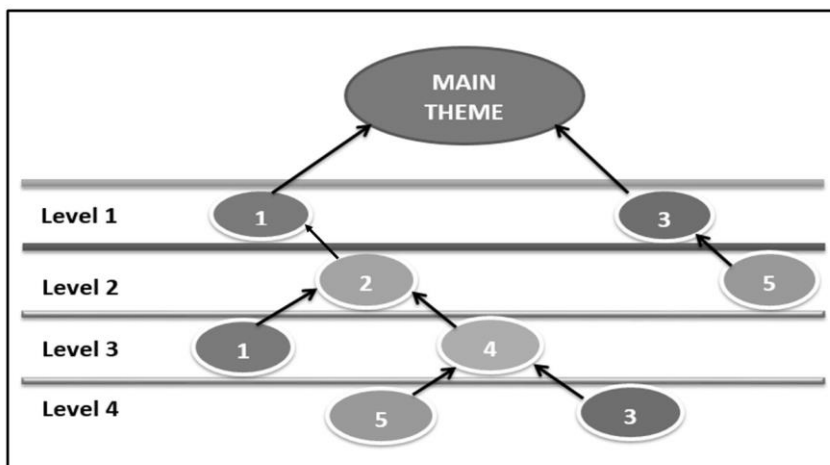


Figure 1

The discussion tree has 4 levels. The main theme is in the root. All comments of the same reviewer have the same number. Reactions are situated on levels 1 – 4

3.2 Attributes of Authority Related to Meta-information

Another approach to the authority attributes selection is based on their extraction of attributes from metadata about the reviewer posts in an online discussion. For example, evaluation of reviewer's authority by other reviewers:

- *Average evaluation of the comment* is represented by the ratio of the sum of all reactions (agree (+) and disagree (-)) on the posts of a given reviewer to the number of all their posts.
- Value of *karma* is represented by a direct ratio of the number of readers who added a like to the total number of readers of the reviewer's posts. Karma is a number from the interval 0 to 200.
- *Number of matched words* with a specific thesaurus.
- *Number of likes* in one online discussion.
- *Total sum of likes*.
- *Number of followers*.
- *Number of followed*.

4 Used Regression Methods

We have used regression analysis methods [12] (linear, logistic and symbolic regression) to learn a model for the recognition of special online reviewers as

authorities or trolls. The linear regression is based on formal definition of a line, which approximates the input data in the best way. It means that predicted values of depending variable (Authority in our case) are the most similar to the real values in the training data. The logistic regression is derived from the linear regression. These two methods vary in the type of variables used. The variable can be continuous or discrete.

- The *continuous variable* takes all values from a given interval of real numbers. For example, temperature from minus to plus values. Linear regression is a suitable method for modelling continuous variables. The continuous variables are used in solving the regression task – how to predict exact value of depended attribute for a new observation.
- The *discrete variable* takes values from a definite list of values. The values are categorical, for example man/woman or authority/non-authority. Logistic regression is a suitable method for modelling the discrete variable. Discrete variables are used in solving the classification task, i.e. how to predict the class for a new observation.

In short, the regression task represents an exact value prediction and classification task represents a class prediction.

The linear and logistic regression appeared to be a natural choice, because we wanted to obtain an estimation function, which can be linear because of the character of the data. We also wanted to learn the decision procedure for distinguishing an authoritative reviewer from a non-authoritative one based on the same data. Our aim was not only to learn the parameters of linear or other implicit functions but also to elicit the appropriate form of an estimation function without any conditions about its form. Therefore, we selected the genetic programming for this task. The genetic programming can not only learn an unexpected function, but can also randomly select only some of the variables (attributes) in training data to form an estimate of the function, thus providing useful information about the importance of authoritative posting attributes.

4.1 Linear Regression

The linear regression is a method for modelling dependency of variable Y (Y =Authority in our case) on the independent variable. It can be a simple $Y=f(x)$ or a multiple regression, when we are modelling the dependency of the variable Y on several independent variables, called predictors, (x_1, x_2, \dots, x_N) (1).

$$Y = f(x_1, x_2, \dots, x_N). \quad (1)$$

The aim is to describe this relation by a suitable mathematical model, for example, by a linear function. The result will be a regression line, which should optimally match the empirical polygon [13]. The linear regression can be represented by the equation (2). Values of parameters (weights of predictors) w_0, w_1, \dots, w_N have to

be found in order to achieve the optimal matching with the point graph (observations). These weights can be calculated using the Ordinary Least Squares method [14] for minimising the sum of square mistakes.

$$y_i = w_0 + w_1x_{i1} + \dots + w_nx_{in} + \varepsilon_i \quad (2)$$

4.2 Logistic Regression

When we use linear regression on data with a discrete variable, there appears to be a problem with the correct prediction. This problem (illustrated in Figure 2) arises when a sudden change (jump) occurs on the Y-axis – change from no reaction to reaction or similarly from non-authority to authority. Linear regression will incorrectly predict part of this change from value “response=0” to value “response=1”. In this case, it is better to use a sigmoid curve obtained by the logistic regression. In Figure 3, we can see that the logistic regression is more suitable for predicting discrete variables.

If our aim is to estimate the value of reviewer’s authoritativeness from a given interval, the linear regression is a more suitable method. However, if we want to classify a web reviewer into the class of authority or non-authority, logistic regression is more suitable. Now the question is how to obtain the sigmoid curve, which is typical for logistic regression from the data. The logistic regression can be considered as an extension of linear regression. A line, i.e. graphical representation of linear regression, is defined by two constants: b_0 (cross point of the line) and b_1 (directional vector). In practice, the line can be used for prediction of value Y depending on X . Predicted value $Y' = b_0 + b_1X$. In the case of logistic regression, the linear line has to be transformed to sigmoid curve. In such case, Y is defined as $Y=1$, if a reviewer is considered an authority, or $Y=0$, if a reviewer is not considered an authority. If p represents probability of the fact, that the given reviewer is an authority, then $(1-p)$ represents probability that the reviewer is not an authority and ratio of p and $1-p$ represents a chance that the reviewer is an authority. It is required for the logarithm of this chance to be linear for predictors X . This is modelled according to [15] by equation (3).

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1X \quad (3)$$

In the case of multiple independent variables X_1, \dots, X_N , the logistic regression can be modelled using the equation (4):

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1X_1 + \dots + b_nX_n. \quad (4)$$

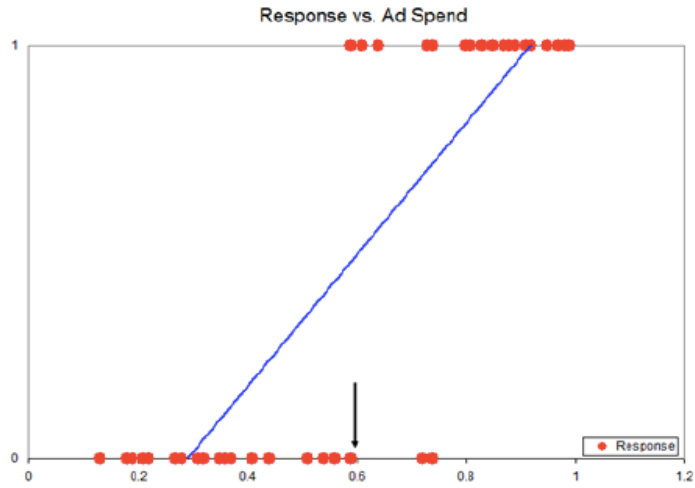


Figure 2

Linear regression in prediction of (no)reaction in dependence on an advertising spending [15]

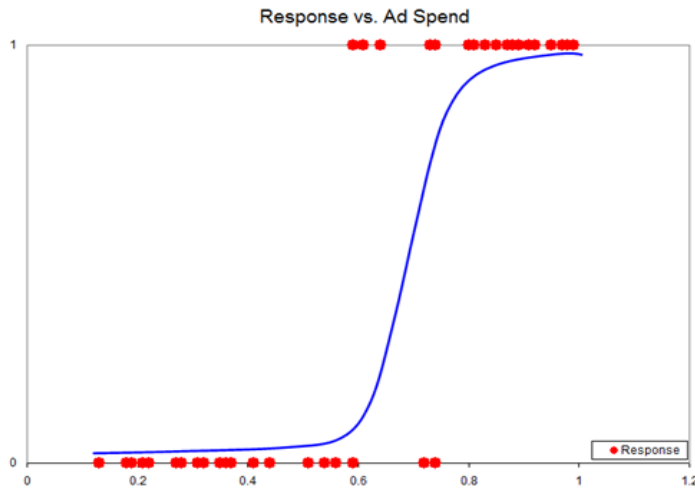


Figure 3

Logistic regression in prediction of (no)reaction in dependence on an advertising spending [15]

According to this model, the probability that a reviewer is an authority ($Y=1$) can be calculated according to the equation (5):

$$p(Y = 1) = \frac{e^{b_0 + b_1 X_1 + \dots + b_N X_N}}{1 + e^{b_0 + b_1 X_1 + \dots + b_N X_N}} \tag{5}$$

After transformation, final model representing sigmoid curve has the following form (6):

$$p(Y = 1) = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + \dots + b_N X_N)}} \quad (6)$$

During the learning of this model, the values of constants (b_0, b_1, \dots, b_N) have to be learned from a training set.

5 Authority Modelling Using the Classic Regression Analysis

5.1 Used Data

We have prepared a set of real data from an online discussion. We have used the same data set in all our experiments with different methods of statistical learning to compare the used techniques. The Authority value was related to reviewers, not to their comments. Therefore, we collected and aggregated all data about each reviewer. After the pre-processing step, the aggregated data had the form illustrated in Table 1. There were 117 reviewers acquired from the portal “www.sme.sk”.

Table 1
Illustration of the data collected about each reviewer

Reviewer	AE	K	NCH	AL	ANR	NP
c1	60	108	26,0	0	1,0	1
c2	80	182	220,0	2	0,5	2
c3	80	171	548,5	3	2,5	2
c4	80	162	57,5	4	0,5	2
c5	50	99	112,5	6	0,0	2
...						

For each reviewer we collected values of the following independent variables – attributes of the web discussion:

- *NP* is the number of posts of a given reviewer.
- *ANR* is the average number of reactions to the author’s posts.
- *AL* is the average number of layers, on which the posts of a reviewer are situated within the discussion tree (Figure 1).
- *NCH* is the average number of characters in a post by the reviewer.
- *K* is the karma of a reviewer in the form of a number from 0 to 200, which represents the activity of a reviewer from the last 3 months.

- *AE* is the average evaluation of the comment, which is available on the web discussion page. The values of *AE* are in the range from 0 to 80.
- The input data matrix was regular. There does not exist any strong correlation between any pair of variables – attributes. It is illustrated in Table 2.

Table 2
Correlation matrix of data from portal “www.sme.sk”

Attributes	K	NCH	AL	ANR	NP	AE
K	1	0.121	0.153	0.078	0.114	0.307
NCH	0.121	1	0.009	-0.024	0.162	0.348
AL	0.153	0.009	1	-0.027	0.354	0.015
ANR	0.078	-0.024	-0.027	1	0.071	-0.050
NP	0.114	0.162	0.354	0.071	1	0.190
AE	0.307	0.348	0.015	-0.050	0.190	1

In the Table 2, we can see values from the interval $\langle +0.01, +0.29 \rangle$ representing a nearly zero correlation or values from the interval $\langle +0.30, +0.49 \rangle$ representing a moderately medium correlation. There are no values representing a strong correlation ($\langle +0.50, +1.00 \rangle$). The types of correlations are presented in [16].

5.2 Linear and Non-linear Regression

The estimation function for the prediction of the value of authority can be represented by the formula (7):

$$Authority = f(NP, ANR, AL, NCH, K, AE) \quad (7)$$

At the beginning, we computed weights of linear and nonlinear regression functions. For computing these weights, we needed to know not only the values of independent variables *NP*, *ANR*, *AL*, *NCH*, *K*, *AE* but also the values of the dependent variable *Authority*. These values were obtained for each reviewer from:

- labelling by a “human expert”,
- labelling by other reviewers of the online discussion – it represents the “wisdom of the crowd”. This labelling represents a sum of information about a labelled example – reviewer, which was obtained from evaluations of this reviewer by other reviewers in the same discussion.

The following regression functions for authority estimation were generated: linear and non-linear functions learned from the “human expert” or “wisdom of the crowd”. All these functions were generated using standard MATLAB functions: “regress” in the case of linear and “lsqnonlin” in the case of non-linear relations. No auxiliary regularization method was used because the input data matrix was regular. All learned functions have been tested using average deviation for the

validation of estimation functions. After transforming the regression problem into a classification problem, other tests were performed using classical precision and recall measures. The results can be found in Table 3.

Table 3

Achieved average deviation, precision and recall of tests of the approach to the authoritativeness identification based on linear and nonlinear regression

Version	DEVIATION		PRECISION		RECALL	
	EXPERT	CROWD	EXPERT	CROWD	EXPERT	CROWD
Linear	17.34	3.29	0.70	0.98	0.67	0.80
Non-linear	18.11	6.56	0.67	0.97	0.67	0.80

Once we have a suitable model for authority estimation, we can provide the information about an exact authority value of some reviewer from a defined interval to web users. The user might not need the exact value of authority but only an indication of whether the reviewer is or is not an authority. Thus, we tried to transform the regression problem to a classification problem by defining a threshold value for the variable Authority. All reviewers with the predicted value of Authority equal or higher than the defined threshold will be identified as authoritative reviewers. For this classification problem, the obvious measures precision and recall were used for the validation of classification to class Authority or Non-authority. For discrimination we used the *Threshold value of Authority = 70* from the interval (0, 100). The detailed results are in [17] and some of them are illustrated in Table 3 to compare them with results achieved by logistic regression and symbolic regression (genetic programming). The best results were achieved by learning a linear function from the “wisdom of the crowd” data.

We have also used RapidMiner Studio (version 7.4) for the creation of predictive models based on regression analysis. We chose this solution to avoid the need to generate an estimation function and use it in combination with the defined threshold to transform the regression problem to a classification problem. RapidMiner disposes of the operator “Classification by regression”. The operator can use linear regression for prediction of a binary variable. But achieved results were very poor. That was the reason why we focused on a different solution using logistic regression, which is suitable just for the prediction of a categorical (binary) variable.

5.3 Logistic Regression

Besides our model based on linear regression, a model based on logistic regression was created using the operator “Logistic Regression” in RapidMiner. We set the

solver to the value AUTO. We used “Split data” for transforming data into training and testing sets in the ratio of 70/30 percentage.

We performed experiments with the model based on linear and logistic regression using 10 fold cross validation. These experiments were done on three data sets: www.sme.sk, www.quora.com, and www.disqus.com using the well-known measures Precision and Recall (see Table 4).

Table 4

Achieved Precision and Recall of tests of the designed approaches to authority classification based on linear and logistic regression

	www.sme.sk		www.quora.com		disqus.com	
Model	Precision	Recall	Precision	Recall	Precision	Recall
Linear regression	0.91	0.97	0.97	0.85	0.86	0.38
Logistic regression	0.96	0.95	0.94	0.91	0.80	0.91

From the results in Table 4, we can hardly say which is better, the model based on linear regression or the model based on logistic regression. Therefore, for easier comparison, we also present results of other measures such as Accuracy, F1 rate and AUC. Table 5 illustrates Accuracy and F₁ rate representing a harmonic average of Precision and Recall.

Table 5

Achieved Accuracy and F1 rate of tests of the designed approaches to authority classification based on linear and logistic regression

	www.sme.sk		www.quora.com		disqus.com	
Model	Accuracy	F ₁	Accuracy	F ₁	Accuracy	F ₁
Linear regression	0.914	0.939	0.929	0.906	0.823	0.527
Logistic regression	0.914	0.955	0.917	0.925	0.887	0.851

The effectiveness of our models was verified by the means of an ROC curve [18] and an AUC value, which represents the area under the ROC curve.

- Values of AUC from the interval $<0, 0.50)$ represent an inappropriate model;
- AUC values from the interval $<0.50, 0.75)$ represent an acceptable but not a very effective model;
- AUC values from the interval $<0.75, 0.92)$ represent a good model for the given data sample;
- AUC values from the interval $<0.92, 0.97)$ represent a very good model;
- AUC values from the interval $<0.97, 1.00)$ represent an excellent model for the given data sample.

The results for linear and logistic regression are presented in Table 6. These results confirm that the model based on logistic regression is more suitable for Authority classification than the model based on linear regression.

Table 6

Achieved results of AUC values of tests of the designed approaches to authority classification based on linear and logistic regression

	www.sme.sk	www.quora.com	disqus.com
Model	AUC	AUC	AUC
Linear regression	0.50	0.50	0.50
Logistic regression	0.98	0.95	0.96

The results of the AUC values for linear regression are 0.50. The reason for this fact could be that linear regression is not primarily designed for the prediction of binary (categorical) attributes. Even the operator “Classification by regression” cannot improve the results of the model based on linear regression. On the other hand, the results of the AUC values for logistic regression evaluate this model as very good or excellent.

6 Genetic Programming as Symbolic Regression

Genetic programming is a modification of genetic algorithms which was designed by John Koza [19]. The chromosomes are represented by character strings in genetic algorithms. In genetic programming, these character strings are replaced by symbols, which can be represented by operators or by more difficult structures such as elementary functions. A chromosome, graphically represented as a syntactic tree, itself represents some function $y=t(x)$. This function assigns a dependent variable y to an independent variable x . For example, the function $t(x) = x(I+x)$ in Figure 4. Let $A = \{(x_i, y_i), i=1, \dots, N\}$ be a training set containing N points. Then, our goal is to find the function $t(x)$ by minimalizing the difference (quadratic or absolute) between the computed value of y and a real value of y from the training set. Particularly, when the sum of all the differences is smallest, we will have found the goal function represented by a syntactic tree. This formulation of the task is close to regression analysis where only the parameters of a given function are learned for a training set of points. The space of functions is finite, and their form changes only based on changing parameters. In genetic programming the space of functions is infinite, so genetic programming represents a generalized approach to symbolic regression. The process of symbolic regression is based on the operations of crossover and mutation. Crossover is illustrated in Figure 5 and Figure 6. Two parents (syntactic trees) were selected as inputs for the operation of crossover. The special nodes emphasized with black circles are called “points of crossover” and

they are selected randomly. After switching these points of crossover, we will obtain the new generation.

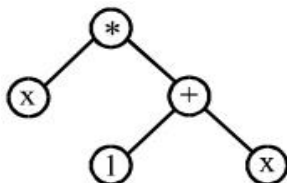


Figure 4

Syntactic tree as a representation of a chromosome corresponding to the function $x(1+x)$ [20]

The process of symbolic regression is based on the operations of crossover and mutation. Crossover is illustrated in Figure 5 and Figure 6. Two parents (syntactic trees) were selected as inputs for the operation of crossover. The special nodes emphasized with black circles are called “points of crossover” and they are selected randomly. After switching these points of crossover, we will obtain the new generation.

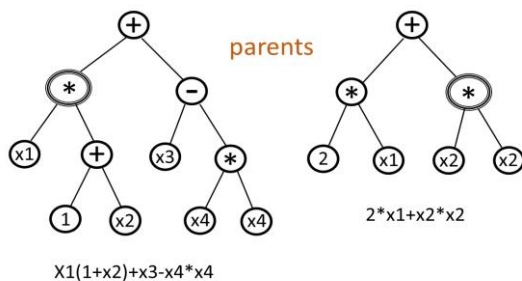


Figure 5

Illustration of parents’ syntactic trees as an input to the operation crossover

The mutation operation (see Figure 7) comprises of replacing a subtree of the original tree with a new randomly generated subtree. This means that some branch of the syntactic tree is randomly changed. Besides these two elementary genetic operators, other secondary operators can be used, for example permutation, editing, encapsulation, decimation and so on [20].

We need to use an individual in the form of a syntactic tree representing a function in genetic programming. The individual can be a computer program formed from functions and terminals. From these illustrations of crossover and mutation operations we can deduce that a random polynomial function can be constructed in the space of nodes with two arguments equivalent to the arithmetic operations of addition, subtraction and multiplication. In case we expand the set of acceptable

nodes by those equivalent to new operations such as division, exponentiation etc., we can express practically any function. Work [21] contains more information about genetic programming.

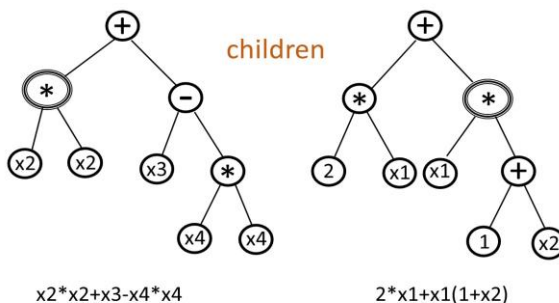


Figure 6

Illustration of childrens' syntactic trees representing a new generation as the result of crossover over two syntactic trees in Figure 5

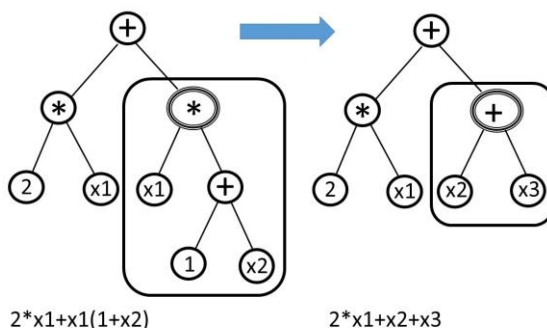


Figure 7

Illustration of mutation of one syntactic tree

Nowadays, genetic programming is often used for different classification problems. The work [22] offers a review of tree-based genetic programming classification methods. It also analyses their strengths and weaknesses in relation to the task of classification. In conclusion, genetic programming can be effectively used for performing the classifier evolution task. On the other hand, long training time and lack of convergence can be considered as its disadvantages. Nevertheless, it can be succesfully used for this task.

7 Authority Modelling Using Genetic Programming

We have prepared an environment in Matlab for experiments with genetic programming to train a model for Authority estimation. At first we did some experiments with the finite although big space of elementary operators (functions). Nevertheless, the generated estimation functions were too complicated, completely unreasonable and it was impossible to interpret them. For this reason, we decided to constrain the space of elementary operators to the following: addition, subtraction, multiplication and exponentiation to the second power. The function regfitness was used for the measure of fitness of individuals in the form: $\text{ind} = \text{regfitness}(\text{ind}, \text{params}, \text{data}, \text{terminals}, \text{varsvals})$.

We have used the data from section 6.1 to train the estimation function. The data were divided into three parts (123, 132 and 231) for three fold cross validation (for example 123: 1st and 2nd parts for training and 3rd part for testing). Six experiments were conducted for both learning from an expert and from wisdom of the crowd. From the six functions generated in the process of learning using genetic programming, the function (8) learned from the “wisdoms of the crowd” on data 132 achieved the best results according Table 7.

$$A = \text{plus}(\text{ANR}, \text{plus}(\text{NC}, \text{plus}(\text{NC}, \text{plus}(\text{AE}, \text{times}(\text{AL}, \text{NC})))))) \quad (8)$$

All learned functions have been tested using classical Precision, Recall and F1 measures. The results can be found in Table 7. The results are promising (between 94 and 100 percent), with the exception of Precision of the models trained on expert labelled data (in average 59 percent) and consequently a lower F1 rate. Overall results for all used methods of regression analysis: linear, non-linear, logistic and symbolic (genetic programming) are presented in Table 8.

Table 7

Achieved results of Precision, Recall and F1 rate of tests of the approach based on genetic programming

Data	Precision		Recall		F1	
	EXPERT	CROWD	EXPERT	CROWD	EXPERT	CROWD
123	0.592	0.937	0.938	1.000	0.726	0.967
132	0.623	0.974	0.979	1.000	0.762	0.987
321	0.562	0.961	0.911	1.000	0.695	0.980
Average	0.592	0.957	0.943	1.000	0.728	0.978

According to the values of the F1 measure (a balanced averaged of Precision and Recall) in Table 8, the best model for authority classification is the model trained by symbolic regression (genetic programming) on data labelled by a crowd. Using logistic regression, we have trained the second best model on the same data. In the case when we only consider the precision measure reflecting the number of false positive classifications, the best model is that learned using linear regression on crowd data. On the other hand, if we only consider the recall measure reflecting

the number of false negative classifications, the best model is the one trained using symbolic and then logistic regression on crowd data.

The genetic algorithms used have brought new information about a real need and suitability of independent attributes of authoritative posting. Table 9 presents the frequency of appearance of these attributes (AE, ANR, AL, NC, NCH and K) in 16 experiments within the space of elementary operators (functions) and 6 experiments with a constrained space of elementary functions in the process of generation of the Authority estimation functions.

Table 8

Achieved results of Precision, Recall and F1 rate of tests in all used methods on the same data

	Precision		Recall		F1	
	EXPERT	CROWD	EXPERT	CROWD	EXPERT	CROWD
Regression	0.70	0.98	0.67	0.80	0.68	0.88
Linear	0.70	0.98	0.67	0.80	0.68	0.88
Non/linear	0.67	0.97	0.67	0.80	0.67	0.87
Logistic	-	0.91	-	0.97	-	0.94
Symbolic	0.59	0.96	0.94	1.00	0.73	0.98

Table 9

Frequency of appearance of the monitored attributes in generated functions using genetic programming within 22 experiments

Attribute	Frequency	%
AE – average evaluation	22	100,00
ANR – average number of reactions	14	63,64
AL – average layer	13	59,09
NC – number of comments	11	50,00
NCH – number of characters	7	31,82
K - karma	7	31,82

For future experiments, we might only use the variables AE, ANR, AL and NC for the generation of the Authority estimation function. A surprising insight was that the “*karma*” of a reviewer doesn’t seem to be very influential in determining his/her authority.

Conclusions

The contribution of this paper is the design of a new approach to the recognition of authority or trolling for special online reviewers. This approach is based on the solution of the problem of Authority identification from a discourse content. A similar solution could be used for Troll classification when attributes of troll posting are defined. The first solution was based on learning the Authority estimation function from the structure of online discussions using linear or non-linear regression. According to the test results, the best solution was a linear regression function trained from “wisdom of the crowd”. The second solution was

the classification model based on logistic regression. According to the test, the logistic regression method is more suitable for the identification of special web reviewers as an Authority. The last solution is a model obtained by symbolic regression in the form of genetic programming. This solution seems to be the most suitable. The results in Table 7 and Table 8 also prove, that learning from data labeled by crowd is more precise, no matter which learning method was used.

Authority identification can be used in many real situations. Mostly when web users search for an authority that can advise them in decision-making or forming an opinion on important topics. The work can have many useful applications, especially in areas where we have to rely on the opinions of reviewers, for example the case of reviewing papers for conferences or for journals.

The decision making process can be influenced by trolling and other types of antisocial online reviewing. For the future, we would like to focus on the modelling of troll reviewers in online discussions within the field of antisocial behavior recognition.

When we talk about sentiment analysis, we usually mean the analysis of opinions and emotions. Emotion analysis can be used in the field of robotics in human-robot interaction. Particularly emotion analysis of a human can be useful for a robot, which communicates with humans to be able to adapt its behavior to the emotional situation of the interaction.

Other future extensions are connected to the semantically enriched detection of special web reviewers using an ontology [23] and involving the topic modeling [24] of the texts of the examined reviews. It would also be useful to involve a neural network as a widely used method into an extended testing.

Acknowledgement

This work was supported by the Slovak Research and Development Agency under No. APVV-017-0267 “Automated Recognition of Antisocial Behaviour in Online Communities”, the contract No. APVV-016-0213 and APVV-015-0731.

References

- [1] Yang, M., Cui, T. Tu, W.: Ordering-sensitive and Semantic-aware Topic Modelling. In: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.
- [2] Markov, I., Baptista, J., Pichardo-Lagunas, O.: Authorship Attribution in Portuguese Using Character N-grams. *Acta Polytechnica Hungarica* Vol. 14, No. 3, 2017, ISSN 1785-8860, 59-78.
- [3] Stamatos, E., Fakotakis, N., Kokkinakis, G.: Automatic Text Categorization in Terms of Genre and Author, *Computational linguistics* 26, 2000, 471–495.

- [4] Mechti, S. Jaoua, M.,Belguith, L.H.,Faiz, R.: Author Profiling Using Style-based Features. In: Notebook for PAN at CLEF, 2013, 1-7.
- [5] Galán-García, P., De La Puerta, J.G., Gómez, C.L., Santos, I., Bringas, P.G.: Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying. Logic Journal of the IGPL, Vol. 24, No. 1, 2014, ISSN 1367-0751, 42-53.
- [6] Chikhaoui, B., Chiazzaro, M., Wang, S., Sotir, M.: Detecting Communities of Authority and Analyzing Their Influence in Dynamic Social Networks. ACM Transactions on Intelligent Systems and Technology, Vol.8, No.6, 2017, article 82, ISSN 2157-6904, 1-28.
- [7] Devi, FK., Yudhoatmojo, SB., Budi, I.: Identification of Opinion Leader on Rumor Spreading in Online Social Network Twitter Using Edge Weighting and Centrality Measure Weighting. In. Proc. of the 12th International Conference on Digital Information Management (ICDIM), Kyushu Univ, Fukuoka, JAPAN, SEP 12-14, 2017, 313-318.
- [8] Bouguessa, M., Ben Romdhane, L.: Identifying Authorities in Online Communities. ACM Transactions on Intelligent Systems and Technology, Vol.6, No.3, article 30, 2015, ISSN 2157-6904, 1-23.
- [9] Ríos, SA., Aguilera, F., Nuñez-Gonzalez, JD., Graña, M.: Semantically Enhanced Network Analysis for Influencer Identification in Online Social Networks. NEUROCOMPUTING, 2017.
- [10] D'Errico, F., Paciello, M., Fida, R., Tramontano, C.: Effect of Affective Priming on Prosocial Orientation through Mobile Application: Differences between Digital Immigrants and Natives. Acta Polytechnica Hungarica Vol. 16, No. 2, 2019, 109 – 128.
- [11] Chavalkova, K.: Authority of a Teacher (in Czech). Philosophic faculty of the University of Pardubice, Pardubice, Czech republic, 2011.
- [12] Chatterjee, S., Simonoff, J.S.: Handbook of Regression Analysis. A John Wiley & Sons, INC. publication, 2013, 1-236.
- [13] Pazman, A., Lacko V.: Lectures from Regression Models(in Slovak). University of Comenius Bratislava, Bratislava, Slovakia, 2012, ISBN 978-80-223-3070-1, 132 ps.
- [14] Pohlman, J.T., Leitner, D.W.: A Comparison of Ordinary Least Squares and Logic Regression. The Ohio Journal of Science, Vol.103, No.5, 2003, 118-125.
- [15] Deshpande, B.: Logistic Regression Digest. http://resources.simafire.com/hs-fs/hub/64283/file-15121709-pdf/pdfs/Simafire_logistic_regression_article_digest.pdf?t=1493427943502 , Accessed 29 November 2019.

-
- [16] Zlacká, A.: Correlation dependence. <http://www.fhvp.unipo.sk/cvt/statistika/zlacka/geoinfo7.pdf>, 2019, Accessed 29 November 2019 (in Slovak).
- [17] Machová, K., Štefaník, J.: Regression Methods in the Authority Identification within Web Discussions. In: Computational Collective Intelligence, LNAI, Vol. LNAI 9329, No. 1, 2015, Springer-Verlag, Berlin Heidelberg, ISSN 0302-9743, 203-212.
- [18] Bortlíček, Z.: ROC Curves. Masarik University of Brno, Science Faculty, Brno, 2018, 1-39.
- [19] Koza, J.R.: Genetic Programming: On the Programming of Computers by means of Natural Selection. MIT Press, Cambridge, MA, 1992.
- [20] Kvasnička, V., Pospíchal, J., Tiňo, P.: Evolutionary Algorithms (5. Genetic Programming). (in Slovak) Publishing House of STU, Bratislava, 2000, 1-215.
- [21] Návrat, P.: Artificial Intelligence: Metaheuristics and Evolutionary Algorithms. (in Slovak) Publishing House of STU, FIIT STU, Bratislava, 2014, 1-418.
- [22] Jabeen, H., Baig, A.R.: Review of Classification Using Genetic Programming. International Journal of Engineering Science and Technology, Vol. 2, No. 2(2010), ISSN 0975-5462, 94-103.
- [23] Machová, K., Vrana, J., Mach, M., Sinčák, P.: Ontology Evaluation Based on the Visualization Methods, Context and Summaries. Acta Polytechnica Hungarica, Vol. 13, No. 4, 2016, 53 –76.
- [24] Smatana, M., Butka, P.: TopicAE: A Topic Modeling Autoencoder. Acta Polytechnica Hungarica, Vol. 16, No. 4, 2019, 67 – 86.