ICPE 2018
International Conference on Psychology and Education

SOFTWARE FOR LEARNING TO SOLVE PROBLEMS OF CLASSIFICATION USING OF MACHINE LEARNING

Vladimir B. Barakhnin (a)*, Olga Yu. Kozhemyakina (b), Elena V. Rychkova (c),
Anna S. Gladkikh (d), Ilya S. Pastushkov (e)
*Corresponding author

(a) Institute of Computational Technologies SB RAS, 6 Lavrentiev Avenue, Novosibirsk, Russia; Novosibirsk State University, 1 Pirogova Street, Novosibirsk, Russia, bar@ict.nsc.ru
(b) Institute of Computational Technologies SB RAS, 6 Lavrentiev Avenue, Novosibirsk, Russia, olgakozhemyakina@mail.ru
(c) Institute of Computational Technologies SB RAS, 6 Lavrentiev Avenue, Novosibirsk, Russia; Novosibirsk State University, 1 Pirogova Street, Novosibirsk, Russia, helen@ict.nsc.ru
(d) Novosibirsk State University, 1 Pirogova Street, Novosibirsk, Russia, mikasa.vol@mail.ru
(e) Institute of Computational Technologies SB RAS, 6 Lavrentiev Avenue, Novosibirsk, Russia; Novosibirsk State University, 1 Pirogova Street, Novosibirsk, Russia, pas2shkov.ilya@gmail.com

Abstract

At present time, the philologists analyze poetic texts in order to identify the various characteristics of the literature work (the information about the verse, the rhythmics of the end of poems, a detailed description of the used rhymes, etc.), which are necessary in the process of study of the author's work. One of the most important characteristics of a poetic text are its style and genre. Currently, the specialists in this field are forced to work with the classification almost manually. The information about the reasons of the provided predictions is not available to the user, although such knowledge is of great importance in further analysis of the results. These problems can be overcome by creating a software system that allows to load the categorized data, and to show a detailed report on the results of the classification on the output display. The purpose of this work is to implement a web application to control the algorithms for automatic determination of styles and genre types of poetic texts. It allows experts in the field of philology to download the necessary data in a convenient way, to choose an automatic classifier and to analyze valuably the result of the classification, based on its justification obtained by the LIME algorithm and its implementation in the ELI5 library. It is worth noting that this decision is not tied to any specific categories of classification, what makes it universal.

© 2018 Published by Future Academy www.FutureAcademy.org.UK

Keywords: Automatic analysis of poetic texts, classification tasks, machine learning.

1. Introduction

Text classification is one of the main tasks of computational linguistics. In this regard, various methods of text classification, as well as the selection of the best algorithms for usage in research and commercial tasks, are of great importance. The modern approach to classification is based on machine learning methods.

At present time, the philologists analyze poetic texts in order to identify the various characteristics of the literature work (the information about the verse, the rhythmics of the end of poems, a detailed description of the used rhymes, etc.), which are necessary in the process of study of the author's work. One of the most important characteristics of a poetic text are its style and genre. Currently, the specialists in this field are forced to work with the classification almost manually.

As for the automatic classification of texts by genres and styles, it is quite difficult for a non-technical specialist to dive into an unknown area. Moreover, despite their widespread adoption, the machine learning models remain mainly "the black boxes". The information about the reasons of the provided forecasts is not available to the user, although such knowledge is of great importance in further analysis of the results. The main problem is the lack of confidence of users to model or prediction, which leads to the rejection of their usage.

2. Problem Statement

The formal definition of the classification task is as follows. Let the set $D = \{d_i\}_{i=1}^n$ be a sampling of objects, where each element is characterized by a feature vector, i.e. $d_i = (f_{i1}, \dots, f_{in})$. The result of object classification d_i is some value from a finite set $T = \{1, \dots, m\}$. In this case, the main purpose of the pattern recognition problem is to find an algorithm that uniquely determines the final value from the set T by the feature vector or gives a vector of estimates of the object's belonging to each class. At the initial stage, a subset of objects with the already known resulting class exists. Such objects make up a training set.

If we apply the above definition to the problem of text classification by styles or genres, we will get that d_i is a poetic text, which is characterized by a set of features (it is not known yet what). Based on these features, the classification algorithm correlates the text with one of the styles or genres. Each style or genre is assigned a unique value from the set $\{1, \dots, m\}$.

Machine learning specialists are not only involved in the construction of various classification algorithms, but also in the development of ways to mapping text objects to real vectors. After selecting a specific model, a detailed selection of algorithm parameters takes place, depending on the characteristics of the problem to be solved. It is the technical specialists with knowledge in this area are engaged in solving these issues.

3. Research Questions

The possibility of using the already developed algorithms by philologists (Barakhnin, Kozhemyakina, & Pastushkov, 2017) is considered in this paper. The problems of such use are manifested due to the lack of understanding of the processes occurring within the classifier. Philologists are unable to answer the questions, what are the algorithms and their parameters are preferable in certain tasks.

There are three main reasons why experts cannot fully use machine learning algorithms to solve practical tasks related to the specifics of the profession:

1. The difficulty in setting up the model. These include such fundamental issues as the choice of a suitable classifier, setting model parameters, preprocessing data additionally. The above issues are dealt with by specialists in the field of machine learning, and the philologist is no able to make any minimal adjustments in the future.

2. Transfer of training and test data. Due to the fact that the classifiers are an implementation of some software interface, then their use requires basic knowledge of programming. Experts in related fields are often lacking. A lot of time is spent on teaching philologists the basics of programming, and on explaining in what format it is necessary to transfer data.

3. Distrust of the results. Trust is critical to effective human interaction with machine learning systems. The user of the model cannot trust only the classification accuracy indicator because he is not aware of the reasons for the obtained results.

4. Purpose of the Study

To overcome the above problems, is need to create a software system that allows to download categorized data, and provides a detailed report on the results of the classification on the output.

The purpose of this work is to implement a web application to control algorithms for automatic determination of styles and genre types of poetic texts.

To achieve this goal, the following tasks were performed:

1. The possible approaches of predictions explanation from automatic classifiers for text data are analyzed;

2. The system requirements to the developed system is formulated;

3. The appropriate software have been selected;

4. Designed and developed software system.

5. Research Methods

5.1. Predictions explanation from classifier

In this paper "predictions explanation" means representation of text artifacts, that provide good understanding of relationship between the word in document and the prediction of the model. Predictions explanation is important condition for users ability to rely on automatic classifiers and use them correctly to obtain solutions for practical tasks. Unreliability of individual predictions is large problem in case of model used for decision making. For instance, predictions obtained by applying machine learning algorithms are incompletely reliable because false results despite of positive or negative can be significantly harmful.

LIME algorithm (Local Interpretable Model-agnostic Explanations) (Ribeiro, Singh, & Guestrin, 2016) has been examined below for further representation of classifier predictions explanation.

Before description of explanation system, it is important to distinguish classifier features from interpreted data representations. Interpreted explanations must have representation which is independent from model features and understandable for users. For instance, possible interpreted representation for text

classifier can be binary vector indicating absence / presence of a word in document although classifier may use more complex text features.

Let is $x \in R^d$ a feature vector of some object, and let is $x' \in \{0,1\}^{d'}$ a binary vector of an object, necessary for interpretation. Class of potentially interpreted models (like a linear models or decision trees) denotes as G and $g \in G$ means certain model that can be easily demonstrated to the user in the form of a text or visual artifact. The scope of g is $\{0,1\}^{d'}$, so g acts on the absence / presence of interpreted components. Since not all are simple to interpret, it is necessary to introduce $\Omega(g)$ as measure of complexity to explain the model g. For example, for decision trees a measure of complexity is the depth of the tree, and for the linear model it is the number of non-zero weights.

Consider $f: R^d \to R$ the explained model, where $f(x)$ represents the probability (or binary identifier) that an x object belongs to a class. Multiple classes have separate explanations, so $f(x)$ is explanation of certain class. Introduce the $\pi_x(z)$ as the measure of proximity between objects x and z, which is necessary to determine the neighborhood x. Finally, $\mathcal{L}(f, g, \pi_x)$ is a measure of how much the wrong model g approaches the model f in the neighborhood defined by π_x. Obtained result is that to achieve the interpretability it is necessary to minimize the function $\mathcal{L}(f, g, \pi_x)$ while $\Omega(g)$ must be sufficiently small. The following formula of LIME algorithm, which reflects the predictions explanation is below:

$$\varepsilon(x) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x) + \Omega(g).$$

Consider G as class of linear models, so $g(z') = w_g * z'$. As a proximity measure function is used function $\pi_x(z) = \exp(-\frac{D(x,z)^2}{\sigma^2})$ which is exponential kernel, defined on some function of the distance D with width σ. For instance, in text analysis cosine similarity (the ratio of the scalar product of vectors and the product of their norms) used as D and for image analysis D is and for images, the Euclidean distance L_2 is used. Thus, the function $\mathcal{L}(f, g, \pi_x)$ takes the following form:

$$\mathcal{L}(f, g, \pi_x) = \sum_{z,z' \in Z} \pi_x(z) * (f(x) - g(z'))^2.$$

5.2. System requirements

The system requirements to the developed system are formulated. These requirements were agreed with expert in the field of philology who face problems with the use of automatic classifiers.

At first user should be able to add the new text corpora. This process contains:

- Entering corpora name for further user orientation;
- Appending known categories and belonged documents. Names of categories are determined by user;
- Appending large amount documents named by user to the system.

In addition to appending new data to the system, the user may need to edit existing data. Therefore, everything that the expert has loaded into the system, it should be able to change. This includes changes in the names of the corpora of texts, categories, documents, as well as the content of the corpora and categories themselves.

The main purpose of the system is to provide recommendations on the classification of documents to the expert. That is, in the end, the user should see not only the results of the classification, but also a tentative explanation of certain conclusions.

5.3. Software

Chosen as a framework for implementation of algorithmic part of web application was the Django (Forcier, Bissex, & Chun, 2008), which encapsulates the solutions for main web-development problems, so writing the application is focused.

For more effective work on user interface the Bootstrap (Spurlock, 2013) framework was chosen. It includes set of HTML and CSS templates for simplification of web interfaces development.

Comparative analysis of DBMS's has conclude the solution to use PostgreSQL (Obe, & Hsu, 2017), that responds main criteria of data storage durability in multi-user application. During design of database structure were selected main entities corresponding input data such as corpora, category, document and etc.

Data transfer format between application and classifier instance on server is JSON (Marrs, 2017). Due to its simplicity this format uses for serializing complex structures which transforms to string values. This format is usable and popular for data transfer between any system components, therefore many programming languages have integrated solutions for data convert. Application supports ability to choose one of implemented classifiers on server.

For representing classifier predictions explanation were used Python library ELI5, that implements LIME algorithm described above. There is special class TextExplainer in ELI5 library, it implements function fit() with two parameters:

- ▪ Document, for which the explanation was built;
- ▪ Classification function, that returns probabilities membership object to category.
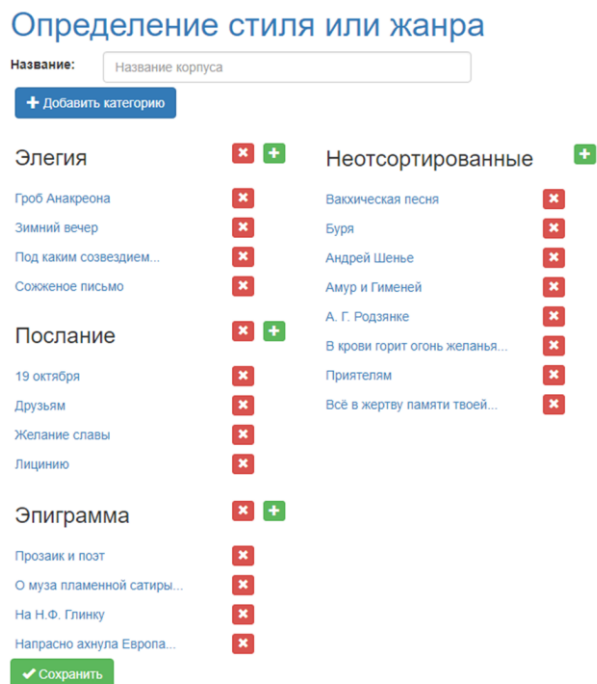
These parameters are defined by user.

6. Findings

On the main screen (Fig. 01) the user is given the opportunity to add a new corpus of text, edit or delete an existing corpus. The classification script, which contains the transition to the selection of the existing algorithm and further waiting for the results from the classifier server, is executed after clicking on the name of corpus.
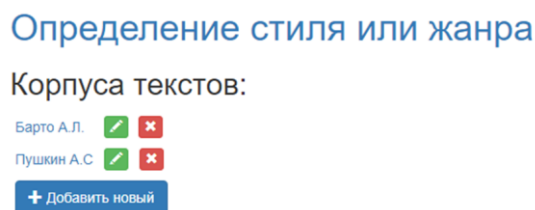
Figure 01. Main screen of the system "Style or genre identification"

The functions of working with data when editing an existing corpus of texts and adding a new corpus are the same, so they are combined on one screen (Fig. 02). By default, each corpus of texts has one non-deleted category "Unsorted", to which the user adds documents for further classification. The remaining categorized documents are then used as a training set.

Определение стиля или жанра

Название: [Название корпуса]

[+ Добавить категорию]

Элегия

- Гроб Анакреона
- Зимний вечер
- Под каким созвездием...
- Сожженое письмо

Послание

- 19 октября
- Друзьям
- Желание славы
- Лицинию

Эпиграмма

- Прозаик и поэт
- О муза пламенной сатиры...
- На Н.Ф. Глинку
- Напрасно ахнула Европа...

[✓ Сохранить]

Неотсортированные

- Ваакхическая песня
- Буря
- Андрей Шенье
- Амур и Гименей
- А. Г. Родзянке
- В крови горит огонь желанья...
- Приятелям
- Всё в жертву памяти твоей...

Figure 02. Example screen with a list of text corpuses after identification

As a result, the user is provided with information about the resulting predictions of the classifier in the form of a table (Fig. 03). In addition, information is available on the explanation of the predictions made for each text after clicking on the document name. In this case, the explanation is a list of words with relative weights. This means that the word on the screen will be highlighted in red or green color of different intensity. The green color indicates that the word has contributed to the document's category. The red color shows the opposite: this word has contributed against this category.

Определение стиля или жанра

Результаты классификации:

Документ	Категория
Ваакхическая песня	Анакреонтика
Буря	Элегия
Андрей Шенье	Элегия
Амур и Гименей	Сказка
А. Г. Родзянке	Послание
В крови горит огонь желанья...	Элегия
Приятелям	Послание
Всё в жертву памяти твоей...	Послание

Figure 03. Example of displaying classification results

7. Conclusion

As a result of the work, the application that allows to minimize the time of training the interaction of an expert philologist with the system was implemented. This is done by providing the expert with

appropriate advice on the style or genre of the poetic text. This solution is universal and easily expandable for different range of classification problems.

The immediate plans include the implementation of the study to identify the most appropriate algorithms and their parameters for the tasks of classification of poetic texts by styles and genres.

Acknowledgments

References

Barakhnin, V., Kozhemyakina, O., & Pastushkov, I. (2017). Automated determination of the type of genre and stylistic coloring of Russian texts. *ITM Web of Conferences*, 10, 02001. doi: 10.1051/itmconf/20171002001.

Forcier, J., Bissex, P., & Chun, W. (2008). *Python Web Development with Django*. Addison-Wesley Professional.

Marrs, T. (2017). *JSON at Work: Practical Data Integration for the Web*. O'Reilly Media Inc

Obe, R., & Hsu, L. (2017). *PostgreSQL: Up and Running: A Practical Guide to the Advanced Open Source Database*. O'Reilly Media Inc.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). ACM.

Spurlock, J. (2013). *Bootstrap: Responsive Web Development*. O'Reilly Media Inc.