2018-9

Building Classifiers with GMDH for Health Social Networks (BD AskaPatient)

John Cardiff
Technological University Dublin

Liliya Akhtyamova
Technological University Dublin

Mikhail Alexandrov
Autonomous University of Barcelona

Recommended Citation

Akhtyamova, L., Alexandrov, M., Cardiff, J., Koshulko, O.: Building Classifiers with GMDH for Health Social Networks (DB AskaPatient). In:,i> Proc. of the Intern. Workshop on Inductive Modelling (IWIM-2018), IEEE, 2018 DOI:10.1109/STC-CSIT.2018.8526655

Building Classifiers with GMDH for Health Social Networks (BD AskaPatient)

Liliya Akhtyamova
and John Cardiff
Institute of Technology Tallaght
Dublin, Ireland
Email: liliya.akhtyamova@postgrad.ittdublin.ie
john.cardiff@it-tallaght.ie

Mikhail Alexandrov
Autonomous University of Barcelona
Barcelona, Spain
Russian Presidential Academy of National Economy and Public Administration
Moscow, Russia
Email: MAlexandrov@mail.ru

***Abstract*—Health social media offer useful data for patients and doctors concerning both various medicines and treatments. Usually, these data are accompanied by their assessments in 5- star scale. But such a detail classification has small usefulness because patients and doctors, first of all, want to know about negative cases and to study in detail the extreme ones. In the paper we build classifiers of texts just for these cases using combined classes as negative, all others and worst, satisfactory, best. For this, we study possibilities of different GMDH-based algorithms and compare them with the results of other methods. The selection of GMDH is provoked by two circumstances: (a) health social media contain significant informative noise, and (b) GMDH is essentially noise-immunity method. The experimental material is the popular health social network Askapatient.**

I. INTRODUCTION

A. Motivation

Social media is a modern phenomenon that has opened absolutely new possibilities for analysis of various aspects of a life of the human society in total or some group of people in particular [1]. The medical domain is presented in various forums, where users discuss both general topics as the state of healthcare system or the specific questions concerning medicine, treatment etc. Such an information could be interesting to various governmental and private institutions. The former has an opportunity to evaluate the reaction of community on the laws and acts concerning healthcare as well as monitor the health condition of citizens and the latter can see a market for medicines for their production.

From the other hand, social media has provoked the development of NLP (Natural Language Processing), namely new models, methods and program systems. Medical domain presented in social media uses traditional approaches of NLP related to opinion mining. But not only these approaches. It considers specific problems related to diseases, treatment, social support of patients, and so on. For example, we may mention here the topic of adverse drug reaction extraction (ADR).

B. Problem settings

In this paper we consider possibilities of GMDH-based algorithms to build useful classifiers for processing texts from HSNs. Speaking useful classifiers we mean classifiers, which allows finding just negative or extreme cases in HSNs. The 1-st classification is grouping with 2 combined classes negative, others and the 2-nd classification is grouping with 3 combined classes full fall, satisfactory, full success. The equivalence is obvious: negative class = (1*,2*), other classes = (3*,4*,5*), full fall = (1*), satisfactory class = (2*, 3*, 4*), full success = (5*). We suppose that often these cases are more interesting for patients and doctors instead of detail classification.

We intend also to study various ways of text parameterization that is a transformation of the dataset to its vectorial form and to put attention for using not only terms but also n-gram of symbols. The latter is still enough rare way of parameterization.

This moment Internet presents various HSNs. In our experiments we use HSN AskaPatient being the typical representative of such type of Health Social Media[1].

The presented research is realized on the platform GMDH Shell (hereinafter GS) including several popular GMDH-based algorithms[2]. This tool has been already used for text classification in different applications: on Forex market [1], in a forecast of crimes [2], in a study of Peruvian Facebook and Twitter related to service and goods [3], in medical diagnostic [4]. Recently GMDH was used to build classifiers for processing well-known database of domestic services Yelp [5]. In this paper GMDH builds classifiers for cases mentioned above, that is for revealing negative and extreme cases.

The contents of the paper are the following: section 1 is the introduction, section 2 describes DB AskPatient, in section 3 we give short info about GMDH and GMDH Shell, section 4 presents the results of experiments, and section 5 concludes the paper.

II. DATABASE

A. General Description

The AskaPatient database consists of 5 fields, which are rating, reason for taking the medication, side effects, comments, gender, age, duration and date added. As the comments usually reflect patients opinions about the drug, we left only this

[1] askapatient.com
[2] https://gmdhsoftware.com/

field for rating prediction purposes. The dataset we retrieved consists of 48088 comments, with 47983 left after removing comments with text field length less than 5 as not fully subjective. The 5-star rating distribution among comments is presented in Table I.

TABLE I
RATINGS DISTRIBUTION

5*	4*	3*	2*	1*
12093(25%)	9152(19%)	8202(17%)	5713(12%)	12823(27%)

With the new class distribution, the class imbalance marginally increased (Table II), especially for 3-class classification problem where classes 1 and 3 almost 3 times as less as class 2. Therefore the baselines are equal 61% for 2 classes and 48% for 3 classes respectively.

TABLE II
DISTRIBUTION OF DOCUMENTS ON 2 COMBINED CLASSES

Contents	Class 1	Class 2	Class 3
2 classes	18536(39%)	29449(61%)	
3 classes	12823(27%)	23069(48%)	12093(25%)

The distribution of number of words among reviews is presented in Table III. It could be observed that there is just a dozen of reviews with the word count exceeded 200 words.

TABLE III
DESCRIPTIVE STATISTICS ON TERM COUNT IN EACH REVIEW

Min number	Max number	Aver. Number	90% percentile
1	823	54.4	200

For the purpose of calculation simplicity, we have chosen 1000 texts for both classification tasks, preserving the class distribution among texts. This lead to small loss in accuracy of models, however, we were able to carry out more experiments trying different modes of GS tool.

B. Parametrization

We have chosen bag-of-words (BOW) as our primary parametrization technique due to its simplicity putting more attention to tuning parameters of models in GS instead of working with parametrization. Stemmed and lemmatized text was converted to BOW n-grams on character and word level. The dictionary size varies between 100-800 terms. We filtered terms which encounter in more than 75% texts. Parametrization was made on word and character level with n-grams of size 1-3 which overall showed themselves better than single terms. The obtained vectors were normalized using l2-norm.

III. CLASSIFIERS

A. GMDH Shell

GMDH Shell is a well-known tool for the following applications:

- time series prognosis (extrapolation),
- function presentation (approximation),
- object classification

including extended possibilities for visualization of results. GMDH-Shell employs the technique of GMDH. At present GMDH-Shell includes 2 classical algorithms with their modifications: Combinatorial GMDH, GMDH-type neural networks. In our research we use the classification option. For this GMDH-Shell uses ne-vs-All method [6], which reduces multi-class classification to binary classification. Each binary classifier is presented here in the form of an equation of dividing surface. Inductive modeling just allows to find the equation of optimal complexity in n-dimensional space of linguistic variables.

One can download the trial version of GMDH-Shell and test it using his/her own data. Universities have the possibility to purchase this product free of charge for teaching purposes.

B. Quality of Classification

For correctly measuring the quality of our model with unbalanced data we use weighted F-score by calculating metrics for each label, and finding their average, weighted by support (the number of true instances for each label). We report accuracy score as well. We hold-out 20% of data as test set. 2-fold cross-validation is used to choose a model.

IV. EXPERIMENTS

A. Options for GMDH Shell

For the experiments we used GMDH-Shell mentioned above. It includes the following possibilities for preprocessing:

- data normalization to a given interval, e.g. [-1.0,1.0] or [0.0, 1.0];
- data transformation with various functions such as square root, cubic root or arctg to suppress or to strengthen small and large values;
- balancing classes using copying for small classes.

In the process of modeling a user can do the following:

- to select one of GMDH-based algorithms, which are combinatorial GMDH, mixed and forward selection, GMDH-type neural networks,
- to limit the total model complexity,
- to assign the form of elements as quasi-linear, quadratic,
- to define the external criterion.

Speaking about post-processing we mean both various form of visualization for result presentation and the procedure of ensembling. The latter is averaging a set of the best models selected by GMDH-Shell. The number of models to be averaged is assigned by a user.

Overall, the investigated parameters of GS are presented in Table IV

Here: Sq means squares, the model includes lineal, pairwise and square members. Rank means the number of features to consider, which keeps some number of most important variables according to the selected ranking algorithm and could lead to dramatic increase in model complexity if pairwise

TABLE IV
OPTIONS FOR GS TUNING

Balance	Ensemble	Form	Complexity	Rank
yes/no	yes/no	lin/sq	20-200	20-300

and square members would be included; this number of final parameters could be reduced by selecting model complexity value.

B. Preliminary Testing

Tuning parameters we were able to choose best parameter combinations, getting some insights:

- balancing impairs results quality;
- data transformation to different forms did not lead to performance increase;
- ensembling in general lead to better results;
- for 2-class classification problem character n-grams perform much better than word n-grams, the opposite is for 3-class classification task;
- less dictionary size for binary problem works better than for 3-class problem;
- complexity of model in about size of dictionary is always best adjustment.

C. Building Classifiers

The result of classification task for 2-class and 3-class are presented in Tables V and VI respectively.

For binary task mixed classifier showed itself better than other classifiers. Overall, the size of ensemble 5 gave better results; its increase impairs accuracy. Small parameter's rank of 20 performed better in this case. In general, character n-grams performed better. The result of 67% F-score outperformed baseline, which is the portion of biggest class by almost 9%.

For 3-class classification again stepwise algorithm has shown itself better. With 64% F-score it outperformed baseline algorithm by 25%.

Overall, stepwise regression algorithms outperformed combinatorial and neural for both tasks and almost any kind of parametrization applied to texts.

TABLE V
THE BEST OPTIONS FOR DIFFERENT ALGORITHMS FROM GS, 2 CLASSES

Methods	Dict.size	Form	Rank	Ensemb.size	F-measure
Combi	150 symb.	lin	20	5	62%
Forward	250 words	lin	300	no	64%
Mixed	150 symb.	sq	20	5	67%
NN	600 words	sq	100	5	64%

TABLE
VI
THE BEST OPTIONS FOR DIFFERENT ALGORITHMS FROM GS, 3 CLASSES

Methods	Dict.size	Form	Rank	Ensemb.size	F-measure
Combi	400 words	lin	30	15	56%
Forward	250 words	sq	20	no	64%
Mixed	250 words	lin	300	no	51%
NN	400 words	lin	20	5	43%

REFERENCES

[1]O. Koshulko, M. Alexandrov, and V. Danilova, "Forecasting Euro/Dollar Rate with Forex News." Springer, Cham, 2014, pp. 148–153.

[2]A. Boldyreva, M. Alexandrov, O. Koshulko, and O. Sobolevskiy, "Internet Queries as a Tool for Analysis of Regional Police Work and Forecast of Crimes in Regions." Springer, Cham, 10 2017, pp. 290–302.

[3]M. Alexandrov, V. Danilova, A. Koshulko, and J. Tejada, "Models for opinion classification of blogs taken from Peruvian Facebook," *Proceedings of the 4th International Conference on Inductive Modeling (ICIM-2013)*, 2013.

[4]O. Kaurova, M. Alexandrov, and O. Koshulko, "Classifiers of Medical Records Presented in Free Text Form (GMDH Shell application)," *International Conference in Inductive Modelling ICIM*, 2013.

[5]M. Alexandrov and G. Skitalinskaya, "Classifiers for Yelp-reviews based on GMDH-algorithms," 2018.

[6]V. S. Stepashko, "Method of Critical Variances as Analytical Tool of Theory of Inductive Modeling," *Journal of Automation and Information Sciences*, vol. 40, no. 3, pp. 4–22, 2008.