FEATURE RANKING VIA EIGENVECTOR CENTRALITY

Ivan Baybuza
Skolkovo Institute of Science and Technology
Ivan.Baybuza@skoltech.ru

Aleksandr Belov
Skolkovo Institute of Science and Technology
Aleksandr.Belov@skoltech.ru

Mariia Kopylova
Skolkovo Institute of Science and Technology
Mariia.Kopylova@skoltech.ru

Miron Kuznetsov
Skolkovo Institute of Science and Technology
Miron.Kuznetsov@skoltech.ru

December 21, 2019

ABSTRACT

The problem of features ranking arises in many practical fields. In this paper, we work with medical data to rank the potential predictors of medical issue called «fibrillation». To do it, we use an Eigenvector Centrality (EC) method, proposed by the work Ranking to Learn: Feature Ranking and Selection via Eigenvector Centrality (2017). The result of this will be delivered to the National Cardiology Center in Moscow and could be reasonable for the future real practice.

1 Introduction

This research is requested by National Cardiology Center in Moscow. The point is that doctors have some scales made by heuristics which they usually use to evaluate the probability of some diagnosis, i.e. "fibrillation". The idea is to try some data analysis methods to understand the importance of different essences by using the history of clinical records. In this work we use a dataset of clinical records provided by National Cardiology Center in Moscow. The main task is to rank this features according of the influence on the target variable "fibrillation". To do it, we use EC method.

2 Task description and data construction

We are provided with big dataset of clinical records from Cardiology Center, including about 167k records of 55k patients. The record usually includes much information about the patient, his/her diagnoses and analysis.

Feature extraction. We extracted more than 35 features from records, including main information (i.e. sex, age), some specific analyses and drugs that were prescribed by a doctor to a patient. The record is a text, which was written by the doctor, so to extract different features we had to use some tricky rules agreed by the specialist from Cardiology Center, i.e. "if "word1" is in the text, then feature "feature1" is True, but if "word2" is also in the text, then False".

Preprocessing. We had to use OneHotEncoding to analyze categorical features, so after preprocessing we produced 88 features.

Also, stemming was used. And finally we had to delete some words and regroup some drugs to have better features

3 Proposed Method by the Paper[1]

The idea of the paper is to create undirected graph $G(V, E)$ based on the features that we have in the initial matrix $X = (x^1, ..., x^n)$ and evaluate the «importance» of each node of this graph. To do this, authors create adjacency matrix A associated with G. In matrix A, as usual, each cell a_{ij} represent a connection between two nodes: $a_{ij} = \phi(x^i, x^j)$,

where x^i and x^j — vector-feature from the initial matrix X. The main idea of this paper is to create the function ϕ, because designing of such function is the core problem in this work. Authors propose such a way to create matrix A:

- Calculate a Fisher criterion for each feature:

$$f_i = \frac{|\mu_{i,1} - \mu_{i,2}|^2}{\sigma_{i,1}^2 + \sigma_{i,2}^2}, \tag{1}$$

  where $\mu_{i,c}$ and $\sigma_{i,c}$ are the mean and variance respectively.
  The idea behind is that the higher f_i, the more discriminative i-feature.

- Calculate mutual information for each feature:

$$m_i = \sum_{z \in x(i)} p(z, y) \log(\frac{p(z, y)}{p(z)p(y)}), \tag{2}$$

  where y is target variable and $p(z, y)$ the joint probability distribution

  Authors use mutual information to obtain a good feature ranking that score high features highly predictive of the class.

- Calculate kernel k:

$$k = (f * m^T), \tag{3}$$

  where f and m are $n \times 1$ column vectors of Fisher criterion and mutual information respectively.

- Calculate a measure, that represent a variance:

$$\Sigma(i, j) = max(\sigma(i), \sigma(j)), \tag{4}$$

  where σ being the standard deviation over a feature.

  The idea before this matrix that it represents the amount of variation of features from average.

- Constructing the adjacency matrix A of the graph G:

$$A = \alpha k + (1 - \alpha)\Sigma \tag{5}$$

  where $\alpha \in (0, 1)$ represents a choice between k and Σ.

The a_{ij} takes into account how distinguishing features i and j are when they are considered together; at the same time, a_{ij} can be considered as the weight of the edge connecting the nodes i and j of the graph.

The next step of the algorithm is to find centrality indicators of each node in the graph G. To do it, they compute eigenvector centrality measure (EC). EC is a vector that is equal to:

$$A^n e = v_0 \tag{6}$$

In other words, authors use Power Method to obtain eigenvector which corresponds to the largest modulo eigenvalue of the matrix A.

In the result, v_0 is a vector which show a rank of each variable from initial matrix X.

4 Results

The EC-FC algorithm proposed in the article was implemented in Python. The algorithm was tested with different initial data.

On the processed dataset the matrix A was constructed, for which using the Power method we found the eigenvector corresponding to the maximum eigenvalue. The Power method for the constructed matrix a converges in several iterations.

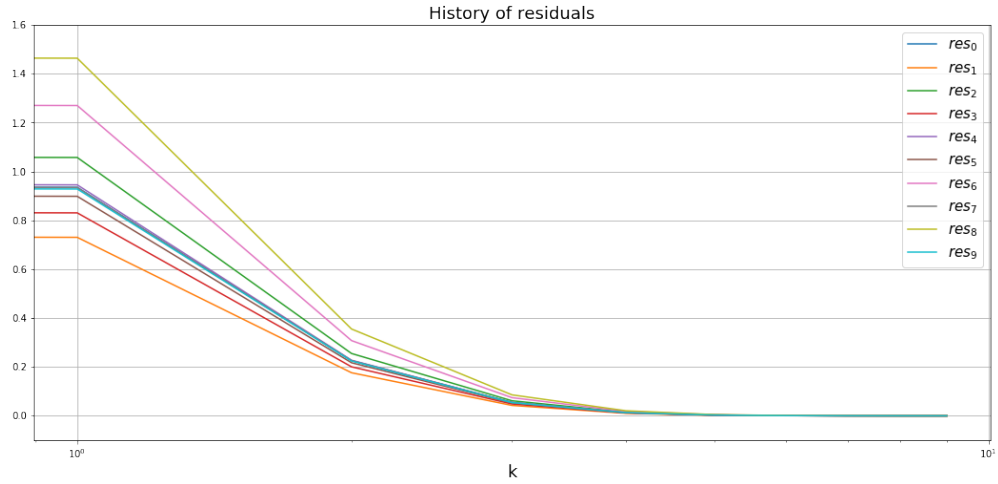

Figure 1: The Power method convergence with different initial states.

With the help of the obtained rank-vector, the most important features were found, which should be studied by specialists.

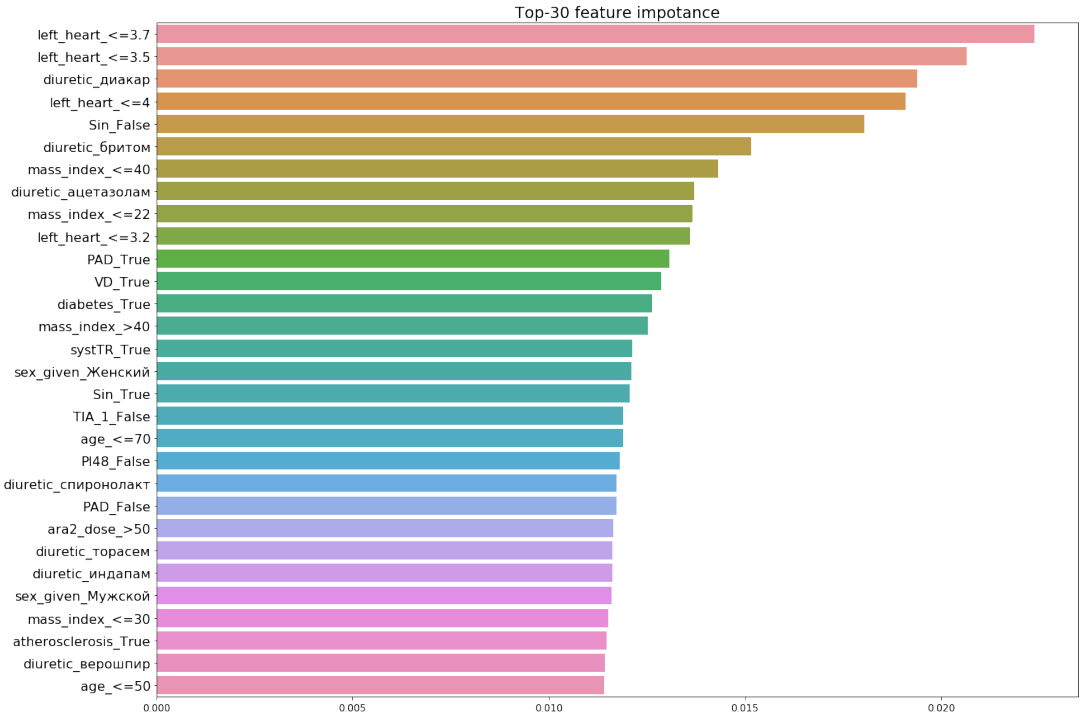

Figure 2: Feature importance.

The main predictors of Atrial fibrillation is:

- Large size of the left heart (> 3.2);
- Large Mass index of a patient (2 degree of obesity and more);
- An absence of Sinus tachycardia;
- A presence of diabetes.

The source codes may be found at:

`https://github.com/albellov/nla_project`

5 Conclusion

In this project we present and implement the idea of the paper: Feature Selection via the Eigenvector centrality measure. The method (supervised) estimates some indicators of centrality identifying the most important features within the graph.

The investigated Eigenvector Centrality method may be applicable to this type of problems and could give some unobvious information. The results should be delivered to the specialists in Cardiology Center to be considered by them.

6 Contribution of team members

Ivan Baybuza:

- coding the method
- presentation
- report

Aleksandr Belov:

- data preprocessing
- coding the method
- report

Mariia Kopylova:

- learning articles with the methods of feature selection
- presentation
- analyze methods

Miron Kuznetsov:

- data and feature extraction
- data preprocessing
- report

References

[1] Source codes: `https://github.com/albellov/nla_project`.

[2] Giorgio Roffo, Simone Melzi, "Ranking to Learn: Feature Ranking and Selection via Eigenvector Centrality", 2017.

[3] Noah Lee, Andrew F. Laine, Jianying Hu, Fei Wang, Jimeng Sun, Shahram Ebadollahi, "Mining electronic medical records to explore the linkage between healthcare resource utilization and disease severity in diabetic patients", 2011

[4] Cheng Guo, Felix Berkhahn, "Entity Embeddings of Categorical Variables", 2016.