Application of machine learning algorithms for the study of Galicia population's social characteristics in the interwar period (1919-1939)

Taras Ustyianovych[1], Nataliia Khymytsia[2]

1. Social Communication and Information Science Department, Lviv Polytechnic National University, UKRAINE, Lviv, S. Bandery street 12, E-mail: ustyk5@gmail.com.

2. Social Communication and Information Science Department, Lviv Polytechnic National University, UKRAINE, Lviv, S. Bandery street 12, E-mail: nhymytsa@gmail.com

Abstract – The machine learning algorithms application on the cliometric data of Galicia's history in the interwar period (1919-1939) was tested. The most accurate algorithms were defined, the «Ukrainian» class was determined in the dataset.

Keywords – demographic process, Galicia, interwar period, data, data processing, cliometrics, machine learning, historical process.

Introduction

New trends in the development of modern Ukrainian humanities more clearly demonstrate the harmonious combination of interdisciplinary and new methodological approach usage. In this sense, the value of those scientific discoveries, which appeal to quantum (quantitative analysis of historical sources) and try to find out hidden patterns and models of historical process functioning, is increasing. On the other hand, the spreading of such phenomenon as Big Data force historians to pay special attention to automated processes of mass historical sources processing, apply modern algorithms, mathematical and statistical modelling methods.

The necessity of comprehensive usage of the new future predicting and the past reconstruction and modelling techniques – machine learning – appears today. That is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. [1] It is especially relevant in historical science because there are a lot of lost information that can be restored using this technique. Efficient machine learning usage can give the probability of some factor with high accuracy (up to 100%), provided that it is correctly used and implemented.

Analysis of researches and publications

The tendency to increasing of research papers about application of machine learning methods on the basis of various data and ways of its development is observed today. M. I. Jordan and T. M. Mitchell explored the practical machine learning usage in many fields of human activity, namely, health care, manufacturing, education, financial modeling, policing, and marketing. Today these fields actively use this method for financial crises, viruses and trends of the modern market forecasting [2]. However, the possibility of project implementation and researching of machine learning process in historical science is not mentioned in the article of these authors. Bottou, Léon, Frank E. Curtis, and Jorge Nocedal described methods of Big Data processing using machine learning. The methods and ways of large-scale data processing optimization were specified [3]. The tendencies to historical art and architecture sights recognition using machine- and deep learning methods are spreading. It allows the computer to recognize objects on photo-data and identify them, find out past monument patterns and styles. Such approach was researched in Andrew Zisserman work [4]. The author emphasized on the relevance of machine learning methods usage for the memorabilia and history of culture studies.

A common area is the processing of mass sources using multidimensional statistical analysis methods. For example, the Russian scientist S.G.Kaschenko, using the variational series analysis methods, sampling and correlation analysis methods, investigated mass sources (statutes, redemption acts, etc.) for the reform of the St. Petersburg, Pskov and Novgorod provinces in 1861 [5, 6]. Belarusian historian A.G. Kochanovsky, on the basis of multidimensional statistical analysis methods of the published results of the census of 1897, land censuses data of 1877 and 1905, military censuses of 1888 and 1900, and other statistical sources, carried out a study of the socio-class structure of the Belarus population in the second half of the XIX century. The researcher identified 20 quantitative features that capture the main social groups proportion in Belarus, and analyzed the relationships between them, using correlation and regression methods. Interpretation of the correlation coefficients matrix allowed him to conclude that the social mobility of the Belarus rural population at the end of the XIX century was weak, and the insignificant capitalization process had influenced on the village social structure [7].

Another approach in application of quantitative methods and information technology is the mass sources processing, using relational and full-text databases. The databases and mass historical sources were managed and processed by Belarusian researchers V.E. Kudryashov and O.L. Lipnitsky. E.A. Pavlova, on the basis of the lists of military formations personnel (1942-1944) created a database about Minsk and Brest regions partizans, which was processed, using Microsoft Excel and DBMS (database management system) Microsoft Access. On the basis of the analysis, the national composition quantitative indicators of partisan detachments were defined, their educational level was determined [8].

New approaches to socio-humanitarian researches, arguing the importance and necessity of machine learning algorithms application for cloimetric data processing are presented in S.Golub, N. Khymytsia[9-11], T.Ustyianovych[12] scientific works. The authors explore historical process modelling questions, application of Big data phenomenon, describe mathematical modelling methods, their advantages, prospects of Big data processing in the domestic historiography.

Application of machine learning algorithms for cliometric data processing

A lot of minorities, who lived in Galicia in interwar period (1919-1939), were singled out by researchers in a plenty of historiographic works. However, persons' nationality, who lived in Western Ukraine, was not mentioned in personal files and other papers, because of the Polish policy. It was caused by the desire to polonize the national minorities much faster, including Ukrainians.

Today the Ukrainians identification problem among other Galician residents of the 1920-1930 is still topical. Application of machine learning algorithms is proposed in order to get accurate probability of persons belonging to a certain state, social status, nationality, to find continuous variables (number of children, salary, etc.), to restore lost data. It allows to predict and model historical process, find out correlationships, regularities between data. These methods will promote historiography popularization, accumulation of new knowledge and information, lost data restore. Prediction and application of algorithms will be done using the existing data, which is quite a lot, so that we can commit the «training» process. In this case the quantitative data characteristics have a great impact on the process – the more data, the more accurate prediction can be got.

In order to get the data, the archival documents of the Central State Historical Archives of Ukraine in Lviv were processed, in particular such funds: «Ruthenian People's institute «Narodnii Dim»», Lviv» [13], «Ukrainian rural workers' socialist union («Sel-rob»), Lviv» [14], «Ukrainian parliamentary representation in the Polish Sejm and Senate, Warsaw [15], «Rutherian rural organization, Lviv» [16]. The collected consolidated data is related to Galicia in the interw period

(1919-1939). Due to the existence of many Ukrainian organizations at that time, the probable Ukrainian and Pole nationality of some persons was defined. Namely, membership in Ukrainian organizations, as a rule, meant belonging to a given nation. The collected data contains 15 characteristics, the most important among them are these: likely nationality, sex, social and professional status, residence. In order to define these features, historical sources and documents that contain cliometric data has been explored. The main goal is to predict persons nationality on the basis of the collected data.

Data analysis and visualization on the dataset has been done. It helps find out relationships between data. For instance, in the sample, persons with the Pole nationality turned out to be older than Ukrainians (Fig.1). The median age both for Ukrainians (42 years) and for Poles (46 years) are shown in the boxplot. The highest educational level coefficient among nationality classes of the sample (Fig.2.) is defined as well.

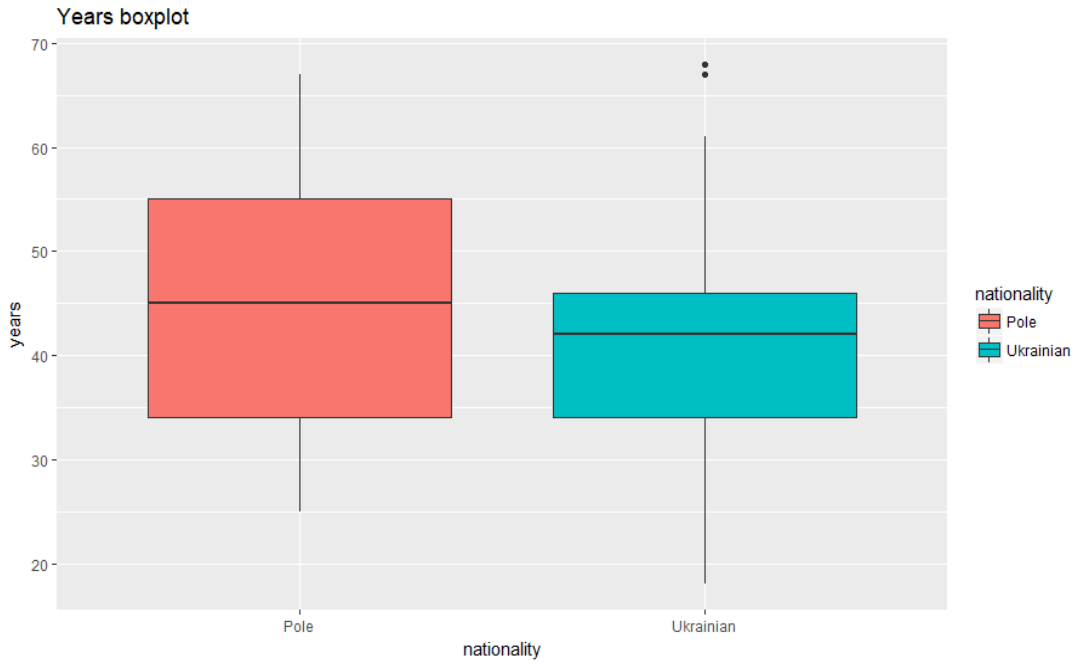

Fig.1. Distribution of that time population age group.

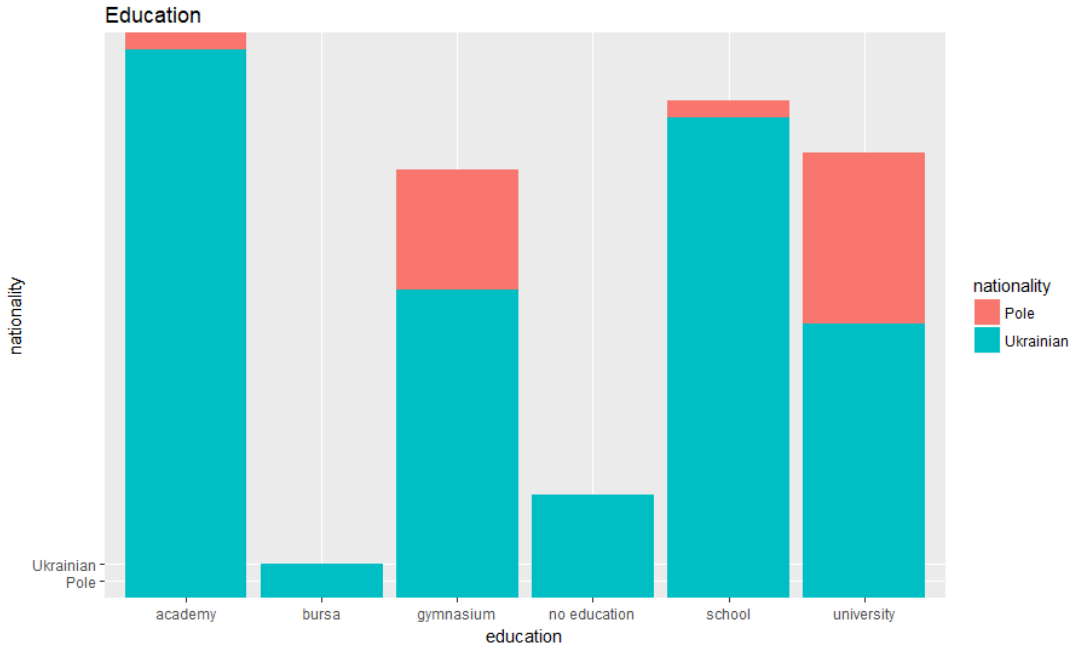

Fig.2. The highest educational level in the two sample classes.

The data is subject to normal distribution. The application of machine learning algorithms was conducted after data analysis. We divided the data into training and test datasets. In the Table 1 the algorithms name, the accuracy of the results based on the training and test samples are specified.

Table 1. Relative error of machine learning algorithms on a dataset.

Algorithms name	Relative error on the training dataset.	Relative error on the test dataset.
Decision tree	18% (0.1833)	10% (0.1)
Logistic Regression	13% (0.1333333)	20% (0.2)
K-nn (k-nearest neighbors)	23% (0.233)	10% (0.1)
Naive Bayes	15% (0.15)	20% (0.2)
Naïve Bayes (improved)	13% (0.1333333)	20% (0.2)

The most accurate result we have got, using the logistic regression and Naïve Bayes (improved) algorithms. The result on the training dataset has more significance, than on the test dataset due to the fact that 80% of all data is belonging to the training sample. The most important attributes are educational level, age, type of the residence, knowledge of languages (Ukrainian and Polish). Every branch consists of the probable class, confidence of belonging to it, and the percentage of sample, which is in the class. The left branches (yes) respond to the class «Pole», whereas right branches (no) respond to the class «Ukrainian». But accuracy of this algorithms has not turned out to be high enough in comparison to others. The reason is that more data with significant characteristics. which will give insights and information for improving the tree, making it clear, is needed to conduct good training process for the sample.

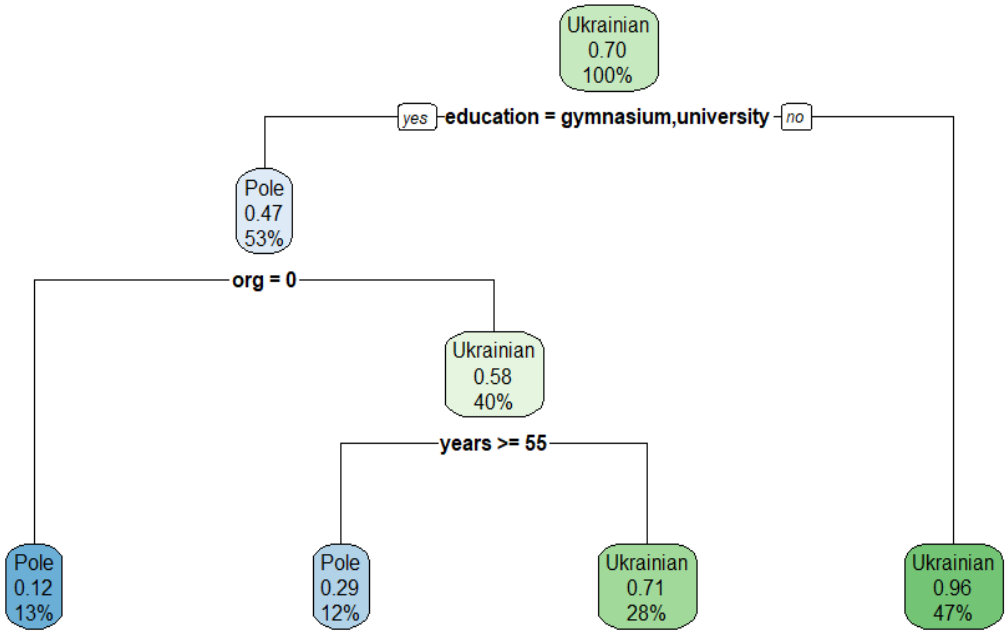

Рис.3. Decision tree based on the historical data.

The least accurate turned out to be K-nn algorithm. The reason is that the algorithm works well only on numerical data, which in this data set has not had a significance in the classification process (age of inhabitants, number of children). Algorithm operation method is the following: an object is assigned to that class, which is the most distributed among k number of the element's neighbor, which classes has been previously defined. The cross-validation process, which allowed to define to best number of k within a specific range, was applied to the algorithm implementation. K-nn algorithm

can be improved using significant numerical data. The best k number is equal to 11 (Fig.4.). Decision making process occurred on the basis of the nearest 11 numerical values analyses on the graph with the axes abscissa and ordinates X and Y.

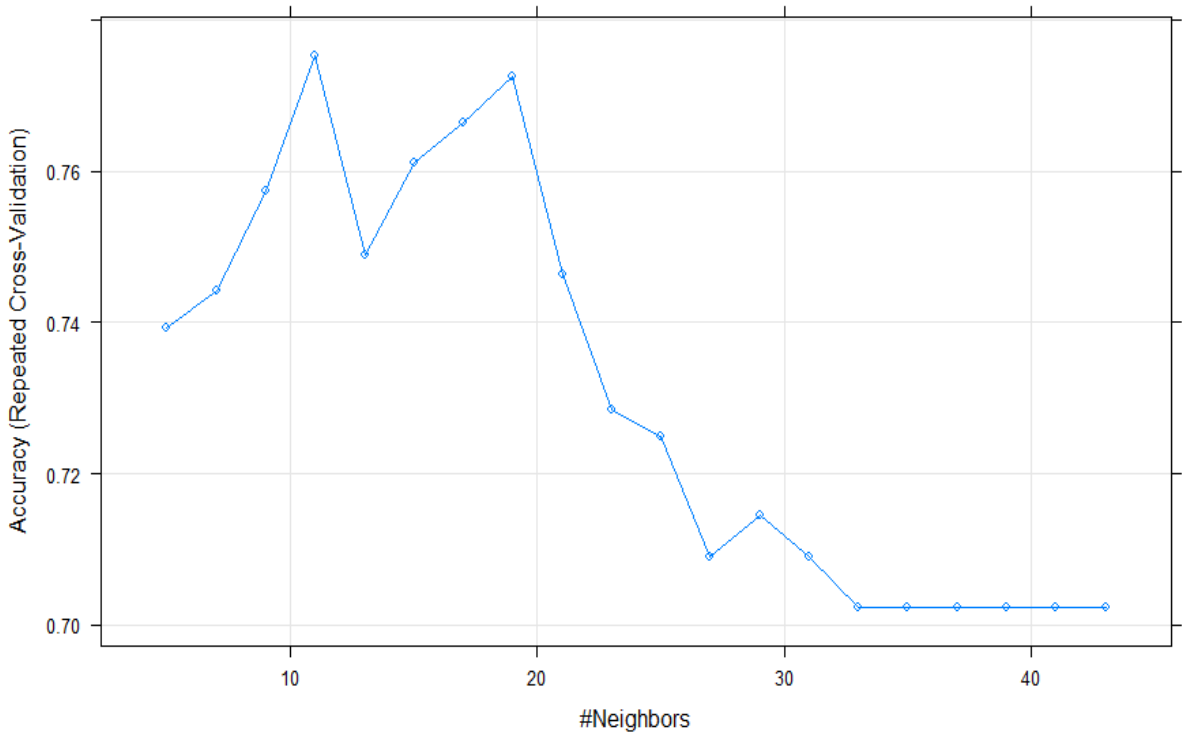

Рис.4. The k-value in the process of applying the algorithm.

Logistic regression and Naïve Bayes (improved) algorithms gave the best result. This is due to the fact that almost all characteristics of the data were taken into account (up to 15), and on the basis of which probable class nationality was modelled.

The main disadvantages can be inaccurate data, lacking of attributes, non-differentiable discontinuous loss functions etc. The necessity of a lot of data is significant, it has great impact on the accuracy and other features of decision process. Algorithm retraining problem (reapplication) can occur due to the small amount of data for training process of some algorithms, the absence of significant numerical values that would impact on the classification process. The main problem is that the historical process in unpredictable. Algorithm can give accuracy up to 100%, however, the actual course of events can significantly vary. Mathematic methods give only the probability of a certain event or process, 100% guarantee that it has happened in a such way will be either unknown or almost uncalculated.

Conclusion

The demographic and socio-economic processes of Galicia in the interwar period (1919-1939) were researched. The main characteristic and relationships between cliometric data were explored. Machine learning algorithms were practically applied to the dataset; the most accurate algorithms were discovered. Cliometric information processing methods, their accuracy optimization methods, namely: increasing the amount of data, adding new features, data consolidation, were defined.

The research can be used for historical process modelling, prediction of the past and course of historical events, in particular for the lost historical data restoration on Galician history of the interwar period, prediction the probable belonging to certain nationality, class or status.

References

[1] R.; Bennett, Forrest H.; Andre, David; Keane, Martin A. (1996). *Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming*. Artificial Intelligence in Design '96. Springer, Dordrecht. pp. 151–170.

[2] M. I. Jordan, T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science* 17 Jul 2015, Vol. 349, Issue 6245, pp. 255-260.

[3] Bottou, Léon, Frank E. Curtis, and Jorge Nocedal. "Optimization methods for large-scale machine learning." *SIAM Review* 60.2 (2018): 223-311.

[4] Zisserman Andrew. *Visual recognition in art using machine learning*. Diss. University of Oxford, 2017.

[5] S. Kashchenko. Reform of February 19, 1861 in North-West Russia: (Quantitative analysis of mass sources): [Proc. allowance]. - M .: Mosgorahiv, 1995.

[6] S. Kashchenko. The Abolition of Serfdom in the Pskov Province: The Experience of Computer Analysis of the Conditions for Implementing Peasant Reform February 19, 1861 - St. Petersburg: St. Petersburg State University Publishing House, 1996.

[7] Kochanowski A.G. Regression analysis in the study of the social history of Belarus in the late XIX century. / A.G. // Kochanowski. Theoretical and methodological problems of historical knowledge. Materials of international scientific conference. - Minsk: BSU RIVSH 2000.

[8] Methods of quantitative analysis of the texts of narrative sources.- M., 2003.

[9] Golyb Sergey, Khymytsya Natalia. The use of multi-level modeling in the cliometric studies process / S. Golyb, N.Khymytsya // Proceedings XIII-th International Conference “Modern Problems of Radio Engineering, Telecommunications and Computer Science” (TCSET'2016) : Lviv, February 23-26, 2016. – Lviv-Slavske, Ukraine. - C. 733-735.

[10] Khymytsia N. Analysis of Computer-based Methods for Processing Historical Information / N. Khymytsia, S. Lisina, O. Morushko, P. Zhezhnych // Advances in Intelligent Systems and Computing: Selected Papers from the International Conference on Computer Science and Information Technologies, CSIT 2017, September 5-9 Lviv, Ukraine, Shakhovska N. (Ed.). – Springer International Publishing: 2017.– Series Volume 1.– p. 365- 368.

[11] Golub S., Khymytsia N. Clinodynamic monitoring using processes of clusterization of historical periods / S. Golub, N. Khymytsia // Proceedings of the 7th International Scientific Conference "Information, Communication, Society": May 17-19, 2018. - Lviv: Lviv Polytechnic Publishing House, 2018. - P.291-292.

[12] Khymytsia N., Ustiyanovich T. Application of Big Data in Historical Science Proceedings / Nataliia Khymytsia, Taras Ustiyanovich // 7th International Academic Conference of Young Scientists “Humanities and Social Sciences 2017” (HSS-2017). — Lviv: Lviv Polytechnic Publishing House, 2017. — Electronic edition on CD-ROM., P. 368-370.

[13] F. 130. Ruthenian People’s institute «Narodnii Dim»», Lviv, affairs 984, 1848-1939. Description. Ru., «yazicie», ukr. language.

[14] F. 351 Ukrainian rural workers' socialist union («Sel-rob»), Lviv, affairs 153, 1926-1932 pp. Description. Ukrainian language.

[15] F. 392. Ukrainian parliamentary representation in the Polish Sejm and Senate, Warsaw, affaris 66, 1919-1938 pp. Description: Ukrainian, Polish, German languages.

[16] F. 394. Rutherian rural organization, Lviv, affairs 13, 1928-1935 pp. Descriptions: Polish, Russian, Ukrainian languages.