Comparison of Machine Learning Methods for Life Trajectory Analysis in Demography

Anna Muratova[1(✉)] , Ekaterina Mitrofanova[1] , and Robiul Islam[1,2]

[1] National Research University Higher School of Economics, Moscow, Russia
{amuratova,emitrofanova}@hse.ru, r_islam@ieee.org
[2] Innopolis University, Innopolis, Russia

Abstract. Nowadays there are representative volumes of demographic data which are the sources for extraction of demographic sequences that can be further analysed and interpreted by domain experts. Since traditional statistical methods cannot face the emerging needs of demography, we used modern methods of pattern mining and machine learning to achieve better results. In particular, our collaborators, the demographers, are interested in two main problems: prediction of the next event in a personal life trajectory and finding interesting patterns in terms of demographic events for the gender feature. The main goal of this paper is to compare different methods by accuracy for these tasks. We have considered interpretable methods such as decision trees and semi- and non-interpretable methods, such as the SVM method with custom kernels and neural networks. The best accuracy results are obtained with a two-channel convolutional neural network. All the acquired results and the found patterns are passed to the demographers for further investigation.

Keywords: Sequence mining · Demographics · Decision trees · SVM · Neural networks · Classification

1 Introduction

Nowadays researchers from different countries have access to a large amount of demographic data, which usually consists of important events, their sequences, and also different features of people, for example, gender, generation, and location, among others. Demographers investigate the relationship between events and identify frequently occurring sequences of events in the life trajectories of people [19]. This helps researchers to understand how the demographic behaviour of different generations in different countries has changed and also allows researchers to track changes in how people prioritise family and work, and compare the stages of growing up of men and women [1,2,10].

There are different works in this sphere. For example, in [3] authors used decision trees to find patterns that discern the demographic behaviour of people

in Italy and Austria, their pathways to adulthood. In [4], the authors discover association rules to detect frequent patterns that show significant variance over different birth cohorts.

Commonly, demographers rely on statistics, but it has some limitations and does not allow for sophisticated sequence analysis. The study of demographic sequences using data mining methods allows for extracting more information, as well as identifying and interpreting interesting dependencies in the data.

Demographers are interested in two main tasks. The first one is the prediction of the next event in personal life trajectories, based on the previous events in their life and different features (for example, gender, generation, location etc.) [20]. The second task is to find the dependence of events on the gender feature, that is, whether the behaviour of men differs from that of women in terms of events and other features [11]. Let us call this task gender prediction. So, the former problem resembles such fundamental problems in machine learning like the next symbol prediction in an input sequence [23], while the latter can be easily recast as a supervised pattern mining problem [27].

The main goal of this paper is to compare different methods, both interpretable and not-interpretable by accuracy. Interpretable methods are good for further interpreting and working with results. Not-interpretable methods allow us to find how accurate the prediction is and to find the best method for these types of data. Among our interpretable methods are decision trees with different event encoding schemes (binary, pairwise, time encoding, and different combinations of these encodings). Among our non-interpretable methods are special kernel variants in the SVM method (ACS, LCS, CP without discontinuities) and neural networks (SimpleRNN, LSTM [13], GRU [7], Convolutional). For all of the methods, we used our scripts and modern machine learning libraries in Python.

As a result of the work, we obtained patterns that are of interest to demographers for further study and interpretation. The best method by accuracy was also found, which is also important for event prediction. Among all of the considered methods, the best method in terms of accuracy is a two-channel Convolutional Neural Network (CNN). The previous results, mainly devoted to pattern mining and rule-based techniques, can be found in our works [11,12,14,16]. Since the considered dataset is unique and was extensively studied only by demographers in a rather descriptive manner, we compare our results on the level of machine techniques' performance and provide the demographers with interesting behavioral patterns and further suggestions on the methods' applicability.

The paper is organised as follows. In Sect. 2, we describe our demographic data. Section 3 contains results obtained with decision trees for the prediction of the next event in a personal life trajectory (or an individual life course) and events that distinguish men and women. Section 4 presents the results using special kernel variants in the SVM method, and Sect. 5 is devoted to the Neural Networks Results (SimpleRNN, LSTM, GRU, CNN). In Sect. 6, a comparison of different classification methods is made, and Sect. 7 concludes the paper.

2   Data Description

The data for the work was obtained from the Research and Educational Group for Fertility, Family Formation, and Dissolution. We used the three-wave panel of the Russian part of the Generations and Gender Survey (GGS), which took place in 2004, 2007, and 2011. After cleaning and balancing the data, it contained the results of 6,626 respondents (3,314 men and 3,312 women). In the database, the dates of birth and the dates of first significant events in respondents' live courses are indicated, such as completion of education, first work, separation from parents, partnership, marriage, the breakup of the first partnership, divorce after the first marriage and birth of a child. We also indicated different personal sociodemographic characteristics for each respondent, such as type of education (general, higher, professional), location (city, town, village), religion (religious person or not), frequency of church attendance (once a week, several times per week, minimum once a month, several times in a year, or never), generation (1930–1969 or 1970–1986), and gender (male or female).

A small excerpt of sequences of demographic events based on the life trajectories of five people is shown in Table 1. Events are arranged in the order of their occurrence. It can be seen that for the first respondent, the events work and separation from parents happened at the same time. These events are indicated in curly braces. Note that sequence 3 contains no events so far due to possibly young age (likewise, generation) of the respondent.

Table 1. An excerpt of life trajectories from a demographic sequence database.

Sequence ID	Demographic sequence
1	$\langle \{work, separation\}, partner, marriage, child \rangle$
2	$\langle work, separation \rangle$
3	$\langle \rangle$
4	$\langle work, partner \rangle$
5	$\langle separation, work, marriage, child \rangle$

3   Decision Trees

3.1   Typical Patterns that Distinguish Men and Women

Let us find different patterns for men and women (gender prediction task) based on previous events in their lives and other sociodemographic features.

We tested decision trees with different event encodings [3]:

1. Binary encoding (or BE), where value "1" means that event has happened in a personal life trajectory and "0" if the event has not happened yet.
2. Time encoding (or TE) with the age in months of when the event happened.

3. Pairwise encoding (or PE), which consists of event pairs coding to mark the type of mutual dependency. If the first event occurred before the second or the second event did not happen yet, then the pair of events is encoded with the symbol"<", if vice versa, then ">", if the events are simultaneous, then "=" and if none of the events has happened yet, then "n".

In addition to these encodings, we also used their different combinations.

Table 2 presents the results of accuracy for decision trees with different event encodings for the gender prediction task. It can be seen that time-based encoding is better than binary and pairwise. Also adding binary or pairwise encodings to time-based encoding slightly lower the accuracy than just using time-based encoding alone. The best accuracy result is obtained for all of the three encodings together with accuracy 0.692.

Table 2. Classification accuracy of different encoding schemes for gender prediction.

Encoding scheme	Classification accuracy
Binary	0.582
Time-based	0.676
Pairwise	0.590
Binary and time-based	0.665
Binary and pairwise	0.592
Time-based and pairwise	0.674
BE, TE, and PE	**0.692**

Let us consider this decision tree since it gives the best accuracy. Table 3 shows the difference between Russian men and women in demographic and socioeconomic spheres. The higher "speed" of reproductive events occurrence in women's life courses indicates the pressure of the "reproductive clocks" over women. During the Soviet era, women who got their first births after the age of 25 were stigmatised and called "older parturient". Men took more time to obtain all the events because they had such obligatory events as military service which made them delaying some other significant events of life.

Table 3. Patterns from the decision tree for men and women for all generations.

Gender	Rules for event sequences	Confidence
Men	First work after 21, marriage after 29.5	70.2%
	Child after 22.9, marriage in 23.9–29.5	69.3%
	Marriage in 23.3–23.9, child after 24.7	67%
Women	Marriage before 20.6 years	77.1%
	First work in 17.7–21, marriage after 29.5	62.8%
	Child before 22.9, marriage in 23.8–29.5	61.2%

Also, we obtained interesting patterns for men and women based only on different features. The main feature which shows the highest difference in patterns is religion. With the probability of 65.9 % if the person is not religious, it is more likely men. Among women the highest number of religious ones is in 1945–1949 with higher education, the probability is 67.6%. Among men the highest number of religious ones is in 1975–1979 with general education, the probability is 65.2%. Also in 1930–1954—the period which embraces Industrialization and the Second World War when the Soviet government declared atheism policy—there are more religious women, however, in 1980–1984—the period preceding ideologic liberalization in 1988—more religious people are among men.

3.2 Prediction of the Next Event in Personal Life Trajectories

Now let us look at the features and events to predict the next event in an individual life course. As in Sect. 3.1, we consider three types of encoding: binary, time-based, and pairwise, as well as all kinds of their combinations.

Table 4 presents results of accuracy for decision trees with different event encodings for the next event prediction task. It can be seen that the best classification accuracy of 0.878 is obtained with the binary encoding scheme. Also adding a pairwise encoding to time-based encoding as well as adding binary encoding to pairwise encoding slightly improves their own accuracy result. The time-based encoding scheme is the lowest by accuracy.

Table 4. Classification accuracy of different encoding schemes for the next life-course event prediction.

Encoding scheme	Classification accuracy
Binary	**0.878**
Time-based	0.359
Pairwise	0.701
Binary and time-based	0.746
Binary and pairwise	0.844
Time-based and pairwise	0.496
BE, TE, and PE	0.750

Let us consider decision tree with binary encoding, since it gives the best accuracy. In Table 5, several patterns, i.e. classification rules, from this decision tree for all generations, both men and women are presented. Note that events in the rules' premises are not indicated in their real order.

From the table, we can see that people tend to find first work after completion of education (event work after education with probability 98.2% vs. event education before work with probability 90.3%).

Table 5. Patterns from the decision tree for the next event's prediction task.

Premise: path in the tree	Conclusion: leaf	Confidence
Education, separation from parents, child birth	Work	98.9%
Work	Education	98.2%
Education, separation from parents, marriage	Child birth	95.7%
Education, child birth	Separation from parents	93.9%
Education	Work	90.3%
Child birth, education, marriage, partner, separation from parents, work	Breakup	78.1%
Work, education	Separation from parents	76.7%
Work, separation from parents, education, marriage, child birth	Divorce	72.9%

Also based on the features only, we obtained that the main feature which shows the highest difference for the last event in a person's life course is education type (general, higher, professional).

4   Using Customized Kernels in SVM

4.1   Classification by Sequences, Features, and the Weighted Sum of Their Probabilities

Using special kernel functions in the SVM method for sequence classification is discussed in [15]. In paper [16] the authors used the following sequence similarity measures: CP (common prefixes), ACS (all common subsequences), and LCS (longest common subsequence). Since demographers are interested in sequences of events without discontinuities (gaps) we derived new formulas, which are the modifications of the original ones [9]. Sequences of events without gaps preserve the right order in which events happened in a person's life.

Let s and t be given sequences and LCSP be the longest common sequence prefix, then similarity measure common prefixes without discontinuities can be calculated as:

$$sim_{CP}(s,t) = \frac{|LCSP(s,t)|}{\max\left(|s|,|t|\right)} \tag{1}$$

Let LCS be the longest common sequence, then similarity measure based on the longest common subsequence without discontinuities is calculated as:

$$sim_{LCS}(s,t) = \frac{|LCS(s,t)|}{\max\left(|s|,|t|\right)} \tag{2}$$

If k is the length of common subsequence, $\Phi(s,t,k)$ is the number of common subsequences of s and t without discontinuities of length k, then similarity measure all common subsequences without discontinuities is calculated as:

$$sim_{ACS}(s,t) = \frac{2\Sigma_{k \leq l}\Phi(s,t,k)}{l(l+1)}, \text{ where} \tag{3}$$

$$l = max(|s|,|t|) .$$

Let us consider special kernel variants in the SVM method (Support Vector Machines). We will combine two methods of classification: by sequences using special kernel functions based on sequence similarity measures without discontinuity CP, ACS, and LCS and by features using the SVM method with default parameters (the kernel function is RBF). This can be done using the probabilities of referring to a certain class (let us consider the case with two classes, men and women), calculated by the SVM method.

Having obtained the probability values for each method, we can perform classification based on the weighted sum of the probabilities of the two methods. Since the methods give different classification accuracies, the final probability of assigning an object $\mathbf{x} = (f,s)$ to a class is calculated by the formula:

$$P(class|\mathbf{x}) = \frac{A_s \cdot P_s(class|\mathbf{x}) + A_f \cdot P_f(class|\mathbf{x})}{A_s + A_f} \tag{4}$$

A_s is the accuracy by sequences, A_f is the accuracy by features, P_s is the conditional probability by sequences, and P_f is the conditional probability by features.

That formula takes into account the accuracy of the method for the final probability calculation. The probability calculated by each method will be included in the final result with a coefficient equal to the method accuracy. Note that probabilistic calibration of classifiers and sampling techniques for imbalanced classes are often used in machine learning applications in various domains, for example, in medicine [24].

The results for pattern prediction that distinguish men and women and for the next event prediction are presented in Table 6. We can see that the highest accuracy for the gender prediction is 0.678, which is obtained with kernel function CP using the weighted sum of probabilities. The weighted sum of probabilities

Table 6. Classification by sequences, features, and weighted sum of their probabilities (for gender and next event prediction).

Gender prediction	CP	ACS	LCS
Accuracy of sequence classification (SVM, custom kernel functions: CP, ACS, LCS)	0.648	0.659	0.490
Accuracy of classification by features (SVM default – RBF)	0.615	0.615	0.615
Accuracy of classification by weighted sum of probabilities	**0.678**	0.670	0.612
Next event prediction	CP	ACS	LCS
Accuracy of sequence classification (SVM, custom kernel functions: CP, ACS, LCS)	0.839	**0.911**	0.807
Accuracy of classification by features (SVM default – RBF)	0.367	0.367	0.367

for the case of next event prediction gives lower results due to the small accuracy of classification by features. The best result for this case is obtained with kernel function ACS with the accuracy 0.911.

4.2 Classification by Features and Categorical Encoding of Sequences

Another possible method of classification by sequences is by transforming each sequence to the feature. After that, existing methods of classification by features could be used.

There are 6,626 sequences in our dataset, where 1,228 sequences are unique. We consider the sequence as a feature taking 1,228 different values. Each unique sequence was encoded as an integer. Then scikit-learn SVM module with default parameters was used for classification.

We obtained the accuracy of 0.716 for the patterns that distinguish men and women and the accuracy of 0.775 for the next event prediction.

5   Neural Network Models

We performed classification using neural networks software Keras[1] with Tensorflow as backend. The simulation was performed on the GPU. Recurrent Neural Network (RNN) allows us to reveal regularities in sequences. Three types of recurrent layers were compared in Keras: SimpleRNN, GRU, and LSTM. All types of recurrent layers showed good performance with a little less accuracy for SimpleRNN.

Table 7. Neural networks performance for gender and next event prediction.

Network	Gender prediction			Next event prediction		
	Accuracy	Variance	Time, s	Accuracy	Variance	Time, s
GRU RNN	0.760	0.000090	14.11	0.930	0.000055	11.01
LSTM RNN	0.759	0.000123	14.54	0.930	0.000042	11.57
SimpleRNN	0.755	0.000135	19.81	0.927	0.000049	13.50
Convolutional	0.762	0.000193	11.31	0.931	0.000041	8.17

For the network with recurrent layer, accuracy 0.760 was obtained for the patterns that distinguish men and women, and 0.930 for the next event prediction.

Also, a two-channel model with a convolutional layer was implemented[2]. A 1D convolutional layer was used for sequences and dense layers for features. We

---

[1] https://keras.io.
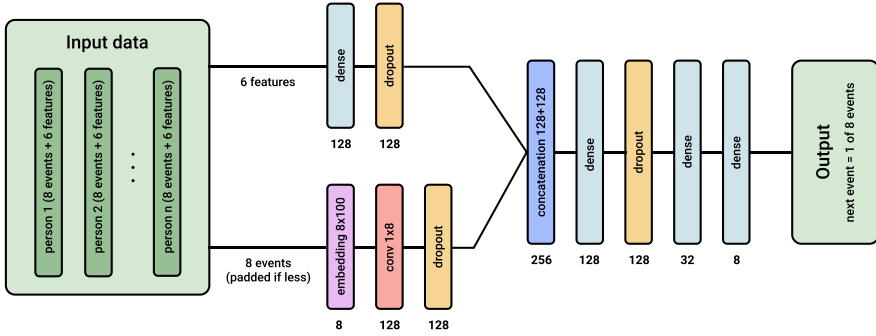[2] https://github.com/anya-m/2CNNSeqDem/.

Fig. 1. The two-channel network structure for the next event prediction problem.

obtained the accuracy 0.762 for the patterns that distinguish men and women and the accuracy 0.931 for the next event prediction. We can see that all implemented layers give high accuracy. For the next event prediction, the accuracy is much higher than for the gender prediction in all cases.

Note that we employ 80-to-20 random cross-validation splits with 10 repetitions and report the averaged results.

The structure of the two-channel network layer for the next event prediction is shown in Fig. 1.

6   Comparison of Methods

The accuracies of all the methods for both tasks, gender prediction, and the next event prediction, are presented in Tables 7 and 8.

Table 8. Comparison of the methods.

Method	Gender prediction	Next event prediction
Decision trees	0.692	0.878
	(BE, TE and PE)	(BE)
SVM with custom kernel function	0.678	0.911
	(CP)	(ACS)
SVM with sequences transformed into features	0.716	0.775
Recurrent neural network	0.760	0.930
Convolutional neural network	**0.762**	**0.931**

From the table, it can be seen that the highest classification accuracy for both cases is obtained with convolutional neural networks, which means that it

is an optimal method (among the considered) for these tasks. Also, the accuracy for the prediction of the next event in personal life trajectories is higher in all of the methods than the accuracy for gender prediction.

7   Conclusion

This work contains results of different machine learning methods for the two demographers' tasks in sequence mining, such as prediction of the next event in an individual life course and finding patterns that distinguish men and women. The interpretable machine learning models [21] such as decision trees are suitable for further interpretation by demographers. The best encodings of events for the cases of gender prediction and the next event prediction are different, so there is no universal best choice.

The SVM method with custom kernels and with sequences transformed into features has approximately the same accuracy for the gender prediction problem, however, for the next event prediction, the resulting accuracy is much higher for the custom kernel function ACS. Although, the prefix-based kernel (CP) in combination with feature-based prediction after their weighting by accuracies gave the best accuracy as well, which shows that starting events in the individual life-course may contain important predictive information. The best accuracy results are obtained with the two-channel convolutional neural network for both cases, especially, with the highest accuracy of 0.931 for the next event prediction. Recurrent neural networks also result in high accuracy, but slightly lower than that of CNN.

Among the future research directions we may outline the following ones:

1. What are the main demographic differences between modern and Soviet generations of the Russian population that machine learning and pattern mining algorithms can capture? Answering this question is very important for demographic theory because it either confirms or disproves a predictive potential of the current ideas about the stadiality of demographic modernisation.
2. Which of the proposed methods so far suits the best the demographer's needs? For example, prefix-based emerging sequential patterns without gaps[3] in terms of pattern structures [5] are good candidates for studies of the transition to adulthood [11]. Other methods and combinations of existing ones appear, which could be of interest for both data scientists and demographers [17, 19].
3. Comparative studies of modern Russian and European generations are useful to prove or deny a hypothesis that Russia still follows a different demographic trajectory than European countries due to its Soviet past, for example, in contrast to Western vs. Eastern Germany [25]. One of the plausible hypotheses is that there exists a lag in about 20–25 years between Russia and European countries [18], in terms of demographic behavior patterns.

---

[3] For emerging patterns in classification setting cf. [8].

4. If like in our studies, neural networks result in high accuracy in different demographic classification problems, they need direct incorporation of interpretable techniques on the level of single events or their itemsets like Shapley value based approaches [6], which are mainly used for separate features on the level of single examples.
5. Further studies of similarity measures [9,22] is needed as well as that of the interplay between the complexity of sequences (cf. turbulence measure in [10]) and their interpetability [11].

Another promising direction, which is often implicitly present in real data science projects but remains unattended in sequence mining research, is outlier detection [26]. For example, in our previous pattern mining studies, we found the following emerging sequence peculiar for men,

$$\langle \{work\}, \{education\}, \{marriage, partner\}, \{divorce, break\text{-}up\} \rangle,$$

but the events divorce and break-up would rarely happen within one month (the used time granule) and they require different preceding events, namely marriage and partnership, which also cannot happen simultaneously. Thus, together with the involved demographers, we realised that there is a misconception of the survey's participants how they treat the terms marriage and partnership (they are not equal); further, we have eliminated the issue by employing an extra loop in the data processing via checking concrete dates and marital statuses.

References

1. Aisenbrey, S., Fasang, A.E.: New life for old ideas: the "second wave" of sequence analysis bringing the "course" back into the life course. Soc. Meth. Res. **38**(3), 420–462 (2010). https://doi.org/10.1177/0049124109357532
2. Billari, F.C.: Sequence analysis in demographic research. Can. Stud. Popul. [Arch.] **28**, 439–458 (2001)
3. Billari, F.C., Fürnkranz, J., Prskawetz, A.: Timing, sequencing, and quantum of life course events: a machine learning approach. Eur. J. Popul. (Revue européenne de Démographie) **22**(1), 37–65 (2006). https://doi.org/10.1007/s10680-005-5549-0
4. Blockeel, H., Fürnkranz, J., Prskawetz, A., Billari, F.C.: Detecting temporal change in event sequences: an application to demographic data. In: De Raedt, L., Siebes, A. (eds.) PKDD 2001. LNCS (LNAI), vol. 2168, pp. 29–41. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44794-6_3
5. Buzmakov, A., Egho, E., Jay, N., Kuznetsov, S.O., Napoli, A., Raïssi, C.: On mining complex sequential data by means of FCA and pattern structures. Int. J. Gen Syst **45**(2), 135–159 (2016). https://doi.org/10.1080/03081079.2015.1072925

6. Caruana, R., Lundberg, S., Ribeiro, M.T., Nori, H., Jenkins, S.: Intelligible and explainable machine learning: best practices and practical challenges. In: Gupta, R., Liu, Y., Tang, J., Prakash, B.A. (eds.) The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2020, pp. 3511–3512. ACM (2020). https://dl.acm.org/doi/10.1145/3394486.3406707

7. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, pp. 1724–1734. ACL (2014). https://doi.org/10.3115/v1/d14-1179

8. Dong, G., Li, J.: Emerging pattern based classification. In: Liu, L., Özsu, M.T. (eds.) Encyclopedia of Database Systems, 2nd edn. Springer, Boston (2018). https://doi.org/10.1007/978-1-4614-8265-9_5002

9. Egho, E., Raïssi, C., Calders, T., Jay, N., Napoli, A.: On measuring similarity for sequences of itemsets. Data Min. Knowl. Discov. **29**(3), 732–764 (2015). https://doi.org/10.1007/s10618-014-0362-1

10. Elzinga, C.H., Liefbroer, A.C.: De-standardization of family-life trajectories of young adults: a cross-national comparison using sequence analysis. Eur. J. Popul. (Revue européenne de Démographie) **23**(3), 225–250 (2007). https://doi.org/10.1007/s10680-007-9133-7

11. Gizdatullin, D., Baixeries, J., Ignatov, D.I., Mitrofanova, E., Muratova, A., Espy, T.H.: Learning interpretable prefix-based patterns from demographic sequences. In: Strijov, V.V., Ignatov, D.I., Vorontsov, K.V. (eds.) IDP 2016. CCIS, vol. 794, pp. 74–91. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-35400-8_6

12. Gizdatullin, D., Ignatov, D., Mitrofanova, E., Muratova, A.: Classification of demographic sequences based on pattern structures and emerging patterns. In: Supplementary Proceedings of 14th International Conference on Formal Concept Analysis, ICFCA, pp. 49–66 (2017)

13. Hochreiter, S., Schmidhuber, J.: LSTM can solve hard long time lag problems. In: Mozer, M., Jordan, M.I., Petsche, T. (eds.) Advances in Neural Information Processing Systems (NIPS), Denver, CO, USA, 2–5 December, vol. 9, pp. 473–479. MIT Press (1996)

14. Ignatov, D.I., Mitrofanova, E., Muratova, A., Gizdatullin, D.: Pattern mining and machine learning for demographic sequences. In: Klinov, P., Mouromtsev, D. (eds.) KESW 2015. CCIS, vol. 518, pp. 225–239. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24543-0_17

15. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N.: Watkins, C: Text classification using string kernels. J. Mach. Learn. Res. **2**, 419–444 (2002). http://jmlr.org/papers/v2/lodhi02a.html

16. Muratova, A., Sushko, P., Espy, T.H.: Black-box classification techniques for demographic sequences: from customised SVM to RNN. In: Tagiew, R., Ignatov, D.I., Hilbert, A., Heinrich, K., Delhibabu, R. (eds.) Proceedings of the 4th Workshop on Experimental Economics and Machine Learning, EEML 2017, Dresden, Germany, 17–18 September 2017, pp. 31–40. CEUR Workshop Proceedings, Aachen (2017). http://ceur-ws.org/Vol-1968/paper4.pdf

17. Piccarreta, R., Studer, M.: Holistic analysis of the life course: methodological challenges and new perspectives. Adv. Life Course Res. (2019). https://doi.org/10.1016/j.alcr.2018.10.004

18. Puur, A., Rahnu, L., Maslauskaite, A., Stankuniene, V., Zakharov, S.: Transformation of partnership formation in eastern Europe: the legacy of the past demographic

divide. J. Comp. Fam. Stud. **43**, 389–417 (2012). https://doi.org/10.3138/jcfs.43.3.389

19. Ritschard, G., Studer, M.: Sequence analysis: where are we, where are we going? In: Ritschard, G., Studer, M. (eds.) Sequence Analysis and Related Approaches. LCRSP, vol. 10, pp. 1–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-95420-2_1

20. Rossignon, F., Studer, M., Gauthier, J.-A., Goff, J.-M.L.: Sequence history analysis (SHA): estimating the effect of past trajectories on an upcoming event. In: Ritschard, G., Studer, M. (eds.) Sequence Analysis and Related Approaches. LCRSP, vol. 10, pp. 83–100. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-95420-2_6

21. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

22. Rysavý, P., Zelezný, F.: Estimating sequence similarity from read sets for clustering next-generation sequencing data. Data Min. Knowl. Discov. **33**(1), 1–23 (2019). https://doi.org/10.1007/s10618-018-0584-8

23. Solomonoff, R.J.: The Kolmogorov lecture the universal distribution and machine learning. Comput. J. **46**(6), 598–601 (2003). https://doi.org/10.1093/comjnl/46.6.598

24. Tomczak, J.M., Zieba, M.: Probabilistic combination of classification rules and its application to medical diagnosis. Mach. Learn. **101**(1–3), 105–135 (2015). https://doi.org/10.1007/s10994-015-5508-x

25. Wahrendorf, M., et al.: Agreement of self-reported and administrative data on employment histories in a German cohort study: a sequence analysis. Eur. J. Popul. (2019). https://doi.org/10.1007/s10680-018-9476-2

26. Wang, T., Duan, L., Dong, G., Bao, Z.: Efficient mining of outlying sequence patterns for analyzing outlierness of sequence data. ACM Trans. Knowl. Discov. Data **14**(5), 62:1–62:26 (2020). https://doi.org/10.1145/3399671

27. Zimmermann, A., Nijssen, S.: Supervised pattern mining and applications to classification. In: Aggarwal, C.C., Han, J. (eds.) Frequent Pattern Mining, pp. 425–442. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07821-2_17