НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Classification Techniques for Demographic Sequences

Anna Muratova
Pavel Sushko
Thomas Espy

Outline

- Aim of the work
- Demographic data
- Results
  - ➢ Formulas of similarity measures of sequences without discontinuities
  - ➢ Variants of a custom kernel in the SVM method
  - ➢ Recurrent neural networks
- Conclusion

Aim of the work

- More useful information from accumulated demographic data can be extracted by applying modern methods of data mining
- Comparison of different methods of classifying demographic data
- The novelty of this work is the use of custom kernel variants in the SVM method

Demographic data

The Data for the work was obtained from the scientific laboratory of socio-demographic policy at HSE and it contains results of a survey of 6626 people (3314 men and 3312 women)

Features:
- Gender (male, female);
- Generation (Soviet 1930-1969 and modern 1970-1986);
- Type of education (general, higher, professional);
- Location (city, town, country);
- Religion (yes, no);
- Frequency of church attendance (several times a week, once a week, at least once a month, several times a year or less);

Also for each person the dates of birth and the dates of significant events in their lives are indicated, such as: partner, marriage, break up, divorce, education, work, separation from parents and birth of a child

Similarity measures

Let *S* and *T* be given sequences.

All common subsequences:

$$sim_{ACS}(S,T) = \frac{\phi(S,T)}{max\{\phi(S),\phi(T)\}}$$

Longest common subsequence:

$$sim_{LCS_{size}}(S,T) = \frac{|LCS(S,T)|}{max\{|S|,|T|\}}$$

where *LCS* is the longest common subsequence.

Similarity measures without discontinuities

Let *S* and *T* be given sequences.

Common prefixes:

$$sim_{CP}(S,T) = \frac{|LCSP(S,T)|}{\max\{|S|,|T|\}}$$

where *LCSP* is the longest common sequence prefix.

Similarity measures without discontinuities

Longest common subsequence:

$$sim_{LCS}(S,T) = \frac{|LCS(S,T)|}{\max\{|S|,|T|\}}$$

All common subsequences:

$$sim_{ACS}(S,T) = \frac{2 \cdot \sum_{k \le l} \phi(S,T,k)}{l(l+1)}$$

$$l = \max\{|S|,|T|\}$$

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*.

Custom kernel in the SVM

- To evaluate classification quality by different methods, the class “gender” for the test sample was used. The initial data was divided into training and test sets in the ratio 80/20.

- Classification with the use of custom kernel functions in the SVM method based on the sequence similarity measures without discontinuities (CP, ACS, LCS).

Parameter	CP	ACS	LCS
Model fitting time, sec	400.97	1580.86	1544.21
Prediction time, sec	98.66	394.20	388.06
Total time, sec	499.62	1975.06	1932.27
Accuracy	0.648	**0.659**	0.490

Look at accuracy for methods comparison, time is shown just for information

Custom kernel in the SVM

- Classification by the SVM method by features.

We use the SVM method with default parameters (kernel function - RBF).

Parameter	Value
Model fitting time, sec	3.62
Prediction time, sec	0.52
Total time, sec	4.14
Accuracy	**0.615**

Custom kernel in the SVM

- Classification by sequences, by features and by weighted sum of probabilities of sequences and features.

Parameter	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 $P = \frac{As \cdot Ps + Af \cdot Pf}{As + Af}$	**0.678**	0.670	0.612

As — Accuracy by Sequences,
Af — Accuracy by Features,
Ps — Probability by Sequences,
Pf — Probability by Features.

Custom kernel in the SVM

- Classification by features and sequence as an additional feature.

Parameter	Classification by sequences only (as features)	Classification by features only	Classification by sequences and features
Number of sequences	6626	0	6626
The number of unique sequences of maximum length (number of features values)	1228	0	1228
Number of initial features	0	5	5
Number of generated features (from sequences)	1	0	1
Model fitting time, sec	4.80	3.62	5.79
Prediction time, sec	0.79	0.52	0.91
Total time, sec	5.59	4.14	6.70
Accuracy	**0.675**	0.615	**0.716**

Recurrent neural networks

- Classification by sequences using recurrent neural networks using Keras and Tensorflow software was made.

Method of classification	**By sequences**			**By features**	**By sequences and features**
	Neural network layers (number)				
Parameter	SimpleRNN(1) Dense(1)	GRU(1) Dense(1)	LSTM(1) Dense(1)	Dropout(1) Dense(3)	SimpleRNN(1) Dense(5) Dropout(3)
The number of events in the sequences (maximum)	8	8	8	0	8
Number of features	0	0	0	5	5
Model fitting time, sec	168.80	452.27	585.73	348.49	418.05
Prediction time, sec	2.28	3.38	3.93	0.68	1.36
Total time, sec	171.07	455.66	585.73	349.17	419.40
Accuracy	0.676	0.672	0.675	0.626	**0.754**

Comparison of classification methods

Methods	Classification by sequences only	Classification by features only	Classification by sequences and features
SVM			
Custom kernel function CP	0.648	0.615	**0.678**
Custom kernel function ACS	0.659	0.615	0.670
Custom kernel function LCS	0.490	0.615	0.612
Using sequences transformed into features	0.675	0.615	**0.716**
Recurrent neural networks (Keras, Tensorflow)			
SimpleRNN, Dense	0.676	0.626	<u>**0.754**</u>
GRU, Dense	0.672	0.626	
LSTM, Dense	0.675	0.626	
Decision trees, time coding			**0.661**

Conclusion

- The best classification results are obtained using the custom kernel function in SVM by transforming sequences into features and even better result with recurrent neural network SimpleRNN
- These two methods take into account events regularities in the sequences, unlike most other methods that can work with features only
- This work can be applied to the analysis of various sequences

Thanks!