SAMARA UNIVERSITY

A Method of Preference and Utility Elicitation By Pairwise Comparisons and its Application to Intelligent Transportation Recommendation Systems

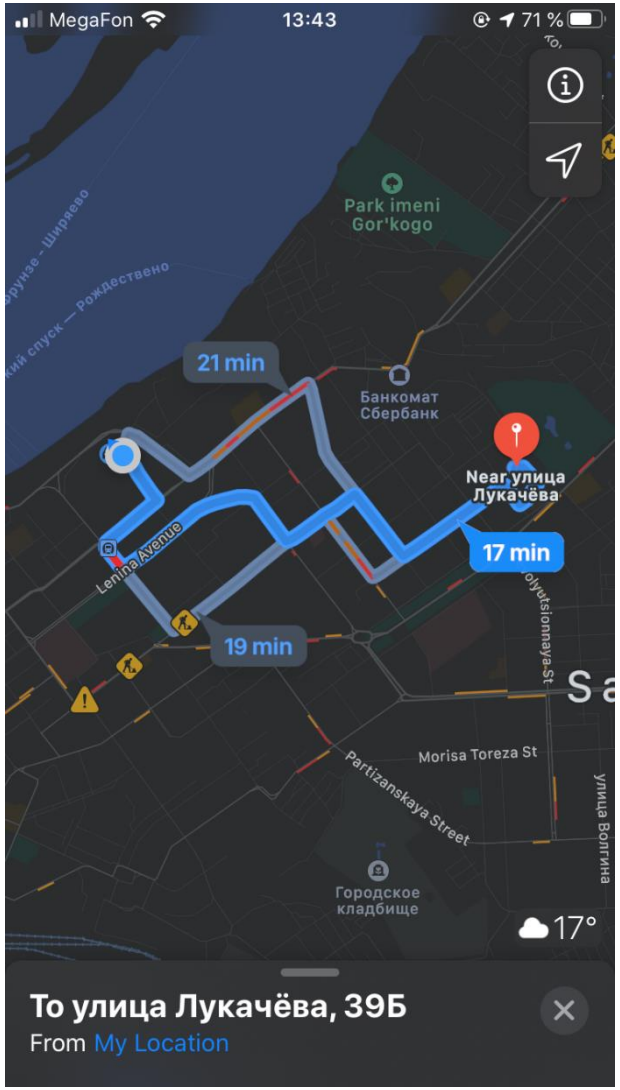
Aleksandr Borodinov, Anton Agafonov, Vladislav Myasnikov

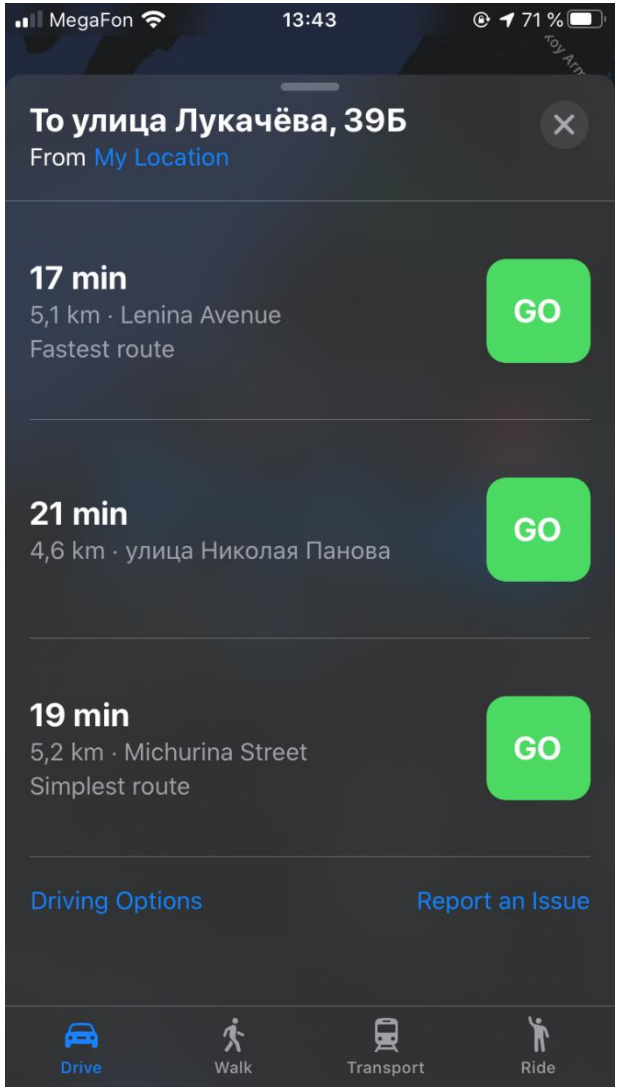
ICIST - 2020

Introduction

Typical tasks in machine learning:

- *label ranking;*
- *instance ranking;*
- ***object ranking.***





Problem Formulation

Let the objects set $\Omega \equiv \{\omega_j\}_{j \in J}$ have an order and/or a strict partial order.

utility function $u: \Omega \to \mathbb{R}$ determines the *absolute preference;*
preference function $p: \Omega \times \Omega \to \mathbb{R}$ determines the *relative preference*.

Information about learning preferences can be defined as follows:
- the values of the desired utility functions for the set of observed objects – *direct information*;
- the results of the pairwise comparison p_{ij} for the subset of observed objects – *indirect information*.

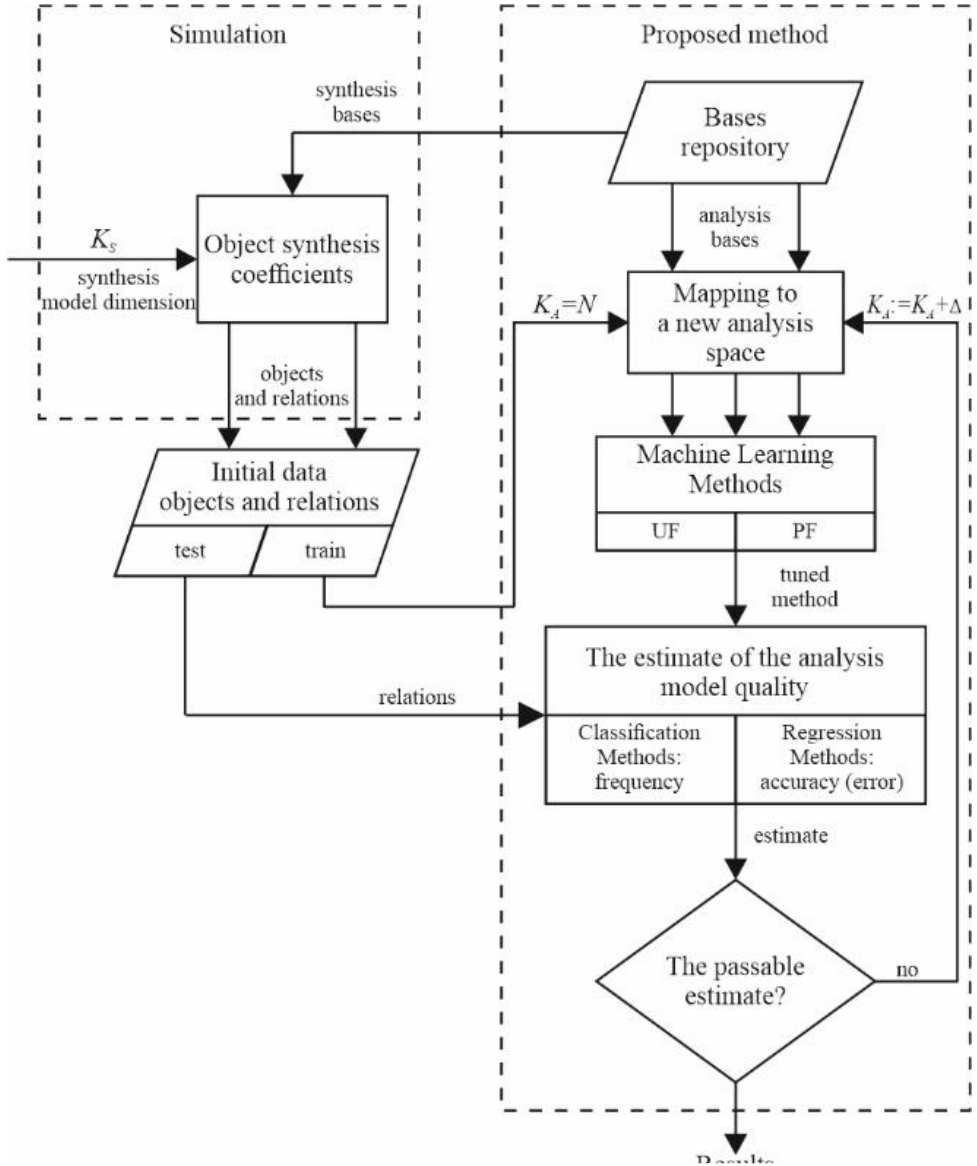
The *indirect information* can be defined in two ways:
- *explicitly*, by the values of the preference function $p(\omega_j, \omega_i)$ or the synthesized preference function by the utility function $p(\omega_j, \omega_i) = u(\omega_j) - u(\omega_i)$;
- *implicitly*, by the symbolic representation of the following form:

$$z_{ij} \equiv z(\omega_i, \omega_j) = \begin{cases} 1, & p(\omega_i, \omega_j) > 0; \\ 0, & p(\omega_i, \omega_j) = 0; \\ -1, & p(\omega_i, \omega_j) < 0. \end{cases}$$

General Method Description

The proposed method:

1) feature values normalization in the range [0,1];
2) selection of a new feature space (basis) *Y*;
3) transformation of the original feature vector **x** into the new feature space *Y* with a higher dimension $K=\dim(Y) \geq N$;
4) building a linear or nonlinear classifier in the feature space *Y*;
5) quality assessment of the building classifier on the test dataset;
6) if the assessment is satisfactory, stop the procedure; otherwise, go to steps 3 or 2 (if all available dimensions of the feature space are already used).



General Method Description

Bases Repository:

1) Original basis:

$$K = \dim(Y) = \dim(X) = N;$$
$$y_n = \varphi_n(\mathbf{x}) = x_n, \quad n = \overline{0, N-1}.$$

2) Polynomial basis:

$$y_k = \varphi_k(\mathbf{x}) = \prod_{n=0}^{N-1} x_n^{k_n}, \quad k = \sum_{n=0}^{N-1} K_0^n k_n$$

3) Fourier basis (harmonic):

$$y_k = \varphi_k(\mathbf{x}) = \prod_{n=0}^{N-1} \cos\left(\pi k_n x_n\right), \quad k = \sum_{n=0}^{N-1} K_0^n k_n$$

Machine Learning Methods (Classifiers) Repository:

1) logistic regression (LR),

2) Fisher's linear discriminant,

3) linear support-vector machine (without kernel),

4) support-vector machine with the radial basis function (SVM-RBF),

5) nearest neighbor method,

6) decision tree,

7) Random Forest (RF).

Experimental Research using Model Data

Comparisons of Fourier Basis and Polynomial Basis
(LR –Logistic Regression, RF – Random Forest)

K_S	K_A	T	Error ($\tilde{d}$)			
			A: Fourier S: polynomial		A: polynomial S: Fourier	
			LR	RF	LR	RF
35	35	10000	0,2864	0,1828	0,0076	0,0145
35	35	50000	0,3341	0,1237	0,0067	0,0083
35	63	10000	0,3059	0,1945	0,0062	0,0141
35	63	50000	0,2633	0,1223	0,0049	0,0084

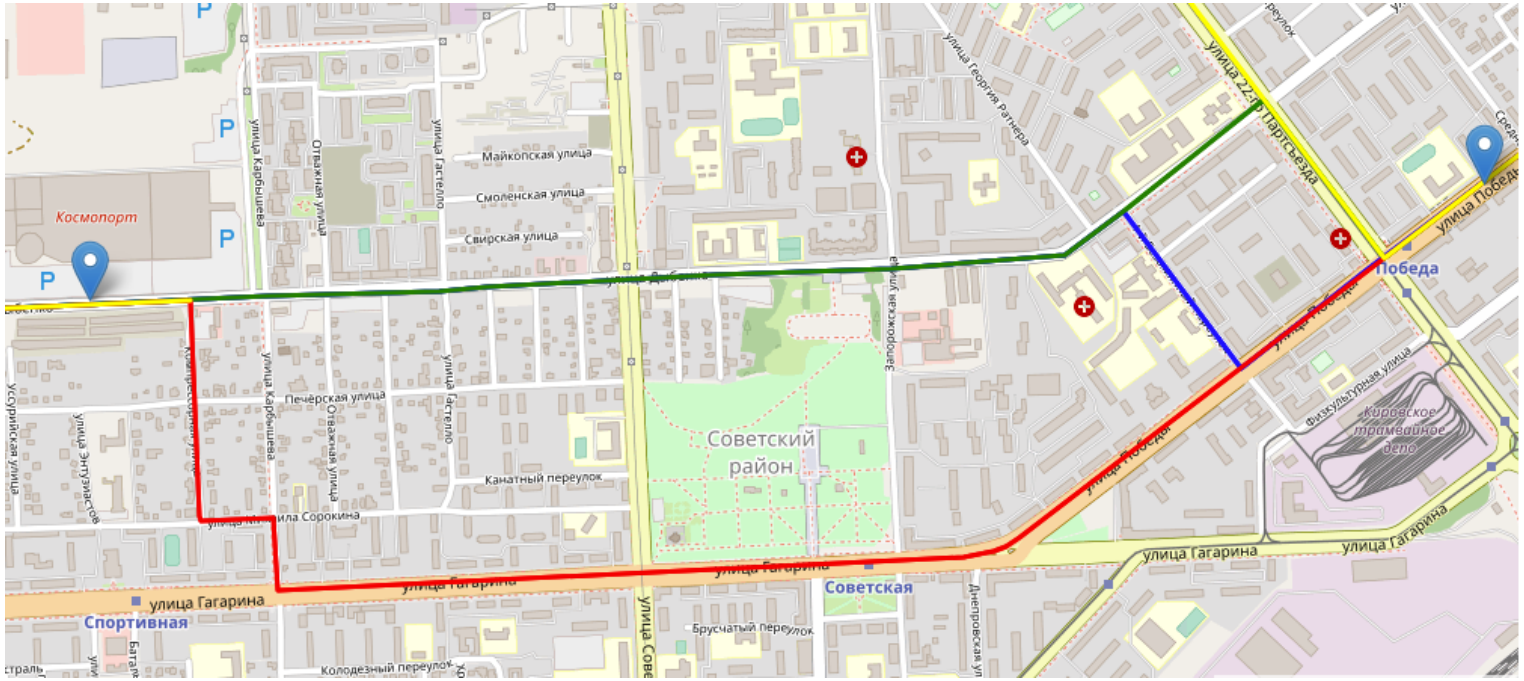
Dependence of Utility Function Reconstruction Error $\tilde{d}$ on the Number of Used Pairwise Comparisons T

T	500	1000	5000	10000	20000	50000
K_S=35 K_A=15	0.012	0.013	0.0128	0.011	0.0111	0.0092
K_S=35 K_A =35	0.03	0.019	0.0048	0.0076	0.00715	0.00668
K_S=35 K_A =63	0.058	0.014	0.006	0.0062	0.00495	0.00495

The Kendall's distance: $d = \left|\left\{\left(i,j\right): z\left(\omega_i,\omega_j\right) \neq z\left(\mathbf{x}(\omega_i),\mathbf{x}\left(\omega_j\right)\right), \quad (i,j) \in \Theta\right\}\right|$;

in the “normalized” form: $\tilde{d} = d \cdot |\Theta|^{-1}$

Experimental Research using Real Data



Cross-validation Procedure Parameters

The size of the training set in the number of made decisions γ	The size of the test set in the number of made decisions
3	40
10	40
20	30
30	20

Experimental Research using Real Data

Features:

1) ratio of the straight line distance between points A and B to the current track length;
2) ratio of the shortest (in distance) track length to the current track length;
3) intensity of the intersections of the current track per 100 meters;
4) ratio of the turns number to the intersections number for the current track;
5) ratio of the minimum intersections number (in all provided tracks between points A and B) to the intersections number for the current track;
6) ratio of the left turns number to the total turns number on the current track;
7) ratio of the minimum travel time (with the maximum permitted speed) to the estimated travel time on the current track;
8) ratio of the shortest (in time) track travel time to the estimated travel time on the current track;
9) ratio of the square root of travel speed variance on the current track to the maximum permitted speed.

Experimental Research using Real Data

Dependence of the Error Probability $\tilde{d}$ and Analysis Space Dimension K_A on the Training Set Size

№ user	γ=3		γ=10		γ=20		γ=30	
	$\tilde{d}$	K_A	$\tilde{d}$	K_A	$\tilde{d}$	K_A	$\tilde{d}$	K_A
1	0.205	11	0.198	13	0.200	15	0.181	13
2	0.225	17	0.205	13	0.194	11	0.168	13
3	0.197	15	0.190	13	0.159	13	0.188	13
4	0.245	13	0.235	11	0.172	11	0.219	11
mean	0.218	14	0.207	12.5	0.181	12.5	0.189	12.5
median	0.215	14	0.202	13	0.183	12	0.185	13

Conclusions:

- the proposed method confirm its effectiveness – the error in all considered problem statements is in the range 0.16-0.25;
- the quality of the constructed solution on real data is expectedly higher for large sizes of the training sets;
- the proposed method is applicable for the system «cold start» mode (γ=3) the efficiency is decreased in about 10-20% for the best quality value (for each user);
- for the considered problem and the selected feature description, the dimension of the new feature space is practically independent of the set size and exceeds the dimension of the original space (9 features) by 30-50%.

SAMARA UNIVERSITY

THANK YOU

The authors would like to thank Geoinformatics and Information Security Department, Samara National Research University for taking part in a survey on data collection.

The work was funded by the Ministry of Science and Higher Education of the Russian Federation (unique project identifier RFMEFI57518X0177).

aaborodinov@yandex.ru