Predicting age and gender from a patient's medical diagnoses

UNIVERSITY OF TARTU
Institute of Computer Science

Taivo Pungas*, Hans Peeter Tulmin
Institute of Computer Science, University of Tartu, *tpungas@ut.ee

IT Akadeemia
toetab Skype™

Conclusions

The methods tested in this study yielded prediction results better than random. However, the accuracy achieved is too low for useful applications.

Restricting analysis to patients with at least some amount of diagnoses may improve results. Ensemble and neural network methods should be tested as well.

Methods

Input data was cleansed and preprocessed as explained in section Data.

Before analysis, PCA was conducted on the diagnoses matrix. The number of principal components was chosen so that the explained variance would be 0.9. This criterion produced a 249-dimensional space, into which the original data was transformed for further analysis.

Software used: R; Python with packages Numpy and MDP.

Results

Predicting gender

An extensive parameter probe was conducted on SVMs with linear and RBF kernels. Comparing models using validation set accuracy, the best model was selected.

The best model (which used an RBF kernel) yielded a test set accuracy of **69.6%**, compared to a naive prediction accuracy of 56.0%. The amount of training data used had a small effect on the accuracy.

Predicting age

Three different methods were tested: SVMs, DBSCAN and SV-regression.

SVM produced near-random results that were not affected by choice of kernel or parameters. DBSCAN only found two clusters out of the expected 8.

SV-regression achieved a test set accuracy of **35%** compared to a naive prediction accuracy of 23%.

Data

The input data, consisting of patients' genders, ages and diagnosis histories, was filtered.

Ages were assigned into 8 bins with edges at 0, 20, 30, 40, 50, 60, 70, 80, 100 years.

The original diagnosis histories, lists of Estonian RHK-10 codes (e.g. 'G01'), were transformed into binary vectors.

After preprocessing, a sample of 48 728 patients was left. This data was partitioned into training (60%), validation (20%) and test (20%) sets.

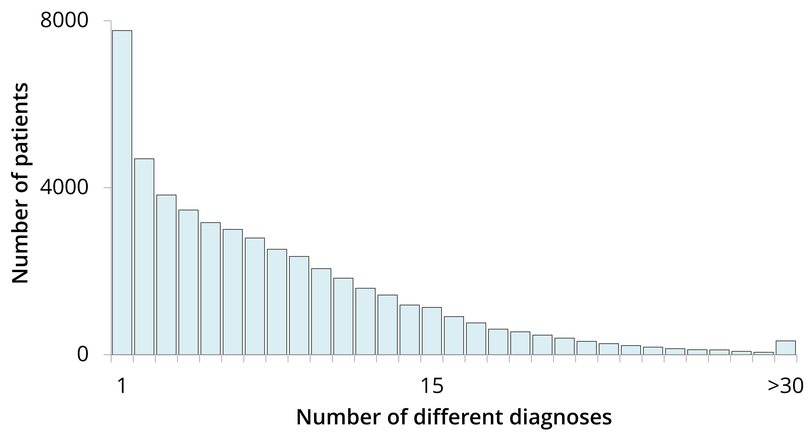

Figure 1. Distribution of diagnosis counts among patients after filtering and preprocessing.