Supervised Learning as a Method to Reduce Clerical Effort

Jörg Feuerhake – Federal Statistical Office of Germany (Destatis)

DESTATIS
Statistisches Bundesamt

The Problem – Irrelevant Enterprises in Crafts Statistics

Crafts Enterprises?

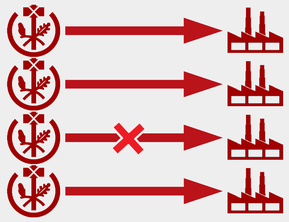
In Germany a set of occupations are subject to compulsory membership in crafts chambers. The Federal Statistical Office compiles statistics on crafts enterprises.

Crafts Statistics and Irrelevant Enterprises

Statistics on crafts are compiled from the business register. Here, membership lists of crafts chambers are linked to enterprises. But there is a Problem ...

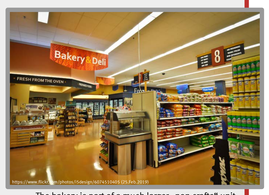
... irrelevant Enterprises

There is a set of units that are not relevant for the crafts statistics because they do not offer crafts products or services directly. Chambers of crafts report them because they themselves cannot identify the units based on their information.



Irrelevant Enterprises – Two Examples

„non crafts" logistics enterprises often keep maintenance units for their fleets



The bakery is part of a much larger „non crafts" unit.

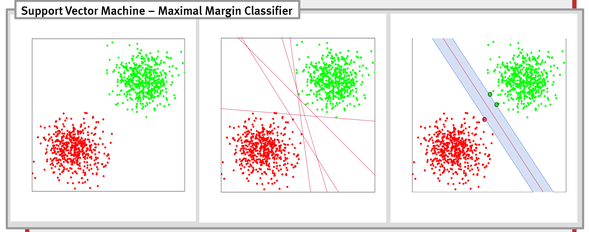
Since there are a lot of reasons, depending on several properties of the enterprise, that determine „irrelevancy" for crafts statistics, the decision used to be made by clerical review.

An annual average of 40.000 Units needs to be checked manually.

How to relieve the staff from the reviewing burden?

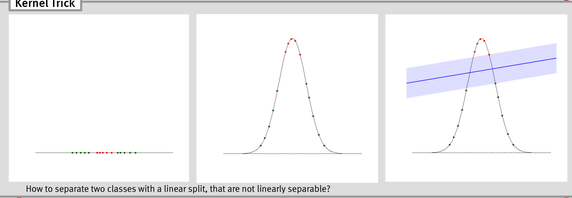
Supervised Learning Methods – Support Vector Machine

Support Vector Machine – Maximal Margin Classifier



SVMs maximize the margin around the separating hyperplane. The decision function is fully specified by a (usually very small) subset of training samples, the support vectors. Maximization is a quadratic programming problem that is easy to solve by standard methods yet tends to consume a lot of computing power.

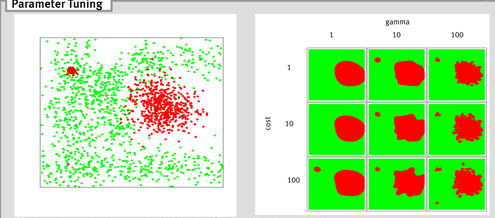
Kernel Trick



How to separate two classes with a linear split, that are not linearly separable?

Patterns that are not linearly separable can be split by transformations of original data to map into new space – the Kernel Trick

Parameter Tuning



The parameters *cost* and *gamma* need to be set so that over-fitting is prevented.

Support Vector Machines classify by finding the widest margin between classes. SVM tend to be very hardware intensive, especially while tuning parameters.

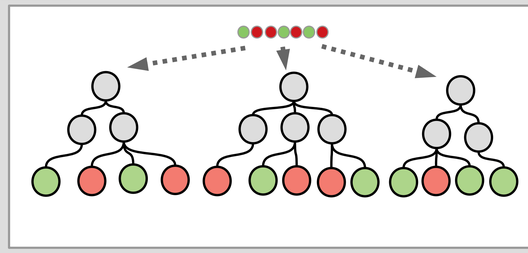
Supervised Learning Methods –Random Forest

Random forests are an ensemble learning method for classification and regression.

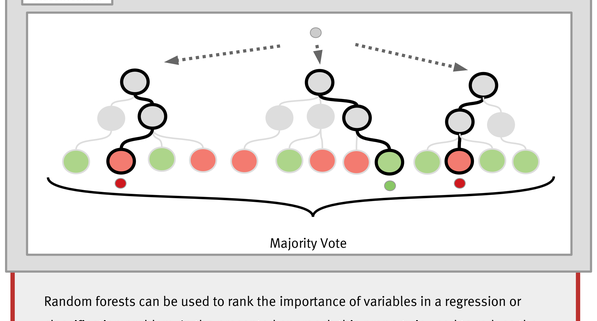
It "grows" a set of uncorrelated decision trees at training time. A unit can then be classified by mode of the classes the forest trees put out.

Random forests correct for over-fitting tendencies to training sets in decision trees.

Training



Classification

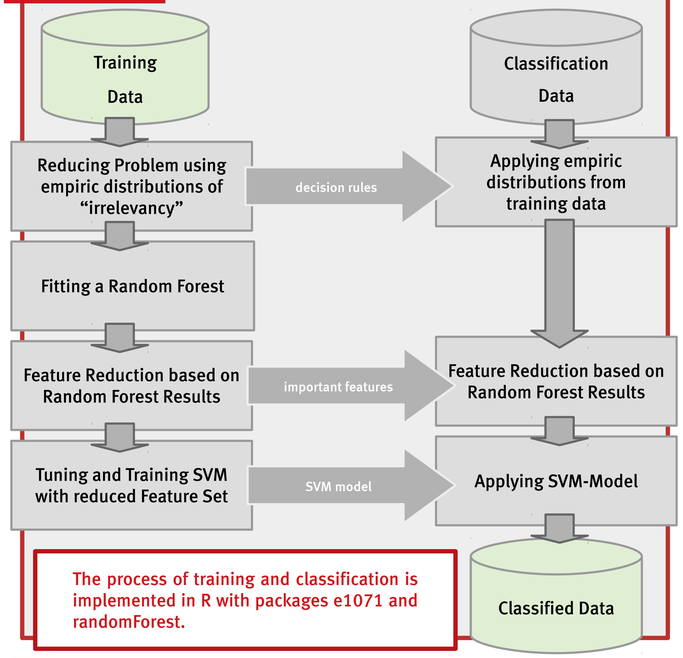


Majority Vote

Random forests can be used to rank the importance of variables in a regression or classification problem. In the presented approach this property is used to reduce the variables of the training set and therefore allowing to tune and train a SVM in the following step.

Random Forest are used in the presented approach to extract the most important features of the training data set. The random forest classification is used as a benchmark for the SVM's performance.

Implementation



The process of training and classification is implemented in R with packages e1071 and randomForest.

Results

Algorithm	Prediction	Original Value	Units	Turnover
SVM	relevant	relevant	*93,1 %*	*65,2 %*
	irrelevant	irrelevant	*3,2 %*	*33,1 %*
	classified correctly:		***96,3 %***	***98,2 %***
	irrelevant	relevant	*0,3 %*	*0,3 %*
	relevant	irrelevant	*3,4 %*	*1,5 %*
	classified incorrectly:		***3,7 %***	***1,8 %***
Random Forest *(as benchmark)*	relevant	relevant	*93,0 %*	*62,8 %*
	irrelevant	irrelevant	*1,4 %*	*20,5 %*
	classified correctly:		***94,4 %***	***83,3 %***
	irrelevant	relevant	*0,4 %*	*2,7 %*
	relevant	irrelevant	*5,2 %*	*14,1 %*
	classified incorrectly:		***5,6 %***	***16,7 %***

The results of the approach are promising. Note the performance of the SVM relative to Random Forest for turnover. The approach is currently being rolled out to production process and does relief clerical burden already.

Data and Patterns – How to Distinguish the Classes?

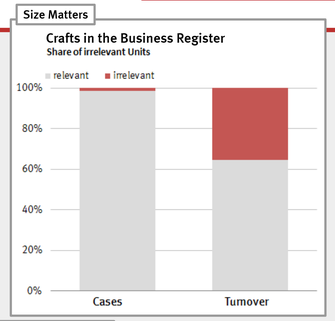
In the business register data from prior periods with results of clerical review exists. Main features in the data set are **turnover**, **number of persons employed** both **subject to social security contribution** and in **marginal employment**, **NACE classes**, **crafts occupations** from the membership lists of crafts chambers. Only turnover and employment data are rationally scaled. Training data had around 600.000 observations initially.

Only 2% of the crafts enterprises are classified as irrelevant. They represent 35% of turnover and 19% of persons employed subject to social security contributions. So the problem is an **imbalanced data** problem and large units seem to be more prone to being irrelevant.

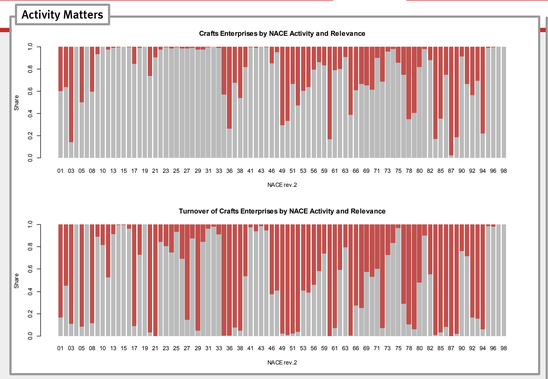
Combinations of NACE classes and crafts occupations are further features that correlate with irrelevancy. There are many combinations that do not have irrelevant cases at all and there are combinations, where almost all of the units are marked irrelevant.

Does the data carry patterns that can be used to train a supervised learning algorithm?

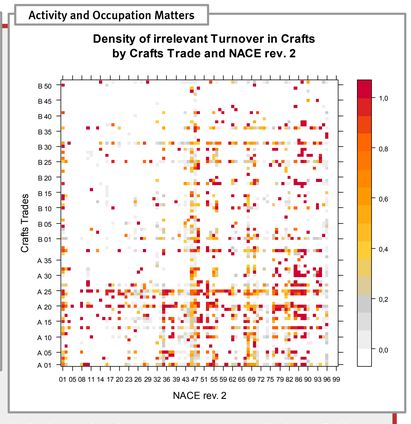
Size Matters



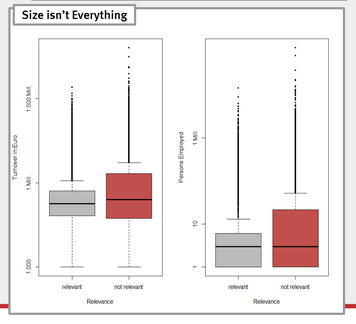
Activity Matters



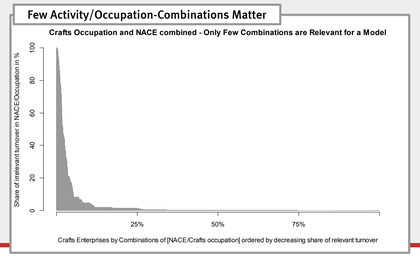
Activity and Occupation Matters



Size isn't Everything



Few Activity/Occupation-Combinations Matter



The exploratory analysis shows, that there are several patterns in the available data. Some of the patterns, like combinations of NACE activity and crafts occupation, can be identified and later on reproduced by rather simple queries on the data. Applying this approach reduces the training data remarkably.

The computing intensive supervised learning methods are applied to the remaining data set with about 70.000 observations. Furthermore by reducing the training data set, the share of irrelevant cases increases to about 7% hence easing the imbalanced data problem.

References

Council of European Union. Regulation (ec) no 1059/2003 of the european parliament and of the council of 26 may 2003 on the establishment of a common classification of territorial units for statistics (nuts), 2003. URL http://data.europa.eu/eli/reg/2003/1059/2018-01-18.

Nace rev.2 statistical classification of economic activities in the european community, 2008. URL https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, pages 144–152, 1992.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. Machine Learning, pages 273–297, September 1995.

Colin Campbell and Yiming Ying. Learning with Support Vector Machines. Morgan & Claypool Publishers, 2011. ISBN 1608456161, 9781608456161.

Ingo Steinwart and Andreas Christmann. Support Vector Machines. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387772413.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning: With Applications in R. Springer Publishing Company, Incorporated, 2014. ISBN 1461471370, 9781461471370.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. CART: Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA, 1984.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29:1189–1232, 2000