Supervised Learning as a Method to Reduce Clerical Effort

Joerg Feuerhake

Keywords: *Support Vector Machines, Random Forests, Business Statistics*

1 Introduction

With the availability of more and more computing power Machine Learning methods become more relevant in the production of statistics. One important field of application is the classification of statistical units based on models trained with units, where the classification is known.
In this paper an approach to classify units from a business database based on prior clerical review is presented. The goal is to remarkably reduce clerical effort in the statistic's production process. Consider the case where a share of roughly 2% of a population about 600.000 units is not relevant for the results of a certain annual statistic. There are several reasons for a unit to become irrelevant for the statistic and the reasons depend on items like size, economic activity and other rationally and nominally scaled variables.
Additionally assume that a unit's relevance in recent periods was controlled by clerical review. So each year all units entering the population or changing in an important variable have to be checked manually. There are on average 40.000 units each year that previously needed clerical review. Thus the staff bound by these reviews was considerable, let alone the training to enable staff members to review cases correctly and the time needed to do the reviews.
In the presented project, methods of supervised learning are applied to achieve the above mentioned goals. Random Forests and Support Vector Machines (SVM) are trained in a combined approach based on populations of prior years to get models that would be able to predict the units that enter the population or change in important variables.

2 Methods

As stated in the introduction, the presented problem is a binary classification problem. A method to decide whether a unit is relevant or irrelevant for a particular statistic needs to be designed. Since classified units exist from prior periods, applying a supervised learning approach seems to be feasible. Furthermore note that only about 2% of the units belong to the "irrelevant" class. So one might have to take measures to deal with imbalanced data.
The datasets from prior periods consist of information about turnover, number of employees, location on NUTS 2 Level [1], economic activity on NACE class level [2] and an occupational classification from authorities that oversee occupational education. Only turnover and number of employees are rationally scaled variables. Concerning these two rationally scaled variables, turnover and number of employees, figure 1 already shows one pattern that might be helpful when classifying new units. Enterprises that are classified "irrelevant" tend to be larger than "relevant" enterprises. A second important pattern is the relation between economic activity and occupational classifier. Some combinations of economic activity and occupational classifier tend to have very dense populations of "irrelevant" enterprises. Even more so if one weights

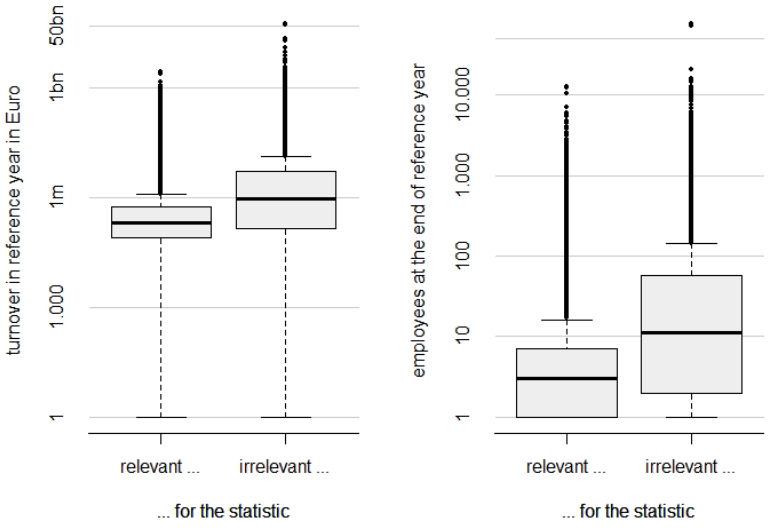

Figure 1: Distribution of Turnover and Number of Employees by Classes

the units with turnover or number of employees. On the other hand, a lot of the combinations have neglectable numbers of irrelevant units.
As a supervised learning algorithm support vector machines were chosen in a first approach. For mathematical details refer to the original works of Boser et. al. (1992) [3] and Cortes and Vapnik (1995) [4]. For an introductory overview refer for example to Campbell/Ying (2011, here: page 1 ff.) [5], Steinwart/Christmann (2008, here: page 13 ff.) [6] and James et al. (2014, here: page 337 ff.) [7].
The project has shown that a naive application of SVM to the above described training data did not yield usable results. This was due to a lack of computing resources, which were not at hand for the task. Since the computing resources could not be increased within the project the initial problem had to be reduced. To achieve that, a three step approach was chosen. In the first step, cases which could be classified relatively easily without machine learning methods would be taken out of the training data. As mentioned above, the combinations of economic activity and occupational classifier is in many cases a good predictor for the relevance of a unit. So units having particular economic activity and occupational classifier combination, where the probability to be irrelevant is either very high or very low, where taken out of the training dataset. The classification of new units belonging to one of theses economic activity/occupational classifier groups was subsequently done based on the previously observed relevance.
The remaining training data consisted of 68.000 units. The first reduction step led to one further improvement. The density of the irrelevant enterprises within the training data increased to about 7% in the training data, hence reducing the imbalanced data problem. Since nominally scaled items need to be represented as dummy variables for the SVM, the training data still consisted of 330 variables. The quality of SVM results strongly depends on parameter tuning. That means, one has to train the SVM with different combinations of parameters. To do this in a acceptably short time the reduction in the first step was not sufficient. Therefore in a second step, the variables are reduced to those with most explanatory value. To achieve this, a random forest was applied to the reduced training data.
Random forest is a tree based classifier. The founding work on the method can be found in Breimann et al. (1984) [8]. Compared to SVMs it tends to need less resources, but may not always yield classification results as good as those SVMs can produce.

Still, random forests can calculate variable importance during the training process. This feature can be used to find the most important variables for classifying purposes in a training data set and it was used to reduce the item set of the training data from 330 to 40. With this second reduction of the training data it was possible to tune and train a SVM and get acceptable results in sufficient time.

3 Results

With the approach described above, it was possible to train a SVM that was used to classify new or significantly changed enterprises for statistical purposes. To get an overview of misclassification ratios of the SVM model, when applied to it's training data, refer to table 1. Note that the over all misclassification rate is 3.7% while weighted with turnover the rate drops to 1.8% for cases that need to be classified by means of SVM.
For the above mentioned goal of minimizing clerical efforts within the statistical production process, the results seemed very promising. Therefore, the approach was integrated in the production process.

Table 1: SVM Classification Results

Prediction	**Original Value**	**Number of Enterprises**	**Turnover**
relevant	relevant	93.1%	65.2%
irrelevant	irrelevant	3.2%	33.1%
correctly classified		96.3%	98.2%
irrelevant	relevant	0.3%	0.3%
relevant	irrelevant	3.4%	1.5%
not correctly classified		3.7%	1.8%

Currently a SVM is used to classify large portions of the units that would have needed clerical intervention. That saves time for staff members to concentrate on larger units. However, the approach depends heavily on the existence of large computing ressources.

4 Conclusions

In the presented case, a machine learning approach consisting of random forest and SVM is able to classify units in a quality that enables staff to shift clerical intervention to the most important units of the population. This enhances the quality of results of the statistic. A drawback of the current approach are the relatively time consuming calculations. A road ahead that is currently checked, is the testing of other algorithms like gradient boosting [9] for the retrieval of the most important variables.
A second point of interest is the interpretability of results. Especially the results of SVMs are quite difficult to interpret, which makes it hard to explain the classification of every single unit to staff members. This tends to influence staff's attitudes towards machine learning approaches negatively. Consequently, ways to improve explanability of results are looked for in order to improve the applicability of SVMs in the production process.

References

[1] 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. http://data.europa.eu/eli/reg/2003/1059/2018-01-18.

[2] 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.

[3] 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.

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

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

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

[7] 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.

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

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