Routledge
Taylor & Francis Group

Economic Research-Ekonomska Istraživanja

ISSN: 1331-677X (Print) 1848-9664 (Online) Journal homepage: http://www.tandfonline.com/loi/rero20

A nonparametric data mining approach for risk prediction in car insurance: a case study from the Montenegrin market

Vladimir Kašćelan, Ljiljana Kašćelan & Milijana Novović Burić

To cite this article: Vladimir Kašćelan, Ljiljana Kašćelan & Milijana Novović Burić (2016) A nonparametric data mining approach for risk prediction in car insurance: a case study from the Montenegrin market, Economic Research-Ekonomska Istraživanja, 29:1, 545-558, DOI: 10.1080/1331677X.2016.1175729

To link to this article: http://dx.doi.org/10.1080/1331677X.2016.1175729

Published online: 13 May 2016.

ECONOMIC RESEARCH-EKONOMSKA ISTRAŽIVANJA, 2016
VOL. 29, NO. 1, 545–558
http://dx.doi.org/10.1080/1331677X.2016.1175729

Routledge
Taylor & Francis Group

A nonparametric data mining approach for risk prediction in car insurance: a case study from the Montenegrin market

Vladimir Kašćelan, Ljiljana Kašćelan and Milijana Novović Burić

Faculty of Economics, University of Montenegro, Podgorica, Montenegro

ABSTRACT

For prediction of risk in car insurance we used the nonparametric data mining techniques such as clustering, support vector regression (SVR) and kernel logistic regression (KLR). The goal of these techniques is to classify risk and predict claim size based on data, thus helping the insurer to assess the risk and calculate actual premiums. We proved that used data mining techniques can predict claim sizes and their occurrence, based on the case study data, with better accuracy than the standard methods. This represents the basis for calculation of net risk premium. Also, the article discusses advantages of data mining methods compared to standard methods for risk assessment in car insurance, as well as the specificities of the obtained results due to small insurance market, such as Montenegrin.

ARTICLE HISTORY
Received 5 July 2013
Accepted 5 February 2016

KEYWORDS
Car insurance; net risk premium; data mining; clustering; support vector regression (SVR); kernel logistic regression (KLR)

JEL CLASSIFICATIONS
C38; C53; C81; G22

1. Introduction

Aggregate claims for a homogeneous car insurance portfolio have long been estimated using pure algorithmic methods to calculate tariffs (Mayer, 2002). This understanding does not allow quantification of uncertainties. Uncertainties can only be determined if we have an underlying stochastic model on which the calculation algorithms can be based.

The generalised linear models (GLS) and other more flexible stochastic models are used in recent studies to predict insurance tariffs on a micro-level, i.e., on level of individual claims (Bortoluzzo, Claro, Caetano, & Artes, 2011; David, 2015; Ohlsson & Johansson, 2010). For these models a major limitation is that the structure is restricted to a linear form, which can be too rigid for real applications. Also, there is a problem with modelling if many explanatory variables are discrete and multi-valued, which is quite common for insurance data-sets. These drawbacks can be overcome by using nonparametric methods from modern statistical machine learning and data mining theory such as support vector regression (SVR) or kernel logistic regression (KLR) (Christmann, 2005). The support vector machine (SVM) has better predictive performance than other techniques, especially for data that exhibit nonlinearity (Tian, Shi, & Liu, 2012).

CONTACT Milijana Novović Burić mnovovic@ac.me

The majority of insurance companies keep the data on history of its operations in a data warehouse. These huge quantities of data hide very important information, which could contribute to easier decision-making and risk assessment in car insurance. Data mining is capable of extracting this important information and it can also justify the investments of insurance companies in data.

Standard methods for risk classification in car insurance are usually based on risk factors such as type of vehicle, age, region, etc. However, the main problem is dispersion of data to large number of classes, and this leads to small number of examples in a class. A data-driven clustering approach to risk classification can provide necessary massiveness and homogeneity within the class as well as the heterogeneity between different classes (Smith, Willis, & Brooks, 2000; Yeo, Smith, Willis, & Brooks, 2001).

This article presents the possibilities, advantages and disadvantages of risk assessment and prediction in car insurance, with application of data mining techniques such as clustering, SVR and KLR.

The second section provides the review of papers dealing with similar issues. Third section defines the concept of risk in car insurance and discusses standard methods for risk assessment and their shortcomings. Data mining techniques, used in this article for risk prediction, are explained in section four. We present the capabilities of these techniques on data of the insurance company Sava Montenegro in section five. We start with a complete data-set which is clustered to homogenous clusters, i.e., clusters with similar amounts of claims. Expected claim sizes for identified clusters are predicted with SVR. For calculation of net risk premium, besides the amount, the probability of claim occurrence is important, too. It is estimated using KLR. In this section we also discussed the obtained results, advantages and disadvantages of the applied data mining methods for risk prediction on a small insurance market such as Montenegrin. Conclusions and future research are discussed in the last section.

2. Related work

There are numerous papers dealing with risk assessment in car insurance.

The heuristic approach implies that the insurance companies categorise policy owners to several different groups depending on the risk factors such as territory, age, sex, type of vehicle, etc. and also on basis of historical data on policies. Some scientific papers already researched this approach. So, for example, Samson and Thomas (1987) selected four factors and categorised each factor to additional three levels, which in total gives 81 (3^4) classes. On each of these classes they assessed the claim sizes using the linear regression. As mentioned above, the main problem in this model is dispersion of data to large number of classes, and this leads to small number of examples in a class. It is known that the main requirement for accuracy in prediction is volume and homogeneity of data within the class as well as the heterogeneity between the classes. With increase in the number of factors being considered, this problem is even more emphasised, because the number of classes increases drastically. Because of this, the approach determined with in advanced defined factors is the limiting one. Yeo et al. (2001) used clustering of 13 variables in their paper. With this method, they got a total of 30 classes containing between 1000 and 20,000 policy owners. The policy owners within a class had very similar amounts of claims, while the average claim sizes between different classes differ significantly. In other words, the conditions of

massiveness and homogeneity were met in their classification. Due to comparison they also used the heuristic method. They have taken only three factors which were divided on five classes. So, they created 125 (5^3) classes with at least 10,000 policies, which were much less discriminatory in relation to claim sizes. It turned out that the prediction results for claim sizes were much better on clusters.

The clustering technique was also used by Williams and Huang (1997) for identification of policy owners with high claim sizes. They combined the clustering with decision tree. In other words, first they obtained the classes using clustering. Then they used the decision trees to generate descriptions of those classes. Smith et al. (2000), suggested that the clusters should be used for prediction of claim sizes by data mining techniques. In order to predict the claim sizes, Chapados et al. (2002) used more sophisticated data mining techniques such as neural networks (NN).

Recent studies have perceived that a mixed discrete-continuous model may be appropriate to estimate claims and risk in insurance data (Christmann, 2004, 2005; Heller, Stasinopoulos, & Rigby, 2006; Parnitzke, 2008). According to Parnitzke (2008), the model explicitly specifies a logit-linear model for the claim occurrence (i.e. claim probability) and linear regression model for the mean claim size. GLMs and more flexible Tweedie's compound Poisson models are often used to construct insurance tariffs (Bortoluzzo et al., 2011; Ohlsson & Johansson, 2010). However, even these more general models still can yield problems in modelling high-dimensional relationships which is quite common for insurance data-sets. Namely, many explanatory variables are discrete which is quite common for insurance data-sets (if there are eight discrete explanatory variables each with eight different values there are approximately $8^8 \approx 16.7$ million interaction terms possible). The best modelling in these circumstances is one which using nonparametric methods from machine learning and data mining such as SVR and KLR (Christmann, 2004, 2005) or tree-based gradient boosting methods (Guelman, 2012; Yang, Qian, & Zou, 2015).

In recent years many papers have dealt with the application of data mining methods for loss cost estimation and risk analysis in insurance (Gepp, Wilson, Kumar, & Bhattacharya, 2012; Liu, Wang, & Lv, 2014; Paglia & Phelippe-Guinvarc'h, 2011).

3. Risk in car insurance and methods for its assessment

Risk assessment is very important for insurance companies. Determination of premium level based on assessed risk (net risk premium), enables insurance company to avoid negative selection, i.e. to lose good clients due to high premiums.

According to Mayer (2002), based on actuarial equivalence principle, net risk premium is equal to expected claim level. According to Renshaw (1994) and Parnitzke (2008), expected claim size is calculated as product of predicted claim size and probability that at least one claim will occur, given in relation (1).

$$E\left(ClaimSize_i\right) = Predicted\ Claim\ Size_i * P\left(ClaimOccur_i\right) \tag{1}$$

Probability for occurrence of at least one claim, in insurance practice, is evaluated with logistic regression (Parnitzke, 2008), given in relation (2).

$$Yi = \ln\left(P\left(ClaimOccurr_i\right)/\left(1 - P\left(ClaimOccur_i\right)\right)\right) = \alpha + \sum_J X_{IJ}\beta_J$$

$$P\left(ClaimOccur_i\right) = \frac{e^{Yi}}{1+e^{Yi}} \tag{2}$$

The claim size can be estimated with linear regression (Parnitzke, 2008), given in relation (3).

$$Predicted\ ClaimSize_i = \beta_0 + \sum_J X_{IJ}\beta_J \tag{3}$$

The X_{IJ} is value of variable which represent risk factors (tariff criteria) and α, β_0, β_J are regression coefficients.

However, the dependencies between risk factors and the claim size, usually is not linear or even monotonic (Christmann, 2004). Classic GLS and more flexible Tweedie's compound Poisson models have lower predictive performance for unknown claim sizes than nonparametric methods. Paglia & Phelippe-Guinvarc'h (2011), compared GLS with nonparametric tree boosted (TB) and NN. They concluded that considered nonparametric methods have better predictive performance (GLS Mean Squared Error [MSE] = 485,685; NN MSE = 473,112; TB MSE = 459,099). Generally, the SVM has better predictive performance than other techniques, especially for data that exhibit nonlinearity (Tian et al., 2012).

According to Christmann (2005), because of the large number of possible values of risk factors, even for data-sets with several million customers, it is not possible to estimate simultaneously all the interaction terms with these classic statistic methods, because the number of interaction terms increases too fast. A nonparametric approach based on a combination of KLR and SVR was able to detect an interesting interaction term and violations of a monotonicity assumption without the necessity that the researcher has to model interaction terms or polynomial terms manually. These methods don't explicitly obtain the intensity of the factor impacts, but the impact can be implicitly shown. Christmann (2005) presented the expected claim size stratified by the age of the main user or by gender and age of the main user. Looking at these dependencies, implicitly can be seen the impact of individual factors and of the interaction terms to the net risk premium.

In standard methods of car insurance, policies are classified based on risk factors such as age, region, type of vehicle, etc. Also, if we take into consideration the bonus-malus classes, which are determined based on history of policies and claims, it is clear that this approach leads to large number of tariff classes. This leads to dispersion of data, so that classes contain very little experiences related to risk and claim sizes.

If the classes have a higher number of policies with similar levels of claim sizes, such classes could be better for prediction of claim sizes, i.e. of premium level. Data clustering can provide necessary massiveness and homogeneity within the class, as well as the heterogeneity between different classes.

4. Description of data mining techniques and tools

Data mining techniques used in this article are clustering, KLR and SVR. In the previous section we explained the advantage of these methods compared to standard methods.

Clustering finds similar groups within the data-set. In this article we used the k-mean clustering method. It forms k-clusters iteratively, using functions for evaluation of clusters distances and their mean values. Mean values are initially set for all clusters. For a specific

data point, distances from the cluster mean values are calculated and the data point is associated to cluster with the smallest distance. In that cluster, mean value is calculated again, taking into consideration this newly added data. The procedure is iteratively repeated, for all data from the initial data-set.

Regression is used for the continuous target (dependent) variable prediction based on the predictor (independent) variables in a data-set. For that purpose, we used SVR and KLR.

Logistic regression is used for prediction of binomial (0,1) or categorical (with limited set of categories) target variable based on the predictor variables, where the data-set is classified to as many classes as the target variable has values.

In the case of binomial target variable, logistic regression predicts a continuous variable *p*, i.e., probability that the target variable value is 1 (success probability). In order to transform the regression into linear form, logistic function ln is used. Logistic regression model is given in the following form (4).

$$Y_i = ln\left(p/\left(1-p\right)\right) = \alpha + \sum_J X_{ij}\beta_J \tag{4}$$

Coefficients are evaluated with method of maximal credibility, which maximises probability p. The method uses iterative calculation of coefficients. When the coefficients are calculated, the probability p can be obtained according to relation (5)

$$p = \frac{e^{Yi}}{1+e^{Yi}} \tag{5}$$

KLR is a nonlinear form of logistic regression, which is obtained by replication of data vector using the kernel function. In this article we used KLR based on fast dual algorithm (Keerthi, Duan, Shevade, & Poo, 2005). Ruping (2003) implemented this algorithm in form of programme MyKLR.

The SVR was introduced by Vapnik (1995), and it is a regression technique that generally produces accurate nonlinear models.

The main concept employed by SVR is that the data vectors, which are not linearly related in the original space, can be mapped to higher or infinite dimensional space (feature space) where their linear relation is possible. In the epsilon SVR (ε–SVR), the goal is to find a hyperplane in feature space that has at most ε deviation from the actually obtained targets y_i for all the training data.

The geometrical margin corresponds to the shortest distance between the closest data points (support vectors) and the hyperplane, and SVR aims to find the hyperplane that minimises this distance. The margin maximisation process increases the generalisability of the support vector machine (SVM). The SVM aims to maximise the accuracy using training data and it also retains sufficient space for the correct prediction of future data. A SVM needs to solve an optimisation problem to find the maximum margin hyperplane, which requires the calculation of the dot product in the feature space.

The mapping of the data vector to the feature space does not have to be defined explicitly. It is sufficient to design it to facilitate the calculation of the dot product in terms of the input space variables, i.e., the dot product derived from the feature space is represented by a kernel function (which meets Mercer's condition) in the input space. This procedure is known as the kernel trick, which allows calculations to be made in the input space instead

of calculating the dot product in the feature space. The most frequently used kernel functions are as follows:

- Linear kernel: $K(x_i,x_j)=x_i^T x_j$
- Radial basis function (RBF) kernel: $K(x_i,x_j)=\exp(-\gamma||x_i-x_j||^2)$
- Polynomial kernel: $K(x_i,x_j)=(x_i^T x_j+1)^d$

Thus, the problem of finding the maximum margin hyperplane is converted into following dual quadratic optimisation problem:

$$max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{6}$$

$$\sum_{i=1}^{n} y_i \alpha_i = 0$$

$$0 \leq \alpha_i \leq C, i = 1, \ldots, n$$

where α_i are Lagrange multipliers, n is the number of training examples and C is a parameter, which is adjusted to trade off margin maximisation against regression error minimisation.

In classic SVR, the proper value for the parameter ε is difficult to determine beforehand. Fortunately, this problem is partially resolved in a new algorithm, nu-SVR (ν - SVR), in which ε itself is a variable in the optimisation process and is controlled by another new parameter $\nu \in (0,1)$. Parameter ν is the upper bound on the fraction of error points or the lower bound on the fraction of points inside the ε -insensitive tube. Thus, a good ε can be automatically found by choosing ν, which adjusts the accuracy level to the data at hand. This makes ν a more convenient parameter than the one used in ε–SVR.

Fan, Chen, and Lin (2005), proposed SVM learner which has been implemented using LIBSVM software since version 2.8 (Chang & Lin, 2011). LIBSVM supports the ε –SVR and ν – SVR. In this article we used that SVM learner.

There are numerous applications of SVM in different areas of economics (Kašćelan, Kašćelan, & Jovanović, 2015; Tian et al., 2012), as well as in insurance (Christmann, 2005; Marin-Galiano & Christmann, 2004).

In this article we used an open source data mining platform Rapid Miner (RM) (www.rapidminer.com), as a tool. We used the data mining platform RM for k-mean clustering, as well as for SVR and KLR. In RM the KLR is implemented as a Java implementation of MyKLR programme. SVR is implemented as a LIBSVM learner. To evaluate the models we used a RM-Split Validation operator (split ratio=0.7), with stratified sampling. It randomly splits up the example set into a training set and a test set.

5. A case study

In this section, we describe one case study of risk assessment in car insurance using data mining techniques. We describe the used data and the results of applied data mining techniques to these data. We make the comparison and discuss the results. Also, we point to advantages and disadvantages of data-driven approach compared to standard methods for risk assessment.

5.1 Description of data

In the research we used motor third party liability data for 2009, 2010 and 2011 from the insurance company Sava Montenegro. The data include 35,521 policies, out of which only 3528 policies are with total claim sizes other than zero. We took the appropriate number of policies without any claims. The size of this data-set is 7285 records. We used the following policies data: region, age, sex, type of vehicle, number of claims per policy, years of policy ownership, insured cases number for a user and average claim size.

In the process of preparation, data were purged. We removed records with unknown values and age is categorised into *Old (over the age of 65)*, *Young (up to the age of 25)* and *Middle (aged 25–65)* age. Policies with extremely low and extremely high average claim sizes are removed. Due to regression method the categorical variables with multiple categories are replaced with dummy (indicator) variables. Some of the parameters which describe the initial data-set are presented in Table 1.

5.2 Clustering of data and estimation of claim size

As initial data-set for claim sizes prediction, we took only policies with one or more claims (3021 records). In order to get homogenous groups of policies this data-set is clustered. The result of clustering is 12 clusters.

Then we applied SVR on such defined clusters. Target variable is Avg Claim Costs, while remaining 10 variables from the initial data-set are predictor variables.

In order to define SVR models, for optimisation of parameters (gamma, C, nu and epsilon) we used a 'grid-search' strategy. We realised it using the RM operator Optimise Parameters (Grid). We divided the data within each cluster into the training and test set in proportion of 70%:30%. Performance vector is obtained from the test data-set (unknown data). The defined SVR models and their predictive performance are presented in Table 2.

The results of the performance vectors, from Table 2, show that relative errors are less than 10% for most of clusters. Cluster 9, with highest claim sizes and small number of policies (only 23), has the largest relative error (14.00% +/- 5.94%).

Table 3 shows results we obtained applying the SVR models to the full clusters (with 30% of unknown data). Average claim costs per clusters are presented in the first column. The model provided deviations lower than 10% for 69% of data, while for 95% of data, deviations were lower than 20%.

Table 1. Initial data-set metadata.

Role	Name	Type	Statistics	Range	Missing
Label	Avg Claim Costs	numeric	avg = 1,338.330 +/– 2,525.931	[100.000; 33,000.000]	0
Regular	Mid Region	integer	avg = 0.600 +/– 0.490	[0.000; 1.000]	0
Regular	North Region	integer	avg = 0.216 +/– 0.412	[0.000; 1.000]	0
Regular	Mid Age	integer	avg = 0.867 +/– 0.340	[0.000; 1.000]	0
Regular	Young	integer	avg = 0.083 +/– 0.276	[0.000; 1.000]	0
Regular	Male	integer	avg = 0.863 +/– 0.344	[0.000; 1.000]	0
Regular	Motorcycle	integer	avg = 0.010 +/– 0.101	[0.000; 1.000]	0
Regular	Car	integer	avg = 0.946 +/– 0.227	[0.000; 1.000]	0
Regular	Number of Claims	integer	avg = 1.081 +/– 0.305	[1.000; 3.000]	0
Regular	Years of Ownership Policy	integer	avg = 2.361 +/– 0.753	[1.000; 3.000]	0
Regular	Number of Insurance Cases	integer	avg = 1.122 +/– 0.422	[1.000; 5.000]	0

Source: Authors' calculations.

Table 2. Support vector regression models and their predictive performance.

Clusters	SVM Parameter Set: (obtained from training set -70% of data)	PerformanceVector (obtained from test set -30% of data)
Cl_0	SVM.svm_type = nu-SVR	root_mean_squared_error: 82.195 +/− 0.000
	SVM.gamma = 1.0	absolute_error: 70.772 +/− 41.802
	SVM.C = 400.06000000000006	relative_error: 8.89% +/− 4.86%
	SVM.kernel_type = rbf	normalised_absolute_error: 0.930
	SVM.nu = 0.6	root_relative_squared_error: 0.937
	SVM.epsilon = 1.0	prediction_average: 790.278 +/− 87.759
Cl_1	SVM.svm_type = nu-SVR	root_mean_squared_error: 45.902 +/− 0.000
	SVM.gamma = 0.2008	absolute_error: 37.006 +/− 27.158
	SVM.C = 200.08	relative_error: 6.67% +/− 4.74%
	SVM.kernel_type = rbf	normalised_absolute_error: 0.698
	SVM.nu = 0.5	root_relative_squared_error: 0.754
	SVM.epsilon = 0.001	prediction_average: 547.902 +/− 60.885
Cl_2	SVM.svm_type = nu-SVR	root_mean_squared_error: 44.099 +/− 0.000
	SVM.gamma = 1.0	absolute_error: 35.720 +/− 25.861
	SVM.C = 600.04	relative_error: 10.53% +/− 7.81%
	SVM.kernel_type = rbf	normalised_absolute_error: 0.939
	SVM.nu = 0.5	root_relative_squared_error: 0.958
	SVM.epsilon = 0.8002	prediction_average: 346.035 +/− 46.036
Cl_3	SVM.svm_type = nu-SVR	root_mean_squared_error: 363.327 +/− 0.000
	SVM.gamma = 0.4006	absolute_error: 292.081 +/− 216.090
	SVM.C = 400.06000000000006	relative_error: 7.50% +/− 5.32%
	SVM.kernel_type = rbf	normalised_absolute_error: 0.826
	SVM.nu = 0.5	root_relative_squared_error: 0.791
	SVM.epsilon = 0.2008	prediction_average: 3866.672 +/− 459.503
Cl_4	SVM.svm_type = nu-SVR	root_mean_squared_error: 962.335 +/− 0.000
	SVM.gamma = 0.4006	absolute_error: 444.481 +/− 853.536
	SVM.C = 1000.0	relative_error: 5.57% +/− 10.04%
	SVM.kernel_type = rbf	normalised_absolute_error: 0.450
	SVM.nu = 1.0	root_relative_squared_error: 0.640
	SVM.epsilon = 0.8002	prediction_average: 6357.550 +/− 1503.147
Cl_5	SVM.svm_type = nu-SVR	root_mean_squared_error: 120.336 +/− 0.000
	SVM.gamma = 0.001	absolute_error: 108.077 +/− 52.916
	SVM.C = 400.06000000000006	relative_error: 9.52% +/− 4.42%
	SVM.kernel_type = rbf	normalised_absolute_error: 1.000
	SVM.nu = 0.7	root_relative_squared_error: 1.000
	SVM.epsilon = 0.8002	prediction_average: 1138.717 +/− 120.387
$Cl_{6.0}$	SVM.svm_type = nu-SVR	root_mean_squared_error: 17.935 +/− 0.000
	SVM.gamma = 0.001	absolute_error: 14.932 +/− 9.935
	SVM.C = 800.0200000000001	relative_error: 11.86% +/− 8.84%
	SVM.kernel_type = rbf	normalised_absolute_error: 1.008
	SVM.nu = 0.6	root_relative_squared_error: 1.003
	SVM.epsilon = 0.6003999999999999	prediction_average: 132.211 +/− 17.874
$Cl_{6.1}$	SVM.svm_type = nu-SVR	root_mean_squared_error: 11.003 +/− 0.000
	SVM.gamma = 1.0	absolute_error: 8.670 +/− 6.775
	SVM.C = 400.06000000000006	relative_error: 3.56% +/− 2.82%
	SVM.kernel_type = rbf	normalised_absolute_error: 0.861
	SVM.nu = 1.0	root_relative_squared_error: 0.921
	SVM.epsilon = 0.6003999999999999	prediction_average: 245.318 +/− 11.946
$Cl_{6.2}$	SVM.svm_type = nu-SVR	root_mean_squared_error: 12.673 +/− 0.000
	SVM.gamma = 1.0	absolute_error: 10.269 +/− 7.427
	SVM.C = 0.1	relative_error: 5.24% +/− 4.10%
	SVM.kernel_type = rbf	normalised_absolute_error: 0.968
	SVM.nu = 0.6	root_relative_squared_error: 0.996
	SVM.epsilon = 0.001	prediction_average: 201.381 +/− 12.718
Cl_7	SVM.svm_type = nu-SVR	root_mean_squared_error: 148.115 +/− 0.000
	SVM.gamma = 0.6003999999999999	absolute_error: 122.093 +/− 83.853
	SVM.C = 200.08	relative_error: 7.45% +/− 5.39%
	SVM.kernel_type = rbf	normalised_absolute_error: 0.807
	SVM.nu = 0.5	root_relative_squared_error: 0.824
	SVM.epsilon = 0.8002	prediction_average: 1658.751 +/− 179.792

(*Continued*)

Table 2. (*Continued*).

Clusters	SVM Parameter Set: (obtained from training set -70% of data)	PerformanceVector (obtained from test set -30% of data)
Cl_8	SVM.svm_type = nu-SVR SVM.gamma = 0.6003999999999999 SVM.C = 1,000.0 SVM.kernel_type = rbf SVM.nu = 0.5 SVM.epsilon = 0.8002	root_mean_squared_error: 240.754 +/– 0.000 absolute_error: 193.130 +/– 143.748 relative_error: 7.49% +/– 4.85% normalised_absolute_error: 0.754 root_relative_squared_error: 0.762 prediction_average: 2467.624 +/– 315.976
Cl_9	SVM.svm_type = nu-SVR SVM.gamma = 0.2008 SVM.C = 1,000.0 SVM.kernel_type = rbf SVM.nu = 0.8 SVM.epsilon = 0.001	root_mean_squared_error: 4,225.478 +/– 0.000 absolute_error: 3,544.845 +/– 2299.725 relative_error: 14.00% +/– 5.94% normalised_absolute_error: 0.723 root_relative_squared_error: 0.774 prediction_average: 24103.856 +/– 5462.292

Source: Authors' calculations.

Policies with an unknown claim size can be joined to the appropriate clusters based on the probability of belonging. This probability can be determined using logistic regression based on the relation (5).

The SVR models in Table 2 give Predicted Claim Size$_i$ inside the clusters. According to relation (1), this size is one of the factors necessary for calculation of expected claim size, i.e., net risk premium. The second factor is probability of claim occurrence, for which the model will be defined in following section.

5.3 Estimation of claim occurrence probability

For prediction of occurrence of at least one claim we used the logistic regression. Target variable is **Number of Claims (0/1)** and its values are 0 or 1 (for policies with one or more claims value is set to 1). For this analysis, we used all 7,284 records. Using the KLR procedure (kernel type='dot', C=1.0 – these parameters are determined using 'grid-search' strategy), we obtained the model in Table 4. Because of the applied linear (dot) kernel, the resulting weight coefficients can be interpreted as the intensity of factor impacts to the target variable.

Table 5 shows the performance of this model. It can be seen from the Confusion Matrix that prediction of claim occurrence has the accuracy of 92.69%. In other words, in the test set out of 643 policies with claims, 47 of them are misclassified. Overall accuracy of the model is 76.70%.

According to relation (1), for calculation of net risk premium it is necessary to have P(ClaimOccur$_i$). This value is calculated based on the model from Table 4 as well as on relation (2).

5.4 Analysis and discussion of the results

Standard method for calculation of premium for motor third party liability in Montenegro is defined by a system of premium tariffs which is adopted by Montenegro National Bureau of Insurers. This system defines eight basic tariff groups, depending on the type of vehicle. Each of these groups is divided on subgroups depending on the engine power, bearing capacity for cargo vehicles, type of transportation for buses, type of trailer and also purpose of special and work vehicles. For each of the subgroups, they have defined three bonus

Table 3. Results of support vector regression models for prediction of claim size.

	Avg Claim Costs per Cluster	No Dev <10%	% Dev <10%	No Dev <20%	% Dev <20%
Cluster 0: 295 items	792.120	179	61%	290	98%
Cluster 1: 472 items	549.469	374	79%	471	100%
Cluster 2: 561 items	345.604	268	48%	485	86%
Cluster 3: 133 items	3916.260	98	74%	127	95%
Cluster 4: 193 items	6227.533	162	84%	182	94%
Cluster 5: 225 items	1150.810	133	59%	225	100%
Cluster6.0: 246 items	135.546	117	48%	205	83%
Cluster 6.1: 229 items	246.330	224	98%	229	100%
Cluster 6.2: 294 items	201.628	258	88%	294	100%
Cluster 7: 192 items	1672.738	158	82%	189	98%
Cluster 8: 158 items	2437.068	124	78%	153	97%
Cluster 9: 23 items	22727.100	5	22%	17	74%
Total number of items: 3,021	1338.330	2100	69%	2,867	95%

Source: Authors' calculations.

Table 4. Model of logistic regression for prediction of claim occurrence.

Weighting Coefficients
Bias (offset): 0.113
w[Years of Ownership Policy] = 2.575
w[Number of Insurance Cases] = −0.389
w[Mid Region] = −0.549
w[North Region] = −0.108
w[Mid Age] = −0.658
w[Young] = 0.113
w[Male] = −0.804
w[Motorcycle] = 0.093
w[Car] = −0.065
w[Buses] = −0.012
w[Special motor vehicles] = −0.005
w[Towing vehicles] = 0.067
w[Trailers] = −0.002
w[Trucks] = −0.046

Source: Authors' calculations.

Table 5. Performance of logistic model for prediction of claim occurrence.

	True 1	True 0	Class Precision (%)
Pred. 1	596	47	92.69
Pred. 0	462	1080	70.04
Class Recall	56.33%	95.83%	Accuracy: 76.70

Source: Authors' calculations.

premium classes, one basic class and three malus classes, i.e., seven premium classes, and all together that produces large number of tariff classes.

Taking into consideration that the Montenegro insurance market is quite small, a very small number of policies will fall into appropriate tariff class. Average claim sizes for certain tariff classes are 0, and this means that there are no policies with claims. The question is, if the small number of policies can provide satisfactory predictions. In these conditions, tariff classes are unusable for estimation of claim size, i.e. for calculation of net risk premium. It became obvious that certain, more sophisticated methods, have to be applied.

Applying clustering method, insurance policies are classified into 12 homogeneous groups (with a similar claim sizes) containing enough data for claim size estimation with high accuracy. With analysis of the prediction results, it can be seen that the majority of the models has accuracy from approximately 80% to 95%.

In practice, the claim size is not always known exactly (if a big accident occurs in December, the exact claim size will often not be known at the end of the year and perhaps not even at the end of the following year). In order to construct a new insurance tariff for the next year, in this case, a statistician will have to use appropriate predictions of the exact claim size. Hence, the empirical distribution of the claim sizes is, in general, a mixture of really observed values and of estimated claim sizes (Christmann, 2004).

SVR models in Table 2 have good predictive performance (relative_error < 10% for most of clusters) on the test data-set for which the claim sizes are unknown (the generality of the model is achieved by adjusting parameter C). So, SVR can be successfully applied to predict the unknown claim sizes on the small data-sets, too.

Model of linear regression on clusters, used by Yeo et al. (2001), provided for 57% of data deviations lower than 10%, while for as much as 90% of data, deviations were lower than 20%. Our model of SVR provided deviations lower than 10% for 69% of data, while for 95% of data deviations were lower than 20% (Table 3). In a previously mentioned paper they used the sample of 146,326 policies with claims. Our data-set contained only 3528 policies with claims. This shows that the method of clustering data and SVR can achieve good results on a small data-set.

The model for prediction of risky policies from Table 4, shows that positive impact on claim occurrence probability have: years of policy ownership, young policy holders, motorcycle and towing vehicles. The model provided accuracy of around 80%, although the percentage of policies which are predicted as risk-free, but they are in a fact risky, is around 30%, and that is quite a high percentage (Table 5). However, with analysis of our data from 2011, we have determined that out of 1182 policies with claims in that year, only 52 policies had claims in previous two years. This means that 1130 out of 1182 (95.6%) of risky polices were recognised as risk-free, based on standard methods for premium calculation. These policies included even bonuses. Models obtained using data mining methods have significantly better accuracy than the risk assessment based on premium tariff tables, which is the standard method for motor third party liability for domestic insurance companies.

Still, our approach can be used not to replace, but to complement, traditional methods, which are used in practice. This approach will be especially important since 2017, when is planned to be introduced premium liberalisation in the Montenegrin car insurance market.

However, this approach also has its own disadvantages. One of the main problems, which we have noticed in this article, is insufficient quantity of data. In other words, on small markets such as Montenegrin, the number of policies is too small to have models with high

accuracy. Also, we noted the lack of certain data which could have significant influence to premium predictions. So, for example, in car insurance, if we would have accurate data about policy owners, such as occupation, wealth, tendency for use of alcohol, condition of health, habits, etc. the prediction itself would be more accurate. Models for prediction of claim occurrence, i.e. for risk classification, classify policies to risky and risk-free. However, within the risky policies there are levels of risk depending on the number of claims. The model which predicts the level of risk, i.e. makes the classification according to number of claims, would be much more useful. But, in small initial data-set, like in our case, this prediction does not provide good results.

So, the quality and availability of data are the most important presumptions for success of data mining process. The other problem that appears is adequacy of applied model, i.e., is it good enough for predictions related with specific data (Pichler, 2014). Selection of appropriate model of certain data-set is precondition for good results of data mining process. Approach of clustering combined with SVM regression has good performance on a small number of policies, such as in our example.

6. Conclusion

In this article we have discussed the methods for risk assessment in car insurance. Standard methods imply classification of policies to large number of tariff classes and calculation of premiums based on them. Using data-driven methods it is possible to get better results in risk assessment and premiums estimation.

On the case study data we proved that nonparametric data mining methods, with better accuracy than the standard methods, can predict claim sizes and occurrence of claims and this represents the basis for calculation of net risk premium. In this approach, the level of premium is not determined based on tariff classes. Using clustering method we classified policies to groups with same level of risk, without fragmentation of data into too high number of small groups. We predicted expected claim size using the SVR method and claim occurrence using KLR. We achieved the prediction accuracy of around 80% or more, on a small data-set, where 30% of the policies have unknown claim sizes.

The main advantage of the proposed approach, even in a small data-set, is its good predictive performance for unknown claim sizes, which is common in car insurance. Also, this research is important for the Montenegrin insurance companies, due to expected premium liberalisation in 2017, because they will be able to use their own methods for risk prediction.

The proposed approach has its drawbacks, which are reflected mainly in the lack of data from small and still underdeveloped market such as Montenegrin.

Some future research could determine how much the use of some other data mining techniques would contribute to better results in car insurance risk assessment, especially on small data-sets, where the dependencies are harder to notice. Analysis and prediction of customer loyalty (churn prediction) in car insurance, using data mining techniques, would also contribute to better risk assessment.

References

Bortoluzzo, A. B., Claro, D. P., Caetano, M. A. L., & Artes, R. (2011). Estimating total claim size in the auto insurance industry: A comparison between tweedie and zero-adjusted inverse gaussian distribution. *BAR-Brazilian Administration Review, 8*, 37–47.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(27), 1–27. Retrieved from http://www.csie.ntu.edu.tw/cjlin/libsvm

Chapados, N., Bengio, Y., Vincent, P., Ghosn, J., Dugas, C., Takeuchi, I., & Meng, L. (2002). *Estimating car insurance premia: A case study in high-dimensional data inference, advances in neural information processing systems* (Vol. 14). Cambridge: MIT Press.

Christmann, A. (2004). An approach to model complex high–dimensional insurance data. *Allgemeines Statistisches Archiv, 88*, 375–396.

Christmann, A. (2005). On a strategy to develop robust and simple tariffs from motor vehicle insurance data. *Acta Mathematicae Applicatae Sinica, English Series, 21*, 193–208.

David, M. (2015, April). Automobile insurance pricing with generalized linear models. *Proceedings in GV-Global Virtual Conference* (No. 1).

Fan, R. E., Chen, P. H., & Lin, C. J. (2005). Working set selection using second order information for training support vector machines. *The Journal of Machine Learning Research, 6*, 1889–1918.

Gepp, A., Wilson, J. H., Kumar, K., & Bhattacharya, S. (2012). A comparative analysis of decision trees vis-à-vis other computational data mining techniques in automotive insurance fraud detection. *Journal of data science, 10*, 537–561.

Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications, 39*, 3659–3667.

Heller, G., Stasinopoulos, M., & Rigby, B. (2006, July). The zero-adjusted inverse Gaussian distribution as a model for insurance claims. *Proceedings of the 21th International Workshop on Statistical Modelling* (226–233).

Kašćelan, L., Kašćelan, V., & Jovanović, M. (2015). Hybrid support vector machine rule extraction method for discovering the preferences of stock market investors: Evidence from Montenegro. *Intelligent Automation & Soft Computing, 21*(4), 1–20.

Keerthi, S. S., Duan, K., Shevade, S. K., & Poo, A. N. A. (2005). Fast dual algorithm for Kernel logistic regression. *Machine Learning, 61*, 151–165.

Liu, Y., Wang, B. J., & Lv, S. G. (2014). Using multi-class AdaBoost tree for prediction frequency of auto insurance. *Journal of Applied Finance and Banking, 4*, 45.

Marin-Galiano, M., & Christmann, A. (2004). *Insurance: An R-program to model insurance data* (No. 2004, 49). Technical Report/Universität Dortmund, SFB 475 Komplexitätsreduktion in Multivariaten Datenstrukturen.

Mayer, U. (2002). *Third party motor insurance in Europe*. Bamberg: University of Bamberg.

Ohlsson, E., & Johansson, B. (2010). *Non-life insurance pricing with generalized linear models*. Berlin: Springer.

Paglia, A., & Phelippe-Guinvarc'h, M. V. (2011). Tarification des risques en assurance non-vie, une approche par modele d'apprentissage statistique [Pricing of risks in non-life insurance, an approach by statistic learning models]. *Bulletin français d'Actuariat, 11*, 49–81.

Parnitzke, T. (2008). A discussion of risk assessment methods for the German automobile insurance industry. *Working papers on risk management and insurance*, 55, University of St. Gallen, Institute of Insurance Economics.

Pichler, A. (2014). Insurance pricing under ambiguity. *European Actuarial Journal, 4*, 335–364.

Renshaw, A. E. (1994). Modeling the claims process in the presence of covariates. *ASTIN Bulletin, 24*, 265–285.

Ruping, S. (2003). myKLR. Retrieved from http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYKLR

Samson, D., & Thomas, H. (1987). Linear models as aids in insurance decision making: The estimation of automobile insurance claims. *Journal of Business Research, 15*, 247–256.

Smith, K. A., Willis, R. J., & Brooks, M. (2000). An analysis of customer retention and insurance claim patterns using data mining: A case study. *Journal of the Operational Research Society, 51*, 532–541.

Tian, Y., Shi, Y., & Liu, X. (2012). Recent advances on support vector machines research. *Technological and Economic Development of Economy, 18*, 5–33.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.

Williams, G. J., & Huang, Z. (1997). Mining the knowledge mine: The hot spots methodology for mining large real world databases. *Lecture Notes in Artificial Intelligence, 1342*, 340–348.

Yang, Y., Qian, W., & Zou, H. (2015). A boosted tweedie compound poisson model for insurance premium. Retrieved from http://arxiv.org/abs/1508.06378

Yeo, A. C., Smith, K. A., Willis, R. J., & Brooks, M. (2001). Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry. *International Journal of Intelligent Systems in Accounting, Finance & Management, 10*, 39–50.