Allgemeines Statistisches Archiv 0, 0–21
© Physica-Verlag 0, ISSN 0002-6018

An approach to model complex high-dimensional insurance data

By Andreas Christmann [*]

Summary: This paper describes common features in data sets from motor vehicle insurance companies and proposes a general approach which exploits knowledge of such features in order to model high-dimensional data sets with a complex dependency structure. The results of the approach can be a basis to develop insurance tariffs. The approach is applied to a collection of data sets from several motor vehicle insurance companies. As an example, we use a nonparametric approach based on a combination of two methods from modern statistical machine learning, i.e. kernel logistic regression and $\varepsilon-$support vector regression.

Keywords: Classification, data mining, insurance tariffs, kernel logistic regression, machine learning, regression, robustness, simplicity, support vector machine, support vector regression. JEL: C14, C13, C10.

1. Introduction

Insurance companies need estimates for the probability of a claim and for the expected sum of claim sizes to construct insurance tariffs. In this paper we consider some statistical aspects for analyzing such data sets from motor vehicle insurance companies.

The main goal of the paper is to propose a general approach to analyze data sets from insurance companies with the emphasis in detecting and modelling a complex structure in the data which may be hard to detect by using traditional methods. Such a complex structure can be a high-dimensional non-monotone relationship between variables. Our approach can be used to complement – not to replace – traditional methods which are used in practice, see *e.g.* Bailey and Simon (1960) or Mack (1997). Generalized linear models are often used to construct insurance tariffs, see McCullagh and Nelder (1989), Kruse (1997) or Walter (1998). Tweedie's compound Poisson model, see Smyth and Jørgensen (2002), is more flexible than the generalized linear model because it contains not only a regression model for the expectation of the response variable Y but also a regression model for the dispersion of Y. However, even this more general model still can yield problems in modelling complex high-dimensional relationships, especially if many explanatory variables are discrete which is quite common for insurance data sets. *E.g.* if there are 8 discrete explanatory variables each with 8 different values millions of interaction terms are possible. One can circumvent some of the drawbacks which classical statistical methods have

Received: / Revised:

[*] This work was supported by the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") and by the Forschungsband DoMuS from the University of Dortmund. I am grateful to Mr. A. Wolfstein and Dr. W. Terbeck from the Verband öffentlicher Versicherer in Düsseldorf, Germany, for making available the data set and for many helpful discussions.

by using methods from statistical machine learning theory such as support vector regression or kernel logistic regression.

In Section 2 the statistical objectives are given. Section 3 describes characteristics of data sets from motor vehicle insurance companies. Section 4 contains the proposed approach. Section 5 briefly describes kernel logistic regression and $\varepsilon-$support vector regression, which both belong to statistical machine learning methods based on convex risk minimization, *cf.* Vapnik (1998). In Section 6 the results of applying the proposed approach to a data set from 15 motor vehicle insurance companies are described. Here, we use a non-parametric approach based on a combination of kernel logistic regression and $\varepsilon-$support vector regression. Section 7 contains a discussion.

2. Statistical objectives

In this section we describe common statistical problems in analyzing data from motor vehicle insurance companies.

An insurance company collects a lot of information for each single policy holder for each year or period. Often dozens or even hundreds of variables are available from each customer. Most of this information belongs to one of the following categories:

- personal information: e.g. name, surname, type of policy, policy number, other insurance policies
- demographic information: e.g. gender, age, place of residence, postal zip code, population density of the region the customer is living in, occupation type
- driver information: e.g. main user, driving distance within a year, car kept in a garage
- family information: e.g. age and gender of other people using the same car
- history: e.g. count and size of previous claims, property damage, physical injury, occurrence of a loss
- motor vehicle: e.g. type, age, strength of engine
- response information: claim (yes/no), number of claims, size and date of claim.

In practice, the claim size is not always known exactly. E.g. 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. Possible reasons are law-suits or the case of physical injuries. In this case, a statistician will have to use more or less appropriate estimations of the exact claim size to construct a new insurance tariff for the next year. Hence, the empirical distribution of the claim sizes is in general a mixture of really observed values and of estimated claim sizes. The mixing proportion is also stochastic.

Further, some explanatory variables can have imprecise values. *E.g.* there is a variable describing the driving distance of a customer within a year. The customer has to choose between some categories, say below 9000 kilometers, between 9000 and 12000 kilometers, between 12000 and 15000 kilometers, etc. There are reasons making it plausible that a percentage of these values are too small, because it is well-known that the premium of an insurance tariff increases for increasing values of this variable.

An insurance company is interested in determining the actual premium (gross premium) charged to the customer. In principle, the actual premium is the sum of the pure premium plus safety loadings plus administrative costs plus the desired profit. In this paper the focus will be on the *pure premium.*

Let n be the number of customers. For each customer i denote the number of claims during a year by n_i, the claim size of claim number j by $y_{i,j}$, and the number of days under risk by t_i, $1 \leq i \leq n$. For each customer compute

$$y_i = \frac{\sum_{j=1}^{n_i} y_{i,j}}{t_i/360} \ ,$$

which we call individual claim amount per year. Denote the corresponding random variables by N_i, $Y_{i,j}$, T_i, and Y_i. We assume here that $Y_{i,j} \geq 0$ almost surely. Useful information of explanatory variables is included in the vector $x_i \in \mathbb{R}^p$. Sometimes we omit the index i.

Our *primary response* is the pure premium

$$p^I(x) := \mathrm{E}(Y|X = x) \, , \quad x \in \mathbb{R}^p.$$

The *secondary response* is the conditional probability

$$p^{II}(x) := \mathrm{P}(Y > 0|X = x) \, , \quad x \in \mathbb{R}^p \, ,$$

that the policy holder will have at least one claim within one year given the information contained in the explanatory variables. Of course, we do not have full knowledge of the conditional distribution of $Y|X = x$.

An estimate $\hat{p}^I(x)$ of the pure premium should have the following four properties.

- It is *fair.* The estimator of the pure premium should be approximately unbiased, i.e. $\mathrm{E}\hat{p}^I(x) - p^I(x) \approx 0$, for the whole population and also in sub-populations of reasonable size.
- It has a *high precision.* One precision criterion is the mean squared error (MSE), which should of course be small. But other measures of precision may also be interesting, e.g. a trimmed version of the MSE or a chi-square statistic of Pearson-type for sub-populations of reasonable size.
- It is *robust.* Moderate violations of the statistical model assumptions and the impact of outliers have only a limited impact on the estimations.
- It fulfills *simplicity.* Very complex tariffs with many interactions terms may only have a reduced practical importance, because the acceptance of such a tariff would be low by many customers and perhaps also by the insurance company.

If a tariff is not fair, there are two cases. A high positive bias is of course bad from the view point of the policy holder, because the estimated premium is too high and he or she will have to pay too much money. But also an insurance company may not be interested in such a bias because there is a danger that the customer will turn to another company. An insurance company will usually avoid the case of a large negative bias. However a small negative bias may be acceptable for the company for a short time period if the company is interested in increasing its share of the market.

3. Characteristics of the data

In this section we describe some of the characteristics of a data set from the Verband öffentlicher Versicherer in Düsseldorf, Germany. The data set contains data from 15 insurance companies in non-aggregated structure. For reasons of data protection and confidentiality, there is no indicator variable describing which customer belongs to which insurance company. Overall, the data set has approximately a size of 3 GB as a compressed SAS data set and contains information from more than 4.6 million policy holders. A single policy holder may contribute more than one case to the data set, *e.g.* if he or she was involved in more than one claim. Using an identification number one can determine for each policy holder, whether he or she had a claim, sum up the claim sizes, and compute the individual claim amount per year, *i.e.* y_i. The data set contains more than 70 explanatory variables; most of them are discrete. Even nowadays, a reasonable statistical analysis approach has to take into account this considerable size and not all software packages can deal with such big data sets.

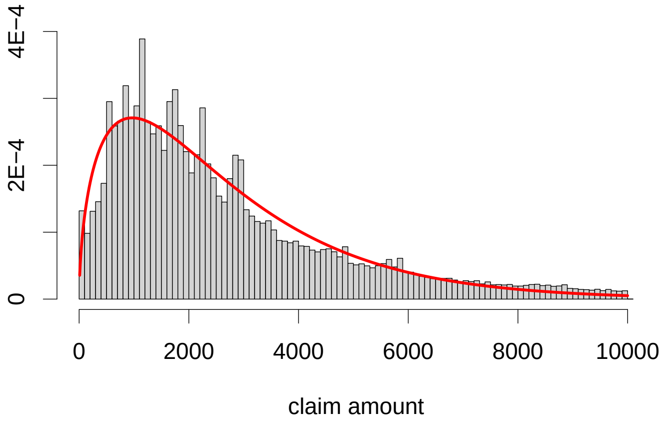
There are approximately $275,000$ claims overall. Approximately 94.37 percent of the customers had no claim, 5.35 percent had one claim, and 0.26 percent had two claims. The others had three or even more claims. Most claims are from the year 2001. Note that therefore some of the claim sizes are probably only estimates. Approximately 5 percent of the claims are older ones (up to 10 years).

An explanatory data analysis (EDA) shows that there are many complex dependencies between various variables. There are empty cells because not all combinations of the discrete explanatory variables are present, and of course, missing values occur as well. In the following we briefly mention some interesting results of the EDA for the (individual) claim amount values y_i. The total mean of the claim amount values taking no explanatory variables into account is approximately 360 EUR. However, the empirical distribution of these values has an atom in zero and is extremely skewed. Table 1 shows that approximately 95% of all policy holders had no claim within the period under consideration. Approximately 2.2% of the policy holders had a positive claim amount up to 2000 EUR, 2.4% had a moderate claim amount of size between 2000 EUR and 10000 EUR. Only 0.4% of the customers had a high claim amount between 10000 EUR and 50000 EUR, but the sum of their claim amount values contribute almost 20% to the

TABLE 1. Description of the individual claim amount.

Claim amount	% obs.	% of total sum	Mean	Median
total			364	0
0	94.9	0	0	0
(0,2000]	2.2	6.7	1097	1110
(2000,10000]	2.4	27.1	4156	3496
(10000,50000]	0.4	19.8	18443	15059
>50000	0.07	46.4	234621	96417

FIGURE 1. Histogram of claim amount. Only positive values less than or equal to 10000 EUR are shown.



overall sum. It is interesting to note that only 0.07% of all policy holders had a claim amount higher than 50000 EUR, but they contribute 46.4% to the grand sum. The maximum claim amount occurred for a customer who was approximately one month under risk and had a claim above 2.5 million EUR. The highest individual claim size was above 6 million EUR.

The distribution of the primary response variable is skewed even if one restricts consideration to the interval $(0, 10000]$ EUR, see Figure 1. The density of the best fitting gamma distribution based on maximum likelihood estimation is shown. This figure draws attention to four peaks in intervals near by 1200, 1800, 2400, and 3000 EUR. Although this is only a univariate description of the data, it is interesting to note that the distance between sequential peaks is of comparable size. One possible explanation is that there are some standard procedures or flat rates to deal with minor or moderate accidents such that claim sizes within these intervals occur with an increased probability. It is clear from Table 1, that extreme claim amount values above 50000 EUR contain a lot of valuable information concerning the pure premium, although the number of extreme claim amount values is small. Methods from extreme value theory can be helpful in this case. Consider

Table 2. Maximum likelihood estimates for the parameters ξ and β of the Generalized Pareto distributions used to model extreme claim amount values higher than 50000 EUR. Confidence intervals at the 95% level are given in parenthesis.

data	$\hat{\xi}$	95% CI	$\hat{\beta}$	95% CI
all	0.804	$(0.743, 0.865)$	50035.42	$(46813.48, 53257.37)$
knot 1	0.887	$(0.788, 0.985)$	53357.66	$(47959.25, 58756.08)$
knot 2	0.635	$(0.541, 0.728)$	51214.80	$(45919.53, 56510.07)$
knot 3	0.852	$(0.720, 0.984)$	44742.58	$(38588.19, 50896.97)$

the Generalized Pareto distribution (GPD) with distribution function

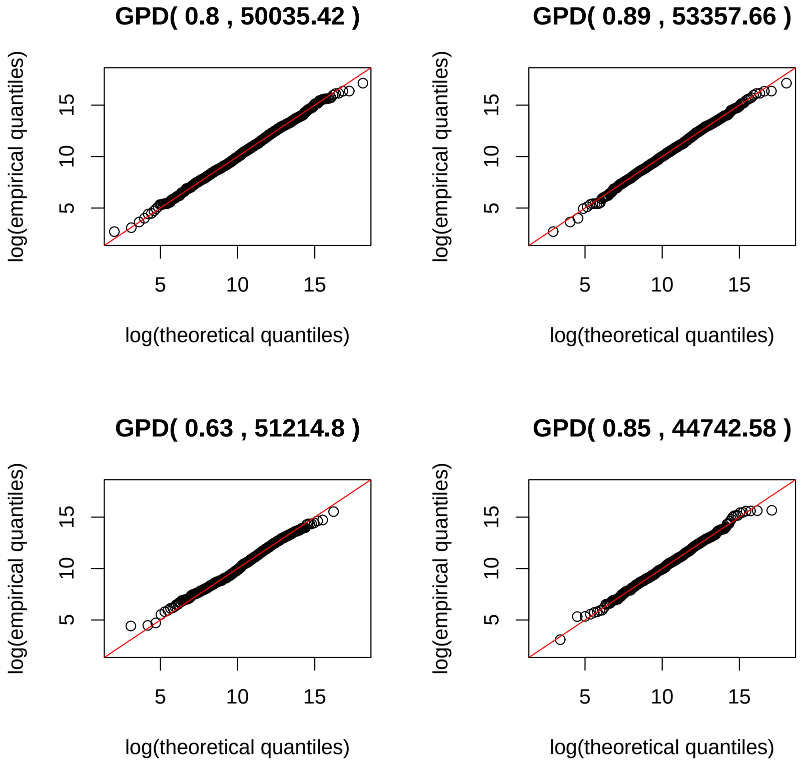
$$G_{\xi,\beta} = \begin{cases} 1-(1+\beta^{-1}\xi x)^{-1/\xi} & \text{if} \quad \xi \neq 0 \\ 1-\exp(-x/\beta) & \text{if} \quad \xi = 0\,, \end{cases}$$

where $x \geq 0$ if $\xi \geq 0$, and $x \in [0, -\xi^{-1}]$ if $\xi < 0$, *cf.* Embrechts *et al.* (1997) and Celebrián *et al.* (2003). The parameter ξ is called the Pareto index. The QQ plots given in Figure 2 show that the GPD can be useful to fit extreme claim amount values. The upper left plot shows a QQ plot on logarithmic scales for all 3343 claim amount values higher than 50000 EUR. The other three plots given in Figure 2 give the corresponding plots for claim amount values higher than 50000 EUR, which belong to the main 3 knots of a regression tree constructed on the basis of a few important explanatory variables. The Pareto index for the data set of knot 2 differs from the Pareto indices from the other two knots, see Table 2. Note that here we treat all extremes in a certain knot as realizations of independent and identically distributed random variables. Of course, one can fit more complex trees, but then the number of data points in the knots will decrease. We do not discuss such trees here, because the focus of this paper is not on the extreme values alone. Beirlant *et al.* (2002) proposed an alternative method to fit extreme claim amounts, see also Teugels (2003).

Figure 3 shows that the determination of the cutoff point can be crucial for the estimate of the Pareto index. Here it was assumed for simplicity, that all extreme values above the specified cutoff point are realizations of independent and identically distributed random variables. The maximum likelihood method was used to fit the GPD model. As before, such plots can also be made for the main knots of a tree. The cutoff point of 50000 EUR is used because it is similar to the cutoff point used by the Verband öffentlicher Versicherer. We like to emphasize that the i.i.d. assumption is an oversimplification for the present data set.

Figure 4 shows the relative frequency that a claim occurred and the claim amount stratified with respect to the age of the main user of the car. The smooth curve is the fit provided by $\varepsilon-$support vector regression, *cf.* Vapnik (1998). There is an interesting non-monotone relationship between the age of the main user and both response variables. It is well-known that young drivers in Germany, say between 18 and 24 years old, have an increased claim frequency. As was expected, the plot also shows an increased

Figure 2. QQ plots with respect to the Generalized Pareto distribution (GPD) for claim amount values higher than 50000 EUR. Upper left: for all such data points ($n = 3343$). Upper right and lower: for data points belonging to the main 3 knots of a regression tree ($n = 1416$, $n = 1175$, and $n = 752$, respectively).



claim frequency for elderly drivers. An interesting subgroup is the age group around 50 years. These main users also show an increased claim frequency and increased claim amount values. One plausible explanation is, that a considerable percentage of main users from this age group have children between 18 and 24 years old, who have already an own driving license but are using the parents' car.

Figure 5 suggests a monotone but non-linear relationship between population density of the region the customer is living in and the claim amount. The regression curve was fitted with $\varepsilon-$support vector regression, *cf.* Vapnik (1998), see also Section 5. However, there is additional geographical information in the data set not explained by the population density alone. A map of the averaged claim amount values computed in regions of Germany based on postal zip codes (not presented here) shows that there is a geographical clustering effect. Typically the claim amount is increased in big German cities, e.g. in Berlin, Hamburg, Munich or Frankfurt, as was to be expected from the previous figure. However, the so-called Ruhr-Area has

FIGURE 3. Impact of the cutoff point on the estimation of the Pareto index assuming an i.i.d. GPD model. Maximum likelihood estimation (solid curve) with pointwise confidence intervals at the 95% level (dotted curve).

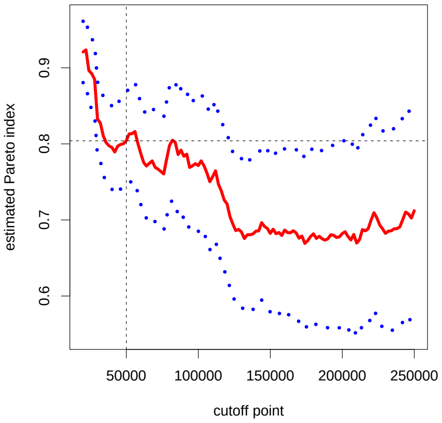
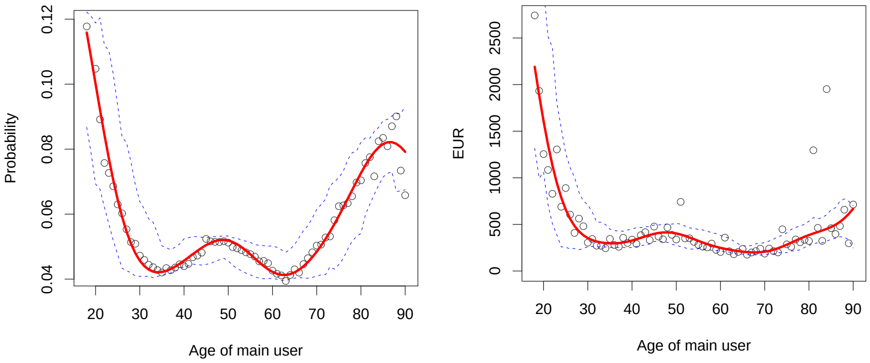


FIGURE 4. Relative frequency of claims and averaged claim amount values stratified by age of the main user. The solid curve joins the fitted values based on $\varepsilon-$support vector regression. The dashed curves show nonparametric bootstrap confidence bands at the 95% level.

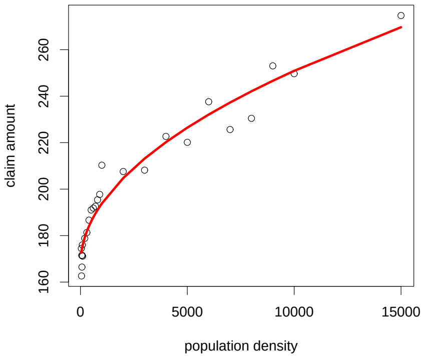


roughly a similar geographical size and also a similar number of inhabitants than Berlin, but the average of the claim amount values is much lower than in Berlin. On the other hand, in the north-east of Germany the mean claim sizes are increased, too, although the population density is relatively low.

4. APPROACH

The following general approach aims at detecting and modelling complex structures in data sets from motor vehicle insurance companies by exploiting certain characteristics of such data sets.

FIGURE 5. Scatterplot of claim amount vs. population density.



1. Most of the policy holders have no claim within a year or a certain period.
2. The claim sizes are extremely skewed to the right, but there is an atom at zero.
3. Extreme high claim amounts are rare events, but contribute enormously to the total sum.
4. There are only imprecise values available for some explanatory variables.
5. Some claim sizes are only estimates.
6. The data sets to be analyzed are huge.
7. There is a complex, high-dimensional dependency structure between variables.

Statistical procedures with good robustness properties are preferable due to the features 3 to 5 in the list. Due to point 7 methods from statistical machine learning are of interest.

As before, let Y_i denote the claim amount of customer i within a year and x_i the vector of explanatory variables. In the first step we construct an additional stratification variable C_i by defining a small number of classes for the values of Y_i with a high amount of interpretability. For example, define a discrete random variable C_i with five possible values:

$$C_i = \begin{cases} 0, & \text{if } \; Y_i = 0 & \text{'no claim'} \\ 1, & \text{if } \; Y_i \in (0, 2000] & \text{'low claim amount'} \\ 2, & \text{if } \; Y_i \in (2000, 10000] & \text{'medium claim amount'} \\ 3, & \text{if } \; Y_i \in (10000, 50000] & \text{'high claim amount'} \\ 4, & \text{if } \; Y_i > 50000 & \text{'extreme claim amount'.} \end{cases}$$

Of course, it depends on the application how many classes should be used and how reasonable boundaries can be defined. We will not address this problem here. In the following the index i is sometimes omitted.

4.1. APPROACH A.

Note that given the information that no claim occurred, it holds that

$$\mathrm{E}(Y|C = 0, X = x) \equiv 0\,. \tag{1}$$

The law of total probability yields

$$\mathrm{E}(Y|X=x) = \mathrm{P}(C>0|X=x) \times \sum\nolimits_{c=1}^{k} \mathrm{P}(C=c|C>0, X=x) \cdot \mathrm{E}(Y|C=c, X=x)\,, \quad (2)$$

and we denote this formula as approach A. Note that in (2) the summation starts with $c = 1$. Hence, it is only necessary to fit regression models to small subsets of the whole data set, see Table 1. However, one has to estimate the conditional probability $\mathrm{P}(C > 0|X = x)$ and the multi-class probabilities $\mathrm{P}(C = c|C > 0, X = x)$ for $c \in \{1, \ldots, k\}$, e.g. by a multinomial logistic regression model or by kernel logistic regression. If one splits the total data set into three subsets for training, validating, and testing, one only has to compute *predictions* for the conditional probabilities and the corresponding conditional expectations for *all* data points. If the estimators of the conditional probabilities and the conditional expectations are consistent, it follows by Slutzky's theorem that the combination of these estimators yields a consistent estimation of $\mathrm{E}(Y|X = x)$ in (2). Usually data sets in the context of motor vehicle insurance are huge, *e.g.* the data set in our application in Section 6 contains data from about 4.6 million customers. Bias reduction techniques applied to the validation data set may be helpful to reduce a possible bias of the estimates.

From our point of view the indirect estimation of $\mathrm{E}(Y|X = x)$ via approach A has practical and theoretical advantages over direct estimation of this quantity. Insurance companies are interested in estimating the terms in (2), because they contain additional information: the probability that a customer has at least one claim (which is our secondary response variable), the conditional probabilities $\mathrm{P}(C = c|C > 0, X = x)$, and the conditional expectations $\mathrm{E}(Y|C = c, X = x)$. The approach circumvents the problem, that most observed claim amount values y_i are 0, but $\mathrm{P}(Y = 0|X = x) = 0$ for many classical approaches based on a gamma or log-normal distribution. A reduction of computation time is possible, because we only have to fit regression models to a small subset (say 5%) of the data set. The estimation of conditional class probabilities for the whole data set is often much faster than fitting a regression model for the whole data set. It is possible, that different explanatory variables show a significant impact on the response variable Y or on the conditional class probabilities for different classes defined by C. This can also result in a reduction of interaction terms. In principle, it is possible to use different variable selection methods for the $k + 1$ classes. This can be especially important for the class of extreme claim amount values: because there may be only some hundreds or a few thousands of these rare events in the data set, it is in general impossible to use all explanatory variables for these data points. Finally, the strategies have the advantage that different techniques can be used for estimating the conditional class probabilities $\mathrm{P}(C = c|X = x)$ and for estimating the expectations $\mathrm{E}(Y|C = c, X = x)$ for different values of C. Examples for reasonable pairs are:

- Multinomial logistic regression + Gamma regression
- Robust logistic regression + semi-parametric regression
- Multinomial logistic regression + ε-Support Vector Regression (ε-SVR)
- Kernel logistic regression (KLR) + ε-Support Vector Regression
- Classification trees + regression trees
- A combination of the pairs given above, where some additional explanatory variables are constructed as a result of classification and regression trees or methods from extreme value theory are applied for the class of extreme claim amount values.

For robust logistic regression see Künsch *et al.* (1989) and Christmann (1994). Of course, $\nu-$SVR (see Schölkopf and Smola, 2002) can be an interesting alternative to ε-SVR.

Even for data sets with several million of customers it is in general not possible to fit simultaneously all high-dimensional interaction terms with classical statistical methods such as logistic regression or gamma regression, because the number of interaction terms increases exponentially fast for nominal explanatory variables.

The combination of kernel logistic regression and $\varepsilon-$support vector regression (*cf.* Section 5), both with an RBF kernel has the advantage that important interaction terms are fitted automatically without the need to specify them manually.

We like to mention that some statistical software packages such as R may run into trouble in fitting multinomial logistic regression models for large and high-dimensional data sets. Three possible reasons are that the dimension of the parameter vector can be quite high, that a data set with many discrete variables recoded into a large number of dummy variables will perhaps not fit into the memory of the computer, and that the maximum likelihood estimate may not exist, *cf.* Christmann and Rousseeuw (2001). To avoid a multinomial logistic regression model one can consider all pairs and then use pairwise coupling, *cf.* Hastie and Tibshirani (1998).

4.2. Alternatives. The law of total probability offers alternatives to (2). The motivation for the following alternative, say approach B, is that we first split the data into the groups 'no claim' vs. 'claim' and then split the data with 'claim' into 'extreme claim amount' and the remaining $k-1$ classes:

$$\begin{aligned} \mathrm{E}(Y|X=x) = \mathrm{P}(C>0|X=x)\ \times& \\ \{\mathrm{P}(C=k|C>0,X=x)\cdot\mathrm{E}(Y|C=k,X=x)+& \\ [1-\mathrm{P}(C=k|C>0,X=x)]\ \times& \\ \sum_{c=1}^{k-1}\mathrm{P}(C=c|0<C<k,X=x)\cdot\mathrm{E}(Y|C=c,X=x)\}\,.& \end{aligned} \tag{3}$$

This formula shares with (2) the property that it is only necessary to fit regression models to subsets of the whole data set. Of course, one can also interchange the steps in the above formula, which results in approach C:

$$\begin{aligned}\mathrm{E}(Y|X=x) = {} & \mathrm{P}(C=k|X=x)\cdot \mathrm{E}(Y|C=k,X=x) \\ & +[1-\mathrm{P}(C=k|X=x)]\cdot\{\mathrm{P}(C>0|C\neq k,X=x)\times \\ & \sum_{c=1}^{k-1}\mathrm{P}(C=c|0<C<k,X=x)\cdot \mathrm{E}(Y|C=c,X=x)\}\,.\end{aligned} \tag{4}$$

Note that two big binary classification problems have to be solved in (4), whereas there is only one such problem in (3).

5. Kernel logistic regression and $\varepsilon-$support vector regression

In this section we briefly describe two modern methods based on convex risk minimization in the sense of Vapnik (1998), see also Schölkopf und Smola (2002).

In statistical machine learning the major goal is the estimation of a functional relationship $y_i \approx f(x_i)+b$ between an outcome y_i belonging to some set $\mathcal{Y}\subseteq\mathbb{R}$ and a vector of explanatory variables $x_i=(x_{i,1},\ldots,x_{i,k})'\in\mathcal{X}\subseteq\mathbb{R}^p$. The function f and the intercept parameter b are unknown. The estimate of (f,b) is used to get predictions of an unobserved outcome y_{new} based on an observed value x_{new}. The classical assumption in machine learning is that the training data (x_i,y_i) are independent and identically generated from an underlying unknown distribution P for a pair of random variables (X_i,Y_i), $1\leq i\leq n$. In applications the training data set is often quite large, high-dimensional and complex. The quality of the predictor $f(x_i)+b$ is measured by some loss function $L(y_i,f(x_i)+b)$. The goal is to find a predictor $f_{\mathrm{P}}(x_i)+b_{\mathrm{P}}$ which minimizes the expected loss, *i.e.*

$$\mathrm{E}_{\mathrm{P}}\,L(Y,f_{\mathrm{P}}(X)+b_{\mathrm{P}})=\min_{f\in\mathcal{F},\,b\in\mathbb{R}}\mathrm{E}_{\mathrm{P}}\,L(Y,f(X)+b), \tag{5}$$

where $\mathrm{E}_{\mathrm{P}}\,L(Y,f(X)+b)=\int L(y,f(x)+b)d\mathrm{P}(x,y)$ denotes the expectation of L with respect to P and $\mathcal{F}$ denotes an appropriate set of functions $f:\mathcal{X}\to\mathbb{R}$. We have $y_i\in\mathcal{Y}:=\{-1,+1\}$ in the case of binary classification problems, $y_i\in\mathcal{Y}:=\{0,1,\ldots,k\}$ for multicategory classification problems, and $y_i\in\mathcal{Y}\subseteq\mathbb{R}$ in regression problems.

As P is unknown, it is in general not possible to solve the problem (5). Vapnik (1998) proposed to estimate the pair (f,b) as the solution of an empirical regularized risk. His approach relies on three important ideas: (1) restrict the class of functions $f\in\mathcal{F}$ to a broad subclass of functions belonging to a certain *Hilbert space*, (2) use a *convex* loss function L to avoid computational intractable problems which are NP-hard, and (3) use a *regularizing term* to avoid overfitting and to decrease the generalization error.

Let $L : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ be an appropriate convex loss function. Estimate (f, b) by solving the following empirical regularized risk minimization problem:

$$(\hat{f}_{n,\lambda}, \hat{b}_{n,\lambda}) = \arg \min_{f \in \mathcal{H},\, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i) + b) + \lambda \|f\|_{\mathcal{H}}^2, \tag{6}$$

where $\lambda > 0$ is a small regularization parameter, $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) of a kernel k, and b is an unknown real-valued offset. The problem (6) can be interpreted as a stochastic approximation of the minimization of the theoretical regularized risk, i.e.:

$$(f_{\mathrm{P},\lambda}, b_{\mathrm{P},\lambda}) = \arg \min_{f \in \mathcal{H},\, b \in \mathbb{R}} \mathrm{E}_{\mathrm{P}}\, L(Y, f(X) + b) + \lambda \|f\|_{\mathcal{H}}^2 \,. \tag{7}$$

In practice, it is often numerically better to solve the dual problem of (6). In this problem the RKHS does not occur explicitly, instead the corresponding kernel is involved. The choice of the kernel k enables the above methods to efficiently estimate not only linear, but also non-linear functions. Of special importance is the Gaussian radial basis function (RBF) kernel

$$k(x, x') = \exp(-\gamma \|x - x'\|^2)\,, \quad \gamma > 0, \tag{8}$$

which is a universal kernel on every compact subset of $\mathbb{R}^d$, *cf.* Steinwart (2001). Note, that γ in (8) describes the distance between two vectors x and x'. A radial basis function kernel offers not only a flexible fit but also a way to spread the risk of a single customer to sub-populations of customers with a similar vector of explanatory variables. The tuning constant γ controls this similarity.

For the case of binary classification, popular loss functions depend on y and (f, b) via $v = y(f(x) + b)$. Special cases are:

- Support Vector Machine (L1-SVM):
  $L(y, f(x) + b) = \max\{1 - y(f(x) + b), 0\}$
- Least Squares (L2-SVM):
  $L(y, f(x) + b) = [1 - y(f(x) + b)]^2$
- Kernel Logistic Regression (KLR):
  $L(y, f(x) + b) = \log(1 + \exp[-y(f(x) + b)])$, *cf.* Wahba (1999)
- AdaBoost:
  $L(y, f(x) + b) = \exp[-y(f(x) + b)]$, *cf.* Freund and Schapire (1996) and Friedman *et al.* (2000).

Kernel logistic regression has the advantage with respect to L1-SVM, that it estimates $\log(\frac{\mathrm{P}(Y=+1|X=x)}{\mathrm{P}(Y=-1|X=x)})$, i.e. $\mathrm{P}(Y = +1|X = x) = (1 + e^{-[f(x)+b]})^{-1}$, such that scoring is possible. Note that L1-SVM 'only' estimates whether $\mathrm{P}(Y = +1|X = x)$ is above or below $\frac{1}{2}$. Bartlett and Tewari (2004) show that there is a conflict in pattern recognition problems between the sparseness of the solution of kernel based convex risk minimization methods and the goal to estimate the conditional probabilities, see also Zhang (2004) for related work. The sparseness of the L1-SVM solution is of course a consequence of the fact that the L1-SVM uses a loss which is exactly equal to zero if $y[f(x)+b)] > 1$.

For the case of regression, Vapnik (1998) proposed the $\varepsilon-$support vector regression ($\varepsilon-$SVR) which is based on the $\varepsilon-$insensitive loss function

$$L_\varepsilon(y, f(x)+b) = \max\left\{0, |y - [f(x)+b]| - \varepsilon\right\},$$

for some $\varepsilon > 0$. Note that only residuals $y - [f(x)+b]$ lying outside of an $\varepsilon-$tube are penalized. Strongly related to $\varepsilon-$support vector regression is $\nu-$support vector regression, *cf.* Schölkopf und Smola (2002).

Christmann and Steinwart (2004) showed that a large class of such convex risk minimization methods with a radial basis function kernel have good robustness properties, e.g. a bounded influence function and a bounded sensitivity curve.

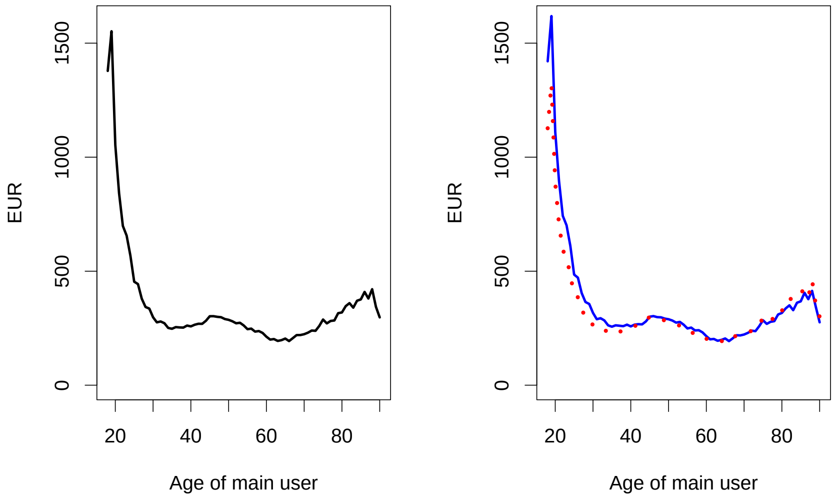
6. Application

This section describes some results for applying the approach A, i.e. (2), to the data set from the Verband öffentlicher Versicherer in Düsseldorf, Germany, see Section 3. We used a nonparametric approach based on the pair (KLR, $\varepsilon-$SVR) to fit the individual claim amount values per year. The following 8 explanatory variables were selected: gender of the main user, age of the main user, driving distance within a year, geographic region, a variable describing whether the car is kept in a garage, a variable for the population density, a variable related to the number of years the customer had no claim, and the strength of the engine.

We used a binary logistic regression model to fit the conditional probabilities that $\mathrm{P}(C > 0|X = x)$ (SAS, PROC LOGISTIC). The classical estimator in binary logistic regression is the maximum likelihood estimator. If the data set has complete separation or quasi-complete separation, the maximum likelihood estimates do not exist. Most statistical software packages do not check whether the maximum likelihood estimates exist. Christmann and Rousseeuw (2001) developed software to check whether the data set has complete or quasi-complete separation. Christmann *et al.* (2002) compared such methods with methods based on Vapnik's support vector machine with a linear kernel. Rousseeuw and Christmann (2003) proposed the hidden logistic regression model, which is strongly related to logistic regression, and investigated estimates which always exist.

The conditional probabilities $\mathrm{P}(C = c|C > 0, X = x)$, $1 \le c \le 4$, were estimated via kernel logistic regression. First, we estimated the probabilities for all pairs $\mathrm{P}(C = j|C \in \{i,j\}, X = x)$, where $1 \le i < j \le 4$. Keerthi *et al.* (2002) developed a fast dual algorithm for kernel logistic regression for the case of pattern recognition. Rüping (2003) implemented this algorithm in the program myKLR. We used myKLR to estimate the probabilities for these pairs. Then the multi-class probabilities $\mathrm{P}(C = c|C > 0, X = x)$, $c \in \{1,2,3,4\}$, were computed with pairwise coupling, *cf.* Hastie and

Figure 6. Results of applying approach A, part I. Left: estimated pure premium stratified by age of the main user. Right: estimated pure premium stratified by gender and age of the main user. Female: dotted. Male: solid.



Tibshirani (1998). The pure premium $\mathrm{E}(Y|C > 0, X = x)$, $c \in \{1, 2, 3, 4\}$, was estimated using $\varepsilon-$support vector regression (svm in the R-package e1071, Leisch et al., 2003). We used the exponential radial basis function kernel both for kernel logistic regression and for $\varepsilon-$support vector regression. Tuning constants were set to the values proposed by Cherkassky and Ma (2004). The data set was split up into subsets: training (50%, $n \approx 2.2E6$), validation (25%, $n \approx 1.1E6$), and test (25%, $n \approx 1.1E6$). Hence there were approximately 55000 positive values of y_i in the validation set and in the test set, respectively.

Here we only give some results for the explanatory variables gender and age of the main user but of course the calculations were done for all 8 explanatory variables simultaneously. For all data points in the test data set, the estimates of the pure premium $\mathrm{E}(Y|X = x)$ and of the conditional probabilities and conditional expectations defined in (2) were computed. Figure 6 and Figure 7 show the averages of these estimates stratified by the age of the main user or by gender and age of the main user.

Figure 6 shows the estimates of $\mathrm{E}(Y|X = x)$ stratified by age have the same sharp peak for young people, a moderate peak around 50 years, and an increase for elderly people which were already visible in Figure 4 from a univariate analysis. There is an interaction term for young people, say for the age group 18 to 24 years, see the right plot in Figure 6. The estimated pure premium for 18 to 20 years old males is approximately 300 EUR higher than for females of the same age. This is a rather big difference, because the base risk is approximately 360 EUR. Our method based on approach A using a combination of the two nonparametric methods kernel logistic regression and $\varepsilon-$support vector regression was able to detect this interaction term automatically although we did *not* model such an interaction term. It was confirmed by the Verband öffentlicher Versicherer, that this interaction term is not an artefact, but typical for their data sets.

However, from our point of view the main strength of approach A becomes visible if one investigates the conditional probabilities and conditional expectations which were fit by (2). Figure 7 shows the conditional probabilities stratified by gender and age of the main user. The probability of a claim, *i.e.* $\mathrm{P}(C > 0|X = x)$, shows again a similar shape than the corresponding curve in Figure 4. However, the conditional probability of a minor claim in the interval $(0, 2000]$ EUR given the event that at least one claim occurred *increases* for people of at least 18 years, *cf.* the subplot for $\mathrm{P}(C = 1|C > 0, X = x)$ in Figure 7. This is in contrast to the corresponding subplots for medium, high or extreme pure premium values, see the subplots for $\mathrm{P}(C = c|C > 0, X = x)$, $c \in \{2, 3, 4\}$ in Figure 7. Especially the last two subplots show, that young people have a two to three times higher probability in producing a high claim amount than more elderly people.

The effect of gender and age of the main user on $\mathrm{E}(Y|C = c, X = x)$, $c \in \{1, \ldots, 4\}$, was also investigated. The impact of these two explanatory variables on the conditional expectation of Y is lower than for the conditional probabilities, and the corresponding plots are omitted here. Nevertheless, it was visible that young people have a higher estimated pure premium than more elderly people, even if one conditions with respect to the class variable C.

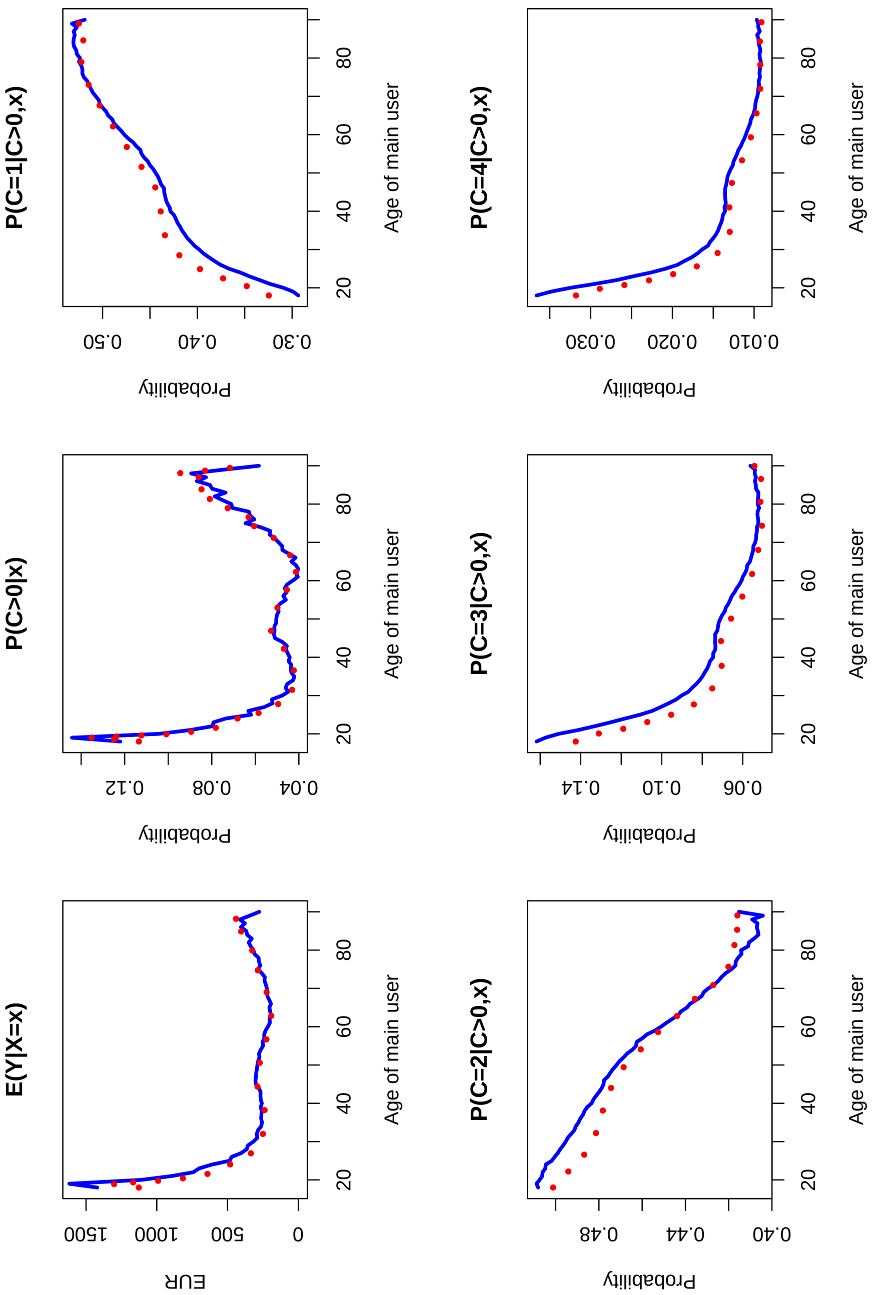
Concluding, people belonging to the age group 18 to 24 years have a higher estimated pure premium and a higher estimated probability to have a claim than more elderly customers. Approach A offers a lot of additional information. The application of the pair kernel logistic regression and $\varepsilon-$support vector regression was able to detect automatically an interaction term for young people with respect to gender and a moderate increased pure premium for main users with an age around 50 years.

7. Discussion

In this paper certain characteristics of complex and high-dimensional data sets from motor vehicle insurance companies were described. We propose to estimate the pure premium in motor vehicle insurance data in an indirect manner. This approach may also be useful in other areas, e.g. credit risk scoring, customer relationship management (CRM) or CHURN analyses.

There are several advantages of this approach in contrast to a straightforward estimation of the pure premium. The approach exploits knowledge of certain characteristics of data sets from motor vehicle insurance companies and estimates conditional probabilities and conditional expectations given the knowledge of an auxiliary class variable C describing the magnitude of the pure premium. This proposal offers additional insight into the structure of the data set, which is not visible with a direct estimation of the pure premium alone. Such additional information can be valuable for insurance

Figure 7. Results of applying approach A, part II. Pure premium and conditional probabilities stratified by gender and age of the main user. Female: dotted. Male: blue.



companies and can also be useful for aspects not related to the construction of insurance tariffs, *e.g.* in the context of direct marketing. Further, different estimation techniques and different variable selection methods can be used for the classes of pure premium defined by the auxiliary variable C.

The application of the proposed approach was illustrated for a large data set containing data from 15 motor vehicle insurance companies from Germany. An explorative data analysis shows that there are complex and nonmonotone dependency structures in this high-dimensional data. We used a nonparametric approach based on a combination of kernel logistic regression and $\varepsilon-$support vector regression. Both techniques belong to the class of statistical machine learning methods based on convex risk minimization, see Vapnik (1998), Schölkopf and Smola (2002), and Hastie *et al.* (2001). Such methods can fit quite complex data sets and have good prediction properties. The combination of these methods 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. Although the combination of kernel based methods from modern statistical machine learning yields some interesting insights into the insurance data set in our application, other pairs of methods to estimate the conditional probabilities and conditional expectations may be more successful for other data sets.

Christmann and Steinwart (2004) showed that a large class of such convex risk minimization methods with a radial basis function kernel have besides many other interesting properties also good robustness properties. Special cases are kernel logistic regression and the support vector machine (L1-SVM). Robustness is an important aspect in analyzing insurance data sets, because some explanatory variables may only be measured in an imprecise manner and some reported values of the claim size are only estimates and not the true values. On the other hand, in contrast to some other areas of applied statistics, extreme high individual claim amount values per year can not be dropped from the data set because this would systematically underestimate the pure premium.

In the literature it is often recommended to determine the tuning constants for kernel logistic regression and for $\varepsilon-$support vector regression, say c_{cost}, λ, γ, and ε, via a grid search or by cross-validation. Such methods can be extremely time-consuming for such a big data set we were dealing with in Section 6. Some preliminary experiments are indicating that these constants can be chosen in a reasonable way if they are determined as the solution of an appropriate optimization problem, e.g. by minimizing the mean squared error of the predictions for the validation data set or by using a chi-squared type statistic similar to the one used by the Hosmer-Lemeshow test for checking goodness-of-fit in logistic regression models, c.f. Hosmer and Lemeshow (1989) or Agresti (1996). The Nelder-Mead algorithm, c.f. Nelder and Mead (1965), worked well in this context for some test data sets and needed much less computation time than a grid search, but a systematic investigation of this topic is beyond the scope of this paper.

Although there were approximately 3300 customers with an individual claim amount above 50000 EUR in the data set we used for illustration purposes, an attractive alternative to $\varepsilon-$support vector regression for this subgroup is extreme value theory, *e.g.* by fitting generalized Pareto distributions of the main knots of a tree or more sophisticated methods.

Now, a brief comparison of the approach described in this paper and well-known methods for constructing insurance tariffs is given. The general approach described in this paper can be used to complement – not to replace – traditional methods which are used in practice to construct insurance tariffs, see *e.g.* Bailey and Simon (1960) or Mack (1997). The goal of our approach is to extract additional information which may be hard to detect or to model by classical methods. Generalized linear models belong to the most successful methods to construct insurance tariffs, see *e.g.* McCullagh and Nelder (1989), Kruse (1997), and Walter (1998). Generalized linear models are flexible enough to construct regression models for response variables with a distribution from an exponential family, e.g. Poisson, Gamma, Gaussian, inverse Gaussian, binomial, multinomial or negative binomial distribution. However, the assumptions of generalized linear models are not always satisfied. Generalized linear models are still parametric models and the classical maximum likelihood estimator in such models is in general not robust, see *e.g.* Christmann (1994, 1998), Cantoni and Ronchetti (2001), and Rousseeuw and Christmann (2003). Further, generalized linear models assume a certain functional relationship between expectation and variance even if one allows that over-dispersion is present. Tweedie's compound Poisson model, see Smyth and Jørgensen (2002), is even more flexible than the generalized linear model because it contains not only a regression model for the expectation of the response variable Y but also a regression model for the dispersion of Y. However, this more general model still yields problems in modelling complex high-dimensional relationships, especially if many explanatory variables are discrete which is quite common for insurance data sets. *E.g.*, if there are 8 discrete explanatory variables each with 8 different values there are approximately $8^8 \approx 16.7$ million interaction terms possible. Usually a data set from a single motor vehicle insurance company in Germany has much less observations. One can circumvent some of the drawbacks which generalized linear models or Tweedie's compound Poisson model have by using nonparametric methods from modern statistical machine learning theory such as support vector regression or kernel logistic regression. A radial basis function kernel offers not only a flexible nonparametric fit but also a way to spread the risk of a single customer to subpopulations of customers with a similar vector of explanatory variables by specifying the constant γ in an appropriate manner. But even if the analyst chooses classical methods such as logistic regression and gamma regression to estimate the conditional probabilities and the conditional expectations in (2), the approach described in this paper may yield additional insight into the structure of the data set, because one can investigate whether these conditional probabilities and conditional expectations differ with respect to subgroups defined by the class variable C.

Bibliography

Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* Wiley, New York.

Bailey, R. A., Simon, L.J. (1960). Two studies in automobile insurance ratemaking. *ASTIN Bulletin* **1** 192–217.

Bartlett, P. L., Tewari, A. (2004). Sparseness vs Estimating Conditional Probabilities: Some Asymptotic Results. Preprint, University of California, Berkeley.

Beirlant, J., de Wet, T., Goegebeur, Y. (2002). Nonparametric Estimation of Extreme Conditional Quantiles. Technical Report 2002-07, Universitair Centrum voor Statistiek, Katholieke Universiteit Leuven.

Cantoni, E., Ronchetti, E. (2001). Robust Inference for Generalized Linear Models. *Journal of the American Statistical Association* **96** 1022–1030.

Celebrián, A.C., Denuit, M., Lambert, P. (2003). Generalized Pareto Fit to the Society of Actuaries' Large Claims Database. *North American Actuarial Journal* **7** 18–36.

Cherkassky, V., Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks* **17** 113–126.

Christmann, A. (1994). Least median of weighted squares in logistic regression with large strata. *Biometrika* **81** 413–417.

Christmann, A. (1998). On positive breakdown point estimators in regression models with discrete response variables. Habilitation thesis, University of Dortmund, Department of Statistics.

Christmann, A., Fischer, P., Joachims, T. (2002). Comparison between various regression depth methods and the support vector machine to approximate the minimum number of misclassifications. *Computational Statistics* **17** 273–287.

Christmann, A., Rousseeuw, P.J. (2001). Measuring overlap in logistic regression. *Computational Statistics and Data Analysis* **37** 65–75.

Christmann, A., Steinwart, I. (2004). On robust properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research* **5** 1007-1034.

Embrechts, P., Klüppelberg, C., Mikosch, T. (1997). *Modelling Extreme Events for Insurance and Finance.* Springer, Berlin.

Freund, Y., Schapire, R. (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the 13th International Conference.* (L. Saitta, ed.), 148–156. Morgan Kaufmann, San Francisco.

Friedman, J., Hastie, T., Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Annals of Statistics* **28** 337–407.

Hastie, T., Tibshirani, R. (1998). Classification by pairwise coupling. *Annals of Statistics* **26** 451–471.

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* Springer, New York.

Hosmer, D.W., Lemeshow, W. (1989). *Applied Logistic Regression.* Wiley, New York.

Keerthi, S.S., Duan, K., Shevade, S.K., Poo, A.N. (2002). A fast dual algorithm for kernel logistic regression. Preprint, National University of Singapore.

Kruse, O. (1997). *Modelle zur Analyse und Prognose des Schadenbedarfs in der Kraftfahrt-Haftpflichtversicherung.* Verlag Versicherungswirtschaft e.V., Karlsruhe.

Künsch, H.R., Stefanski, L.A., Carroll, R.J. (1989). Conditionally Unbiased Bounded-Influence Estimation in General Regression Models, With Applications to Generalized Linear Models. *Journal of the American Statistical Association* **84** 460–466.

Leisch, F. et al. (2003). R package e1071. http://cran.r-project.org.

Mack, T. (1997). *Versicherungsmathematik.* Verlag Versicherungswirtschaft e.V., Karlsruhe.

McCullagh, P., Nelder, J.A. (1989). *Generalized linear models, 2nd ed.*. Chapman & Hall, London.

Nelder, J. A., Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal* **7** 308–313.

R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, R Foundation for Statistical Computing.

Rousseeuw, P.J., Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics & Data Analysis* **43** 315–332.

Rüping, S. (2003). myKLR - kernel logistic regression. Department of Computer Science, University of Dortmund,
`http://www-ai.cs.uni-dortmund.de/SOFTWARE`.

SAS Institute Inc. (2000). *SAS/STAT User's Guide, Version 8.* Cary, NC: SAS Institute Inc..

Schölkopf, B., Smola, A. (2002). *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge.

Smyth, G. K., Jørgensen, B. (2002). Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling. *ASTIN Bulletin* **32** 143–157.

Steinwart, I. (2001). On the Influence of the Kernel on the Consistency of Support Vector Machines. *Journal of Machine Learning Research* **2** 67–93.

Teugels, J.L. (2003). Reinsurance Actuarial Aspects. EURANDOM, Report 2003-006.

Vapnik, V. (1998). *Statistical Learning Theory.* Wiley, New York.

Wahba, G. (1999). Support Vector Machines, Reproducing Kernel Hilbert Spaces and the Randomized GACV. In *Advances in Kernel Methods - Support Vector Learning* (B. Schölkopf, C.J.C. Burges, A.J. Smola, eds.), 69–88. MIT Press, Cambridge, MA.

WALTER, J.T. (1998). *Zur Anwendung von Verallgemeinerten Linearen Modellen zu Zwecken der Tarifierung in der Kraftfahrzeug-Haftpflichtversicherung.* Verlag Versicherungswirtschaft e.V., Karlsruhe.

ZHANG, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics* **32** 56-134.

Andreas Christmann
University of Dortmund
Department of Statistics
44221 Dortmund
Germany
christmann@statistik.uni-dortmund.de