Accident Analysis and Prevention 101 (2017) 97–106

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Selected aspects of prior and likelihood information for a Bayesian classifier in a road safety analysis

Marzena Nowakowska

Faculty of Management and Computer Modelling, Kielce University of Technology, Al. Tysiąclecia Państwa Polskiego 7, 25-314 Kielce, Poland

ARTICLE INFO

Article history:
Received 28 December 2015
Received in revised form 21 October 2016
Accepted 16 January 2017

Keywords:
Bayesian logistic regression
Informative prior
Likelihood knowledge
Balanced sample
Model assessment
Crash severity

ABSTRACT

The development of the Bayesian logistic regression model classifying the road accident severity is discussed. The already exploited informative priors (method of moments, maximum likelihood estimation, and two-stage Bayesian updating), along with the original idea of a Boot prior proposal, are investigated when no expert opinion has been available. In addition, two possible approaches to updating the priors, in the form of unbalanced and balanced training data sets, are presented. The obtained logistic Bayesian models are assessed on the basis of a deviance information criterion (DIC), highest probability density (HPD) intervals, and coefficients of variation estimated for the model parameters. The verification of the model accuracy has been based on sensitivity, specificity and the harmonic mean of sensitivity and specificity, all calculated from a test data set. The models obtained from the balanced training data set have a better classification quality than the ones obtained from the unbalanced training data set. The two-stage Bayesian updating prior model and the Boot prior model, both identified with the use of the balanced training data set, outperform the non-informative, method of moments, and maximum likelihood estimation prior models. It is important to note that one should be careful when interpreting the parameters since different priors can lead to different models.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

In traffic safety analyses, statistical methods have always played a dominant role. In particular the frequentist (also called classical) modelling developed for years have resulted in the variety of models that describe and explain accident phenomena (see for example excellent elaborations in Huang and Abdel-Aty, (2010), Savolainen et al. (2011), and Hughes et al. (2015)). The research activity has been widened and enriched by a non-classical statistical approach adopting another philosophy originating from the Bayes theorem. In the Bayesian regression modelling, parameters are not constant values but random variables subject to certain posterior distributions derived from earlier (prior) knowledge on the parameters and from updating the knowledge by the information taken from empirical data (likelihood function). Multiplying the prior and the likelihood function leads to the posterior probability of the parameters (Box and Tiao, 1992; Congdon, 2006). The progress in the methodology is possible owing to numerical techniques: sampling methods applied to Monte Carlo Marcov chain generation procedures.

Two issues arise when the Bayesian regression model is developed. One is the formulation of prior distributions for the model parameters. The other one is the delivery of the likelihood information that updates the prior.

A prior distribution is commonly built on a researcher's belief. Therefore, the opinion is that any prior distribution is possible (Lee et al., 2009; Ma and Kockelman, 2006). However, the proper choice of the prior knowledge for the Bayesian regression model is important, which was signalized for example by Lee et al. (2009) and Pei et al. (2011, 2012). Some other works concerning the issue were also undertaken (Yu and Abdel-Aty, 2013; Heydari et al., 2014). Generally, three types of prior distributions are considered: non-informative, semi-informative and informative.

A non-informative prior expresses the lack of prior knowledge. Mostly, in this case a flat (diffuse, vague) distribution is assumed. A common choice for the majority of models is the normal distribution with zero value of a mean and a big standard deviation. In recent years there has been a variety of traffic safety analyses in which non-informative prior distributions for Bayesian model parameters were applied, either because of the lack of prior knowledge or because no additional value was obtained from informative priors in comparison with non-informative priors (e.g. Mitra and Washington, 2007; Lord and Miranda-Moreno, 2008; Huang et al., 2008; Schneider et al., 2009; El-Basyouny and Sayed, 2009; Haque

E-mail address: spimn@tu.kielce.pl

http://dx.doi.org/10.1016/j.aap.2017.01.009
0001-4575/© 2017 Elsevier Ltd. All rights reserved.

et al., 2010; El-Basyouny and Sayed, 2012; Aguero-Valverde, 2013; El-Basyouny et al., 2014; Islam and El-Basyouny, 2015).

To deal with the lack of prior knowledge as well as to eliminate some weaknesses of non-informative priors (Lord and Miranda-Moreno, 2008; Heydari et al., 2014), semi-informative prior distributions are suggested (El-Basyouny and Sayed, 2009). There is no universal rule except for taking a comparatively smaller standard deviation than in the non-informative prior when the normal distribution is assumed. For example, the value of 100 was used by Strauss et al. (2013) and Heydari et al. (2014).

A typical proposal for informative prior distributions is the use of expert knowledge, i.e. an opinion of a professional that knows a considered problem very well. Yet, such expert opinions are not always suitable or applicable. All the more so, the expert knowledge that is good for some cases is not necessarily good for others. The prior informative belief strongly depends on the choice of a research polygon, like for example a country, a country region, a road category (divided, undivided, high or small traffic volume), the type of an area (built-up, non built-up, highly congested like in a city), road segments (a road network, intersection or non intersection road segments, or even a chosen road segment) – to mention only the peak of the iceberg. A systematic analytical approach concerning the informative priors was presented by Yu and Abdel-Aty (2013). The authors considered several propositions of informative priors for safety performance functions. Some of their suggestions were applied to the Bayesian models in Ahmed et al. (2014) and commented in El-Basyouny et al. (2014). A large number of prior distributions for model parameters was thoroughly discussed by Heydari et al. (2014) in relation to a certain safety performance function when the models were developed in the conditions of limited data.

Choosing informative prior distributions for Bayesian classification models is seldom observed in traffic safety research works, particularly for logistic regression, although such model types are widely used. In the vast majority of cases, non-informative priors are assumed for the regression model parameters. If non-informative priors are used, then the influence of the prior distributions on the posterior distributions is usually very small and the results of the Bayesian regression model are numerically similar to that of the classical model estimation. However, the choice of prior distribution may play an important role in obtaining the posterior distributions, especially if the data are limited. Therefore, an appropriate choice of the priors can lead to some powerful properties for the logistic Bayesian models, including the improvement of the classification capability. Thus, the investigation of the prior distributions for the logistic Bayesian model classifying the road accident severity has been undertaken in the paper. The author's own idea of defining the informative prior, called the Boot approach, is presented and discussed together with the other prior propositions.

Updating prior information (using likelihood function) is the second issue in the Bayesian regression model. This is delivered by a training data set. A raw data set with the original proportion of the values of the road accident severity is frequently used to build the logistic model. Fatal accident observations are often extremely rare in such files. In consequence, the classifier exhibits a weak (or a very weak) classification of the accident fatality – an important category in a traffic safety research. However, because of the qualitative character of the response variable, it is possible to compose a training sample so as to reduce the negative influence of big differences in the proportion of the response value on the model quality. Such an approach has been adopted in the paper – stratified sampling and forcing the assumed proportions of the accident severity values are employed, which has strengthened the influence of rare categories.

The whole study is inspired by the above mentioned work by Yu and Abdel-Aty (2013) – the solutions proposed there have been adopted in developing the priors for the logistic regression model. Moreover, a personal prior-related idea as well as the use of unbalanced and balanced training data sets were investigated, and also the assessment of the models based on an independent data set was applied in the study.

2. Methodology

Logistic models are commonly used in classification problems in traffic safety analyses for their ease and flexibility. Here, such a model is employed to define the influence of at-fault driver's features together with a road specification on the accident severity.

Various domains are assumed for the road accident severity in many research works. It is determined by classification system country regulations (e.g. five categories in the USA and four categories in Poland and some other EU countries). In the vast majority of cases, fatal accidents occur most rarely and then there are serious accidents. Rare or very rare category occurrence in a training data set may lead to difficulties in explaining and predicting – statistical models can sharply underestimate the probability of rare events despite a good or very good total classification quality (Kubat and Matwin, 1997; Larose, 2006). This is especially important when a rare value is an event (the category on which a research is focused) and when the model results in a small value of sensitivity measure (the level of event classification quality). Such a concern appears in the road traffic safety research in which fatality or serious injury is the event. One method of solving the problem is the aggregation of the rare categories. The method has been applied in a variety of investigations. Some of them used other approaches than that of binary logistic regression in modelling tasks (e.g.: Jung et al., 2010; Pei et al., 2012; De O˜na et al., 2013; Yu and Abdel-Aty, 2014; Kwon et al., 2015). But there are also works in which such models were successfully adopted, notwithstanding ordinal effects of the two-valued accident severity response variable (e.g.: Vaez and Laflamme, 2005; Huang et al., 2008; Olszewski et al., 2015; Mujalli et al., 2016; Tay, 2016). Hence, in the study:

- the aggregation method was utilised by joining two accident categories: fatal and serious (which, in fact, is an incapacitating injury), in this way assuming the highest level of crash severity with respect to both the nature of the data set chosen for the analysis and the Polish accident severity classification,
- the binary logistic regression model was developed (taking into account also its simplicity).

Thus, the response variable $Y = AcSvr$ has two values: LA – light accident (considered as a failure) and FSA – fatal or serious accident (considered as a success). All input variables $\boldsymbol{X}$ are qualitative – the road specification is represented by the road number $RdNr$ and the driver's features are represented by: the incorrect behaviour Bhv, the age group $AgGrp$, the intoxication by alcohol or other substances Alh, and the gender $Gndr$. The conditional probability $\mathrm{P}(AcSvr = FSA \mid \boldsymbol{X})$ is the argument of a link function (logit) in the investigated models:

$$\operatorname{logit}(\mathrm{P}(AcSvr = FSA|\boldsymbol{X})) = \ln\left(\frac{\mathrm{P}(AcSvr = FSA|\boldsymbol{X})}{1 - \mathrm{P}(AcSvr = FSA|\boldsymbol{X})}\right) = \boldsymbol{\beta} \cdot \boldsymbol{X} \quad (1)$$

where $\boldsymbol{\beta}$ denotes the vector of the model parameters and $\boldsymbol{X}$ is the vector of input variables including unity taken for an intercept in the linear combination $\boldsymbol{\beta} \cdot \boldsymbol{X}$. Following the Bayesian idea, each parameter $\boldsymbol{\beta}_\mathrm{i}$ is a random variable that could have variety of values according to its probability density function expressed by the posterior distribution.

Download English Version:

https://daneshyari.com/en/article/4978728

Download Persian Version:

https://daneshyari.com/article/4978728

Daneshyari.com