Expert Systems With Applications 93 (2018) 104–117

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Feature selection for continuous aggregate response and its application to auto insurance data

Suyeon Kang, Jongwoo Song*

Department of Statistics, Ewha Womans University, 52 Ewhayeodaegil, Seodaemun-gu, Seoul, 03760, Republic of Korea

ARTICLE INFO

Article history:
Received 29 May 2017
Revised 15 September 2017
Accepted 1 October 2017
Available online 4 October 2017

Keywords:
Aggregate data
Feature selection
Auto insurance
Tariff classification
Risk assessment

ABSTRACT

This paper presents new feature selection algorithms for aggregate data analysis. Data aggregation is commonly used when it is not appropriate to model the relationship between a response and explanatory variables at an individual-level. We investigate substantial challenges in analysis for aggregate data. Then, we propose a groupwise feature selection method that addresses (i) the change in dataset depending on the selection of predictor variables, (ii) the presence of potential missing responses, and (iii) the suitability of model selection criteria when comparing models using different datasets. In application to real auto insurance data, we find a set of important predictors to classify the policyholders into some homogeneous risk groups. Our results clearly demonstrate the potential of the proposed feature selection method for aggregate data analysis in terms of flexibility and computational complexity. We expect that the proposed algorithms would be further applied into a wide range of decision-making tasks using aggregate data as they are applicable to any type of data.

1. Introduction

Data aggregation has been popular in social and physical sciences (Clark & Avery, 1976). In particular, researchers in epidemiology and non-life insurance industries commonly use aggregate data for their research. The reason for growing interest in grouped or aggregate data is partly that increasing emphasis on data privacy and confidentiality leads to restrictions to access complete micro data and partly that not all covariates are useful to predict a target variable at an individual-level (Li et al., 2007; Moineddin & Urquia, 2014). We often work with aggregate data when the number of individual-level observations is large and aggregate data approach do not lead to significant loss of information compared to non-aggregate data approach (Tse, 2009). For example, Roux (2004) studied group-level factors in epidemiology, based on the fact that some disease determinants cannot be conceptualized as individual-level attributes. Another example is non-life insurance, one of the most data-intensive industries. Traditionally, insurance companies divide their policyholders into different risk groups via data aggregation (risk classification) and set different premiums based on the group risk levels. This method is called a tariff analysis in non-life insurance and has a long history in analyzing non-life insurance data (Ohlsson & Johansson, 2010). Such an approach can significantly simplify data representation, data storage, and computation (de Jong & Heller, 2008).

Many studies have compared aggregate and non-aggregate data approaches, focusing on the impact of data aggregation on estimating regression coefficients and predicting a target variable based on this regression model (Caudill & Jackson, 1993; Clark & Avery, 1976; Lang & Gottschalk, 1996; Li et al., 2007; Moineddin & Urquia, 2014). For example, Lang and Gottschalk (1996) discussed the efficiency loss when fitting aggregate data to estimate coefficients compared to the case when fitting non-aggregate data, using ordinary least squares (OLS) regression. They defined the relative efficiency as the ratio of the variance of individual-level coefficient estimators to that of group-level coefficient estimators and concluded that this value hinges on the correlations between explanatory variables. When the explanatory variables are orthogonal, there is no efficiency loss from using aggregate data. Although they are not orthogonal, if the values within each group of the aggregate data are similar, the loss in efficiency is small. Li et al. (2007) extended this study to logistic regression and examined the effect of data aggregation on modeling a binary response. Their conclusion is the same with the OLS regression case; the efficiency loss depends on the correlation between explanatory variables and the ratio of the within-group variation to the between-group variation. When aggregating data, we can employ a variety of aggregation procedures and it may affect the regression coefficients. For example, geographers may be interested in grouping observations based on their spatial proximity (Blalock, 1964;

* Corresponding author.
E-mail addresses: korea92721@naver.com (S. Kang), josong@ewha.ac.kr (J. Song).

https://doi.org/10.1016/j.eswa.2017.10.007
0957-4174/

Clark & Avery, 1976). This type of aggregation method is a systematic aggregation, which may produce relatively efficient but biased estimates. On the other hand, random aggregation is likely to yield inefficient, albeit unbiased estimates (Cramer, 1964). Therefore, finding an ideal aggregation procedure is not simple. From previous studies, we can see that important factors for successful aggregate data analysis are finding significant explanatory variables whose correlations are not large and finding the optimal aggregation method that makes observations in the same group as homogeneous as possible. To this end, we attempt to develop a feature selection algorithm for successful aggregate data analysis, considering above-mentioned important factors. To illustrate this algorithm, we will analyze a real auto insurance data. This data would be a good example because it consists of large number of both observations and explanatory variables, as well as data aggregation is commonly used in this industry.

Generally, an insurance company keeps a huge amount of data that is associated with its policies. Such data may contain demographic characteristics of policyholders and the properties of insured objects, which are cars in auto insurance applications. We can extract useful and important information from the data for actuarial decision-making and risk assessment. A policy in non-life insurance is an agreement between an insurance company and a policyholder: If the policyholder pay a fee, the premium, the company compensate for unpredictable losses and damages on an insured object, which are encountered during the policy period. Therefore, it is important for the insurance company to predict the amount of loss that is transferred from the policyholder to the company. Based on the expected loss, the company can set different pure premiums. This process is essentially based on aggregate data, which is subdivided into different risk groups using explanatory variables in the data. For example, if two policyholders have the same age, sex, and previous accident history, they will be given the same insurance premium. As the insurance market has saturated, this ratemaking process becomes more important. Christmann (2005) argued that estimation of pure premium should have the following properties: Fairness, high precision, robustness, and simplicity. For example, the estimated premiums should be fair because high premiums make the company less competitive and low premiums make the company have low profit or even negative profit. If we find important risk factors (explanatory variables) and construct homogeneous risk groups, we can have a good ratemaking regression model that satisfies the above properties. Accordingly, the total amount of loss can be reduced. Like this insurance application, we sometimes need to predict a target based on aggregate data. In this paper, we study the properties of aggregate data and propose an efficient feature selection method, focusing on analyzing real auto insurance data. Our method is simple, fast and flexible, and can yield high prediction accuracy by identifying important predictors. We will give more details for the advantages of our method in the following sections.

In summary, the main contributions of this work can be summarized as follows:

- Although data aggregation is commonly used in many fields, there is few study focusing on how to select important predictors in this kind of data. Some studies analyzed their data in aggregate form, but they only used the fixed number of predictors (Frees, 2009) or just pointed out some difficulties that hamper to analyze such data. In this paper, we overcome these critical difficulties.
- As pre-processing for an effective data aggregation, we propose a merging and splitting procedure for all types of predictors. It is one of major issues for data aggregation, but there is no clear solution in the literature.
- Data aggregation and inverse data aggregation procedures are computationally intensive and the most time-consuming process in our feature selection method. We significantly reduce the total computation time by developing efficient algorithms to implement these two procedures. Accordingly, we can find an optimal set of predictors within short computation time.

The remainder of this paper is organized as follows. In Section 2, we canvass major difficulties in aggregate data analysis and review some papers related to this study. In Section 3, we suggest a proper strategy to solve these problems and develop several algorithms to apply this strategy to real data. In Section 4, the proposed algorithms and other two existing methods are applied to analyze real auto insurance data. We identify important rating factors on risk assessment in auto insurance and compare prediction accuracy of the three methods. Finally, we conclude this paper and give suggestions for further research in Section 5.

2. Related work

Data aggregation is used in a wide range of fields such as business, science, industry, and medicine. Examples in insurance include loss ratio prediction and fair premium setting. A target for decision-making in insurance can be claim frequency, claim severity, pure premium, and loss ratio. For modeling these targets, a variety of statistical methods have been developed.

Let us first look at non-life insurance applications. Traditionally, insurance companies divide their policyholders into several risk groups. This method uses combinations of rating factors, which is likely to affect policyholders' risk level. For example, Samson and Thomas (1987) used four rating factors with three levels for each of them, resulting in total $3^4 = 81$ groups, and estimated claim costs for each of these groups using a linear regression model. Other examples of the grouped data approach, compared to the ungrouped data approach, can be seen in de Jong and Heller (2008) and Tse (2009). Although these studies mainly focused on the ungrouped data approach, we can see simple examples of the grouped data approach for both continuous and binary responses. As pointed out in these studies, one disadvantage of the grouped approach is that the number of total groups grows rapidly whenever we consider additional rating factor.

More recently, a new grouping method, a clustering approach, has been employed and successfully used in risk classification (Bassi & Hernandez, 1997; Hanagandi et al., 1996; Smith et al., 2000; Williams & Huang, 1997; Yeo et al., 2001). For example, Williams and Huang (1997) used a k-means clustering to develop initial groups of policyholders and identified high-claiming policyholders in auto insurance. Smith et al. (2000) and Yeo et al. (2001) extended the use of clustering from identification of specific groups to prediction of claim costs for the groups. For example, Yeo et al. (2001) repeatedly applied a k-means clustering to Australian auto insurance dataset until the constructed clusters had moderate size. Finally, 30 different policyholders' groups are constructed using 13 predictors. To show the superiority of this method over traditional grouped data approach, Yeo et al. (2001) also considered the approach of Samson and Thomas (1987), and created $5^3 = 125$ different groups of policyholders using only 3 predictors among 13 predictors with 5 levels of each predictor. Yeo et al. (2001) reported that their method yielded higher prediction accuracy than the conventional one. They also argued that the proposed method overcome the disadvantage of the traditional method, a limitation on the number of groups, and finally can use all information of 13 predictors.

In the sense that such clustering approach is free from considering all combinations of the predictors, clustering method allows more predictors to be considered than traditional one. This clearly