MODELING OF DISCRETE QUESTIONNAIRE DATA WITH DIMENSION REDUCTION

Š. Jozová[*], *E. Uglickich*[†], *I. Nagy*[*†], *R. Likhonina*[†]

Abstract: The paper deals with the task of modeling discrete questionnaire data with a reduced dimension of the model. The discrete model dimension is reduced using the construction of local models based on independent binomial mixtures estimated with the help of recursive Bayesian algorithms in the combination with the naive Bayes technique. The main contribution of the paper is a three-phase algorithm of the discrete model dimension reduction, which allows to model high-dimensional questionnaire data with high number of explanatory variables and their possible realizations. The proposed general solution is applied to the traffic accident questionnaire analysis, where it takes the form of the classification of the accident circumstances and prediction of the traffic accident severity using the currently measured discrete data. Results of testing the obtained model on real data and comparison with theoretical counterparts are demonstrated.

Key words: *questionnaire data analysis, dimension reduction, binomial mixture, recursive Bayesian mixture estimation, accident severity*

Received: December 16, 2021 **DOI:** 10.14311/NNW.2022.32.002
Revised and accepted: February 28, 2022

1. Introduction

Modeling discrete data is an important task in the analysis of questionnaires to be filled in in case of traffic accidents. The questionnaires of a closed format (e.g., multiple choice, rating scale, Likert scale, etc.) produce sets of discrete valued data to be analyzed. In general, analysis of such data is impacted by the uncertainty of answers collected without a direct interaction with respondents, limited number of options to answer, which may not suit everyone [38], measurement errors [10], missing data [31], non-representative samples [65], imbalanced data [24], etc.

In questionnaires on traffic accidents, the discrete observations that need to be analyzed are primarily the severity of the accident and the circumstances in which the accident took place [1, 48, 49, 51, 66]. The circumstances put together a multivariate discrete explanatory variable, and describing the relationship between

[*]Šárka Jozová; Ivan Nagy; Faculty of Transportation Sciences, Czech Technical University, Na Florenci 25, 11000 Prague, Czech Republic, E-mail: jozovsar@fd.cvut.cz

[†]Evženie Uglickich – corresponding author; Ivan Nagy; Raissa Likhonina; Department of Signal Processing, Institute of Information Theory and Automation of the CAS, Pod vodárenskou věží 4, 18208 Prague, Czech Republic, E-mail: uglickich@utia.cas.cz, nagy@utia.cas.cz, likhonina@utia.cas.cz, https://orcid.org/0000-0003-1764-5924

them and severity of the accident is a key issue of the questionnaire analysis. In this case, when modeling the severity as the target variable, the uncertainty is enhanced due to the enormous dimension of the probability function conditioned by all individual circumstances, each of which has its own number of realizations. This results in low probabilities of the accident severity levels in the model. This means that the use of categorical models is not beneficial for this task due to operating with such dimensions and necessity of keeping the information in unreduced tables. In addition, the traffic accident questionnaires often contain mainly nominal data, which complicates the dimension reduction of the discrete model. Therefore, a way to reduce the dimension of the conditional probability function not leading to the loss of information from the questionnaires is searched for.

1.1 Literature review

Approaches to discrete questionnaire data analysis found in the literature range from trivial using a proportion estimation and hypothesis testing to advanced classification methods upon the specific task to be solved. In the context of modeling a binary target variable (e.g., crash or non-crash accident), the fundamental approaches are generalized linear models (GLM), namely, the logistic, probit as well as gompit regressions [5, 8, 30, 34, 70] originally directed at the classification of continuous explanatory variables, but applicable to indicator-based discretized explanatory data [6]. For a multinomial target variable (e.g., slight, serious or fatal accident), GLMs are distinguished upon the type of a variable among

(i) the multinomial logit regression, e.g., [5, 70] intended primarily for nominal target variables, but also utilized for Likert-type scaled or other rate scaled ordinal data [23],

(ii) the cumulative logit model for ordinal target variables [6, 30] or multinomial logit regression ignoring the knowledge of ordering the target values [5, 23, 34] and

(iii) the Poisson and negative binomial regression models [6, 23, 30] for count target variables as well as their zero-inflated versions [47].

For ordinal questionnaire data, the use of linear regression techniques [5] along with multiple indicators [23, 42] as well as analysis of ordinal Likert scale variables using structural equation modeling [11] can be also met. The item response theory with the models using a logit link function for ordinal [13, 57] and nominal [17, 57] variables along with the Mokken scaling analysis [68] should be also mentioned.

The publications close to the presented paper are devoted to Bayesian categorical analysis, e.g., [5, 21, 40, 52] and a group of clustering and classification methods, which can be used for discrete data. They include data mining techniques, such as decision tree [64], Bayesian networks [56], neural networks [4], k-nearest neighbors [22], fuzzy rules [16], naive Bayes classifiers [14, 25], genetic algorithms and model-based methods. The latter include the use of discrete mixture models such as latent class and Rasch mixture models [5], Poisson and negative binomial mixtures [21], mixtures of Poisson regressions [5, 21], mixtures of logistic regressions for binary data [21], Poisson-gamma and beta-binomial models [5, 21] as well as

Dirichlet mixtures [18, 45]. The estimation of the mentioned mixtures is solved primarily using the iterative expectation-maximization (EM) algorithm [28]. Algorithms of the recursive estimation of categorical mixtures with conjugate prior Dirichlet distributions, which create the basis for the proposed approach are elaborated in [39, 40, 52].

The model dimension reduction is solved in the literature using various effective techniques [12]. The approach proposing a way of reducing model dimension relatively close to the presented paper is independent component analysis (ICA) [20], which extracts independent components from a linear mixture of non-Gaussian data using maximum likelihood estimation [59]. In this area, a number of efficient ICA-based algorithms can be found. Fast fixed point ICA algorithm [35] focused on separating complex values and linearly mixed source data. Its version based on the Chebyshev-Pade approximation was proposed by [29]. The paper [7] investigated the ICA algorithm in the combination with principal component analysis and reproducibility stability approach to handle with dimensions of non-Gaussian sources. The copula based ICA algorithm can be found in [61]. The study [71] considered the method of ranking and averaging ICA by reproducibility (RAICAR) aimed at Gaussian and non-Gaussian sources. The work [62] proposed the ICA-based feature extraction method with the help of machine learning algorithms. Temporal ICA technique separating global noise and global signals can be found in [27]. Other modifications of ICA include a combination of ICA and kernel methods [36], a hybrid approach combining ICA and hierarchical clustering [55], probabilistic ICA [15], faster ICA under orthogonal constraint [2] and sparse Gaussian ICA method [3].

This paper proposes a systematic approach to modeling discrete questionnaire data, where the probability function dimension is reduced via:

(i) looking for numerous clusters in the data space of individual explanatory variables and describing each of them by a mixture of binomial distributions under condition of the mixtures' independence,

(ii) constructing the local target variable models on the detected clusters and

(iii) obtaining the classification discrete model of the target variable with the help of the naive Bayes technique.

For this purpose, the models of the explanatory variables in the form of independent binomial mixtures are estimated using recursive Bayesian algorithms based on [40, 41, 58]. Presented for normal models in [58], they were extensively elaborated for categorical distributions by [40], normal mixtures with the static pointer [41] and with dynamic pointer in [53]. The publication [52] generalized the recursive dynamic mixture estimation approach to various types of distributions with reproducible statistics. Here, this approach is derived for binomial mixtures. Unlike the well-known expectation-maximization algorithm [28], Variational Bayes approach [69] and Markov chain Monte Carlo techniques [26], the adopted methodology is free of iterative computations and related concerns about the convergence of algorithms. This allows the algorithms to run online based on the permanently measured explanatory data entering the classification model, which is constructed

using the results of the mixture-based analysis of the data combined with the naive Bayes principle [25].

The proposed approach is applied to the traffic accident questionnaire analysis, where the solution takes the form of the classification of the accident circumstances and prediction of the traffic accident severity using the currently measured discrete data.

The layout of the paper is organized as follows: Section 2 formulates the problem and gives necessary preliminaries. The general solution in the form of the three-phase approach of modeling questionnaire data is presented in Section 3. Its application to the traffic accident severity prediction is demonstrated in Section 4, which provides the model specification, practical mixture initialization, results and discussion. Conclusions and open problems can be found in Section 5.

2. Problem formulation and preliminaries

2.1 Problem formulation

Let us observe a multimodal system, which produces realizations of the discrete target variable y_t and discrete explanatory vector $[x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(n)}]'$ at discrete time instants t. The target variable is a scalar $y_t \in \{1, 2, \ldots, N_y\}$. The entries of the explanatory vector are $x_t^{(i)} \in \{0, 1, 2, \ldots, N_x^{(i)}\}$, where the superscript $(i) \in \{1, \ldots, n\}$ relates to the ith individual entry. The target and explanatory variables can be measured for a period of time up to $t = T$, while further for $t > T$, only the vector $[x_t^{(1)}, \ldots, x_t^{(n)}]'$ is observed and thus y_t should be predicted. In this paper, the analyzed data are defined as the realizations of the target and explanatory variables.

The aim is to construct and estimate the classification model of the target variable y_t for the time $t > T$ using the data available for $t \leq T$.

The basis for deriving the model is the use of the joint probability function (pf) $f(y_t, x_t^{(1)}, \ldots, x_t^{(n)})$ covering all relationships between the variables involved. This joint pf can be factorized according to the chain rule [58] as

$$f(y_t, x_t^{(1)}, \ldots, x_t^{(n)}) = f(x_t^{(1)}, \ldots, x_t^{(n)}) f(y_t | x_t^{(1)}, \ldots, x_t^{(n)}), \tag{1}$$

where the first marginal pf describes the space of the explanatory variables $x_t^{(i)}$ and the second conditional one is the predictive (classification) pf. Here, the pf $f(x_t^{(1)}, \ldots, x_t^{(n)})$ analyzes the explanatory variables with the goal of extracting the decisive information carried by them using local models on the explanatory data space. The classification model $f(y_t | x_t^{(1)}, \ldots, x_t^{(n)})$ represents the prediction of the target variable on these local models.

The basic problem in constructing the classification model is its enormous dimension, when it is based on the categorical distribution. For instance, a categorical model describing a target variable $y_t \in \{1, 2, 3\}$ depending on a single explanatory variable $x_t \in \{1, 2, 3, 4\}$ has the dimension 4×3 and has 12 parameters. In the case of two explanatory variables $x_t^{(1)} \in \{1, 2, 3, 4\}$ and $x_t^{(2)} \in \{1, 2, 3, 4, 5\}$, the dimension of the model becomes 20×3 and has 60 parameters. It is not hard to notice that

for a three-variate variable containing, e.g., $x_t^{(1)} \in \{1,2,3,4\}$, $x_t^{(2)} \in \{1,2,3,4,5\}$ and $x_t^{(3)} \in \{1,2,3,4,5\}$, the dimension of the categorical model grows up to 100×3 with 300 parameters. Thus, the categorical model is not only huge, but also over-parameterized and evidently not suitable for questionnaire data analysis tasks.

The solution proposed in this paper is based on two following features. The first one is to construct the pf $f(x_t^{(1)}, \ldots, x_t^{(n)})$ in the form of a product of independent models of individual explanatory variables $x_t^{(i)}$; the second one is to use a non-categorical distribution to describe components of the explanatory variables – here, specifically the binomial one.

Independence assumption A severe assumption of the independence of explanatory variables seems to be too strong[1]. That is why a model in the form of independent mixtures of individual explanatory variables is introduced. The reason is as follows: Mixture models will be beneficial in creating the local models on the explanatory data space, and due to this local modeling, the assumption of the independence necessary for using the naive Bayes principle will be more realistic from a practical point of view.

Binomial distribution The choice of the binomial distribution for the independent mixture components in the model $f(x_t^{(1)}, \ldots, x_t^{(n)})$ has the following reasons:

- It has finite number of realizations (similarly as the categorical distribution, but unlike the geometric one, Poisson, etc.).
- Its pf is able to cover a broad range of shapes: growing, decreasing or a shape of a hill with an arbitrary positioned top. Thus, the mixture of binomial components can be very close to the general categorical distribution.
- With a fixed number of components, the binomial distribution has the only one parameter – the probability of success. Thus, the probabilities of the values of this pf are bound together and formed just by one parameter. This property is similar to continuous distributions.

2.2 Preliminaries

The recursive parameter estimation of single binomial and categorical models will be used for the derivation of the classification model. That is why their estimation is briefly given below.

2.2.1 Binomial model

A single binomial model describing the individual explanatory variable $x_t^{(i)}$ (omitting here the superscript (i) for the simplicity) has the form of the pf

$$f\left(x_t = k|p\right) = \left(\begin{array}{c} N_x \\ x_t \end{array} \right) p^{x_t} \left(1-p\right)^{N_x - x_t}, \tag{2}$$

[1] A possible way to avoid this assumption could be the use orthogonalization in the space of $x_t^{(i)}$, which seems to promising. However, it is not discussed here.

where $k \in \{0, 1, 2, \ldots, N_x\}$ is a value of the variable x_t from the set of its possible realizations with the fixed known number N_x and p is an unknown parameter expressing the probability of success. The parameter p can be recursively estimated based on the Bayes rule [44, 58, 60] as follows:

$$f\left(p|x(t)\right) \propto f\left(x_t = k|p\right) f\left(p|x(t-1)\right), \tag{3}$$

where the denotation $x(t-1)$ means the collection of all observations of x_t up to the time $t-1$ including prior knowledge. Using the likelihood function computed on the available data measured up to the time $t-1$, the prior distribution $f\left(p|x(t-1)\right)$ to be substituted into (3) is derived as follows:

$$f\left(p|x(t-1)\right) \propto p^{S_{t-1}} \left(1-p\right)^{\kappa_{t-1}N_x - S_{t-1}}, \tag{4}$$

where the prior statistics S_{t-1} and κ_{t-1} are straightforward

$$S_{t-1} = \sum_{\tau=1}^{t-1} x_\tau, \quad \kappa_{t-1} = t-1. \tag{5}$$

Similarly to the general approach of the recursive Bayesian estimation of distributions of the exponential family [40], it can be seen that after the substitution of the prior distribution $f\left(p|x(t-1)\right)$ with the statistics (5) and model (2) into the Bayes rule (3), the statistics of the posterior pf are recursively updated as follows:

$$S_t = S_{t-1} + x_t, \quad \kappa_t = \kappa_{t-1} + 1, \tag{6}$$

where the statistics S_0 and κ_0 can also express prior knowledge. The updated statistics (6) are used for the recursive computation of the point estimate of the parameter p denoted by $\hat{p}_t$ based on measured data up to the time t, i.e.,

$$\hat{p}_t = \frac{S_t}{\kappa_t N_x}, \tag{7}$$

which is the result identical to the maximum likelihood estimation of the binomial distribution obtained via the derivation of the likelihood function, see, e.g., [19]. These straightforward derivations create the basis for the mixture estimation algorithm presented in Section 3.

2.2.2 Categorical model

A categorical model of the discrete random variable $c_t \in \{1, 2, \ldots, N_c\}$ is the pf $f\left(c_t = j|\alpha\right)$ with $j \in \{1, 2, \ldots, N_c\}$, which has the following form

$$\begin{array}{cccccc} c_t & 1 & 2 & \cdots & N_c \\ f\left(c_t = j|\alpha\right) & \alpha_1 & \alpha_2 & \ldots & \alpha_{N_c} \end{array} \tag{8}$$

where α_j are the probabilities of the value j of the variable c_t and $\alpha = \{\alpha_j\}_{j=1}^{N_c}$. According to [40], the recursive estimation of the parameter α is based on the Bayes rule and the conjugate prior Dirichlet distribution $f\left(\alpha|x(t-1)\right)$. Its statistics denoted by $\nu_t = \{\nu_{j;t}\}_{j=1}^{N_c}$ is updated as follows [40]:

$$\nu_{j;t} = \nu_{j;t-1} + 1, \tag{9}$$

starting from initial statistics chosen randomly. The normalized updated statistics gives the point estimates of α

$$\hat{\alpha}_{j;t} = \frac{\nu_{j;t}}{\sum_{l=1}^{N_c} \nu_{l;t}}. \tag{10}$$

3. Discrete model with reduced dimension

The solution to the formulated problem is proposed in three subsequent phases: (i) the explanatory variable analysis, (ii) the local model construction and (iii) the classification. The phases are subsequently presented below.

3.1 Explanatory variable analysis

In this section, the analysis of n explanatory variables is discussed. Each individual variable $x_t^{(i)}$ is modeled by a mixture of $N_c^{(i)}$ binomial components (2), i.e., it has its own number of components along with corresponding parameters denoted by the superscript (i). Due to the assumed independence of individual mixtures, each i-th explanatory variable $x_t^{(i)}$ will be modeled separately and the superscript (i) can be omitted for the sake of simplicity.

The switching of N_c binomial components (2) of each x_t is described by the discrete variable $c_t \in \{1, 2, \ldots, N_c\}$, which is called the pointer [41]. Values of each pointer indicate the active component generating the data of the individual explanatory variable x_t at time t. The pointer has the categorical distribution with the model (8).

With the used denotations, formally, for $j \in \{1, 2, \ldots, N_c\}$, the j-th binomial component (2) is conditioned by the pointer as well, i.e., it is

$$f\left(x_t = k | p, c_t = j\right) \tag{11}$$

and the parameter p is now the N_c-dimensional vector $[p_1, p_2, \ldots, p_{N_c}]'$, where each p_j is the probability of success corresponding to the j-th component of the variable x_t.

3.1.1 Recursive binomial mixture estimation

To describe the explanatory variables x_t using the introduced pfs (8) and (11), it is necessary to estimate the unknown parameters p and α along with the pointer c_t. The binomial mixture estimation algorithm is derived on the basis of the recursive Bayesian estimation theory for categorical models [40] with the static pointer [41]. To apply the similar recursive approach to the binomial pfs, the following schema of the mixture estimation is used:

$$f\left(p, \alpha, c_t = j | x(t)\right) \propto f\left(x_t = k, p, \alpha, c_t = j | x(t-1)\right)$$

$$= \underbrace{f\left(x_t = k | p, c_t = j\right)}_{(11)} \underbrace{f\left(p | x(t-1)\right)}_{(4)} \underbrace{f\left(c_t = j | \alpha\right)}_{(8)} \underbrace{f\left(\alpha | x(t-1)\right)}_{Dirichlet}. \tag{12}$$

Here, the Bayes rule is applied to the joint probability distribution of the unknown variables p, α and c_t, which is decomposed into the product of the models (8) and (11) along with the corresponding prior distributions [40]. The independence of the parameters p and α as well as of the pointer c_t and p is assumed.

In order to obtain the pointer estimate at time t, the relation (12) should be subsequently marginalized over all of the unknown parameters. As a result, the posterior pf of the pointer is got by means of the integrals over the entire definition spaces p^* and α^* of the variables involved [40, 41]

$$f\left(c_t = j|x(t)\right) = \int_{\alpha^*} \int_{p^*} f\left(x_t = k|p, c_t = j\right) f\left(p|x(t-1)\right) \\ f\left(c_t = j|\alpha\right) f\left(\alpha|x(t-1)\right) \, dp \, d\alpha, \tag{13}$$

where the first integral uses the point estimates of parameters and gives the approximation called the proximity (see also [37, 52, 54]) of the actual explanatory data item to each of its components at time t. The second integral provides the stationary prediction of the pointer [40]. The inside of the integrals represents the update of the statistics for estimating the relevant parameters. Finally, the result of the relation (13) is the pointer distribution in the form of the weighting vector $w_t = [w_{1;t}, \ldots, w_{N_c;t}]$, where each weight $w_{j;t}$ expresses the probability of the activity of each j-th component at time t, i.e., of the classification of the current data into the j-th component. Specifically, the weighting vector $w_t = [w_{1;t}, \ldots, w_{N_c;t}]$ is obtained with the help of the *entry-wise* multiplication of the proximities vector denoted by m and the vector of the last point estimates $\hat{\alpha}_{t-1}$ [40]

$$\tilde{w}_t = m \, \hat{\alpha}_{t-1} \tag{14}$$

and its subsequent normalization as follows:

$$w_{j;t} = \frac{\tilde{w}_{j;t}}{\sum_{k=1}^{N_c} \tilde{w}_{k;t}}, \quad j = \{1, 2, \ldots, N_c\}, \tag{15}$$

which forms the final weighting vector $w_t \; \forall j = \{1, 2, \ldots, N_c\}$. Each of the proximities in the vector m in (14) is the result of substituting the last point estimate $\hat{p}_{j;t-1}$, the current data entry x_t and the number of trials N_x into the component (11)

$$m_j = \mathcal{B}\left(x_t, N_x, \hat{p}_{j;t-1}\right), \tag{16}$$

which gives the proximity of the data item to the j-th binomial component.

Based on the adopted methodology [40, 41] and similarly to (6), the update of the statistics of the j-th component is

$$S_{j;t} = S_{j;t-1} + w_{j;t} x_t, \quad \kappa_{j;t} = \kappa_{j;t-1} + w_{j;t}, \tag{17}$$

starting with the prior statistics of the components (see Section 2.2.1). Based on measured data up to the time t, the actualized statistics (17) are used for computing the point estimates of the component parameters via (7).

The parameter α of the pointer model (8) is estimated according to (9) and (10) using weighted data in the update [40, 41]

$$\nu_{j;t} = \nu_{j;t-1} + w_{j;t}. \tag{18}$$

The point estimate of the pointer c_t is the index j of the maximum entry $w_{j;t}$ of the weighting vector w_t at time t

$$\hat{c}_t = \arg \max_{j \in \{1,\ldots,N_c\}} [w_{1;t}, \ldots, w_{N_c;t}], \tag{19}$$

which indicates the active j-th component generating the actual data item x_t.

The binomial mixture estimation listed in (12)-(19) is applied individually to all of the n explanatory variables, i.e., it works with scalars only. This phase is summarized in Algorithm 1.

Algorithm 1

{Initialization (for $t = 1$)}
for all $i \in \{1, 2, \ldots, n\}$ **do**
  Set the number of components $N_c^{(i)}$.
  for all $j \in \left\{1, 2, \ldots, N_c^{(i)}\right\}$ **do**
    Set the initial statistics $S_{j;t}^{(i)}$, $\kappa_{j;t}^{(i)}$ and $\nu_{j;t}^{(i)}$.
    Use them to compute the point estimates according to (7) and (10).
  end for
end for
{Explanatory variable analysis)}
for $t = 2, 3, \ldots, T$ **do**
  for all $i \in \{1, 2, \ldots, n\}$ **do**
    Load the data item $x_t^{(i)}$.
    for all $j \in \left\{1, 2, \ldots, N_c^{(i)}\right\}$ **do**
      Substitute the last point estimate $\hat{p}_{j;t-1}^{(i)}$, the current entry $x_t^{(i)}$ and the number of trials $N_x^{(i)}$ into the component (11) and compute the proximity of the data item to the j-th binomial component (16).
    end for
    Obtain the weighting vector $w_t^{(i)}$ [40] according to (14) and (15).
    for all $j \in \left\{1, 2, \ldots, N_c^{(i)}\right\}$ **do**
      Update the statistics according to (17) and (18).
      Re-compute the point estimates via (7) and (10).
    end for
    Obtain the point estimate of $c_t^{(i)}$ according to (19) and declare the active component for classifying data.
  end for
end for

The result of this phase is the clustered data of each explanatory variable.

3.2 Local model construction

This phase of the approach is the construction of local models on clusters detected on the explanatory data space.

First, the marginal pf of the target variable $f(y_t = q)$ with $q \in \{1, 2, \ldots, N_y\}$ is constructed with the help of the normalized histogram of the data y_t available for the time $t \leq T$. Secondly, the measurements of each explanatory variable x_t within each its component detected in Section 3.1 are collected together with those values of the target variable y_t that were observed simultaneously. The measurements are gathered in the form of contingence tables, existing for each component of each explanatory variable. The tables are normalized over columns in order to obtain the conditional pfs $f(x_t = k|y_t = q, c_t = j)$. In this way, the pfs are constructed locally only on the data x_t and y_t corresponding to each other in the components. The dimensions of these pfs are given only by the number N_y of possible realizations of the target variable y_t due to the collection of the scalars x_t. Algorithm 2 summarizes steps of this phase.

Algorithm 2

{Local model construction}
Normalize the histogram of all the measurements $q = \{1, 2, \ldots, N_y\}$ of y_t for the time $t \leq T$ to obtain the pf $f(y_t = q)$.
for all $i \in \{1, 2, \ldots, n\}$ **do**
  for all $j \in \left\{1, 2, \ldots, N_c^{(i)}\right\}$ **do**
    Construct the j-th contingence table with all the values of $x_t^{(i)}$ in rows and y_t in columns up to $t = T$.
    Normalize it over columns to obtain the pf $f\left(x_t^{(i)} = k^{(i)}|y_t = q, c_t^{(i)} = j\right)$.
  end for
end for

The constructed local models will be used for designing the classification model using the naive Bayes technique in the next phase.

3.3 Classification

This phase of the solution focuses on determining the value of the target variable y_t at each time instant $t > T$, where it is not observed and only explanatory variables are measured. Here, the two previous phases are combined using the actual weights of the learned components along with the local pfs constructed in Section 3.2.

To this end, the proximities of the current explanatory data to their components are calculated using the final (i.e., from time $t = T$) point estimates $\hat{p}_{j;t}$ according to (16). They are used to determine the actual weights of the components similarly to (14) and (15), but again with the final point estimates of the parameter α. Then, having the actual weights of each component of each explanatory variable, the weighted average of their local models from Section 3.2 is calculated for each i-th x_t

$$f\left(x_t^{(i)} = k^{(i)}|y_t = q\right) = \sum_{j=1}^{N_c^{(i)}} w_{j;t}^{(i)} f\left(x_t^{(i)} = k^{(i)}|y_t = q, c_t^{(i)} = j\right). \tag{20}$$

Finally, the predictive pf of the target variable, which classifies the explanatory data, is obtained using the results (20) for each $i \in \{1, 2, \ldots, n\}$, marginal pf $f(y_t = q)$ from Section 3.2 and the naive Bayes approach [25] as follows:

$$f\left(y_t = q | x_t^{(1)}, \ldots, x_t^{(n)}\right) = f(y_t = q) \prod_{i=1}^{n} f\left(x_t^{(i)} = k^{(i)} | y_t = q\right). \tag{21}$$

It is conditioned by the whole vector $[x_t^{(1)}, \ldots, x_t^{(n)}]'$ containing all the individual explanatory variables $x_t^{(i)}$. This result is the N_y-dimensional vector of probabilities of each of the possible realizations of y_t based on the explanatory data measured at time t. Using (21), the point prediction of the target variable is trivial

$$\hat{y}_t = \arg\max_q f\left(y_t = q | x_t^{(1)}, \ldots, x_t^{(n)}\right), \quad q \in \{1, 2, \ldots, N_y\}. \tag{22}$$

Algorithm 3 summarizes this phase.

Algorithm 3

{Classification}
for $t = T+1, T+2, \ldots$ **do**
  for all $i \in \{1, 2, \ldots, n\}$ **do**
    Measure the data item $x_t^{(i)}$.
    for all $j \in \left\{1, 2, \ldots, N_c^{(i)}\right\}$ **do**
      Compute the proximity via (16) using $\hat{p}_{j;T}^{(i)}$.
    end for
    Determine the weighting vector $w_t^{(i)}$ according to (14) and (15) using $\hat{\alpha}_T^{(i)}$.
    Compute the weighted average of the components (20).
    Compute the predictive pf (21).
    Determine the prediction of y_t according to (22).
  end for
end for

In this way, due to the proposed three-phase approach, the classification is solved using the local models on the explanatory data space, which are created by the independent mixtures. This allows to reduce computations during the construction of the classification model up to operating only with scalars instead of the categorical pf of enormous dimension. This is the main contribution of the presented paper.

4. Traffic accident severity prediction

This section demonstrates the application of the proposed three-phase approach to the analysis of traffic accident questionnaires with the aim of modeling and predicting the accident severity. The implementation was prepared in a free and open source software Scilab (www.scilab.org) providing a powerful programming environment for engineering and scientific applications.

4.1 Data

The data set from questionnaires on reported road accidents in Greater Manchester in North West England publicly available under Open Government Licence (© Crown copyright and database right 2021) was used for testing the presented approach. The data set is available at the following link (last accessed 17 August 2021): https://data.gov.uk/dataset/25170a92-0736-4090-baea-bf6add82d118/gm-road-casualty-accidents-full-stats19-data. The data set provides 40 238 records of the accident circumstances reported from 2010 to 2019 along with the severity of the accidents. 19 discrete accident circumstances were used as the explanatory variables belonging to the vector $[x_t^{(1)}, \ldots, x_t^{(19)}]'$. The sets of their original possible realizations including nominal values and raw numerical irregular denotations can be found in Appendix.

The target variable y_t was the accident severity with the realizations {"Fatal" $\equiv 1$, "Serious" $\equiv 2$, "Slight" $\equiv 3$}.

To make the data suitable for binomial components, the raw explanatory data records were recoded so that the set of possible realizations of each explanatory variable contained a zero value. The final sets of recoded possible realizations of those variables for which this procedure was necessary are shown in Tab. I. The rest of them remained raw.

Explanatory entry	Recoded possible realizations	Explanatory entry	Recoded possible realizations
$x_t^{(1)}$	$\{0, 1, 2, 3, 4\}$	$x_t^{(7)}$	$\{0, 1, 2, 5, 6, 8\}$
$x_t^{(2)}$	$\{0, 1, 2, \ldots, 12, 18\}$	$x_t^{(8)}$	$\{0, 10, 20, 30, 40, 50\}$
$x_t^{(3)}$	$\{0, 1, 2, \ldots, 9, 12, 13, 16, 18, 28\}$	$x_t^{(14)}$	$\{0, 1, \ldots, 6\}$
$x_t^{(4)}$	$\{0, 1, 2, \ldots, 6\}$	$x_t^{(15)}$	$\{0, 1, \ldots, 8\}$
$x_t^{(5)}$	$\{0, 1, 2, \ldots, 23\}$	$x_t^{(16)}$	$\{0, 1, \ldots, 4\}$
$x_t^{(6)}$	$\{0, 1, 2, \ldots, 6\}$	$x_t^{(19)}$	$\{0, 1\}$

Tab. I *Recoded sets of possible realizations of explanatory variables.*

4.2 Accident severity model

Using the binomial mixture given by (8) and (11) along with Algorithm 1, the individual accident circumstances were distributed among the components of their mixtures, where each mixture had its own number of components. The scheme of the application of Algorithm 1 to this part of the task is shown in Fig. 1, where the resulting mixtures of the circumstances are indicated.

Thereafter according to Algorithm 2, the estimated mixture components of the circumstances were used to construct the contingence tables expressing their relationship to the reported severity. Their normalization provided probabilities of the individual accident circumstances conditioned by the severity of the accidents.

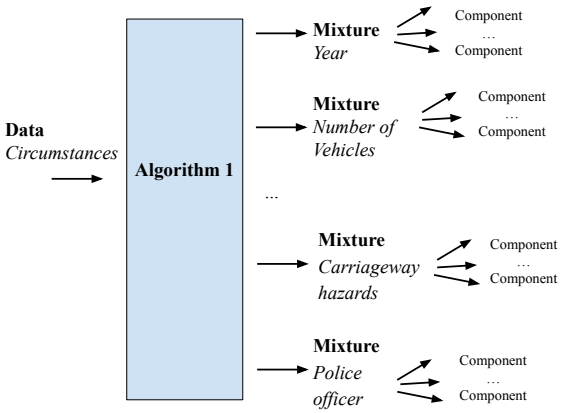

Fig. 1 *The scheme of the accident circumstances analysis.*

Similarly, the marginal probabilities of the severity were obtained from the reported data, see the scheme in Fig. 2.

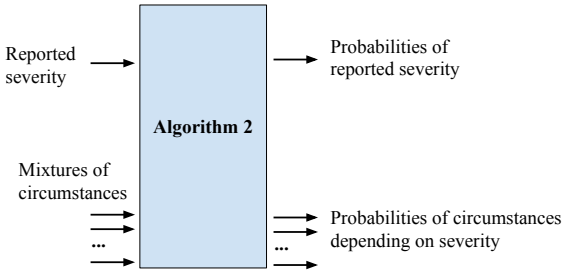

Fig. 2 *The scheme of the construction of the accident circumstances local models.*

Finally, using Algorithm 3, the predictive pf (21) gave the model of the accident severity, where the severity depends on all the accident circumstances actually measured in real time. It is

$$f\,(\text{severity}|\text{all real-time circumstances}) = f\,(\text{reported severity}) \times \prod_{i=1}^{19} f\,(\text{individual circumstance}|\text{reported severity})\,, \tag{23}$$

which involves 19 scalar models simultaneously due to Algorithm 3. The maximal probability of this pf provides the actual severity value, see the scheme in Fig. 3.

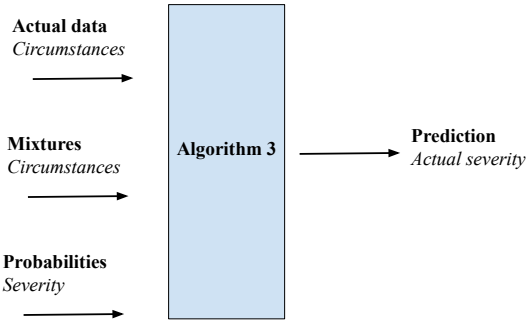

Fig. 3 *The accident severity prediction.*

4.3 Mixture initialization

The initialization of the binomial mixture components for starting Algorithm 1 was based on the analysis of histograms of reported accident circumstances. The histograms of each of the 19 circumstances were analyzed from the multimodality point of view in order to guess the number of components and their initial location. For example, Fig. 4 demonstrates the histograms of 18 000 randomized data of 3 selected circumstances taken from the data set for the aim of their mixture initialization: $x_t^{(7)}$ – road or carriageway type with two guessed components, $x_t^{(9)}$ – junction detail with four components and $x_t^{(13)}$ – pedestrian crossing physical facilities with three initialized components. The initial component statistics of all of the circumstances were expertly set as follows:

$$\begin{gathered}
S_{t-1}^{(1)} = [0.2\ \ 0.8]',\ S_{t-1}^{(2)} = [0.1\ \ 0.5\ \ 0.9]',\ S_{t-1}^{(3)} = [0.1\ \ 0.5\ \ 0.9]',\ S_{t-1}^{(4)} = [0.6\ \ 0.2\ \ 0.8]', \\
S_{t-1}^{(5)} = [0.1\ \ 0.4\ \ 0.6\ \ 0.9]',\ S_{t-1}^{(6)} = [0.05\ \ 0.6]',\ S_{t-1}^{(7)} = [0.4\ \ 0.6]',\ S_{t-1}^{(8)} = [0.2\ \ 0.5\ \ 0.99]', \\
S_{t-1}^{(9)} = [0.6\ \ 0.9\ \ 0.45\ \ 0.1]',\ S_{t-1}^{(10)} = [0.1\ \ 0.5\ \ 0.99]',\ S_{t-1}^{(11)} = [0.05\ \ 0.5\ \ 0.99]', \\
S_{t-1}^{(12)} = 0.8,\ S_{t-1}^{(13)} = [0.99\ \ 0.2\ \ 0.01]',\ S_{t-1}^{(14)} = [0.01\ \ 0.6]',\ S_{t-1}^{(15)} = [0.1\ \ 0.9]', \\
S_{t-1}^{(16)} = [0.01\ \ 0.1]',\ S_{t-1}^{(17)} = [0.01\ \ 0.5]',\ S_{t-1}^{(18)} = [0.01\ \ 0.5]',\ S_{t-1}^{(19)} = [0.2\ \ 0.8]'.
\end{gathered} \tag{24}$$

The initial statistics $\kappa_{j;t-1}^{(i)}$ were set equal to one for all of the components of all the circumstances. The pointer model statistics $\nu_{j;t-1}^{(i)}$ were initialized uniformly.

4.4 Results

For the validation of the proposed three-phase approach (TPA), 20 000 data records of the accident circumstances and severity were taken randomly from the entire data set 10 times. 18 000 of them were used each time for the two first phases of the approach in order to learn the models via Algorithms 1 and 2, while 2 000 measurements remained for testing the prediction using Algorithm 3. With these

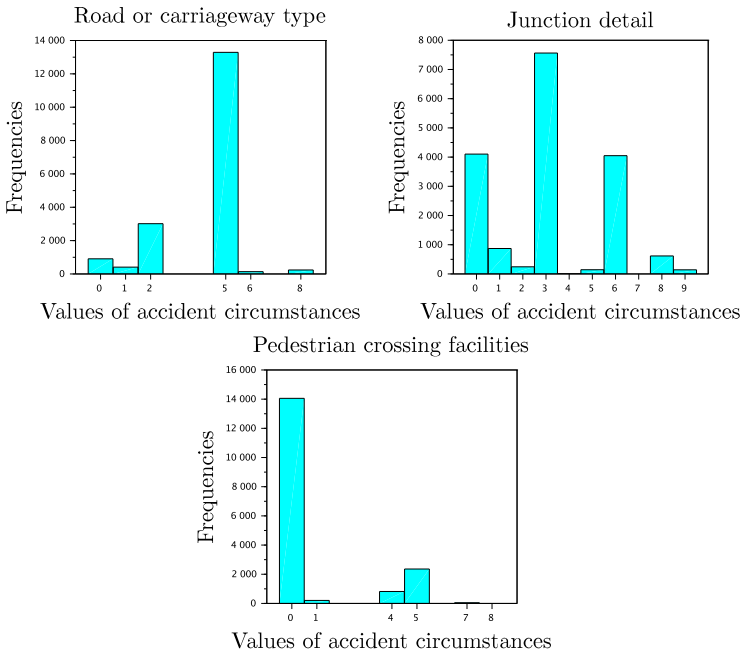

Fig. 4 *Histograms of the selected reported circumstances used for the mixture initialization.*

randomized data sets, a series of experiments were performed using fixed initialized mixtures. Their results were evaluated using the following criteria:

- The successful running of the first phase of TPA using Algorithm 1 is decisive for the overall validation of the approach. Here, the success strongly depends on the multimodal nature of the accident circumstances and their relationship to the reported severity along with the correct initialization. Therefore, the aim of the experiments was first to determine whether components of the circumstances that are the basis for predicting the severity of the accident were detected in the data space.

- Algorithm 2 of TPA is the trivial offline processing of the data to construct local models of circumstances depending on the accident severity, see Fig. 2. Hence, the resulting local models should express the multimodal nature of the circumstances, e.g., by changing the dominating probabilities in components.

- Finally, the prediction accuracy obtained as a result of running Algorithm 3 of TPA should be compared with theoretical counterparts. For this aim, the well-known reliable algorithms were chosen, implemented in a free and open-source data analytics, reporting and integration platform KNIME (www.knime.com), namely, decision tree (DT) [64, 67], k-nearest neighbors (k-NN) [9, 22], fuzzy rules (FR) [16] and naive Bayes (NB) [14]. The quantitative

evaluation of the prediction results of all the methods is given by means of the cross-validation with the use of 10 randomized data sets.

4.4.1 Detected components of accident circumstances

Fig. 5 shows the components of selected circumstances that were detected using Algorithm 1 of TPA in one of the randomized data sets. To save space, the most important circumstances with a dominant influence on the accident severity were expertly selected from all the explanatory variables used. In Fig. 5 each of these circumstances is plotted against the accident severity to show the two-dimensional components in the data space. For the visibility, the data values are jiggled. It can be seen that all the circumstances have their number of components. For example, the day of a week (left top plot), speed limit, junction control and pedestrian crossing physical facilities (middle plots) have three components. The accident time (middle top plot) has four components and the rest of the displayed circumstances have two. Fig. 5 emphasizes the independence of the mixtures of the circumstances by plotting their components in different colors: they are not related to each other. Their multimodal relationship with the accident severity, which is common for all of them, can be seen. TPA Algorithm 1 puts together data values that belong to the same mode, and the component is then a representative of these data values.

Moreover, the algorithm helps determine the combinations of variables, which were observed in the connection with different severity values in different components, and analyze the circumstances, which led to the maximum number of accidents and especially fatalities. Speaking about the day of a week (the left top plot in Fig. 5), according to the number of data in its components detected by Algorithm 1, the highest number of accidents (2 527) belongs to the combination of the slight accidents (see the severity value 3 in the y-axis) and Friday (value 5 in the x-axis) in the third component denoted by blue '□'. As for the fatal accidents (the severity value 1), the most frequent record was Saturday in the third component – 43 values. As regards the accident time, the most frequently reported combination (1 451 data records) was the slight accidents about 16 hours in the third component denoted by blue '·' in the middle top plot. The same time and component were obtained for the 20 fatal accidents.

The right top plot of Fig. 5 shows the components of the road or carriageway type, where most accidents correspond to the second component denoted by red '○'. Their maximum is 11 265 values, which is the combination of the slight severity and single carriageway (value 5 in the x-axis). The highest number of fatal accidents (149) was observed under the same conditions. For the speed limit (left middle plot of Fig. 5), the maximum number of data (12 438) corresponds to the 30 km/h speed limit and slight accidents in the first component marked by magenta '□'. The fatal accidents' maximum was 135 in the same component and speed limit.

In the middle plot of the middle row of Fig. 5, the components of the junction control are demonstrated. Here, the maximum number of accidents was 8 748 slight accidents reported under the condition of the give way or uncontrolled junction (value 4 in the x-axis) in the third component marked by magenta '∗'. The highest number of fatal accidents was 88 under the same conditions. 12 055 slight accidents and 140 fatalities were obtained, when a physical pedestrian crossing facility was

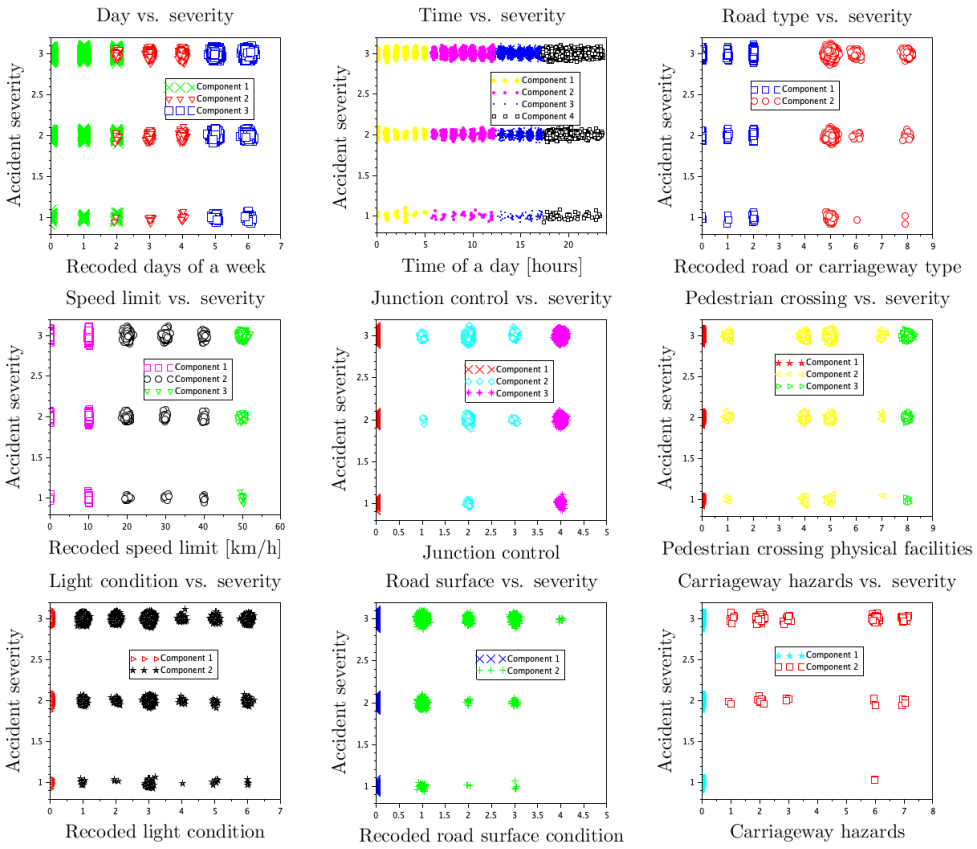

Fig. 5 *The two-dimensional components of the selected accident circumstances.*

not available within 50 m. This is shown in the first component (red '⋆') in the right middle plot.

The results of Algorithm 1 for the light condition (left bottom plot) report that the maximum number of accidents was 8 915 received for the combination of the slight severity and the daylight, which corresponds to the first component denoted by red '▷'. The fatal severity maximum was 91 in the same component during the daylight. For the road surface condition (middle bottom plot), the obtained results show the maximum of 10 302 slight accidents and 138 fatalities on a dry surface, which is the first component (blue '×'). The right bottom plot provides the components of the carriageway hazards. Here, the maximum number of data (15 280) was obtained for the slight severity without reported hazards, see the first component marked by cyan '⋆'. 194 fatal accidents belong to the same component under the same conditions of the hazards.

To summarize these results, it can be stated that most accidents over all the circumstances and their components in the data set were of the slight severity. The greatest influence on their number was obtained in the case of the carriageway haz-

ards, speed limit, pedestrian crossing physical facilities and road type in descending order, as it is shown in Fig. 6 (left). As regards the fatalities, their maximum occurrence was detected by Algorithm 1 for the case, when the observable hazards on the road were not reported. Over all the components, it was most affected by the carriageway hazards, road or carriageway type, pedestrian crossing facilities, road surface and speed limit, see Fig. 6 (right).

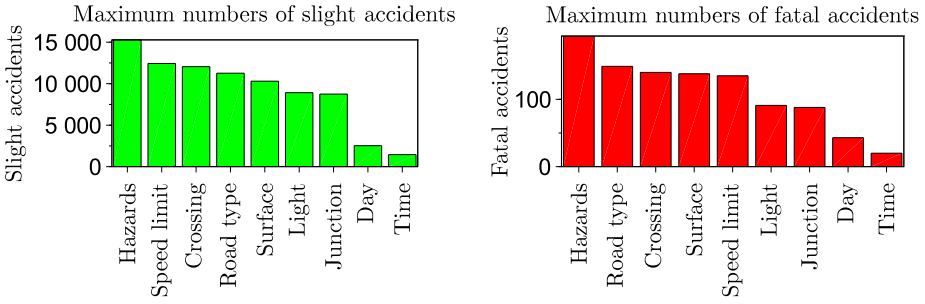

Fig. 6 *The maximum numbers of accidents with the corresponding circumstances in descending order.*

4.4.2 Local models of circumstances depending on severity

Here, to save space, the constructed local models of one of the circumstances obtained using the contingence tables according to TPA Algorithm 2 are shown. For this aim, the carriageway hazards located at the first position influencing the occurrence of the accidents in Fig. 6 were chosen. Since the selected circumstance had two components detected, the available measurements were gathered into two contingence tables with six values of the hazards in the rows and three values of the severity in the columns. They were normalized over columns to obtain the probabilities of the hazards depending on the severity. The two resulting local models can be found in Tab. II, which shows that they are different.

	Slight severity	Serious severity	Fatal severity	Slight severity	Serious severity	Fatal severity
No hazards	1	1	1	3.333e-09	4.762e-10	7.299e-11
Dislodged vehicle load	5.102e-11	4.272e-12	6.535e-13	0	0.0952381	0.0729927
Other object	5.102e-11	4.272e-12	6.535e-13	0	0.4761905	0.5328467
Previous accident	5.102e-11	4.272e-12	6.535e-13	0	0.0952381	0.0948905
Pedestrian	5.102e-11	4.272e-12	6.535e-13	1	0.1428571	0.1824818
Animal	5.102e-11	4.272e-12	6.535e-13	0	0.1904762	0.1167883

Tab. II *Local models 1 (left) and 2 (right) of the carriageway hazards conditioned by the accident severity.*

In the left table, the dominating probabilities correspond to the absence of any hazards in the carriageway and are the same for each value of the accident severity. The rest of the probabilities rely on the initialization. However, in the right table, the maximum probability refers to the presence of a pedestrian in the carriageway conditioned by the slight severity, while the second highest probability is related to other object in carriageway under the condition of the fatal severity. Since the models are different, it means that the hazards depending on the severity switch between two working modes. The local models were obtained for all the circumstances and their multimodal behavior depending on the severity was confirmed.

4.4.3 Accident severity prediction accuracy

Tab. III presents the results of the cross-validation of the compared methods on 10 data sets, where the prediction accuracy was computed as a percentage of correctly predicted severity values. It can be seen that TPA provided the average improvement in accuracy 7.01 %, 1.55 %, 2.51 % and 6.06 % compared to DT, k-NN, FR and NB respectively with the second lowest standard deviation.

	TPA, [%]	DT, [%]	k-NN, [%]	FR, [%]	NB, [%]
Data set 1	86.85	78.75	85.00	82.87	80.05
Data set 2	85.75	78.55	84.50	82.13	79.55
Data set 3	85.00	78.10	83.10	83.67	80.40
Data set 4	85.25	78.10	83.50	82.52	80.40
Data set 5	84.65	78.55	83.65	83.98	79.60
Data set 6	84.80	77.60	83.20	82.08	77.45
Data set 7	85.45	80.00	84.15	82.41	80.70
Data set 8	86.60	79.50	84.65	83.38	79.10
Data set 9	85.55	77.40	83.75	84.32	78.05
Data set 10	85.00	78.20	83.90	82.42	79.00
Average accuracy	85.49	78.48	83.94	82.978	79.43
Standard deviation	0.736	0.799	0.628	0.805	1.06

Tab. III *The percentage of the correct accident severity predictions.*

The accident severity predictions of the compared methods were also tested to see if they originate from the same distribution with the help of the Kruskal-Wallis test [43] in MATLAB [50]. The obtained p-value 2.3127e-47 was less than the significance level of 0.05, which indicates that the differences among the tested predictions are statistically significant. The post-hoc Bonferroni test [32, 33] was used to test the null hypothesis that there is no statistical difference in each pair of the compared predictions. It showed that only the k-NN and FR predictions had no statistically significant difference between them with the p-value higher than 0.05. The rest of the pairs had the p-values lower than the significance level, which means that their predictions are significantly different. The obtained confidence intervals of the post-hoc Bonferroni test can be found in Fig. 7, where TPA with

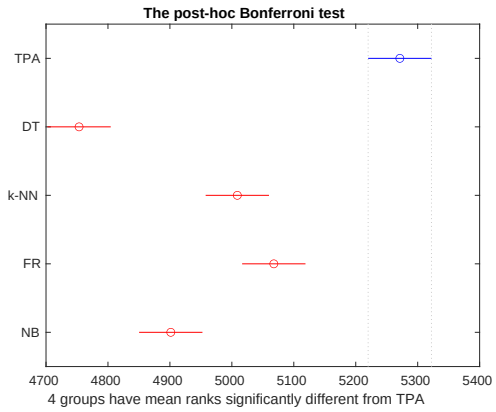

Fig. 7 *The post-hoc Bonferroni test confidence intervals.*

its improved accuracy has mean ranks significantly different from the rest of the methods.

As regards the average computational time necessary for a set of 20 000 data, TPA needed 100.08 seconds for all the three algorithms including the initialization. 0.01 % of the average time was necessary for the initialization, 90.65 % for Algorithm 1, 3.99 % for Algorithm 2 and 5.35 % for the prediction with Algorithm 3, which was 6 seconds on average. The average computational time of the rest of the methods was as follows: DT 2 seconds, k-NN 3 seconds, FR 127.2 seconds and NB 50 seconds. However, it should be emphasized that the proposed method has a completely different approach to computations including the real time phase of predicting, which is missing in the other discussed algorithms. Therefore, the computational time in this case, in contrast to the prediction accuracy, can hardly be a comparable indicator of the algorithm performance.

4.5 Discussion

The main aim of the experimental part of the paper was to test the presented general solution of modeling discrete questionnaire data with a reduced model dimension on real traffic accident data and predict the traffic accident severity. The aim has been successfully accomplished. The reduction of the model dimension was achieved due to (i) the description of the explanatory variables by independent mixture models and (ii) the use of the binomial distribution of mixture components. Without the proposed reduction, the accident severity model with three severity levels conditioned by 19 accident circumstances is a categorical distribution table with 7.779e16 rows and 3 columns. This enormous dimension is extremely complicated for the probabilistic approach because of the uncertainty bringing by the high number of variables with different large numbers of realizations. Due to TPA combining the recursive mixture estimation and naive Bayes procedure, the computations are performed on 19 scalar models. This significantly facilitates the calculation of probabilities of individual severity values for the aim of prediction.

Three levels of the accident severity were predicted, combining the information from the estimated binomial components and the actually measured circumstances. It was shown that the severity values vary among the components of the accident circumstances. This information covered by learning the models is critical for recognizing the components during the real time prediction of the severity. The performed experiments showed improvements in the prediction accuracy in the comparison with four existing methods. It is worth noting the improved accuracy compared with the original NB method creating a basis of Algorithm 3, which took the fourth place in Tab. III in contrast to TPA, which showed the best results.

In addition, TPA was beneficial in an area that was not originally the primary focus of the application. The identified components contributed to the analysis of the influence of circumstances on the maximum occurrence of the individual levels of accident severity. This TPA feature can be useful for determining combinations of real-time potential accident circumstances that lead to an increased risk of fatal and serious accidents, as it was discussed in [63]. Moreover, the mixture-based analysis from Algorithm 1 performed preliminarily on prior data can be applied for selecting explanatory variables for the predictive model, see, e.g., [46].

However, it should be stated that the reported improvements were obtained at the expense of extending the average computational time, where TPA took the fourth place. Nevertheless, the possibility of predicting the accident severity taking into account circumstances measured on the road in real time is an essential advantage of TPA, which can make this compromise acceptable.

A practical application of the traffic accident questionnaire analysis is expected in the area of road safety intelligent systems aimed at monitoring and evaluating real time risks of accidents. However, the general solution is not limited by the presented area of application and can be required in social fields, medicine, etc.

The limitation of the approach is the assumption of the multimodal relationship between the target and explanatory variables. The general suitability of the explanatory data to binomial mixtures is important as well. The minimum requirement to them is a higher number of possible realizations.

5. Conclusion

The paper presents a three-phase approach to modeling discrete questionnaire data with a reduced dimension of the model. The model dimension is reduced in three phases (i) using the recursive Bayesian estimation of independent mixtures of binomial distributions describing entries of the explanatory multivariate variable, (ii) constructing local models of the target variable on detected clusters of the explanatory data and (iii) obtaining the classification model of the target variable with the help of the naive Bayes technique. The proposed data-based probabilistic approach combines the offline analysis of previously available observations and online prediction, which enables feeding actually measured data in the algorithm. The model validation using real data and comparison with well-known algorithms has shown improvements in classifying discrete traffic accident data and predicting the accident severity, which confirms the efficiency of the proposed approach and its perspectives in terms of future research.

Due to the proposed reduction of dimension, the presented approach allows to model discrete questionnaire data with many explanatory variables and simultaneously a high number of their realizations. This is the main contribution of the paper. A significant advantage is also the mixture initialization, which is based on a visual analysis of prior data sets naturally available for Algorithms 1 and 2. It is fixed for all TPA runs with the same data.

The problems that remain open in the discussed area concern primarily: (i) the description of explanatory variables by a single mixture model, which can bring the possibility of combinations of realizations in components; (ii) the choice of other distributions (optionally mixed) of the mixture components; (iii) the automatization of the initialization task and (iv) the investigation of the continuous case of the presented approach using the multimodal relationship of the target and explanatory variables.

In general, the proposed approach can be used as a promising tool in the field of large-scale practical applications dealing with the analysis of data from questionnaires, which allows not to limit the dimension of the discrete model that could lead to the loss of information to be analyzed.

Appendix

The sets of the original possible realizations of the 19 discrete accident circumstances including their nominal values and raw numerical irregular denotations are as follows:

- $x_t^{(1)}$ – year $\in \{2010, 2011, \ldots, 2019\}$;
- $x_t^{(2)}$ – number of vehicles $\in \{1, 2, \ldots, 13, 16, 19\}$ after removing outliers taking less than 99th percentile;
- $x_t^{(3)}$ – number of casualties $\in \{1, 2, \ldots, 14, 17, 19, 29\}$ after removing the outliers;
- $x_t^{(4)}$ – day of a week $\in \{1, 2, \ldots, 7\}$ from Sunday to Saturday respectively;
- $x_t^{(5)}$ – output time in the format HH:MM;
- $x_t^{(6)}$ – 1st road class $\in \{1, 2, \ldots, 7\}$;
- $x_t^{(7)}$ – road or carriageway type $\in \{$"Roundabout" $\equiv 1$, "One way street" $\equiv 2$, "Dual carriageway" $\equiv 3$, "Single carriageway" $\equiv 6$, "Slip road" $\equiv 7$, "Unknown" $\equiv 9\}$;
- $x_t^{(8)}$ – speed limit $\in \{20, 30, 40, 50, 60, 70\}$;
- $x_t^{(9)}$ – junction detail $\in \{$ "Not at or within 20 meters of junction" $\equiv 0$, "Roundabout" $\equiv 1$, "Mini roundabout" $\equiv 2$, "T or staggered junction" $\equiv 3$, "Slip road" $\equiv 5$, "Crossroads" $\equiv 6$, "Junction more than four arms (not RAB)" $\equiv 7$, "Using private drive or entrance" $\equiv 8$, "Other junction" $\equiv 9\}$;

- $x_t^{(10)}$ – junction control $\in$ {"Not at or within 20 meters of junction" $\equiv$ 0,"Authorized person" $\equiv$ 1, "Automatic traffic signal" $\equiv$ 2, "Stop sign" $\equiv$ 3, "Give way or uncontrolled" $\equiv$ 4};

- $x_t^{(11)}$ – 2nd road class $\in \{0, 1, 2, \ldots, 7\}$;

- $x_t^{(12)}$ – pedestrian crossing human control $\in$ {"None within 50 meters" $\equiv$ 0, "Control by school crossing patrol" $\equiv$ 1, "Control by other authorized person" $\equiv$ 2};

- $x_t^{(13)}$ – pedestrian crossing physical facilities $\in$ {"No physical crossing facility within 50m" $\equiv$ 0, "Zebra crossing" $\equiv$ 1, "Pelican, puffin, toucan or similar non-junction pedestrian light crossing" $\equiv$ 4, "Pedestrian phase at traffic signal junction" $\equiv$ 5, "Footbridge or subway" $\equiv$ 7, "Central refuge – no other controls" $\equiv$ 8};

- $x_t^{(14)}$ – light condition $\in$ {"Daylight" $\equiv$ 1, "Darkness: street lights present and lit" $\equiv$ 4, "Darkness: street lights present but unlit" $\equiv$ 5, "Darkness: no street lighting" $\equiv$ 6, "Darkness: street lighting unknown" $\equiv$ 7};

- $x_t^{(15)}$ – weather condition $\in$ {"Fine without high winds" $\equiv$ 1, "Raining without high winds" $\equiv$ 2, "Snowing without high winds" $\equiv$ 3, "Fine with high winds" $\equiv$ 4, "Raining with high winds" $\equiv$ 5, "Snowing with high winds" $\equiv$ 6, "Fog or mist – if hazard" $\equiv$ 7, "Other" $\equiv$ 8, "Unknown" $\equiv$ 9};

- $x_t^{(16)}$ – road surface condition $\in$ {"Dry" $\equiv$ 1, "Wet / Damp" $\equiv$ 2, "Snow" $\equiv$ 3, "Frost / Ice" $\equiv$ 4, "Flood (surface water over 3cm deep)" $\equiv$ 5};

- $x_t^{(17)}$ – special conditions at site $\in$ {"None" $\equiv$ 0, "Auto traffic signal out" $\equiv$ 1, "Auto traffic signal partially defective" $\equiv$ 2, "Permanent road signing or marking defective or obscured" $\equiv$ 3, "Roadworks" $\equiv$ 4, "Road surface defective" $\equiv$ 5, "Oil or diesel" $\equiv$ 6, "Mud" $\equiv$ 7};

- $x_t^{(18)}$ – carriageway hazards $\in$ {"None" $\equiv$ 0, "Dislodged vehicle load in carriageway " $\equiv$ 1, "Other object in carriageway" $\equiv$ 2, "Involvement with previous accident" $\equiv$ 3, "Pedestrian in carriageway – not injured" $\equiv$ 6, "Any animal in carriageway (except ridden horse)" $\equiv$ 7};

- $x_t^{(19)}$ – did a police officer attend the scene and obtain the details for this report? $\in$ {"Yes" $\equiv$ 1, "No" $\equiv$ 2}.

Acknowledgement

This work was supported by the project Arrowhead Tools, the project number ECSEL 826452 and MSMT 8A19009.

References

[1] ABDEL-ATY M.A., HASSAN H.M., AHMED M., AL-GHAMDI A.S. Real-time prediction of visibility related crashes. Transportation Research Part C: Emerging Technologies, 2012, 24, pp. 288–298.

[2] ABLIN P., CARDOSO J.-F., GRAMFORT A. Faster ICA under orthogonal constraint, 2017, arXiv preprint arXiv:1711.10873.

[3] ABRAHAMSEN N., RIGOLLET P. Sparse gaussian ICA, 2018 , arXiv preprint arXiv:1804.00408.

[4] AGGARWAL C.C. Neural Networks and Deep Learning: A Textbook. Springer, 2018.

[5] AGRESTI A. Categorical Data Analysis. 3rd Ed. John Wiley & Sons, 2012.

[6] AGRESTI A. An Introduction to Categorical Data Analysis. 3rd Ed. Wiley, 2018.

[7] AKKALKOTKAR A., BROWN K.S. An algorithm for separation of mixed sparse and gaussian sources. PloS one, 2017, 12(4), pp. e0175775.

[8] ALLISON P.D. Logistic Regression Using SAS: Theory and Application. 2nd Ed. SAS Institute, 2012.

[9] ALTMAN N.S. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 1992, 46(3), pp. 175–185.

[10] ALWIN D.F. Margins of Error: a Study of Reliability in Survey Measurement. Wiley-Interscience, 2007.

[11] AWANG Z., AFTHANORHAN A., MAMAT M. The Likert scale analysis using parametric based Structural Equation Modeling (SEM). Computational Methods in Social Sciences, 2016. 4, pp. 13–21.

[12] AYESHAU S., HANIF M.K., TALIB R. Overview and comparative study of dimensionality reduction techniques for high dimensional data. Information Fusion, 2020, 59, pp. 44–58.

[13] BARTOLUCCI F., BACCI S., GNALDI M. Statistical Analysis of Questionnaires: A Unified Approach Based on R and Stata. Chapman & Hall/CRC, Boca Raton, 2016.

[14] BASHA S.M., RAJPUT D.S., POLURU R.K., BHUSHAN S.B., BASHa S.A.K. Evaluating the performance of supervised classification models: decision tree and naïve Bayes using KNIME. International Journal of Engineering & Technology, 2018, 7(4.5), pp. 248–253.

[15] BECKMANN C.F., SMITH S.M. Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Transactions on Medical Imaging, 2004, 23(2), pp. 137–152.

[16] BERTHOLD M.R., WISWEDEL B., GABRIEL T.R. Fuzzy logic in KNIME – modules for approximate reasoning. International Journal of Computational Intelligence Systems, 2013, 6(1), pp. 34–45.

[17] BOCK R.D. The nominal categories model. In: W. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory. New York: Springer, 1997, pp. 33–50.

[18] BOUGUILA N., ELGUEBALY W. Discrete data clustering using finite mixture models. Pattern Recognition, 2009, 42(1), pp. 33–42.

[19] COLLANI E., DRÄGER K. Binomial Distribution Handbook for Scientists and Engineers. Springer Science+Business Media, LLC. Birkhäuser Boston, 2001.

[20] COMON P. Independent component analysis, A new concept? Signal Processing, 1994, 36(3), pp. 287–314.

[21] CONGDON P. Bayesian Models for Categorical Data. John Wiley & Sons, 2005.

[22] COVER T.M., HART P.E. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967, 13 1), pp. 21–27.

[23] FALISSARD B. Analysis of Questionnaire Data with R. Chapman & Hall/CRC, Boca Raton, 2012.

[24] FERNÁNDEZ A., GARCÍA S., GALAR M., PRATI R.C., KRAWCZYK B., HERRERA F. Learning from Imbalanced Data Sets. Springer, 2018.

[25] FORSYTH D. Applied Machine Learning. Springer, 2019.

[26] FRÜHWIRTH-SCHNATTER S. Finite Mixture and Markov Switching Models, 2nd Edition, Springer New York, 2006.

[27] GLASSER M.F., COALSON T.S., BIJSTERBOSCH J.D., HARRISON S.J., HARMS M.P., ANTICEVIC A., ESSEN D.C.V., SMITH S.M. Using temporal ICA to selectively remove global noise while preserving global signal in functional MRI data. Neuroimage, 2018, 181, pp. 692–717.

[28] GUPTA M.R., CHEN Y. Theory and use of the EM algorithm. Foundations and Trends in Signal Processing, 2011, 4(3), pp. 223–296.

[29] HE X.-S., HE F., HE A.L. Super-gaussian BSS using fast-ICA with Chebyshev-Pade approximant. Circuits, Systems, and Signal Processing, 2018, 37(1), pp. 305–341.

[30] HEERINGA S.G., WEST B.T., BERGLUNG P.A. Applied Survey Data Analysis. Chapman & Hall/CRC, 2010.

[31] HEYMANS M.W., EEKHOUT I. Applied Missing Data Analysis with SPSS and (R) Studio, Heymans and Eekhout: Amsterdam, 2019, https://bookdown.org/mwheymans/bookmi/.

[32] HOCHBERG Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika, 1988, 75(4), pp. 800–802.

[33] HOCHBERG Y., TAMHANE A.C. Multiple Comparison Procedures, Wiley, New York, 1987.

[34] HOSMER D.W., LEMESHOW S. Applied Logistic Regression. 2nd Ed. Wiley-Interscience, 2000.

[35] HYVARINEN A. Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks, 1999, 10(3), pp. 626–634.

[36] INCE H., TRAFALIS T.B. A hybrid forecasting model for stock market prediction. Economic computation and economic cybernetics studies and research, 2017, 51(3), pp. 263–280.

[37] JOZOVÁ Š., UGLICKICH E., NAGY I. Bayesian mixture estimation without tears. In: Proceedings of the 18th International Conference on Informatics in Control, Automation and Robotics - ICINCO, ISBN 978-989-758-522-7, 2021, pp. 641–648.

[38] KAPLAN R.M., SACCUZZO D.P. Psychological Testing: Principles, Applications, and Issues. 9th Ed. Cengage Learning, 2017.

[39] KÁRNÝ M. Recursive estimation of high-order Markov chains: Approximation by finite mixtures. Information Sciences, 2016, 326, pp. 188–201.

[40] KÁRNÝ M., BÖHM J., GUY T. V., JIRSA L., NAGY I., NEDOMA P., TESAŘ L. Optimized Bayesian Dynamic Advising: Theory and Algorithms. Springer-Verlag, London, 2006.

[41] KÁRNÝ M., KADLEC J., SUTANTO E.L. Quasi-Bayes estimation applied to normal mixture. In: 3rd European IEEE Workshop on Computer-Intensive Methods in Control and Data Processing, Rojíček, J., Valečková, M., Kárný, M., Warwick, K. (Eds.), 1998, September, Prague, Czech Republic, 1998, pp. 77–82.

[42] KEITH T.Z. Multiple Regression and Beyond. An Introduction to Multiple Regression and Structural Equation Modeling. 3rd Ed. Routledge, New York and London, 2019.

[43] KRUSKAL W.H., WALLIS W.A. Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 1952, 47(260), pp. 583–621.

[44] LEE P.M. Bayesian Statistics. An Introduction, 4th ed., John Willey & Sons, 2012.

[45] LI Y., SCHOFIELD E., GÖNEN M. A tutorial on Dirichlet process mixture modeling. Journal of Mathematical Psychology, 2019, 91, pp. 128–144.

[46] LIN L., WANG Q., SADEK A.W. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction, Transportation Research Part C: Emerging Technologies, 2015, 55, pp. 444–459.

[47] LONG J.S., FREESE J. Regression Models for Categorical Dependent Variables Using Stata. 3rd Ed. Stata Press, 2014.

[48] MANNERINGE F.L., BHAT C.R. Analytic methods in accident research: Methodological frontier and future directions. Analytic Methods in Accident Research, 2014, 1, pp. 1–22.

[49] MANNERING F.L., SHANKAR V., BHAT C.R. Unobserved heterogeneity and the statistical analysis of highway accident data, Analytic Methods in Accident Research, 2016, 11, pp. 1–16.

[50] MATLAB 9.10.0. (R2021a), Natick, Massachusetts: The MathWorks Inc., 2021.

[51] MOLAN A.M., REZAPOUR M., KSAIBATI K. Investigating the relationship between crash severity, traffic barrier type, and vehicle type in crashes involving traffic barrier. Journal of Traffic and Transportation Engineering (English Edition), 2020, 7(1), pp. 125–136.

[52] NAGY I., SUZDALEVA E. Algorithms and Programs of Dynamic Mixture Estimation. Unified Approach to Different Types of Components. SpringerBriefs in Statistics. Springer International Publishing, 2017.

[53] NAGY I., SUZDALEVA E., KÁRNÝ M., MLYNÁŘOVÁ T. Bayesian estimation of dynamic finite mixtures. International Journal of Adaptive Control and Signal Processing. 2011, 25(9), pp. 765–787.

[54] NAGY I., SUZDALEVA E., PECHERKOVÁ P. Comparison of various definitions of proximity in mixture estimation. In: Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics - Volume 1: ICINCO, ISBN 978-989-758-198-4, 2016, pp. 527–534.

[55] NASCIMENTO M., SILVA F.F., SÁFADI T., NASCIMENTO A.C.C., FERREIRA T.E.M., BARROSO L.M.A., AZEVEDO C.F., GUIMARÃES S.E.F., SERÃO N.V.L. Independent component analysis (ICA) based-clustering of temporal RNA-seq data. PloS one, 2017, 12(7), pp. e0181195.

[56] NEAPOLITAN R.E. Learning Bayesian Networks. Pearson, 2019.

[57] NERING M.L., OSTINI R. Handbook of Polytomous Item Response Theory Models. Routledge, 2010.

[58] PETERKA V. Bayesian system identification, in Eykhoff, P. (Ed.), Trends and Progress in System Identification. Oxford, Pergamon Press, 1981, pp. 239–304.

[59] PHAM D.T., GARAT P. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. IEEE Transactions on Signal Processing, 1997, 45(7), pp. 1712–1725.

[60] PUZA B. Bayesian Methods for Statistical Analysis, ANU Press, 2017.

[61] RAHMANISHAMSI J., DOLATI A., AGHABOZORGI M.R. A copula based ICA algorithm and its application to time series clustering. Journal of Classification., 2018, 35(2), pp. 230–249.

[62] RADÜNTZ T., SCOUTEN J., HOCHMUTH O., MEFFERT B. Automated EEG artifact elimination by applying machine learning algorithms to ICA-based features. Journal of Neural Engineering, 2017, 14(4), pp. 46004.

[63] RETALLACK A., OSTENDORF B. Current understanding of the effects of congestion on traffic accidents. International Journal of Environmental Research and Public Health. 2019, 16(18), pp. 3400.

[64] SALZBERG S.L. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Mach Learn, https://doi.org/10.1007/BF00993309, 1994, 16, pp. 235–240.

[65] SARIS W.E. Design, Evaluation, and Analysis of Questionnaires for Survey Research, 2nd Ed. (Wiley Series in Survey Methodology), Wiley, 2014.

[66] SAVOLAINEN P.T., MANNERING F.L., LORD D., QUDDUS M.A. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. Accident Analysis & Prevention, 2011, 43(5), pp. 1666–1676.

[67] SHAFER J., AGRAWAL R., MEHTA M. SPRINT: A scalable parallel classifier for data mining. In: Proceedings of the 22nd VLDB Conference, Mumbai, India, 1996, pp. 544–555.

[68] SIJTSMA K., VAN DER ARK L.A. A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. British Journal of Mathematical and Statistical Psychology, 2017, 70, pp. 137–158.

[69] ŠMÍDL V., QUINN A. The Variational Bayes Method in Signal Processing. Springer, 2005.

[70] TANG W., HE H., TU X.M. Applied Categorical and Count Data Analysis. Chapman and Hall/CRC, 2012.

[71] YANG Z. , LA CONTE S., WENG X., HU X. Ranking and averaging independent component analysis by reproducibility (raicar). Human Brain Mapping, 2008, 29(6), pp. 711–725.