Pattern Structures for Risk Group Identification

Natalia V. Korepanova and Sergei O. Kuznetsov

National Research University Higher School of Economics, Moscow, Russia,
nkorepanova@hse.ru, skuznetsov@hse.ru

Abstract. Today personalized medicine is one of the most popular interdisciplinary research field, risk group identification being one of its most important tasks. Even though the first attempts to estimate the effect of patient's characteristics on the outcome were proposed in statistics in the middle of the twentieth century, it is still an open question how to explore such effects properly. In this paper we propose a trial version of the approach to risk group specification based on pattern structures and competing risk estimation, and discuss further steps of research on its performance and specificity.

Keywords: pattern structures, competing risks, risk group stratification

1 Introduction

Risk group identification is one of the task of personalized medicine. Often clinicians are required to define or to specify risk groups of patients to change treatment protocols properly. The task becomes more complicated if treatment outcome variable is not binary or numerical, but censored with several possible events. Such outcome is typical for most of the oncological diseases. Medical statistics calls the risk of occurrence of such event as competing risk and has a collection of techniques for their estimation and comparison [1]. However, there are almost no good techniques for risk groups specification when we deal with this type of outcome. To solve this problem, we presented an idea of approach to risk group identification when patients are described by nominal and/or numerical features with censored multi-event outcome. The mining process is based on pattern structures construction [2–5] with several adjustments which make it possible to apply the general approach to the dataset on acute lymphoblastic leukemia from ALL-MB-2008 trial [6] to obtain some promising results demonstrating the potential of the proposed approach.

The paper is organized as follows. In section 2 the theoretical basics of pattern structures and competing risks essential for the proposed approach are presented. Section 3 is devoted to the proposed approach. In section 4 we briefly describe results of application of the proposed approach to the data on acute lymphoblastic leukemia. Section 5 concludes the paper.

2 Preliminaries

2.1 Pattern Structures

In this section we give an introduction to pattern structures and discuss some of their applications.

Let G be a set (of objects), $(D, \sqcap)$ be a meet-semilattice (of all possible object descriptions), and $\delta : G \to D$ be a mapping. Then $(G, (D, \sqcap), \delta)$ is called a *pattern structure*, provided that the set $\delta(G) = \{\delta(g) \mid g \in G\}$ generates a complete subsemilattice $(D_\delta, \sqcap)$ of $(D, \sqcap)$, i.e. every subset X of $\delta(G)$ has an infimum $\bigsqcap X$ in $(D, \sqcap)$. Elements of D are called *patterns* and are naturally ordered by subsumption relation $\sqsubseteq$: $c \sqsubseteq d \Leftrightarrow c \sqcap d = c$, where $c, d \in D$. If $(G, (D, \sqcap), \delta)$ is a pattern structure we define the derivation operators which form a Galois connection between the powerset of G and $(D, \sqcap)$ as:

$$\begin{array}{ll} A^\diamond = \bigsqcap_{g \in A} \delta(g) & \text{for } A \subseteq G \\ d^\diamond = \{g \in G \mid d \sqsubseteq \delta(g)\} & \text{for } d \in D \end{array} \tag{1}$$

The pairs (A, d) satisfying $A \subseteq G$, $d \in D$, $A^\diamond = d$, and $A = d^\diamond$ are called *pattern concepts* of $(G, (D, \sqcap), \delta)$, with *pattern extent* A and *pattern intent* d. Pattern concepts are ordered with respect to set inclusion on extents. The ordered set of pattern concepts makes a lattice, called *pattern concept lattice*. Operator $(\cdot)^{\diamond\diamond}$ is an algebraical closure operator on patterns, since it is idempotent, extensive, and monotone.

If objects are described by binary attributes from set M, then $D = \wp(M)$, the powerset of M, and $\delta(g)$ is prime operator $(\cdot)'$ in the context (G, M, I): $\delta(g) = \{m \in M \mid gIm\}$, and $d_1 \sqcap d_2 = d_1 \cap d_2$ where $d_1, d_2 \in D$. So, subsumption corresponds to set inclusion: $d_1 \sqsubseteq d_2 \Leftrightarrow d_1 \sqcap d_2 = d_1 \Leftrightarrow d_1 \cap d_2 = d_1 \Leftrightarrow d_1 \subseteq d_2$.

If objects are described by k nominal features $\{\psi_1, ..., \psi_k\}$, we assume that ψ_i takes values from $\{1, ..., l_i\}$, where $l_i \in \mathbb{N}$, for $i \in \{1, ..., k\}$. Nominal features can be transformed into binary attributes in many different ways. In this paper we will assume just one variant of transformation. For each ψ_i we construct l_i binary attributes $\{\beta_i^1, ..., \beta_i^{l_i}\}$ such that $\beta_i^j : G \to \{0, 1\}$ and $\beta_i^j(g) : g \to \psi_i(g) \neq j$ where $g \in G, j = 1, ..., l_i$. As a result we get $\sum_{i=1,...,n} l_i$ binary attributes. Patterns on such binary attributes define subsets of values of nominal features. For instance, if a pattern contains binary attributes β_i^1 and β_i^4 and no other β_i^j then in terms of nominal feature it will look like $\psi_i \in \{2, 3, 5, ..., l_i\}$.

To operate with numerical features *interval pattern structures* [3–5] can be applied. If objects are described by n numerical features, we can represent them as a set of functions $\{\varphi_1, \varphi_2, ..., \varphi_n\}$, such that $\varphi_i : G \to \mathbb{R}$ for $i = 1, ..., n$. For each feature φ_i we construct an interval attribute $\alpha_i : G \to [\mathbb{R}, \mathbb{R}]$ such that if $\varphi_i(g) = x$ for $g \in G$, then $\alpha_i(g) = [x, x]$, where $x \in \mathbb{R}$. Then each object is described by a n-dimensional tuple of intervals. Let a and b be tuples of n intervals, so $a = \langle [v_i, w_i] \rangle_{i=1,...,n}$ and $b = \langle [x_i, y_i] \rangle_{i=1,...,n}$, where $v_i, w_i, x_i, y_i \in \mathbb{R}\ \forall i = 1, ..., n$. In this case operator $\sqcap$ is defined as follows:

$$a \sqcap b = \langle [v_i, w_i] \rangle_{i=1,...,n} \sqcap \langle [x_i, y_i] \rangle_{i=1,...,n} = \langle [v_i, w_i] \sqcap [x_i, y_i] \rangle_{i=1,...,n} \tag{2}$$

where $[v_i, w_i] \sqcap [x_i, y_i] = [\min(v_i, x_i), \max(w_i, y_i)]$. Hence, subsumption on interval tuples is defined as:

$$\begin{array}{l} a \sqsubseteq b \Leftrightarrow [v_i, w_i] \sqsubseteq [x_i, y_i]_{i=1,...,n} \Leftrightarrow [v_i, w_i] \sqcap [x_i, y_i] = [v_i, w_i]_{i=1,...,n} \Leftrightarrow \\ \Leftrightarrow [\min(v_i, x_i), \max(w_i, y_i)] = [v_i, w_i]_{i=1,...,n} \Leftrightarrow [v_i, w_i] \supseteq [x_i, y_i]_{i=1,...,n} \end{array} \quad (3)$$

For example, $\langle [2,6], [4,5] \rangle \sqsubseteq \langle [3,4], [5,5] \rangle$ as $[2,6] \sqsubseteq [3,4]$ and $[4,5] \sqsubseteq [5,5]$.

In a case when objects have both nominal and numerical features, we associate the set of binary attributes with each nominal feature and an interval attribute with each numerical one. Assume there are k binary and n interval attributes, then every $d \in D$ can be represented as $d = \langle \alpha, \beta \rangle$ where α is an interval tuple of length n, and β is a subset of binary attributes. If $d_1, d_2 \in D$, $d_1 = \langle \alpha_1, \beta_1 \rangle$, and $d_2 = \langle \alpha_2, \beta_2 \rangle$ operator $\sqcap$ can be set as $d_1 \sqcap d_2 = \langle \alpha_1 \sqcap \alpha_2, \beta_1 \sqcap \beta_2 \rangle$ where operator $\sqcap$ for interval tuples and the sets of binary attributes is defined above. The subsumption is also defined by subsumption on interval tuples and the sets of binary attributes:

$$\begin{array}{c} d_1 \sqsubseteq d_2 \Longleftrightarrow d_1 \sqcap d_2 = d_1 \Longleftrightarrow \langle \alpha_1, \beta_1 \rangle \sqcap \langle \alpha_2, \beta_2 \rangle = \langle \alpha_1, \beta_1 \rangle \Longleftrightarrow \\ \Longleftrightarrow \alpha_1 \sqcap \alpha_2 = \alpha_1, \beta_1 \sqcap \beta_2 = \beta_1 \Longleftrightarrow \alpha_1 \sqsubseteq \alpha_2, \beta_1 \sqsubseteq \beta_2 \end{array} \quad (4)$$

In our application we consider below objects are patient IDs, and object descriptions are tuples of binary and interval attribute values.

2.2 Competing Risks

In this paper the outcome variable is represented by a pair (T, C), where $C \in \{0, 1, ..., p\}$ and $T \in \mathbb{R}_+$. Here $C = i, i \in \{1, ..., p\}$ corresponds to a type of event (for instance, relapse or death in case of cancer-study data), and in this case T is the time from the beginning of observation to the event occurrence. When $C = 0$ corresponds to censorship (observation had ended before any event occurred), T is the time from the beginning of observation to the moment of censorship. Then the *survival function* is defined as $S(t) = \mathbb{P}(T > t)$, the probability of being free from any event at time t [1, Chapter 2]. However, if we aim at estimation of the risk of a specific event i (event of interest) in the presence of other competing events we are more interested in the *cumulative incidence function (CIF)*, defined as $F_i(t) = \mathbb{P}(T \leq t, C = i)$, the probability that event i occurs before time t [1, Chapter 4, p. 55]. Let $t_1 < t_2 < ... < t_r$ be observed unique uncensored time points, d_{ij} be the number of events of type i observed at time t_j, n_j be the number of patients for whom T is not less than t_j. Then, first, we non-parametrically estimate the survival function through Kaplan-Meier estimator [7, 8]:

$$\hat{S}(t) = \prod_{t_j \leq t} \left[1 - \frac{\sum_{i=1}^{p} d_{ij}}{n_j} \right] \quad (5)$$

Second, we estimate the CIF of the event of interest i as [1, Chapter 4, p. 56]:

$$\hat{F}_i(t) = \sum_{t_j \leq t} \frac{d_{ij}}{n_j} \hat{S}(t_{j-1}) \quad (6)$$

To test the confidence of the difference in risk of event of interest occurrence between M non-overlapping groups of patients, we use *Gray's test* for multiple groups [9] [1, Chapter 5, p. 74]. The main subject of it is to test the null hypothesis $H_0 : F_i^m(t) = F_i(t), m = 1, ..., M$, where $F_i^m(t)$ is the CIF of event i in group m, and F_i is the CIF of event i without groups specification. The general form of the score for group m is

$$z_m = \int_0^{t_r} W_i^m(t)\{\gamma_i^m(t) - \gamma_i(t)\}dt \tag{7}$$

where t_r is the maximum observed time, $\gamma_i^m(t) = \frac{F_i^m(t)'}{1-F_i^m(t)}$ is the hazard of event i in group m, $\gamma_i(t) = \frac{F_i(t)'}{1-F_i(t)}$ is the hazard of event i without groups specification, and $W_i^m(t)$ is a weight function. The test statistics is a quadratic form

$$Z\Sigma^{-1}Z^t \sim \chi^2_{M-1} \tag{8}$$

where $Z = (z_1, ..., z_{M-1})$ and Σ is the corresponding covariance matrix.

For $M = 2$ only z_1 needs to be computed. Frequently, $W_i^1(t)$ is assumed to be equal to $R_i^1(t)$, an adjusted number of individuals at risk. Let n_j^1 be the number of patients in group 1 for whom T is not less than t_j. We put $R_{ij}^1 = \hat{R}_i^1(t_j) = n_j^1 \frac{1-\hat{F}_i^1(t_{j-1})}{\hat{S}^1(t_{j-1})}$, where $\hat{S}^1(t_{j-1})$ is Kaplan-Meier estimation of survival function in group 1 and $\hat{F}_i^1(t_{j-1})$ is the estimation of the CIF of event i in group 1. Then the score is

$$z_1 = \sum_{t_j \in \{t_1,...,t_r\}} R_{ij}^1 \Big(\frac{d_{ij}^1}{R_{ij}^1} - \frac{d_{ij}^1 + d_{ij}^2}{R_{ij}^1 + R_{ij}^2} \Big), \quad \frac{z_1}{\widehat{Var}(z_1)} \sim \chi_1^2 \tag{9}$$

where d_{ij}^1 and d_{ij}^2 are the number of events i at time t_j in groups 1 and 2, respectively.

3 Risk Group Identification Procedure

In this section the procedure of identification of groups of patients with high or low risk of the event of interest is proposed. Here we assume that patients are described by numerical and/or nominal features, which can be transformed into ordinal and binary attributes, respectively. The most obvious way of group identification is the exhaustive search among all possible relevant descriptions. However, several group descriptions may correspond to the same subset of patients, therefore it is reasonable to search them only among pattern intents, which we also call *closed descriptions*, as all descriptions of any subset of patients are subsumed by the corresponding pattern intent. Hence, the search process can be considerably reduced. Further we will see that the reduction of the search is also important in terms of multiple testing. To construct all closed descriptions object-wise version of Close-by-One (CbO) algorithm [11] can be applied.

It is also possible to use InClose2 [10] after the appropriate scaling of interval features.

As the number of closed descriptions is theoretically exponential in the number of patients and the number of unique values of all features, it may be still necessary to curtail the search. The possible solution is to construct closed descriptions only from data on patients who experienced exactly the event of interest, but to perform all further steps on the whole dataset.

We may also consider a closed description as a rule which splits the whole set of patients into two parts: those who satisfy and not satisfy this description. It allows us to perform Gray's test for every closed description with two parts of the corresponding split as groups. However, it may be also reasonable to perform additional pre-selection steps before.

The idea of pre-selection is to define one or several measures of difference between CIFs of two parts of a split different from Gray's test statistics. If $F_{cd}(t)$ and n_{cd} are the CIF of the event of interest and the number of patients satisfying a closed description and $F_r(t)$ and n_r are the CIF of the event of interest and the number of the remaining patients, you can find two examples of difference measures below:

– Absolute difference of total CIFs × Entropy:

$$\left|\hat{F}_{cd}(\infty) - \hat{F}_r(\infty)\right|\Big[- \frac{n_{cd}}{n_{cd}+n_r} \log\big(\frac{n_{cd}}{n_{cd}+n_r}\big) - \frac{n_r}{n_{cd}+n_r} \log\big(\frac{n_r}{n_{cd}+c_r}\big)\Big] \tag{10}$$

– Absolute t-test statistics:

$$\frac{\left|\hat{F}_{cd}(\infty) - \hat{F}_r(\infty)\right|}{\sqrt{\frac{\widehat{Var}(\hat{F}_{cd}(\infty))}{n_{cd}} + \frac{\widehat{Var}(\hat{F}_r(\infty))}{n_r}}} \tag{11}$$

As we want to obtain groups with high or low risk of the event of interest we attempt to maximize difference measures chosen for pre-selection. The main problem here is choosing the cutoff of pre-selection. A possible solution would be computing the value of difference measure for every closed description, sorting computed values in ascending order and setting cutoff at the point from which the values of measure increase faster than before it. So, for further analysis we retain only closed descriptions with the value of difference measure not less than the chosen cutoff value.

After all pre-selection steps we compute Gray's test p-value for all remaining closed descriptions. Taking into account multiple testing correction we select only closed descriptions with confident difference in CIF between two parts of the corresponding splits. Let us call them *confident descriptions*. Then from all confident descriptions we retain only those which are not subsumed by other confident descriptions.

To present the obtained descriptions in a better way, one can remove uninformative attributes from the descriptions, such as interval attributes taking the whole attribute range, and transform binary attributes to the subsets of values of nominal features. Finally, when several similar splits (i.e. closed descriptions) are obtained we try to combine them through intersection and unification.

4 Experiments

The proposed procedure was applied to the data of MB-ALL-2008 study [6]. We tried to specify relapse risk groups separately for patients from standard risk group (SRG) and intermediate risk group (ImRG) (more than 1000 patients each). Patients are described by 5 nominal and 4 numerical features. The outcome is represented by the censored variable with 5 possible events, one of which is a relapse (the event of interest).

As for SRG, closed descriptions were constructed on the relapse patient data. Closed descriptions with the small inside or outside number of patients (less than 50) were excluded beforehand as the corresponding splits are too unbalanced, which may badly affect the value of Gray's test statistics. So, we started from 8383 closed descriptions for SRG. After that they were pre-selected with the use of two difference measures provided in section 4 as difference measure examples. The cutoff vales were set manually after looking at the graphs in Figure 1.

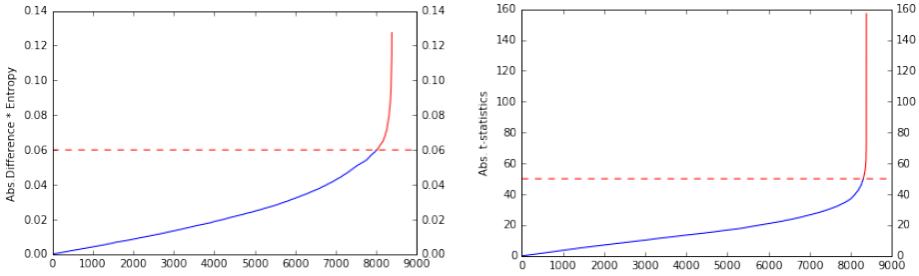

Fig. 1. Sorted values of absolute total CIF difference time entropy on the left picture and sorted value of t-statistics on the right picture for SRG. Dashed lines are the chosen cutoffs.

Applying both measures with chosen cutoffs only 18 closed descriptions were selected and corresponding splits were tested with Gray's test. Among 18 performed Grey's tests p-value of 14 tests was less than corrected level of confidence 2.8×10^{-3}. Here and further Bonferroni correction [12] was applied. Subsumption removing resulted in 3 quite similar descriptions, which were finally transformed into one split of all SRG patients into two groups with Gray's test p-value 7.8×10^{-11}:

1. High risk (201 patients): (age is not less than 11 years) OR (age $\in [4, 11)$ in years AND spleen enlargement is more than 2 cm)
2. Low risk (959 patients): others.

CIF estimations with 95% confidence intervals are shown in Figure 2. The p-value of Gray's test is very small and may satisfy even more strict multiplicity

corrections. For instance, if we apply Bonferroni correction for all closed descriptions the difference between CIFs will remain confident. This definitely looks like a promising result.

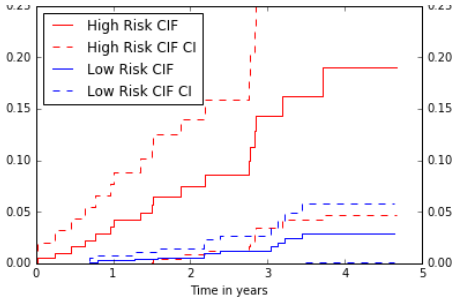

Fig. 2. The final split of SRG into high- and low-relapse-risk groups.

As far as ImRG concerns, 588178 closed descriptions, for which the inner and outer number of patients were not less than 100, were constructed. After preselection with two measures mentioned above, the number of candidate closed descriptions decreased to 4599. Performing Gray's test with Bonferroni corrected p-value 1.1×10^{-5} we cut down the number of descriptions to 61, and after subsumption removal we resulted in 12 confident closed descriptions. Most of the descriptions did not differ a lot, hence after their transformation and combination we resulted in the split of ImRG into 3 groups:

1. High risk (97 patients): (age is not less than 6 years old) AND (neuroleukemia OR white blood cell count is larger than 100/nl)
2. Low risk (575 patients): no neuroleukemia AND white blood cell count is not larger than 100/nl AND age is less than 6 years old
3. Medium risk (388 patients): others.

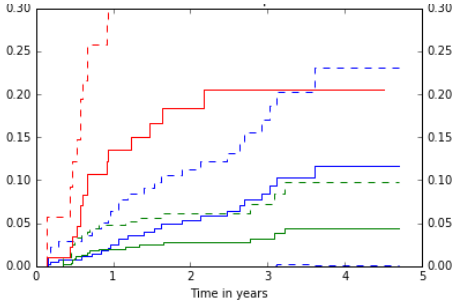

Fig. 3. The final split of ImRG into high-, medium- and low-relapse-risk groups.

You can find the illustration of CIF estimations with 95% confidence intervals in Figure 3. Gray's test p-value for 3 groups is 3×10^{-4}, pairwise Gray's test p-values: low-medium – 1.1×10^{-2}, medium-high – 1.3×10^{-3}, low-high – 1.8×10^{-8}. So, the significance of the difference in relapse risk between found groups of ImRG patients is questionable, except the difference between the low- and high-risk groups.

5 Conclusion

In this paper we proposed the first version of the procedure which realizes the approach to risk group specification based on pattern structures. This approach was applied to data with nominal and/or numerical features and censored outcome with several possible events and one event of interest. The reason for using pattern structures and closed descriptions comes from the idea that, first, definitions of risk groups should be interpretable and, second, not a greedy optimization, but global one should be realized. However, the general procedure still employs several heuristics that allow us to reduce the global search. The open questions that remain are the following: what is the best decision for multiple testing problem in terms of the proposed approach? What should be done with multiple solutions given by overlapping groups? Although some promising results were obtained, a comparison to other approaches has to be performed.

Acknowledgments

The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100'.

References

1. Pintilie, M. Competing risks: a practical perspective. John Wiley & Sons (2006).
2. Ganter, B., Kuznetsov, S.O. Pattern Structures and Their Projections. In: Proc. Stumme, G., Delugach, H. (eds.) 9th International Conference on Conceptual Structures (ICCS 2001). Lecture Notes in Artificial Intelligence, vol. 2120, pp. 129–142, Springer (2001)
3. Kuznetsov, S.O. Pattern Structures for Analyzing Complex Data. In: Sakai H. et al. (eds.) Proc. 12th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC 2009). Lecture Notes in Artificial Intelligence, vol. 5908, pp. 33–44, Springer (2009)
4. Kaytoue, M., Kuznetsov, S.O., Napoli, A., Duplessis, S. Mining Gene Expression Data with Pattern Structures in Formal Concept Analysis. Information Sciences, 10(181), 1989–2001, Elsevier, New York (2011)

5. Kuznetsov, S.O. Scalable Knowledge Discovery in Complex Data with Pattern Structures. In: Maji, P., Ghosh, A., et al. (eds.) Proc. 5th International Conference Pattern Recognition and Machine Intelligence (PReMI'2013). Lecture Notes in Computer Science, vol. 8251, pp. 30–41, Springer (2013)
6. Karachunskiy, A., Roumiantseva, J., Lagoiko, S. ALL-MB-2008. Dmitry Rogachev National Research Center of Pediatric Hematology, Oncology and Immunology (2011)
7. Kaplan, E.L., Meier, P. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, 282(53), 457–481 (1958)
8. May, W.L. Kaplan-Meier Survival Analysis. In: Encyclopedia of Cancer, pp. 1590-1593, Springer Berlin Heidelberg (2009)
9. Gray R.J. A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. The Annals of Statistics, 16(3), 1141–1154 (1988).
10. Andrews, S. "A 'Best-of-Breed' Approach for Designing a Fast Algorithm for Computing Fixpoints of Galois Connections". Information Sciences 295 (2015)
11. Kuznetsov, S. O., Obiedkov, S. A. Algorithm for the Construction of the Set of All Concepts and Their Line Diagram, Preprint MATH-Al-05, TU-Dresden (2000)
12. Miller, R. G. Jr. Simultaneous Statistical Inference. New York: Springer-Verlag (1991)