Survival Analysis and Forecasting Methods Based on Ensembles of Regularities

Dedovets M.S[1]., Senko O.V. [2]

1) "Arstel - Consulting" ,

2)Dorodnicyn Computer Center of Russian Academy of Sciences, Russia, 119991, Moscow, Vavilova, 40, senkoov@mail.ru

Method of regularities search in tasks of survival or reliability analysis is discussed that is based on optimal partitioning of prognostic variable space. Also method of survival curves evaluating for individual objects is described that is based on collective solutions by sets of regularities found with the help of optimal partitioning.

Introduction. The methods of recognition or forecasting based on collective solutions by ensembles of regularities now are successfully used in variety of applied task. Numerous experiments showed that they are effective in many tasks with high dimensionality of data. The goal of present paper is development of new method using ensembles of regularities for survival analysis in medicine and for similar reliability or duration analysis in engineering or economics. All these tasks will be referred to as survival tasks for object of arbitrary type. The first section describes method of regularities search that is based on optimal partitioning of prognostic variables space. This technique is modification of optimal valid partitioning method [1,3] for survival analysis. In the second section method for evaluating of survival curves for individual objects is described that is based on collective solutions by sets of regularities found with the help of optimal partitioning.

Optimal partitioning method for survival analysis. Optimal partitions of prognostic variables space are searched inside apriori fixed partitions families by initial training information $\tilde{S}_0 = \{(\alpha_1, t_1, \mathbf{x}_1), \ldots, (\alpha_m, t_m, \mathbf{x}_m)\}$. The aim of partitioning is achievement of maximal difference between survival in neighbor subregions (elements of partitions). The partition family is defined as the set of partitions with limited number of elements that are constructed by the same procedure. The uni-dimensional and two-dimensional families are considered. The uni-dimensional families includes partitions of admissible intervals of single variables. The simplest Family I includes all partitions with two elements that are divided by one boundary point. The more complex Family II includes all partitions with no more than three elements that are divided

by two boundary points. The two-dimensional Family III includes all partitions of two-dimensional admissible areas with no more than four elements that are separated by two boundary lines parallel to coordinate axes. The quality of partition or divergence between survival in groups that are induced by its elements are described with the help of quality functional $\mathrm{F}(R,\tilde{S}_0)$.

Let $g_{km}(t,\tilde{S})$ is Kaplan-Mayer estimate [4] of survival curve in group of patients $\tilde{S}$. In other words $g_{km}(t,\tilde{S})$ is estimate of probability $\Pr(t'<t)$ where t' is death time of object from the same general set as objects from $\tilde{S}$. Suppose some partition R of prognostic variables space induce partition $\{\tilde{S}_1,\ldots,\tilde{S}_k\}$ of training set $\tilde{S}_0$. Let $\rho[g_{km}(t,\tilde{S}'),g_{km}(t,\tilde{S}'')]=\int_0^T[g_{km}(t,\tilde{S}')-g_{km}(t,\tilde{S}'')]^2\,dt$ is distance between Kaplan-Meyer estimates of survival curves by sets $\tilde{S}',\tilde{S}''$, where T maximal time of observation.. Then quality functional $\mathrm{F}(R,\tilde{S}_0)$ value for some partition R is calculated as $\mathrm{F}(R,\tilde{S}_0)=\sum_{i=1}^{k}\{\left|\tilde{S}_i\right|\rho[g_{km}(t,\tilde{S}_i),g_{km}(t,\tilde{S}_0)]\}$, where $\left|\tilde{S}_i\right|$ is number of objects in set $\tilde{S}_i$.

The optimal partition R_{opt} inside some family is searched by calculating all possible values of $\mathrm{F}(R,\tilde{S}_0)$ and selecting partition corresponding maximal value. Found regularities (optimal partitions) are statistically verified with the help of permutation test (PT) that is based on testing basic null hypothesis that survival is fully independent on involved explanatory variables. The optimal value of quality functional $\mathrm{F}(R,\tilde{S}_0)$ is used as PT statistics. Let optimal partition was found for dataset $\tilde{S}_0=\{(\alpha_1,t_1,\mathbf{x}_1),\ldots,(\alpha_m,t_m,\mathbf{x}_m)\}$. Let $\mathrm{F}(R_{opt},\tilde{S}_0)$ is the optimal value of quality functional. To evaluate statistical validity of discovered regularity set of N random permutations $\{\pi_1,\ldots,\pi_{\mathrm{N}}\}$ is calculated with the help of random numbers generator. Initial dataset $\{(\alpha_1,t_1,\mathbf{x}_1),\ldots,(\alpha_m,t_m,\mathbf{x}_m)\}$ and permutations $\{\pi_1,\ldots,\pi_{\mathrm{N}}\}$ give rise to permuted datasets $\{\tilde{S}_1^r,\ldots,\tilde{S}_{\mathrm{N}}^r\}$, where $\tilde{S}_j^r=\{(\alpha_{\pi_j(1)},t_{\pi_j(1)},\mathbf{x}_1),\ldots,(\alpha_{\pi_j(m)},t_{\pi_j(m)},\mathbf{x}_m)\}$. For each dataset $\tilde{S}_{\pi_j}^r$ from $\{\tilde{S}_1^r,\ldots,\tilde{S}_{\mathrm{N}}^r\}$ optimal partition is searched inside the same family for the same variable (variables) and by optimizing the same quality functional that were previously used in case of $\tilde{S}_0$. Let $\mathrm{N}_{gt}[F(R_{opt},\tilde{S}_0)]$ is the number of datasets in $\{\tilde{S}_1^r,\ldots,\tilde{S}_{\mathcal{N}}^r\}$ for which

$F(R_{opt}, \tilde{S}_*^r) > F(R_{opt}, \tilde{S}_0)$. The ratio $N_{gt}[F(R_{opt}, \tilde{S}_0)]/N$ is used as estimate of PT p-value.

Usually from 1000 to 5000 permutations are used to evaluate p-value with the help of PT. Important advantageous of permutation tests are absence of apriori suppositions about probability distributions. There are no demands also to size of data sets.

Discussed method was realized and tested at dataset of oncology patients.

Collective method for individual survival curves calculating.

It is often important not only evaluate influence of some variables on survival but also to estimate the hazard (or risk) of death, or other event of interest, for individual object, given all prognostic variables set.. Such task is usually solved with the help of survival models. The most widespread of them is Cox proportional hazards model. However existing models demands rather strong assumptions about probability distribution. In this section we try to show that approach based on ensembles of regularities search also allows to make multifactor hazard estimates for individual object. Let $\tilde{Q}$ is set of subregions of prognostic variables space that were found with the help of optimal portioning**.** Suppose that we try to estimate survival curve for object s^* by its vector description $\mathbf{x}^*$. Let $\mathbf{x}^*$ belongs to intersection of subregion $q_1, \ldots, q_r$ from system $\tilde{Q}$. We suggest to evaluate probability $\Pr(t > t_0)$ for object s^* as weighted sum:

$$\Pr(t' > t) = \frac{\sum_{i=1}^{r} wei_i g_{km}(t, \tilde{S}_i)}{\sum_{i=1}^{r} wei_i}$$, where wei_i is so called "weight " of subregion q_i

By analogy to a method of statistically weighed syndromes [6] "weight" is calculated under the formula: $wei(\tilde{S}_i) = \frac{|\tilde{S}_i|}{|\tilde{S}_i| + 1} \frac{1}{\hat{d}_i}$, where $\hat{d}_i = \sqrt{\frac{\sum_{s \in \tilde{S}_i} \rho[g_{km}(t, \tilde{S}_i), g_{km}(t, s)]}{|\tilde{S}_i|}}$ - a "dispersion"

$\rho[g_{km}(t, \tilde{S}_i), g_{km}(t, s)]$ - distance from survival curve for sample $\tilde{S}_i$, to survival curve of an object s from this sample. In case of object s with known "death" moment t^* survival "curve" is step function equal 1 at $[0, t^*)$ and 0 at $[t^*, T]$. But we face the following problem: how to work with censored objects (i.e. objects from the training sample for which object was alive at moment of last observation and time of death is unknown). It is offered to reduce task step survival "curves" restoration for such objects to set of recognition tasks with 2 classes. Classes are defined with the help of several boundary points for time and uncensored objects are used as training sets. Results of recognition can be verified with the help of cross validation and partly at the set of censored objects. State of censored object (dead or alive) is established according

recognition results.

Conclusion

So method of survival analysis that is based on based on optimal partitioning of prognostic variables space with using permutation test for verification and method of survival curves evaluating for individual objects. Methods do not demand apriori assumptions about probability distributions of variables. Methods may be used in variety of tasks in medicine, engineering and economics.

Acknowledgements The work was supported by RFFI grants 08-01-90427, 08-07-00437.

References

1) Sen'ko O.V., Kuznetsova A.V. (1998). The use of partitions constructions for stochastic dependencies approximation. Proceedings of the International conference on systems and signals in intelligent technologies. Minsk (Belarus), pp. 291-297.

2) Senko O.V., Kuznetsova A.V., Kropotov D.A. (2003). The Methods of Dependencies Description with the Help of Optimal Multistage Partitioning. **Proceedings** of the 18th International Workshop on Statistical Modelling Leuven, Belgium, 2003, pp. 397-401.

3) Oleg V.Senko and Anna V. Kuznetsova The Optimal Valid Partitioning Procedures . Statistics on the Internet ***http://statjournals.net/***, April, 2006

4) Kaplan E.L., Meier P. Nonparametric estimation from incomplete observations //J.Amer.Stat.Assoc. 1958, v.53, P.457-481.

5) Cox D.R. Regression models and life tables.//J.R. Statist. Soc., B. p. 34-187

6) Kuznetsov V.V., Senko O.V. , Kuznetsova A.V. et. all. Recognition of fuzzy systems by method of statistically weighed syndromes and its use for immune and hematologic norm and chronical pathology.//Chemical Physics, 1996, v.15, №1, p.