Conclusion

Within the framework of the Bayesian learning theory, we analyzed a classifier generalization ability for the recognition on finite set of events. It was shown that the obtained results can be applied for classification tree pruning. Numeric experiments showed that the Bayesian pruning has at least the same efficiency or better than standard reduced error pruning, and at the same time is more resistant to overtraining.

Acknowledgements

This work was supported by the Russian Foundation of Basic Research, grant 04-01-00858a

Bibliography

[1] Lbov, G.S., Startseva, N.G., *About statistical robustness of decision functions in pattern recognition problems.* Pattern Recognition and Image Analysis, 1994. Vol 4. No.3. pp.97-106.

[2] Berikov V.B., Litvinenko A.G. *The influence of prior knowledge on the expected performance of a classifier.* Pattern Recognition Letters, Vol. 24/15, 2003, pp. 2537-2548.

[3] UCI Machine Learning Database Repository. http://www.ics.uci.edu/~mlearn/ MLRepository.html

[4] Quinlan, J.R. *C4.5: Programs for Machine Learning.* Morgen Kaufmann, San Mateo, CA, 1989.

[5] Berikov V.B. *A priori estimates of recognition quality for discrete* features. Pattern Recognition and Image Analysis. V. 12, N 3, 2002. pp. 235-242.

Author's Information

Vladimir Berikov – Sobolev Institute of Mathematics SD RAS, Koptyug pr.4, Novosibirsk, Russia, 630090; e–mail: berikov@math.nsc.ru

EXTREME SITUATIONS PREDICTION BY MULTIDIMENSIONAL HETEROGENEOUS TIME SERIES USING LOGICAL DECISION FUNCTIONS[1]

Svetlana Nedel’ko

Abstract: A method for prediction of multidimensional heterogeneous time series using logical decision functions is suggested. The method implements simultaneous prediction of several goal variables. It uses deciding function construction algorithm that performs directed search of some variable space partitioning in class of logical deciding functions. To estimate a deciding function quality the realization of informativity criterion for conditional distribution in goal variables' space is offered. As an indicator of extreme states, an occurrence a transition with small probability is suggested.

Keywords: multidimensional heterogeneous time series analysis, data mining, pattern recognition, classification, statistical robustness, deciding functions.

ACM Classification Keywords: G.3 Probability and Statistics: Time series analysis; H.2.8 Database Applications: Data mining; I.5.1 Pattern Recognition: Statistical Models

[1] The work is supported by RFBR, grant 04-01-00858-a

Introduction

The specifics of multidimensional heterogeneous time series analysis consists in simultaneous prediction of several goal variables. But the most of known algorithms construct decision function for each goal variable separately. Such approach looses some information about features interdependencies [Mirenkova, 2002].

The next problem is strong increasing of dimensionality when analysing window length increases. So one has to either simplify decision functions class or make the window shorter.

The problem of insufficient sample appears much more essential [Raudys, 2001] when rare events are to be predicted.

In this work an algorithm of prediction multidimensional heterogeneous time series based on finding certain partitioning that maximizes informativity criterion [Lbov, Nedel'ko, 2001] for matrix of transitions between partitioning areas. This allows to avoid increasing complexity when a window get longer, but prediction looses accuracy.

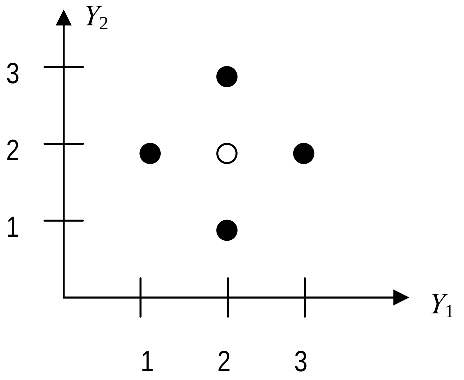

Fig. 1.

Extreme situations are characterised by low number of precedents in a period under observation. Therefore, one need statistically robust methods of multidimensional heterogeneous time series forecast.

It might be interesting also to predict events having only a few precedents or may be no precedents at all. In this case it seems to be impossible to forecast extreme situations themselves, but one could catch changing a probabilistic model of time series and consider this as an indicator of abnormal process behaviour.

Problem Definition

Let a random n-dimensional process $Z(t)=(Z_1(t), \ldots, Z_n(t))$ with discrete time be given. Features may include both continuous and discrete (with ordered or unordered values) ones. Suppose that for a time moment t values of n variables depend on its values in previous l time moments, i. e. on a window of length l.

The most algorithms for prediction multidimensional time series use replacement of time series sample by a sample in form of data table. This is made via new notation: goal values are designated as $Y_j(t)=Z_j(t)$, and previous values (prehistory) as $X_j(t)=Z_j(t\text{-}1)$, $X_{j+n}(t)=Z_j(t\text{-}2)$, ..., $X_{j+n(l\text{-}1)}(t)=Z_j(t\text{-}l)$ $j = 1, \ldots, n$.

Now any time series realization $Z(t)$, $t=\overline{1,T}$, may be represented like a sample $v=\left\{\left(x^i, y^i\right) \middle| \, i=\overline{1,N}\right\}$, where $N=T-l$ — the sample size. Here $y^i=\left(y_1^i,\ldots,y_j^i,\ldots,y_n^i\right)$, $y^i_j = Y_j(i)$, $x^i=\left(x_1^i,\ldots,x_j^i,\ldots,x_m^i\right)$, $x^i_j = X_j\ (i+l)$, $m = nl$ — predictor space dimensionality. Note that the first l time moments have no prehistory of length l.

Such notation allows using a data mining methods to predict each feature separately. They may be for example classification or regression analysis methods in logical decision functions [Lbov, Startseva, 1999]. But this approach neglects features interdependence, so it is possible to construct an examples where separate decision functions give incompatible forecast [Mirenkova, 2002].

Let's consider an example that shows the weakness of separate feature forecast. Suppose two discrete features are given and probabilistic measure on them is like shown on figure 1. Each of black points has probability 0,25; another points have probability zero. Methods those make decision for every feature separately give predicted value marked by white circle. But such value combination will never occur.

This example shows necessity in methods constructing a decision rule for all features together because interdependencies are important. One need also to use decision in form of an area (in the example such area contains four black points), but not a single point.

In this work, we suggest not to separate features onto X and Y but to build partitioning in space Z directly.

Quality Criterion

Let's introduce quality criterion for decision in form of areas if goal features space. Such type criteria were proposed in [Rostovtsev, 1978].

It's suitable now to consider again separately D_X – space of predictors and D_Y – a goal features space. Let $P(E_Y)$ and $P(E_Y|x)$ be unconditional and conditional measures for $E_Y \subseteq D_Y$. Suppose a set $B_Y = \{E_Y^d \subseteq D_Y \mid d = 1,...,k\}$ of non-intersected areas to be given. Then quality criterion will be $K(B_Y) = \sum_{d=1}^{k}\left(P(E_Y^d|x) - P(E_Y^d)\right)$. Optimal decision in x will be $B_Y^* = \arg\max K(B_Y)$.

Quality criterion for conditional probabilistic measure may be defined as $K(P[D_Y|x]) = \max_{B_Y} K(B_Y)$.

This criterion is some kind of distance between conditional by given x and unconditional measures on goal features space. There are known modifications those use uniform distribution instead of unconditional one.

If B_Y is a partitioning of D_Y one needs to use modified criterion:

$$K'(B_Y) = \sum_{d=1}^{k}\left| P(E_Y^d \mid x) - P(E_Y^d)\right|. \qquad (1)$$

It differs in taking absolute values.

When the distribution is unknown and we have a sample only we can't estimate criterion for each $x \in D_X$, so need to build some partitioning λ of D_X.

Then $K(\lambda) = \sum_{E_X \in \lambda} K(P[D_Y|E_X]) \cdot P(E_X)$ will be integral decision quality criterion.

All probabilities in expression may be estimated on sample.

Algorithm

Suggested algorithm makes partitioning directly in space $D_Z = \prod_{j=1}^{n} D_j$, where D_j – a set of feature Z_j all values.

Since partitioning $\lambda = \{E^i \in D_Z \mid i = \overline{1,k}\}$ was fixed initial time series $Z(t)$ may be represented by one symbolic sequence $\beta(t) \in \{\beta^i \mid i = \overline{1,k}\}$, where β_i – a symbol correspondent to area E^i, and $\beta(t) = \beta_i$ when $Z(t) \in E^i$.

Criterion (1) may be applied to transition matrix of process $\beta(t)$:

$$K'(\lambda) = \sum_{i_0=1}^{k}\cdots\sum_{i_l=1}^{k}\left| p_{i_0 \ldots i_l} - \left(\sum_{j_0=1}^{k} p_{j_0 i_1 \ldots i_l}\right)\left(\sum_{j_1=1}^{k}\cdots\sum_{j_l=1}^{k} p_{i_0 j_1 \ldots j_l}\right)\right|, \qquad (2)$$

where $p_{i_0 \ldots i_l} = \mathrm{P}\left(\bigwedge_{\tau=0}^{l}\left(\beta(t-\tau) = \beta^{i_\tau}\right)\right) = \mathrm{P}\left(\bigwedge_{\tau=0}^{l}\left(Z(t-\tau) \in E^{i_\tau}\right)\right)$ — the probability of given prehistory of length l.

To obtain sample estimation of the criterion need to replace $p_{i_0 \ldots i_l}$ by $N_{i_0 \ldots i_l}/N$ – a rate of prehistory appearance in the sample.

Transition probabilities for partitioning areas are a kind of multi-variant decision functions [Lbov, Nedel'ko, 2001].

Note that a partitioning λ may be constructed in any appropriate class, e. g. by linear discriminating functions or by logical deciding functions (decision trees).

Logical Decision Functions

For constructing a partitioning λ we shall use algorithm LRP [Lbov, Startseva, 1999] that builds a decision tree. This algorithm was designed first for classification task and applied then for various tasks of data analysis by using special quality criteria.

The algorithm builds a partitioning onto multidimensional intervals. Here an interval is a set of neighbour values when order is defined or any subset of values if feature values are unordered. Multidimensional interval is a Cartesian product of intervals.

Algorithm LRP makes sequential partitioning the space D onto given number of areas.

Since partitioning $\left\{E^1,...,E^i,...,E^s\right\}$, $E^i \subseteq D$, was constructed on step $s-1$, on step s the algorithm goes over the all areas and selects one that being split by all possible ways onto two sub-areas provides criterion maximum. Then these sub-areas replace initial area and the process is repeated until k areas been produced.

The partitioning may be represented by decision tree. Each non-terminal node ω is correspondent to some predicate $P^\omega \equiv \left(z_j \in E_j^\omega\right)$, $E_j^\omega \subseteq D_j$. Each terminal node corresponds to an area of the partitioning λ.

Rare Events Prediction

Extreme situations are characterised by low number of precedents in a sample. Therefore, statistical robustness of the methods used is especially actual. Proposed method of multidimensional heterogeneous time series prediction provide high robustness.

Nevertheless, it may be not enough if there are only several precedents. Moreover, it might be interesting to predict events having no precedents.

Obviously, in this case reliable prediction is impossible, but one could try to mark time moments where extreme situation is probable. One of indicators for such time moments may be changing a probabilistic model of time series.

Since we represent initial time series by correspondent Markov chain, all related mathematical results are available. So, a moment of changing a probabilistic model can be revealed.

Another indicator of process abnormality might be occurring in correspondent symbolic chain a transition with small probability.

Conclusion

Methods of simultaneous prediction the all variables of multidimensional heterogeneous time series allows using features interdependence information in comparison with method of separate constructing a decision function for each feature. It's possible also to build decision based on partitioning initial features space that decreases algorithm complexity. As quality criterion the method uses transition matrix informativity that was introduced.

The method proposed represents initial time series by correspondent Markov chain that allows avoiding great increasing complexity when considered prehistory length increases. This is especially important for predicting rare events. Such representation also allows applying all mathematical results related to Markov chains.

To predict time moments when extreme situations have higher probability here was suggested using changes in probabilistic model of time series.

Bibliography

[Lbov, Startseva, 1999] Lbov G.S., Startseva N.G. Logical deciding functions and questions of statistical stability of decisions. Novosibirsk: Institute of mathematics, 1999. 211 p. (in Russian).

[Rostovtsev, 1978] P. S. Rostovtsev. Typology constructing algorithm for big sets of social-economy information. // Models for aggregating a social-economy information. Proceedings, publ. IE and SPP SB AS USSR, 1978. (in Russian).

[Lbov, Nedel'ko, 2001] G.S. Lbov, V.M. Nedel'ko. A Maximum informativity criterion for the forecasting several variables of different types. // Computer data analysis and modeling. Robustness and computer intensive methods. Minsk, 2001, vol 2, p. 43–48.

[Raudys, 2001] Raudys S., Statistical and neural classifiers, Springer, 2001.

[Mirenkova, 2002] S. V. Mirenkova (Nedel'ko). A method for prediction multidimensional heterogeneous time series in class of logical decision functions // Artificial Intelligence, No 2, 2002, p. 197–201. (in Russian).

Author's Information

Svetlana Valeryevna Nedel'ko – Institute of Mathematics SB RAS, Laboratory of Data Analysis, 630090, pr. Koptyuga, 4, Novosibirsk, Russia, e-mail: nedelko@math.nsc.ru

EVALUATING MISCLASSIFICATION PROBABILITY USING EMPIRICAL RISK[1]

Victor Nedel'ko

***Abstract**: The goal of the paper is to estimate misclassification probability for decision function by training sample. Here are presented results of investigation an empirical risk bias for nearest neighbours, linear and decision tree classifier in comparison with exact bias estimations for a discrete (multinomial) case. This allows to find out how far Vapnik–Chervonenkis risk estimations are off for considered decision function classes and to choose optimal complexity parameters for constructed decision functions. Comparison of linear classifier and decision trees capacities is also performed.*

***Keywords**: pattern recognition, classification, statistical robustness, deciding functions, complexity, capacity, overtraining problem.*

***ACM Classification Keywords**:I.5.1 Pattern Recognition: Statistical Models*

Introduction

One of the most important problems in classification is estimating a quality of decision built. As a quality measure, a misclassification probability is usually used. The last value is also known as a risk. There are many methods for estimating a risk: validation set, leave-one-out method etc. But these methods have some disadvantages, for example, the first one decreases a volume of sample available for building a decision function, the second one takes extra computational resources and is unable to estimate risk deviation. So, the most attractive way is to evaluate a decision function quality by the training sample immediately.

But an empirical risk or a rate of misclassified objects from the training sample appears to be a biased risk estimate, because a decision function quality being evaluated by the training sample usually appears much better than its real quality. This fact is known as an overtraining problem.

To solve this problem in [Vapnik, Chervonenkis, 1974] there was introduced a concept of capacity (complexity measure) of a decision rules set. The authors obtained universal decision quality estimations, but these VC–estimations are not accurate and suggest pessimistic risk expectations.

For a case of discrete feature in [Nedel'ko, 2003] there were obtained exact estimations of empirical risk bias. This allows finding out how far VC–estimations are off.

The goal of this paper is to extrapolate the result on continuous case including linear and decision tree classifiers.

[1] The work is supported by RFBR, grant 04-01-00858-a