Criteria of a Selection of Attributes: Minimum of Mistakes Versus FRiS-function

Nikolay Zagoruiko, Irina Borisova, Olga Kutnenko

Institute of Mathematics of SD RAS, pr. Koptyg, 4, Novosibirsk, 630090, Russia

zag@math.nsc.ru

Abstract. *For an estimation of informativeness of separate attributes or their subsystems it is offered to use average value normalized functions of a rival similarity (FRiS) objects of training sample to the pattern. This criterion differs from criterion in the form of number of correctly recognized objects higher connection with results of recognition of control sample, a greater noise stability, an opportunity to estimate suitability of the chosen attributes and reliability of recognition of control object.*

Keywords

Function of rival similarity, minimum of mistakes, feature selection, suitability of attributes.

1 Introduction

For estimation the informativeness of attributes the share of m_1 objects of training sample (U), correctly recognized in a mode cross-validation is usually used. The long-term practice of use of such criterion of informativeness shows its serious lacks. Very often attributes highly appreciated by U-criterion, show bad results at recognition of control sample. The reasons of this phenomenon consist that the U-criterion does not consider character of distribution of training sample and reacts only to that fact that the object has appeared in borders of the own pattern or another's.

Fisher suggests to consider features of a situation more full: it is necessary to estimate remoteness of expectation of patterns (m_1 and m_2) from each other and dispersions (d_1 and d_2) these patterns. His criterion $Q = |m_1 - m_2| / (d_1 + d_2)$ gives more solvent estimations of the informativeness attributes. However we shall apply it only to cases of normal distribution of patterns. Many modern tasks deal with samples in which the number of objects is less than dimension of space of attributes. In these conditions to estimate the statistical moments of distribution it is inconvenient.

It would be desirable to have criterion which would reflect Fisher's basic idea, but would not be focused on this or that law of distribution of sample. As such criterion it is offered to use function of rival similarity (FRiS).

In the first section this function and its properties is described. In the second section advantage of FRiS-criterion in comparison with other criteria informativeness of attributes is shown. In the third section application of FRiS-criterion for an estimation of suitability of attributes is described. In summary it is underlined use FRiS for the decision of other Data Mining problems.

2. Function of rival similarity (FRiS)

At recognition of control object Z this or that function of similarity of object with standards of all patterns S_i, $i=1, 2, \ldots, K$ is usually used. The object Z is considered belonging that pattern S_i, similarity to which standard has appeared maximal. In the literature there are tens variants of measures of similarity [1]. The size, inversely proportional to distance r_i from object up to standard S_i is most often used.

It is possible to note two lacks of the majority of these measures. First, they have absolute character while in human perception of a category of similarity have relative character. Secondly, they do not consider a distribution of those patterns to which the control object is compared.

Really, the answer to a question of type « Is similar or not similar? », « close or far? », etc. depends on the answer to a question « In comparison with what (whom)? ». In some measures of similarity it is considered, and in recognition the competitive situation is considered. So in the k nearest neighbor rule (*kNN*) the decision on a belonging of object *Z* to first pattern is accepted not in that case when the distance r_1 is "little" but when it less distances r_2 till a rival pattern. Hence, to estimate similarity of object *Z* for the first pattern, it is necessary to know distance not only up to it, but also up to the nearest competitor, and to compare these distances in a scale of the order.

It has appeared, that it is possible to take advantage of knowledge of sizes r_1 and r_2 more effectively if to consider not only the relation of the order between them, but also stronger relations. For example, size

$$F_1 = (r_2 - r_1)/(r_1 + r_2) \quad \ldots \ldots (1)$$

will characterize similarity of object *Z* with the first pattern in a competition to the second pattern in a scale of differences. Thus value of function of rival similarity *F* varies within the limits of from +1 up to-1. If the control object *Z* coincides with the standard of the first pattern, r_1=0 and F1=1. Similarity Z with the standard of the second pattern thus will be equal *F2 =-1*. At distances $r_1 = r_2$ values *F1=F2=0* that specifies border between patterns. In points of border the object is equally similar and not similar to these competing patterns. Function *F* has relative character and will well be agreed with mechanisms of perception of similarity which the person uses.

To consider features of distribution of patterns to which the object Z is compared, it would be necessary to normalize according to Fisher distances r_i on dispersions of patterns. But it was above marked, that a estimation of dispersion it is possible not always. In this connection features of distributions are offered to be estimated in size d_i average distance between all pairs objects of an pattern. Then distance from object *Z* up to the standard of pattern S_i we shall consider equal $R_i = r_i / d_i$, and function of rival similarity *F* of signs a following kind:

$$F = (R_2 - R_1)/(R_2 + R_1) \quad \ldots \ldots (2)$$

Let's note one more feature of human perception of similarity. Between objects the person prefers to not notice small distinctions and considers these objects "similar". At increase in distinctions he starts to react to them and at achievement of some threshold of distinctions comes to conclusion, that objects "are various". Such characteristic of perception is characteristic, in particular, for acoustical system of the person. In this connection at recognition of speech signals Akaiki measure is applied: $f=a/(a+r_i^2)$. Function f looks like, similar to mirror display of symbol *S*. The steepness of bends of function *f* can be changed in size of a constant *a*. It is interesting, that the function reflecting ability of an environment to resist to destroying influences has a similar kind.

That function of rival similarity had the S-shaped form we shall transform the formula (2) to such kind:

$$F = 1 - 2x, \text{ where } x = R_1 / (R_1 + R_2) \quad \text{….. (3)}$$

Having added to *F* size $w = bSin(4\pi x)$, we shall receive:

$$F=1-2x+w \text{ … … (4)}$$

Let's limit values of a constant *b* to limits (*0-0,14*). At *b=0* we shall receive the dependence described by the formula (2). If to take *b> 0,14* the size of similarity can accept values more than 1, that is not comprehensible. On fig. 1 variants of functions of similarity are presented at different parameters *b* and *d*.

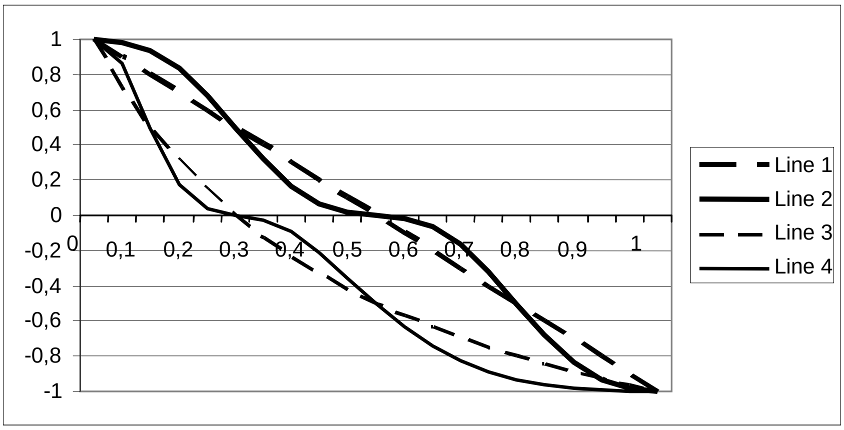

Fig. 1. A kind of function of rival similarity at different values of parameters b and d.
A line 1: b=0, $d_1=d_2$. A line 2: b=0.14, $d_1=d_2$.. A line 3: b=0, $d_1=3d_2$. A line 4: b=0.14, $d_1=3d_2$.

Function of rival similarity (FRiS) possesses following useful properties: reflects features of human perception of similarity; reflects relative character of a category "similarity"; considers features of distribution of objects; adaptable to any kind of distribution of objects; accepts values in a range from +1 up to-1.

The offered function has appeared useful to the decision of many tasks of Data Mining: for automatic classification (clustering), construction of decision rules, censoring of samples and others. In the given work its use as criterion for an estimation informativeness and suitability of attribute spaces is considered.

In the experiments described in following sections, the most simple kind of function *F* presented by the formula (1) and line 1 was used.

3. Comparison the criteria of informativeness of attributes.

Let there are two patterns, presented in training sample by the objects and standards. If patterns are divided by linear borders the estimation of informativeness, found by criterion *U*, will not depend on distances between objects inside of patterns and distance from objects up to dividing border. Unlike it, average value of function of rival similarity of objects with the standards (*Fs*) will depend on these factors. Those objects which settle down close to the standards and are considerably removed from dividing border, have higher value of function *F*, than the peripheral objects close to border.

We checked this statement by experimental comparison of three criteria of informativeness: functions of similarity (*F*), Fisher's criterion (*Q*) and criterion of correct recognition of training sample (*U*). Modeling initial data consisted of 200 objects of two patterns (on 100 objects of each pattern) in 100-dimensional space. Attributes were generated so that they possessed different informativeness. As a result of 30 attributes were to some extent informative, and the others 70 attributes were generated by the random-number generator and were obviously no informative. Under this table algorithm AdDel [2] the most informative subsystems of dimension *n* (from 1 up to 22) got out. For training casually got out on 35 objects of each pattern were used. On the control the others 130 objects were shown.

To reliability of recognition of control sample at use of criteria U, Fs and Q, average on 10 experiments, are shown on fig. 2. From them it is visible, that the attributes chosen by criterion *U*, worse chosen by criteria *Q* and *Fs*. It is possible to explain it to that measures Q and Fs is better consider features of distribution of objects, than a measure U. Advantages of criterion *Fs* above Fisher's *Q* criterion speak that measure *Q* is guided by linear decision functions, and a measure Fs - on more powerful class of piece-linear dividing borders.

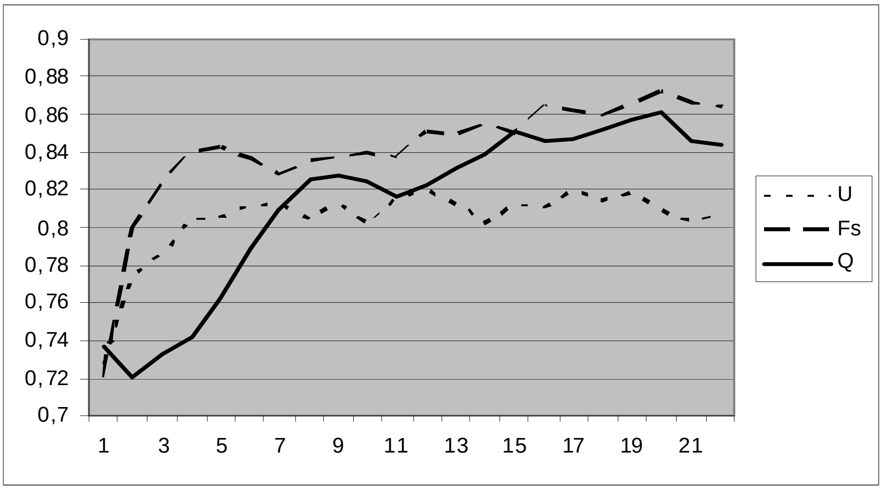

Fig. 2. Results of a choice of subsystems of attributes at use of three criteria: on number of correctly recognized objects (*U)*, on function of similarity (*Fs*) and on Fisher's *Q* criterion.

Criteria *U* also *Fs* were investigated on stability to handicaps. For this purpose the initial table of the previous experiment was distorted by gauss noise of different intensity. At each noise level (from 0,05 up to 0,3) the best subsystems by these criteria got out. Results are presented in figure 3 from which it is visible, that the criterion *Fs* is steadier, than criterion *U*. Results on the control show a high degree of correlation of criterion *Fs* with the results received on training. It testifies about high prognostic properties of criterion *Fs.*

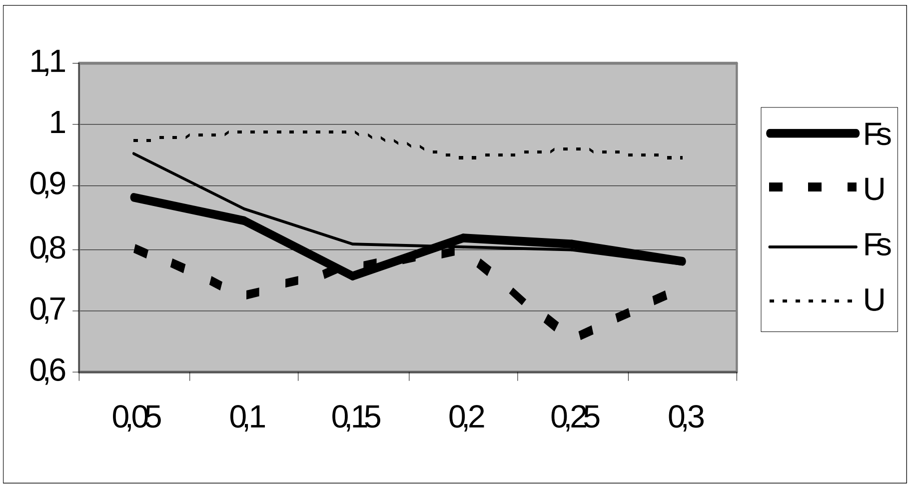

Fig. 3. Results of training and recognition by criteria *U* and *Fs* at different levels of noise. Thin lines - training, fat - the control.

The described experiments confirm essential advantages of criterion informativeness *Fs* in comparison with other known criteria.

4. Estimation of "suitability" of the chosen subsystems

In 1933 A.N. Kolmogorov has published work [3] in which has paid attention to problem a choice of a subset informative attributes among the big number of initial attributes. Here business is not only in the big labor input of this task. In the case a small number of objects and a lot of attributes, among them there can be successful combinations of rustling attributes on which training sample is recognized well. Check of these attributes on control sample will show that they were not suitable for reliable decision-making. Last years the urgency of a problem of an estimation of suitability, a no randomness of attributes for pattern recognition has strongly increased. Real tasks, for example, in genetics in which tens objects of training sample it is described by tens thousand attributes began to meet. By criterion *U* it is possible to find tens subsystems from a small number of attributes (from 3 up to 10) on which training sample is recognized very well. How among them to choose subsystems which will be suitable for recognition of control objects?

We offer a next way for decision of this problem. Let the subsystem from *n* attributes is found under the training table consisting from *N* of attributes ($n << N$) and *M* of the objects divided on two patterns. The choice of a subsystem was done by criterion *Fs.* Its value for this subset is equal Fs^*.

By the random-number generator we create the table of the same size $N*M$, casually we divide objects into two classes and by same criterion *Fs* the best *n*-dimension subsystem it is chosen. Value of criterion of this subsystem will be equally Fs'. Having repeated *T* time procedure of generation of casual tables and a choice of subsystems from *n* attributes, we shall receive *T* estimations of their quality. Among them we shall find a subsystem having the maximal estimation of criterion $Fs'(max)$.

If it has appeared, that $Fs^* < Fs'(max)$ the subsystem chosen by us on the real table is not suitable. It is not better than any casual subsystem of attributes. If $Fs^* > Fs'(max)$ the subsystem of attributes chosen by us can be considered as suitable for further use.

On the table of data which was used in the previous experiments, from 100 attributes the subsystem of 20 attributes has been chosen with criterion $Fs^* = 0.87$. On ten casual tables of the same size 100*200 best 20-dimension subsystems had values *Fs* from 0.61 till 0.67. Values *Fs* for the subsystems found under the initial table, lie considerably above this corridor and consequently can be considered not casual, suitable for recognition of control objects.

5 Conclusion

Carried out researches allow drawing following conclusions:

1. For an estimation the informativeness of attributes it is necessary to use not quantity of correctly recognized objects of training sample (*U*), but average value of function *Fs* of similarity of training objects with standards of the patterns.
2. The values of a measure *Fs* received on the training table and on a series of casual tables of the same size, allow receiving an estimation of suitability chosen attributes for recognition of control sample.

6 Acknowledgements

This work was supported by Russian Fond of Basic Researches, Grant № 05-01-00241

References

[1] Voronin Yu.A. The fundamentals the theory of similarity. Novosibirsk: Edition of Computer Centre, 1989.

[2] Zagoruiko N.G. Applied methods of the data and knowledge mining. Novosibirsk: Edition of Institute of Mathematics the Siberian Branch of the Russian Academy of Science, 1999.

[3] Kolmogorov A.N. To a question on suitability of the forecast formula founded by statistical way. - Factory laboratory. 1933. №1. pp. 164-167.