THE STABLE INTEGRAL INDICATORS CONSTRUCTION USING A SUPPORT SET OF THE OBJECTS[1]

Vadim Strijov, Tatiana Kazakova
Computing Centre of the Russian Academy of Sciences
Vavilov st. 40, 119993, Moscow, Russia
strijov@ccas.ru

Abstract. This paper describes an integral indicator construction algorithm. The integral indicator is a linear combination of the object features. The features are in the linear scale. Outliers in the features are assumed. So the problem of stable integral indicators construction arises. To construct a stable integral indicator a special-defined subset of the objects is selected. A non-supervised type of algorithm is used to make the integral indicator. The proposed algorithm was applied to construct the integral indicator of the foodstuff pollution level in the Russian regions.

Keywords: integral indicator, stable estimation, principal components analysis, regression.

1. Introduction

Assume each object is described as a row-vector of the feature values. The features are in the linear scale. An indicator of an object is a scalar defined by a convolution of the feature values of the object. For the set of the objects denote a vector of indicators of the objects as an integral indicator. Define an indicator as a linear combination of the object features [1,2]. Define the optimal weights of the features by an information criterion. In this paper the criterion of the maximal value of information by S. R. Rao [3] is used. This criterion is described in the next part of the paper. However this criterion does not provide an adequate comparison of the objects if there outlier values in the feature space are observed. Experts, who define the set of the objects, assume all objects to be comparable and expect an adequate integral indicator to be computed. However if several feature values are much bigger than the average, then the linear model will be affected since the outlier objects bring a big influence to the indicator. In fact they can determine the result even more than the rest of the objects. Exclusion of such objects can drastically change value of the indicator. Some times this operation changes as much as the ordinal-scaled representation of the integral indicator.

A number of techniques to construct stable integral indicators for the objects of the linear scales were proposed [4]. The goal of this paper is to introduce a stable integral indicator construction algorithm. It deals with the set of the objects, which contains outliers. This algorithm discovers a set of the support objects and then compute an indicator for the whole set of the objects. The key idea is to split the set of the objects into two subsets. The first one is a support set and the second one contains only outliers. The probability of an object belongs to the support set is calculated. When support set is constructed the algorithm computes feature weights using support set only. The principal component analysis is used to obtain the optimal feature weights. At the next step the obtained weights are used to make the integral indicator for the whole set of the objects.

2. Integral Indicator Construction

We have a sample set of m objects and a set of n features. This defines a matrix $A \in R^{m \times n}$, where an element a_{ij} is j-th feature value for i-th object. A row vector $\mathbf{a}_i = \langle a_{i1}, ..., a_{in} \rangle$ is the description of i-th object.

The indicator for an object is the linear combination

[1] This research is supported by the RFBR, grant 04-01-00401.

$$q_i = \sum_{j=1}^{n} w_j g_j(a_{ij}), \quad (1)$$

where g_j is a function, which maps the feature values into a unified scale:

$$g_j : a_{ij} \mapsto (a_{ij} - \min_i a_{ij})(\max_i a_{ij} - \min_i a_{ij})^{-1}, \; i = 1,...,m; j = 1,...,n. \quad (2)$$

If in (2) the denominator is zero for some j, then j-th feature can not be used and so it should be excluded from consideration. Without limitation of the applicability assume the following. If some value of given feature for *i*-th object is not less than value of the feature for *k*-th object then an integral indicator for *i*-th object is not less than an integral indicator for *k*-th one. It is the condition of monotonicity. From this condition and (2) it follows that the feature weights are positive. Since an integral indicator is expected to be invariant for linear combinations, define an additional conditions for the feature weights: $\sum_{j=1}^{n} w_j^2 = 1$.

The matrix A has to be prepared so that the conditions above are satisfied. The matrix has to fit the concept "the bigger the better". It means that an expert expect an object with bigger feature values has bigger indicator. An object of the maximal indicator is considered to be the best as well as a feature of the maximal weight be the most important.

One of the results of the non-supervised integral indicator construction algorithm is a vector of the feature weights $\mathbf{w} = \langle w_1,...,w_n \rangle^T$. To obtain weights the Principal Components Analysis [5] is used. So the indicator satisfies the maximal value of information criterion, developed by S. R. Rao [3]. According to this criterion the maximal value of information is achieved when sum of Euclidian distances between the object descriptions $\{\mathbf{a}_i\}$ and their projections $\{\mathbf{p}_i\}$, i=1,…,n, to the first principal component is minimal.

To find the first principal component, it is required to find linear combinations $Z^T = WA^T$, ($WW^T = I$) of column-vectors of the matrix A, such that the column-vectors $\mathbf{z}_1,...,\mathbf{z}_n$ of the matrix Z have the maximal variances, $\max \sum_{j=1}^{n} Dz_j$. Rows of the matrix W are eigenvectors of the covariance matrix $\Sigma = A^T A$. Therefore the integral indicator $\mathbf{q}$ is the projection of row-vectors of the matrix A to the first principal components, $\mathbf{q} = A\mathbf{w}$, where $\mathbf{w}$ is the row-vector of the matrix W^T, which corresponds to the maximal eigenvalue of the covariance matrix Σ.

3. Stable Integral Indicators Construction

Consider regularization techniques, which can be used to construct stable integral indicators to compare with. A. M. Shurygin described two methods of covariance matrix Σ regularization [6]. The first method is the ridge-regression $\Sigma_{r\beta} = \Sigma + \beta I$, where β is a regularization coefficient. The second method is the diagonal regression $\Sigma_{d\nu} = (1-\nu)\Sigma + \nu \cdot diag(\Sigma)$, where $\nu \in [0,1]$ is a regularization coefficient. It was proved that the second method produces more robust results.

However the regularization techniques cause information loss. Try to keep the maximal information value of given matrix A using the support set of the objects. To make the support let one should do the following.

Denote the set of the object descriptions $S_0 = \{\mathbf{a}_1,...,\mathbf{a}_m\}$. Let $\mathbf{S} = \{S_1,...,S_l\}$ be a set of all subsets of S_0, where the number of elements is $l = 2^m$. The proposed approach calculates the most informative linear predictor using S_ξ, then calculates the feature weights $\mathbf{w}_\xi = \mathbf{w}(S_\xi)$ using principal components analysis and returns the indicator $\mathbf{q}_\xi = A\mathbf{w}_\xi \in R^m$. Denote $\overline{S}_\xi$ the supplement

S_ξ to S_0. Exclude two trivial sets with the number of the elements $\#S_\xi = 1$ and $\#S_\xi = m$ from consideration. Assume feature values are independent random variables with a Gaussian distribution.

Then $p_\xi = P(\mathbf{a}_i \in S_\xi)$ is the probability of some object from S_0 belongs to S_ξ and $\overline{p}_\xi$ is the probability of some object from S_0 belongs to $\overline{S}_\xi$. Point in $\mathbf{S}$ the support set S_ξ, which brings the maximal value of the criterion $f_\xi = p_\xi / \overline{p}_\xi$.

Denote σ_ξ, $\overline{\sigma}_\xi$ the cumulative variances of the projections $\mathbf{p}_i$ of $\mathbf{a}_i$ in S_ξ or in $\overline{S}_\xi$ to the first principal components. Denote $n_\xi, \overline{n}_\xi, n_0$ the numbers of elements in the sets $S_\xi, \overline{S}_\xi, S_0$. The cumulative variance $\sigma^2(S_0)$ of the whole set is the sum of variances of the projections $\mathbf{p}_i$ from S_ξ or $\overline{S}_\xi$ weighted by probabilities of the corresponded objects are in S_ξ or $\overline{S}_\xi$:

$$\sigma^2(S_0) = p_\xi^2 \sigma^2(S_\xi) + p_\xi^2 \sigma^2(\overline{S}_\xi) = \frac{p_\xi^2 \sigma_\xi^2}{n_\xi} + \frac{\overline{p}_\xi^2 \overline{\sigma}_\xi^2}{\overline{n}_\xi}. \qquad (3)$$

To calculate the function f_ξ it is necessary to minimize the variance $\sigma^2(S_0)$. Since (3) should satisfy the condition $n_\xi + \overline{n}_\xi = n_0$ let us use the Lagrange method. So

$$L = \sigma^2(S_0) + \lambda(n_\xi + \overline{n}_\xi - n_0) = \frac{p_\xi^2 \sigma_\xi^2}{n_\xi} + \frac{\overline{p}_\xi^2 \overline{\sigma}_\xi^2}{\overline{n}_\xi} + \lambda(n_\xi + \overline{n}_\xi - n_0),$$

where λ is the Lagrange coefficient. The partial derivatives of L should be equal to zero:

$$\partial L / \partial n_1 = -p_1 \sigma_1^2 / n_1^2 + \lambda = 0, \quad \partial L / \partial \lambda = n_1 + n_2 - n = 0.$$

Therefore $p_1 \sigma_1 = n_1 \sqrt{\lambda}$. The last two expressions give the following results $n\sqrt{\lambda} = p_1\sigma_1 + p_2\sigma_2$ and $p_1 = n_1(p_1\sigma_1 + p_2\sigma_2)/n\sigma_1$. For the other derivative the similar expression for $\overline{p}_\xi$ is to be solved. Then the criterion $f_\xi = p_\xi / \overline{p}_\xi$ equals

$$\frac{p_\xi}{\overline{p}_\xi} = \frac{n_\xi \overline{\sigma}_\xi}{\overline{n}_\xi \sigma_\xi}. \qquad (4)$$

So the probability of the object $\mathbf{a}_i$ belongs to the support set S_ξ is proportional to the number of the objects n_ξ in this set and it is inverse proportional to the standard deviation σ_ξ of projection $\mathbf{p}_i$. The support set brings the maximal value to the criterion f_ξ. It defines the feature weights $\mathbf{w}_\xi$ according to the principal component analysis of S_ξ and so the integral indicator of the whole set of the objects is $\mathbf{q}_\xi = A\mathbf{w}_\xi$.

4. Numerical Experiment

A comparative analysis of the foodstuff polluted with quicksilver for the Russian regions was performed. An indicator was assigned to each region. It shows the pollution level of the foodstuff. Three groups of foodstuff were considered: meat foods, dairy products and bakery. The measured data from 29 regions was processed.

Normalization of the data was performed in the following way. In each region the quicksilver measurements were performed for each group of the foodstuff. The value of a_{ij} is a pollution value of j-th product in i-th region. This value is a ratio of 0.9-quintile of the quicksilver measurement

series for j-th product to the maximum permissible concentration of quicksilver in j-th product.

The proposed algorithm discovers the support set S_ξ, calculates the features weights $\mathbf{w}_\xi = \mathbf{w}(S_\xi)$ and constructs the stable integral indicator. The algorithm consists of three steps: the kernel of the support set definition, the support set searching and the integral indicator computation.

1. Find the center of the kernel set of the objects. To do that, calculate averages of the features for all objects of S_0. Remove the most distant (in Euclidian metric) object from S_0. Repeat this step until the last vector remains. This vector is the center of the kernel set. Assign 2/3 of the objects, which have the less distance from the center as the kernel of the support set.

2. Split S_0 into two subsets S_ξ and $\overline{S}_\xi$ so that the first one contains the kernel of the support set and the second one contains investigated outliers. Calculate the criterion function $f_\xi = p_\xi / \overline{p}_\xi$ for each split ξ. Select the set S_ξ, which brings the maximal value of the criterion.

3. The objects descriptions of the selected support set S_ξ define the matrix «object–feature» A_ξ. Use it to make the covariance matrix $\Sigma = A_\xi^T A_\xi$. The first eigenvector of the covariance matrix Σ defines feature weights $\mathbf{w}_\xi$. It gives integral indicator $\mathbf{q}_\xi = A\mathbf{w}_\xi$.

As the result the following outliers were discovered. The raw object descriptions contain the outlier values of the second feature (diary produce) in the regions: Karelia, St. Petersburg and Moscow. There outlier values in all three features in Karelia region were discovered. These regions were excluded to make the support set.

Table 1. Weights of the features before and after the proposed algorithm usage

Features\Algorithm	No regularization, $\mathbf{w}_1$	Regularization, $\mathbf{w}_2$	Support set, $\mathbf{w}_3$
Meat foods	0.0204	0.2264	0.4693
Diary produce	0.9983	0.7687	0.7706
Bakery	0.0548	0.5982	0.4312

Table 1 shows the feature weights calculated by three algorithms. The first one uses the principal components analysis without regularization. The second algorithm uses the principal components analysis with the diagonal regularization of the covariance matrix. There were shown that the diagonal regularization gives better results than the ridge-regression. The last algorithm is the principal components analysis for the support set. In the first case outliers increase the weight of the second feature significantly. The integral indicator depends on the diary produce only in the case of the 1-st algorithm. In the 3-rd case all features have almost equal impact to the indicator.

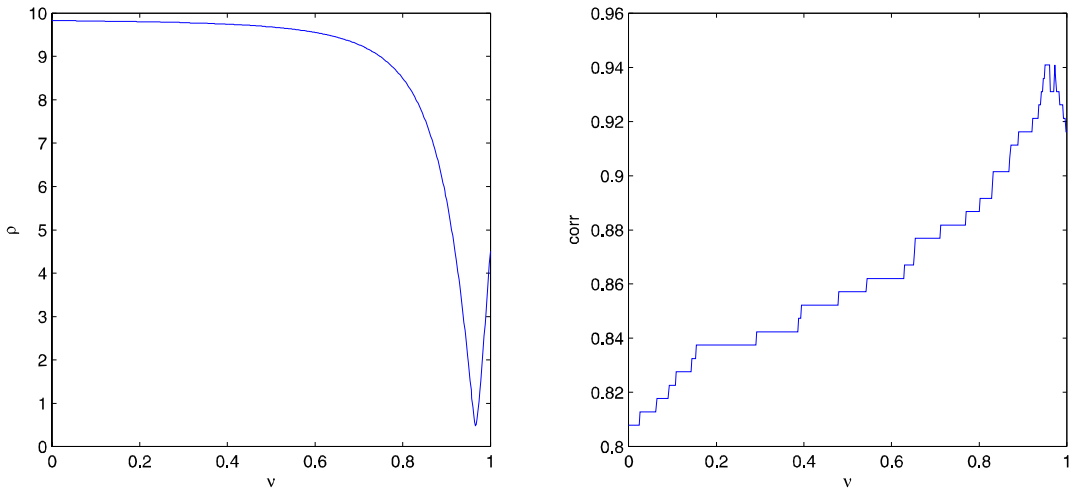

Fig. 1. Euclidian distance and rank correlation between the regularized integral indicator and the stable integral indicator

Diagonal regularization algorithm obtains an adequate indicator, too. However this algorithm is still affected by outliers. On the picture 1(left) the Euclidean distance $\rho = \|\mathbf{q}_2 - \mathbf{q}_3\|$ is displayed. Here $\mathbf{q}_2$ is the indicator for the diagonal regularization and $\mathbf{q}_3$ is the indicator for proposed algorithm. Regularization parameter denoted as ν. When $\nu = 0.9660$, the distance ρ is minimal. The picture 1(right) displays the Kendall rank correlation coefficient for the indicators $\mathbf{q}_2$ and $\mathbf{q}_3$. The maximal value of the correlation is 0.94. The rank correlation coefficient was shown due to its independency on indicator monotone transformation. It ignores outliers. Obviously, the regularization techniques can not ignore outliers completely.

Table 2. The integral indicators with no regularization and the integral indicators based on the support set (part of the table)

Region \ Foodstuff	$\mathbf{q}_1$	$r(\mathbf{q}_1)$	$\mathbf{q}_3$	$r(\mathbf{q}_3)$
Arkhangelskaya oblast’	0.5367	19	0.8356	23
Khabarovskiy kray	0.7986	21	0.6165	19
…	…	…	…	…
Vladimirskaya oblast’	0.0324	12	0.3577	14
Krasnodarskiy kray	0.0449	16	0.1578	10

The algorithm with no regularization computes an integral indicator, which significantly depends on the fact the set contains outliers or it does not. Rank correlation coefficient of the indicators for 1-th (no regularization) and 3-rd (proposed) algorithms is 0.82. It means these algorithms make 37 pairs of the investigated objects in different order. In the columns $\mathbf{q}_1$ and $\mathbf{q}_3$ of the Table 2 values of such pairs are shown. Columns $r(\mathbf{q}_1)$ and $r(\mathbf{q}_3)$ show region numbers according to the rank ordering.

5. Conclusions

In this paper the stable integral indicator construction algorithm is described. It was proposed to select the support set of the objects that brings the maximal value to the special criterion. The designed criterion is the ratio of the probabilities. It is proportional to the probability of an object

belongs to the support set and inverse proportional to the probability of an object is in the outlier set. The proposed algorithm is an alternative to regularization algorithms. Its advantage is it excludes outliers from consideration completely. The proposed algorithm was used to construct the integral indicator of the foodstuff pollution levels in the Russian regions.

References

1. Aivzian, S.A., Mkhitarian, V.S. Applied Statistics and essential econometrics. Moscow, UNITI, 1998. P. 363.
2. Orlov, A.I. State of art of the expert estimations theory. Moscow, Zavodskaya Laboratoriya. 1996 (1). P 61.
3. Rao, S.R. Linear Statistics Methods and Applications. Moscow, Nauka. 1968. P. 530–533.
4. Strijov, V., Shakin, V. Index construction: the expert-statistical method. Environmental research, engineering and management. 2003. No.4(26), P.51–55.
5. Jolliffe, I. T. Principal Component Analysis, 2nd ed., Springer. 2002.
6. Shurygin, A.M. Applied stochastic: robustness, estimations and forecast. Moscow, Finansy I Statistica. 2000. P. 99.