Is On-Line Data Analysis Safety? Pitfalls Steaming from Automated Processing of Heterogeneous Environmental Data and Possible Solutions

Jiří Jarkovský, Ladislav Dušek, Eva Janoušová

Is on-line data analysis safety? Pitfalls steaming from automated processing of heterogeneous environmental data and possible solutions.

Jiří Jarkovský[1], Ladislav Dušek[1], Eva Janoušová[1]

[1] Institute of Biostatistics and Analyses, Masaryk University, Kamenice 3, Brno, Czech Republic

Abstract. The current situation in environmental monitoring is characterized by increasing amount of data from monitoring networks together with increasing requirements on joining of these data from various sources in comprehensive databases and their usage for decision support in environmental protection and management. The automated analysis of such a heterogeneous datasets is a complicated process, rich in statistical pitfalls. There is a number of methods for multivariate classification of objects, e.g. logistic regression, discriminant analysis or neural networks; however, most of commonly used classification techniques have prerequisites about distribution of data, are computationally demanding or their model can be considered as "black box". Keeping these facts in mind, we attempted to develop a robust multivariate method suitable for classification of unknown cases with minimum sensitivity to data distribution problems; and thus, suitable for routine use in practice.

Keywords: classification, nonparametric, multivariate analysis, heterogeneous data.

1 Introduction

The current situation in environmental monitoring is characterized by increasing amount of data from monitoring networks together with increasing requirements on joining of these data from various sources in comprehensive databases and their usage for decision support in environmental protection and management. The important part of these requirements is demand on automated on-line analysis of data with immediate delivery of results.

The automated analysis of such a heterogeneous datasets is a complicated process, especially in case of multivariate analysis. The common tasks and their pitfalls in automated analysis are as follows.

Descriptive statistics of measured concentrations and sampling sites characteristics: i) Pitfalls: unfulfilled prerequisites of parametric descriptive statistics can easily lead to unrealistic results and automated testing and taking these prerequisites into account is extremely problematic; ii) Solutions: there is a well-accepted alternative of parametric descriptive statistics , i.e. nonparametric statistics.

Statistical tests of differences in measured values between/among groups of sampling sites or relationships between measurements: i) Pitfalls: unfulfilled

prerequisites of parametric tests lead to biased or incorrect results; ii) Solutions: nonparametric tests can be computed instead of parametric testing.

Classification of newly added samples or sampling sites into defined classes of environmental quality based on multivariate analysis of reference dataset: i) Pitfalls: most of commonly used classification techniques have prerequisites about distribution of data, are computationally demanding or their model can be considered as "black box"; Solutions: nonparametric models can be the solution but they are not common and well developed.

There are a number of methods for multivariate classification of objects, e.g. logistic regression, discriminant analysis or neural networks; however, these also have their problems, e.g. prerequisites of normality and absence of outliers for discriminant analysis [1]. Moreover, the methods should be used in a routine way in monitoring, i.e. without proper analysis of problems concerning the data. Keeping these facts in mind, we attempted to develop a robust multivariate method suitable for classification of unknown cases with minimum sensitivity to data distribution problems; and thus, suitable for routine use in practice.

2 Suggested methodology

The suggested methodology of the classification of unknown cases into categories of reference data for automated procedure should be as simple and robust as possible.
The simplest and the most objective measure of object association in multivariate space is their distance; thus, we decided to build our method on an analysis of a distance matrix among objects.

Now, selection of proper distance metrics is the first task in designing the method. We have adopted Gower distance metrics [2]; however, any multivariate distance metrics suitable for given data could be used. Concerning environmental monitoring data, there are some advantages in Gower metrics:

Continuous, binary or categorical parameters may be incorporated in computation: binary data is computed by coefficient – agreement and disagreement of values forming distance 0 or 1 respectively; categorical data is computed in the same way. Distance of objects according to continuous data is weighted to i) a parameter range in the data file or ii) an externally provided parameter range, i.e. difference in parameter values of objects is divided by parameter range to obtain partial metrics ranging from 0 to 1.

As noted above, parameters are weighted to their range, i.e. the influence of parameter absolute value is removed.

The final distance metrics ranges from 0 to 1 and could be interpreted easily.

Parameters in computation could be weighted according to expert knowledge or results of preliminary analysis. The final metrics takes the following form:

$$D\ (x_1, x_2) = \frac{\sum_{j=1}^{p} w_j d_{12j}}{\sum_{j=1}^{p} w_j}, \quad (1)$$

where D is a distance between objects x_1 and x_2, d_{12j} is a partial distance of objects x_1 and x_2 associated with parameter j (there are 1..p parameters; partial metrics associated with parameter ranges from 0 to 1) and w_j is a weight of parameter j ranging from 0 to 1.

Every homogeneous category of reference data could be characterised by its position in the multivariate space; and also, by its multivariate variability. Position of the reference category centroid (based on the median of continuous data and modus of binary/categorical data) exhibits representative of this group; multivariate radius of group provides the measure of its variability (in fact 95% percentile of radius is used in our computation to remove the influence of outliers). The distance of an unknown case to the centroid (**D**) is compared to the percentile of the reference category range (**R**). This ratio measures the extent to which an unknown case differs from objects incorporated in the reference category – see figure 1. Due to the fact that reference categories are not probably multivariate spheres we had to add a safety measure reflecting the real multivariate shape of the reference data. There are two parameters incorporated in the computation: the distance of an unknown case to the nearest neighbour in the reference group (**N**) and the measure of intragroup distances (**I**) within the reference group. The measure of intragroup distances is taken as median length of the MST branches (minimal spanning tree, [3]) of objects in the reference group. The following formula gives the measure of distance of an unknown case to the reference group x (U_x) in multiplies of the reference group x radius weighted for multivariate shape of this group.

$$U_x = \frac{abs(D + N - I)}{R}. \quad (2)$$

This computation could be also expressed as a probability of case U belongs to group x:

$$P(U_x) = \frac{1}{U_x} \times 100. \quad (3)$$

Where values over 100 % (i.e. objects inside the reference group) are truncated to 100 %. In the first step of the analysis, $P(U_x)$ is computed for all reference groups x=1..n and probability of unknown case belongs to a particular group is weighted as follows:

$$PW(U_x) = \frac{P(U_x)}{P(U_1) + P(U_2) + \cdots + P(U_n)}. \quad (4)$$

In the second step of the analysis, U_x or $P(U_x)$ based on case characteristics is adopted for assessing distance/similarity of an unknown case from/to a particular reference group. The main output is the probability of assigning a locality into the reference category based on case characteristics, i.e. to which reference category the evaluated case belongs.

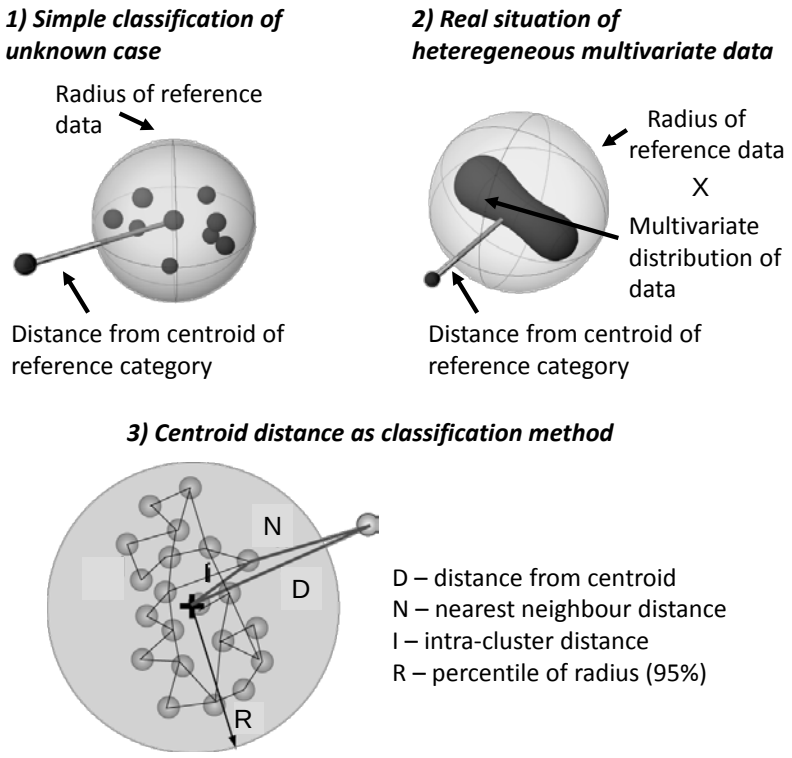

Fig. 1. Centroid distance method classification

3 Results of methodology testing

The presented methodology was tested on real datasets of 300 reference localities on river network thorough the whole Czech Republic. First, the localities were divided into 8 homogeneous clusters using k-means clustering. The clusters were based on parameters of natural heterogeneity (ecoregion, Strahler order, main river basin, width and depth of the stream, distance from well and altitude), the importance of factors

and mutual clusters position was validated using principal component analysis. The analyses were performed using Statistica for Windows (StatSoft, Inc., 2005).

The classification methods applied on data were: i) The novel method ("centroid distance") mentioned above; ii) Discriminant analysis; iii) Classification tree and iv) Neural network.

The dataset with the given groups of localities was divided into two files which were used for cross validation in these analyses. The following results of application of the models on independent cross validation datasets were obtained: i) Centroid distance: correct classification 91.3%; ii) Discriminant analysis: 87.6%; iii) Classification tree: 94.7%; iv) Neural network: 93.7%.

The results suggest that the developed methodology has similar predictive power as the commonly used methods or even better than some of them (discriminant analysis).

4 Conclusion

The presented methodology is a robust nonparametric classification method suitable for automated computing in heterogeneous environmental datasets. The predictive power of the method is comparable to commonly used parametric classification methods but without their extensive prerequisites and with simple interpretation of the classification model based on multivariate distances of objects.

Acknowledgments. This study was supported by FP7 project TaToo "Tagging Tool based on a Semantic Discovery Framework".

References

1. Legendre, P. and Legendre, L.:Numerical ecology. Elseviere Science BV, Amsterdam (1998)
2. Gower, J.C.: A general coefficient of similarity and some of its properties. Biometrics 27, pp. 857–871 (1971)
3. Prim, R.C.: Shortest connection networks and some generalizations. Bell Syst Tech J 36, pp. 1389–1401 (1957)