PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1169 2007
TAKSONOMIA 14
Klasyfikacja i analiza danych – teoria i zastosowania

Krzysztof Jajuga
Akademia Ekonomiczna we Wrocławiu

FROM MULTIVARIATE DISTRIBUTION TO DATA ANALYSIS – MODEL BASED CLUSTERING

1. Approaches in data analysis

The methods of statistical data analysis, including multivariate statistical analysis, have undergone the significant development since the well known paper by Tukey [1962]. Very many approaches have been proposed, suitable for different types of data. The most important classification contains two general approaches:

– descriptive (data-analytic) approach;

– stochastic approach.

Sometimes one links descriptive approach to exploratory data analysis and stochastic approach to confirmatory data analysis, but this may not be true in general situation. Particularly, confirmatory data analysis can be performed also with the use of non-stochastic methods, provided the underlying hypothesis to be confirmed (or rejected) is given.

It is worth to indicate that both approaches are model based. In the stochastic approach, the underlying model is multivariate distribution. In the descriptive approach the underlying model is usually particular characteristic of multivariate data set, for example: location vector (including mean vector), scatter matrix (including covariance matrix), regression function, principal components, etc.

The models used in descriptive approach have two important features. The first one is the fact, that characteristics of data set can be regarded as „analogies" of the parameters of multivariate distribution, particularly multivariate normal distribution. For example, location vector is descriptive analogue of mean vector of multivariate random variable.

The second important feature is the fact that characteristics of data set can be obtained through the solution of the optimization problem. For example:

– the location vector is the solution of the following minimization problem:

$$\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{v})^{\mathrm{T}}(\mathbf{x}_i - \mathbf{v}) \rightarrow \min, \tag{1}$$

– the regression function is the solution of the following minimization problem:

$$\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{x}_i\beta)^{\mathrm{T}}(\mathbf{y}_i - \mathbf{x}_i\beta) \rightarrow \min. \tag{2}$$

We now turn to stochastic approach. Here the model is obtained by the estimation of the parameters of this distribution. However, there are two possible ways of to analyze multivariate distribution. The first one is the classical way, where all parameters of multivariate distribution are estimated jointly by deriving likelihood function from multivariate density function and performing maximization problem.

The second way is through the use of copula functions. Then we get the decomposition of the multivariate distribution into two components, namely univariate distributions and the copula function linking these marginal distributions. Copula function reflects the dependence between the components of the random vector. The decomposition of multivariate distribution function is given in Sklar theorem [Sklar 1959], in the following formula:

$$F(x_1,...,x_m) = C(F_1(x_1),...,F_m(x_m)), \tag{3}$$

where:

F – the multivariate distribution function;

F_i – the distribution function of the i-th marginal distribution;

C – copula function.

So the copula function is the distribution function of the multivariate uniform distribution. One can also invert the presentation given by (3) to get:

$$C(u_1,...,u_m) = F(F_1^{-1}(u_1),...,F_m^{-1}(u_m)). \tag{4}$$

Therefore, to get multivariate distribution, one has to apply copula function to univariate distributions.

The other notion related to copula function is copula density. It is given as:

$$c(u_1,...,u_m) = \partial^m C(u_1,...,u_m), \tag{5}$$

$$f(x_1,...,x_m) = c(F_1(x_1),...,F_m(x_m)) \cdot f_1(x_1) \cdot ... \cdot f_m(x_m), \tag{6}$$

where:

f – the multivariate density function,

f_i – the univariate density function,

c – copula density.

As it can be seen, the analysis of multivariate distribution function is conducted by separating the analysis of marginal univariate distributions from the analysis of the dependence.

It is worth to mention, that multivariate normal distribution can be obtained by apply so called Gaussian (normal) copula function to univariate normal distributions. The Gaussian copula is given as:

$$C(u_1,...,u_m) = \Phi^m(\Phi^{-1}(u_1),...,\Phi^{-1}(u_m)), \tag{7}$$

where:

Φ^m – the distribution function of m-variate normal distribution;

Φ – the distribution function of univariate normal distribution.

Two other possibilities lead to multivariate distributions different from multivariate normal distribution:

– when Gaussian copula is applied to non-normal univariate distributions;

– when other than Gaussian copula is applied to normal univariate distributions.

Two particularly important – from the practical point of view – features of copula functions are:

– their flexibility allows to obtain multivariate distributions of non-elliptical shapes;

– the number of parameters to be estimated grows linearly with the increase of the dimensionality.

From the point of view of statistical inference, the basic problem is the estimation of the parameters of multivariate distribution by maximum likelihood method. The log-likelihood function is given as:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log c(F_1(x_{i1}),..., F_m(x_{im})) + \sum_{i=1}^{n}\sum_{j=1}^{m} \log(f_j(x_{ij})). \tag{8}$$

One of the basic estimation algorithms is performed in two steps:

1. First step is maximum likelihood estimation of the parameters of the marginal distributions (for each j), through the maximization of the following function:

$$l(\boldsymbol{\theta}_j) = \sum_{i=1}^{n} \log(f_j(x_{ij})). \tag{9}$$

2. Second step is maximum likelihood estimation of the parameters of copula function (given estimates obtained in the first step), through the maximization of the following function:

$$l(\alpha) = \sum_{i=1}^{n} \log c(F_1(x_{i1}),..., F_m(x_{im})). \tag{10}$$

Of course, the particular solution of maximum likelihood estimation depends on copula density function.

2. Model based clustering

One of the most common approaches used in clustering is the so-called model based clustering. In this approach a particular model is specified for each cluster. In the descriptive case, model is given as certain data-analytic characteristic of the cluster, for example location vector and scatter matrix, regression, principal component, etc. Then particular objective goodness-of-fit function is minimized, like in

the formula (1) or (2). We can get for example some extended version of k-means method by solving the following conditional minimization problem

$$\sum_{j=1}^{K}\sum_{i\in C_j}(\mathbf{x}_i-\mathbf{v}_j)^T\mathbf{M}_j(\mathbf{x}_i-\mathbf{v}_j)\to\min, \tag{11}$$

$$\left|\mathbf{M}_j\right|=b_j,\qquad j=1,\ldots,m. \tag{12}$$

Here we minimize the sum of Mahalanobis distances from the cluster centers, under condition of fixed volume of each cluster.

Now we will consider model based clustering in stochastic approach.

Here the basic assumption is that multivariate data can be considered as a sample drawn from a population consisting of a number of classes (subpopulations), denoted by K, and a particular multivariate distribution is a model for each class. There are two common approaches in such stochastic model based clustering:

– classification likelihood approach (e.g. [Scott, Symons 1971]);

– mixture approach (e.g. [Wolfe 1970]).

In the classification likelihood approach the likelihood function for n observations is given as:

$$L(\theta|\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)=\prod_{i=1}^{n}f(\mathbf{x}_i|\boldsymbol{\theta})=\prod_{i=1}^{n}f_{\gamma_i}(\mathbf{x}_i|\boldsymbol{\theta}_{\gamma_i}),$$
$$\gamma_i=j\Leftrightarrow\mathbf{x}_i\in\Pi_j. \tag{13}$$

Assuming that the number of parameters for each class is equal to s, we get the total number of parameters to be estimated equal to $Ks+n$. The estimation of parameters is performed by iterative algorithm where for given assignment of observations to classes the parameters are estimated and then the assignment (classification) is updated.

In the mixture approach the likelihood function for n observations is given as:

$$L(\boldsymbol{\theta}|\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)=\prod_{i=1}^{n}f(\mathbf{x}_i|\boldsymbol{\theta})=\prod_{i=1}^{n}\left(\sum_{j=1}^{K}P_jf_j(\mathbf{x}_i|\boldsymbol{\theta}_j)\right). \tag{14}$$

Assuming that the number of parameters for each class is equal to s, we get the total number of parameters to be estimated equal to $Ks+K-1$. It can be proved that maximum likelihood estimates of prior probabilities, class parameters and posterior probabilities in a mixture approach can be obtained through the following equations (after taking derivatives of log-likelihood function):

$$\hat{P}_j=\frac{1}{n}\sum_{i=1}^{n}\hat{p}(j|\mathbf{x}_i), \tag{15}$$

$$\sum_{i=1}^{n}\hat{p}(j|\mathbf{x}_i)\nabla\hat{\boldsymbol{\theta}}_j[\log f_j(\mathbf{x}_i|\hat{\boldsymbol{\theta}}_j)]=0, \tag{16}$$

$$\hat{p}(j|x_i) = \frac{\hat{P}_j f_j(\mathbf{x}_i|\hat{\boldsymbol{\theta}}_j)}{\sum_{l=1}^{K} \hat{P}_l f_l(\mathbf{x}_i|\hat{\boldsymbol{\theta}}_l)}. \tag{17}$$

The estimation of parameters is performed by iterative algorithm where for given posterior probabilities the estimation of parameters (prior probabilities and class parameters) is performed and then posterior probabilities are updated.

The particular models for classes (clusters) depend on the choice of multivariate distribution. As one can expect, the most popular are models where multivariate normal distribution is assumed. Banfield and Raftery [Banfield, Raftery 1993] showed for the classification likelihood approach that some well-known deterministic and stochastic criteria for clustering can be derived from multivariate normal model. Of course, this model is suitable for the cluster of elliptical shape (generally: hyperellipsoidal shape).

Now we move to the proposal to apply copula function in model based clustering. This will be generalization of approach proposed by Banfield and Raftery. The idea is rather simple, since the model for each cluster is given through multivariate distribution function decomposed according to Sklar theorem. Therefore for each cluster we have the following distribution function and density function:

$$F_j(x_1,...,x_m) = C_j(F_{j1}(x_1),...,F_{jm}(x_m)), \tag{18}$$

$$f_j(x_1,...,x_m) = c_j(F_{j1}(x_1),...,F_{jm}(x_m)) \cdot f_{j1}(x_1) \cdot ... \cdot f_{jm}(x_m). \tag{19}$$

As one can see, the model for each cluster "consists of" separate models for each component of a random vector and model for the dependence between these components.

By introducing copula model given in (19) into classification likelihood approach (formula (13)) and mixture approach (formula (14)) we get new proposals for these two approaches of model based clustering. Of course, the particular model depends on the choice of copula function. In any case, to estimate parameters of the models one should apply iterative algorithm. Now we will present such algorithms for both, classification likelihood approach and mixture approach.

3. Algorithm – classification likelihood approach

1. Start from initial classification.
2. In each iteration:
   - estimate parameters of the distribution in each class using (19) by two step estimation (parameters of marginal distributions, parameters of dependence);
   - calculate density for each observation given each class;

– update classification by assigning each observation to the class of highest density.

3. Iterate until classification does not change.

Algorithm – mixture approach

1. Start from initial posterior probabilities
2. In each iteration:

– estimate parameters of the distribution for each class using (16) and (19) by two step estimation (parameters of marginal distributions, parameters of dependence) and estimate prior probabilities using (15);

– calculate new posterior probabilities using (17).

3. Iterate until posterior probabilities do not change significantly.

Of course, the studies should be performed as far as the performance of the proposed methods and algorithms is concerned.

References

Banfield J.D, Raftery A.E. (1993), *Model-Based Gaussian and Non-Gaussian Clustering*, „Biometrics" 49, s. 803-828.

Scott A.J., Symons M.J. (1971), *Clustering Methods Based on Likelihood Ratio Criteria*, „Biometrics" 27, s. 387-397.

Sklar A. (1959), *Fonctions de repartition à n dimensions et leurs marges*, Publications de l'Institut de Statistique de l'Université de Paris, 8, s. 229-231.

Tukey J.W. (1962), *The Future of Data Analysis*, „Annals of Mathematical Statistics" 33, s. 1-67.

Wolfe J.H. (1970), *Pattern Clustering by Multivariate Mixture Analysis*, „Multivariate Behavioral Research" 5, s. 329-350.

OD ROZKŁADÓW WIELOWYMIAROWYCH DO ANALIZY DANYCH – KLASYFIKACJA OPARTA NA MODELU

Streszczenie

W artykule przedstawione są pewne nowe propozycje metod klasyfikacji mającej u podstaw model dla klasy. Są to uogólnienia klasycznego podejścia Banfielda i Raftery'ego z 1993 r., uzyskane na skutek wprowadzenia funkcji połączeń i podania algorytmów w podejściu klasyfikacyjnym i mieszankowym.