Iwona Bąk

Influence of feature selection methods on classification sensitivity based on the example of a study of Polish voivodship tourist attractiveness

Folia Oeconomica Stetinensia 13(21)/2, 134-145

2013

VERSITA

Folia Oeconomica Stetinensia
DOI: 10.2478/foli-2013-0017

Wydział Nauk Ekonomicznych i Zarządzania
Uniwersytetu Szczecińskiego

INFLUENCE OF FEATURE SELECTION METHODS ON CLASSIFICATION SENSITIVITY BASED ON THE EXAMPLE OF A STUDY OF POLISH VOIVODSHIP TOURIST ATTRACTIVENESS

Iwona Bąk, Ph.D.

Department of Application of Mathematics in Economics
Faculty of Economics
West Pomeranian University of Technology, Szczecin
Janickiego 31, 71-101 Szczecin, Poland
e-mail: iwona.bak@zut.edu.pl

Received 21 October 2013, Accepted 22 December 2013

Abstract

The purpose of this article is to determine the influence of various methods of selection of diagnostic features on the sensitivity of classification. Three options of feature selection are presented: a parametric feature selection method with a sum (option I), a median of the correlation coefficients matrix column elements (option II) and the method of a reversed matrix (option III). Efficiency of the groupings was verified by the indicators of homogeneity, heterogeneity and the correctness of grouping. In the assessment of group efficiency the approach with the Weber median was used. The undertaken problem was illustrated with a research into the tourist attractiveness of voivodships in Poland in 2011.

Keywords: feature selection method, classification sensitivity, Weber median, tourist attractiveness.

JEL classification: C38, L83.

Introduction

At the very beginning each multi-dimensional comparative analysis it is necessary to define the objects of the comparison and a set of features which widely characterize the properties of these objects, i.e. the diagnostic features. The results of such research greatly depend on the correctness of such selection, regardless of the methods and techniques used during the further phase of the research[1]. The criteria for selection of the features can be divided into two groups: metaphorical and formally-statistical[2]. In the former approach, such elements are taken into the set of diagnostic features that are regarded as the most important characteristics of the compared objects in the light of the researcher's knowledge of the analysed phenomena. In the second approach the feature selection is made in the way of processing and analysing statistical information by means of adequate formal procedures[3]. The best is a two-phase feature selection procedure where both approaches are used simultaneously. The first step is to create an initial feature list on the basis of the researcher's own working hypothesis (as a result of their knowledge of the research subject and the widely understood theory of economy) and their collaboration with representatives of proper scientific fields (experts)[4]. In the second phase the list is being reduced through formal methods with regards to the statistical properties of the primary features.

The purpose of this article is to determine the influence of various methods of selecting diagnostic features on the classification efficiency. Three options of feature selection are presented: the parametrical feature selection method with a sum (option I), a median of the correlation coefficients matrix column elements (option II) and the method of a reversed matrix (option III). The linear assignment of voivodships and defining typological groups of objects was conducted by means of a method based on the Weber median vector. The efficiency of the groupings was verified with the indicators of homogeneity, heterogeneity and focus points correctness, where the role of the gravity centers was played by the Weber median. The undertaken problem was illustrated by a research on tourist attractiveness of voivodships in Poland in 2011.

1. Research materials and methods

Initially, 26 diagnostic features were proposed for the research, characterizing the tourist attractiveness, which consists of: environmental values, the level of tourist development, transport accessibility and the level of environmental pollution[5]. The National Statistical Office's

data, which were made available at the Local Data Bank, were used in the research (www.stat.gov.pl). For the analysis, the following set of diagnostic features was used:

X_1 – forestation rate (in %),
X_2 – the share of legally protected land in the whole area (in %),
X_3 – the length of hard surface roads in km per 10 thousand people,
X_4 – the number of people per 1 post office,
X_5 – the number of main telephone lines 1000 people,
X_6 – the number of apartments in thousands per 1000 people,
X_7 – the number of shops per 1000 people,
X_8 – the number of gas stations per 1000 people,
X_9 – the general number of permanent marketplaces per 1000 people,
X_{10} – the number of subjects entered into the REGON registry per 10 thousand people,
X_{11} – the number of people per one hospital bed in general hospitals,
X_{12} – the number of people per 1 generally accessible pharmacy,
X_{13} – the number of people per 1 library,
X_{14} – the number of books in libraries per 1000 people,
X_{15} – the number of people per 1 seat in permanent cinemas,
X_{16} – the number of museums, including their departments, per 1000 people,
X_{17} – the number of people per 1 seat in theatres and musical institutions,
X_{18} – the number of tourist mass accommodation centers per 1000 people,
X_{19} – accommodation places in tourist sites per 1000 people,
X_{20} – financial investments per fixed assets used for environmental protection per 1 inhabitant,
X_{21} – the number of people using water treatment plants in % of the general population,
X_{22} – emission of gas air pollutants in general per 1 km^2,
X_{23} – emission of dust air pollutants in general per 1 km^2,
X_{24} – suppressed or neutralized gas pollutants in devices for pollution reduction in % of produced pollution
X_{25} – waste produced per 1 km^2,
X_{26} – industrial and communal wastewater treated in % of the wastewater needing treatment.

After defining and gathering data concerning the initial set of features, proper verification actions are usually performed against two most important criteria[6]:

1. Variability– the features should be diverse, i.e. effectively discriminating the objects. To assess the variability, a diversity coefficient, calculated from the formula, is used:

$$V_j = \frac{S_j}{\bar{x}_j} \tag{1}$$

where: $\bar{x}_j$ – arithmetic mean of X_j, S_j value – standard deviation of j^{th} feature, $j = 1, 2, \ldots, m$, m – feature count.

2. Correlation – two strongly correlated features carry similar information; therefore one of them is redundant. For this reason, the correlation indicators of all the features should be taken into account, and then, the most suitable verification method should be applied to eliminate features most similar to others. The starting point here is to create a matrix of feature correlations:

$$R = \begin{bmatrix} 1 & r_{12} & \ldots & r_{1m} \\ r_{21} & 1 & \ldots & r_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ r_{m1} & r_{m2} & \ldots & 1 \end{bmatrix} \tag{2}$$

where r_{jk} – the Person linear correlation coefficient of the j^{th} and k^{th} feature.

One of the most commonly used in practice discrimination methods of features based on a correlation coefficients matrix is a parametric method, as proposed by Z. Hellwig[7]. However, this method has two essential drawbacks[8]:

- it is sensitive to values, that stand out, which means, that a high correlation coefficient can be, to a large degree, a result of its correlations with just one feature,
- it only accounts for direct links of a feature to other features, while it does not include indirect links.

To increase the immunity of this method results to values that stand out, the sum of the **R** matrix first column (row) elements can be replaced with their median in the first step. The second drawback can be eliminated by using the inverse matrix method. It involves creating an inverse matrix of the **R** matrix, as follows:

$$R^{-1} = \begin{bmatrix} \widetilde{r}_{11} & \widetilde{r}_{12} & \dots & \widetilde{r}_{1m} \\ \widetilde{r}_{21} & \widetilde{r}_{22} & \dots & \widetilde{r}_{2m} \\ \dots & \dots & \dots & \dots \\ \widetilde{r}_{m1} & \widetilde{r}_{m2} & \dots & \widetilde{r}_{mn} \end{bmatrix}, \tag{3}$$

where:

$$\widetilde{r}_{jk} = \frac{(1)^{j+k} \det(R_{kj})}{\det(R)} \tag{4}$$

$\det(\mathbf{R})$ – determinant of the matrix $\mathbf{R}$, $\mathbf{R}_{kj}$ – indicates a matrix created from the matrix after removing from it the j^{th} row and k^{th} column ($j, k = 1, 2, \dots, m$).

The diagonal elements of the $\mathbf{R}^{-1}$ matrix take up the values from the $[1, \infty)$ range. Those of them, which exceed the maximum set level $\widetilde{r}_0$ (often it is set to $\widetilde{r}_0 = 10$) indicate a faulty numerical conditioning of the **R** matrix. Such features, for which $\left|\widetilde{r}_{jj}\right| > \widetilde{r}_0$ should thus be eliminated.

The linear assignment of Polish voivodships and defining typological groups of objects was conducted using the method based on the Weber median vector[9]. The positional option of the linear object assignment takes a different standardization formula, compared to the classical approach, based on a quotient of the feature value deviation from the proper coordinate of the Weber median and a weighed absolute median deviation, using the Weber median[10]:

$$z_{ij} = \frac{x_{ij} - \theta_{0j}}{1{,}4826 \cdot m\widetilde{a}d(X_j)}, \tag{5}$$

where: $\theta_0 = (\theta_{01}, \theta_{02}, ..., \theta_{0m})$ is the Weber median, $m\widetilde{a}d(X_j)$ is the absolute median deviation, in which the distance from the features to the Weber vector is measured, i.e.: $m\widetilde{a}d(X_j) = \underset{i=1,2,...,n}{med} \left|x_{ij} - \theta_{0j}\right|$ ($j = 1, 2, \dots, m$). The aggregate measure is calculated with the formula:

$$\mu_i = 1 - \frac{d_i}{d_-}, \tag{6}$$

where: $d_- = med(\boldsymbol{d}) + 2{,}5\ mad(\boldsymbol{d})$, where $d = (d_1, d_2, \dots, d_n)$ is a distance vector calculated with the formula: $d_i = \underset{j=1,2,...m}{med} \left|z_{ij} - \varphi_j\right|$ $i = 1, 2, \dots, n$, $\phi_j = \max_{i=1,2,...,n} z_{ij}$ – the coordinated of the development pattern vector, which constitute of the maximum values of the normalized features.

The assignment of objects with a positioning measure is the basis for a division of objects into four classes. The most commonly used grouping method in the positioning scope

is called the *three medians method*. It involves indicating a median of vector coordinates $\mu = (\mu_1, \mu_2, ..., \mu_n)$, which is denoted $med(\mu)$, then dividing the population of objects into two groups: those, for which the measure values exceed the median and are higher than it. Next the indirect medians are defined as: $med_k(\mu) = \underset{i:\Gamma_i \in \Omega_k}{med}(\mu_i)$, where k = 1,2.

This way the following groups of objects are created:

– Group I: $\mu_i > med_1(\mu)$,

– Group II: $med(\mu) < \mu_i \leq med_1(\mu)$,

– Group III: $med_2(\mu) < \mu_i \leq med(\mu)$,

– Group IV: $\mu_i \leq med_2(\mu)$.

The last stage of the taxonomic analysis is to check the quality of objects grouping. The methods of grouping lead to such a classification of objects into groups, where the objects belonging to the same group are most similar to each other (as high homogeneity of object groups as possible), and the objects belonging to different groups should be as different as possible (heterogeneous). To assess the quality of classification the measures of homogeneity and heterogeneity of groups are used, involving the concept of a group gravity centre and the distance from it. In this work an approach was taken, where the center of gravity of a group was replaced with a Weber median of its elements. In the homogeneity assessment of the formed groups the following measure was used[11]:

$$hm_6^* mx = \max_{k=1,2,..p} hm_6^*(P_k) \tag{7}$$

where:

$hm_6^*(P_k) = \underset{i:\Gamma_i \in P_k}{med} \delta(\Gamma_i, \Gamma_{0k})$ – median of P_k group objects distances from its Weber's median vector,

$\Gamma_{0k} = (\theta_{1P_k}, \theta_{2P_k}, ..., \theta_{mP_k})$ – Weber's median vector calculated for the P_k group, k = 1, 2, …, p,

p – number of focus points obtained at a certain level of group formation.

In the heterogeneity assessment the following measure was used:

$$ht_6^* mn = \min_{k=1,2,..p} ht_6^*(P_k) \tag{8}$$

where $ht_6^*(P_k) = \underset{\substack{i=1,...,p \\ i \neq k}}{med} \delta(\Gamma_i, \Gamma_{0k})$ – a median of distances between the Weber median of a group and analogical vectors for other groups.

In the assessment of group formation correctness a complex measure was used, in the following form:

$$ct_6^* = \frac{hm_6^* mx}{ht_6^* mn}. \tag{9}$$

2. Research results

In the first step, where the features were chosen for a taxonomic study a discrimination criterion was set, expressed with a variability coefficient. Those following features, for which the variability coefficient did not exceed 10%, were excluded from the research: X_6, X_7, X_{11}, X_{12}, X_{26}.

In the next step a reduction of potential diagnostic features was made, according to three options. The first two options involved the Hellwig parametric method: with a sum (option I) and median of correlation coefficients matrix column elements (option II), the third option concerns the reverse matrix method. Hereby the following sets of diagnostic features were distinguished:

for option I: X_1, X_2, X_3, X_5, X_{15}, X_{16}, X_{17}, X_{19}, X_{20}, X_{21}, X_{22};

for option II: X_1, X_2, X_5, X_{14}, X_{16}, X_{20}, X_{21}, X_{22}, X_{24};

for option III: X_1, X_2, X_4, X_5, X_{10}, X_{14}, X_{16}, X_{20}, X_{24}, X_{25}.

A classification of voivodships was made using the obtained sets of diagnostic features by determining for this purpose the positioning taxonomic measures based on Weber's median. The results are presented in Table 1.

Table 1. The ranking of Poland's Voivodships in 2011in the scope of their tourist attractiveness

Voivodship	Option I		Option II		Option III	
	measure value	deposit	measure value	deposit	measure value	deposit
1	2	3	4	5	6	7
Dolnośląskie	0.0897	13	0.0670	14	0.0099	15
Kujawsko-pomorskie	0.2458	8	0.2270	8	0.2424	8
Lubelskie	–0.0813	16	0.0386	15	0.0171	14
Lubuskie	0.3836	2	0.3683	1	0.4610	1
Łódzkie	0.2301	9	0.2109	9	0.2179	10
Małopolskie	0.1184	12	0.0965	13	0.1426	12
Mazowieckie	0.1643	10	0.1435	11	0.2364	9
Opolskie	0.1627	11	0.1418	12	0.2674	7
Podkarpackie	0.2669	7	0.3440	2	0.3038	5
Podlaskie	0.3434	3	0.3271	3	0.1447	11
Pomorskie	0.3214	5	0.3045	6	0.3161	4
Śląskie	0.0614	14	0.0380	16	–0.0209	16

1	2	3	4	5	6	7
Świętokrzyskie	0.3228	4	0.3059	5	0.2827	6
Warmińsko-mazurskie	0.0530	15	0.1450	10	0.0636	13
Wielkopolskie	0.2736	6	0.2731	7	0.3661	2
Zachodniopomorskie	0.3913	1	0.3072	4	0.3380	3

Source: own calculations.

As Table 1 shows, the alignments of voivodships using the aforementioned options of feature selection are not uniform and in some cases they vary significantly. To determine if the tested objects are aligned in a compatible way Spearman ranks correlation coefficients were calculated (Table 2). These coefficients take values within the [–1.1] range. The closer their value is to 1 or –1, the stronger the studied relation is[12].

Table 2. Spearman ranks correlation coefficients calculated for the ranks of Voivodships according to the taxonomic development measures obtained from the three options of feature selection

Options	I	II	III
I	1.0000	0.8941	0.8118
II	0.8941	1.0000	0.7853
III	0.8118	0.7853	1.0000

Source: own calculations.

High coefficient values indicate a good compatibility of voivodships linear alignment, regardless of the variances in the positions of some voivodships, e.g. Podkarpackie Voivodship in the option I alignment is ranked 7th, in option II it is ranked 2nd. Quite significant differences can be noticed in case of such voivodships as: Podlaskie (option I and II – position 3, option III – position 11), Wielkopolskie (option I – position 6, option III – position 2) and Opolskie (option I – position 11, option III – position 7). Only the Kujawsko-Pomorskie Voivoship has a constant position in all the rankings.

The taxonomic development measures replace the description of studied objects containing many features with one aggregate value. Aside the object alignment, it also allows dividing them into groups of a similar development level. Using the three median method, the set of voivodships was divided into four groups, containing objects similar in the scope of studied criterion – the tourist attractiveness (Table 3).

Table 3. Results of the voivodship grouping according to their tourist attractiveness

Groups	Option I	Option II	Option III
Group I	Zachodniopomorskie, Lubuskie, Podlaskie, Świętokrzyskie	Lubuskie, Podkarpackie, Podlaskie, Zachodniopomorskie	Lubuskie, Wielkopolskie, Zachodniopomorskie, Pomorskie
Group II	Pomorskie, Wielkopolskie, Podkarpackie, Kujawsko--pomorskie	Świętokrzyskie, Pomorskie, Wielkopolskie, Kujawsko--pomorskie	Podkarpackie, Świętokrzyskie, Opolskie, Kujawsko-pomorskie
Group III	Łódzkie, Mazowieckie, Opolskie, Małopolskie	Łódzkie, Warmińsko-mazurskie, Mazowieckie, Opolskie	Mazowieckie, Łódzkie, Podlaskie, Małopolskie
Group IV	Dolnośląskie, Śląskie, Warmińsko-mazurskie, Lubelskie	Małopolskie, Dolnośląskie, Lubelskie, Śląskie	Warmińsko-mazurskie, Lubelskie, Dolnośląskie, Śląskie

Source: own calculations.

The obtained groups varied from each other in terms of voivodships belonging to them, regardless the fact, that the contents of some of the classes were partially the same. Generally, the difference between two different classifications of the same object did not exceed one focus point, but sometimes bigger variations occured. For example, according to the first and second option, the Podlaskie Voivodship belonged to the focus point I, while the third option classified it into Group III. To determine the efficiency of the obtained groups, they underwent verification by determining the homogeneity, heterogeneity and correctness indicators (Table 4).

Table 4. Assessment measures of object grouping correctness

Indicators	Option I	Option II	Option III
Homogeneity of groups	**12.3919**	13.3523	33.2724
Heterogeneity of groups	7.6069	9.1433	**35.6634**
Correctness of groups	1.6290	1.4603	**0.9330**

Source: own calculations.

While analyzing the results concerning the sensitivity of grouping, it can be concluded that as far as homogeneity of groups is concerned, the best result was obtained for option I. Nevertheless, the classification based on the set of features received from the reverse matrix method (option III) yielded much better results in the scope of heterogeneity and correctness of grouping.

In the classification conducted according to the third option, the following voivodships are members of the best, first group: Lubuskie, Wielkopolskie, Zachodniopomorskie and Pomorskie. They present favorable values of the means, in comparison to the general means, concerning the following features: forestation rate, the number of main telephone lines per 1000 people,

number of subjects entered into the Regon registry per 10 thousand people, waste produced per 1 km^2.

A good situation in the second group of Voivodships is identified in case of such feature mean values as: share of legally protected land in the whole area (in %), number of books in libraries per 1000 people, financial investments per fixed assets used for environmental protection per 1 inhabitant, waste produced per 1 km^2. The disadvantaging values include the means concerning the number of main telephone lines per 1000 people and the number of entities entered into the Regon registry per 10 thousand people.

The low tourist attractiveness of the voivodships belonging to the third group results mostly from their low forestation rate and the little amount of suppressed or neutralized gas pollutants in devices for pollution reduction in % of produced. A positive influence on the matter in study is exerted by: the number of people per 1 post office and the number of museums, including their departments, per 1000 people.

The fourth group consists of voivodships, for which most of the features take negative values in comparison to the means from the entire country. A negative influence on the tourist attractiveness in this class comes from low mean values related to: the share of legally protected land in the whole area in general, the number of books in libraries per 1000 people, the number of entities entered into the Regon registry per 10 thousand people, the number of museums, including their departments, per 1000 people, financial investments per fixed assets used for environmental protection per 1 inhabitant. Moreover, in this class the largest amount of waste produced per 1 km^2 was observed.

Conclusions

This paper presents an attempt to answer the question regarding the influence of different methods of diagnostic features selection on the sensitivity of classification. In this research, three selection methods were used: two options of a parametric method (with a sum and median of correlation coefficients matrix column elements) and the reverse matrix method. The created sets of diagnostic features were used for the classification of Polish voivodships according to their tourist attractiveness. The obtained ratings varied between each other, not many voivodships had similar positions in the ratings and only one object (Kujawsko-Pomorskie Voivodship) was on the same position in all the three ratings. On the basis of the obtained rankings typological groups of voivodships were created. In each classification, four groups were created and the sensitivity of the obtained divisions was studied on the basis of the indicators of homogeneity,

heterogeneity and grouping correctness, where the role of gravity centers was played by the Weber median. The indicators defining the grouping quality indicate that a group using a set of features obtained through the method of a reverse correlation coefficients matrix gave better results in the scope of heterogeneity and grouping correctness. However, as far as homogeneity of groups is concerned, the best result was obtained for the first option, i.e. construction of a taxonomic development measure, on the basis of a set of features formed with a parametric method of feature selection, with a sum of correlation coefficients matrix column elements.

Summarizing, it can be concluded that application of taxonomic development measures based on different diagnostic features selection methods provides non-identical results in the ranking and grouping of objects in question.

Notes

[1] Panek (2009), p. 16.

[2] Grabiński (1992), p. 43.

[3] Nowak (1990), p. 23.

[4] Gatnar, Walesiak (2004), p. 320.

[5] Rapacz (2004), p. 57.

[6] Młodak (2006), pp. 28–32.

[7] Nowak (1990), pp. 28–30; Panek (2009), pp. 21–22.

[8] Młodak (2006), p. 31.

[9] Weber's median is a multi-dimensional generalization of the classical concept of the median. This vector minimizes the sum of Euclidean distances from the data points representing the considered objects, so is a kind of "middle" one, but it is also immune to the presence of outliers (Młodak 2006).

[10] Młodak (2006), pp. 136–137.

[11] Ibidem, pp. 138–141.

[12] Hozer (1998), p. 224; Luszniewicz, Słaby (2003), p. 291.

References

Gatnar, E. & Walesiak, M. (Eds.), (2004). *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wrocław: Wydawnictwo Akademii Ekonomicznej we Wrocławiu.

Grabiński, T. (1992). *Metody taksonometrii*. Kraków: Akademia Ekonomiczna w Krakowie.

Hozer, J. (Ed.) (1998). *Statystyka. Opis Statystyczny*. Szczecin: Katedra Ekonometrii i Statystyki, Uniwersytet Szczeciński, Stowarzyszenie Pomoc i Rozwój.

Luszniewicz, A. & Słaby, T. (2003). *Statystyka z pakietem komputerowym STATISTICA PL. Teoria i zastosowania*. Warszawa: Wydawnictwo C.H. Beck.

Młodak, A. (2006). *Analiza taksonomiczna w statystyce regionalnej*. Warszawa: Difin.

Nowak, E. (1990). *Metody taksonomiczne w klasyfikacji obiektów społeczno-gospodarczych*. Warszawa: PWE.

Panek, T. (2009). *Statystyczne metody wielowymiarowej analizy porównawczej*. Warszawa: Oficyna Wydawnicza SGGW.

Rapacz, A. (Ed.) (2012). *Wyzwania współczesnej polityki turystycznej. Problemy funkcjonowania rynku turystycznego*. Prace Naukowe AE we Wrocławiu nr 258, Wrocław: Wydawnictwo UE we Wrocławiu.