Comparison of Classical Dimensionality Reduction Methods with Novel Approach Based on Formal Concept Analysis

Eduard Bartl[1], Hana Rezankova[2], and Lukas Sobisek[2]

[1] Department of Computer Science, Faculty of Science,
Palacky Univeristy, Olomouc, Czech Republic
eduard.bartl@upol.cz
[2] Department of Statistics and Probability,
University of Economics, Prague, Czech Republic
{hana.rezankova,lukas.sobisek}@vse.cz

Abstract. In the paper we deal with dimensionality reduction techniques for a dataset with discrete attributes. Dimensionality reduction is considered as one of the most important problems in data analysis. The main aim of our paper is to show advantages of a novel approach introduced and developed by Belohlavek and Vychodil in comparison of two classical dimensionality reduction methods which can be used for ordinal attributes (CATPCA and factor analysis). The novel technique is fundamentally different from existing ones since it is based on another kind of mathematical apparatus (namely, Galois connections, lattice theory, fuzzy logic). Therefore, this method is able to bring a new insight to examined data. The comparison is accompanied by analysis of two data sets which were obtained by questionnaire survey.

Keywords: dimensionality reduction, discrete data, factor analysis, formal concept analysis, fuzzy logic, matrix decomposition, principal component analysis.

1 Introduction

Nowadays, in many areas (such as engineering, computer science, biology or economics) we are facing a problem of efficient processing of the large datasets. Typical scenario is that we accomplish an experiment, questionnaire survey or some kind of test, and as a result we gain a large tabular dataset. The rows of such a table correspond to objects (e.g. respondents' answers or observations), while the columns correspond to examined attributes. Inside the table there are stored attribute values for all objects. We can also interpret the attributes as random variables taking on the values from certain domain.

The number of attributes that are examined on every object is called the dimensionality of the dataset. In many practical situations, the dimensionality of the dataset is very high. Dimensionality reduction methods are able to transform a high-dimensional space of attributes to a lower-dimensional space. The

J.T. Yao et al. (Eds): RSKT 2011, LNCS 6954, pp. 26–35, 2011.

problem of dimensionality reduction has been studied extensively in the past few decades and there are mainly two reasons for such a interest. First, despite of our increased computational capability, the high-dimensional data is hard to process efficiently. Moreover, dimensionality reduction techniques enables us to understand given data in easier way.

In the paper we focus on dimensionality reduction methods based on various types of matrix decompositions. We only consider methods which can be applied to datasets with discrete (i.e. non-continuous) attribute values. Examples of discrete attributes are binary attributes which take on two values (e.g. correct/incorrect, married/not-married), ordinal attributes which take on the values from the ordered set (e.g. "bad" $\leq$ "neutral" $\leq$ "good"), or nominal attributes whose values are represented by unordered categories (for instance, "mathematics", "physics", "history").

The basic methods for dimensionality reduction of the attribute value vectors characterizing the analyzed objects are principal component analysis (PCA) and factor analysis (FA). These methods suppose linear relationships between original quantitative attributes and transform the original vectors to the new ones characterized by new latent attributes. The aim of PCA is to find a real dimension of vectors. It goes from the covariance matrix. To reduce the dimensionality of the original matrix, this one is transformed to the new coordinate system by an orthogonal linear transformation.

For other types of attributes and relationships some other methods have been developed. Multidimensional scaling (MDS) is generalization of factor analysis. While factor analysis investigates relationships between attributes and is based on the correlation matrix, MDS can be based on any matrix which expressed relationships between either attributes or objects. For ordinal attributes, coefficients of rank correlation can be used. Non-metric MDS (NMMDS) is an alternative approach in which only the order of values are considered.

Another approach to dimensionality reduction in case of categorical attributes is their transformation to quantitative attributes. This is the basis of categorical principal component analysis (CATPCA) which can include nonlinear relationships between attributes.

Distinct assumptions are related with latent class (LC) models. We can mention LC Cluster models and LC DFactor models as examples. In the former, the model supposes a single nominal latent attribute with the number of categories equal to the number of attribute groups. In the latter, the model can contain more latent attributes (dichotomous or ordinal) called discrete factors. In both cases, response attributes (indicators) can be nominal, ordinal, continuous, and/or counts.

The main contribution of this paper is to show comparison of just mentioned classical methods with an approach which based on novel view how we can grasp the problem of matrix decomposition. Classical methods are briefly described in Section 2.1, while the explanation of the new approach is given in Section 2.2. Chapter 3 focuses on using all described methods to two real datasets obtained by questionnaire survey.

2 Dimensionality Reduction Methods

In this section we briefly describe two classical methods and a novel approach of dimensionality reduction.

2.1 Classical Methods

CATPCA. The CATPCA method transforms categorical attributes (both nominal and ordinal) to quantitative attributes by means of optimal scaling. This optimization leads to obtaining optimal principal components. The iterative process begins by assignment of a random value (object score) to each object. Let us denote the matrix of object scores by the symbol $\mathbf{X}$. Then the matrix $\mathbf{X}_w$ of weighted object scores ($\mathbf{X}_w = \mathbf{WX}$) is created under the following relationships:

$$\mu^T \mathbf{M}_* \mathbf{WX} = 0 \text{ and } \mathbf{X}^T \mathbf{M}_* \mathbf{WX} = n_w m_W \mathbf{I},$$

where μ is the vector of expected values, $\mathbf{M}_* = \sum_i \mathbf{M}_j$ ($\mathbf{M}_j$ is a diagonal matrix with the elements $m_{(j)ii}$ expressing weights v_j of individual attributes for each object; if the weight is not specified, then $v_j = 1$), $\mathbf{W}$ denotes a diagonal matrix with the elements w_i expressing the weights of individual objects (for the non-weighted objects $w_i = 1$), n_w is the sum of object weights, m_w is the sum of attribute weights.

FA. The factor analysis model is based on the correlation matrix. We can write it in the form $\mathbf{X} = \mu + \mathbf{\Gamma F} + \mathbf{E}$, where μ is the vector of expected values, $\mathbf{F}$ denotes the k-dimensional random vector of common factors F_i, $\mathbf{\Gamma}$ is the matrix of factor loadings (p x k), and $\mathbf{E}$ is the p-dimensional vector specific factors ε_i (p is the number of original attributes and k is the number of factors). One supposes that the following assumptions are satisfied: $E(F_i) = 0$, $E(\varepsilon_i) = 0$, $\mathrm{Cov}(\mathbf{F}, \mathbf{E}) = 0$, $\mathrm{Cov}(\mathbf{F}) = \mathbf{I}$, and $\mathrm{Cov}(\mathbf{E}) = \mathbf{\Psi}$ is a diagonal matrix.

The factors F_i are interpreted by means of correlation with original attributes. The correlation matrix can be written in the form $\mathbf{P}_{XF} = \mathbf{D}^{-\frac{1}{2}} + \mathbf{\Gamma}$, where $\mathbf{D}$ is a diagonal matrix with elements expressing the variance of the original attributes.

2.2 The Novel Method

The novel method of dimensionality reduction introduced in [4] can be characterized by the following points.

1. *The attributes take on the values from a bounded scale which is equipped with particular operations.* The meaning of these operations (i.e. the way how we compute with attribute values) is based on the theory of fuzzy logic in narrow sense (see e.g. [7]). The bounded scale of attribute values is called complete residuated lattice and it is often denoted by L. Binary operations defined on L are supremum $\vee$, infimum $\wedge$, multiplication $\otimes$ and its residuum $\rightarrow$ (multiplication and residuum are connected via adjointness property, for more

details see [2]). If we consider ordinal attributes with values from linearly ordered unit interval (i.e. $L = [0, 1]$), then supremum and infimum coincide with maximum and minimum, respectively, multiplication is left-continuous t-norm (e.g. usual product of real numbers), and residuum can be derived from the multiplication using adjointness property.

2. *Input dataset is interpreted as relational data.* We consider a tabular data, where X denotes the set of object, and Y denotes the set of attributes that take on values from complete residuated lattice L. In terms of fuzzy logic, I is a fuzzy relation between sets X and Y, i.e. I is a mapping $X \times Y \to L$. We consider fuzzy relations as a particular case of fuzzy sets (see [2,8,10]). Therefore, using standard fuzzy set notation we write $I \in L^{X \times Y}$. The value $I_{ij} \in L$ (in i-th row and j-th column of the matrix I) is degree to which i-th object has j-th attribute. To sum up, with a slight abuse of the notation, we identify the matrix I representing tabular data with fuzzy relation $I \in L^{X \times Y}$.
3. *The problem of dimensionality reduction is transformed to the problem of matrix decomposition.* Technically, for an $n \times m$ matrix I we try to find an $n \times k$ matrix A and a $k \times m$ matrix B such that $I = A \circ B$, where $\circ$ is a particular composition operator and the inner dimension k is as small as possible. Again, we identify matrix A with fuzzy relation $A \in L^{X \times K}$, and matrix B with fuzzy relation $B \in L^{K \times Y}$ (K is a set with k elements). The composition operator is defined as follows:

$$(A \circ B)_{ij} = \bigvee_{l=1}^{k} A_{il} \otimes B_{lj}.$$

   In practice, we usually do not need the exact factorization. Instead, it is sufficient to find an approximate decomposition $I \approx A \circ B$ which even makes the inner dimension smaller. Elements of the set K can be thought of as factors which are able to explain original data stored in I. This justifies our need to push the inner dimension k as much as possible. The meaning of factorizing matrices can be then described as follows: A_{il} is a degree to which l-th factor applies to i-th object, and B_{lj} is degree to which j-th attribute is a manifestation of l-th factor.
4. *The problem of finding factors is solved using a particular method of analysis of relational data called Formal Concept Analysis (FCA).* This technique was initiated by Wille in the paper [9]. The central notion in FCA is a formal concept inspired by Port-Royal logic. The formal concepts represent interesting clusters which can be found in the data. From the point of view of Port-Royal logic, the formal concept is a couple $\langle A, B \rangle$ consisting of an extent $A \in L^X$ (fuzzy set of objects covered by the concept) and an intent $B \in L^Y$ (fuzzy set of attributes covered by the concept). The extents can be mathematically described as fixpoints of a closure operator ${}^{\uparrow\downarrow} : L^X \to L^X$ consisting of two adjoint operators ${}^{\uparrow} : L^X \to L^Y$ and ${}^{\downarrow} : L^Y \to L^X$ (for more details, see [2,6]). Similarly, the intents are fixpoints of a closure operator ${}^{\downarrow\uparrow} : L^Y \to L^Y$. Set of all formal concepts is denoted by $\mathcal{B}(X, Y, I)$ and

together with subsethood ordering of extents (or, equivalently, intents) forms a complete lattice that is called concept lattice. In the end of this item, let us mention that formal concepts have a nice geometrical meaning, particularly, they form rectangular-shaped patterns in the input table (for more details, refer to [4,3]).

The core of the novel method is based on the idea that formal concepts play the role of factors. Namely, suppose a set $\mathcal{F} = \{\langle C_1, D_1\rangle, \ldots, \langle C_k, D_k\rangle\} \subseteq \mathcal{B}(X, Y, I)$ of several formal concepts. We denote by $A_\mathcal{F}$ an $n \times k$ matrix such that l-th column of this matrix coincides with vector C_l (extent of l-th formal concept from $\mathcal{F}$). Similarly, by $B_\mathcal{F}$ we denote an $k \times m$ matrix in which l-th row coincides with vector D_l (intent of l-th formal concept from $\mathcal{F}$). It has been shown in [4] that decomposition using formal concept is universal, i.e. for every I there exists a set $\mathcal{F} \subseteq \mathcal{B}(X, Y, I)$ of formal concepts such that $I = A_\mathcal{F} \circ B_\mathcal{F}$. In addition to that, formal concepts are optimal factors. Formally, if $I = A \circ B$ with inner dimension equal to k, then there exists a set $\mathcal{F} \subseteq \mathcal{B}(X, Y, I)$ of formal concepts such that $I = A_\mathcal{F} \circ B_\mathcal{F}$ and $|\mathcal{F}| \leq k$ (i.e. number of formal concepts, which serve as factors, is not greater than inner dimension k of the given decomposition $I = A \circ B$).

Using geometrical interpetation of formal concepts, the problem of finding $\mathcal{F} \subseteq \mathcal{B}(X, Y, I)$ such that $I = A_\mathcal{F} \circ B_\mathcal{F}$ can be reduced to the problem of finding the smallest set of formal concepts (rectangular-shaped patterns) covering all non-zero values in given tabular data (due to lack of space, we just refer to [3] for more information). If we need not the exact decomposition $I = A_\mathcal{F} \circ B_\mathcal{F}$ we can take only few formal concepts $\mathcal{F}' \subseteq \mathcal{F}$. In this case, we obtain approximate decomposition $I \approx A_{\mathcal{F}'} \circ B_{\mathcal{F}'}$, i.e. formal concepts from $\mathcal{F}'$ cover the non-zero values in given tabular data just partly.

3 Applications

3.1 Analyzed Dataset

For illustration, we analyzed two real datasets obtained on the basis of a questionnaire survey. They concern perception of the policemen in the Czech Republic by young people (survey from 2006, 356 respondents). The first dataset (named Typical policeman) includes 24 ordinal attributes characterizing a typical policeman and the second one (named Ideal policeman) includes the same number of attributes characterizing an ideal policeman. Respondents' answers are coded from 1 to 7 (the value 1 means the most positive level, the value 7 the most negative level). Four and five factors obtained by traditional factor analysis are characterized in [5].

3.2 Analyses Using CATPCA and FA

For the comparison of the novel approach with classical methods modified for ordinal attributes, we chose categorical principal component analysis (CATPCA) and factor analysis (FA) based on Kendall's coefficient of the rank correlation. We used the SPSS system for these analyses.

CATPCA. On the basis of our previous experiments results and for the reason of comparability of different methods, we realized the analysis for four dimensions. As a result we obtained for example the values of these dimensions, percentages of explained variability for each dimension, and component loadings graphs for selected dimensions.

These graphs for the second and fourth dimensions (for the reason of the best resolution of individual attributes) are displayed in Fig. 1. However each combination of dimensions gives a little distinct view on the relationships between variables. In this type of graph, the smaller angle means the greater similarity of attributes. However, we do not get any information on the level of the answers, if positive or negative features are predominant.

Results for dataset Typical policeman: Four dimensions explain almost 64% of variance. In Fig. 1 (left) we can identify some very similar attributes, e.g. y_5 and y_6 (hardness and power), y_1, y_2 and y_3 (attributes expressing ambitious level, fastness, and activity), or y_7 and y_9 (friendliness and kindness).

Results for dataset Ideal policeman: Four dimensions explain almost 60% of variance. Contrary of the previous case, we can see in Fig. 1 (right) that attributes y_2 and y_5 (fastness and hardness) are close. Further, attributes y_4, y_6 and y_8 (bravery, power and cleverness) are very similar. One pair is also created by attributes y_9 and y_{11} (kindness and fairness).

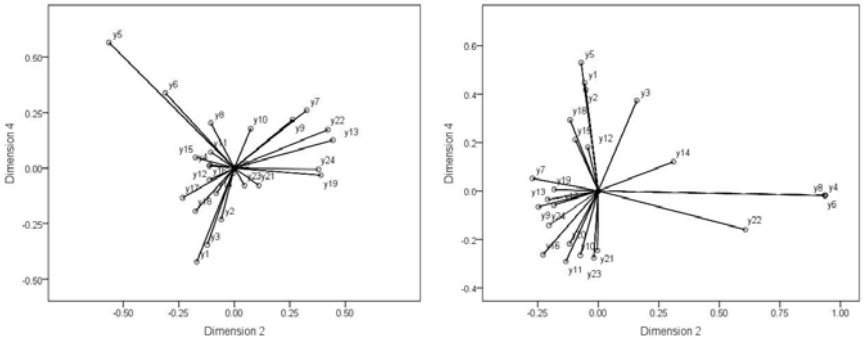

Fig. 1. CATPCA: dataset Typical policeman (left) and Ideal policeman (right)

FA. In this case we also realized the analysis for four factors. We applied the Varimax rotation. As a result we obtained for example the values of component loadings, percentages of explained variability for each factor, and component graphs for selected factors. These graphs for the second and fourth components are displayed in Fig. 2.

Results for dataset Typical policeman: Four factors explain more than 59% of variance. The relationships are less evident but some attributes are also close, e.g. y_7 and y_9 (friendliness and kindness). We can identify groups of variables according to quadrants. For example attributes y_{13}, y_{19}, y_{21}, y_{23} and y_{24} express moral features.

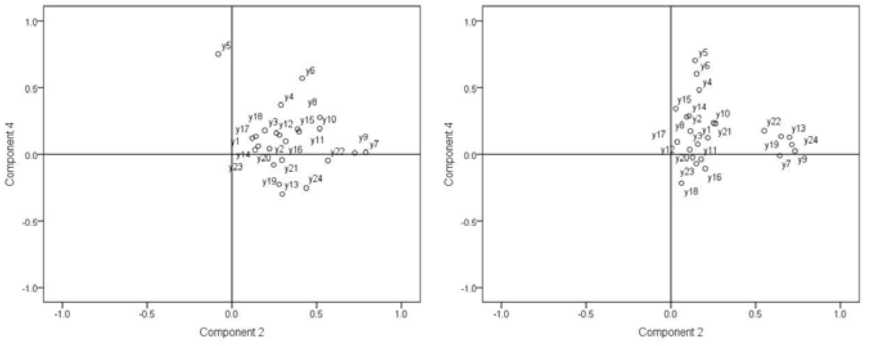

Fig. 2. FA: dataset Typical policeman (left) and Ideal policeman (right)

Results for dataset Ideal policeman: Four factors explain more than 50% of variance. In this case we can distinguish several groups of attributes. On one hand, we can see attributes y_4, y_5 and y_6 (bravery, hardness and power, i.e. features characteristic for a man) in the top part of the graph, on the other hand there is a separate group of attributes on the right side. There are attributes y_7, y_9, y_{13}, y_{19}, y_{22} and y_{24} which concerns moral features and human relationships.

3.3 Analysis Using the Novel Method

First of all, we need to choose a bounded scale of attribute values with appropriate operations. For the purpose of analysis of given datasets we use so-called 7-element Łukasiewicz chain, i.e. complete residuated lattice $L = \{0, \frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, 1\}$, where $\vee = \max$, $\wedge = \min$, $a \otimes b = \max(0, a+b-1)$, and $a \to b = \min(1, 1-a+b)$ for all $a, b \in L$.

Since attributes values and coded respondents' answers are different, we need to make a simple adjustment in the preprocessing stage: $I_{ij} = \frac{1}{6} \cdot (I'_{ij} - 1)$, where $I'_{ij} \in \{1, 2, \ldots, 7\}$ is j-th coded answer of i-th respondent, and $I_{ij} \in L$ is corresponding normalized attribute value. For instance, coded answer "3" of attribute "ambitious-lazy" with the meaning "rather ambitious" is adjusted to the attribute value $\frac{2}{6} \in L$.

Results for Dataset Typical Policemen. As an output of the algorithm based on the novel method (see [3]) we obtain a collection of factors $\mathcal{F}$, as described in Section 2.2. The typical behaviour of this algorithm for exact decomposition $I = A_{\mathcal{F}} \circ B_{\mathcal{F}}$ is that the number of factors $|\mathcal{F}|$ is greater than the number of all attributes $|Y|$. Particularly, in the case of dataset Typical policeman the algorithm computes 48 factors which explain input data precisely.

On the other hand, first few factors computed by the algorithm give us a very good approximation. Particularly, the first factor explains about 20% of the data, first 7 factors explain about 50% of the data, and first 17 factors explain about 75% of the data (in terms of denotation used in Section 2.2, $\mathcal{F}'$ consisting of

first 17 factors computed by the algorithm cover 75% of the input data, which means that the three quarters of the cells in matrices $A_{\mathcal{F}'} \circ B_{\mathcal{F}'}$ and I contain the same values). This phenomena directly relates to the fact that the algorithm firstly find the factors covering an input data in maximal way (i.e. the algorithm computes the factors according their importance).

We can lucidly depict every factor $F_i \in \mathcal{F}$ in terms of its extent and intent. Because both extent and intent are fuzzy sets, we can draw them using a graph (see [8], [10]). In our case, x-axis denotes objects or attributes. While y-axis denotes degree to which F_i applies to particular object, or degree to which a particular attribute is a manifestation of F_i. The first factor F_1 is shown in Fig. 3–4.

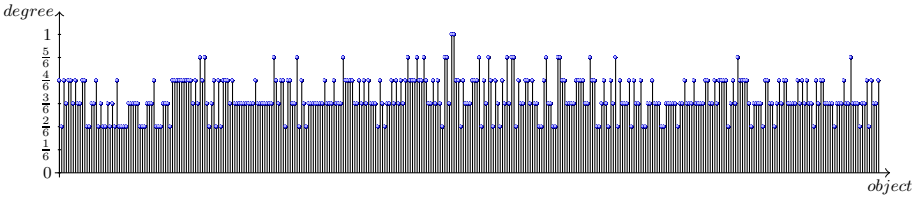

Fig. 3. Dataset Typical policeman: extent of the factor F_1

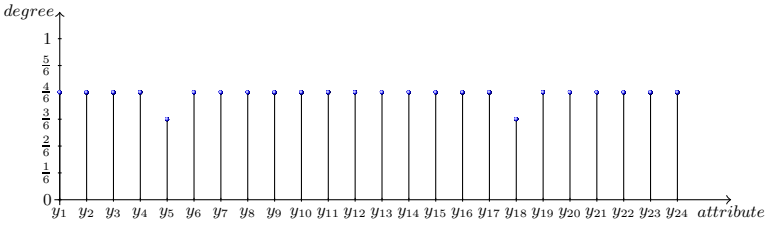

Fig. 4. Dataset Typical policeman: intent of the factor F_1

Now, we describe the first factor verbally. From Fig. 4 we can see that the degree to which every attribute is a manifestation of F_1 is rather high (in most cases $\frac{4}{6}$). So factor F_1 represents overall bad personality traits of a typical policeman: rather lazy, unfriendly, unfair, rude and so on. Moreover, Fig. 4 shows that F_1 applies to most of objects in high degree. In other words, many respondents characterized a typical policeman by this factor.

Results for Dataset Ideal Policeman. Similarly as in the case of the dataset Typical Policeman, 45 factors explain all the data. What is different is that only the first factor explains about 50% of the data. Furthermore, first 9 and 19 factors explain about 75% and 90% of the data, respectively. So compare to the dataset Typical policeman, we need very few factors to make considerably good approximation.

First factor F_1 is depicted in Fig. 5 and 6. Attributes y_5, y_7, y_9, y_{13}, y_{19} and y_{24} are manifestations of the first factor in very high degree (except the

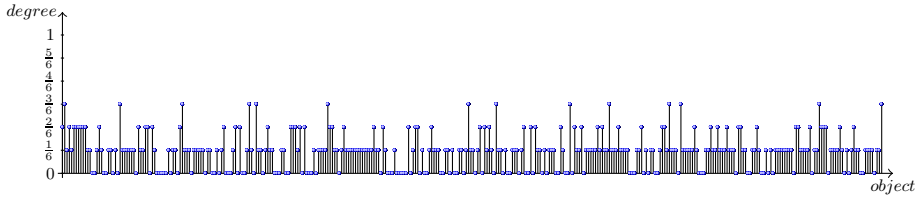

Fig. 5. Dataset Ideal policeman: extent of the factor F_1

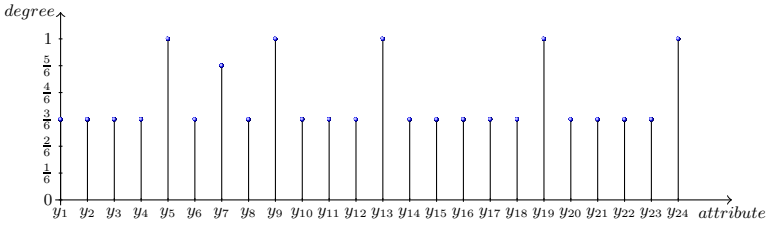

Fig. 6. Dataset Ideal policeman: intent of the factor F_1

attribute y_7, this degree is equal to 1). Using these attributes we can say that the first factor describes a policeman who is nonauthoritative (attribute y_5: hard-soft) and who has very bad communication skills (for instance, attributes y_7: friendly-unfriendly, y_9: kind-rude or y_{13}: peaceful-violent). Since this factor applies to all respondents in low degree (mostly in degree $\frac{1}{6}$), we can argue, that authoritativeness and communication skills are the most desired qualities of a policeman.

The interpretation of other factors for the datasets Typical policeman and Ideal policeman can be made in similar way, and it will be shown in extended version of this paper.

4 Conclusions and Future Work

In this paper we analyzed the datasets Typical policeman and Ideal policeman using categorical principal component analysis, factor analysis and using the new method based on formal concept analysis. All methods gives us the meaningful factors reducing the dimensionality of the input datasets. Since the factors in the novel method are extent-intent-based, this method is able to describe in what degree a particular factor is applicable to all respondents. Such feature can be viewed as one of the advantages of the novel method.

In terms of approximate decomposition, the new approach gives us two different results for both datasets. We need 7 factors in order to explain about 50% of dataset Typical policemen, but only 1 factor explaining 50% of the dataset Ideal policeman. This distinction leads us to the conclusion that the respondents have similar conception of the personality traits of an ideal policeman.

Future work will include analyzing other datasets using the novel method in order to obtain deeper insight to the practicability of this new approach.

Acknowledgement. Supported by Grant No. 202/10/0262 of the Czech Science Foundation.

References

1. Belohlavek, R.: Algorithms for fuzzy concept lattices. In: Proc. Fourth Int. Conf. on Recent Advances in Soft Computing, pp. 67–68 (2002)
2. Belohlavek, R.: Fuzzy Relational Systems: Foundations and Principles. Kluwer Academic/Plenum Publishers, New York (2002)
3. Belohlavek, R., Vychodil, V.: Factor analysis of incidence data via novel decomposition of matrices. In: Ferré, S., Rudolph, S. (eds.) ICFCA 2009. LNCS, vol. 5548, pp. 83–97. Springer, Heidelberg (2009)
4. Belohlavek, R., Vychodil, V.: On Boolean factor analysis with formal concepts as factors. In: Int. Conf. Soft Computing and Intelligent Systems & Int. Symposium on Intelligent Systems, pp. 1054–1059 (2006)
5. Moulisova, M.: Investigation of policeman perception. Kriminalistika 42(1), 56–71 (2009) (in Czech)
6. Ganter, B., Wille, R.: Formal concept analysis. Mathematical Foundations. Springer, Berlin (1999)
7. Hájek, P.: Metamathematics of Fuzzy Logic. Kluwer Academic, Dordrecht (1998)
8. Klir, G.J., Yuan, B.: Fuzzy Sets and Fuzzy Logic. Theory and Applications. Prentice-Hall, Englewood Cliffs (1995)
9. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival, I. (ed.) Ordered Sets, pp. 445–470 (1982)
10. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)