STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

FACTORIZATION METHODS OF BINARY, TRIADIC, REAL AND FUZZY DATA

CYNTHIA VERA GLODEANU

Abstract. We compare two methods regarding the factorization problem of binary, triadic, real and fuzzy data, namely Hierarchical Classes Analysis and the formal concept analytical approach to Factor Analysis. Both methods search for the smallest set of hidden variables, called factors, to reduce the dimensionality of the attribute space which describes the objects without losing any information. First, we show how the notions of Hierarchical Classes Analysis translate to Formal Concept Analysis and prove that the two approaches yield the same decomposition even though the methods are different. Finally, we give the generalisation of Hierarchical Classes Analysis to the fuzzy setting. The main aim is to connect the two fields as they produce the same results and we show how the two domains can benefit from one another.

1. Introduction and Problem Setting

In this article we compare two methods of factorization: Formal concept analytical approach to Factor Analysis presented in [3] and Hierarchical Classes Analysis introduced in [6]. Both methods were generalised to the factorization of triadic data. We have generalised the factorization through Formal Concept Analysis for the triadic case in [8]. The triadic version of Hierarchical Classes Analysis was introduced in [10] and an even more general case in [5]. As we will see in the following, for binary and triadic data, the two methods both use formal concepts as factors and yield the same results. We also compare the two approaches for real data sets. Unfortunately, there is no more a one-to-one correspondence between the two. The formal concept analytical approach uses fuzzy concepts and performs better than Real-Valued Hierarchical Classes Analysis. Therefore, it was promising to generalize the latter to the fuzzy setting, which we also present in this article.

Received by the editors: March 31, 2011.

2010 *Mathematics Subject Classification.* 62H25, 03G10.

1998 *CR Categories and Descriptors.* I.5.3 [**Pattern Recognition**]: Clustering – *Algorithms*; I.5.1 [**Pattern Recognition**]: Models – *Fuzzy set*.

Key words and phrases. Formal Concept Analysis, Hierarchical Classes Analysis, Factor Analysis.

2. Dyadic Factorization

Formal Concept Analysis [7] has as the underlying structure the notation of a *formal context* $\mathbb{K} = (G, M, I)$ consisting of two sets G (objects) and M (attributes) and a binary relation $I \subseteq G \times M$. Then $(g, m) \in I$ means that the object g has the attribute m. The relation I is called the *incidence relation* of the context. For $A \subseteq G$ and $B \subseteq M$ the *derivation operators* are defined as

$$\begin{aligned} A^{\shortmid} &:= \{m \in M \mid (g, m) \in I \text{ for all } g \in A\}, \\ B^{\shortmid} &:= \{g \in G \mid (g, m) \in I \text{ for all } m \in B\}. \end{aligned}$$

A *concept* of $\mathbb{K}$ is a tuple (A, B) with $A \subseteq G$ and $B \subseteq M$ such that $A^{\shortmid} = B$ and $B^{\shortmid} = A$. All the objects from A have all the attributes from B in common and the attributes from B apply to all the objects from A. As in Philosophy, the *extent* A contains the objects which fall under the concept's meaning and the *intent* B includes attributes which apply to all the object of the extent. Finite small contexts can be represented through cross tables. The rows of the table are named after the objects and the columns after the attributes. The row corresponding to the object g and the column corresponding to the attribute m contains a cross if and only if $(g, m) \in I$. Concepts ordered by the inclusion form complete lattices, see [7].

The formal concept analytical approach to Factor Analysis was presented in [3] and searches for the smallest subset of formal concepts which covers the incidence relation of the context. Working with binary matrices, a $p \times q$ binary matrix W is decomposed into the Boolean matrix product $P \circ Q$ of a $p \times n$ binary matrix P and an $n \times q$ binary matrix Q with n as small as possible. The Boolean matrix product $P \circ Q$ is defined as $(P \circ Q)_{ij} := \bigvee_{l=1}^{n} P_{il} \cdot Q_{lj}$, where $\bigvee$ denotes the maximum and $\cdot$ the product. The matrix P has as columns the characteristic vectors of the extents and the matrix Q has as rows the characteristic vectors of the intents from the concepts contained in the factorization. Then, the matrices W, P and Q represent an object-attribute, object-factor and factor-attribute relationship, respectively. That the factorization has indeed the smallest number of factors follows from the maximality of formal concepts, i.e., formal concepts correspond to maximal rectangles full of crosses in the cross table representation of a formal context.

Example 1. *Suppose we have a context with patients as objects and symptoms as attributes. Then, the factors would be the diseases the patients have. The matrix P associates each patient the disease he/she suffers from and the matrix Q associates each disease the symptoms it causes. Therefore, the factors have a verbal description and they can be potentially more fundamental than the original attributes.*

Hierarchical Classes Analysis was developed in [6] and it addresses the same factorization problem as discussed above with the same matrix product. However, the mathematical formalisation is slightly different. We give directly the translation of the notation into Formal Concept Analysis and just the definitions for objects. The ones for attributes can be done analogously. In a formal context (G, M, I) two objects $g_1, g_2 \in G$ are called *equivalent* iff $g_1^{\downarrow} = g_2^{\downarrow}$. The set $[g_1] := \{g \in G \mid g^{\downarrow} = g_1^{\downarrow}\}$ is called the *object class* of the object $g_1 \in G$. The object set which corresponds to an attribute class can be decomposed into object classes such that their size is maximal and their number minimal. These objects are then called *object bundles*. An *object* (*attribute*) *bundle* is the extent (intent) of some concept. In Hierarchical Classes Analysis the matrices P and Q contain the object and attribute bundles, respectively. We have presented the comparison for the dyadic case between these two approaches to the factorization problem in [9].

In [3] a greedy approximation algorithm was considered because the factorization problem is NP-hard, but can compute factorizations with up to 15 bundles.

3. Triadic Factorization

The triadic approach to Formal Concept Analysis was introduced by R. Wille and F. Lehmann in [11]. A *triadic context* is defined as a quadruple (K_1, K_2, K_3, Y) where K_1, K_2 and K_3 are sets and Y is a ternary relation between K_1, K_2 and K_3. The elements of K_1, K_2 and K_3 are called *(formal) objects, attributes* and *conditions*, respectively, and $(g, m, b) \in Y$ is read: *the object g has the attribute m under the condition b*. A *triconcept* of (K_1, K_2, K_3, Y) is a triple (A_1, A_2, A_3) with $A_i \subseteq K_i$ $(i \in \{1, 2, 3\})$ that is maximal with respect to component-wise set inclusion.

In [8] we have generalized the factorization problem presented in [3] to the triadic case.[1] We define a *triadic factorization* as the smallest set $\mathcal{F}$ of triconcepts such that they cover the ternary incidence relation Y of the triadic context. In [8] we have proved that every Boolean $3d$-matrix can be decomposed into the $3d$-product of three binary matrices and that by using triconcepts as factors we obtain the smallest possible factorization. The proofs are based on the fact that the triconcepts are maximal rectangular boxes in a triadic context.

[1] During the revision period of [8] it turned out there is yet unpublished work of R. Belohlavek and V. Vychodil dealing with the same subject [4].

The triadic version of Hierarchical Classes Analysis, called *Indclas*, was presented in [10]. The main difference to the dyadic version consists in working with three bundle matrices instead of two. Then, once again, every notation from Indclas can be translated into the language of Triadic Concept Analysis and the three bundles correspond to the intents, extents and modi, respectively.

4. Real and Fuzzy Factorization

In the last part we compare the factorization of real valued data through the formal concept analytical approach to Factor Analysis and Hierarchical Classes Analysis. The first, uses fuzzy concepts and the second one bundles, and an association matrix containing the real values. Because the fuzzy factorization yields fewer factors, we propose the generalisation of Hierarchical Classes Analysis to the fuzzy case.

There are many approaches to Fuzzy Formal Concept Analysis, however, we consider the method developed independently by Pollandt [13] and Belohlavek [1] as the standard one. A triple (G, M, I) is called a *fuzzy formal context* if $I : G \times M \to L$ is a fuzzy relation between the sets G and M and L is the support set of some residuated lattice. The fuzzy relation I assigns to each $g \in G$ and each $m \in M$ a truth degree $I(g, m) \in L$ to which the object g has the attribute m. A *fuzzy concept* is a tuple of the form $(A, B) \in L^G \times L^M$.

In [2] the formal concept analytical approach to Factor Analysis was generalised to the fuzzy setting. All the results from the dyadic case can be translated into the fuzzy case, i.e., a *fuzzy factorization* is the smallest subset of fuzzy concepts, such that they cover the fuzzy relation in the fuzzy context.

The disjunctive Hiclas-R model was presented in [12]. It implies the decomposition of a $p \times q$ matrix W with integer entries from $V = \{1, \ldots, v\}$ in a binary $p \times n_1$ *object bundle matrix* P, a binary $q \times n_2$ *attribute bundle matrix* Q and a rating-valued $n_1 \times n_2$ *core matrix* T which takes n_3 different non-zero values, where $n_3 \leq v$. The *equivalence relations* is defined analogously to the binary Hiclas model. The *association relation* is given by $W_{ij} = \bigvee_{h=1}^{n_1} \bigvee_{k=1}^{n_2} P_{ih} \cdot Q_{jk} \cdot T_{hk}$ for all $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, q\}$. Object i is associated with attribute j at the maximum value of association indicated by the core matrix T for the pair of bundles which contain object i and attribute j.

The core matrix also allows association of an object bundle with more attribute bundles. The association relation is not binary any more, it contains integer entries, which represent the value of association between an object and an attribute bundle. On the other hand, the fuzzy concepts contain the values of association in their membership values for each object and attribute.

Such a decomposition has a natural interpretation since the factors are fuzzy concepts. The factorization through fuzzy concepts is a more parsimonious method. First of all, because it does not require a third matrix, namely the core matrix. Second, the fuzzy approach yields in general a smaller number of factors than the bundle decomposition, due to the properties of the t-norm.

The factorization through fuzzy concepts is not possible in the setting of Hierarchical Classes Analysis, however weaker structures provide optimal solutions. We call (A, B) a *fuzzy preconcept* if and only if $A \subseteq B^{\shortmid}$ ($\Leftrightarrow B \subseteq A^{\shortmid}$, where $^{\shortmid}$ are the fuzzy derivation operators). The fuzzy preconcept (A, B) is called *fuzzy protoconcept* if and only if $(B^{\shortmid}, A^{\shortmid})$ is a fuzzy concept of (G, M, I). We will be searching for the smallest subset of fuzzy protoconcepts which covers the fuzzy relation in the fuzzy context. Due to the properties of the t-norms it is possible to choose the fuzzy protoconcept of a fuzzy concept such that they both yield the same maximal rectangle. With these remarks, we are able to generalize all the notation from Hierarchical Classes Analysis into the fuzzy setting.

Definition 1. *Let (G, M, I) be a fuzzy context and L the support set of some residuated lattice. Two fuzzy objects $g_1(a), g_2(b) \in G \times L$ are **equivalent** if and only if $g_1(a)^{\shortmid} = g_2(b)^{\shortmid}$. Equivalent objects form an **object class**. For two objects $g_1(a), g_2(b) \in G \times L$ we call g_1 **hierarchically below** g_2, written $g_1(a) \leq g_2(b)$, if and only if $g_1(a)^{\shortmid} \subseteq g_2(b)^{\shortmid}$.*

Note that an object can be hierarchically below itself for different values, i.e., $g_1(a), g_1(b) \in G \times L$ may yield $g_1(a) \leq g_1(b)$.

As in the other models of Hierarchical Classes Analysis, we build object and attribute bundle matrices and define for them the matrix product.

Definition 2. *An **object bundle** is a subset $g_{i_1}(a_{j_1}), ..., g_{i_n}(a_{j_n})$ of fuzzy objects such that $g_{i_1}^{\shortmid}(a_{j_1}) \subseteq ... \subseteq g_{i_n}^{\shortmid}(a_{j_n})$. An object bundle is **associated to** an attribute bundle if and only if they form a protoconcept together. For the matrix representation of a fuzzy context with n bundles and associated object bundle matrix P and attribute bundle matrix Q, the **fuzzy matrix product** is given by $(P \circ Q)_{ij} := \bigvee_{l=1}^{n} P_{il} \otimes Q_{lj}$.*

That is, we compute the t-norm multiplication between each element of the l-th column of P with each element of the l-th row of Q for each $l \in \{1, \ldots, n\}$ and take the maximum over these products.

Compared to the fuzzy factorization with fuzzy concepts this method is more laborious, since the number of fuzzy protoconcepts is much bigger than the number fuzzy concepts.

5. Conclusion

The main aim of this paper is to connect two fields with another and show how they can benefit from each other. The formal concept analytical approach to Factor Analysis and Hierarchical Classes Analysis can be connected trough the factorization problem. We compared these two methods regarding dyadic, triadic and real data. Concerning the first two data types there is a one-to-one correspondence between the two methods. Due to reasons of parsimony and interpretability we developed the fuzzy approach to Hierarchical Classes Analysis.

References

[1] R. Belohlávek, *Fuzzy Relational Systems: Foundations and Principles*, Systems Science and Engineering, Kluwer Academic/Plenum Press, 2002.

[2] R. Belohlávek and V. Vychodil, *Factor analysis of incidence data via novel decomposition of matrices*, in Formal Concept Analysis: 7th International Conference, ICFCA 2009, S. Ferré and S. Rudolph, eds., vol. 5548 of Lecture Notes in Artificial Intelligence, 2009, pp. 83–97.

[3] ———, *Discovery of optimal factors in binary data via a novel method of matrix decomposition*, Journal of Computer and System Sciences, 76 (2010), pp. 3–20.

[4] ———, *Optimal factorization of three-way binary data*, in GrC, Hu X., L. T. Y., R. V., G.-B. J., L. Q., and B. A., eds., 2010, pp. 61–66.

[5] E. Ceulemans, I. V. Mechelen, and I. Leenen, *Tucker3 hierarchical classes analysis*, Psychometrika, 68 (2003), pp. 413–433.

[6] P. De Boeck and S. Rosenberg, *Hierarchical classes: model and data analysis*, Psychometrika, 53 (1988), pp. 361–81.

[7] B. Ganter and R. Wille, *Formale Begriffsanalyse: Mathematische Grundlagen*, Springer, Berlin, Heidelberg, 1996.

[8] C. Glodeanu, *Triadic factor analysis*, in Concept Lattices and Their Applications 2010, M. Kryszkiewicz and S. Obiedkov, eds., 2010, pp. 127–138.

[9] ———, *Factorization with hierarchical classes analysis and with formal concept analysis*, 9th International Conference on Formal Concept Analysis, LNAI 6628, (2011).

[10] I. Leenen, I. V. Mechelen, P. De Boeck, and S. Rosenberg, *Indclas: A three-way hierarchical classes model*, Psychometrika, 64 (1999), pp. 9–24.

[11] F. Lehmann and R. Wille, *A triadic approach to formal concept analysis.*, in ICCS, G. Ellis, R. Levinson, W. Rich, and J. F. Sowa, eds., vol. 954 of Lecture Notes in Computer Science, Springer, 1995, pp. 32–43.

[12] I. V. Mechelen, I. Lombardi, and E. Ceulemans, *Hierarchical classes modeling of rating data*, Psychometrika, 72 (2007), pp. 475–488.

[13] S. Pollandt, *Fuzzy-Begriffe*, Springer, 1997.

Technische Universität Dresden, 01062 Dresden, Germany
E-mail address: Cynthia_Vera.Glodeanu@mailbox.tu-dresden.de