UNIVERSITY OF OSTRAVA
FACULTY OF SCIENCE
DEPARTMENT OF MATHEMATICS

METHODS USING UNCERTAINTY IN THE ANALYSIS OF LARGE SCALE DATA SETS

Ph.D. THESIS

AUTHOR: Mgr. Iva Koplíková
SUPERVISOR: Prof. Ing. Vilém Novák, DrSc.

2016

OSTRAVSKÁ UNIVERZITA V OSTRAVĚ
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATIKY

METODY VYUŽÍVAJÍCÍ NEURČITOST PŘI ANALÝZE ROZSÁHLÝCH DAT

DOKTORSKÁ DISERTAČNÍ PRÁCE

AUTOR: Mgr. Iva Koplíková
VEDOUCÍ PRÁCE: Prof. Ing. Vilém Novák, DrSc.

2016

Prohlašuji, že předložená práce je mým původním autorským dílem, které jsem vypracovala samostatně. Veškerou literaturu a další zdroje, z nichž jsem při zpracování čerpala, v práci řádně cituji a jsou uvedeny v seznamu použité literatury.

Ostrava . . . .

(podpis)

Čestné prohlášení:
Já, níže podepsaná studentka, tímto čestně prohlašuji, že text mnou odevzdané závěrečné práce v písemné podobě i na CD nosiči je totožný s textem závěrečné práce vloženým v databázi DIPL2.

Ostrava . . . .
(podpis)

Beru na vědomí, že tato doktorská disertační práce je majetkem Ostravské univerzity (autorský zákon Č. 121/2000 Sb., §60 odst. 1), bez jejího souhlasu nesmí být nic z obsahu práce publikováno.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně Ostravské univerzity.

Ostrava . . . .

(podpis)

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Ing. Vilém Novák, DrSc., for his continuous support, valuable and constructive suggestions and motivation. In addition, I would like to thank to all my colleagues from the Institute for Research and Applications of Fuzzy Modeling for really kind co-operation.

Ostrava, 2016 Iva Koplíková

Summary

The aim of my thesis is a contribution to one of fields of data mining, i.e. the linguistic association analysis. The main idea is from numerical data set to identify valid, novel, potentially useful, and ultimately understandable knowledge. The new knowledge is called "association". Therefore we represent associations in natural language we speak about linguistic association analysis. The main advantage of linguistic associations is in a hight understandability, furthermore the found linguistic associations can be interpreted as standard fuzzy IF-THEN rules.

The goal of the thesis is to suggest the new mathematical model with more specific results and represented it in the known algorithm. Moreover, in this thesis, the fuzzy confirmation measures are employed with several properties that enable a further work with found associations. Further these theoretical knowledge is implemented into one of well-known algorithms.

In the thesis, the basic information of data mining is summarized. We will specialize in a research of an association analysis. Theme motivation, thesis structure and contributions are described in Chapter 1. The elementary concepts of the theory of fuzzy modeling are established in Chapter 2.

The main part of thesis is elaborated in Sections 3, 4 and 5. In first one we present the original mathematical model published by V. Novák. The original mathematical model is modified. We obtain the same results as well as more specific results. At the end of the chapter the comparison of models is shown. The chapter Properties induced by fuzzy confirmation measures (Chapter 4) studies three pairs of fuzzy confirmation measures (support and confidence degrees) with the respect to axioms and inference rules that are used in database design as well as properties that are motivated by analogous properties that are used in GUHA method or in the classic Apriori algorithm. On the base of the extended mathematical model and above mentioned properties the modified Apriori algorithm is constructed in Chapter 5.

Keywords: Data mining, linguistic association, association rule, reduction rule, confirmation measure.

Anotace

Cílem této práce je jedna z oblastí dobývaní dat, a to lingvistická asociační analýza. Hlavní myšlenkou asociační analýzy je identifikovat z numerických dat nové, platné, potenciálně užitečné a pochopitelné znalosti. Novou znalost budeme nazývat “asociace” a jelikož ji prezentujeme v přirozeném jazyce, jedná se o lingvistickou asociační analýzu. Hlavní výhodou zmíněné analýzy je její pochopitelnost, jelikož nalezené asociace můžeme interpretovat formou IF-THEN pravidel.

V této disertační práci je navržen nový matematický model, který kromě výsledků dosažených v původním modelu, obsahuje také více specifické výsledky. V práci jsou použity fuzzy konfirmační míry s několika vlastnostmi, které umožňují další práci s nalezenými asociacemi. Tyto teoretické znalosti jsou implementovány do jednoho z dobře známých algoritmů.

V práci jsou shrnuty základní informace o dobývání dat. Motivace, struktura práce a přínosy jsou popsány v Kapitole 1. Základní pojmy teorie fuzzy modelování jsou zavedeny v Kapitole 2.

Hlavní část práce je zpracována v kapitolách 3, 4 a 5. V první z nich prezentujeme původní matematický model, který byl publikován V. Novákem. Původní matematický model je modifikován a díky němu získáváme stejné a současně více specifické výsledky. Na konci kapitoly můžeme vidět srovnání těchto modelů. V kapitole Vlastnosti indukované fuzzy konfirmačními mírami (Kapitola 4) se zaměřujeme na tři páry fuzzy konfirmačních měr (definovaných pomocí stupně podpory a stupně spolehlivosti) s ohledem na axiomy a odvozovací pravidla, která se používají při návrhu databází, stejně jako vlastnosti, které jsou motivovány obdobnými vlastnostmi používajícími se v GUHA metodě nebo v klasickém Apriori algoritmu. Na základě rozšířeného matematického modelu a výše uvedených vlastností je popsán v Kapitole 5 modifikovaný Apriori algoritmus.

Klíčová slova: Dobývání dat, jazyková asociace, asociační pravidlo, redukční pravidlo, konfirmační míra.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data.

In the world, the amount of data is increasing. We would like to mine some knowledge from data or look for patterns in data. We investigate ways how found patterns can be sought automatically, identified, validated, and used for prediction. In other words, the data mining is about solving problems by analysing data already present in databases.

The data mining is used in a practise in the sphere of scientific research as well as in the commercial sector. There exist many applications, for example, in medical and pharmaceutical area (diagnostic methods and the development and testing of medicaments), marketing and sales area (automatic web page design, recommendations for new purchases, cross selling) but also in bioinformatics, counter-terrorism and for analysis of the data collected by NASA from space observation and evaluation of geophysical data.

The Knowledge Discovery in Databases (KDD) process is commonly defined with the stages:

1. In the *selection*-step the significant data gets selected or created. Only relevant information is selected, and also meta data or data that represents background knowledge.

2. Before applying data mining an appropriate data preparation is important. Relevant elements of the provided data have to be detected and filtered out. This phase is called the *pre-processing* phase. To detect knowledge the effective

main task is to pre-process the data properly and not only to apply data mining tools. Elements of the pre-processing span the cleaning of wrong data, the treatment of missing values and the creation of new attributes.

3. The *transformation* phase of the data may result in a number of different data formats, since variable data mining tools may require variable formats. The data also is manually or automatically reduced.

4. In the *data mining* phase, the data mining task is approached. There exist many techniques for data mining. The output of this step is detected patterns, novel relations, predictions, etc.

5. The *interpretation* of results of data mining process reveals whether or not the result is interesting. This is why this step is also called evaluation.

The another view is the Cross Industry Standard Process for Data Mining (CRISP-DM) which defines six phases (see Figure 1.1):

1. *Business Understanding*
   This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.

2. *Data Understanding*
   The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

3. *Data Preparation*
   The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

4. *Modeling*
   In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

5. *Evaluation*
   At this stage in the project you have built a model (or models) that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

6. *Deployment* Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. Even if the analyst deploys the model it is important for the customer to understand up front the actions which will need to be carried out in order to actually make use of the created models.

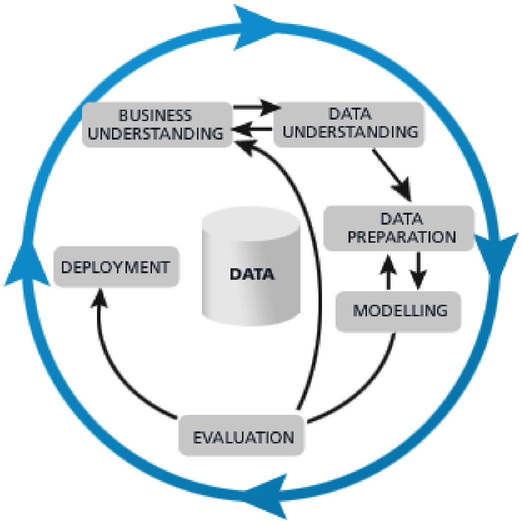

Figure 1.1: The six phases of CRISP-DM.

There exists many techniques of data mining. Below the mostly used approaches are meant.

Decision trees or also called Top Down Induction of Decision Trees (TDIDT). At the begining is one node and then other subnodes are specialized. Trees are constructed by the divide-and-conquer algorithm. The aim is to search for tree that is consistent to training data. These decision trees are mostly use to classification rules.

Clustering is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. The most known algorithm is k-means clustering algorithm. It was first used by James MacQueen in 1967 though the idea goes back to Hugo Steinhaus in 1957. At the beginning k initial points are chosen to represent initial cluster centers. The remaining data are assigned to the nearest one initial cluster center. The iteration continues until there are no changes in the cluster. The another clustering algorithm is DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise) that was proposed by Martin Ester, Hans-Peter Kriegel, Jrg Sander and Xiaowei Xu in 1996. It finds a number of clusters starting from the estimated density distribution of corresponding nodes. DBSCAN is one of the most common clustering algorithms.

Classification describes a task of generalizing known structure to apply to new data. It is separate-and-conquer technique because it identifies a rule that covers instances in the class (and excludes ones not in the class), separates them out, and continues on those that are left.

Regression is a data mining (machine learning) technique used to fit an equation to a model the data with the least error. The regression functions are used to determine the relationship between the dependent variable and one or more independent variables. The dependent variable is the one whose values you want to predict, whereas the independent variables are the variables that you base your prediction on. Most often used types of regression models are linear, polynomial, and logistic regression.

Association Rule Learning is one of the most frequent used techniques. It was described by R. Atrawal and R. Srikant in [1]. Association rules search for relationships between variables. For example, the known market basket analysis discovers and understands customer purchasing behavior. A supermarket collects data about customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. More detailed about the association rule learning is in Chapter 3.1.

Data mining is the process of analysing data from different perspectives and summarizing it into useful, valid, novel information in large scale data sets. It allows users to analyse data from many different dimensions or angles, categorize it, and summarize the relationships identified. The process of data mining has attracted a lot of research interest in the last two decades. It should be mentioned that the first data mining method was the GUHA method presented in [9] even earlier. Probably because of a different terminology (the author of [9] did not use the term "data mining") the GUHA method is not well known and some results were forgotten and then rediscovered in the nineties (e.g. [18], [27]). For more precise information on the GUHA method we refer to [11] and references therein.

This thesis is a contribution to the theoretical foundations of data mining and partially extends the use of the GUHA method. We follow a direction that was recently developed by V. Novák in several papers (c.f., e.g. [22] and [19]). Within the novel Novák's approach a method for searching for so-called linguistic associations was elaborated ([23]). This method is based on the GUHA method and results of formal fuzzy logic ([20]) and allows us to mine linguistic associations of the form

IF the area of the base of a cylinder is big AND the height of this cylinder is also big THEN the volume of this cylinder is big.

The advantage of this approach is especially the high understandability of founded associations since they are presented in natural language. Additionally, it should be also mentioned that found linguistic associations can be interpreted as standard fuzzy IF-THEN rules (see [6] and references therein). Further, any data mining procedure working with categorical or logical data can be applied to Novák's mathematical model of linguistic expressions and predications.

We consider three the most commonly used confirmation measures using of which was mathematically justified in [6] (see also [8] for further information). For each of considered confirmation measures we study special properties motivated by so-called Armstrong axioms that, among other things, are used for database design (see [2]) and are also valid in fuzzy attribute logic developed, e.g. in [3]. This logic can be applied to similar data sets and, under some assumptions, establishes a complete and sound system of associations. Thus it was a natural question under what conditions we can obtain similar relations in ordinary fuzzy association analysis. The remaining properties are motivated by properties that are used, for example, in GUHA method ([11]) or in known Apriori algorithm ([1, 5]).

We demonstrate how to apply results from [15] and [16], and some specific properties of the model of evaluative linguistic expressions (see [14] and [23]) into known Apriori algorithm. Our algorithm might be computationally more complex than the original one as we use more complex model of linguistic expressions. But we also suggest several ways allowing us to possibly reduce number of tested associations. Some of them are based on properties of fuzzy confirmation measures and, consequently, can be used in the original algorithm as well. Additionally, this is the first algorithm where linguistic associations can be mined without transforming to the non-fuzzy case. Additionally, the proposed algorithm is the first case where hierarchical structure of evaluative linguistic expressions is taken into consideration and this feature can substitute some preprocessing steps. It should be emphasized that the same ideas can be used in other models of the same kind.

1.2 Thesis structure

The structure of thesis is following. In Chapter 1 the motivation and goals of this thesis are introduced. In Chapter 2 the basic concepts and definitions are described.

In Chapter 3 the original mathematical model published by V. Novák is modified. At the end of the chapter the models are compared. In Chapter 4, three pairs of fuzzy confirmation measures (support and confidence degrees) with the respect to axioms and inference rules that are used in database design as well as properties that are motivated by analogous properties. These properties are used in GUHA method or in the classical Apriori algorithm. On the base of the extended mathematical model and above mentioned properties the modified Apriori algorithm is constructed in Chapter 5.

1.3 My contribution

The contribution of Chapter 3 is to show two mathematical representations of linguistic expressions based on common notions of fuzzy mathematics, namely on a fuzzy partition and covering, respectively. The purpose of this chapter is to show that such representations are possible and the second one based on fuzzy covering extends the model developed by V. Novák et al. in [23] in a natural way.

Additionally, due to using standard notions of fuzzy modeling we also extend the applicability of the method. The original model from [23] is suggested such that any data mining methods using crisp decompositions (working with logical or categorical data) can be used. In this chapter we use the same approach in order to demonstrate our results. However, our model is suggested such that other data mining techniques using fuzzy partitions ([25]) and coverings (c.f. [4], [12]) can be applied.

In Chapter 4 three the most commonly used fuzzy confirmation measures are considered - see [6] and [8] for justification of their existence. For each pair of considered confirmation measures (i.e., support and confidence measures) we study several properties. Six of them are motivated by so-called Armstrong axioms that, among other things, can be used for database design (see e.g. [2]) and are also valid in fuzzy attribute logic developed e.g. in [3]. This logic can be applied to data sets similar to ours and, under some additional assumptions, establishes a complete and sound system of associations. Thus, it was a natural question under what conditions we can obtain similar relations in ordinary fuzzy association analysis. Further properties we have decided to study are motivated by properties that can be used in current methods of association analysis - for example, in GUHA method ([11]) or in known Apriori algorithm ([1] and also [5]).

We explain that some properties remain valid when we use our fuzzy confirmation measures. But we have also obtained some negative results and we demonstrated that the situation can be improved if some additional (expert) knowledge is applied to our properties. We would like to stress that our results are provided separately either for support or confidence measure if necessary.

The Chapter 5 demonstrates how results from [15] and [16] can be applied. Some specific properties of the model of evaluative linguistic expressions (see [14] and [23]) are implemented into known Apriori algorithm. Our algorithm might be computationally more complex than the original one as we use more complex model of linguistic expressions. But we also suggest several ways allowing us to possibly reduce number of tested associations. Some of them are based on properties of fuzzy confirmation measures and, consequently, can be used in the original algorithm as well. Additionally, this is the first algorithm where linguistic associations can be mined without transforming to the non-fuzzy case. Additionally, the proposed algorithm is the first case where hierarchical structure of evaluative linguistic expressions is taken into consideration and this feature can substitute some preprocessing steps. It should be emphasized that the same ideas can be used in other models of the same kind.

Chapter 2

Basic concepts of the theory of fuzzy modeling

At the beginning of this chapter the basic notions of fuzzy mathematics are described. Then we deal with special concepts for linguistic associations.

2.1 Basic concepts

By $\mathbb{N}$ we denote the set of natural numbers, the set of real numbers is denoted by $\mathbb{R}$.

Definition 1 A *fuzzy set* in the universe U (notation $A \underset{\sim}{\subset} U$) is a function from the universe U to the closed unit interval $[0,1]$, i.e., $A : U \to [0,1]$. The function A is called a *membership function* of the fuzzy set A and the value $A(x)$ is a *membership degree* of an element $x \in U$.

Definition 2 A *support* $\mathrm{Supp}(A)$ of a given fuzzy set A is usually defined as

$$\mathrm{Supp}(A) = \{x \in U \mid A(x) > 0\}.$$

Definition 3 A fuzzy set $A \underset{\sim}{\subset} U \subseteq \mathbb{R}$ is *convex*, if

$$A(\lambda x + (1-\lambda)y) \geq A(x) \ \wedge \ A(y)$$

for all $x, y \in U$ and for all $0 \leq \lambda \leq 1$, where the symbol $\wedge$ denotes the minimum of values $A(x)$ and $A(y)$.

Definition 4 Let $A \underset{\sim}{\subset} [a,b]$ and $\alpha \in [0,1]$. Then the set

$$A_\alpha = \{x | x \in [0,1], A(x) \geq \alpha\}$$

is called the α–cut.

It should be mentioned that if A is upper semi-continuous then every α–cut is a closed subset of $[a, b]$.

Definition 5 A *t–norm* is binary function $\otimes : [0,1] \times [0,1] \longrightarrow [0,1]$ satisfying these conditions:

$$\begin{aligned} x \otimes y = y \otimes x, &\quad \text{(commutativity)},\\ x \otimes (y \otimes z) = (x \otimes y) \otimes z, &\quad \text{(associativity)},\\ \text{if } x \leq y \text{ then } x \otimes z \leq y \otimes z, &\quad \text{(monotonicity)},\\ 0 \otimes x = 0 \text{ and } 1 \otimes x = x &\quad \text{(boundary condition)}. \end{aligned}$$

Definition 6 For given t–norm exists a corresponding *t–conorm*. It is binary function $\oplus : [0,1] \times [0,1] \longrightarrow [0,1]$ given by

$$x \oplus y = 1 - (1 - x) \otimes (1 - y)$$

Example 1 *The familiar t–norms are:*

1. minimum *given by* $x \otimes y = \min\{x, y\}$*,*
2. product t–norm *is* $x \otimes y = x \cdot y$ *(the ordinary product of real numbers),*
3. Łukasiewicz t–norm *is defined* $x \otimes y = \max\{0, x + y - 1\}$*.*

Example 2 *The corresponding t–conorms are:*

1. maximum or Gödel t–conorm *given by* $a \oplus b = \max\{a, b\}$*,*
2. product t–conorm or probabilistic sum *is* $a \oplus b = a + b - a \cdot b$*,*
3. Łukasiewicz t-conorm or bounded sum *is defined* $a \oplus b = \min\{1, a + b\}$*.*

Definition 7 A two-dimensional *copula* c is a mapping $c : [0,1] \times [0,1] \longrightarrow [0,1]$ satisfying the following three conditions:

1. $c(u, 0) = c(0, u) = 0$ for every $u \in [0,1]$,
2. $c(u, 1) = c(1, u) = u$ for every $u \in [0,1]$,
3. $c(u_2, v_2) - c(u_1, v_2) - c(u_2, v_1) + c(u_1, v_1) \geq 0$ for every $u_1, u_2, v_1, v_2 \in [0,1]$ satisfying $u_1 \leq u_2$, $v_1 \leq v_2$.

We can see that minimum and Łukasiewicz t–norm are examples of copula. Moreover Łukasiewicz t–norm is the smallest copula.

Definition 8 An *implication operator* $\rightarrow: [0,1] \times [0,1] \longrightarrow [0,1]$ is a generalization of the material implication if it satisfies, for $x, y, x_0, y_0 \in [0,1]$,

1. $x \rightarrow y \leq x_0 \rightarrow y$ for $x_0 \leq x$,
2. $x \rightarrow y \leq x \rightarrow y_0$ for $y \leq y_0$,
3. $1 \rightarrow y = y$,
4. $0 \rightarrow 0 = 1$.

Example 3 *For below meaning t–norms (from Example 1) there exist fuzzy implications (more detailed in [24]).*

1. Gödel implication

$$x \rightarrow y = \begin{cases} 1, & \textit{if } x \leq y, \\ y, & \textit{if } x > y. \end{cases}$$

2. Product implication

$$x \rightarrow y = \begin{cases} 1, & \textit{if } x \leq y, \\ \frac{y}{x}, & \textit{if } x > y. \end{cases}$$

3. Łukasiewicz implication

$$x \rightarrow y = \begin{cases} 1, & \textit{if } x \leq y, \\ 1 - x + y, & \textit{if } x > y. \end{cases}$$

In this thesis we mainly work with the standard Łukasiewicz MV-algebra $\mathcal{L}$ (more detailed in [24]) as an algebra of truth values. This is the algebra

$$\mathcal{L} = \langle [0,1], \vee, \wedge, \otimes, \rightarrow, 0, 1 \rangle,$$

where $\vee$ (resp. $\wedge$) is an operation of *supremum* (resp. *infimum*) and, for any $a, b \in [0,1]$, $a \otimes b = 0 \vee (a+b-1)$ is *Łukasiewicz conjunction* and $a \rightarrow b = 1 \wedge (1-a+b)$ is *Łukasiewicz implication*. The operations $\wedge$, $\vee$ and $\rightarrow$ interpret logical connectives `AND`, `OR` and the implication, respectively. The negation $\neg$ is defined pointwise by $\neg A(x) = 1 - A(x)$ for any $x \in [0,1]$. For a justification of the choice of $\mathcal{L}$ we refer to [23] and references therein.

In the case when the universe U is a closed interval $[a,b]$, $a,\ b \in \mathbb{R}$, we may define a fuzzy partition. We say that a finite system of fuzzy sets $A_1, \ldots, A_n \underset{\sim}{\subset} [a,b]$ forms a *fuzzy partition* of $[a,b]$ if there are points $x_i \in [a,b]$, $i = 1, 2, \ldots, n$, and $n \geq 2$, $a = x_1 < x_2 < \ldots < x_n = b$, called *nodes* and the following five conditions are satisfied for $k = 1, \ldots, n$:

1. $A_k : [a,b] \to [0,1]$, $A_k(x_k) = 1$,
2. $A_k(x) = 0$ if $x \notin (x_{k-1}, x_{k+1})$, where for the uniformity of denotation, we put $x_0 = a$ and $x_{n+1} = b$,
3. $A_k(x)$ is continuous,
4. $A_k(x)$, $k = 2, \dots, n$, monotonically increases on $[x_{k-1}, x_k]$ and $A_k(x), k = 1, \dots, n-1$, monotonically decreases on $[x_k, x_{k+1}]$,
5. For all $x \in [a,b]$
$$\sum_{k=1}^{n} A_k(x) = 1.$$
6. $A_k(x_k - x) = A_k(x_k + x)$, for all $x \in [0,h], k = 2, \dots, n-1, n > 2$,
7. $A_{k+1}(x) = A_k(x-h)$, for all $x \in [a+h, b], k = 2, \dots, n-2, n > 2$.

If the nodes $x_1, \dots, x_n$ are equidistant i.e., $x_k = a + h(k-1)$, $k = 1, \dots, n$ where $h = (b-a)/(k-1)$ and the properties 6. and 7. are met then we call the fuzzy partition *uniform* and talk about *uniform basic functions*.

Definition 9 A convex fuzzy set A is called a *fuzzy number* if it satisfies conditions 1.– 4. of the preceding definition (for $A := A_k$) for some points $x_{k-1} < x_k < x_{k+1}$.

A *fuzzy covering* of an interval $[a,b]$ is a system $\{A_i\}_{i=1}^{n}$ such that there are $n+1$ points $a = x_1 < x_2 < \dots < x_n < x_{n+1} = b$ and the following conditions hold for $k = 1, 2, \dots, n$

1. $A_k(x) : [a,b] \longrightarrow [0,1]$ is continuous,
2. $A_k(x) = 0$ if $x \notin (x_{k-1}, x_{k+2})$, where for the uniformity of denotation, we put $x_0 = a$ and $x_{n+2} = b$,
3. $A_k(x) = 1$ for $x \in [x_k, x_{k+1}]$,
4. $A_k(x)$ monotonically increases on $[x_{k-1}, x_k]$ and $A_k(x)$, $k = 1, \dots, n-1$, monotonically decreases on $[x_{k+1}, x_{k+2}]$.

2.2 Evaluative linguistic expressions and predications

In this section we define several notions that are necessary for mining linguistic associations. First of all we introduce a few fundamental notions of the theory of evaluative linguistic expressions. This approach was initiated by V. Novák in [19] and then developed in several subsequent papers (see [23] and also [22]). A special attention should be focused also to [21] where the theory of evaluative linguistic expressions was elaborated using higher order fuzzy logic.

The evaluative linguistic expressions are natural language expressions, such as "significantly large, extremely big, roughly expensive, more or less thin", etc. They can be used in the data-mining process especially for user–friendly presentation of discovered associations. We distinguish several types of *evaluative linguistic expressions*:

- ⟨atomic evaluative expression⟩ := *Small* (briefly, *Sm*), *Medium* (*Me*), *Big* (*Bi*).
- ⟨pure evaluative expression⟩ :=
  ⟨linguistic hedge⟩ ⟨atomic evaluative expression⟩ .

A *linguistic hedge* was introduced by L. Zadeh in [28]. It is a special adverb modifying the meaning of a given atomic evaluative expression. In paper [23] the following linguistic hedges are considered: *Extremely* (*Ex*), *Significantly* (*Si*), *Very* (*Ve*), *More or Less* (*ML*), *Roughly* (*Ro*), *Quite Roughly* (*QR*) and *Very Roughly* (*VR*). We distinguish linguistic hedges with a *narrowing effect* (*Ex*, *Si* and *Ve*) or *widening effect* (*ML*, *Ro*, *QR* and *VR*).

Linguistic hedges are constructed by using continuous functions $\nu_{a,b,c} : [0,1] \longrightarrow [0,1]$ (horizon shifts) where $a < b < c$ are parameters, $\nu_{a,b,c}(y) = 0$ for $y \leq a$, $\nu_{a,b,c}(y) = 1$ for $c \leq y$ and it is increasing otherwise, i.e.,

$$\nu_{a,b,c}(y) = \begin{cases} 1, & c \leq y, \\ 1 - \frac{(c-y)^2}{(c-b)(c-a)}, & b \leq y < c, \\ \frac{(y-a)^2}{(b-a)(c-a)}, & a \leq y < b, \\ 0, & y \leq a. \end{cases}$$

Further evaluative linguistic expressions are the following:

- ⟨fuzzy number⟩ := ⟨linguistic hedge⟩ ⟨numeral⟩,
  where *numeral* is a name of a given real number,

- $\langle$negative evaluative expression$\rangle$:= `not` $\langle$atomic evaluative expression$\rangle$,
- $\langle$specifying evaluative expression$\rangle$:=
  $\langle$atomic evaluative expression$\rangle$ `but` $\langle$negative evaluative expression$\rangle$.

A *compound evaluative linguistic expression* is of the form

- $E := \texttt{AND}_{i\in I}\, C_i, \quad$ where $C_i := \texttt{OR}_{k\in K_i}(A_k)$
- $F := \texttt{OR}_{j\in J} D_j, \quad$ where $D_j := \texttt{AND}_{l\in L_j}(B_l)$

and A_k and B_l for index sets I, J, K_i and L_j are above defined evaluative linguistic expressions. The term `AND` (resp. `OR`) represents commonly used linguistic connective "and" (resp. "or").

In this paragraph, we outline the main concepts of semantics of evaluative expressions. For more precise information on described notions we refer to [23]. One of established notions is the *context* (*possible world*). From the point of view of logic understood the context is a state of the world at a given time moment and place.

Definition 10 *The context is a mapping $w : [0, 1] \longrightarrow [v_L, v_R]$ and is specified by a triplet $\langle v_L, v_M, v_R\rangle$, where $v_L, v_M, v_R \in \mathbb{R}$ and $v_L < v_M < v_R$. The value v_L denotes the smallest value (a left border) and v_R means the highest value (a right border) that makes a sense to consider. The value v_M signifies a mean value (a center value).*

For simplicity, we write $x \in w$ instead an element x belongs to the interval $[v_L, v_R]$ $(x \in [v_L, v_R])$.

Intension of a linguistic expression is an abstract construction which conveys a property denoted by the expression and it is invariant towards change of the context.

Definition 11 *The intension is a function*

$$Int(A) : W \longrightarrow F(U),$$

where A is linguistic expression, W is set of contexts and $F(U)$ is set of fuzzy sets in the universe U.

This means that the intension is a function from the set of all possible contexts to the set of fuzzy sets. In other words, the intension of evaluative expression is a representation of the property that the given expression determines.

We can model intensions of evaluative expressions in this way:

$$Sm_\nu(z) = \nu(LH(z)),$$

$$Me_\nu(z) = \nu(MH(z)),$$

$$Bi_\nu(z) = \nu(RH(z)),$$

where $\nu \in \mathbf{Hf}$ interprets ⟨linguistic hedge⟩, $z \in [0,1]$ and linear functions $LH, MH, RH : [0,1] \longrightarrow [0,1]$ are defined by

$$LH(z) = (\frac{0,5 - z}{0,5})^\star,$$

$$MH(z) = (\frac{z}{0,5})^\star \wedge (\frac{1-z}{0,5})^\star,$$

$$RH(z) = (\frac{z - 0,5}{0,5})^\star$$

where the star $\star$ means cut of all the values to interval $[0,1]$. Then we put

$$Int(\langle\text{linguistic hedge}\rangle)small = Sm,$$

$$Int(\langle\text{linguistic hedge}\rangle)medium = Me,$$

$$Int(\langle\text{linguistic hedge}\rangle)big = Bi.$$

The next important notion is an *extension* of a linguistic expression. That is a class of objects that are determined by its intension in a given context (possible world) and so, it changes whenever the context (time, place) is changed.

Definition 12 *The extension is a fuzzy set in context $w \in W$*

$$Ext_w(A) = Int(A)(w) \underset{\sim}{\subseteq} U.$$

It follows that for any context is derived one extension, i.e. fuzzy set. One intension specifies a class of extensions.

Example 4 *We assume the expression "young age" then "be young" is name of intension that is independent on time moment and place. First, the context has to be specify and then the extension can be determined. If we consider an age of turtles, the context is the interval $[0,200]$. Then the young turtle is given by the extension (fuzzy set) on interval $[0,80]$. But if we consider age of people with the context $[0,120]$ then the young person is the extension (fuzzy set) on interval $[0,25]$.*

We will distinguish between evaluative predications and expressions. An *evaluative linguistic predication* is an evaluative linguistic expression considered in a certain context (possible world), i.e. a sentence of the form

⟨noun phrase X⟩ is A,

where A denotes an evaluative linguistic expression. Such a linguistic predication will be denoted by $A(X)$.

For example, the word *tall* is an evaluative linguistic expression but the sentence "a man is tall" (or, equivalently, the term "a tall man") forms an example of a linguistic predication. A precise mathematical representation of above mentioned notions is elaborated in [23].

According to an ordering $\preceq$ ([23]) given by what a given linguistic hedge expresses in natural language, i.e.

$$\underbrace{Ex \preceq\ Si\ \preceq Ve}_{\text{narrowing effect}} \preceq \langle\text{empty hedge}\rangle \preceq \underbrace{ML \preceq Ro \preceq QR \preceq VR}_{\text{widening effect}}, \tag{2.2.1}$$

where ordering $\preceq$ means that all values in some context, that are extremely small (or big), are also significantly small (or big), etc.

Later we will need an ordering of evaluative linguistic expressions and predications in each context $w \in W$. First we define an ordering on the set of linguistic expressions and then we naturally shift this ordering to the set of linguistic predications. We define an ordering $Ev_{\nu_1} \preceq Ev_{\nu_2}$.

Definition 13 *We suppose that the set* ***Hf*** *is partially ordered by the specificity relation $\preceq$. The ordering induces a partial ordering on evaluating expressions defined by*

$$Ev_{\nu_1} \preceq Ev_{\nu_2} \quad \textit{iff} \quad \nu_1 \leq \nu_2, \tag{2.2.2}$$

where Ev is either of Sm, Me or Bi and $\leq$ is pointwise ordering of functions.

Definition 14 *The position ordering $\lessdot$ corresponds to the position of the evaluating expressions on the scale, i.e.*

$$Sm_{\nu_1} \lessdot Me_{\nu_2} \lessdot Bi_{\nu_3} \tag{2.2.3}$$

where ν_1, ν_2, ν_3 are arbitrary.

From the ordering $\preceq$ and $\lessdot$ we introduce *natural (partial) ordering.*

Definition 15 *The natural (partial) ordering of evaluative expressions as lexicographic ordering is*

$$Ev_1 \lessapprox Ev_2,$$

where first we order Ev_1, Ev_2 according to $\preceq$ (Definition 13), and then according to $\lessdot$ from (Definition 14).

Example 5 *For Si Sm and Me we obtain Si Sm $\preceq$ Me by Definition 14 since the first (resp. the second) linguistic expression contains the expression small (resp. medium), i.e. Si Sm $\lessapprox$ Me. But if we consider Si Sm and Sm then these sets cannot be ordered by Definition 14 since both mentioned expressions contain the atomic expression small. Thus we order separately evaluative linguistic expressions describing small, medium and big values by Definition 13, i.e. Si Sm $\preceq$ Sm and it means Si Sm $\lessapprox$ Sm.*

At the end of this chapter we want to show a picture of fuzzy sets representing extensions of certain evaluative linguistic expressions elaborated in [23]. Here we can see that the context $[a_j, b_j]$ is covered by fuzzy sets that are represented by atomic evaluative linguistic predications *Sm*, *Me*, *Bi* and by evaluative linguistic predications *Si Sm*, *ML Sm* and *ML Me*.

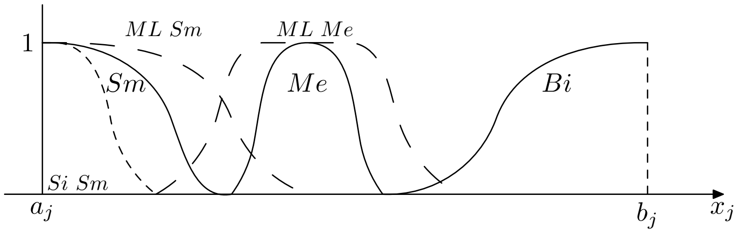

Figure 2.1: Fuzzy sets that are extensions of the corresponding evaluative linguistic expressions described in [23].

Chapter 3

Extended mathematical model of evaluative linguistic expressions

In this chapter we contribute to a part of data mining allowing us to search for associations among attributes that can be expressed in a form of natural language sentences. The theoretical background and also a method for mining such associations was published recently in [23]. We elaborated other mathematical representations of the model presented in the mentioned paper in order to extend its applicability.

This chapter is a contribution to the theoretical foundations of data mining and partially extends the use of the GUHA method. We follow a direction that was recently developed by V. Novák in several papers (c.f., e.g. [19] and [22]). Within the novel Novák's approach a method for searching for so-called *linguistic associations* was elaborated ([23]). This method is based on the GUHA method and results of formal fuzzy logic ([20]) and allows us to mine linguistic associations of the form

IF the area of the base of a cylinder is big AND the height of this cylinder is also big THEN the volume of this cylinder is big.

The advantage of this approach is especially the high understandability of founded associations since they are presented in natural language. Additionally, it should be also mentioned that found linguistic associations can be interpreted as standard fuzzy IF-THEN rules (see [6] and references therein). Further, any data mining procedure working with categorical or logical data can be applied to Novák's mathematical model of linguistic expressions and predications. However, this mathematical model has some disadvantages if we use it for mining of linguistic associations. We found them when we tested this method ([7]). The problems are indicated in Section 3.2. To cope with them we present another two mathematical representations of linguistic expressions based on common notions of fuzzy mathematics, namely on a fuzzy partition and covering, respectively. The purpose of this chapter is to show that such representations are possible and the second one based on fuzzy covering extends the model developed by V. Novák et al. in [23] in a natural way.

Additionally, due to using standard fuzzy notions we also extend the applicability of the method. The original model from [23] is suggested such that any data mining methods using crisp decompositions (working with logical or categorical data) can be used. In this chapter we use the same approach in order to demonstrate our results. However, our model is suggested such that other data mining techniques using fuzzy partitions ([25]) and coverings (c.f. [4], [12]) can be applied.

Organization of this chapter is the following. In Section 3.1, we present and discuss two possible mathematical models used for association mining. A method for mining of linguistic associations is presented further in Section 3.2 and Section 3.3 is devoted to a short discussion on reduction rules. Finally, before a short concluding section (Section 3.5) an example demonstrating our method is presented (Section 3.4).

3.1 Mathematical models and mining of linguistic associations from numerical data

We present a data mining process that is applied to real-valued two-dimensional table. Thus, we have a data-set of the following form

$$\begin{array}{c|cccc}
 & X_1 & X_2 & \dots & X_k \\
\hline
o_1 & a_{11} & a_{12} & \dots & a_{1k} \\
o_2 & a_{21} & a_{22} & \dots & a_{2k} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
o_m & a_{m1} & a_{m2} & \dots & a_{mk}
\end{array}$$

where any real number $a_{ij} \in \mathbb{R}$ is a value of jth *attribute* (property) X_j on ith *object* (observation, transaction) o_i. The set of all objects $\mathcal{D}_o = \{o_i | i = 1, \dots, m\}$ is called row set.

To apply our method we have to specify a *context* w_j for each attribute X_j (see the definition on Page 24). We consider a_{ij} from a context w_j, i.e. $a_{ij} \in [a_j, b_j] \subset \mathbb{R}$ where values of a closed interval $[a_j, b_j]$ are the left and right values of the context. The values a_j, b_j are chosen by a real meaning of an attribute X_j or by what we want to represent by linguistic expressions.

Example 6 *Assume that X_j represents the height of people in centimeters and, in* *the data set,* $a_{ij} \in [190, 195]$ *for any* $i = 1, 2, \ldots, m$*. Then the common choice* ($a_j = \min\{a_{1j}, a_{2j}, \ldots, a_{mj}\}$ *and* $b_j = \max\{a_{1j}, a_{2j}, \ldots, a_{mj}\}$*) is not suitable since a 190,5 cm tall person would be named* small *after using the model from Figure 2.1. A reasonable choice can be* $a_j = 150,\ b_j = 200$ *in this case.*

Below we work with fixed context $[a_j, b_j]$. The extension of evaluative linguistic expression is modelling by fuzzy set in given context.

The first mathematical representation of linguistic expressions (*Model I*) is based on the fact that the context of each attribute X_j is covered by a fuzzy partition $P_j := \{A_{i,j}\}_{i=1}^{n_j}$. This assumption is quite general and is common for several fuzzy data mining techniques (c.f. [25]).

The number n_j of fuzzy sets $A_{i,j}$ contained in P_j need not be fixed since there are several possibilities how to define relevant evaluative linguistic expression represented by members of P_j. The choice of n_j, resp. of evaluative linguistic expressions (predications) could be defined without knowledge of the data set, but it should be chosen according to what is represented by the attribute X_j and how are values of X_j distributed in its context. Then fuzzy sets $A_{i,j}$ of P_j are divided into three subsets S_j, M_j, B_j representing extensions of *small*, *medium* and *big* values, respectively. Then (see the next paragraph) we can find a suitable linguistic expression for any member of S_j, M_j or B_j.

Our goal is to extend the mathematical model of evaluative linguistic expressions elaborated in [23]. Therefore linguistic expressions associated with fuzzy set $A_{i,j}$ consist mostly of specifying linguistic expressions (see Example 7 below). This choice is motivated by our latter effort to reconstruct fuzzy sets representing linguistic expressions introduced in [23] (see (3.1.1)). It should be stressed that the linguistic representation in Example 7 is not unique. We only show the way (using specifying fuzzy sets representing linguistic expressions) how to do this. Another linguistic representation is shown in Section 3.4.

Example 7 *If $n_j = 21$ we can obtain the following fuzzy sets representing extensions of the following linguistic expressions:*

$A_{1,j} \sim \mathit{Ex\ Sm}$,

$A_{2,j} \sim \mathit{Si\ Sm}$ `but not` $\mathit{Ex\ Sm}$,

$A_{3,j} \sim \mathit{Ve\ Sm}$ `but not` $\mathit{Si\ Sm}$,

$A_{4,j} \sim \mathit{Sm}$ `but not` $\mathit{Ve\ Sm}$,

$A_{5,j} \sim \mathit{ML\ Sm}$ `but not` Sm,

$A_{6,j} \sim \mathit{Ro\ Sm}$ `but not` $\mathit{ML\ Sm}$,

$A_{7,j} \sim \mathit{QR\ Sm}$ `but not` $\mathit{Ro\ Sm}$,

$A_{8,j} \sim \mathit{VR\ Sm}$ `but not` $\mathit{QR\ Sm}$,

$A_{9,j} \sim \mathit{FN1}$,

$A_{10,j} \sim \mathit{Lo\ Me}$,

$A_{11,j} \sim \mathit{Ex\ Me}$,

$A_{12,j} \sim \mathit{Hi\ Me}$,

$A_{13,j} \sim \mathit{FN2}$,

$A_{14,j} \sim \mathit{VR\ Bi}$ `but not` $\mathit{QR\ Bi}$,

$A_{15,j} \sim \mathit{QR\ Bi}$ `but not` $\mathit{Ro\ Bi}$,

$A_{16,j} \sim \mathit{Ro\ Bi}$ `but not` $\mathit{ML\ Bi}$,

$A_{17,j} \sim \mathit{ML\ Bi}$ `but not` Bi,

$A_{18,j} \sim \mathit{Bi}$ `but not` $\mathit{Ve\ Bi}$,

$A_{19,j} \sim \mathit{Ve\ Bi}$ `but not` $\mathit{Si\ Bi}$,

$A_{20,j} \sim \mathit{Si\ Bi}$ `but not` $\mathit{Ex\ Bi}$,

$A_{21,j} \sim \mathit{Ex\ Bi}$,

where $A \sim Sm$ means fuzzy set A representing extension of the linguistic expression Small. FN1, FN2 denote fuzzy sets associated with linguistic descriptions of a fuzzy number (for instance, more or less 5 *see definition on page 22) and linguistic expressions Hi Me and Lo Me denote linguistic expressions Higher Medium and Lower Medium, respectively. For medium values of the attribute X_j it is necessary to use linguistic expressions that are not considered in [22]. The reason for this is that natural language contains only a few linguistic expressions describing values that lie between typically* small *(resp.* big*) and typically* medium *values. For completeness, $S_j = \{A_{1,j}, A_{2,j}, ..., A_{8,j}\}$, $M_j = \{A_{9,j}, A_{10,j}, \ldots, A_{13,j}\}$ and $B_j = \{A_{14,j}, A_{15,j}, \ldots, A_{21,j}\}$.*

We denote by L_{K_j}, $K_j \in \{S_j, M_j, B_j\}$, the system of convex fuzzy sets $B \underset{\sim}{\subsetneq} w_j$ of the form $B := \sum_{i \in J} A_{i,j}(x),\ x \in w_j$ where $\{A_{i,j}\}_{i \in J}$ is a subsystem of K_j and $\sum$ denotes a pointwise summation of fuzzy sets. By using specifying evaluative linguistic expressions and pure evaluative ones, we can obtain suitable linguistic expressions for each convex fuzzy set from L_{K_j}. The idea how to represent such sets linguistically is very simple and natural.

If we consider Example 7 we can demonstrate how to obtain suitable linguistic expressions for *small* values at first. For instance, $A_{1,j} \sim \mathit{Ex\ Sm}$ and $A_{2,j} \sim \mathit{Si\ Sm}$ `but not` $\mathit{Ex\ Sm}$. *Significantly small* values of a given attribute are expressed by $A_{1,j} + A_{2,j}$ where $+$ denotes a common (pointwise) addition of real functions. Since $A_{1,j}$ and $A_{2,j}$ are members of the fuzzy partition P_j, it is consistent with commonly used mathematical representation of evaluative linguistic expressions (see

Figure 2.1). Similarly, in Model I we can obtain

$$\begin{aligned}
A_{1,j} + A_{2,j} &\sim \textit{Si Sm}, \\
A_{2,j} + A_{3,j} &\sim \textit{Ve Sm} \text{ but not } \textit{Ex Sm}, \\
A_{1,j} + A_{2,j} + A_{3,j} &\sim \textit{Ve Sm}, \\
A_{1,j} + \ldots + A_{4,j} &\sim \textit{Sm}, \qquad (3.1.1) \\
A_{1,j} + \ldots + A_{5,j} &\sim \textit{ML Sm}, \\
A_{1,j} + \ldots + A_{6,j} &\sim \textit{Ro Sm}, \\
A_{1,j} + \ldots + A_{7,j} &\sim \textit{QR Sm}, \\
A_{1,j} + \ldots + A_{8,j} &\sim \textit{VR Sm}, \\
&\text{etc.}
\end{aligned}$$

The following Figure 3.1 demonstrates the mentioned construction.

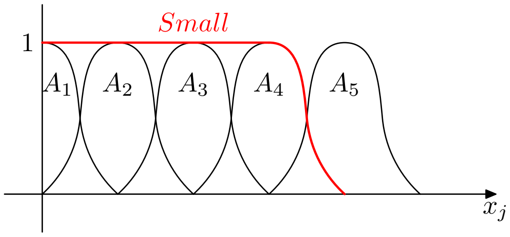

Figure 3.1: Scheme of the construction of the extension of the evaluative linguistic expression Small.

Similarly we obtain evaluative linguistic expressions for *big* values of the attribute X_j. The situation concerning *medium* values is different and, as we have indicated above, we use a little bit different evaluative expressions. For instance, in Model I we can obtain the following expressions:

$$
\begin{aligned}
A_{9,j} + A_{10,j} &\sim \mathit{ML}\ \mathit{Me}\ \texttt{but not}\ \mathit{ML}\ \mathit{Hi}\ \mathit{Me},\\
A_{10,j} + A_{11,j} &\sim \mathit{ML}\ \mathit{Lo}\ \mathit{Me}\ \texttt{but not}\ \mathit{FN1},\\
A_{11,j} + A_{12,j} &\sim \mathit{ML}\ \mathit{Hi}\ \mathit{Me}\ \texttt{but not}\ \mathit{FN2},\\
A_{12,j} + A_{13,j} &\sim \mathit{ML}\ \mathit{Hi}\ \mathit{Me}\ \texttt{but not}\ \mathit{Ex}\ \mathit{Me},\\
A_{9,j} + A_{10,j} + A_{11,j} &\sim \mathit{ML}\ \mathit{Lo}\ \mathit{Me},\\
A_{10,j} + A_{11,j} + A_{12,j} &\sim \mathit{Me},\\
A_{11,j} + A_{12,j} + A_{13,j} &\sim \mathit{ML}\ \mathit{Hi}\ \mathit{Me},\\
A_{9,j} + \ldots + A_{12,j} &\sim \mathit{ML}\ \mathit{Me}\ \texttt{but not}\ \mathit{FN2},\\
A_{10,j} + \ldots + A_{13,j} &\sim \mathit{ML}\ \mathit{Me}\ \texttt{but not}\ \mathit{FN1},\\
A_{9,j} + \ldots + A_{13,j} &\sim \mathit{ML}\ \mathit{Me}.
\end{aligned}
$$

It is also demonstrated on the following Figure 3.2

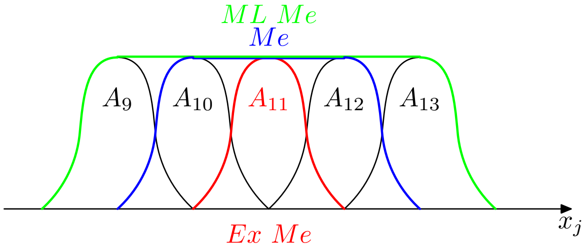

Figure 3.2: Extensions of evaluative expressions characterizing medium values.

Using of fuzzy partitions is often required in fuzzy data mining methods (c.f. [25]). Model I is suitable for such methods and, after using such methods, founded associations can be represented in natural language. When we use this model, it is necessary to keep in mind that there are some limitations. The main problem comes from mathematical representation of above mentioned linguistic expressions. When we use Model I (i.e., a fuzzy partition to decompose the context of any attribute), the linguistic connective `OR` is seemingly represented as the pointwise summation $((\mathit{Ex}\ \mathit{Sm})\ \texttt{OR}\ (\mathit{Si}\ \mathit{Sm}\ \texttt{but not}\ \mathit{Ex}\ \mathit{Sm}) \sim A_{1,j} + A_{2,j} \sim \mathit{Si}\ \mathit{Sm})$ which need not be suitable. In this case (if `OR` is represented by a t-conorm $\min(a+b,1)$) the conjunction `AND` is represented by Łukasiewicz conjunction. But this conjunction is quite restrictive if it is used for mining of linguistic associations ([23]), especially if many attributes are considered. Because if we consider Łukasiewicz conjunction $a \otimes b = 0 \vee (a+b-1)$ then we can see that value $a+b$ has to be higher than 1 to get non zero value, expecially if we consider more than two attributes.

To overcome this problem we have decided to choose another mathematical model (*Model II*). Thus, we use a fuzzy covering C_j to cover the context of a given attribute X_j. Shapes of fuzzy sets representing evaluative linguistic expressions are distinct from those used in Model I, but their names can be the same. For instance, Example 7 demonstrates evaluative linguistic expressions represented by 21 members of C_j.

Additionally, by the choice of C_j, we can use standard logical connectives `AND`, `OR` interpreted by $\wedge$ and $\vee$ to obtain additional evaluative linguistic expressions. Namely,

$$(Ex\ Sm)\ \texttt{OR}\ (Si\ Sm\ \texttt{but not}\ Ex\ Sm) \sim A_{1,j} \vee A_{2,j} \sim Si\ Sm.$$

On the following picture Figure 3.3 shapes of these fuzzy sets and their additional evaluative linguistic expression are shown

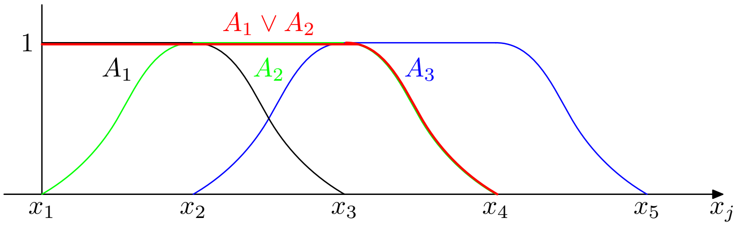

Figure 3.3: Shapes of fuzzy sets used in Model II.

It is obvious that the shape of the fuzzy set $A_{1,j} \vee A_{2,j}$ is commonly used for a representation of the evaluative linguistic expression *significantly small* (see Figure 2.1). Thus, similarly as for P_j we consider a decomposition of C_j into S_j, M_j and B_j and we specify evaluative linguistic expressions separately for fuzzy sets from S_j, M_j and B_j, i.e., *small*, *medium* and *big* values of a given context. For *small* (and similarly for *big*) values we obtain

$$\begin{aligned}
A_{1,j} \vee A_{2,j} &\sim \mathit{Si}\ \mathit{Sm}, \\
A_{2,j} \vee A_{3,j} &\sim \mathit{Ve}\,\mathit{Sm} \text{ but not } \mathit{Ex}\,\mathit{Sm}, \\
A_{1,j} \vee A_{2,j} \vee A_{3,j} &\sim \mathit{Ve}\ \mathit{Sm}, \\
A_{1,j} \vee \ldots \vee A_{4,j} &\sim \mathit{Sm}, \\
A_{1,j} \vee \ldots \vee A_{5,j} &\sim \mathit{ML}\ \mathit{Sm}, \\
A_{1,j} \vee \ldots \vee A_{6,j} &\sim \mathit{Ro}\ \mathit{Sm}, \\
A_{1,j} \vee \ldots \vee A_{7,j} &\sim \mathit{QR}\ \mathit{Sm}, \\
A_{1,j} \vee \ldots \vee A_{8,j} &\sim \mathit{VR}\ \mathit{Sm}, \\
&\text{etc.}
\end{aligned}$$

Analogously we obtain evaluative linguistic expressions for *medium* values

$$\begin{aligned}
A_{9,j} \vee A_{10,j} &\sim \mathit{ML}\ \mathit{Me}\ \texttt{but not}\ \mathit{ML}\ \mathit{Hi}\ \mathit{Me}, \\
A_{10,j} \vee A_{11,j} &\sim \mathit{ML}\ \mathit{Lo}\ \mathit{Me}\ \texttt{but not}\ \mathit{FN1}, \\
A_{11,j} \vee A_{12,j} &\sim \mathit{ML}\ \mathit{Hi}\ \mathit{Me}\ \texttt{but not}\ \mathit{FN2}, \\
A_{12,j} \vee A_{13,j} &\sim \mathit{ML}\ \mathit{Hi}\ \mathit{Me}\ \texttt{but not}\ \mathit{Ex}\ \mathit{Me}, \\
A_{9,j} \vee A_{10,j} \vee A_{11,j} &\sim \mathit{ML}\ \mathit{Lo}\ \mathit{Me}, \\
A_{10,j} \vee A_{11,j} \vee A_{12,j} &\sim \mathit{Me}, \\
A_{11,j} \vee A_{12,j} \vee A_{13,j} &\sim \mathit{ML}\ \mathit{Hi}\ \mathit{Me}, \\
A_{9,j} \vee \ldots \vee A_{12,j} &\sim \mathit{ML}\ \mathit{Me}\ \texttt{but not}\ \mathit{FN2}, \\
A_{10,j} \vee \ldots \vee A_{13,j} &\sim \mathit{ML}\ \mathit{Me}\ \texttt{but not}\ \mathit{FN1}, \\
A_{9,j} \vee \ldots \vee A_{13,j} &\sim \mathit{ML}\ \mathit{Me}.
\end{aligned}$$

Ordering of evaluative linguistic expressions and predications is described on page 26. In the following example we suppose an ordering of evaluative linguistic expressions defined in Example 7.

Example 8 *For* $C \sim \mathit{Si}\ \mathit{Sm}\ \texttt{but not}\ \mathit{Ex}\ \mathit{Sm}$ *and* $D \sim \mathit{Me}$ *we obtain* $C \lessapprox D$ *by* (2.2.3) *since the first (resp. the second) linguistic expression contains the expression* small *(resp.* medium*). But if* $C \sim \mathit{Si}\ \mathit{Sm}\ \texttt{but not}\ \mathit{Ex}\ \mathit{Sm}$, $D \sim \mathit{Sm}$, *then these sets cannot be ordered by* (2.2.3) *since both mentioned expressions contain the expression* small. *By (*(2.2.2)*) we get* $C \lessapprox D$. *If* $C \sim \mathit{Si}\ \mathit{Sm}\ \texttt{but not}\ \mathit{Ex}\ \mathit{Sm}$, $D \sim \mathit{Sm}\ \texttt{but not}\ \mathit{Si}\ \mathit{Sm}$, *then these sets cannot be ordered.*

3.2 Mining linguistic associations

In this section we present one of methods for searching for linguistic associations. The process presented in this section can be applied to both Models I and II. However we will continue only with Model II since it forms a natural extension of the model published in [23] and naturally follows ideas of the GUHA method (e.g., see [10] and references therein). In addition to this, using of Model I is analogous. As we mentioned in the preceding section, we search for linguistic associations in the data set of the form

$$\begin{array}{c|cccc} & X_1 & X_2 & \dots & X_k \\ \hline o_1 & a_{11} & a_{12} & \dots & a_{1k} \\ o_2 & a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ o_m & a_{m1} & a_{m2} & \dots & a_{mk} \end{array} .$$

The first step of this method consists of replacing numerical values a_{ij} in the data set by the most suitable evaluative linguistic expressions (i.e., by fuzzy sets representing them). This is done by function $Perc$ that is defined separately for the context w_j of each attribute X_j. Thus, we obtain k components $Perc_j : R \times w_j \longrightarrow L(C)$, $j = 1, 2, \dots, k$, of the function $Perc$ that are defined by the user of the data mining process and $L(C)$ is the set of evaluative linguistic expressions. Usually, the most specific and informative evaluative linguistic predications are assigned to $a_{ij} \in w_j$, $i = 1, 2, \dots, m$.

For instance, functions $Perc_j$ can be given by $Perc_j(a_{ij}, w_j) = \tilde{A}_{i,j} := A_{l,j}$ if $A_{l,j} \in C_j$ is the unique fuzzy set from C_j satisfying $A_k(a_{ij}) = 1$. Otherwise, we choose $A_{l,j}$ from $\{A \in C_j \,|\, A(a_{ij}) = 1\}$ that represents evaluative linguistic predications of the most narrow sense (see (2.2.1)).

In [23], the function $Perc_0$ assigns to each context w_j and to each element a_{ij} an evaluating expression with intension $\tilde{A}_{i,j}$, where $\tilde{A}_{i,j}$ is the most specific (sharpest) in the sense of the natural ordering $\lessapprox$ defined on page 26, and $a_{ij} \in w_j$ is typical in the extension $A_{l,j}$ in given context. To be typical means that the membership degree $A_{l,j}(a_{ij})$ is greater than some reasonable threshold α (we usually put $\alpha = 0,9$ or even $\alpha = 1$).

Thus, we obtain a transformed data set of the form

$$\begin{array}{c|cccc} & X_1 & X_2 & \dots & X_k \\ \hline o_1 & A_{1,1} & A_{1,2} & \dots & A_{1,k} \\ o_2 & A_{2,1} & A_{2,2} & \dots & A_{2,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ o_m & A_{m,1} & A_{m,2} & \dots & A_{m,k} \end{array},$$

where A_{ij} denote evaluative linguistic predications (see page 25). Now we can look for dependencies between given disjoint sets of attributes $\{Y_o\}_{o=1}^{q}$, $\{Z_p\}_{p=1}^{r} \subseteq \{X_j\}_{j=1}^{k}$. We search for simpler *linguistic associations* of the form

$$E(\{Y_o\}_{o=1}^{q}) \Rightarrow F(\{Z_p\}_{p=1}^{r}), \tag{3.2.1}$$

where E, F are compound evaluative predications containing only the connective `AND`. For simplicity we will write only $E \Rightarrow F$ instead of (3.2.1). An example of possible linguistic representation of this association can be: *"IF the area of the base of a cylinder is big AND the height of this cylinder is also big THEN the volume of this cylinder is big."* The left and right side of (3.2.1) is called the *antecedent* and *consequent*, respectively.

A symbol $\Rightarrow$ represents an implication, resp. a relationship between an antecedent and a consequent. This can be described by the so-called *quantifier* (see [11]). The quatifier $\Rightarrow$ in (3.2.1) characterizes validity of the association. By $\Rightarrow (a, b, c, d)$ we denote the fact that the four fold table quantifier is computed from the four fold table with components a, b, c, d. There are many quantifiers that can be computed in such a way. For a comprehensive list of commonly used four-fold table quantifiers we refer to [17] where these quantifiers are compared mutually by several user oriented criteria. We would like also to point out to paper [26] where several classes of four fold table quantifiers were studied from the logical point of view. Probably the most often used quantifiers are the following

- $\Rightarrow := \Rightarrow_x$ a *symmetric associational quantifier.* This quantifier is valid if $ad > bc$.

- $\Rightarrow := \sqsubset_r^{\gamma}$ a *binary multitudinal quantifier*. This quantifier is taken as true if $a > \gamma(a+b)$ and $\frac{a}{m} > r$, where $\gamma \in [0,1]$ is a *confidence degree* and $r \in [0,1]$ is a *support degree*,

where parameters a, b, c and d are elements from the standard four-fold table (more detail in [26, 11]).

	F	`not` F
E	a	b
`not` E	c	d

For completeness, a is a number of positive occurrences of E as well as F - that is, a number of objects (rows) such that attributes $\{Y_o\}_{o=1}^{q}$ are evaluated via the function $Perc_j$ by the same linguistic expression E and, at the same time, the attributes $\{Z_p\}_{p=1}^{r}$ are evaluated by the respective linguistic expression F. The numbers b is a number of positive occurrences of E and negative occurrences of F. c is number of negative occurrences of E and positive occurrences of F and d is a number of negative occurrences of E as well as F.

Example 9 *We use linguistic expressions from Example 7 in all attributes. If we look for an association* IF the attribute $\bar{X}_1$ is small but not very small AND $\bar{X}_2$ is extremely big THEN Y_1 is higher medium *where* X_2, X_3, X_6 *are attributes considered in data set and* $\bar{X}_1 = X_3$, $\bar{X}_2 = X_2$ *and* $Y_1 = X_6$. *The positive occurrence of this association in the row* o_j *means that* $Perc_3(a_{j,3}) = A_{j_3,3}$, $Perc_2(a_{j,2}) = A_{j_2,2}$ *and* $Perc_6(a_{j,6}) = A_{j_6,6}$ *where* $A_{j_3,3}$, $A_{j_2,2}$ *and* $A_{j_6,6}$ *are fuzzy sets of the covering* C_j *representing linguistic expression* small but not very small*,* extremely big *and* higher medium *in contexts* w_3, w_2 *and* w_6*, respectively. Analogously,* b *is a number of positive occurrences of* E *but negative occurrences of* F*,* c *is a number of negative occurrences of* E *and positive occurrences of* F *and* d *is a number of negative occurrences of* E *and negative occurrences of* F*.*

In the rest of this section we present some limitations of the method presented in [23]. First, using only common linguistic hedges (*more or less*, *significantly* etc.) could cause some problems in the data mining process ([24, p.7]). For instance, when we mine linguistic associations that contain linguistic expression *more or less medium* but do not contain linguistic expression *medium*, we are not able to distinguish whether founded linguistic associations deal with transactions having either rather smaller or rather bigger values.

Analogous problem appears when we intend to present mined associations. When presenting associations dealing with values that are *more or less medium* `but not` *medium* the reader should obtain more specified knowledge on rather smaller or rather bigger values. Our representation allows this.

Further, there are some facts that is necessary to reflect during the presentation of founded associations. If the method (using fuzzy sets on Figure 2.1, the function $Perc$ and some four-fold table quantifiers) presented in this chapter is used, we can obtain associations describing *small* values. It is necessary to keep in mind that such associations deal with values that are *small* but are not, for instance, *very small* in the classical meaning. Using fuzzy coverings or decompositions together with specifying linguistic expressions overcomes this limitation.

3.3 Reduction of rules

We meet with the problem that there exist large number of mined hypothesis (associations). For further work with them and for clarity it is better to reduce the set of hypothesis.

The ordering $\lessapprox$ of linguistic predications together with specific properties of a chosen quantifier $\Rightarrow$ can be used for a reduction of rules as it is suggested in [23], [26] and [10]. A reduction of rules must be clear enough from the point of view of the user of the data mining procedure and is used, for instance, to simplify the output of the data mining procedure (to reduce the number of presented linguistic associations) or to decrease the number of actually tested associations.

Definition 16 *Let A, B, C, D be linguistic predications and $\Rightarrow$ be a quantifier. We denote by*

$$(A \Rightarrow B) \ \vdash \ (C \Rightarrow D)$$

that the association $(C \Rightarrow D)$ follows *from $(A \Rightarrow B)$. This means that the validity of $(A \Rightarrow B)$ implies the validity of $(C \Rightarrow D)$.*

The quantifier $\Rightarrow$ (see page 37) characterizes validity (truth) of the association in the data. In the case of a symmetric associational quantifier ($\Rightarrow := \Rightarrow_x$) taken as true if $ad > bc$. Or in case of a binary multitudinal quantifier ($\Rightarrow := \sqsubset_r^\gamma$), where $\gamma,\ r \in [0;1]$. It taken as true, if $a > \gamma(a+b)$ and $\frac{a}{m} > r$.

The next statement was inspired by a theorem proved in [23] for $\Rightarrow = \sqsubset_r^\gamma$.

Theorem 1 *Let A, B, C be linguistic predications and $\Rightarrow$ be an implicational quantifier. Then*

(i) If $B \lessapprox B'$ then $(A \Rightarrow B) \vdash (A \Rightarrow B')$,

(ii) $(A \Rightarrow B) \vdash (A \Rightarrow B\ \texttt{OR C})$.

PROOF:

(i) Let a, b, c, d (resp. a', b', c', d') be elements of the four fold table of the association $A \Rightarrow B$ (resp. $A \Rightarrow B'$). By the assumption of the theorem, $\Rightarrow (a, b, c, d)$ is valid.

Since $B \lessapprox B'$, we detone a set of all objects having the property $A(\{Y_o\}_{o=1}^q)$ as M_A and similarly M_B, $M_{A\ \texttt{AND}\ B}$, etc. Their cardinality is denoted by $|\cdot|$. Then we obtain

$$M_{A\ \texttt{AND}\ B'} = M_{A\ \texttt{AND}\ B} \subset (M_{A\ \texttt{AND}\ B'}, M_{A\ \texttt{AND}\ B}),$$

$$M_{A \texttt{ AND } \neg B} = M_{A \texttt{ AND } \neg B'} \subset (M_{A \texttt{ AND } B'}, M_{A \texttt{ AND } B}).$$

Hence $a = |M_{A \texttt{ AND } B}| \leq a' = |M_{A \texttt{ AND } B'}|$ and $b' = |M_{A \texttt{ AND } \neg B'}| \leq b = |M_{A \texttt{ AND } \neg B}|$.

Thus, by the definition of the implicational quantifier $r \leq a \leq a'$ and $\gamma < \frac{a}{a+b} \leq \frac{a'}{a'+b'}$, then $\Rightarrow (a', b', c', d')$ is also valid.

(ii) It easily follows from (i) and from the choice of the ordering of linguistic associations $\precapprox$.

□

We can apply Theorem 1 to reduction on the following artificial associations:

Example 10 *We assume data dealing with the measurement of Number of cars, Temperature, Wind and Concentration of NO2. From data we can obtain following associations in the form of IF-THEN rules.*

(i) *IF the Number of cars is big but not very big AND the Temperature is upper medium AND the Wind is more or less small but not small THEN the Concentration of NO2 is medium.*

(ii) *IF the Number of cars is big but not very big AND the Temperature is upper medium AND the Wind is more or less small but not small THEN the Concentration of NO2 is more or less medium.*

We can see that IF-THEN rules (i) and (ii) are similar, the different is in the consequent. From Definition 14 we know that $Me \precapprox ML\ Me$. It follows from Theorem 1 that the linguistic association (ii) logically follows from linguistic association (i). Thus, only the association (i) will be stated in the set of presented linguistic associations.

3.4 Testing

At the beginning of this section we introduce data we used for our experiment. The original model elaborated in [23] and Model II are used for mining of linguistic associations.

The chosen data were downloaded from a public web page*). The data are a subsample of 500 observations from a data set that originate from a study where air

*) http://lib.stat.cmu.edu/

pollution at a road is related to traffic volume and meteorological variables, collected by the Norwegian Public Roads Administration. In the original data set there are 8 columns of attributes for any observation. The response variable (Y_NO2) consists hourly values of the logarithm of the concentration of NO2, measured at Alnabru in Oslo, Norway, between October 2001 and August 2003. The data description is following (see Tab. 3.1).

Attribute	Context	Measurement	Description
Y_ NO2	[1.2, 6.2]	particles	Concentration of NO2
NoCar	[4.1, 8.2]		Number of cars per hour
Temp	[-18.6, 19.3]	degree C	Temperature
Wind	[0.3, 9.5]	meters/second	Wind
TempDiff	[-5.4, 3.9]	degree C	Temperature difference
WindDir	[2, 343]	degrees between [0,360]	Wind direction
Hour	[1, 23]		Hour of day
DayNumb	[32, 582]		Day number

Table 3.1: Data description.

We searched dependencies of concentration of NO2 on NoCar, Temp and Wind. A software utility LAM (Linguistic Associations Mining)[†)] was developed in our institute by A. Dvořák, H. Habiballa, V. Novák, I. Perfilieva and V. Pavliska. It is used for searching for linguistic associations or associations using fuzzy numbers and F-transform in numerical data. First, user have to specify input - a table of numerical data and variables. Second, various parameters can be set, e.g., parameters r and γ, type of quantifier, range of evaluative linguistic expressions used, etc. At first we applied the original model and then Model II.

In both models, we replaced numerical data by appropriate evaluative expressions according to their meaning. Then, in the original model, the function $Perc_0$ (see page 36) assigns to each context w_j and to each element $a_{ij} \in w_j$ one of linguistic expressions *Ve Sm*, *Sm*, *ML Sm*, *Me*, *ML Me*, *ML Bi*, *Bi*, and *Ve Bi*. The fuzzy sets corresponding with these linguistic expressions are elaborated in [23] (see also Figure 2.1).

Second, we searched dependencies between attributes by using a binary multitudinal quantifier. The software LAM analyzed the data set, tested 343 hypotheses and found 39 linguistic associations satisfying inequalities of binary multitudinal quantifier with parameters $\gamma = 0.2$, $r = 0.005$ (see page 37).

[†)] For more information look at http://irafm.osu.cz/

Similarly we proceeded when we looked for linguistic associations by using Model II. We replaced numerical data using the function $Perc$ by the following evaluative linguistic expressions (see Figure 3.4) used in Model II.

$$\begin{aligned}
&A_{1,j} \sim \mathit{Ve\ Sm},\\
&A_{2,j} \sim \mathit{Sm} \text{ but not } \mathit{Ve\ Sm},\\
&A_{3,j} \sim \mathit{ML\,Sm} \text{ but not } \mathit{Sm},\\
&A_{4,j} \sim \mathit{Lo\ Me},\\
&A_{5,j} \sim \mathit{Me},\\
&A_{6,j} \sim \mathit{Up\ Me},\\
&A_{7,j} \sim \mathit{ML\,Bi} \text{ but not } \mathit{Bi},\\
&A_{8,j} \sim \mathit{Bi} \text{ but not } \mathit{Ve\ Bi},\\
&A_{9,j} \sim \mathit{Ve\ Bi}.
\end{aligned}$$

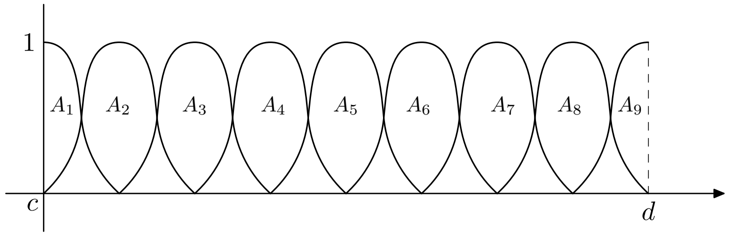

Figure 3.4: The most simple nontrivial mathematical model with fuzzy sets of Model II.

It is possible to construct evaluative linguistic expressions used also in the original model (see Figure 2.1)

$$\begin{gathered}
A_{1,j} \vee A_{2,j} \sim \mathit{Sm},\\
A_{1,j} \vee A_{2,j} \vee A_{3,j} \sim \mathit{ML\ Sm},\\
A_{4,j} \vee A_{5,j} \vee A_{6,j} \sim \mathit{ML\ Me},\\
A_{7,j} \vee A_{8,j} \vee A_{9,j} \sim \mathit{ML\ Bi},\\
A_{8,j} \vee A_{9,j} \sim \mathit{Bi}.
\end{gathered}$$

Software LAM tested 398 hypotheses and found 22 associations. More hypotheses were tested because more fuzzy sets covering the context of each attribute were used. In the rest of this section we explain relations between linguistic associations found by the original model and Model II, respectively.

Now we show six associations found by the original model and Model II.

The following three linguistic associations were found by the original model:

1. IF the NoCar is *big* AND the Temp is *more or less medium* AND the Wind is *more or less small* THEN the Concentration of NO2 is *more or less big.*

2. IF the NoCar is *big* AND the Temp is *more or less small* AND the Wind is *more or less medium* THEN the Concentration of NO2 is *more or less medium.*

3. IF the NoCar is *very big* AND the Temp is *more or less medium* AND the Wind is *more or less small* THEN the Concentration of NO2 is *more or less medium.*

The following three linguistic associations were found by Model II:

$1'$. IF the NoCar is *big but not very big* AND the Temp is *upper medium* AND the Wind is *more or less small but not small* THEN the Concentration of NO2 is *more or less big but not big.*

$2'$. IF the NoCar is *big but not very big* AND the Temp is *more or less small but not small* AND the Wind is *lower medium* THEN the Concentration of NO2 is *upper medium.*

$3'$. IF the NoCar is *very big* AND the Temp is *lower medium* AND the Wind is *more or less small but not small* THEN the Concentration of NO2 is *upper medium.*

Since the same parameters ($\gamma = 0.2$, $r = 0.005$) were used in both models for each association obtained by Model II we can find the corresponding association obtained by the original model. Such pairs of associations ($1, 1'$), ($2, 2'$), ($3, 3'$) represent the same knowledge from the point of view of the chosen quantifier. We emphasize that different definitions of $Perc$ and $Perc_0$ are used in our models.

In the original model, the function $Perc_0$ (its definition is in [23]) assigns to each context w_j and to each element $a_{ij} \in w_j$ one of linguistic expressions *Ve Sm*, *Sm*, *ML Sm*, *Me*, *ML Sm*, *ML Bi*, *Bi*, *Ve Bi* (see Fig. 2.2). This linguistic expression is the most specific (sharpest) in the sense of the ordering $\preceq$ (see Page 26).

They provide the same (crisp) decomposition of considered contexts but intervals of this decomposition represent different linguistic variables. For instance, *big* values in the original model are the same as *big but not very big* values in Model II etc.

Consequently associations 1 and $1'$ (2 and $2'$ etc.) represent the same dependence (knowledge). It is easy to see that the way of presentation of associations found by Model II is longer but intuitively they represent more precise knowledge. For example, if we consider the context $[0, 100]$ in the original model, the function $Perc_0$ assigns the intervals $(27, 39)$ and $(42, 57]$ (resp. $[39, 42]$) to the linguistic expression *ML Me* (resp. *Me*). This does not correspond to common shape of the fuzzy set representing the linguistic expression *ML Me* (see Figure 2.1) - hence the user expects that the association containing the linguistic expression *ML Me* gives the information about interval $(27, 57]$.

From this point of view, *more or less medium* values of a given attribute do not contain *medium* ones in the original model. By using Model II we are able to specify whether we deal with *lower medium* or *higher medium* values. Consequently, they present more precise knowledge.

As we demonstrated above, for 22 linguistic associations obtained by Model II we can find 22 corresponding associations obtained by the original model. The remaining linguistic associations found by the original model cannot be found by the method presented in this thesis. By using different definitions for $Perc$ and $Perc_0$, we have the following situation – *more or less medium* values of the original model are represented by both *lower medium* and *higher medium* values of Model II. The following example demonstrates that such associations need not be found directly by Model II.

Example 11 *If we look for associations by the original model we can obtain the linguistic association containing the linguistic expression ML Me for a certain attribute. Let elements of four fold table of this association be* $a = 12,\ b = 24$*. Then the inequalities* $a > \gamma(a+b)$*,* $\frac{a}{m} > r$ *are satisfied for* $\gamma = 0.2$*,* $r = 0,005$*. As we could see above, this linguistic association can be expressed in Model II by the linguistic associations containing the linguistic expression Lo Me (resp. Hi Me) with* $a = 10$*,* $b = 90$ *(resp.* $a = 2$*,* $b = 10$*). But these elements of four fold table do not satisfy the inequality* $a > \gamma(a + b)$ *of the multitudinal quantifier with the parameter* $\gamma = 0.2$*.*

We emphasize that the existence of the linguistic expression *ML Me* in the original model (see 1 and $1'$ etc.) does not imply that the relevant linguistic associations containing expressions *Lo Me* and *Hi Me* are not found by Model II. If we want to obtain the rest of linguistic associations we need to use more sophisticated method.

3.5 Conclusion

In this chapter we have demonstrated that fuzzy partitions or coverings can be used for mining of linguistic associations in data sets. We have also introduced linguistic expressions (the specifying ones) that allow this. Advantages of our method are the following - it provides more accurate knowledge to the user, it naturally extends the method recently developed in [23] and that it extends applicability of mining of linguistic associations since many other methods (based on fuzzy partitions or coverings) can be used for such mining. Additionally, at the end of this chapter, we have indicated one of possible ways of our future research, namely Theorem 1.

Chapter 4

Properties induced by fuzzy confirmation measures

Fuzzy associations (3.2.1) are evaluated using appropriate confirmation measures. There are several ways how to choose confirmation measures that determine linguistic associations. One of the best known methods for searching linguistic associations is GUHA method ([11]). Its confirmation measures (called quantifiers) are computed from relevant four-fold tables. To construct such tables crisp partitions (induced by relevant fuzzy sets) of contexts of considered attributes are required ([11], [23] etc.). However, there are also other possibilities due to which we can work directly with fuzzy sets carrying linguistic labels.

For instance, in [6], the problem of choosing reasonable fuzzy confirmation measures was studied systematically and choices of various confirmation measures were justified especially in connection with a certain and very natural partition of the row set $\mathcal{D}_o$ (see page 29). The partition of $\mathcal{D}_o$ is given by fuzzy sets $S_+, S_-, S_\pm$ and the condition

$$S_+(o_i) + S_-(o_i) + S_\pm(o_i) = 1, \text{ for any } o_i \in \mathcal{D}_o, \tag{4.0.1}$$

where $S_+(o_i), S_-(o_i), S_\pm(o_i)$ denotes a *positive*, *negative* and *irrelevant* evaluation, respectively, of each row $o_i \in \mathcal{D}_o$ of a given rule (3.2.1). Such a partition of $\mathcal{D}_o$ can be of the form

$$\begin{aligned} S_+(o_i) &:= E(o_i) \otimes F(o_i), \\ S_-(o_i) &:= 1 - (E(o_i) \to F(o_i)), \\ S_\pm(o_i) &:= 1 - E(o_i) \end{aligned} \tag{4.0.2}$$

where $E(o_i)$ (resp. $F(o_i)$) means a membership degree of o_i in the fuzzy set E (resp. F) representing antecedent (resp. succedent). A t–norm $\otimes$ is so-called *copula* (see page 20). From conjunction (4.0.1) and (4.0.2) an implication operator → is given by $a \to b \ = \ (1-a)+(a\otimes b)$. Under these assumptions, partition (4.0.2) guarantees (4.0.1) for any possible rule (resp. association) of the form (3.2.1). Additionally, the authors of [6] justified how such partition induces meaningful fuzzy confirmation measures.

Definition 17 *For partition* (4.0.2)*, we introduce the following (*t–norm-based*) support measure of* (3.2.1) *in the data set* $\mathcal{D}$

$$supp_t(E \Rightarrow F) := \sum_{o_i \in \mathcal{D}_o} E(o_i) \otimes F(o_i). \tag{4.0.3}$$

Remark 1 *However, the problem in [6] can be further specified. For instance, when we require the implication operator → to be a self implication (i.e., $a \to a = 1$), the authors obtained that ⊗ has to be the t–norm ($x \otimes y = \min\{x, y\}$), or, when we require → to be the strong implication (i.e., $a \to b = n(a) \otimes b$ for a strong negation $n(a)$) the solution is given only by the product t–norm ($x \otimes y = xy$). For completeness we note that the partition is again given by (4.0.2).*

The authors of [13] discussed also *gradual* (resp. *certainty*) rules. Such rules are of the form "The more the property E is true, the more the conclusion F is true". In that case, another definition of partition of $\mathcal{D}_o$ for association $E \Rightarrow F$ was considered

$$\begin{aligned} S_+(o_i) &:= E(o_i) \otimes (E(o_i) \to F(o_i)), \\ S_-(o_i) &:= E(o_i) \otimes (1-(E(o_i) \to F(o_i))), \\ S_\pm(o_i) &:= 1 - E(o_i). \end{aligned} \tag{4.0.4}$$

For this partition, condition (4.0.1) is satisfied only for the product t–norm.

Definition 18 *The (*implication-based*) support measure is given for (3.2.1)*

$$supp_c(E \Rightarrow F) := \sum_{o_i \in \mathcal{D}_o} E(o_i) \cdot (E(o_i) \to F(o_i)), \tag{4.0.5}$$

where → represents any generalized implication*.*

Definition 19 *An* implication operator $\mathcal{I} : I \times I \longrightarrow I$ *for* $I = [0, 1]$ *is a generalization of the material implication if it satisfies, for* $x, y, x_0, y_0 \in I$*,*

(I1) $\mathcal{I}(x, y) \leq \mathcal{I}(x_0, y)$ *for* $x_0 \leq x$*,*

(I2) $\mathcal{I}(x, y) \leq \mathcal{I}(x, y_0)$ *for* $y \leq y_0$*,*

(I3) $\mathcal{I}(1, y) = y$*,*

(I4) $\mathcal{I}(0, 0) = 1$.

In order to keep preceding notation we put $\to := \mathcal{I}$. Sometimes, $x \to y$ denotes also a *product implication* that is equal

$$x \to y = \begin{cases} 1, & x \leq y, \\ \frac{y}{x}, & x > y. \end{cases}$$

Definition 20 *Let $\otimes$ is a continuous t–norm and $\to$ is derived from that t–norm through residuation. Then (*minimum-based*) support measure is given by*

$$supp_m(E \Rightarrow F) := \sum_{o_i \in \mathcal{D}_o} \min\{E(o_i), F(o_i)\}. \tag{4.0.6}$$

Definition 21 *For support measures (4.0.3), (4.0.5) and (4.0.6), relevant* confidence *measures are defined by*

$$conf_j(E \Rightarrow F) := \frac{supp_j(E \Rightarrow F)}{\sum_{o_i \in \mathcal{D}_o} E(o_i)} \tag{4.0.7}$$

for $j \in \{t, c, m\}$.

Note that (4.0.7) cannot be strictly greater than 1 for any association $E \Rightarrow F$.

Definition 22 *Let r and γ are given support and confidence thresholds. We say that the rule $E \Rightarrow F$ is* valid *if $supp_j(E \Rightarrow F) \geq r$ and $conf_j(E \Rightarrow F) \geq \gamma$.*

Further, for given rules $E_1 \Rightarrow F_1$ and $E_2 \Rightarrow F_2$, $E_1 \Rightarrow F_1 \vdash_s E_2 \Rightarrow F_2$ denotes the fact that $supp(E_1 \Rightarrow F_1) \leq supp(E_2 \Rightarrow F_2)$. Similarly, $E_1 \Rightarrow F_1 \vdash_c E_2 \Rightarrow F_2$ denotes the fact that $conf(E_1 \Rightarrow F_1) \leq conf(E_2 \Rightarrow F_2)$ and finally $E_1 \Rightarrow F_1 \vdash E_2 \Rightarrow F_2$ means that $E_1 \Rightarrow F_1 \vdash_j E_2 \Rightarrow F_2$ for $j = \{s, c\}$. Analogous notation we can also use for sets of associations - the expression

$$A \Rightarrow B,\ C \Rightarrow D \vdash E \Rightarrow F$$

means that $E \Rightarrow F$ is valid, i.e., either $supp(A \Rightarrow B) < supp(E \Rightarrow F)$ and $conf(A \Rightarrow B) < conf(E \Rightarrow F)$ or $supp(C \Rightarrow D) < supp(E \Rightarrow F)$ and $conf(C \Rightarrow D) < conf(E \Rightarrow F)$.

Let us remark that associations of the form $C \Rightarrow C$ are valid. A considering such associations is very natural, their confidence degree has to be equal to 1. Consequently, the validity of this association implies that the linguistic expression represented by C has a sufficiently large support.

The following lemma will be used several times in this chapter. It describes rather natural property but we have put it to this chapter for the sake of completeness.

We say that data sets $\mathcal{D}_l, l = 1, 2, \ldots, r$, are of the same type if they have the same attributes $X_i, i = 1, 2, \ldots, k$, possessing the same contexts $w_i, i = 1, 2, \ldots, k$, and fuzzy sets evaluating various linguistic terms of attributes $X_i, i = 1, 2, \ldots, k$, are also the same. Let m_l denote the number of objects in each $D_l, l = 1, 2, \ldots, r$.

For such data sets we can define another data set $\mathcal{D} := \odot_{l=1}^{r} \mathcal{D}_l$ called direct join by joining data tables $\mathcal{D}_l$ to the unique one. Then, for example, for $\mathcal{D}_1$ and $\mathcal{D}_2$; $\mathcal{D}$ has $m = m_1 + m_2$ objects, the first m_1 objects of $\mathcal{D}$ comes from $\mathcal{D}_1$, the following m_2 objects of $\mathcal{D}$ comes from $\mathcal{D}_2$, etc. The following lemma claims that the validity of a given rule in each particular data set ensures the validity of the rule in the direct join of such data sets. Let us stress that this lemma is independent of the choice of support measure.

Lemma 1 *Let $\mathcal{D}_l, l = 1, 2, \ldots, r$, be data sets of the same type and let a rule $(A \Rightarrow B)$ is valid in each $\mathcal{D}_l$. Then $(A \Rightarrow B)$ is also valid in $\mathcal{D} := \odot_{l=1}^{r} \mathcal{D}_l$.*

PROOF: Consider any support measure $supp_j(A \Rightarrow B)$ and the confidence measure $conf_j(A \Rightarrow B)$ for $j \in \{t, c, m\}$. By $(\mathcal{D}_o)_l$ we denote rows of $\mathcal{D}$ coming from $\mathcal{D}_i$. Then we use $supp^l(A \Rightarrow B)$ and $conf^l(A \Rightarrow B)$ to denote that the confidence measures are counted just by using rows $(\mathcal{D}_o)_l$. According to our assumptions, for given confidence degree γ and support degree r, we have

$$supp^l(A \Rightarrow B) = \sum_{o_i \in (D_o)_l} A(o_i) \otimes B(o_i) = r_l \geq r \tag{4.0.8}$$

and

$$conf^l(A \Rightarrow B) = \frac{\sum_{o_i \in (D_o)_l} A(o_i) \otimes B(o_i)}{\sum_{o_i \in (D_o)_l} A(o_i)} = \frac{r_l}{\sum_{o_i \in (D_o)_l} A(o_i)} = \gamma_l \geq \gamma \tag{4.0.9}$$

for any $l = 1, 2, \ldots, m$. To simplify the proof put $\mathcal{A}_l = \sum_{o_i \in (D_o)_l} A(o_i)$. By the definition of $\mathcal{D}$ and (4.0.8) we immediately have

$$supp(A \Rightarrow B) = \sum_{l=1}^{r} supp^l(A \Rightarrow B) = \sum_{l=1}^{r} r_l \geq r \cdot l \geq r. \tag{4.0.10}$$

As regards the confidence degree, (4.0.9) implies that $\gamma \cdot \mathcal{A}_l \leq r_l$ for any $l = 1, 2, \ldots, r$ and hence

$$\sum_{l=1}^{r} \gamma \cdot \mathcal{A}_l = \gamma \sum_{l=1}^{r} \cdot \mathcal{A}_l \leq \sum_{l=1}^{r} r_l$$

for any $l = 1, 2, \ldots, r$. Consequently, by the choice of $\mathcal{D}, r_l$ and $\mathcal{A}_l$,

$$conf(A \Rightarrow B) = \frac{\sum_{l=1}^{r} r_l}{\sum_{l=1}^{r} \mathcal{A}_l} \geq \gamma.$$

This and (4.0.10) finishes this proof. □

Remark 2 *We can also use the last lemma in the following way - in order to check the validity of the rule* $(A \Rightarrow B)$ *in* $\mathcal{D}$ *it is sufficient to decompose the data set* $\mathcal{D}$ *to smaller data sets* $\mathcal{D}_i$ *and to check the validity of* $(A \Rightarrow B)$ *in each particular* $\mathcal{D}_i$*.*

4.1 Using additional knowledge

Let us introduce a set $\mathcal{E}$ of associations (i.e., the set of *expert knowledge*) that can be provided to the data mining process. We would like to emphasize that linguistic and mathematical representation is the same for associations from $\mathcal{E}$ as well as for associations we want to mine in a given data set.

Note that we need not to specify the inner structure of such expert associations (i.e., associations from $\mathcal{E}$). For a given unknown association $E \Rightarrow F$ we would like to test, associations from $\mathcal{E}$ (associations from $\mathcal{E}$ are denoted by a symbol *, i.e., $(A \Rightarrow^* B) \in \mathcal{E}$) can describe information between the antecedent and succedent part of $E \Rightarrow F$ as well as between attributes contained either in the antecedent or succedent part of $E \Rightarrow F$, respectively. We would like to stress that it makes sense to deal only with confidence measures of associations from $\mathcal{E}$.

We would like to emphasize that, within this chapter, associations from $\mathcal{E}$ are assumed to be *fully* valid in the dataset $\mathcal{D}_o$ (see page 29), i.e. we assume $conf_i(A \Rightarrow^* B) = 1$. According to the choice of a confidence measure, we can obtain some additional information.

Definition 23 For t–norm-based confidence measures (resp. for the minimum-based)

$$conf_i(A \Rightarrow^* B) = 1$$

where $(A \Rightarrow^* B) \in \mathcal{E}$ and $A(o_i) \leq B(o_i)$, $o_i \in \mathcal{D}_o$.

Remark 3 *When an implication-based confidence measure is considered, we can obtain the same condition provided* $\rightarrow$ *is a residuated implication of some* t*–norm. But if* $\rightarrow$ *is a generalized implication then only* $B(o_i) = 1$ *for any* $o_i \in \mathcal{D}_o$ *can be assumed.*

4.2 Properties

We study the following properties where A, B, C and D are evaluative linguistic predications and the relation $\vdash$ means that if the association on the left-hand side is true in the data set then the association on the right-hand side is necessarily also true on the basis of support and confidence degree. In this subsection we analyse general assumptions. On the base of these assumptions it is possible to claim, the validation of associations on the left side implies validation of association on the right side. In this way we can obtain further valid associations. Vice versa it is possible to reduce set of associations without loss of information on the base of such properties or to use properties for faster process of verification of associations.

P1 $(A \texttt{ OR } B) \Rightarrow A$,
P2 $A \Rightarrow B,\ (B \texttt{ OR } C) \Rightarrow D \vdash (A \texttt{ OR } C) \Rightarrow D$,
P3 $A \Rightarrow B \vdash (C \texttt{ AND } A) \Rightarrow (C \texttt{ AND } B)$,
P4 $(A \Rightarrow B),\ (A \Rightarrow C) \vdash (A \Rightarrow (B \texttt{ OR } C))$,
P5 $A \Rightarrow (B \texttt{ OR } C) \vdash A \Rightarrow B$,
P6 $A \Rightarrow B,\ B \Rightarrow C \vdash A \Rightarrow C$,
P7 $A \Rightarrow B,\ C \Rightarrow D \vdash (A \texttt{ OR } C) \Rightarrow (B \texttt{ OR } D)$,
P8 $(A \texttt{ AND } B) \Rightarrow (C \texttt{ AND } D) \vdash (A \texttt{ AND } B \texttt{ AND } D) \Rightarrow C$.

In [2] the authors study axioms and inference rules used in database design. It should be also mentioned that the same rules are valid also in fuzzy attribute logic elaborated e.g. in [3]. These axioms and inference rules are described by Properties P1-P6.

The last two properties - Properties P7 and P8 are motivated by analogous properties that are used e.g. in GUHA method ([11]) or in the classic Apriori algorithm (see [1] and references therein). Below, we will discuss all properties P1 - P8.

Properties P1 and P2

As regards Property P1, it has been explained in [16] that this property need not be satisfied in general.

The symbol $\vdash_s$ (resp. $\vdash_c$) means that formulae on the right sight is valid with support degree (resp. confidence degree) higher or equal than formulae on the left side.

Thus let us study

$$A \Rightarrow B,\ (B \texttt{ OR } C) \Rightarrow D \vdash (A \texttt{ OR } C) \Rightarrow D.$$

We explained in [16] that, for t–norm-based measures,

$$A \Rightarrow B,\ (B \texttt{ OR } C) \Rightarrow D \not\vdash_s (A \texttt{ OR } C) \Rightarrow D$$

and also

$$A \Rightarrow B,\ (B\ \texttt{OR}\ C) \Rightarrow D \nvdash_c (A\ \texttt{OR}\ C) \Rightarrow D$$

where the relation $\nvdash_s$ (resp. $\nvdash_c$) means that this rule is not valid for a given threshold of support measure (resp. confidence measure).

The same negative results we obtain also for the minimum-based and implication-based confirmation measures - see the next two simple examples.

Example 12 *Consider a dataset with three objects and fuzzy sets A, B, C, D with values:*

	A	*B*	*C*	*D*
o_1	*0.8*	*1*	*0.4*	*1*
o_2	*0.7*	*0.8*	*0.4*	*0.1*
o_3	*0.6*	*0.7*	*0.8*	*0.4*

Then we immediately obtain $supp_m(A \Rightarrow B) = 2.1 > supp_m((B\,\texttt{OR}\,C) \Rightarrow D) = 1.5 > supp_m((A\,\texttt{OR}\,C) \Rightarrow D) = 1.3$. Moreover, $conf_m(A \Rightarrow B) = 1 > conf_m((B\,\texttt{OR}\,C) \Rightarrow D) = 0.57 > conf_m((A\,\texttt{OR}\,C) \Rightarrow D) = 0.56$. Additionally, this example contradicts

$$A \Rightarrow^* B,\ (B\ \texttt{OR}\ C) \Rightarrow D \vdash (A\ \texttt{OR}\ C) \Rightarrow D.$$

Example 13 *Consider a dataset with three rows and fuzzy sets A, B, C, D with values:*

	A	*B*	*C*	*D*
o_1	*0.2*	*1*	*0.8*	*0.2*
o_2	*0.1*	*0.8*	*1*	*0.1*
o_3	*0.7*	*0.8*	*0.8*	*0.2*

Then, for $x \to y := \max\{1-x, y\}$, we get $supp_c(A \Rightarrow B) = 0.85 > supp_c((B\,\texttt{OR}\,C) \Rightarrow D) = 0.49 > supp_c((A\,\texttt{OR}\,C) \Rightarrow D) = 0.46$. Moreover, we obtain $conf_c(A \Rightarrow B) = 0.85 > conf_c(B\,\texttt{OR}\,C) \Rightarrow D) = 0.17 > conf_c((A\,\texttt{OR}\,C) \Rightarrow D) = 0.16$, i.e., it contradicts also

$$A \Rightarrow^* B,\ (B\ \texttt{OR}\ C) \Rightarrow D \vdash (A\ \texttt{OR}\ C) \Rightarrow D.$$

Analogously it could be demonstrated that using $A \Rightarrow^* B$ need not be justified for general t–norm-based confirmation measures.

Property P3

Lemma 2 *([16])Consider a dataset $\mathcal{D}_o$ such that $B(o_i) < C(o_i) < A(o_i)$ is satisfied for no row o_i. Then, for minimum–based confidence measure $conf_m$, we have*

$$A \Rightarrow B \vdash_c (C\ \texttt{AND}\ A) \Rightarrow (C\ \texttt{AND}\ B).$$

PROOF: By our assumptions, for given confidence restraint γ, we have

$$conf_m(A \Rightarrow B) = \frac{\sum_{o_i \in \mathcal{D}_o} min\{A(o_i), B(o_i)\}}{\sum_{o_i \in \mathcal{D}_o} A(o_i)} \tag{4.2.1}$$

is greater than or equal to γ. We want to prove

$$conf_m((C \texttt{ AND } A) \Rightarrow (C \texttt{ AND } B)) = \frac{\sum_{o_i \in \mathcal{D}_o} min\{A(o_i), B(o_i), C(o_i)\}}{\sum_{o_i \in \mathcal{D}_o} min\{A(o_i), C(o_i)\}} \tag{4.2.2}$$

cannot be smaller than (4.2.1). According to Lemma 1 (see also Remark 2), we may decompose $\mathcal{D}_o$ into four disjoint subdatasets according to subsequent row inequalities

(i) $C(o_i) < A(o_i), B(o_i)$,

(ii) $A(o_i) \leq B(o_i) \leq C(o_i)$,

(iii) $B(o_i) < A(o_i) \leq C(o_i)$,

(iv)) $A(o_i) \leq C(o_i) < B(o_i)$.

Because the cases (i), (ii) and (iv) lead to (4.2.2) = 1, we obtained that (4.2.1) is smaller than or equal to (4.2.2) on these subdatasets. It remains to explain the case (iii). But then (4.2.2) = $conf(A \Rightarrow B)$ and this concludes this proof. □

To ensure the validity of this rule it is necessary to require more.

Corollary 1 *For minimum–based confidence degree we have*

$$A \Rightarrow^* B \vdash_c (C \texttt{ AND } A) \Rightarrow (C \texttt{ AND } B).$$

In the rest of this subsection we deal with an exact modification of Property P3 that can hold for support measures (4.0.6), (4.0.3) and (4.0.5). However, it can be shown that it does not hold for confidence measures (Example 14).

Lemma 3 *Let us consider the t–norm-based support measure given by* (4.0.3). *Then*

$$(A \texttt{ AND } C) \Rightarrow (B \texttt{ AND } C) \vdash_s A \Rightarrow B.$$

PROOF: By the assumption, we have $supp_t((A \texttt{ AND } C) \Rightarrow (B \texttt{ AND } C)) = \sum_{o_i \in D_o} A(o_i) \otimes C(o_i) \otimes B(o_i) \otimes C(o_i)$ and it is smaller than $supp_t(A \Rightarrow B) = \sum_{o_i \in D_o} A(o_i) \otimes B(o_i)$ and therefore $supp_t((A \texttt{ AND } C) \Rightarrow (B \texttt{ AND } C)) \leq supp_t(A \Rightarrow B)$. □

From this lemma the next corollary easily follows.

Corollary 2 *Let us consider the minimum-based support measure given by* (4.0.6)*. Then*

$$(A \texttt{ AND } C) \Rightarrow (B \texttt{ AND } C) \vdash_s A \Rightarrow B.$$

Lemma 4 *Let us consider the implication-based support measure given by* (4.0.5) *and the product implication. Then*

$$(A \texttt{ AND } C) \Rightarrow (B \texttt{ AND } C) \vdash_s A \Rightarrow B.$$

PROOF: We consider the implication-based support measure (4.0.5) with the product implication $\rightarrow$. Put $D_1 := \{o_i \in \mathcal{D}_o \mid A(o_i) \leq B(o_i)\}$ and $D_2 := \mathcal{D}_o \setminus D_1$ and consider two expressions:

$$A(o_i) \cdot (A(o_i) \rightarrow B(o_i)) \tag{4.2.3}$$

and

$$(A(o_i)C(o_i)) \cdot (A(o_i)C(o_i) \rightarrow B(o_i)C(o_i)). \tag{4.2.4}$$

For any $o_i \in D_1$ we easily obtain

$$A(o_i)C(o_i) = (4.2.4) \leq (4.2.3) = A(o_i). \tag{4.2.5}$$

Consequently, since $supp_c$ of $(A \texttt{ AND } C) \Rightarrow (B \texttt{ AND } C)$ and $A \Rightarrow B$ is counted as the sum of (4.2.4) and (4.2.3), respectively, we get that (4.2.5) implies

$$(A \texttt{ AND } C) \Rightarrow (B \texttt{ AND } C) \vdash_s A \Rightarrow B \tag{4.2.6}$$

on the set D_1.

Analogously, we can use an analogous argument for the set D_2, since we clearly obtain

$$B(o_i)C(o_i) = (4.2.4) \leq (4.2.3) = B(o_i)$$

for any $o_i \in D_2$. □

In the following example it is shown that this rule can not be proved for confidence measures.

Example 14 *We consider a dataset consisting of one object with the following values of fuzzy sets* A, B, C*:* $A(o_1) = 0.9$*,* $B(o_1) = 0.5$*,* $C(o_1) = 0.1$*. Then* $conf_m((A \texttt{ AND } C) \Rightarrow (B \texttt{ AND } C)) = 1$ *and it is greater than* $conf_m(A \Rightarrow B) = 5/9$*.*

Or when we consider a dataset consisting of two objects with values of fuzzy sets A, B, C*:* $A(o_1) = 0.9$*,* $B(o_1) = 0.5$*,* $C(o_1) = 0.1$*,* $A(o_2) = 0.6$*,* $B(o_2) = 0.8$*,* $C(o_2) = 0.2$*. Then* $conf_c((A \texttt{ AND } C) \Rightarrow (B \texttt{ AND } C)) = 0.8$ *and it is greater than* $conf_c(A \Rightarrow B) = 0.73$*.*

Remark 4 *Here and in the subsequent counterexamples we consider the product implication for implication-based confirmation measures - i.e., the residuated implication induced by the product t–norm.*

Property P4

Let us study the rule

$$A \Rightarrow B,\ A \Rightarrow C \vdash A \Rightarrow (B\ \texttt{OR}\ C).$$

As we can see from the following lemma, the validity of this property is straightforward.

Lemma 5 (P4) *([16])Let us consider confirmation measures given by* (4.0.3), (4.0.5), (4.0.6) *and* (4.0.7)*. Then*

$$A \Rightarrow B,\ A \Rightarrow C \vdash A \Rightarrow (B\ \texttt{OR}\ C).$$

The proof of this lemma is based on the proof of Theorem 1 (see page 39).

Remark 5 *Clerly, since Property P*4 *is valid in general, it can be used also in connection with the expert knowledge (i.e, associations from $\mathcal{E}$) we consider in our task. Thus, e.g.,*

$$A \Rightarrow^* B,\ A \Rightarrow C \vdash A \Rightarrow (B\ \texttt{OR}\ C).$$

Property P5

In this subsection we focus on the property

$$A \Rightarrow (B\ \texttt{OR}\ C) \vdash A \Rightarrow B.$$

In [16] some examples demonstrate that this property does not hold in general for various confirmation measures.

According to examples from [16] we can claim, for all confirmation measures considered in this paper, that

$$A \Rightarrow (B\ \texttt{OR}\ C) \not\vdash_s A \Rightarrow B.$$

and

$$A \Rightarrow (B\ \texttt{OR}\ C) \not\vdash_c A \Rightarrow B.$$

However, it can be seen from the subsequent lemma we can specify some additional assumptions in order to ensure the validity of Property P5.

Lemma 6 (P5) *Let us consider the minimum-based confirmation measures* (4.0.6) *and* (4.0.7)*. Then*

$$A \Rightarrow (B\ \texttt{OR}\ C), C \Rightarrow^* B \vdash A \Rightarrow B.$$

PROOF: By our assumptions we have, for support and confidence thresholds r and γ,

$$supp_m(A \Rightarrow B \texttt{ OR } C) := \sum_{o_i \in \mathcal{D}_o} \min\{A(o_i), \max\{B(o_i), C(o_i)\}\} \geq r \tag{4.2.7}$$

and

$$conf_m(A \Rightarrow B \texttt{ OR } C) := \frac{\sum_{o_i \in \mathcal{D}_o} \min\{A(o_i), \max\{B(o_i), C(o_i)\}\}}{\sum_{o_i \in \mathcal{D}_o} A(o_i)} \geq \gamma. \tag{4.2.8}$$

We want to prove

$$supp_m(A \Rightarrow B) := \sum_{o_i \in \mathcal{D}_o} \min\{A(o_i), B(o_i)\} \geq r \tag{4.2.9}$$

and

$$conf_m(A \Rightarrow B) := \frac{\sum_{o_i \in \mathcal{D}_o} \min\{A(o_i), B(o_i)\}}{\sum_{o_i \in \mathcal{D}_o} A(o_i)} \geq \gamma. \tag{4.2.10}$$

According to Section 4.1, $C \Rightarrow^* B$ implies that $C(o_i) \leq B(o_i)$ for any $o_i \in \mathcal{D}_o$. Having this in mind, it is clear that (4.2.7) is equal to (4.2.9) and also (4.2.8) is equal to (4.2.10). □

The following counterexample shows that the lemma above can not be constructed for t–norm-based and implication-based confirmation measures.

Example 15 *Consider dataset with one object and fuzzy sets A, B, C with values - $A(o_1) = 0.5,\ B(o_1) = 0.1,\ C(o_1) = 0.9$.*

Then $supp_t(A \Rightarrow (B \texttt{ OR } C)) = 0.455$, resp. $conf_t(A \Rightarrow (B \texttt{ OR } C)) = 0.91$, is strictly greater than $supp_t(A \Rightarrow B) = 0.05$, resp. $conf_t(A \Rightarrow B) = 0.01$.

Further, $supp_c(A \Rightarrow (B \texttt{ OR } C)) = 0.5$, resp. $conf_c(A \Rightarrow (B \texttt{ OR } C)) = 1$, is strictly greater than $supp_c(A \Rightarrow B) = 0.1$, resp. $conf_c(A \Rightarrow B) = 0.2$.

Property P6

In this subsection we consider the property

$$A \Rightarrow B,\ B \Rightarrow C \vdash A \Rightarrow C. \tag{4.2.11}$$

The authors of [16] demonstrated in counterexamples that this property is not valid in general in the set of mined associations, thus

$$A \Rightarrow B,\ B \Rightarrow C \not\vdash_s A \Rightarrow C$$

and

$$A \Rightarrow B,\ B \Rightarrow C \nvdash_c A \Rightarrow C \tag{4.2.12}$$

for all support measures (4.0.3), (4.0.5) and (4.0.6), respectively. Additionally, there are the examples demonstrating that requiring some additional assumptions (for example $A \Rightarrow A,\ B \Rightarrow B,\ C \Rightarrow C$) need not lead to the validity of Property P6.

For completeness, we can prove a lemma claiming that by using some expert knowledge we can reasonably use Property P6.

Lemma 7 *([16]) Let us consider confirmation measures given by* (4.0.7) *and* (4.0.3), (4.0.5), (4.0.6)*. Then*

$$A \Rightarrow B,\ B \Rightarrow^* C \vdash A \Rightarrow C.$$

PROOF: First we consider t–norm-based confirmation measures. By our assumptions we have, for support and confidence thresholds r and γ,

$$supp_t(A \Rightarrow B) = \sum_{o_i \in \mathcal{D}_o} A(o_i) \otimes B(o_i) \geq r$$

and

$$conf_t(A \Rightarrow B) = \frac{\sum_{o_i \in \mathcal{D}_o} A(o_i) \otimes B(o_i)}{\sum_{o_i \in \mathcal{D}_o} A(o_i)} \geq \gamma.$$

We want to prove

$$supp_t(A \Rightarrow C) = \sum_{o_i \in \mathcal{D}_o} A(o_i) \otimes C(o_i) \geq r$$

and

$$conf_t(A \Rightarrow C) = \frac{\sum_{o_i \in \mathcal{D}_o} A(o_i) \otimes C(o_i)}{\sum_{o_i \in \mathcal{D}_o} A(o_i)} \geq \gamma.$$

According to Section 4.1, $B \Rightarrow^* C$ implies that $B(o_i) \leq C(o_i)$ for any $o_i \in \mathcal{D}_o$. Having this in mind, it is clear that

$$supp_t(A \Rightarrow B) = \sum_{o_i \in \mathcal{D}_o} A(o_i) \otimes B(o_i) \leq supp_t(A \Rightarrow C) = \sum_{o_i \in \mathcal{D}_o} A(o_i) \otimes C(o_i)$$

and also

$$conf_t(A \Rightarrow B) = \frac{\sum_{o_i \in \mathcal{D}_o} A(o_i) \otimes B(o_i)}{\sum_{o_i \in \mathcal{D}_o} A(o_i)} \leq conf_t(A \Rightarrow C) = \frac{\sum_{o_i \in \mathcal{D}_o} A(o_i) \otimes C(o_i)}{\sum_{o_i \in \mathcal{D}_o} A(o_i)}.$$

This finishes the proof for t–norm-based confirmation measures. For minimum-based confirmation measures the proof would be analogous. When we consider implication-based confirmation measures (4.0.5) and (4.0.7), we can analogously use conditions $B(o_i) \leq C(o_i)$ or $B(o_i) = 1$ for any $o_i \in \mathcal{D}_o$. By using the monotonicity of the ordinary product and a chosen implication operator, we immediately obtain the required ordering for the support measure. Finally, it is trivial to finish the proof for the implication-based confidence measures. □

Remark 6 *As an easy corollary of Lemma 7 we can see that an ordinary transitivity $(A \Rightarrow^* B,\ B \Rightarrow^* C \vdash A \Rightarrow^* C)$ is preserved in the set $\mathcal{E}$. On the other side, the property $(A \Rightarrow^* B,\ B \Rightarrow C \vdash A \Rightarrow C)$ need not be valid in general.*

Lemma 8 *For measures* (4.0.3), (4.0.5), (4.0.6) *and relevant* (4.0.7)

$$A \Rightarrow B',\ B \Rightarrow^* C \vdash A \Rightarrow C,\ \text{whenever } B' \preceq B.$$

PROOF: We consider t–norm-based support measure (4.0.3). We obtain $supp_t(A \Rightarrow B') = \sum_{o \in D_o} A(o) \otimes B'(o)$. $B \Rightarrow^* C$ implies that $B(o) \leq C(o)$ for any $o \in \mathcal{D}_o$. From $B' \preceq B$ it holds $B'(o) \leq B(o) \leq C(o)$. Consequently, we obtain directly from (4.0.3) and (4.0.7) that $supp_t(A \Rightarrow B') \leq supp_t(A \Rightarrow C)$ (resp. $conf_t(A \Rightarrow B') \leq conf_t(A \Rightarrow C)$).

As minimum-based confirmation measures are a special case of t–norm-based ones, it remains to finish this proof for implication-based confirmation measures (4.0.6). We obtain $supp_c(A \Rightarrow B') = \sum_{o \in D_o} A(o)(A(o) \rightarrow B'(o))$. As above we have $B'(o) \leq B(o) \leq C(o)$. Consequently, we obtain from (4.0.6) and (4.0.7) that $supp_c(A \Rightarrow B') \leq supp_c(A \Rightarrow C)$ (resp. $conf_c(A \Rightarrow B') \leq conf_c(A \Rightarrow C)$). □

Property P7

We can return to the original motivation ([6]) of establishing confirmation measures (4.0.7), (4.0.3), (4.0.5) and (4.0.6). We use the partition of $\mathcal{D}_o$ given by fuzzy sets $S_+, S_-, S_\pm$, i.e. a *positive*, *negative* and *irrelevant* part of the rule $E \Rightarrow F$ (notation $S_+(E \Rightarrow F)$, $S_-(E \Rightarrow F)$ and $S_\pm(E \Rightarrow F)$), respectively. Note that each $S_i(E \Rightarrow F)$, $i \in \{+, -, \pm\}$, can be seen as a fuzzy set on $\mathcal{D}_o$. In [6] confirmation measures (4.0.7), (4.0.3), (4.0.5) and (4.0.6) were established in order to satisfy

$$supp(E \Rightarrow F) = \sum_{o_i \in \mathcal{D}_o} S_+(E \Rightarrow F)(o_i)$$

and

$$conf(E \Rightarrow F) = \frac{\sum_{o_i \in \mathcal{D}_o} S_+(E \Rightarrow F)(o_i)}{\sum_{o_i \in \mathcal{D}_o} (S_+(E \Rightarrow F)(o_i) + S_-(E \Rightarrow F)(o_i))}$$

for given partitions satisfying (4.0.1).

It is easy to see from the last two expressions that having two valid associations $E_1 \Rightarrow F_1$, $E_2 \Rightarrow F_2$ with "disjoint" positive parts ensures the validity of $(E_1\, \texttt{OR}\, E_2) \Rightarrow (F_1\, \texttt{OR}\, F_2)$ whenever the connective `OR` is represented by a pointwise maximum.

Therefore, we can work with a rule

$$A \Rightarrow B,\ C \Rightarrow D \vdash (A \texttt{ OR } C) \Rightarrow (B \texttt{ OR } D)$$

for fuzzy sets A, C with disjoint supports. Generally, the following results for P7 and its special case $(C \Rightarrow C)$ can be provided.

Lemma 9 (P7)*([16]) Let us consider the t–norm-based support measure, resp. minimum-based support measure given by* (4.0.5)*, resp.* (4.0.6)*. Then*

$$A \Rightarrow B,\ C \Rightarrow D \vdash_s (A \texttt{ OR } C) \Rightarrow (B \texttt{ OR } D),$$

PROOF: From the definition of t–norm-based support measure it is easy to see that

$$\begin{aligned} \mathit{supp}_t(A \Rightarrow B) = \sum_{o_i \in \mathcal{D}_o} A(o_i) \otimes B(o_i) \leq \\ \leq \sum_{o_i \in \mathcal{D}_o} (A \texttt{ OR } C) \otimes (B \texttt{ OR } D) = \mathit{supp}_t((A \texttt{ OR } C) \Rightarrow (B \texttt{ OR } D)). \end{aligned}$$

For minimum-based support measure the proof would be analogous. □

Corollary 3 *([16]) For t–norm-based support measure, resp. minimum-based support measure we have*

$$A \Rightarrow B,\ C \Rightarrow C \vdash_s (C \texttt{ OR } A) \Rightarrow (C \texttt{ OR } B).$$

Lemma 10 (P7) *Let us consider the implication-based support measure given by* (4.0.5)*. Then*

$$A \Rightarrow B,\ C \Rightarrow D \vdash_s (A \texttt{ OR } C) \Rightarrow (B \texttt{ OR } D) \tag{4.2.13}$$

PROOF: Let us consider the implication-based support measure (4.0.5) with the product implication → and the following decomposition of $\mathcal{D}_o$ into $\mathcal{D}'_1, \mathcal{D}''_1, D''_2, D'_{21}$ and D'_{22} – $\mathcal{D}_1 := \{o_i \in \mathcal{D}_o \,|\, C(o_i) \leq D(o_i)\}$, $\mathcal{D}_2 := \mathcal{D}_o \setminus D_1$ and $\mathcal{D}'_1 := \{o_i \in \mathcal{D}_1 \,|\, A(o_i) \oplus C(o_i) \leq B(o_i) \oplus D(o_i)\}$, $\mathcal{D}''_1 := \mathcal{D}_1 \setminus \mathcal{D}'_1$, $\mathcal{D}'_2 := \{o_i \in \mathcal{D}_2 \,|\, A(o_i) \oplus C(o_i) \leq B(o_i) \oplus D(o_i)\}$ and $\mathcal{D}''_2 := \mathcal{D}_2 \setminus \mathcal{D}'_2$ and finally $\mathcal{D}'_{21} := \{o_i \in \mathcal{D}'_2 \,|\, A(o_i) \leq B(o_i)\}$ and $\mathcal{D}'_{22} := D'_2 \setminus \mathcal{D}'_{21}$. Finally, by $\oplus$ we denote a t–conorm of the product t–norm.

Let us study expressions

$$A(o_i) \cdot (A(o_i) \rightarrow B(o_i)), \tag{4.2.14}$$

$$C(o_i) \cdot (C(o_i) \rightarrow D(o_i)), \tag{4.2.15}$$

$$(A(o_i) \oplus C(o_i)) \cdot (A(o_i) \oplus C(o_i) \to B(o_i) \oplus D(o_i)). \qquad (4.2.16)$$

Then $(A(o_i) \oplus C(o_i)) \cdot (A(o_i) \oplus C(o_i) \to B(o_i) \oplus D(o_i)) = A(o_i) \oplus C(o_i)$ (or $B(o_i) \oplus D(o_i)$) on set $\mathcal{D}'_1$ (or $\mathcal{D}''_1$). In both cases we have

$$C(o_i) \cdot (C(o_i) \to D(o_i)) \leq (A(o_i) \oplus C(o_i)) \cdot (A(o_i) \oplus C(o_i) \to B(o_i) \oplus D(o_i)) \quad (4.2.17)$$

since $C(o_i) \cdot (C(o_i) \to D(o_i)) = C(o_i)$ for any $o_i \in \mathcal{D}_1$. Similarly, (4.2.17) holds also on $\mathcal{D}''_2$ because, for $o_i \in \mathcal{D}''_2$

$$\begin{aligned} C(o_i) \cdot (C(o_i) \to D(o_i)) = D(o_i) \leq B(o_i) \oplus D(o_i) < A(o_i) \oplus C(o_i) = \\ = (A(o_i) \oplus C(o_i)) \cdot (A(o_i) \oplus C(o_i) \to B(o_i) \oplus D(o_i)). \end{aligned}$$

Analogously, for any $o_i \in D'_{21}$

$$\begin{aligned} &A(o_i) \cdot (A(o_i) \to B(o_i)) = \\ &\quad = A(o_i) \leq A(o_i) \oplus C(o_i) = (A(o_i) \oplus C(o_i)) \cdot (A(o_i) \oplus C(o_i) \to B(o_i) \oplus D(o_i)), \end{aligned}$$

and for any $o_i \in D'_{22}$

$$\begin{aligned} A(o_i) \cdot (A(o_i) \to B(o_i)) = B(o_i) < A(o_i) \leq A(o_i) \oplus C(o_i) = \\ = (A(o_i) \oplus C(o_i)) \cdot (A(o_i) \oplus C(o_i) \to B(o_i) \oplus D(o_i)). \end{aligned}$$

Consequently, $A(o_i) \cdot (A(o_i) \to B(o_i)) \leq$ (4.2.16) holds for any $o_i \in \mathcal{D}'_2$. Since (4.2.17) holds for any $o_i \in (\mathcal{D}_o \setminus \mathcal{D}'_2) = \mathcal{D}_1 \cup \mathcal{D}''_2$, we obtain (4.2.13) directly from the definition of (4.0.5). □

From Lemma 10 we get the following corollary on the special case of Property P7.

Corollary 4 *([16]) For the implication-based support measure with the product implication we have*

$$A \Rightarrow B,\ C \Rightarrow C \vdash_s (C \texttt{ OR } A) \Rightarrow (C \texttt{ OR } B).$$

As regards the confidence measures, the following example demonstrates that Property P7 need not be proved for the minimum-based confidence measure. But it can be proven for the special case of Property P7 in the next lemma.

Example 16 *Consider minimum-based confidence measure and take a dataset consisting of three rows. Let fuzzy sets A, B, C, D be defined by $A(o_1) = 0.9$, $B(o_1) = C(o_1) = D(o_1) = 0.1$, $A(o_2) = B(o_2) = C(o_2) = D(o_2) = 0.9$, $A(o_3) = B(o_3) = D(o_3) = 0.1$ and $C(o_3) = 0.9$.*

Then $conf_m(A \Rightarrow B) = conf_m(C \Rightarrow D) = 11/19$ and this expression is greater then $conf_m(A \texttt{ OR } C \Rightarrow B \texttt{ OR } D) = 11/27$.

Lemma 11 (P7) *([16]) Let us consider the minimum-based confidence measure given by* (4.0.7). *Then*

$$A \Rightarrow B,\ C \Rightarrow C \vdash_c (C \texttt{ OR } A) \Rightarrow (C \texttt{ OR } B).$$

PROOF: By our assumptions we have,

$$conf_m(A \Rightarrow B) := \frac{\sum_{o_i \in \mathcal{D}_o} \min\{A(o_i), B(o_i)\}}{\sum_{o_i \in \mathcal{D}_o} A(o_i)}$$

and

$$conf_m(C \Rightarrow C) := \frac{\sum_{o_i \in \mathcal{D}_o} \min\{C(o_i), C(o_i)\}}{\sum_{o_i \in \mathcal{D}_o} C(o_i)}.$$

We want to prove

$$conf_m((C \texttt{ OR } A) \Rightarrow (C \texttt{ OR } B)) := \frac{\sum_{o_i \in \mathcal{D}_o} \min\{(C(o_i) \texttt{ OR } A(o_i)),\ (C(o_i) \texttt{ OR } B(o_i))\}}{\sum_{o_i \in \mathcal{D}_o} (C(o_i) \texttt{ OR } A(o_i))}.$$

- For $A(o_i) \leq B(o_i) \leq C(o_i)$ (resp. $B(o_i) \leq A(o_i) \leq C(o_i)$, or $A(o_i) \leq C(o_i) \leq B(o_i)$ or $C(o_i) \leq A(o_i) \leq B(o_i)$) where $o_i \in \mathcal{D}_o$ we have

$$\frac{\sum_{o_i \in \mathcal{D}_o} \min\{C(o_i), C(o_i)\}}{\sum_{o_i \in \mathcal{D}_o} C(o_i)} \leq \frac{\sum_{o_i \in \mathcal{D}_o} \min\{(C(o_i) \texttt{ OR } A(o_i)),\ (C(o_i) \texttt{ OR } B(o_i))\}}{\sum_{o_i \in \mathcal{D}_o} (C(o_i) \texttt{ OR } A(o_i))}.$$

- Otherwise, i.e., for $B(o_i) \leq C(o_i) \leq A(o_i)$ (resp. $C(o_i) \leq B(o_i) \leq A(o_i)$) where $o_i \in \mathcal{D}_o$ we have

$$\frac{\sum_{o_i \in \mathcal{D}_o} \min\{A(o_i), B(o_i)\}}{\sum_{o_i \in \mathcal{D}_o} A(o_i)} \leq \frac{\sum_{o_i \in \mathcal{D}_o} \min\{(C(o_i) \texttt{ OR } A(o_i)),\ (C(o_i) \texttt{ OR } B(o_i))\}}{\sum_{o_i \in \mathcal{D}_o} (C(o_i) \texttt{ OR } A(o_i))}.$$

□

Property P8

The last property is the condition

$$(A \texttt{ AND } B) \Rightarrow (C \texttt{ AND } D) \vdash (A \texttt{ AND } B \texttt{ AND } D) \Rightarrow C.$$

It can be easily proven that this property can be valid for t–norm-based confirmation measures, and hence also for the minimum-based ones.

Lemma 12 (P8) *([16]) Let us consider the t–norm-based confirmation measures given by* (4.0.3)*and* (4.0.7). *Then*

$$(A \texttt{ AND } B) \Rightarrow (C \texttt{ AND } D) \vdash (A \texttt{ AND } B \texttt{ AND } D) \Rightarrow C.$$

PROOF: Since the linguistic AND is represented by a given t–norm $\otimes$, it follows directly from the associativity of $\otimes$ that

$$(A(o_i) \otimes B(o_i)) \otimes (C(o_i) \otimes D(o_i)) = (A(o_i) \otimes B(o_i) \otimes D(o_i)) \otimes (C(o_i))$$

for any $o_i \in \mathcal{D}_o$. Hence, by the choice of $supp_t$, we immediately obtain

$$supp_t((A \texttt{ AND } B) \Rightarrow (C \texttt{ AND } D) = supp_t((A \texttt{ AND } B \texttt{ AND } D) \Rightarrow C).$$

Consequently also

$$conf_t((A \texttt{ AND } B) \Rightarrow (C \texttt{ AND } D) \leq conf_t((A \texttt{ AND } B \texttt{ AND } D \Rightarrow C)$$

since $A(o_i) \otimes B(o_i) \geq A(o_i) \otimes B(o_i) \otimes D(o_i)$ for each $o_i \in \mathcal{D}_o$. □

Corollary 5 (P8) *([16]) Let us consider the minimum-based confirmation measures given by* (4.0.6)*and* (4.0.7)*. Then*

$$(A \texttt{ AND } B) \Rightarrow (C \texttt{ AND } D) \vdash (A \texttt{ AND } B \texttt{ AND } D) \Rightarrow C.$$

Remark 7 *Since Property P8 is valid in general, it would be superfluous to study how to use associations from $\mathcal{E}$.*

However, for implication-based confirmation measures, the negative answer can be obtained - see the next example.

Example 17 *([16]) Consider data with attributes represented by fuzzy sets A, B, C, D having values $A(o_i) = B(o_i) = 0.1$ and $C(o_i) = D(o_i) = 0.2$.*

Then, for the product t–norm and its residuated implication, we obtain

$$supp_c((A \texttt{ AND } B) \Rightarrow (C \texttt{ AND } D)) = 0.01 \geq supp_c((A \texttt{ AND } B \texttt{ AND } D) \Rightarrow C) = 0.002.$$

4.3 Summary

In this chapter we sketch obtained results for particular confirmation measures.

For minimum-based confirmation measures we have demonstrated that some rules (P1, P2, P3, P5, P6, P7) are not valid in general. However, when we can modify some of them (P3) or specify some conditions (P7) or expert knowledge (P4 and P6) in order to guarantee the validity of the considered rule. Finally, P4 and P8 are always valid.

For t–norm-based confirmation measures we have got that Properties P1, P2, P3, P5, P6 and P7 are not valid in general. Similarly as above, we can specify some conditions (for P7) or some expert knowledge (for P4 and P6) in order to get their validity. As above, P4 and P8 are valid.

Finally we consider implication-based confirmation measures. For such measures, Properties P1, P2, P3, P5, P6 and P7 cannot be used in general. On the other side, P4 and P8 are always valid and for other rules some additional knowledge (for P4 and P6) or assumptions (for P7) can guarantee their validity.

4.4 Experiment

At the end of this chapter we devise a simple example demonstrating how the mentioned results can be used in the data mining process. We use a dataset entitled NO2 downloaded from the web page: http://lib.stat.cmu.edu/modules.php. For mining of associations we used the program LAMWin32[*)].

Our tools are the following (for details see [23])

- a model of evaluative linguistic expressions (more precisely, Model I),
- the implicational quantifier with parameters $r \geq 0.005$ and $\gamma \geq 0.2$.

The next two tables show some of found linguistic associations. The first two columns of Table 4.1 represent associations of the form "*Hour* $\Rightarrow$ *Temp*" and "*Temp* $\Rightarrow$ *Y_NO2*", etc.

IF	THEN	IF	THEN
Hour is	*Temp* is	*Temp* is	*Y_NO2* is
$ML\ Me$	$ML\ Me.$	$ML\ Me$	$ML\ Me.$
Me	$ML\ Sm.$	$ML\ Sm$	$ML\ Me.$
$Ve\ Sm$	$ML\ Sm.$	$ML\ Sm$	$ML\ Me.$

Table 4.1: Found linguistic associations via LAMWin32.

These tables demonstrate that, e.g., Property P6 is suitable for simplification of the data mining process. We can simplify the data mining process provided we have a suitable set $\mathcal{E}$ possessing associations from the right side of the first table. Then

[*)] For more information look at http://irafm.osu.cz/

IF	THEN
Hour is	*Y_NO2* is
$ML\ Me$	$ML\ Me.$
Me	$ML\ Me.$
$Ve\ Sm$	$ML\ Me.$

Table 4.2: New linguistic associations obtaining Property 6.

it would be sufficient to mine only for associations from the left side of that table.
For example, in the first rows we can see the associations
"IF *Hour* is $ML\ Me$ THEN *Temp* is $ML\ Me$."
"IF *Temp* is $ML\ Me$ THEN *Y_NO2* is $ML\ Me$."
Then we immediately obtain another association
"IF *Hour* is $ML\ Me$ THEN *Y_NO2* is $ML\ Me$".

Chapter 5

Modified APRIORI Algorithm

5.1 Introduction

There are more mathematical models of evaluative linguistic predications. Namely, the original one from [23] and the novel one [14] we work with here. Thus, linguistic predications can be represented by a fuzzy covering $\mathcal{P} := \{A_{i,j}\}$ of a chosen universe U (for details we refer to [14]). The latter model allows us to work with specifying evaluative linguistic expressions containing the formula "*but not*". In this chapter, we use only an example of evaluative linguistic predications with linguistic hedges *more or less* or *very* – see following Example 18 and Figure 5.1. The model from Example 18 is the most simple nontrivial mathematical model of evaluative linguistic predications. Note that later $|X_j|$ denotes a *number of fuzzy sets* $A_{i,j}$ from $\mathcal{P}_j$.

Example 18 *We consider the attribute (resp. property, variable) X_j. For simplicity we denote the attribute X_j without index and this model we can consider for all attributes X_j, $j = 1, \ldots, k$. The attribute X is given on certain interval $[c, d]$ whose covering $\mathcal{P}$ contains 9 fuzzy sets $\{A_i\}$, (i.e., $A_i : [c, d] \longrightarrow [0, 1]$ and $|X| = 9$)*

$$\begin{array}{lll} A_1 \sim Ve\ Sm, & A_2 \sim Sm \text{ but not } Ve\ Sm, & A_3 \sim ML\ Sm \text{ but not } Sm, \\ A_4 \sim Lo\ Me, & A_5 \sim Me, & A_6 \sim Hi\ Me, \\ A_7 \sim Ve\ Bi, & A_8 \sim Bi \text{ but not } Ve\ Bi, & A_9 \sim ML\ Bi \text{ but not } Bi. \end{array}$$

For medium values we introduce special linguistic hedges. Concepts Lo Me ("lower medium") and Hi Me ("higher medium") are more naturale.

For every attribute X we can define sets $P^1(X) := \{A_i\}$ and, for various indexes l

$$P^l(X) = \left\{A \in \mathcal{P}([c, d]) \,|\, A = \mathtt{OR}_{i=1}^{l} A_i, \text{ where } A_i \in P^1(X)\right\},$$

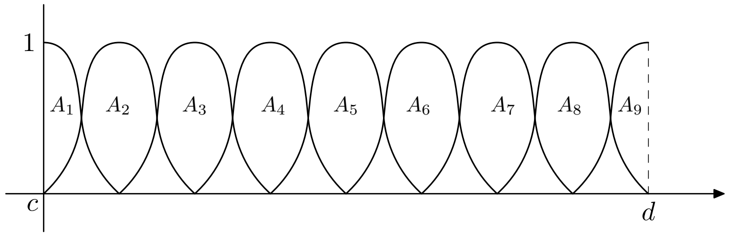

Figure 5.1: Fuzzy sets representing evaluative linguistic expressions in Example 18.

where OR can be represented by a relevant t–conorm (see page 20). Further, for every context we distinguish subsystems representing *small, medium* and *big* values, respectively, i.e., we can have $P^l_{Sm}(X), P^l_{Me}(X), P^l_{Bi}(X)$ for various indexes l. Finally, we put $\mathcal{P}(X) = \bigcup_l P^l(X)$.

Example 19 *For Example 18 it makes sense to consider only $l = 1, 2, 3$. Then $P^1_{Sm}(X) = \{A_1, A_2, A_3\}$, $P^1_{Me}(X) = \{A_4, A_5, A_6\}$ and $P^1_{Bi}(X) = \{A_7, A_8, A_9\}$, where each $A_i \in P(X)$ represents a suitable evaluative linguistic predication. For instance, for* small *values we have*

$$\begin{aligned} A_1 \texttt{ OR } A_2 &\sim Sm \in P^2_{Sm}(X), & A_1 &\sim Ve\ Sm,\\ A_2 \texttt{ OR } A_3 &\sim ML\ Sm \textit{ but not } Ve\ Sm \in P^2_{Sm}(X), & A_2 &\sim Sm \textit{ but not } Ve\ Sm,\\ A_1 \texttt{ OR } A_2 \texttt{ OR } A_3 &\sim ML\ Sm \in P^3_{Sm}(X), & A_3 &\sim ML\ Sm \textit{ but not } Sm. \end{aligned}$$

For medium *and* big *values we construct P^l_{Me}, P^l_{Bi} analogously (for details see [14]).*

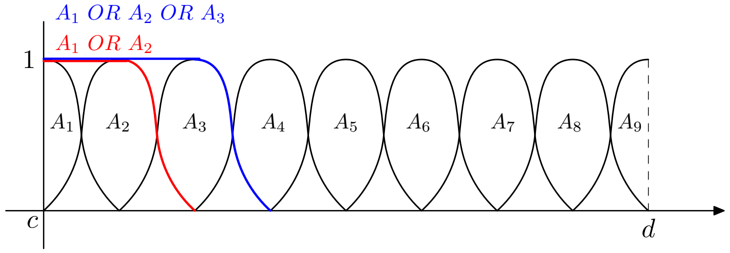

Figure 5.2: Subsystems representing small values in Example 18.

It is possible to construct either simpler or more comlex mathematical models than the one from Examples 18 and 19, but in this contribution we work only with this one as it is the most simple nontrivial mathematical model of evaluative linguistic predications.

A *specificity ordering* $\preceq$ is an ordering of fuzzy sets interpreting evaluative linguistic predications. We denote by $A' \preceq A$ the fact that, for each $x \in X$, $A'(x) \leq A(x)$ holds.

Example 20 *Let A, A' denote fuzzy sets from Examples 18 and 19. If $A' \sim Ve\ Sm$ and $A \sim Sm$ then $A' \preceq A$.*

In our task we consider a numerical data set in the form of two-dimensional table $\mathcal{D}$ (see page 29),

	X_1	X_2	$\dots$	X_k
o_1	a_{11}	a_{12}	$\dots$	a_{1k}
o_2	a_{21}	a_{22}	$\dots$	a_{2k}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
o_m	a_{m1}	a_{m2}	$\dots$	a_{mk}

where an element of table is a real number $a_{ij} \in \mathbb{R}$ ($e_{ij} = [o_i]_j$), it is a value of jth attribute (property) X_j measured on ith *object* (observation, transaction) o_i. Let $\mathcal{D}_o$ denote the set of rows (resp. objects) of $\mathcal{D}$.

Now contexts of all attributes must be specified. Mathematically, for $j = 1, 2, \dots, m$, a *context* of any attribute X_j is a closed interval $[c_j, d_j]$. Any context should be set by the expert (user) which is more natural. When contexts are specified, one can work with fuzzy sets $\mathcal{P}(X_j)$ introduced above.

Example 21 *Consider 10 objects in an attribute* Age *with values* $\{28, 45, 67, 32, 56, 70, 43, 73, 33, 72\}$. *Then the context of the attribute* Age *must be given by the expert as* $[0, 110]$. □

Our goal is to search for dependencies between given disjoint sets of attributes $\{Y_o\}_{o=1}^q$, $\{Z_p\}_{p=1}^r \subseteq \{X_j\}_{j=1}^k$. We look for unknown linguistic associations of the form

$$E(\{Y_o\}_{o=1}^q) \Rightarrow F(\{Z_p\}_{p=1}^r), \tag{5.1.1}$$

($E \Rightarrow F$ in short) where E, F are conjunctive evaluative linguistic predications, i.e., predications of the form

$$E = \mathtt{AND}_{o=1}^q (Y_o\ is\ S_o),\ S_o \in \mathcal{P}(Y_o), \tag{5.1.2}$$

and $\Rightarrow$ denotes a relationship between E and F. This relationship can be given by chosen confirmation measures introduced in Chapter 4 (for more details we refer to [6]).

Below we also work with so-called itemsets. A *k–itemset* T is a set of k ordered pairs (o, S_o) where any $o \in \{1, 2, \dots, m\}$ denotes an index for which $S_o \in \mathcal{P}(Y_o)$. Clearly, see the next example, there exists a one-to-one correspondence between linguistic predications (5.1.2) and p–itemsets. Consequently, we can identify p–itemsets with relevant linguistic predications - for instance, the cardinality of a p–itemset T can be considered as a cardinality of E, (3.2.1) can be thought as $T \Rightarrow R$ where T is a p–itemset and R is a q–itemset, respectively, and so on.

Example 22 *Assume that linguistic predications are defined in all attributes. Then an expression "X_2 is* very small `AND` *X_5 is* big but not very big*" can be represented by 2–itemset $\{(2, E_2), (5, E_5)\}$, where $E_2 \sim$ Ve Sm and $E_5 \sim$ Bi but not Ve Bi, respectively.*

Similarly, it is easy to see that specificity ordering of fuzzy sets (see Page 66 or can be extended to the set of itemsets in a very natural way. Namely, for a p–itemset $T = \{(i, E_i)\}_{i \in I}$ and q–itemset $R = \{(j, F_j)\}_{j \in J}$ we denote $T \preceq R$ if $I \subseteq J$ and $E_i \preceq F_i$ for any $i \in I$. Finally, we can specify an operator `C` representing cardinality of a given p–itemset (resp. (5.1.2)).

Definition 24 *An operator `C` is cardinality of a given p–itemset and it is defined by*

$$\mathtt{C}\,(E)(o) = \mathtt{AND}_{o=1}^{q} A_o([o_i]_o) \tag{5.1.3}$$

for any $o \in \mathcal{D}_o$.

The Apriori algorithm is one of the best known algorithm used for searching for associations. In the first step of this algorithm frequent itemsets are discovered. Then candidate associations are generated and tested by chosen confidence measure. In this chapter we demonstrate how to implement our purposed mathematical model into this algorithm. The computational complexity of the proposed algorithm is higher, however our algorithm allows to adapt mined association to the data set and hence, in some way, substitutes some preprocessing steps.

The aim of this section is twofold. Firstly, we demonstrate how the properties described above and background knowledge can be implemented into the Apriori algorithm (e.g., [1]). Secondly, we suggest an implementation of our model of evaluative linguistic expressions.

5.2 The Algorithm

The proposed algorithm is the following:

INPUT:

Data description - notation:

m . . . the number of objects,
k . . . the number of attributes,
$\mathcal{D}_o$ . . . the set of objects,
X_j . . . the jth attribute $j = 1, \ldots, k$,
a_{ij} . . . the value of jth attribute measured on ith object.

What is specified by the user:

supp_p . . . the support measure (p is one of t, m, c) (see Chapter 4),
$\mathit{supp\_\ min}$ . . . minimal support threshold,
$\mathit{conf\_\ min}$ . . . minimal confidence threshold and a suitable *linguistic description*,
$[c_j, d_j]$ . . . the context of attribute X_j,
$\mathcal{P}(X_j)$ . . . fuzzy covering $\{A_{jl}\}_{l=1}^{|X_j|}$ on X_j consisting of fuzzy sets A_{jl},
where $|X_j|$ is the number of fuzzy sets $P^1(X_j) := \mathcal{P}(X_j)$
(e.g., see Example 19),
$\mathcal{E}$ . . . the set of associations representing background knowledge.

Other symbols:

C_r . . . sets of candidate r–itemsets,
(e.g., $C_1 = \{(j, A_{jl}) \,|\, \forall j = 1, \ldots, k;\ A_{jl} \in P^1(X_j)\}$), (we start with $C_r = \emptyset$),
L_r . . . sets of large r–itemsets (we start with $L_r = \emptyset$),
$\mathcal{A}$ . . . the set of found associations (we start with $\mathcal{A} = \emptyset$),
$\tilde{\mathcal{A}}$ . . . the set of derived found associations.
$\mathcal{E}$ . . . the set of background knowledge (in the form of found associations).
In the next we work with t-itemsets of knowledge
$\tilde{\mathcal{E}}$ . . . the set of derived background knowledge.

OUTPUT: The set of linguistic associations $\mathcal{A}$.

STEP 1: Construct a set $C_1 := \{(j, A_{jl}) \,|\, A_{jl} \in P^1(X_j),\ j = 1, 2, \ldots, m\}$ of all 1–itemsets and, for each $t \in C_1$, compute (see (5.1.3))

$$\mathit{count}(t) = \sum_{o \in D_o} \mathtt{C}\,(t)(o). \tag{5.2.1}$$

STEP 2: For each $t \in \mathcal{E}$ we consider all possible itemsets t' satisfying $t \preceq t'$ and we put background knowledge formed by appropriate itemset t' into $\tilde{\mathcal{E}}$.

STEP 3: If $t \notin \mathcal{E}$ and $t \notin \tilde{\mathcal{E}}$ check the *count*(t) of each $t \in C_1$:

- **(a)** If *count*$(t) \geq$ *supp*_ *min* then put t into L_1.
- **(b)** If *count*$(t) <$ *supp*_ *min* and $t \in P^1_Q(X_j)$*) (Q is one of Sm, Me, Bi), then check all $t' \in P^2_Q(X_j)$, satisfying $t \preceq t'$. If *count*$(t') \geq$ *supp*_ *min* for such t', put $t' \in L_1$. Otherwise, check all $t'' \in P^3_Q(X_j)$, satisfying $t \preceq t''$. If *count*$(t'') \geq$ *supp*_ *min* for such t'', put $t'' \in L_1$.

STEP 4: Set $r = 1$.

STEP 5: As in the original algorithm, to generate C_{r+1} from large r–itemsets, i.e., use r–itemsets from L_r. The only difference is that we have to deal with linguistic expressions which can be ordered by specificity ordering. In order to keep cardinalities of r–subitemsets, every generated $(r+1)$–itemset the most broad expressions mentioned in r–itemsets.

Example 23 *Let only pairs* $\{t_1, t_2\}$, $\{t_1, t_3\}$, $\{t_1, t_4\}$ *and* $\{t_2, t_3\}$ *be in* L_2. *Then only* $\{t_1, t_2, t_3\} \in C_3$ *while* $\{t_1, t_2, t_4\} \notin C_3$ *(resp.* $\{t_2, t_3, t_4\}$, $\{t_1, t_3, t_4\}$*) because* $\{t_2, t_4\} \notin L_2$ *(resp.* $\{t_2, t_4\}$, $\{t_3, t_4\} \notin L_2$*).*

STEP 6: For any $t \in C_{r+1}$ do the following:

- **(a)** Compute *count*(t) by (5.2.1).
- **(b)** If *count*$(t) \geq$ *supp*_ *min*, put t in L_{r+1}.
- **(c)** If *count*$(t) <$ *supp*_ *min* we have to consider "broader" linguistic expressions in every attribute as in STEP 3(b). For example, if

$$t = \{(u_1, A_{u_1}), (u_2, A_{u_2}), \ldots, (u_i, A_{u_i}), \ldots, (u_{r+1}, A_{u_{r+1}})\} \tag{5.2.2}$$

and (u_i, A_{u_i}) is such that $A_{u_i} \in P^h_Q(X_j)$, $Q \in \{Sm, Me, Bi\}$ and $h \leq 2$, then instead of A_{u_i} we take all $A'_{u_i} \in P^{h+1}_Q(X_j)$. Thus, we check (5.2.1) for $(r+1)$–itemset

$$t' = \{(u_1, A_{u_1}), (u_2, A_{u_2}), \ldots, (u_i, A'_{u_i}), \ldots, (u_{r+1}, A_{u_{r+1}})\}.$$

If *count*$(t') \geq$ *supp*_ *min* then we do not check $\tilde{t} \in C_{r+1}$ for which $t' \preceq \tilde{t}$ and put $t' \in L_{r+1}$. But we have to check other possible combinations

*) Here and below we use the fact that k–itemsets can be identified with elements of fuzzy covering.

of "broader" expressions in this step as well as for $h := h + 1$ (if it is possible).

(d) For any $t \in L_{r+1}$ we may assume that elements of t are ordered by their cardinalities. I.e., for (5.2.2) we assume

$$\mathit{count}\{(u_i, A_{u_i})\} \leq \mathit{count}\{(u_k, A_{u_k})\} \text{ whenever } u_k \leq u_i.$$

STEP 7: If $L_{r+1} = \emptyset$ and $r \geq 2$ then follow the next steps.

STEP 8: Set $w = 1$.

STEP 9: Choose element $t \in L_r$ of the form

$$t = \{(u_1, A_{u_1}), (u_2, A_{u_2}), \ldots, (u_i, A_{u_i}), \ldots, (u_r, A_{u_r})\}. \tag{5.2.3}$$

The r–itemset t can be decomposed into $t'(w)$ and $t''(w) := t \setminus t'(w)$ where w denotes that t' consists of w elements. For instance,

$$t'(w) = \{(u_1, A_{u_1}), (u_2, A_{u_2}) \ldots, (u_w, A_{u_w})\},$$

$$t''(w) = \{(u_{w+1}, A_{u_{w+1}}), (u_{w+2}, A_{u_{w+2}}) \ldots, (u_r, A_{u_r})\}.$$

Clearly, such decomposition defines an association $a(w) := t'(w) \Rightarrow t''(w)$. For all w–itemsets $t'(w)$ we do the following steps.

STEP 10: If $\mathit{conf}_p(a(w)) < \mathit{conf\_ min}$, then $t'(w)$ is replaced by w–itemset $\tilde{t}$ possessing "broader" expressions stepwise in $1, 2$ up to w elements (i.e., for $i = 1, \ldots, w$ $(i, A_i) \in t'(w)$ implies that there exists $(i, \tilde{A}_i) \in \tilde{t}$ such that $A_i \preceq \tilde{A}_i$) an association $a(w) := \tilde{t} \Rightarrow t''(w)$ is checked instead of $a(w)$. (As in STEP 6 (c) - all possible combinations of "broader" expressions have to be considered here.

For $i = 1, 2, \ldots, w$ put $A_{u_i} := A'_{u_i}$ that $A_{u_i} \preceq A'_{u_i}$ and repeat STEP 10.

If A'_{u_i} does not exist then $i := i + 1$ and repeat STEP 10.

If none association $\tilde{t} \Rightarrow t''(w)$ can be constructed then choose different $t'(w) \subseteq t$ and repeat this step. If this is not possible and $w < r$, put $w := w + 1$ and repeat this step with another $t'(w) \subseteq t$. If $w = r$, then $L_r := L_r \setminus t$ and go to STEP 9 whenever $L_r \neq \emptyset$. In the latter case, $r := r - 1$ and go to STEP 8.

STEP 11: If $\mathit{conf}_p(a(w)) \geq \mathit{conf\_ min}$ then

(a) put $a(w)$ into set $\mathcal{A}$.

(aa) put all $\tilde{a}(w)$ into set $\tilde{\mathcal{A}}$, where $\tilde{t}''(w) = t''(w)$ and $\tilde{t}'(w)$ is "broader" expression of the association $\tilde{a}(w)$ satisfying $t'(w) \preceq \tilde{t}'(w)$.

For confirmation measures (4.0.3) and (4.0.5) and relevant (4.0.7) we can use the following reduction tools.

(b) It follows from Property 8 that, for z={w+1, . . . , r-1}, $a(z)$ is valid if the antecedent of $a(z)$ contains $t'(w)$ therefore put $a(z)$ in $\tilde{\mathcal{A}}$.

(c) It follows from Property 8 that all associations $a(w)$:= $t'(w) \Rightarrow \tilde{t}$ are valid whenever $t''(w) \preceq \tilde{t}$. Therefore put $a(w)$ in $\tilde{\mathcal{A}}$.

(d) According to (4.0.7) it is clear that elements of $t'(w)$ can be replaced by elements of t with lower cardinality and the validity is not corrupted.

Example 24 *We consider* 3*–itemset* $\{t_1, t_2, t_3\}$ *where* $\mathtt{C}\,(t_3) \leq \mathtt{C}\,(t_2) \leq \mathtt{C}\,(t_1)$ *and we obtain* $a(1) := (t_1 \Rightarrow t_2 \texttt{ AND } t_3) \in \mathcal{A}$*, then* $(t_2 \Rightarrow t_1 \texttt{ AND } t_3)$ *and* $(t_3 \Rightarrow t_1 \texttt{ AND } t_2)$ *are also valid associations.*

For all confirmation measures from Chapter 4:

(e) Lemmas 7 and 8 and associations from $\mathcal{E}$ might be applied in this step as well.

STEP 12: If all associations $a(w)$ generated from t were checked and $w < r$, put $w := w + 1$ and go back to STEP 9. If $w = r$, then $L_r := L_r \setminus t$ and go to STEP 9 whenever $L_r \neq \emptyset$. In the next, $r := r - 1$ and go to STEP 8. In the latter case $r = 1$ then it means the end of the algorithm.

In this part we provide analysis of the complexity of our algorithm.

Lemma 13 *Let us consider* m *attributes where* $|X_j|$ *means a number of fuzzy sets of an attribute* X_j *covering its context* $[c_j, d_j]$*. Then total number of all the possible associations is*

$$\sum_{j=2}^{k} \left(\sum_{i=1}^{j-1} \binom{j}{i} \cdot \sum_{r_1=1}^{j} \sum_{r_2=r_1+1}^{j} \cdots \sum_{r_j=r_{j-1}+1}^{j} |X_{r_1}| \cdot |X_{r_2}| \cdots |X_{r_k}| \right). \tag{5.2.4}$$

If we want to obtain a number of associations included "broader" fuzzy sets, we only increase the number of fuzzy sets for each attribute. The proof of the lemma is given by a mathematical induction.

PROOF: For 2–itemset we consider one attribute in the antecedent (one in the succedent) and all possible fuzzy sets for each attribute, then we obtain the number of investigated associations in the form

$$\binom{2}{1} \cdot \sum_{r_1} = 1^k \left(\sum_{r_2=r_1+1}^{k} |X_{r_1}| \cdot |X_{r_2}| \right).$$

We suppose that for j–itemset we have the number of investigated associations given by (5.2.4). We verify the number of investigated associations for $(j+1)$–itemset.

For 2–itemset is the number given in the previous paragraph. For 3–itemset we consider one as well as two attributes in antecedent, then we obtain

$$\left(\binom{3}{1}+\binom{3}{2}\right)\cdot\sum_{r_1=1}^{k}\left(\sum_{r_2=r_1+1}^{k}\left(\sum_{r_3=r_2+1}^{k}|X_{r_1}|\cdot|X_{r_2}|\cdot|X_{r_3}|\right)\right).$$

Hence, for $(k+1)$–itemset we obtain

$$\sum_{i=1}^{j}\binom{j+1}{i}\cdot\sum_{r_1=1}^{j+1}\left(\sum_{r_2=r_1+1}^{j+1}\cdots\left(\sum_{r_{j+1}=r_j+1}^{j+1}|X_{r_1}|\cdot|X_{r_2}|\cdots|X_{r_{j+1}}|\right)\right).$$

The total number of all associations (i.e., 2–itemsets, 3–itemsets, $\cdots$, $(j+1)$–itemsets) is the sum of the numbers of all possible itemsets then we get the formula

$$\sum_{j=2}^{k}\left(\sum_{i=1}^{j}\binom{j+1}{i}\cdot\sum_{r_1=1}^{j+1}\sum_{r_2=r_1+1}^{j+1}\cdots\sum_{r_{j+1}=r_j+1}^{j+1}|X_{r_1}|\cdot|X_{r_2}|\cdots|X_{r_{j+1}}|\right).$$

In this way we prove that Lemma 13 is valid. □

Corollary 6 *If we consider k attributes where $|X_j|>2$ means a number of fuzzy sets of an attribute X_j covering its context $[c_j,d_j]$ (where we work with small, medium and big values). Then the number of attribute X_j is bigger more*

$$\sum_{i=1}^{\frac{|X_j|}{3}-1}\frac{|X_j|}{3}-i,$$

i.e., the total sum of all possible associations is given by (5.2.4) where $|X_j|:=|X_j|+\sum_{i=1}^{\frac{|X_j|}{3}-1}$.

Example 25 *Consider 3 attributes ($k=3$) with the mathematical model considered in Example 18 where*

(a) *the narrowest fuzzy sets, i.e., $|X_1|=|X_2|=|X_3|=9$. Then the number of all possible associations according to (5.2.4) is $2\cdot 9\cdot 9+6\cdot 9\cdot 9\cdot 9=4862$.*

(b) *the number of fuzzy sets with "broader" fuzzy sets describing in Example 18 is $|X_1|=|X_2|=|X_3|=18$. Then the number of all possible associations according to (5.2.4) is $2\cdot 3\cdot 18\cdot 18+6\cdot 18\cdot 18\cdot 18=36936$.*

According to Lemma 4.2 if the association

"IF X_1 is small THEN X_2 is very big"

is a valid association then following associations are valid associations as well

"IF X_1 is small THEN X_2 is big,"

"IF X_1 is small THEN X_2 is more or less big."

In this way, if all associations with the narrowest fuzzy sets are valid associations then we need not investigate $2 \cdot 4862$, i.e., 9724 associations. Thus we need not investigate 26% from all possible associations.

According to Lemma 4.2 if there exists a valid association in this form "IF X_1 is A THEN X_2 is B AND *X_3 is C" where A, B, C are fuzzy sets of appropriate attributes, then we need not investigate these two associations:*

"IF X_1 is A AND *X_2 is B THEN X_3 is C",*

"IF X_1 is A AND *X_3 is C THEN X_2 is B".*

Moreover, if there exist B', B'', C', C'' that $B \preceq B' \preceq B''$ and $C \preceq C' \preceq C''$ then following associations are also valid associations according to Lemma 5

"IF X_1 is A AND *X_2 is B THEN X_3 is C'",*
"IF X_1 is A AND *X_2 is B THEN X_3 is C'''",*
"IF X_1 is A AND *X_3 is C THEN X_2 is B'",*
"IF X_1 is A AND *X_3 is C THEN X_2 is B''".*

Chapter 6

List of the author's publications

- I. Tomanová and J. Kupka, *Implementation of Background Knowledge and Properties Induced by Fuzzy Confirmation Measures in Apriori Algorithm*, International Joint Conference CISIS12-ICEUTE12-SOCO12 Special Sessions (189), 2013.

- J. Kupka and I. Tomanová, *Dependencies among attributes given by fuzzy confirmation measures*, Expert Systems with Applications (39), 9, p.7591–7599, 2012.

- J. Kupka and I. Tomanová, *Some dependencies among attributes given by fuzzy confirmation measures*, Proc. of the LFA-EUSFLAT'2011, France, p.498–505, 2011.

- J. Kupka and I. Tomanová, *Some Extensions of Mining of Linguistic Associations*, Neural Network World (20), p.27–44, 2010.

- M. Štěpnička, V. Pavliska, V. Novák, I. Perfilieva, L. Vavříčková and I. Tomanová, *Time Series Analysis and Prediction Based on Fuzzy Rules and the Fuzzy Transform*, Proceedings of IFSA World Congress/EUSFLAT Conference. Lisabon: Universidade Técnica de Lisboa, p. 483–488, 2009.

Bibliography

[1] R. Agrawal and R. Srikant, *Fast algorithms for mining association rules*, vol. 1215, pp. 487–499, Citeseer, 1994.

[2] W.W. Armstrong, *Dependency structures of database relationships*, Proceedings of IFIP 1974 (1974), 580–583.

[3] R. Bělohlávek and V. Vychodil, *Fuzzy attribute logic over complete residuated lattices.*

[4] Y. L. Chen and C. H. Weng, *Mining fuzzy association rules from questionnaire data*, Knowledge-Based Systems **22** (2009), 46–56.

[5] M. Delgado, N. Marn, D. Sánchez, and M.-A. Vila, *Fuzzy association rules: General model and applications*, IEEE Transactions on Fuzzy Systems **11** (2003), 214–225.

[6] D. Dubois, E. Hüllermeier, and H. Prade, *A systematic approach to the assessment of fuzzy association rules*, Data mining and Knowledge Discovery **13** (2006), 167–192.

[7] A. Dvořák and J. Kupka, *Linguistic associations mining with the lam software*, Technical report 15, IRAFM, University of Ostrava (2008).

[8] D. H .Glass, *Fuzzy confirmation measures*, Fuzzy Sets and Systems **159** (2008), 475–490.

[9] P. Hájek, *The question of a general concept of the guha method*, Kybernetika (1968), 505–515.

[10] P. Hájek, *Logics for data mining (guha rediviva)*, Neural Network World **10** (2000), 301–311.

[11] P. Hájek and T. Havránek, *Mechanizing hypothesis formation*, Mathematical foundations for a general theory, Berlin/Heidelberg/New York: Springer-Verlag, 1978.

[12] T. P. Hong, K. Y. Lin, and S. L. Wang, *Fuzzy data mining for interesting generalized association rules*, Fuzzy Sets and Systems **138** (2003), 255–269.

[13] E. Hüllermeier, *Implication based fuzzy association rules*, Proc. of the 5th European conference on principles and practice of knowledge discovery in databases, PKDD-01 Freiburg, Germany: Springer-Verlang (France), 2001, pp. 241–252.

[14] J. Kupka and I. Tomanová, *Some extensions of mining of linguistic associations*, Neural Network World **20** (2010), 27–44.

[15] ______, *Some dependencies among attributes given by fuzzy confirmation measures*, Proc. of the LFA-EUSFLAT'2011 (France), 2011, pp. 498–505.

[16] J. Kupka and I. Tomanová, *Dependencies among attributes given by fuzzy confirmation measures*, Expert Systems with Applications **39** (2012), no. 9, 7591–7599.

[17] P. Lenca, P. Meyer, B. Vaillant, and S. Lallich, *On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid*, European Journal of Operation Research **184** (2008), 610–626.

[18] R. J. Miller and Y. Yang, *Association rules over interval data*, ACM SIGMOND **26** (1997), no. 2, 452–461.

[19] V. Novák, *Towards formalized integrated theory of fuzzy logic*, Fuzzy Logic and Its Applications to Engineering, Information Sciences, and Intelligent Systems, Kluwer, Dordrecht (1995), 353–363.

[20] V. Novák, *On fuzzy type theory*, Fuzzy Sets and Systems (2005), no. 149, 235–273.

[21] V. Novák, *Fuzzy logic theory of evaluative expressions and comparative quantifiers*, Proceedingd of 11th International Conference IPMU (Paris), vol. 2, 2006, pp. 1572–1579.

[22] V. Novák, *A comprehensive theory of trichotomous evaluative linguistic expressions*, Fuzzy Sets and Systems **22** (2008), no. 159, 2939–2969.

[23] V. Novák, I. Perfilieva, A. Dvořák, Q. Che., Q. Wei, and P. Yan, *Mining pure linguistic associations from numerical data*, International Journal of Approximate Reasoning **48** (2008), no. 1, 4–22.

[24] V. Novák, I. Perfilieva, and J. Močkoř, *Mathematical principles of fuzzy logic*, Kluwer Academic Publishers, Boston, 1999.

[25] I. Perfilieva, V. Novák, and A. Dvořák, *Fuzzy transform in the analysis of data*, International Journal of Approximate Reasoning **48** (2008), no. 1, 36–46.

[26] J. Rauch, *Logic of association rules*, Applied Intelligence **22** (2005), 9–28.

[27] R. Srikant and R. Agrawal, *Mining quantitative association rules in large relational tables*, Proceedings of the ACM SIGMOD International Conference on Management of Data (Montreal, Canada), vol. 48, 1996, pp. 1–12.

[28] L. A. Zadeh, *The concept of a linguistic variable and its application to approximate reasoning i, ii, iii*, Information Sciences **8–9** (1975), 199–257 , 301–357, 43–80.