Mining for Patterns Based on Contingency Tables by *KL-Miner* – First Experience

Jan Rauch[1,3], Milan Šimůnek[2,4], and Václav Lín[3]

[1] EuroMISE Centrum – Cardio,
[2] Department of Information Technology,
[3] Department of Information and Knowledge Engineering,
[4] Laboratory for Intelligent Systems,
University of Economics, Prague
nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic
rauch@vse.cz, simunek@vse.cz, xlinv05@vse.cz

Summary. A new datamining procedure called *KL-Miner* is presented. The procedure mines for various patterns based on evaluation of two–dimensional contingency tables, including patterns of statistical or information theoretic nature. The procedure is a result of continued development of the academic system LISp-Miner for KDD.

Key words: Data mining, contingency tables, system LISp–Miner, statistical patterns

1 Introduction

Goal of this paper is to present first experience with data mining procedure *KL-Miner*. The procedure mines for patterns of the form

$$R \sim C/\gamma \ .$$

Here R and C are categorial attributes, the attribute R has *categories* (possible values) $r_1, \ldots, r_K$, the attribute C has categories $c_1, \ldots, c_L$. Further, γ is a Boolean attribute.

The *KL-Miner* works with data matrices. We suppose that R and C correspond to columns of the analyzed data matrix. We further suppose that the Boolean attribute γ is somehow derived from other columns of the analyzed data matrix and thus that it corresponds to a Boolean column of the analyzed data matrix.

The intuitive meaning of the expression $R \sim C/\gamma$ is that the attributes R and C are in relation given by the symbol $\sim$ when the condition given by the derived Boolean attribute γ is satisfied.

The symbol $\sim$ is called *KL-quantifier* . It corresponds to a condition imposed by the user on the contingency table of R and C. There are several restrictions that the user can choose to use (e.g. minimal value, sum over the table, value of the χ^2 statistic, and other).

We call the expression $R \sim C/\gamma$ a *KL-hypothesis* or simply *hypothesis*. The KL-hypothesis $R \sim C/\gamma$ is *true* in the data matrix $\mathcal{M}$ if the condition corresponding to the KL-quantifier $\sim$ is satisfied for the contingency table of R and C on the data matrix $\mathcal{M}/\gamma$. The data matrix $\mathcal{M}/\gamma$ consists of all rows of data matrix $\mathcal{M}$ satisfying the condition γ (i.e. of all rows in which the value of γ is `TRUE`).

Input of the *KL-Miner* consists of the analyzed data matrix and of several parameters defining a set of potentially interesting hypotheses. Such a set can be very large. The *KL-Miner* automatically generates all potentially interesting hypotheses and verifies them in the analyzed data matrix. The output of the *KL-Miner* consists of all hypotheses that are true in the analyzed data matrix (i.e. supported by the analyzed data). Some details about input of the *KL-Miner* procedure are given in Sect. 2. KL-quantifiers are described in Sect. 3.

The *KL-Miner* procedure is one of data mining procedures of the LISp-Miner system that are implemented using bit string approach [4]. It means that the analysed data is represented using suitable strings of bits. Software modules for dealing with strings of bits developed for the LISp-Miner system are used for *KL-Miner* implementation, see Sect. 4.

Let us remark that the *KL-Miner* is a GUHA procedure in the sense of the book [1]. Therefore, we will use the terminology introduced in [1]. The potentially interesting hypotheses will be called *relevant questions*, and the hypotheses that are true in the analyzed data matrix will be called *relevant truths*. Also, the use of the term *quantifier* for the symbol $\sim$ in the expression $R \sim C/\gamma$ is inspired by [1]. The cited book contains rich enough theoretical framework to build a formal logical theory for the *KL-Miner*–style data mining; however, we will not do it here. Furthermore, *KL-Miner* is related to a GUHA procedure from the 1980's called *CORREL*, see [2]. Let us also remark that the analysis of contingency tables by *KL-Miner* is similar in spirit to that of the procedure *49er* [7].

2 Input of *KL-Miner*

Please recall that the *KL-Miner* mines for hypotheses of the form $R \sim C/\gamma$, where R and C are categorial attributes, γ is a Boolean attribute, and $\sim$ is a KL-quantifier. The attribute R is called the *row attribute*, the attribute C is called the *column attribute*.

Input of the *KL-Miner* procedure consists of

- the analyzed data matrix