Decision Rules and Dependencies

Zdzisław Pawlak

Institute for Theoretical and Applied Informatics
Polish Academy of Sciences
ul. Bałtycka 5, 44-100 Gliwice, Poland
and
University of Information Technology and Management
ul. Newelska 6, 01-447 Warsaw, Poland
e-mail: zpw@ii.pw.edu.pl

Abstract. We proposed in this paper to use some ideas of Jan Łukasiewicz, concerning independence of logical formulas, to study dependencies in databases.

1 Introduction

This paper concerns the application of some ideas given by Jan Łukasiewicz in [1], in connection with his study of logic and probability – to data mining and data analysis. The relationship between implication and decision rules is formulated and studied along the lines proposed by the author in [2, 3]. Moreover, the independence of propositional functions, introduced by Łukasiewicz, is generalized and used to characterization of decision rules – leading to a new look on dependencies in databases. The proposed approach seems to give a new tool to discovering patterns in data.

2 Decision rules

Let U be a non empty finite set, called the *universe* and let Φ , Ψ be logical formulas. The meaning of Φ in U, denoted by $|\Phi|$, is the set of all elements of U, that satisfies Φ in U. The truth value of Φ denoted $val(\Phi)$ is defined as $card|\Phi|/card(U)$.

A *decision rule* is an expression $\Phi \to \Psi$, read if Φ *then* Ψ, where Φ and Ψ are referred to as *conditions* and *decisions* of the rule, respectively.

The number $supp(\Phi,\Psi) = card(|\Phi \wedge \Psi|)$ will be called the *support* of the rule $\Phi \to \Psi$. We will consider non void decision rules only, i.e., rules such that $supp(\Phi,\Psi) \neq 0$.

With every decision rule $\Phi \to \Psi$ we associate its *strength* defined as

$$str(\Phi,\Psi) = \frac{supp(\Phi,\Psi)}{card(U)}.$$

Moreover, with every decision rule $\Phi \to \Psi$ we associate the *certainty factor* defined as

$$cer(\Phi, \Psi) = \frac{str(\Phi, \Psi)}{val(\Phi)} \tag{1}$$

and the *coverage factor* of $\Phi \to \Psi$

$$cov(\Phi, \Psi) = \frac{str(\Phi, \Psi)}{val(\Psi)}, \tag{2}$$

where $val(\Phi) \neq 0$ and $val(\Psi) \neq 0$.

If a decision rule $\Phi \to \Psi$ uniquely determines decisions in terms of conditions, i.e., if $cer(\Phi, \Psi) = 1$, then the rule is *certain*, otherwise the rule is *uncertain*.

If a decision rule $\Phi \to \Psi$ covers all decisions, i.e., if $cov(\Phi, \Psi) = 1$ then the decision rule is *total*, otherwise the decision rule is *partial*.

Immediate consequences of (1) and (2) are:

$$cer(\Phi, \Psi) = \frac{cov(\Phi, \Psi) val(\Psi)}{val(\Phi)}, \tag{3}$$

$$cov(\Phi, \Psi) = \frac{cer(\Phi, \Psi) val(\Phi)}{val(\Psi)}. \tag{4}$$

Note, that (3) and (4) are Bayes' formulas. This relationship first was observed by Łukasiewicz [1].

3 Decision rules and inference rules

Let $\Phi \to \Psi$ be a decision rule. We have

$$val(\Psi) = \frac{val(\Phi) cer(\Phi, \Psi)}{cov(\Phi, \Psi)} = \frac{str(\Phi, \Psi)}{cov(\Phi, \Psi)} \tag{5}$$

and

$$val(\Phi) = \frac{val(\Psi) cov(\Phi, \Psi)}{cer(\Phi, \Psi)} = \frac{str(\Phi, \Psi)}{cer(\Phi, \Psi)}. \tag{6}$$

Formulas (5) and (6) are direct consequences of (3) and (4), respectively and consequently they are Bayes' rules, too.

It is easily seen that formulas resemble well known *modus ponens* and *modus tollens* inference rules, which have the form

$$\begin{array}{lll} \textit{if} & \Phi \to \Psi & \textit{is true} \\ \textit{and} & \Phi & \textit{is true} \\ \hline \textit{then} & \Psi & \textit{is true} \end{array}$$

and

$$\begin{array}{lll} \textit{if} & \Phi \rightarrow \Psi & \textit{is true} \\ \textit{and} & \sim \Psi & \textit{is true} \\ \hline \textit{then} \sim \Phi & & \textit{is true} \end{array}$$

respectively.

Inference rules allow us to obtain true consequences from true premises. In reasoning about data (data analysis) the situation is slightly different. Instead of true propositions we consider propositional functions, which are true to a "degree", i.e., they assume truth values which lie between 0 and 1, in other words, they are probable, not true [1].

Let us formulate this idea more exactly.

We can write

$$\begin{array}{llll} \textit{if} & \Phi \rightarrow \Psi & & \\ \textit{and} & \Phi & \textit{is true to a degree} & val(\Phi) \\ \hline \textit{then} & \quad \Psi & \textit{is true to a degree} & val(\Psi) = \alpha val(\Phi) \end{array}$$

and

$$\begin{array}{llll} \textit{if} & \Phi \rightarrow \Psi & & \\ \textit{and} & \quad \Psi & \textit{is true to a degree} & val(\Psi) \\ \hline \textit{then} & \Phi & \textit{is true to a degree} & val(\Phi) = \alpha^{-1} val(\Psi) \end{array}$$

where

$$\alpha = \frac{cer(\Phi, \Psi)}{cov(\Phi, \Psi)}.$$

The above inference rules can be considered as counter-parts of *modus ponens* and *modus tollens* for data analysis.

4 Independence in decision rules

Independence of logical formulas considered in this section first was proposed by Łukasiewicz [1].

Let $\Phi \rightarrow \Psi$ be a decision rule. Formulas Φ and Ψ are independent on each other if

$$str(\Phi, \Psi) = val(\Phi)val(\Psi).$$

Consequently

$$\frac{str(\Phi, \Psi)}{val(\Phi)} = cer(\Phi, \Psi) = val(\Psi),$$

and

$$\frac{str(\Phi, \Psi)}{val(\Psi)} = cov(\Phi, \Psi) = val(\Phi).$$

If

$$cer(\Phi, \Psi) > val(\Psi),$$

or

$$cov(\Phi, \Psi) > val(\Phi),$$

then Φ and Ψ *depend positively* on each other. Similarly, if

$$cer(\Phi, \Psi) < val(\Psi),$$

or

$$cov(\Phi, \Psi) < val(\Phi),$$

then Φ and Ψ *depend negatively* on each other.

Let us observe that relations of independency and dependences are symmetric ones, and are analogous to that used in statistics.

Example 1. Let $U = \{1, 2, \ldots, 6\}$, $x \in U$ and let Φ_1 denote "x is divisible by 2", Φ_0 – "x is not divisible by 2". Similarly, Ψ_1 stands for "x is divisible by 3" and Ψ_0 – "x is not divisible by 3". Because there are 50% elements divisible by 2 and 50% elements not divisible by 2 in U, therefore we have $val(\Phi_1) = 1/2$ and $val(\Phi_0) = 1/2$. Similarly, $val(\Psi_1) = 1/3$ and $val(\Psi_0) = 2/3$, respectively.
The situation is presented in Fig. 1.

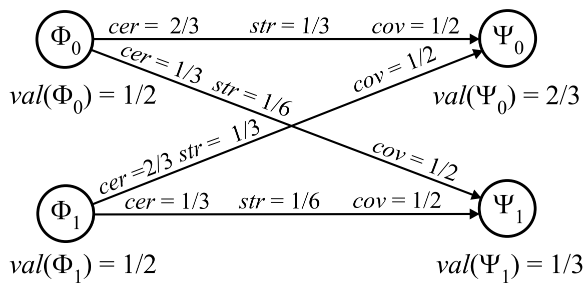

Fig. 1. Divisibility by "2" and "3"

Formulas Φ_0 and Ψ_0, Φ_0 and Ψ_1, Φ_1 and Ψ_0, Φ_1 and Ψ_1 are pair-wise independent on each other, because, e.g., $cer(\Phi_0, \Psi_0) = val(\Psi_0)(cov(\Phi_0, \Psi_0) = val(\Phi_0))$. □

Example 2. Let $U = \{1, 2, \ldots, 8\}$, $x \in U$ and Φ_1 stand for "x is divisible by 2", Φ_0 – "x is not divisible by 2", Ψ_1 – "x is divisible by 4" and Ψ_0 – "x is not

divisible by 4". As in the previous example $val(\Phi_0) = 1/2$ and $val(\Phi_1) = 1/2$; $val(\Psi_0) = 3/4$ and $val(\Psi_1) = 1/4$ because there are 75% elements not divisible by 4 and 25% divisible by 4 in U.
The situation is shown in Fig. 2.

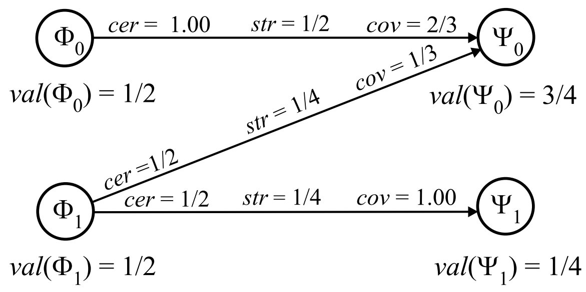

Fig. 2. Divisibility by "2" and "4"

The pairs of formulas Φ_0 and Ψ_0, Φ_1 and Ψ_0, Φ_1 and Φ_1 are dependent. Pairs of formulas Φ_0 and Ψ_0, Φ_1 and Ψ_1 are positively dependent on each other, because $cer(\Phi_0, \Psi_0) > val(\Psi_0)(cov(\Phi_0, \Psi_0) > val(\Phi_0))$ and – $cer(\Phi_1, \Psi_1) > val(\Psi_1)(cov(\Phi_1, \Psi_1) > val(\Phi_1))$. Formulas Φ_1 and Ψ_0 are negatively dependent on each other, because $cer(\Phi_1, \Psi_0) < val(\Psi_0)(cov(\Phi_1, \Psi_0) < val(\Phi_1))$. □

Example 3. Consider a population in which 20% are blond, 80% are dark haired, 40% have blue eyes and 60% have hazel eyes. The relationship between color of hair and eyes is shown in Fig. 3.

It can be seen that blond hair and blue eyes are positively dependent on each other, as well as dark hair and hazel eyes. However, dark hair and blue eyes, and negatively dependent on each other in this population. □

5 Dependency factor

For every decision rule $\Phi \to \Psi$ we define a *dependency factor* $\eta(\Phi, \Psi)$ defined as

$$\eta(\Phi, \Psi) = \frac{cer(\Phi, \Psi) - val(\Psi)}{cer(\Phi, \Psi) + val(\Psi)} = \frac{cov(\Phi, \Psi) - val(\Phi)}{cov(\Phi, \Psi) + val(\Phi)}.$$

It is easy to check that if $\eta(\Phi, \Psi) = 0$, then Φ and Ψ are independent on each other, if $-1 < \eta(\Phi, \Psi) < 0$, then Φ and Ψ are negatively dependent and if $0 < \eta(\Phi, \Psi) < 1$ then Φ and Ψ are positively dependent on each other. Thus the dependency factor expresses a degree of dependency, and can be seen as a counterpart of correlation coefficient used in statistics.

For example, for situation presented in Fig. 1 we have: $\eta(\Phi_0, \Psi_0) = 0$, $\eta(\Phi_0, \Psi_1) = 0$, $\eta(\Phi_1, \Psi_1) = 0$, and $\eta(\Phi_1, \Psi_0) = 0$. However, for Fig. 2 we have $\eta(\Phi_0, \Psi_0) = 1/7$, $\eta(\Phi_1, \Psi_0) = -1/5$ and $\eta(\Phi_1, \Psi_1) = 1/3$. The meaning of the above results is obvious.

For example 3 results are shown in Fig. 3.

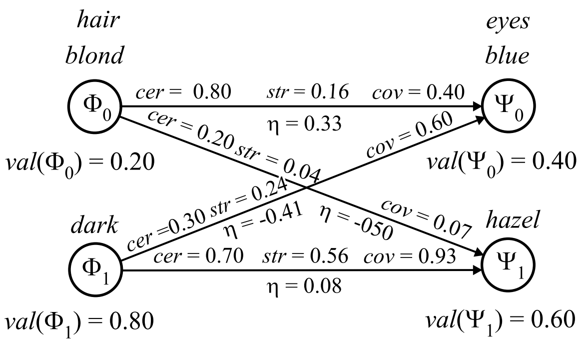

Fig. 3. Correlation between color of hair and eyes

Another dependency factor has been proposed in [4].

6 Summary

We proposed in this paper a new look on dependencies in databases based on some ides of Łukasiewcz proposed in his study of logic and probability.

Acknowledgment

Thanks are due to Professor Andrzej Skowron for critical remarks.

References

1. Łukasiewicz, J.: Die logishen Grundlagen der Wahrscheinilchkeitsrechnung. Kraków (1913), in: L. Borkowski (ed.), Jan Łukasiewicz – Selected Works, North Holland Publishing Company, Amsterdam, London, Polish Scientific Publishers, Warsaw (1970) 16-63
2. Pawlak, Z.: In Pursuit of Patterns in Data Reasoning from Data – The Rough Set Way, in: J.J. Alpigini et al. (eds.), Lecture Notes in Artificial Intelligence 2475 (2002) 1-9
3. Pawlak, Z.: Probability, Truth and Flow Graphs, in: RSKD – International Workshop and Soft Computing, ETAPS 2003, A. Skowron, M. Szczuka (eds.), Warsaw (2003) 1-9
4. Słowiński, R., Greco, S.: A note on dependency factor (manuscript).