Data mining and flow graphs

Zdzisław Pawlak
University of Information Technology and Management
01-447 Warsaw, ul. Newelska 6
e-mali: zpw@ii.pw.edu.pl

Abstract

Searching for patterns in databases is of utmost importance in data mining in recent years. Many methods have been developed and used in this domain (see e.g., [1]). In this talk we will present a new approach to this end, based on information flow analysis in flow networks. The proposed approach consists in representing data structure in a data base in a form of a directed, acyclic graph. Nodes of the graph represent formulas describing features of data, whereas branches are to be interpreted as implications (decision rules), representing relationships between data features. It is revealed the relationship between data can be described as information flow in the graph. This leads to a new class of flow networks, different to that proposed by Ford and Fulkerson [2].

In the talk the notion of a flow graph will be defined and some its basic properties will be shown.

With every decision rule three coefficients: *strength, certainty* and *coverage* are associated. The coefficients are numbers from the interval $< 0, 1 >$. The strength expresses how strongly the decision rule is supported by the data in a data base about our universe of interest; the certainty expresses how strongly we can trust the decision rule, in view of the data available and the coverage represents the degree to what the decision is related to the decision rule with regard to data. Mathematically the certainty and the coverage coefficients can be interpreted as conditional probabilities or as relative truth values. But in our setting they will be interpreted in terms of deterministic flow intensity throughout the branches of an abstract flow network.

Basic mathematical properties of these coefficients will be shown and their interpretation in terms of decision analysis will be studied.

Now let us define the above concepts more formally.

A *flow graph* is a directed, acyclic, finite graph $G = (N, \mathcal{B}, \phi)$, where N is a set of *nodes*, $\mathcal{B} \subseteq N \times N$ is a set of *directed branches*, $\phi : \mathcal{B} \to R^+$ is a *flow function* and R^+ is the set of non-negative reals.

The set of all *inputs* and *outputs* of a node x are $I(x) = \{y \in N : (y, x) \in \mathcal{B}\}$ and $O(x) = \{y \in N : (x, y) \in \mathcal{B}\}$. $I(G) = \{x \in N : I(x) = \emptyset\}$ and $O(G) = \{x \in N : O(x) = \emptyset\}$ denote input and output of the flow graph G, respectively.

If $(x, y) \in B$ then $\phi(x, y)$ is a *troughflow* from x to y. $\phi_+(y) = \sum_{x \in I(y)} \phi(x, y)$, $\phi_-(x) = \sum_{y \in O(x)} \phi(x, y)$ denote an *inflow* and *outflow* of a node y and x, respectively.

$phi_+(G) = \sum_{x \in I(G)} \phi_-(x)$, $\phi_-(G) = \sum_{x \in O(G)} \phi_+(x)$ are *inflow* and *outflow* a the graph G, respectively.

We assume that for any internal node x, $\phi_+(x) = \phi_-(x) = \phi(x)$ and $\phi_+(G) = \phi_-(G) = \phi(G)$. The *strength* (*normalized trough flow* of (x, y)) is defined as $\sigma(x, y) = \frac{\phi(x,y)}{\phi(G)}$.

Obviously, $0 \leq \sigma(x, y) \leq 1$. The strength of a branch expresses simply the percentage of a total flow through the branch.

The *certainty* and the *coverage* of (x, y) are defined as $cer(x, y) = \frac{\sigma(x,y)}{\sigma(x)}$, and the $cov(x, y) = \frac{\sigma(x,y)}{\sigma(y)}$, respectively.

The certainty and the coverage factors describe flow distribution among inputs and outputs of nodes. It is worthwhile to mention the coefficient have been for a long time used in machine learning and data bases, but in fact these coefficients have been first used by J. ukasiewicz in connection with his study of logic, probability and Bayes' theorem [3].

The below properties are immediate consequences of definitions given in the preceding section.

$$\sum_{y \in O(x)} cer(x, y) = 1 \tag{1}$$

$$\sum_{x \in I(y)} cov(x, y) = 1 \tag{2}$$

$$cer(x, y) = \frac{cov(x, y)\sigma(y)}{\sigma(x)} \tag{3}$$

$$cov(x, y) = \frac{cer(x, y)\sigma(x)}{\sigma(y)} \tag{4}$$

It is easily seen that the above properties have a probabilistic flavor, particularly, equations (3) and (4) are Bayes formulas. However, in our case the properties can be interpreted without referring to their probabilistic character. They simply describe some features of steady flow in a flow network, i.e., flow distribution among branches in the network.

Finally, the application of the introduced conception to patterns discovery in databases will be examined and a numerical example will be used to illustrate the ideas considered.

The proposed approach is a continuation of the authors research on data mining (see e.g., [4,5], and was inspired by Jan ukasiewiczs study on logic and probability (see [3]).

References

1. M. Berthold, D.J. Hand, Intelligent Data Analysis - An Introduction. Springer-Verlag, Berlin, Heidelberg, New York, 1999.

2. L.R. Ford, D.R. Fulkerson, Flows in Networks. Princeton University Press, Princeton. New Jersey, 1962.

3. J. ukasiewicz, Die logishen Grundlagen der Wahrscheinilchkeitsrechnung. Krakw (1913), in: L. Borkowski (ed.), Jan ukasiewicz Selected Works, North Holland

Publishing Company, Amsterdam, London, Polish Scientific Publishers, Warsaw, 1970.

4. Z. Pawlak, Probability, Truth and Flow Graphs, In: RSKD Rough Sets in Knowledge Discovery, Proceedings, A. Skowron, M. Szczuka (eds.) Warsaw, 2003, pp. 1-9.

5. Z. Pawlak, Flow Graphs and Decision Algorithms, In: Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Proceedings, G. Wang, Y. Yao and A. Skowron (eds.) Lecture Notes in Artificial Intelligence, 2639, Springer, 2003, pp. 1-10.