Flow Graphs and Decision Algorithms

Zdzisław Pawlak

University of Information Technology and Management
ul. Newelska 6, 01-447 Warsaw, Poland
and
Chongqing University of Posts and Telecommunications
Chongqing, 400065, P.R. China
zpw@ii.pw.edu.pl

Abstract. In this paper we introduce a new kind of flow networks, called flow graphs, different to that proposed by Ford and Fulkerson. Flow graphs are meant to be used as a mathematical tool to analysis of information flow in decision algorithms, in contrast to material flow optimization considered in classical flow network analysis. In the proposed approach branches of the flow graph are interpreted as decision rules, while the whole flow graph can be understood as a representation of decision algorithm. The information flow in flow graphs is governed by Bayes' rule, however, in our case, the rule does not have probabilistic meaning and is entirely deterministic. It describes simply information flow distribution in flow graphs. This property can be used to draw conclusions from data, without referring to its probabilistic structure.

1 Introduction

The paper is concerned with a new kind of flow networks, called flow graphs, different to that proposed by Ford and Fulkerson [3]. The introduced flow graphs are intended to be used as a mathematical tool for information flow analysis in decision algorithms, in contrast to material flow optimization considered in classical flow network analysis.

In the proposed approach branches of the flow graph are interpreted as decision rules, while the whole flow graph can be understood as a representation of decision algorithm.

It is revealed that the information flow in flow graphs is governed by Bayes' formula, however, in our case the rule does not have probabilistic meaning and is entirely deterministic. It describes simply information flow distribution in flow graphs, without referring to its probabilistic structure.

Despite Bayes' rule is fundamental for statistical reasoning, however it has led to many philosophical discussions concerning its validity and meaning, and has caused much criticism [1], [2]. In our setting, beside a very simple mathematical form, the Bayes' rule is free from its mystic flavor.

G. Wang et al. (Eds.): RSFDGrC 2003, LNAI 2639, pp. 1–10, 2003.

This paper is a continuation of some authors' ideas presented in [6], [7], [8], where the relationship between Bayes' rule and flow graphs has been introduced and studied.

From theoretical point of view the presented approach can be seen as a generalization of Łukasiewicz's ideas [4], who first proposed to express probability in logical terms. He claims that probability is a property of propositional functions, and can be replaced by truth values belonging to the interval <0,1>. In the flow graph setting the truth values, and consequently probabilities, are interpreted as flow intensity in branches of a flow graph. Besides, it leads to simple computational algorithms and new interpretation of decision algorithms.

The paper is organized as follows. First, the concept of a flow graph is introduced. Next, information flow distribution in the graph is defined and its relationship with Bayes' formula is revealed. Further, simplification of flow graphs is considered and the relationship of flow graphs and decision algorithms is analyzed. Finally, statistical independence and dependency between nodes is defined and studied.

All concepts are illustrated by simple tutorial examples.

2 Flow Graphs

A flow graph is a *directed, acyclic*, *finite* graph $G = (N, B, \sigma)$, where N is a set of *nodes*, $B \subseteq N \times N$ is a set of *directed branches,* $\sigma : B \to <0,1>$ is a *flow function.*

Input of $x \in N$ *is the set* $I(x) = \{y \in N: (y,x) \in B\}$; *output* of $x \in N$ is defined as $O(x) = \{y \in N: (x,y) \in B\}$ and $\sigma(x,y)$ is called the *strength* of (x,y).

Input and *output* of a graph G, are defined as $I(G) = \{x \in N : I(x) = \emptyset\}$, $O(G) = \{x \in N : O(x) = \emptyset\}$, respectively.

Inputs and outputs of G are *external nodes* of G; other nodes are *internal nodes* of G.

With every node x of a flow graph G we associate its *inflow* and *outflow* defined as $\sigma_+(x) = \sum_{y \in I(x)} \sigma(y,x)$, $\sigma_-(x) = \sum_{y \in O(x)} \sigma(x,y)$, respectively.

We assume that for any internal node x, $\sigma_+(x) = \sigma_-(x) = \sigma(x)$, where $\sigma(x)$ is a *troughflow* of x.

An inflow and an outflow of G are defined as $\sigma_+(G) = \sum_{x \in I(G)} \sigma_-(x)$, $\sigma_-(G) = \sum_{x \in O(G)} \sigma_+(x)$, respectively.

Obviously $\sigma_+(G) = \sigma_-(G) = \sigma(G)$, where $\sigma(G)$ is a *troughflow* of G. Moreover, we assume that $\sigma(G) = 1$.

The above formulas can be considered as *flow conservation equations* [3].

3 Certainty and Coverage Factors

With every branch of a flow graph we associate the *certainty* and the *coverage factors* [9], [10].

The *certainty* and the *coverage* of (x, y) are defined as $cer(x, y) = \frac{\sigma(x, y)}{\sigma(x)}$ and $cov(x, y) = \frac{\sigma(x, y)}{\sigma(y)}$, respectively, where $\sigma(x)$ is the *troughflow* of x. Below some properties, which are immediate consequences of definitions given above are presented:

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

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

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

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

Obviously the above properties have a probabilistic flavor, e.g., equations (3) and (4) are Bayes' formulas. However, these properties can be interpreted in deterministic way and they describe flow distribution among branches in the network.

Notice that Bayes' formulas given above have a new interpretation form which leads to simple computations and gives new insight into the Bayesian methodology.

Example 1: Suppose that three models of cars x_1, x_2 and x_3 are sold to three disjoint groups of customers z_1, z_2 and z_3 through four dealers y_1, y_2, y_3 and y_4.

Moreover, let us assume that car models and dealers are distributed as shown in Fig. 1.

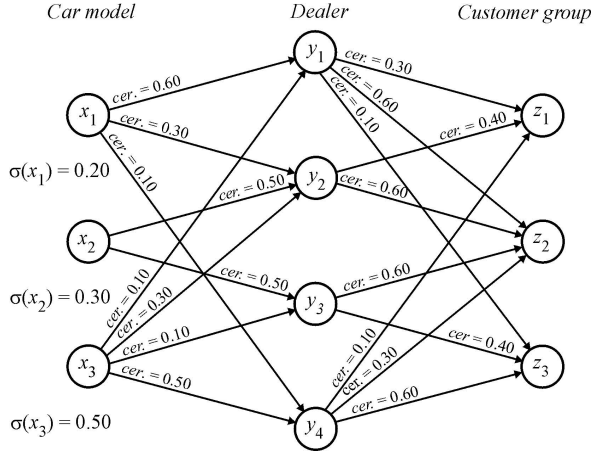

Fig. 1. Cars and dealers distribution

Computing strength and coverage factors for each branch we get results shown in Figure 2.

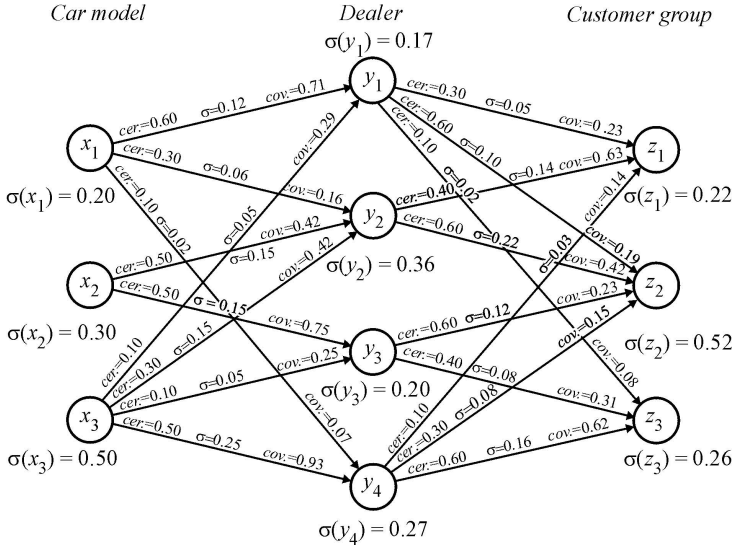

Fig. 2. Strength, certainty and coverage factors

4 Paths and Connections

A (*directed*) *path* from x to y, $x \neq y$ is a sequence of nodes $x_1,\ldots,x_n$ such that $x_1 = x$, $x_n = y$ and $(x_i, x_{i+1}) \in B$ for every i, $1 \leq i \leq n$-1. A path $x \ldots y$ is denoted by $[x,y]$.
The *certainty* of a path $[x_1, x_n]$ is defined as

$$cer[x_1, x_n] = \prod_{i=1}^{n-1} cer(x_i, x_{i+1}), \tag{5}$$

the *coverage* of a path $[x_1, x_n]$ is

$$cov[x_1, x_n] = \prod_{i=1}^{n-1} cov(x_i, x_{i+1}), \tag{6}$$

and the *strength* of a path $[x, y]$ is

$$\sigma[x, y] = \sigma(x)\, cer[x, y] = \sigma(y)\, cov[x, y]. \tag{7}$$

The set of all paths from x to y ($x \neq y$) denoted $< x, y >$, will be called a *connection* from x to y. In other words, connection $< x, y >$ is a sub-graph determined by nodes x and y.

The *certainty* of connection $< x, y >$ is

$$cer < x, y >= \sum_{[x,y] \in < x,y >} cer[x, y], \tag{8}$$

the *coverage* of connection $< x, y >$ is

$$cov < x, y >= \sum_{[x,y] \in < x,y >} cov[x, y], \tag{9}$$

and the *strength* of connection $< x, y >$ is

$$\sigma < x, y >= \sum_{[x,y] \in < x,y >} \sigma[x, y]. \tag{10}$$

Let x, y ($x \neq y$) be nodes of G. If we substitute the sub-graph $< x, y >$ by a single branch (x, y) such that $\sigma(x, y) = \sigma < x, y >$, then $cer(x, y) = cer < x, y >$, $cov(x, y) = cov < x, y >$ and $\sigma(G) = \sigma(G')$, where G' is the graph obtained from G by substituting $< x, y >$ by (x, y).

Example 1 (cont). In order to find how car models are distributed among customer groups we have to compute all connections among cars models and consumers groups. The results are shown in Fig. 3.

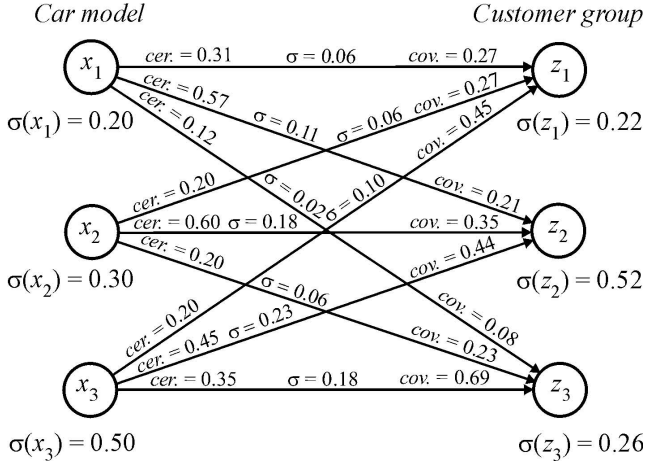

Fig. 3. Relation between car models and consumer groups

For example, we can see from the flow graph that consumer group z_2 bought 21% of car model x_1, 35% – of car model x_2 and 44% – of car model x_3. Conversely, for example, car model x_1 is distributed among customer groups as follows: 31% cars bought group z_1, 57% – group z_2 and 12% – group z_3.

5 Decision Algorithms

With every branch (x, y) we associate a decision rule $x \rightarrow y$, read *if x then y*; x will be referred to as a *condition*, whereas y – *decision* of the rule. Such a rule is characterized by three numbers, $\sigma(x, y)$, $cer(x, y)$ and $cov(x, y)$.

Thus every path $[x_1, x_n]$ determines a sequence of decision rules $x_1 \to x_2$, $x_2 \to x_3, \ldots, x_{n-1} \to x_n$.

From previous considerations it follows that this sequence of decision rules can be interpreted as a single decision rule $x_1 x_2 \ldots x_{n-1} \to x_n$, in short $x^* \to x_n$, where $x^* = x_1 x_2 \ldots x_{n-1}$, characterized by

$$cer(x^*, x_n) = cer[x_1, x_n], \tag{11}$$

$$cov(x^*, x_n) = cov[x_1, x_n], \tag{12}$$

and

$$\sigma(x^*, x_n) = \sigma(x_1)\, cer[x_1, x_n] = \sigma(x_n)\, cov[x_1, x_n]. \tag{13}$$

Similarly, every connection $< x, y >$ can be interpreted as a single *decision rule* $x \to y$ such that:

$$cer(x, y) = cer< x, y >, \tag{14}$$

$$cov(x, y) = cov< x, y >, \tag{15}$$

and

$$\sigma(x, y) = \sigma(x) cer< x, y > = \sigma(y) cov< x, y > . \tag{16}$$

Let $[x_1, x_n]$ be a path such that x_1 is an input and x_n an output of the flow graph G, respectively. Such a path and the corresponding connection $< x_1, x_n >$ will be called *complete*.

The set of all decision rules $x_{i_1} x_{i_2} \ldots x_{i_{n-1}} \to x_{i_n}$ associated with all complete paths $[x_{i_1}, x_{i_n}]$ will be called a *decision algorithm* induced by the flow graph.

The set of all decision rules $x_{i_1} \to x_{i_n}$ associated with all complete connections $< x_{i_1}, x_{i_n} >$ in the flow graph, will be referred to as the *combined decision algorithm* determined by the flow graph.

Example 1 (cont.). The decision algorithm induced by the flow graph shown in Fig. 2 is given below:

Rule no.	*Rule*	*Strength*
1)	$x_1 y_1 \to z_1$	0.036
2)	$x_1 y_1 \to z_2$	0.072
3)	$x_1 y_1 \to z_3$	0.012
.....		
20)	$x_3 y_4 \to z_1$	0.025
21)	$x_3 y_4 \to z_2$	0.075
22)	$x_3 y_4 \to z_3$	0.150

For the sake of simplicity we gave only some of the decision rules of the decision algorithm. Interested reader can easily complete all the remaining decision rules. Similarly we can compute certainty and coverage for each rule.

Remark 1. Due to round-off errors in computations, the equalities (1)...(16) may not be satisfied exactly in these examples.

The combined decision algorithm associated with the flow graph shown in Fig. 3, is given below:

Rule no.	*Rule*	*Strength*
1)	$x_1 \to z_1$	0.06
2)	$x_1 \to z_2$	0.11
3)	$x_1 \to z_3$	0.02
4)	$x_2 \to z_1$	0.06
5)	$x_2 \to z_2$	0.18
6)	$x_2 \to z_3$	0.06
7)	$x_3 \to z_1$	0.10
8)	$x_3 \to z_2$	0.23
9)	$x_3 \to z_3$	0.18

This decision algorithm can be regarded as a simplification of the decision algorithm given previously and shows how car models are distributed among customer groups.

6 Independence of Nodes in Flow Graphs

Let x and y be nodes in a flow graph $G = (N, \mathtt{B}, \sigma)$, such that $(x,y) \in \mathtt{B}$.

Nodes x and y are *independent* in G if

$$\sigma(x, y) = \sigma(x)\,\sigma(y). \tag{17}$$

From (17) we get

$$\frac{\sigma(x, y)}{\sigma(x)} = cer(x, y) = \sigma(y), \tag{18}$$

and

$$\frac{\sigma(x, y)}{\sigma(y)} = cov(x, y) = \sigma(x). \tag{19}$$

If

$$cer(x,y) > \sigma(y), \tag{20}$$

or

$$cov(x, y) > \sigma(x), \tag{21}$$

then y *depends positively* on x in G.

Similarly, if

$$cer(x, y) < \sigma(y), \tag{22}$$

or

$$cov(x, y) < \sigma(x), \tag{23}$$

then y *depends negatively* on x in G.

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

Example 1 (cont.). In flow graphs presented in Fig. 2 and Fig. 3 there are no independent nodes, whatsoever.

However, e.g. nodes x_1,y_1 are positively dependent, whereas, nodes y_1,z_3 are negatively dependent.

Example 2. Let $X = \{1,2,\ldots,8\}$, $x \in X$ and let a_1 denote "x is divisible by 2", a_0 – "x is not divisible by 2". Similarly, b_1 stands for "x is divisible by 3" and b_0 – "x is not divisible by 3". Because there are 50% elements divisible by 2 and 50% elements not divisible by 2 in X, therefore we assume $\sigma(a_1) = ½$ and $\sigma(a_0) = ½$. Similarly, $\sigma(b_1) = ¼$ and $\sigma(b_0) = ¾$ because there are 25% elements divisible by 3 and 75% not divisible by 3 in X, respectively.

The corresponding flow graph is presented in Fig. 4.

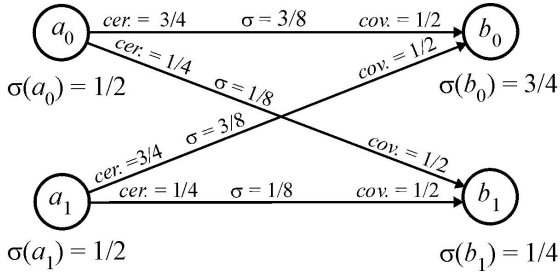

Fig. 4. Divisibility by "2" and "3"

The pair of nodes (a_0,b_0), (a_0, b_1), (a_1,b_0) and (a_1,b_1) are independent, because, e.g., $cer(a_0,b_0) = \sigma(b_0)$ ($cov(a_0,b_0) = \sigma(a_0)$).

Example 3. Let $X = \{1,2,\ldots,8\}$, $x \in X$ and a_1 stand for "x is divisible by 2", a_0 – "x is not divisible by 2", b_1 – "x is divisible by 4" and b_0 – "x is not divisible by 4". As in the previous example $\sigma(a_0) = ½$ and $\sigma(a_1) = ½$; $\sigma(b_0) = ¾$ and $\sigma(b_1) = ¼$ because there are 75% dements not divisible by 4 and 25% divisible by 4 – in X.

The flow graph associated with the above problem is shown in Fig. 5.

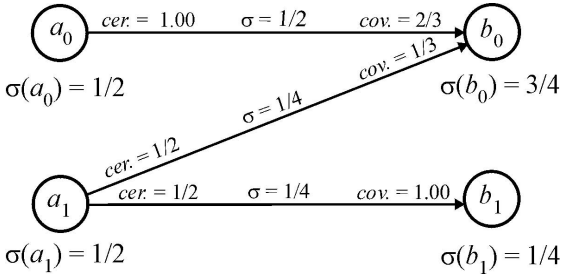

Fig. 5. Divisibility by "2" and "4"

The pairs of nodes (a_0,b_0), (a_1,b_0) and (a_1,b_1) are dependent. Pairs (a_0,b_0) and (a_1,b_1) are positively dependent, because $cer(a_0,b_0) > \sigma(b_0)$ ($cov(a_0,b_0) > \sigma(a_0)$) and – $cer(a_1,b_1) > \sigma(b_1)$ ($cov(a_1,b_1) > \sigma(a_1)$). Nodes (a_1,b_0) are negatively dependent, because $cer(a_1,b_0) < \sigma(b_0)$ ($cov(a_1,b_0) < \sigma(a_1)$).

For every branch $(x,y)\in B$ we define a *dependency factor* $\eta(x,y)$ defined as

$$\eta(x,y)=\frac{cer(x,y)-\sigma(y)}{cer(x,y)+\sigma(y)}=\frac{cov(x,y)-\sigma(x)}{cov(x,y)+\sigma(x)}. \tag{24}$$

Obviously $-1\le\eta(x,y)\le 1$; $\eta(x,y)=0$ if and only if $cer(x,y)=\sigma(y)$ and $cov(x,y)=\sigma(x)$; $\eta(x,y)=-1$ if and only if $cer(x,y)=cov(x,y)=0$; $\eta(x,y)=1$ if and only if $\sigma(y)=\sigma(x)=0$.

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

For example, in the flow graph presented in Fig. 4 we have: $\eta(a_0,b_0)=0$, $\eta(a_0,b_1)=0$, $\eta(a_1,b_0)=0$ and $\eta(a_1,b_1)=0$. However, in the flow graph shown in Fig. 5 we have $\eta(a_0,b_0)=1/7$, $\eta(a_1,b_0)=-1/5$ and $\eta(a_1,b_1)=1/3$.

The meaning of the above results is obvious.

7 Conclusions

In this paper a relationship between flow graphs and decision algorithms has been defined and studied. It has been shown that the information flow in a decision algorithm can be represented as a flow in the flow graph. Moreover, the flow is governed by Bayes' formula, however the Bayes' formula has entirely deterministic meaning, and is not referring to its probabilistic nature. Besides, the formula has a new simple form, which essentially simplifies the computations.

This leads to many new applications and also gives new insight into the Bayesian philosophy.

Acknowledgement. Thanks are due to Professor Andrzej Skowron for critical remarks.

References

1. Bernardo, J. M., Smith, A. F. M.: Bayesian Theory. Wiley series in probability and mathematical statistics. John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore (1994)
2. Box, G.E.P., Tiao, G.C.: Bayesian Inference in Statistical Analysis. John Wiley and Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore (1992)
3. Ford, L.R., Fulkerson, D.R.: Flows in Networks. Princeton University Press, Princeton. New Jersey

4. Łukasiewicz, J.: Die logischen Grundlagen der Wahrscheinlichkeitsrechnung. Kraków (1913). In: L. Borkowski (ed.), Jan Łukasiewicz – Selected Works, North Holland Publishing Company, Amsterdam, London, Polish Scientific Publishers, Warsaw (1970)
5. Greco, S., Pawlak, Z., Słowiński, R.: Generalized Decision Algorithms, Rough Inference Rules, and Flow Graphs, in: J.J. Alpigini *et al*. (eds.), Lecture Notes in Artificial Intelligence 2475 (2002) 93–104
6. 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
7. Pawlak, Z.: Rough Sets, Decision Algorithms and Bayes' Theorem. European Journal of Operational Research 136 (2002) 181–189
8. Pawlak, Z.: Decision Rules and Flow Networks (to appear)
9. Tsumoto, S., Tanaka, H.: Discovery of Functional Components of Proteins Based on PRIMEROSE and Domain Knowledge Hierarchy, Proceedings of the Workshop on Rough Sets and Soft Computing (RSSC-94), 1994: Lin, T.Y., and Wildberger, A.M. (eds.), *Soft Computing*, SCS (1995) 280–285.
10. Wong, S.K.M., Ziarko, W.: Algorithm for Inductive Learning. Bull. Polish Academy of Sciences 34, 5–6 (1986) 271–276