Flow Graphs – a New Paradigm for Data Mining and Knowledge Discovery

Zdzisław Pawlak
Institute of Theoretical and Applied Informatics, Polish Academy of Sciences
ul. Bałtycka 5, 44 100 Gliwice, Poland
and
Warsaw School of Information Technology, ul. Newelska 6, 01-447 Warsaw, Poland
zpw@ii.pw.edu.pl

ABSTRACT

In this paper we propose a new approach to data (mining) and knowledge discovery based on information flow distribution study in a flow graph. Flow graphs introduced in this paper are different from those proposed by Ford and Fulkerson for optimal flow analysis and they model rather, e.g., flow *distribution* in a network, than the optimal flow. The flow graphs considered in this paper are not meant to physical media (e.g., water) flow analysis, but to information flow examination in decision algorithms. It is revealed that flow in the flow graph is governed by Bayes' rule, but the rule has entirely deterministic interpretation, not referring to its probabilistic roots. Besides, decision algorithm induced by the flow graph and dependency between conditions and decisions of decision rules are defined and studied. This idea is based on statistical concept of dependency but in our setting it has deterministic meaning.

Keywords: flow graph, data mining, knowledge discovery

1. INTRODUCTION

In this paper we propose a new approach to data (mining) and knowledge discovery based on information flow distribution study in a flow graph.

Flow graphs introduced in this paper are different from those proposed by Ford and Fulkerson [1] for optimal flow analysis and they model rather, e.g., flow *distribution* in a network, than the optimal flow.

The flow graphs considered in this paper are not meant to physical media (e.g., water) flow analysis, but to information flow examination in decision algorithms. To this end branches of a flow graph can be interpreted as decision rules. With every decision rule (i.e. branch) three coefficients are associated, the *strength*, *certainty* and *coverage factors*.

This coefficient have been used under different names in data mining (see e.g., [2], [3]) but they were used first by Łukasiewicz [4] in his study of logic and probability.

This interpretation, in particular, leads to a new look on Bayes' theorem. Let us also observe that despite Bayes' rule fundamental role in statistical inference it has led to many philosophical discussions concerning its validity and meaning, and has caused much criticism [5], [6], [7].

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

First we introduce basic concepts of the proposed approach, i.e., flow graph and its fundamental properties. It is revealed that flow in the flow graph is governed by Bayes' rule, but the rule has entirely deterministic interpretation, not referring to its probabilistic roots. Besides decision algorithm induced by the flow graph and dependency between conditions and decisions of decision rules are defined and studied. This idea is based on statistical concept of dependency but in our setting it has deterministic meaning. Simple tutorial examples is used to illustrate how the introduced ideas work in data mining. The presented ideas can be used, as a new tool for data mining, and knowledge discovery. Besides, it also throw a new light on the concept of probability.

2. FLOW GRAPHS

2.1 Overview

In this part the fundamental concepts of the proposed approach are defined and discussed. In particular flow graphs, certainty and coverage factors of branches of the flow graph are defined and studied. Next these coefficient are extended to paths and some classes of sub-graphs, called connections. Further a notion of a fusion of a flow graph is defined.

Further dependences of flow are introduced and examined. Finally dependency factor (correlation coefficient) is defined.

2.2 Basic Concepts

A flow graph is a *directed, acyclic, finite* graph $G = (N, \mathcal{B}, \varphi)$, where N is a set of *nodes*, $\mathcal{B} \subseteq N \times N$ is a set of *directed branches,* $\varphi : \mathcal{B} \to R^+$ is a *flow function* and R^+ is the set of non-negative reals. *Input* of a node $x \in N$ *is the set* $I(x)=\{y \in N: (y,x) \in \mathcal{B}\}$; *output* of a node $x \in N$ is defined as $O(x) = \{y \in N: (x,y) \in \mathcal{B}\}$.We will also need the concept of *input* and *output* of a graph G, defined, respectively, as follows: $I(G) = \{x \in N : I(x) = \emptyset\}$, $O(G) = \{x \in N : O(x) = \emptyset\}$. Inputs and outputs of G are *external nodes* of G; other nodes are *internal nodes* of G.

If $(x,y) \in \mathcal{B}$ then $\varphi(x,y)$ is a *throughflow* from x to y.

With every node x of a flow graph G we associate its *inflow*

$$\varphi_+(x) = \sum_{y \in I(x)} \varphi(y,x) , \tag{1}$$

and *outflow*

$$\varphi_-(x) = \sum_{y \in O(x)} \varphi(x,y) , \tag{2}$$

Similarly, we define an inflow and an outflow for the whole flow graph, which are defined as

$$\varphi_+(G) = \sum_{x \in I(G)} \varphi_-(x) , \tag{3}$$

$$\varphi_-(G) = \sum_{x \in O(G)} \varphi_+(x) . \tag{4}$$

We assume that for any internal node x, $\varphi_+(x) = \varphi_-(x) = \varphi(x)$, where $\varphi(x)$ is a *throughflow* of node x.

Obviously, $\varphi_+(G) = \varphi_-(G) = \varphi(G)$, where $\varphi(G)$ is a *throughflow* of graph G.

The above formulas can be considered as *flow conservation equations* [1].

Example. We will illustrate basic concepts of flow graphs by an example of a group of 1000 patients put to the test for certain drug effectiveness.

Assume that patients are grouped according to presence of the disease, age and test results, as shown in Fig. 1.

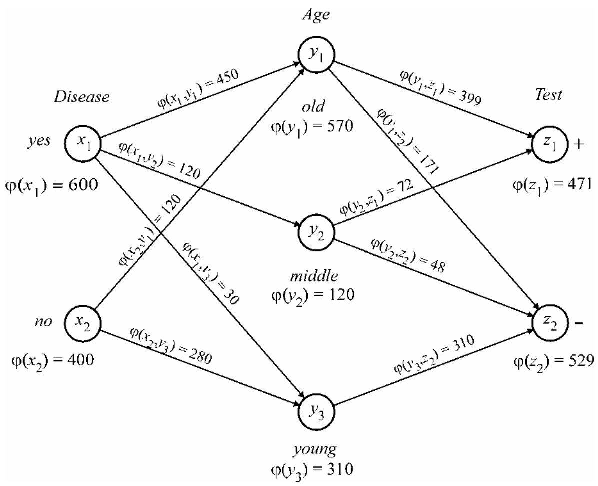

Fig. 1. Flow graph

E.g., $\varphi(x_1) = 600$ means that these are 600 patients suffering from the disease, $\varphi(y_1) = 570$ means that these are 570 old patients $\varphi(z_1) = 471$ means that 471 patients have positive test result; $\varphi(x_1, y_1) = 450$ means that these are 450 old patients which suffer from disease etc.

Thus the flow graph gives clear insight into the relationship between different groups of patients.

Let us now explain the flow graph in more details.

Nodes of the flow graph are depicted by circles, labeled by x_1, x_2, y_1, y_2, y_3, z_1, z_2. A branch (x, y) is denoted by an arrow from node x to y. E.g., branch (x_1, z_1) is represented by an arrow from x_1 to z_1, inputs of node y_1 are nodes x_1 and x_2, outputs of node x_1 are nodes y_1, y_2 and y_3.

Inputs of the flow graph are nodes x_1 and x_2 and x_3, whereas outputs of the flow graph are nodes z_1 and z_2.

Nodes y_1, y_2 and y_3 are internal nodes of the flow graph. The throughflow of the branch (x_1, y_1) is $\varphi(x_1, y_1) = 450$. Inflow of node y_1 is $\varphi_+(y_1) = 450 + 120 = 570$. Outflow of node y_1 is $\varphi_-(y_1) = 399 + 171 = 570$.

Inflow of the flow graph is $\varphi(x_1) + \varphi(x_2) =$ 600 + 400 = 1000, and outflow of the flow graph is $\varphi(z_1) + \varphi(z_2) =$ 471 + 529 = 1000.

Throughflow of node $y_1 = \varphi(y_1) = \varphi(x_1, y_1) + \varphi(x_2, y_1)$ $= \varphi(y_1, z_1) + \varphi(y_2, z_2) = 570$. □

We will define now a *normalized flow graph*. A normalized flow graph is a *directed, acyclic*, *finite* graph $G = (N, \mathcal{B}, \sigma)$, where N is a set of *nodes*, $\mathcal{B} \subseteq N \times N$ is a set of *directed branches* and $\sigma: \mathcal{B} \to$ <0,1> is *a normalized flow* of (x, y) and

$$\sigma(x,y) = \frac{\varphi(x,y)}{\varphi(G)} \tag{5}$$

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

In what follows we will use normalized flow graphs only, therefore by flow graphs we will understand normalized flow graphs, unless stated otherwise.

With every node x of a flow graph G we associate its *inflow* and *outflow* defined as

$$\sigma_+(x) = \frac{\varphi_+(x)}{\varphi(G)} = \sum_{y \in I(x)} \sigma(y,x) , \tag{6}$$

$$\sigma_-(x) = \frac{\varphi_-(x)}{\varphi(G)} = \sum_{y \in O(x)} \sigma(x,y). \tag{7}$$

Obviously for any internal node x, we have $\sigma_+(x) = \sigma_-(x) = \sigma(x)$, where $\sigma(x)$ is a *normalized throughflow* of x.

Moreover, let

$$\sigma_+(G) = \frac{\varphi_+(G)}{\varphi(G)} = \sum_{x \in I(G)} \sigma_-(x) \,, \tag{8}$$

$$\sigma_-(G) = \frac{\varphi_-(G)}{\varphi(G)} = \sum_{x \in O(G)} \sigma_+(x) \,. \tag{9}$$

Obviously, $\sigma_+(G) = \sigma_-(G) = \sigma(G) = 1$.

Example (cont.). The normalized flow graph of the flow graph presented in Fig. 1 is given in Fig. 2.

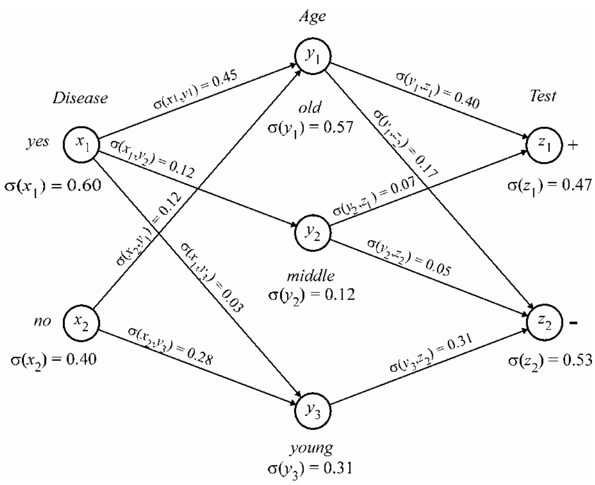

Fig. 2. Normalized flow graph

In the flow graph e.g., $\sigma(x_1) = 0.60$, that means that 60% of total inflow is associated with input x_1. The strength $\sigma(x_1, y_1) = 0.45$ means that 45% of total flow flows through the branch (x_1, y_1) . etc. □

Let $G = (N, \mathscr{B}, \sigma)$ be a flow graph. If we invert direction of all branches in G, then the resulting graph $G = (N, \mathscr{B}', \sigma')$ will be called an *inverted* graph of G. Of course the inverted graph G' is also a flow graph and all inputs and outputs of G become inputs and outputs of G', respectively.

2.3 Certainty and Coverage Factors

With every branch (x, y) of a flow graph G we associate the *certainty* and the *coverage factors*.

The *certainty* and the *coverage* of (x, y) are defined as

$$cer(x, y) = \frac{\sigma(x, y)}{\sigma(x)} \,, \tag{10}$$

and

$$cov(x, y) = \frac{\sigma(x, y)}{\sigma(y)} \,. \tag{11}$$

respectively.

Evidently, $cer(x, y) = cov(y, x)$, where $(x, y) \in \mathscr{B}$ and $(y, x) \in \mathscr{B}'$.

Example (cont.). The certainty and the coverage factors for the flow graph presented in Fig. 2 are shown in Fig. 3.

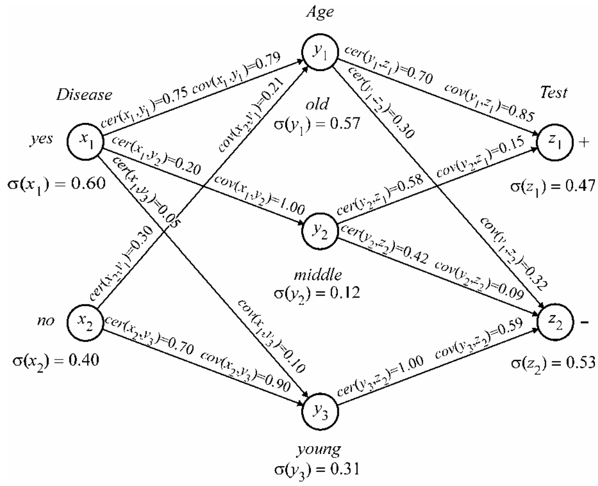

Fig. 3. Certainty and coverage

E.g., $cer(x_1, y_1) = \dfrac{\sigma(x_1, y_1)}{\sigma(x_1)} = \dfrac{0.45}{0.60} = 0.75$,and $cov(x_1, y_1) = \dfrac{\sigma(x_1, y_1)}{\sigma(y_1)} = \dfrac{0.45}{0.57} \approx 0.21$. □

Example (cont.). The inverted flow graph of the flow graph from Fig. 3 is shown in Fig. 4.

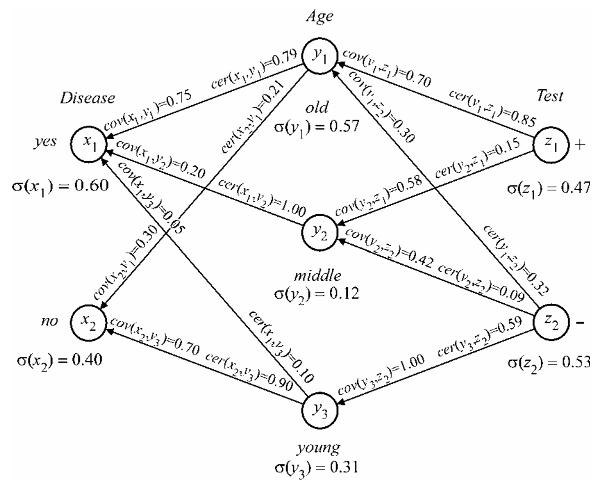

Fig. 4. Inverted flow graph

□

Below some properties of certainty and coverage factors, which are immediate consequences of definitions given above, are presented:

$$\sum_{y \in O(x)} cer(x, y) = 1 \,, \tag{12}$$

$$\sum_{x \in I(y)} cov(x, y) = 1 , \tag{13}$$

$$\sigma(x) = \sum_{y \in O(x)} cer(x, y)\sigma(x) = \sum_{y \in O(x)} \sigma(x, y) , \tag{14}$$

$$\sigma(y) = \sum_{x \in I(y)} cov(x, y)\sigma(y) = \sum_{x \in I(y)} \sigma(x, y) , \tag{15}$$

$$cer(x, y) = \frac{cov(x, y)\sigma(y)}{\sigma(x)} , \tag{16}$$

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

Obviously the above properties have a probabilistic flavor, e.g., equations (14) and (15) have a form of total probability theorem, whereas formulas (16) and (17) are Bayes' rules. However, these properties in our approach are interpreted in a deterministic way and they describe flow distribution among branches in the network.

2.4 Paths, Connections and Fusion

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

The *certainty* of the path $[x_1\ldots x_n]$ is defined as

$$cer[x_1 \ldots x_n] = \prod_{i=1}^{n-1} cer(x_i, x_{i+1}) , \tag{18}$$

the *coverage* of the path $[x_1\ldots x_n]$ is

$$cov[x_1 \ldots x_n] = \prod_{i=1}^{n-1} cov(x_i, x_{i+1}) , \tag{19}$$

and the *strength* of the path $[x\ldots y]$ is

$$\sigma[x...y] = \sigma(x)\, cer[x...y] = \sigma(y)\, cov[x...y]. \tag{20}$$

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

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

$$cer < x, y > \; = \sum_{[x\ldots y] \in <x,y>} cer[x \ldots y] , \tag{21}$$

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

$$cov < x, y > \; = \sum_{[x\ldots y] \in <x,y>} cov[x \ldots y] , \tag{22}$$

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

$$\sigma < x, y > \; = \sum_{[x\ldots y] \in <x,y>} \sigma[x \ldots y] = \sigma(x) cer < x, y >= \sigma(y) \operatorname{cov} < x, y > . \tag{23}$$

Let x, y ($x \neq y$) be nodes of G. If we substitute simultaneously every for the sub-graph $< x, y >$ of a given flow graph G, where x and y are input and output nodes of G respectively, by single branch (x, y) such that $\sigma(x, y)$, then in the resulting graph G', called the *fusion* of G, we have $cer(x, y) = cer < x, y >$, $cov(x, y) = cov < x, y >$ and $\sigma(G) = \sigma(G')$.

Example (cont.). In the flow graph presented in Fig. 3 for the path $p = [x_1, y_1, z_1]$ we have $cer(p) = 0.75 \times 0.70 \approx 0.53$, $cov(p) = 0.85 \times 0.79 \approx 0.67$.

The connection $< x_1, z_1 >$ in the flow graph consists of paths $[x_1, , y_1, z_1]$ and $[x_1, y_2, z_1]$. This connection is shown in Fig. 5.

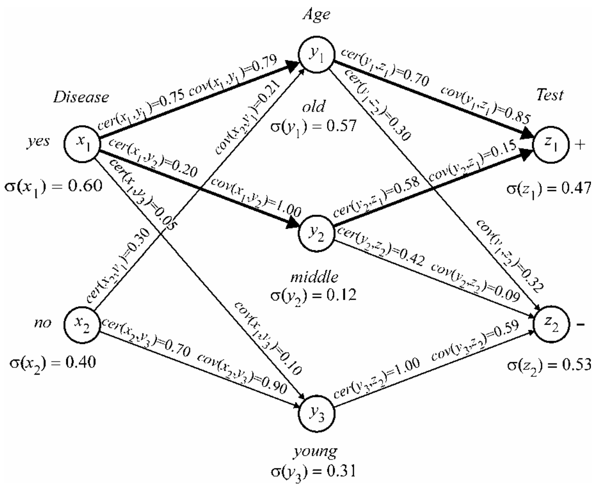

Fig. 5. Connection $< x_1, z_1 >$

For this connection we have $cer <x_1, z_1> = 0.75 \times 0.70 + 0.20 \times 0.58 \approx 0.64$; $cov<x_1,z_1> = 0.85 \times 0.79 + 0.15 \times 1.00 \approx 0.82$.

The strength of the connection x_1, z_1 is $0.64 \times 0.60 \approx 0.82 \times 0.47 \approx 0.38$.

Connections $< x_1, z_2 >$, $< x_2, z_1 >$ and $< x_2, z_1 >$ are presented in Fig. 6, Fig. 7 and Fig. 8, respectively.

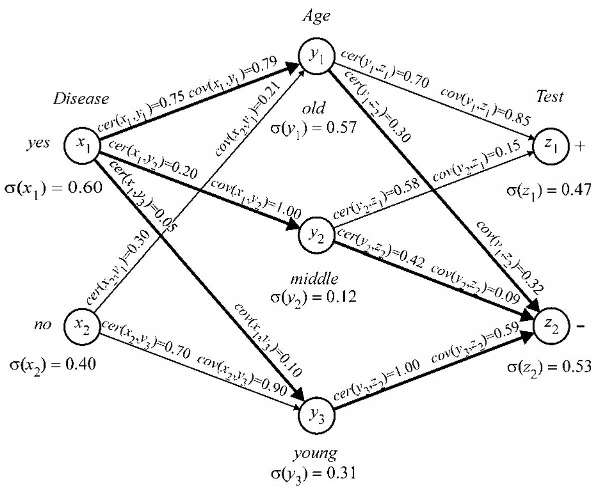

Fig. 6. Connection $< x_2, z_1 >$

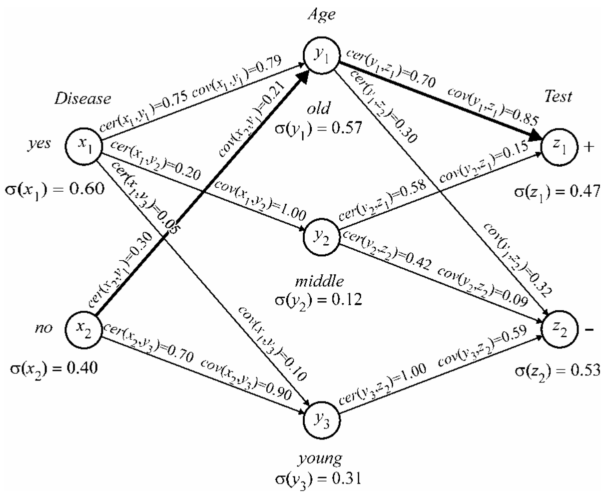

Fig. 7. Connection $< x_2, z_3 >$

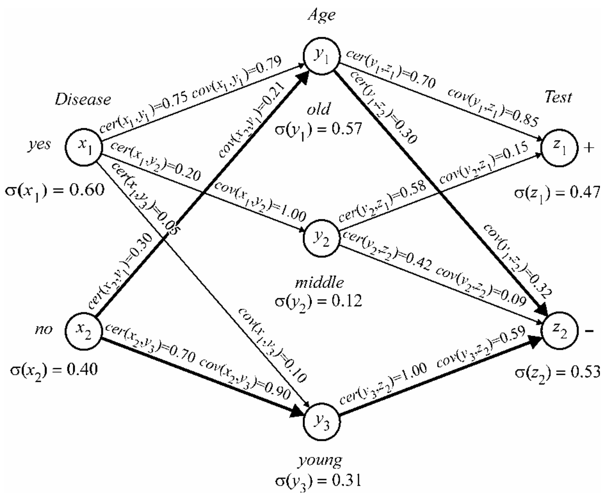

Fig. 8. Connection $< x_1, z_3 >$

□

Example (cont.). The fusion of the flow graph shown in Fig. 3 is given in Fig. 9.

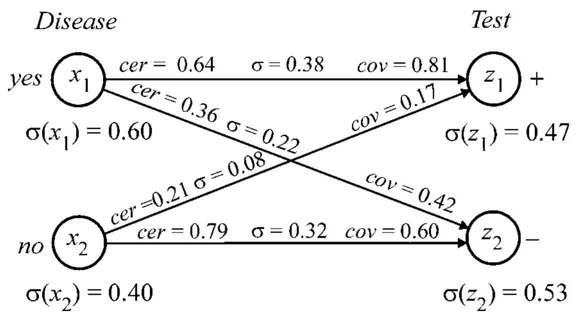

Fig. 9. Fusion of a flow graph

The fusion of a flow graph gives information about the flow distribution between input and output of the flow graph, i.e., it leads to the following conclusions:

- if the disease is present then the test result is positive with certainty 0.64,
- it the disease is absent then the test result is negative with certainty 0.79.

Explanation of test results is as follows:

- if the test result is positive then the disease is present with certainty 0.81,
- if the test result is negative then the disease is absent with certainty 0.60. □

2.5 Dependences in Flow Graphs

Let x and y be nodes in a flow graph $G = (N, \mathcal{B}, \sigma)$, such that $(x,y)\in\mathcal{B}$. Nodes x and y are *independent* in G if

$$\sigma(x,y) = \sigma(x)\,\sigma(y). \tag{24}$$

From (21) we get

$$\frac{\sigma(x,y)}{\sigma(x)} = cer(x,y) = \sigma(y)\,, \tag{25}$$

and

$$\frac{\sigma(x,y)}{\sigma(y)} = cov(x,y) = \sigma(x). \tag{26}$$

If

$$cer(x,y) > \sigma(y), \tag{27}$$

or

$$cov(x,y) > \sigma(x), \tag{28}$$

then x and y are *positively dependent* on x in G. Similarly, if

$$cer(x,y) < \sigma(y), \tag{29}$$

or

$$cov(x,y) < \sigma(x), \tag{30}$$

then x and y are *negatively dependent* in G.

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

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

$$\eta(x,y) = \frac{cer(x,y)-\sigma(y)}{cer(x,y)+\sigma(y)} = \frac{cov(x,y)-\sigma(x)}{cov(x,y)+\sigma(x)}. \tag{31}$$

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.

Example (cont.). Dependency factors for the flow graph shown in Fig. 9 are given Fig. 10.

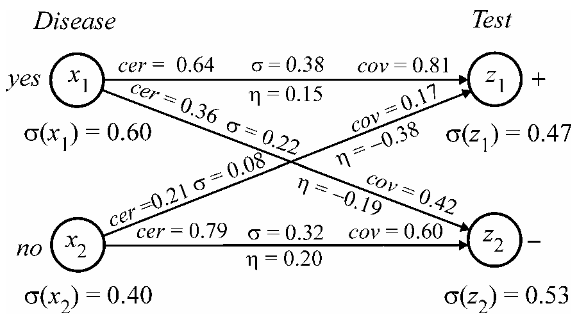

Fig. 10. Dependencies in a flow graph

Thus, there is positive dependency between presence of the disease and positive test result as well as between absence of disease and negative test result. However there is much stronger negative dependency between presence of the disease and negative test result or similarly – between absence of the disease and positive test result. □

2.6 Flow Graph and Decision Algorithms

Flow graphs can be interpreted as decision algorithms [8].

Let us assume that the set of nodes of a flow graph is interpreted as a set of logical formulas. The formulas are understood as propositional functions and if *x* is a formula, then $\sigma(x)$ is to be interpreted as a truth value of the formula. Let us observe that the truth values are numbers from the closed interval <0, 1>, i.e., $0 \le \sigma(x) \le 1$.

According to [4] these truth values can be also interpreted as probabilities. Thus $\sigma(x)$ can be understood as flow distribution ratio (percentage), truth value or probability. We will stick to the first interpretation.

With every branch (x, y) we associate a decision rule $x \to y$, read *if x then y*; *x* will be referred to as *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 $[x_1 \ldots x_n]$ determines a sequence of decision $x_1 \to x_2$, $x_2 \to x_3, \ldots, x_{n-1} \to x_n$.

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

$$cer(x^*, x_n) = \frac{\sigma(x^*, x_n)}{\sigma(x^*)}, \qquad (32)$$

$$cov\,(x^*, x_n) = \frac{\sigma(x^*, x_n)}{\sigma(x_n)}, \qquad (33)$$

and

$$\begin{aligned}\sigma(x^*, x_n) = \sigma(x_1) cer[x_1 \ldots x_n] = \sigma(x_n) = \\ = \mathrm{cov}[x_1 \ldots x_n]\end{aligned}, \qquad (34)$$

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

If $x \to y$ is a decision rule then we say that condition and decision of the decision rule are independent if *x* and *y* are independent, otherwise condition and decision of the decision rule are dependent (positively or negatively).

To measure the degree of dependency between condition and decision of the decision rule $x \to y$ we can use the dependency factor $\eta(x, y)$.

Thus every decision rule beside strength, certainty and coverage factor can be also characterized by the degree of dependency between its condition and decision. This measure can be used as a new tool for data mining in pursuit of patterns in data.

Example (cont.). The decision algorithm induced by the flow graph shown in Fig. 3 is given below.

	certainty	coverage	strength
$x_1, y_1 \to z_1$	0.71	0.67	0.32
$x_1, y_1 \to z_2$	0.29	0.25	0.14
$x_1, y_2 \to z_1$	0.58	0.15	0.07
$x_1, y_2 \to z_2$	0.42	0.09	0.05
$x_1, y_3 \to z_2$	0.05	0.06	0.03
$x_2, y_1 \to z_1$	0.67	0.18	0.08
$x_2, y_1 \to z_2$	0.33	0.08	0.04
$x_2, y_3 \to z_2$	1.00	0.53	0.28

The corresponding flow graph is presented in Fig. 11.

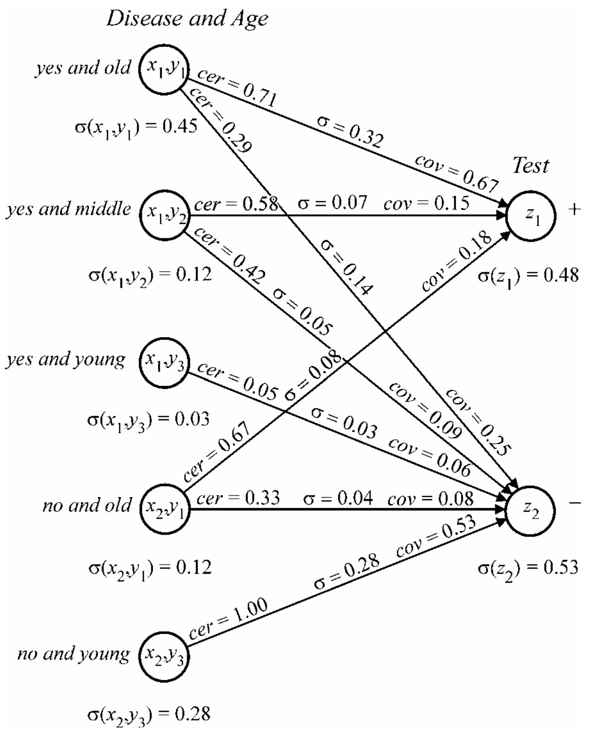

Fig. 11. Flow graph for the decision algorithm

From the decision algorithm we can see e.g., that 71% ill and old patients have positive test result, whereas 100% young healthy patients have negative test results. We can also conclude that positive test result have mostly ill and old patients and negative test result display mostly young healthy patients.

For the above decision rules dependency factors are $\eta \approx 0.19$ and $\eta \approx 0.31$ respectively. That means that the relationship between young healthy patients and negative test results is more substantial then - between ill old patients and positive test result.

CONCLUSIONS

We propose in this paper a new approach to knowledge representation and data mining, based on flow analysis in a new kind of flow networks.

We advocate in this paper to represent relationships in data by means of flow graphs. Flow in the flow graph is meant to capture structure of data rather than to describe any physical material flow in the network. It is revealed the information flow in the flow graph is governed by Bayes' formula, however the formula can interpreted in entirely deterministic way, without referring to its probabilistic character. This representation allows us to study different relationships in data and can be used as a new mathematical tool for data mining.

Summing up

- flow graphs can be used to knowledge representation,
- flow distribution represents relationships in data,
- flow conservation is described by Bayes' formula,
- Bayes' formula has deterministic interpretation.

REFERENCES

1. L. R. Ford, D. R. Fulkerson, *Flows in Networks*. Princeton University Press, Princeton. New Jersey, 1962.
2. S. Tsumoto, H. Tanaka, 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, pp. 280-285.
3. S. K. M. Wong, W. Ziarko, Algorithm for inductive learning. *Bull. Polish Academy of Sciences* 34, 5-6, 1986, pp. 271-276.
4. J. Łukasiewicz, 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.
5. J. M. Bernardo, A. F. M. Smith, *Bayesian Theory*. Wiley series in probability and mathematical statistics. John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore, 1994.
6. G. E. P. Box, G.C. Tiao, *Bayesian Inference in Statistical Analysis*. John Wiley and Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore, 1992.
7. R. Swinburne (ed.), Bayes' Theorem, Oxford University Press, 2002.
8. Z. Pawlak, Rough Sets, Decision algorithms and Bayes' theorem. *European Journal of Operational Research* 136, 2002, pp. 181-189.