 ISBN 978-80-87927-04-5, IISES

MERT BAL
YILDIZ TECHNICAL UNIVERSITY, TURKEY

AYSE DEMIRHAN
YILDIZ TECHNICAL UNIVERSITY, TURKEY

USING FLOW GRAPHS IN DATA MINING

Abstract:

Databases are widely used in data processes and each day their sizes are getting larger. In order to access to the data stored in growing databases and to use them, new techniques are developed to discover the knowledge automatically. Data mining techniques may be used to find the useful knowledge with analyzing and discovering the data. Data mining is the search for the relations and the rules, which help us to make estimations about the future from large-scale databases, using computer programs. Data mining is a process that uses the existing technology and acts as a bridge between data and logical decision-making.
The knowledge discovery from the databases is the determination of different patterns, and defining them in a meaningful, short and unique manner. Knowledge discovery allows using necessary systematical data to obtain the useful patterns from a large database. Knowledge discovery for decision-making processes and market estimations plays an important role in supplying necessary information to business in databases. There are various methods that have been used in data mining such as support vector machines, artificial neural networks, decision trees, genetic algorithms, Bayesian networks, flow graphs etc. Flow graphs proposed by Pawlak are efficient and useful graphical tools that are used in data mining in order to analyze and represent knowledge. In this study; the mathematical background of flow graphs that was proposed by Pawlak will be examined and then an example will be given.

Keywords:

Flow Graph, Data Mining, Decision Algorithms, Knowledge Discovery

JEL Classification: C44

1. INTRODUCTION

Databases are widely used in data processes and each day their sizes are getting larger. In order to access to the data stored in growing databases and to use them, new techniques are developed to discover the knowledge automatically. Data mining techniques may be used to find the useful knowledge with analyzing and discovering the data. Data mining is the search for the relations and the rules, which help us to make estimations about the future from large-scale databases, using computer programs. Data mining is a process that uses the existing technology and acts as a bridge between data and logical decision-making.

The knowledge discovery from the databases is the determination of different patterns, and defining them in a meaningful, short and unique manner. Knowledge discovery allows using necessary systematical data to obtain the useful patterns from a large database. Knowledge discovery for decision-making processes and market estimations plays an important role in supplying necessary information to business in databases. There are various methods that have been used in data mining such as support vector machines, artificial neural networks, decision trees, genetic algorithms, Bayesian networks, flow graphs etc. Flow graphs proposed by Pawlak are efficient and useful graphical tools that are used in data mining in order to analyze and represent knowledge. In this study; the mathematical background of flow graphs that was proposed by Pawlak will be examined and then an example will be given.

2. DATA MINING AND KNOWLEDGE DISCOVERY IN DATABASES

With the increase in the usage of databases and their volumes, the organizations now have a problem about getting benefit from those data. The incapability of the conventional query and reporting tools before big data stacks, the search for new approaches continues under the subject “Knowledge Discovery in Databases – KDD”. In KDD process, Data Mining, which consists of establishment of the model and evaluation of that model, is the most important part. This importance results that KDD and data mining terms are used synonymously by many researchers. The main steps to follow in knowledge discovery in databases are as follows:

1) Definition of the Problem,
2) Preparing the Data,
3) Establishment of the Model and Evaluating that Model,
4) Using the Model,
5) Monitoring the Model.

There are many algorithms to apply data mining process. The main reason of that, some technologies result better than the others do in certain conditions and subjects. In the core of data mining lays the process of establishment of the model that represents a data set. The process of establishment of the model that represents a data set is generic for all data mining products, but the method for model establishment is not generic. In data mining process, various flexible calculation methods can be used such as Rough sets, Bayesian networks, artificial neural networks, decision trees, genetic algorithms, fuzzy sets and inductive logic programming. Data mining processes are used to determine the pattern types that may exist in data mining tasks. Generally, data mining tasks can be divided into two categories as descriptive and predictive mining tasks. Descriptive mining tasks characterize the general specifications of the data residing in database. Predictive mining, on the other hand, makes inferences from the existing data in order to predict (Han & Kamber, 2001). Data mining tasks and the pattern types that are used in the fields such as association rules, classification, clustering, summarizing, prediction, time series analysis, sequence analysis and visualization, respectively. The main concepts and terms about flow graphs will be given in Section 3 and an example will be given in Section 4.

3. FLOW GRAPHS

Flow graph introduced by Pawlak in his paper (Pawlak, 2002a) is a mathematical model to represent and analyze knowledge in databases. Pawlak's flow graph depicts dependent relation among data in virtue of information flow distribution. Given a flow graph, the dependent relation between nodes, i.e., knowledge, can be accurately calculated by the flow capacity passed through them. Due to its close relationship with probability theory, flow graph has been investigated with several theories (e.g., rough sets, decision systems, Bayes' theorem and granular computing). Moreover, the information flow distributions in flow graph are accord with Bayes' formula and abide by flow conservation equations (Pawlak, 2002b)- (Liu.et.al., 2009). The main concepts and terms of flow graph are given below.

3.1. Basic Concepts

A flow graph is a directed, acyclic, finite graph G=(N, B, σ), where N is a set of nodes, $B \subseteq N \times N$ is a set of directed branches, σ: B→<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 σ(x,y) is called the strength of (x,y). Input and output of a graph G, are defined as $I(G)=\{x \in N: I(x)=ø\}$, $O(G)=\{x \in N: O(x)=ø\}$, 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

$$\sigma_+(G)=\sum_{x\in I(G)}\sigma_-(x), \qquad (1)$$

and

$$\sigma_-(G)=\sum_{x\in O(G)}\sigma_+(x), \text{ respectively.} \qquad (2)$$

Obviously $\sigma_{+}(G)=\sigma_{-}(G)=\sigma(G)$, where $\sigma(G)$ is a troughflow of G. Moreover, we assume that $\sigma(G)=1$. (Pawlak, 2003a)

3.2. Properties of Flow Graphs

With every branch of a flow graph we associate the certainty and coverage factors.

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

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

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

Respectively, where $\sigma(x)$ is the normalized trough flow of x, defined by $\sigma(x)=\sum_{y\in O(x)}\sigma(x,y)=\sum_{y\in I(x)}\sigma(y,x)$. Immediate consequences of definitions given above are:

$$\sum_{y\in O(x)} cer(x,y)=1, \tag{4}$$

$$\sum_{x\in I(y)} \operatorname{cov}(x,y)=1, \tag{5}$$

$$cer(x,y)=\frac{\operatorname{cov}(x,y)\sigma(y)}{\sigma(x)}, \tag{6}$$

$$\operatorname{cov}(x,y)=\frac{cer(x,y)\sigma(x)}{\sigma(y)} \tag{7}$$

Explicitly the above properties have a probabilistic character, e.g., equations (5) and (6) can be interpreted as Bayes' formulas. However, these properties can be interpreted in deterministic way and they describe flow distribution among branches in the network (Pawlak, 2003b).

3.3. Paths and Connections

For many applications we will need generalization of the strength, the certainty and coverage factors. A (directed) path from x to y, $x \neq y$ denoted [x,y], is a sequence of nodes $x_1,\ldots\ldots,x_n$ such that x_1= x, x_n=y and $(x_i, x_{i+1}) \in B$ for every i, $1 \le i \le n-1$.The certainty of a path $[x_1,x_n]$ is defined as:

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

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

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

the strength of a path [x,y] is

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

The set of all paths from x to y (x $\neq$ y) denoted (x,y), will be called a connection from x toy. In other words, connection (x,y) is a sub-graph determined by nodes x and y. We will also need extension of the above coefficients for connections (i.e., sub-graphs determined by nodes x and y) as shown in what follows:

The certainty of connection (x,y) is

$$cer(x, y) = \sum_{[x,y] \in <x,y>} cer[x, y],$$

the coverage of connection (x,y) is

$$\operatorname{cov}(x, y) = \sum_{[x,y] \in <x,y>} \operatorname{cov}[x, y],$$

the strength of connection (x,y) is

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

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 $cer(x, y) = cer < x, y >$, $\operatorname{cov}(x, y) = \operatorname{cov} < x, y >$ and $\sigma(G) = \sigma(G')$, where G' is the graph obtained from G by substituting <x,y> by (x,y) (Pawlak, 2004).

3.4. Dependences in Flow Graphs

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

$$\sigma(x, y) = \sigma(x)\sigma(y). \tag{8}$$

From (8) we get

$$\frac{\sigma(x, y)}{\sigma(x)} = cer(x, y) = \sigma(y), \tag{9}$$

and

$$\frac{\sigma(x, y)}{\sigma(y)} = \operatorname{cov}(x, y) = \sigma(x). \tag{10}$$

If

$$cer(x, y) > \sigma(y), \tag{11}$$

or

$$\mathrm{cov}(x, y) > \sigma(x), \tag{12}$$

then x and y positively dependent on x in G.

Similarly, if

$$cer(x, y) < \sigma(y), \tag{13}$$

or

$$\mathrm{cov}(x, y) < \sigma(x), \tag{14}$$

then x and y are negatively dependent in G. (Pawlak, 2005)

3.5. Flow Graph and Decision Algorithms

Flow graphs can be interpreted as decision algorithms. The most general case of this correspondence has been considered in (Greco et al., 2002). 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<$\sigma(x)$<1 (Pawlak, 2006). 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 $\mathrm{cov}(x, y)$. Thus every path [x_1..........x_n] determines a sequence of decision $x_1 \to x_2$, $x_2 \to x_3$,….., $x_{n-1} \to x_n$. From previous considerations it follows that this sequence of decision rules can be interpreted as a single decision rule $x_1 x_2 .... x_{n-1} \to x_n$, in short $x^* \to x_n$, where $x^* = x_1 x_2 .... x_{n-1}$, characterized by

$$cer(x^*, x_n) = \frac{\sigma(x^*, x_n)}{\sigma(x^*)}, \tag{15}$$

$$\mathrm{cov}(x^*, x_n) = \frac{\sigma(x^*, x_n)}{\sigma(x_n)} \tag{16}$$

and

$$\sigma(x^*, x_n) = \sigma(x_1) cer[x_1 .... x_n] = \sigma(x_n) = \mathrm{cov}[x_1 .... x_n] \tag{17}$$

(http://www.bcpw.bg.pw.edu.pl/Content/1956/fgnpdmkd.pdf).

 ISBN 978-80-87927-04-5, IISES

4. AN APPLICATION

This application has been modified from the example in Pawlak (2005) article. Let us consider that there are 3 different plants and 3 different products are produced and the quality of these products is considered as high and poor. We will show the plants with x_1, x_2 and x_3; products with y_1, y_2 and y_3 and product quality with z_1, z_2. We will try to determine the relationship among plants, product and product quality. Firstly, the flows on flow graph are computed. The computed values are shown below in Figure 1.

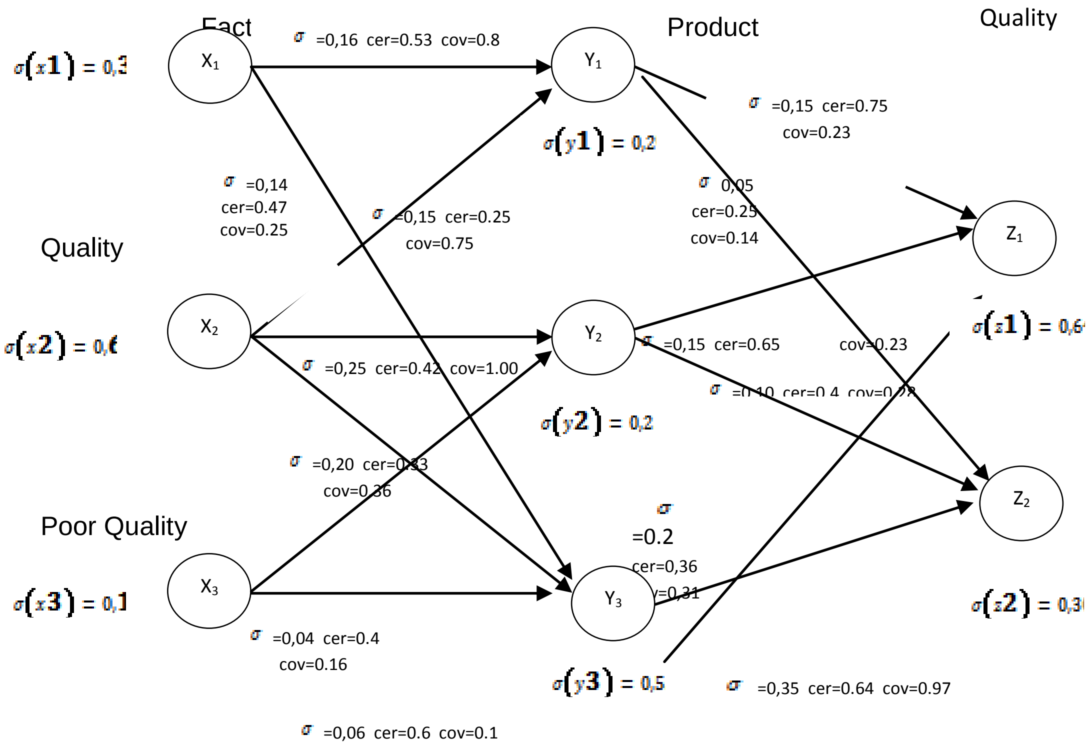

Figure 1: Relationship between factory, product and quality

We will compute the corresponding fusion in order to determine the relationship between plants and product quality. It is shown below in Figure 2.

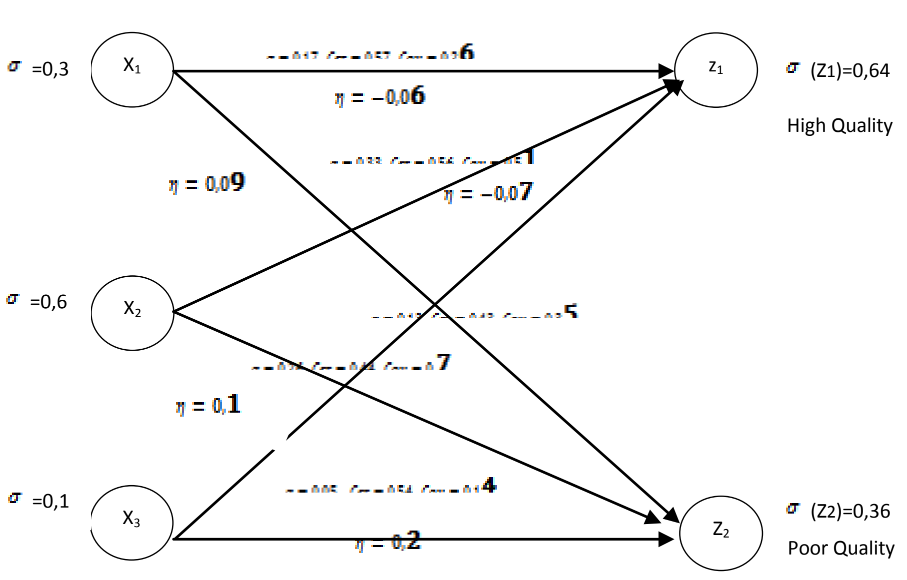

Figure 2: Fusion between factory and quality

It can be seen from dependency factor that all the decision rules except $x_3 \to z_2$ rule have rather low dependency factor. The dependency factor for $x_3 \to z_2$ decision rule is $\eta = 0{,}2$. x_3 plant has the highest ratio for producing low quality products but also the corresponding decision rule has a weakness strength $(\sigma = 0{,}05)$. It has % 5 support value and it means that this plant is the worst one and the quality of its products among all is the worst. The corresponding decision rules are shown below in Table 1.

Table 1: Decision Rules and Its Strength

Rule No	Rules	Strength
1	$X_1Y_1 \to Z_1$	0,3x0,53x0,75= 0,1192
2	$X_1Y_1 \to Z_2$	0,3x0,53x0,25= 0,0397
3	$X_1Y_3 \to Z_1$	0,3x0,47x0,36= 0,0507
4	$X_1Y_3 \to Z_2$	0,3x0,47x0,64= 0,0902
5	$X_2Y_1 \to Z_1$	0,6x0,25x0,75= 0,1125
6	$X_2Y_1 \to Z_2$	0,6x0,25x0,25= 0,0375
7	$X_2Y_2 \to Z_1$	0,6x0,42x0,6 = 0,1512
8	$X_2Y_2 \to Z_2$	0,6x0,42x0,4 = 0,1008
9	$X_2Y_3 \to Z_1$	0,6x0,33x0,36 = 0,0713
10	$X_2Y_3 \to Z_2$	0,6x0,33x0,64 = 0,1267
11	$X_3Y_2 \to Z_1$	0,1x0,4x0,6 = 0,0240
12	$X_3Y_2 \to Z_2$	0,1x0,4x0,4 = 0,0160
13	$X_3Y_3 \to Z_1$	0,1x0,6x0,36 = 0,0216
14	$X_3Y_3 \to Z_2$	0,1x0,6x0,64 = 0,0384

5. CONCLUSION

The size of data stored in data bases have been increasing rapidly. Exploring the useful and efficient patterns in this database has an importance for strategic decision making process. There are various data mining methods that can be used in data analysis such as decision trees, artificial neural network, support vector machines, Bayesian networks and flow graphs. One of these methods is flow graphs that are proposed by Pawlak is a graphical method that can be used in knowledge discovery and knowledge representation. Flow graphs are one of the most efficient methods in rule-induction. In this study, the mathematical background of flow graphs is analyzed and an application is given.

REFERENCES

Han, J., Kamber, M.(2001). *Data mining: concepts and techniques*, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers.

Greco, S., Pawlak, Z., Slowinski, R.(2002).*Generalized decision algorithms, rough inference rules and flow graphs*. In J,J. Alpigini, J.F.Peters, A.Skowron, N.Zhong(Ed.). Rough Sets and Current Trends in Computing. Lecture Notes in Artificial Intelligence 2475, 93-104, Springer-Verlag Berlin

Pawlak, Z.(2002a).Decision algorithms, bayes' theorem and flow *graphs, European Journal of Operational Research,* 3(6), 181-189.

Pawlak, Z.(2002b).*The rough set view on bayes' theorem*. In Pae,N.R, Sugeno,M.(Ed.), AFSS2002, LNCS, 2275, 106-120, Springer Heidelberg.

Pawlak, Z.(2003a). *Flow graphs and decision algorithms*, G.Wang et al.(Ed.). RSFDGrC 2003, LNAI 2639, 1-10, Springer-Verlag Berlin Heidelberg.

Pawlak, Z.(2003b). Probability, truth and flow graph, *Int. Workshop on Rough Set in Knowledge Discovery and Soft Computing, Electronic Notes in Theoretical Computer Science*, 82(4), 1-9.

Pawlak, Z.(2004).Decisions rules and flow networks, *European Journal of Operational Research*, 154, 184-190.

Pawlak, Z.(2005).*Flow graphs and data mining*, J.F. Peters and A.Skowron (Ed.).Transactions on Rough Sets III, LNCS, 3400, 1-36, Springer-Verlag Berlin Heidelberg.

Pawlak, Z.(2006). *Decision trees and flow graphs*, S.Greco et al.(Ed.).RSCTS 2006, LNAI 4259, 1-11, Springer-Verlag Berlin Heidelberg

Lui, H., Sun, J., Zheng, H., and Lui, L.(2009). *Extended Pawlak's flow graphs and information theory*, Gavrilova, M.L; et al.(Ed.). Transactional on Computational Science, LNCS, 5540, 220-236, Springer-Verlag Berlin Heidelberg.

Flow Graphs-a New Paradigm for Data Mining and Knowledge Discovery. (2014, March 26). Retrieved from http://www.bcpw.bg.pw.edu.pl/Content/1956/fgnpdmkd.pdf.