Classification of Financial Conditions of the Enterprises in Different Industries of Ukrainian Economy Using Bayesian Networks

Oleksandr Chernyak[1], Yevgen Chernyak[2]

[1] Department of Economic Cybernetics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, e-mail: chernyak@univ.kiev.ua
[2] Department of International Economics, Taras Shevchenko National University of Kyiv , Kyiv, Ukraine, e-mail: evgenius206@mail.ru

Abstract. In this work the analysis of branches of Ukrainian economy was done, particularly average financial parameters were found. For each parameter the boundaries were determined which divide enterprises into 5 parts and allow making more detailed ratings. The ratings were made by each parameter and then the aggregate rating was found. The analysis of indices interrelation was made using Bayesian network (BN). The coefficient of partial correlation in BN was used to analyze the interrelation of indices. This subject-matter was developed for Ministry of Industrial Policy of Ukraine. We recommend to use cascade naive Bayes model in financial planning.

Keywords: financial indices, bankruptcy, Bayesian networks, naïve Bayes, partial correlation.

1 Introduction

Each industry of economy is characterized by numerous features which distinguish one particular industry from a variety of others, for instance such features are length of operating cycle, requirement in available funds, tax policy of the state etc. The peculiarity of every industry causes the difference in major financial indices. That is why defining indices standards, their average values within the industry is an important issue, which helps to describe the place of each enterprise in the industry and also to compare industries with each other.

Setting the problem of standardizing of the financial indices estimation in frames of industries at once raises a question about the necessity of calculating the bankruptcy probability for each industry separately. To be mentioned, defining bankruptcy probability following problems are faced: a) the fact of bankruptcy is influenced not only by quantitative but also by qualitative indices like the possibility of getting preferential crediting, support of the state, uninteresting of creditors to confess a debtor to be a bankrupt; b) inadequate statistics of bankruptcies (procedure

In: M. Salampasis, A. Matopoulos (eds.): Proceedings of the International Conference on Information and Communication Technologies

for Sustainable Agri-production and Environment (HAICTA 2011), Skiathos, 8-11 September, 2011.

of bankruptcy stretches on a few years and fact of confession a bankrupt becomes separate from the beginning of problems what could have been foreseen before by the changes of financial indices); c) absence of adequate, representative base of bankruptcies, which would allow estimating probability of bankruptcy within industries.

An estimation of arising of overdue payments probability from the side of enterprise would be more precisely, as problems which are described above level with the estimation of non-fulfillment of creditor liabilities.

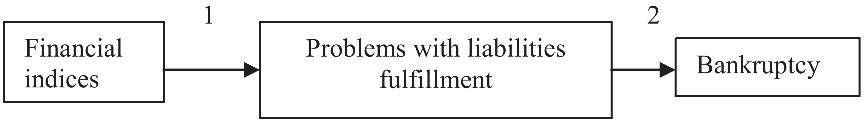

Fig. 1. Interaction financial indices and bankruptcy.

Qualitative indices influence the stage 2 in a greater measure than stage 1 (Fig. 1). A fact and a term of payment delay is accurately fixed by credit organizations. Statistics of overdue debt is collected by credit organizations, delays in payments happen more frequently, so that estimation of probability for every industry is more exact.

Therefore we stress on adequacy and possibility of estimation on the stage 1 and mention that during the transition from the first stage to the second one accuracy is being lost, and that is why estimation of the link “financial indices – bankruptcy” is considered to be purposeless.

2 Criterion of Choosing the Standards

The only assumption we will use to make the analysis is that we know the direction of index influence, in other words, increasing of the index influences positively the enterprise state or contrariwise. The standard values of index can be based on following considerations:

a) Through the influence of index on a resulting index (investigation of different variants both negative and positive: fact of overdue debt, bankruptcy debt, increase of net income). Recommended value of index would be the one which guarantees fulfillment of obligations with certain probability.

b) Finding average value within the industry, medians or division into several groups of sorted index values (more than 2) and finding average value for every group. This approach is similar to rating; some part receives the highest rate and the other lowest. Moreover, it is convenient to follow the indices moving from one group to other and afterwards to check stability of a model.

The disadvantage of the first variant is difficulty to work with the correlated indices because we have to define which of them exactly influences the result. The exclusion of the strongly correlated indices from a model will not deprive us of possibility to estimate standards for them. For example, we will have to use one of the indices of liquidity only. The disadvantage of the second variant is a risk that industry is in the phase of recession/growth and we will not get the standard values, but correspondingly decreased (increased). The best way would be to compare the results which were found by two methods and exactly to estimate in what parts the whole set is divided by probability found by first method and what probabilities we will get for the indices were found by dividing the set into equal parts.

3 Breaking on the Branch

Companies were divided into the industries according to The Classifier of Kinds of Economic Activities (CKEA). But the way of fragmentation of CKEA was different from the standard approach. We tried to pick out specific industries. For example, insurance was picked out of financial sector, pharmacy – out of chemical industry. Such method turned out to be appropriate, that was proved by the difference between indicators.

We tried to provide the fragmentation as accurately as possible to be sure that the company's activity is the same that is in the industry. For example, how production of metal should be divided from production of metal products, wholesale trade and subsidiary services? Trade and subsidiary services may differ much one from another. But at the same time it is inappropriate to combine them in one industry. Therefore, companies were divided into the next classes: extraction, production, engineering industry, wholesale trade, retail trade, rent and services.

Finally, we have got the following distribution of all the enterprises (376151) into the industries: Auto – 9 384, Building – 41 831, Building materials – 12 271, Power engineering – 4 427, Cafe and hotels – 10 400, Municipal service – 6 208, Culture and education – 10 602, Wooding – 11 656, Medicine – 5 446, Metallurgy -5 469, Real estate – 30 671, Fuel – 8 134, Polygraphy – 6 537, Cattle breeding – 6 473, Textile – 6 396, Telecommunications – 14 568, Transport – 12 684, Tourism – 4 978, Pharmacy – 5 481, Media – 3 529, Food industry – 28 058, Chemical – 7 061, Wholesale trade – 50 019, Retail – 27 121, Machinery construction – 9 685, Financial services – 15 685, Insurance – 726, Non-financial services – 16 892, Law – 3 759.

4 Dividing into Groups with the Further Purpose to Make Ratings

Now we will determine the average indices (see Table 1).

Table 1. Financial indices for the enterprises

	Name	Definition
1.	Moment liquidity ratio:	$ML = A_{HL} / L_c$
2.	Current ratio:	$CR = A_l / L_c$
3.	General liquidity ratio:	$GL = A_w / L_c$
4.	Current assets to equity ratio	$CA = (Eq - Anon\_current) / Eq$
5.	Independence coefficient:	$IC = OF / Eq$
6.	Return on assets:	$R(a) = (NP \cdot 12)/(AA \cdot N)$
7.	Return on sales:	$R(s) = (NI \cdot 12)/(NP \cdot N)$
8.	Inventory turn(days):	$IT = N \cdot 30 \cdot ICavg / NP$
9.	Debtors accounts turn(days):	$DT = N \cdot 30 \cdot ARavg / NP$
10.	Creditors accounts turn (days):	$CT = N \cdot 30 \cdot APavg / NP$
11.	Capital assets depreciation:	$D(ca) = D / OC$
12.	The proportion of capital assets and goods in process in total assets:	$CAinA = (CA + G) / A$

A_{HL} – high-liquidity assets, which consist of cash, their equivalents and current financia investmens; L_c – current liabilitis which consist of short-term credits and accounts with creditors; A_l – liquid assets which consist of high-liquidity assets, accounts receivable and billss of exchange received; A_w – working assets; E_q - equity; *Anon_current* – non-current assets; *OF*- obtained funds; *Eq*- equity; *NP* – net profit; *N* – number of monthes in period; *NI* – net income; *AA* – average value of assets is calculated as (assets at the beginning of period + assets at the end of period)/2; *ICavg* – average value of inventoryis calculated as (inventory a the beginning of a period+inventory at the end of a period)/2; *ARavg* – the average sum of the accounts receivable is calculated as (accounts receivable at the beginning of a period + accounts receivable at the end of a period)/2; *APavg* – the average sum of accounts payable is calculated as (accounts payable at the beginning of a period+accounts payable at the end of a period)/2; *OC* – original cost of capital assets; *D* – depreciation; *CA*-capital assets; *G*-goods-in-process; *A*-assets (see definitions in Van Horne and Wachowicz, (2008) or Stickney et al., 2010).

The period for *NI*, *NP*, *AA* , *IT*, *DT* , *CT* is quarter.

The differences between the branch indices showed the necessity of the work which was done. The short-term indices them selves don't allow to estimate the enterprises adequately, their place in the whole field. The values of each index were divided by quantity into 5 equal groups (see Table 2).

Table 2. Fragment division (value average of diapasons).

Food-industry	ML	CR	GL	CA	IC	R(a)	R(s)	IT
100%	5	10	15	10	50	1	1	500
80%	0,115	0,996	2,058	1,000	2,545	0,045	0,025	203,774
60%	0,022	0,508	1,154	0,967	0,557	0,002	0,005	108,812
40%	0,003	0,221	0,883	0,397	0,074	0,000	0,000	52,493
20%	0,000	0,045	0,426	-0,009	-1,155	-0,025	-0,024	18,246
0%	0,000	0,000	0,000	-10,000	-50,000	-1,000	-1,000	1,000
Insurance	ML	CR	GL	CA	IC	R(a)	R(s)	IT
100%	5	10	15	10	50	1	1	500
80%	2,627	5,479	7,569	0,981	0,281	0,188	0,300	28,630
60%	1,224	2,451	3,429	0,689	0,058	0,041	0,091	8,203
40%	0,408	1,333	1,554	0,184	0,008	0,004	0,031	4,418
20%	0,049	0,515	0,757	0,000	0,001	0,000	0,000	2,323
0%	0,000	0,000	0,000	-10,000	-50,000	-1,000	-1,000	1,000

It gives the possibility to determine the position of an enterprise by each of the parameters more precisely. In this table we can see that 20% (after filtered of information) enterprises of food industry have high value of *ML* in range [0,115; 5], also 20% enterprises of insurance industry have high value of *ML* in range [2,627; 5]. 40% enterprises of food industry have low value of *ML* in range [0; 0,003], also 20% of insurance industry have low value of *ML* in range [0; 0,049]. We recommend use this information in comparative analysis and determination position in industry.

After making the division for each enterprise by all the parameters the ratings were made (0 means error, 2-6 according to the value of parameter: the less parameter is the bigger the rating is, 1 was used for errors testing and isn't applied as a rating estimation). In this work there were considered both those coefficients which increase is positive for an enterprise (return on assets, absolute liquidity) and those, which increase is negative (depreciation, stock turn). For making the general rating it's necessary to make transformation so that the increase of the rating estimation by all the parameters will cause increase of the general rating. Let's convert the rating estimation of the parameters, which increase is positive by the following formula: $R^* = 8 - R$. This transformation leads to $2 \to 6$, $3 \to 5$, $4 \to 4$, $5 \to 3$, $6 \to 2$.

Below is given the rating of three branches enterprises (Fig. 2):

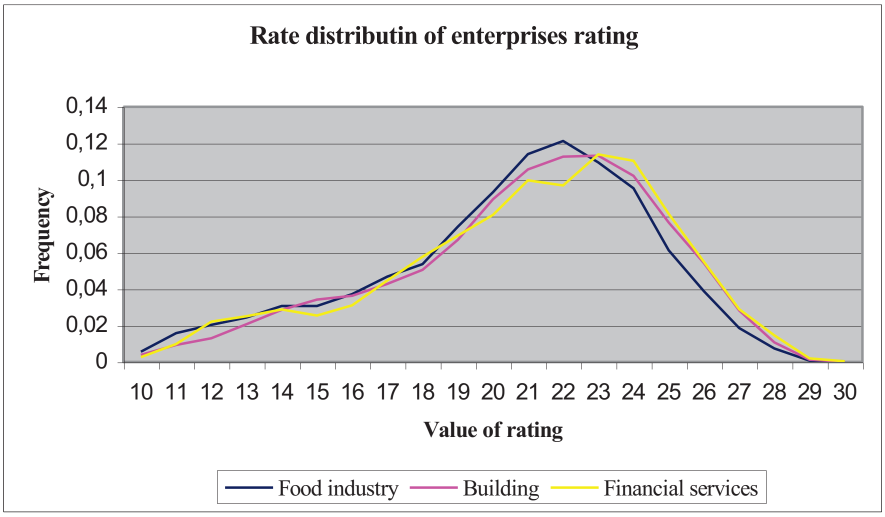

Fig. 2. Rate distribution of enterprises rating.

As a result we have a distribution close to even (it was expected because the coefficients with the least correlation values were chosen for this rating). The similarity for different branches is the evidence of the proposed method adequacy and gives the possibility to compare enterprises from different fields by means of this rating. For making rating 5 parameters were used: *GL*, *IC*, *R(a)*, *CT*, *D(ca)*. While forming the rating the following indices were transformed: *GL*, *IC*, *R(a)*; so the higher the value of *R* is, the more risk there is for solvency in the future. Visual similarity of distributions causes a question about the similar connection between the values notwithstanding the branch. The more detailed research of the parameters influence using Bayesian networks will be given further.

5 Construction of Bayesian Network

Bayesian networks are used for modelling subject domains which are characterized by uncertainty. BNs are often used for the classification problem (Friedman et al., 1997). There are the direction of using Bayesian networks in economics: bankruptcy prediction (Sun and Shenoy, 2007), early warning of bank failures (Sarkar and Sriram, 2001), credit risk modeling (Pavlenko, Chernyak, 2010), portfolio risk analysis and others.

Now we calculate the coefficients of correlation among the variables. In the Table 3 represented values of the coefficients of correlation among the variables. Colored cells represent coefficients of correlation which $|\rho| \geq 0{,}1$.

Table 3. Value of the coefficients of correlation.

	ML	CR	GL	CA	IC	R(a)	R(s)	IT	DT	CT	D(ca)	CAinA
ML	1	0,49	0,43	-0,09	0,11	0,17	0,14	-0,09	-0,14	-0,24	0,01	-0,02
CR		1,00	0,70	-0,05	0,19	0,19	0,21	-0,14	0,22	-0,23	0,02	-0,19
GL			1,00	-0,09	0,26	0,26	0,28	0,10	0,06	-0,28	0,06	-0,21
CA				1,00	-0,52	-0,08	-0,09	0,08	0,08	0,07	0,24	-0,44
IC					1,00	0,23	0,19	0,01	0,03	0,04	-0,08	-0,09
R(a)						1,00	0,89	-0,01	-0,03	-0,11	-0,04	-0,09
R(s)							1,00	0,03	0,06	-0,09	-0,06	-0,06
IT								1,00	0,16	0,25	0,03	-0,14
DT									1,00	0,29	-0,05	-0,19
CT										1,00	-0,02	-0,11
D(ca)											1,00	-0,31
CAinA												1,00

According to the Table 3 results the connection graph was built (Fig. 3). On this graph *R*-rating is the value of the 0-level. *ML*, *CR*, *GL*, *IC*, *R(a)*, *R(s)*, *CT*, *IT*, *D(ca)* are the first-level values (on the graph *ML*, *GL* are imaged not on the same level with the other values of the first-level for the better visual perception and for showing the influence of this value on the other, their interdependency).The second-level indices (*DT*, *CAinA*, *CA*) have the biggest influence on the turnover indices (*CT*, *IT*) and liquidity (*ML*, *CR*, *GL*). We chose $|\rho| = 0{,}1$ to be the level of link value.

In case if the influence of some index (eliminating the other indicators influence) on rating is inessential (absolute value of partial correlation is less then 0,1) then this index will be moved from the first- level to the second and then its influence on the first-level indices will be estimated. If some index of the second-level will influence all the first-level linked indices inessential then it will be moved into the third-level. While moving into the lower level we “break” only the links with the indices of the upper level (while moving the index into the second-level only the link with the rating is broken).The following are the values of partial correlations for indices, which are linked on the graph (Table 4):

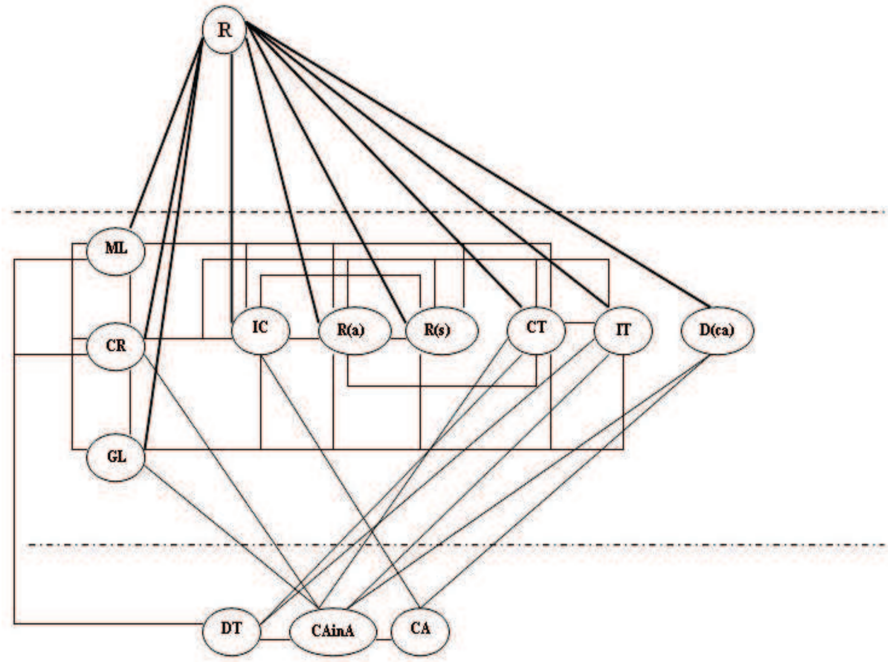

Fig. 3. Dependences among the variables ($|\rho| \geq 0{,}1$).

Table 4. Partial correlations (first-level indices).

corr(R;ML|CR)	-0,38	corr(R;CR|ML)	-0,05	corr(R;GL|ML)	-0,09	corr(R;ML|GL)	-0,39
corr(R;ML|IC)	-0,49	corr(R;IC|ML)	-0,15	corr(R;GL|R(s))	-0,21	corr(R;R(s)|GL)	-0,20
corr(R;ML|R(a))	-0,47	corr(R;R(a)|ML)	-0,36	corr(R;GL|IC)	-0,23	corr(R;IC|GL)	-0,12
corr(R;ML|R(s))	-0,48	corr(R;R(s)|ML)	-0,19	corr(R;GL|R(s))	-0,21	corr(R;R(s)|GL)	-0,20
corr(R;ML|CT)	-0,48	corr(R;CT|ML)	0,01	corr(R;GL|IT)	-0,36	corr(R;IT|GL)	0,52
corr(R;CR|GL)	-0,11	corr(R;GL|CR)	-0,12	corr(R;GL|CT)	-0,25	corr(R;CT|GL)	0,04
corr(R;CR|R(s))	-0,22	corr(R;R(s)|CR)	-0,22	corr(R;R(a)|R(s))	-0,16	corr(R;R(s)|R(a))	0,08
corr(R;CR|IC)	-0,24	corr(R;IC|CR)	-0,14	corr(R;R(a)|GL)	-0,34	corr(R;GL|R(a))	-0,33
corr(R;CR|R(a))	-0,22	corr(R;R(a)|CR)	-0,36	corr(R;R(a)|CT)	-0,39	corr(R;CT|R(a))	0,08
corr(R;CR|R(s))	-0,22	corr(R;R(s)|CR)	-0,22	corr(R;IC|R(a))	-0,10	corr(R;R(a)|IC)	-0,36
corr(R;CR|IT)	-0,23	corr(R;IT|CR)	0,45	corr(R;IC|R(s))	-0,14	corr(R;R(s)|IC)	-0,24
corr(R;CR|CT)	-0,25	corr(R;CT|CR)	0,05	corr(R;IT|CT)	0,45	corr(R;CT|IT)	0,00

According to the given calculations we come to the conclusion that the influence *CT* on *R* is inessential so this index should be moved into the second-level. Colored cells show insignificant correlations (absolute value of partial correlation is less then 0,1).

Now we only have to calculate the partial correlations between the first and second-level taking into account translation of *CT* into the second-level. Before moving *CT* we have following result (Table 5).

Table 5. Partial correlations (second-level indices).

corr(CR;DT|CT)	0,30	corr(CR;CT|DT)	-0,31
corr(CR;DT|CAinA)	0,19	corr(CR;CAinA|DT)	-0,16
corr(D(ca);CAinA|CA)	-0,22	corr(D(ca);CA|CAinA)	0,23
corr(D(ca);CA|CAinA)	0,12	corr(D(ca);CAinA|CA)	-0,22
corr(IT;DT|CAinA)	0,13	corr(IT;CAinA|DT)	-0,12
corr(CT;DT|CAinA)	0,28	corr(CT;CAinA|DT)	-0,05

Here we may conclude that the influence of *CAinA* on *CT* is inessential. After moving *CT* we have following result (Table 6).

Table 6. Partial correlations (second-level indices, after moving CT).

corr(CR;DT|CT)	0,30	corr(CR;CT|DT)	-0,31
corr(CR;CAinA|CT)	-0,23	corr(CR;CT|CAinA)	-0,27
corr(IT;DT|CT)	0,09	corr(IT;CT|DT)	0,22
corr(IT;CAinA|CT)	-0,12	corr(IT;CT|CAinA)	0,25
corr(GL;CAinA|CT)	-0,25	corr(GL;CT|CAinA)	-0,32
corr(ML;DT|CT)	-0,07	corr(ML;CT|DT)	-0,21

We come to the conclusion that the link between *IT* and *DT*, *ML* and *DT* is absent. As a result we get the following links (Fig. 4 – cascaded naïve Bayes model):

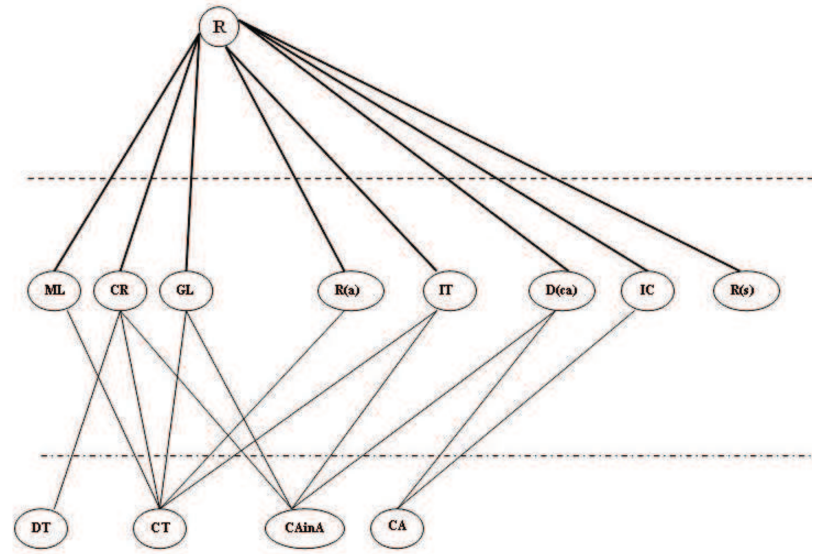

Fig.4. The structure for the cascade naïve Bayes model.

In the article (Sun and Shenoy, 2007) it was proposed to set the value level 0,1 analogically. Finding bigger threshold value of ρ, the influence of the second-level indices on first-level indices was confirmed, it didn't lead to any changes in the graph structure.

We recommend using cascade naive Bayes model while making financial planning. For example, an enterprise seeks to minimize the risk of insolvency – it should seek to decrease/increase the correspondent index (depending on the correlation sign), taking into consideration that the first-level indices are influenced by the second-level indices. Measure and character of the influence have to be compared using the following tables of conditional probabilities (Tables 7, 8):

Table 7. Conditional probabilities of insolvency depending of moment liquidity ratio.

ML		Rating				
		0	High	Medium	Low	Sum
	Error	10,61%	10,49%	4,29%	0,18%	25,58%
	High	0,00%	0,47%	7,29%	7,08%	14,84%
	Medium-High	0,00%	0,87%	8,42%	5,61%	14,91%
	Medium	0,00%	1,15%	9,83%	3,91%	14,89%
	Low-Medium	0,00%	2,59%	13,24%	2,64%	18,47%
	Low	0,00%	3,62%	7,16%	0,55%	11,32%
	Sum	10,61%	19,18%	50,24%	19,97%	100,00%

Table 8. Conditional probabilities of insolvency depending of return on assets.

ML		Rating				
		0	High	Medium	Low	Sum
	Error	10.61%	10.49%	4.29%	0.18%	25.58%
	High	0.00%	0.47%	7.29%	7.08%	14.84%
	Medium-High	0.00%	0.87%	8.42%	5.61%	14.91%
	Medium	0.00%	1.15%	9.83%	3.91%	14.89%
	Low-Medium	0.00%	2.59%	13.24%	2.64%	18.47%
	Low	0.00%	3.62%	7.16%	0.55%	11.32%
	Sum	10.61%	19.18%	50.24%	19.97%	100.00%

Tables of conditional probabilities are very useful when we have incomplete information. For example, value of *ML* – is known (high – level) and other information – absent.

$$P(R = high / ML = high) = \frac{0{,}0047}{0{,}0047 + 0{,}0729 + 0{,}0708} = 0{,}032\ , \tag{1}$$

$$P(R = medium / ML = high) = \frac{0{,}0729}{0{,}0047 + 0{,}0729 + 0{,}0708} = 0{,}49\ , \tag{2}$$

$$P(R = low / ML = high) = \frac{0{,}0708}{0{,}0047 + 0{,}0729 + 0{,}0708} = 0{,}478\ , \tag{3}$$

$$P(R(a) = high / ML = high) = 0{,}032 \cdot 0{,}1918 + 0{,}49 \cdot 0{,}5024 + 0{,}478 \cdot 0{,}1997 = 0{,}3476\ . \tag{4}$$

6 Conclusions

The main idea of this research is to demonstrate the differences between the financial indices for different industries. The analysis of indices interrelation was made using Bayesian network. The coefficient of partial correlation in BN was used to analyze the interrelation of indices. While making ratings there was made an assumption about the independence of the distribution form in which the rating frequency is described for all enterprises from branch.

The explanation of the inadequacy of the bankruptcy probability estimation is given (especially in terms of Ukrainian economy). The bigger accuracy of the solvency estimation is pointed out. The assumption is made about keeping the coefficients proportions in discriminatory models of solvency estimation notwithstanding the branch.

This subject-matter is being developed for Ministry of Industrial Policy with the purpose of temporary revelation of the enterprises subordinate to these Ministry financial problems.

References

1. Friedman, N., Geiger, D., Goldszmidt, M. (1997) Bayesian network classifiers. Machine Learning, 29, p.131-163.
2. Pavlenko, T., Chernyak, O. (2010) Credit risk modeling using bayesian networks. International Journal of Intelligent Systems, 25, N4, p.326 – 344.
3. Sarkar, S., Sriram, R.S. (2001) Bayesian models for early warning of bank failures. Management science, 47, N 11, p.1457-1475.
4. Stickney, C.P., Weil, R.L., Francis, J. (2010) Financial Accounting: An Introduction to Concepts, Methods and Uses. 13th Edition. South-Western: Cengage Learning.
5. Sun, L., Shenoy, P.P. (2007) Using Bayesian networks for bankruptcy prediction: Some methodological issues. European Journal of Operation Research, 180, p.738-753.
6. Van Horne, J.C., Wachowicz, J.M. (2008) Fundamentals of Financial Management. 12th Edition. Lebanon, Indiana, USA: FT Prentice Hall.