STUDIA INFORMATICA
Nr 1(12) Systems and information technology 2009

Supporting investment decisions using data mining methods

Witold Sysiak[1], Jędrzej Trajer[1], Monika Janaszek[1]
[1] Chair of Fundamentals of Engineering,
Warsaw Agricultural University (SGGW),
ul. Nowoursynowska 166, 02-787 Warsaw

Abstract. This paper presents an application of *k*-means clustering in preliminary data analysis which preceded the choice of input variables for the system supporting the decision about stock purchase or sale on capital markets. The model forecasting share prices issued by companies in the food-processing sector quoted at the Warsaw Stock Exchange was created in STATISTICA 7.1. It was based on neural modeling and allowed for the assessment of changes direction in securities values (increase, decrease) and generates the quantitative forecast of their future price.

Keywords. Data mining, decision support, *k*-means clustering, neural networks

1 Introduction

An important advantage of computer systems supporting decisions taken at capital markets is the elimination of subjectivity, which is often a cause of wrong decisions. Output signals, generated by programs, depend mostly on historical data. The general problem encountered constructing such models seems to be the selection of factors determining forecast quantities and formalizing the description of these interdependences. Thus, using advanced computer methods for making predictions about stock prices on the stock exchange is justified as there is a lot of information accumulated over a short period of time (indices, prices, volumes, political factors) that influence share prices. The fact that such a huge set of data is formed in a relatively short period of time results in so called 'information noise' which is difficult to analyze. Data mining methods may prove to be helpful in solving such kind of problems (Hand et al. 2005) because they are dedicated for the exploration of huge data sets.

Defining the tendency of price changes is important as it determines the decision about buying or selling stocks on the market floor at the right moment. Evaluation of securities made it possible to assess the prospect of future profits. The crucial problem for stock-exchange speculators is choosing a direction of investment. It determines future profits or losses.

Intensive development of economy and progress in automatic processing of data coming from capital markets caused increasing of mathematical models' interest. Such models may be applied in prediction of specific capital markets' behaviors. In this domain neural networks seem to be very attractive tool since they don't require any assumptions of data distribution which are very difficult to verify in case of financial analysis. Moreover they prove to be useful for modeling nonlinear and dynamic effects as well as incomplete data sets. Neural networks are also very effective in finding relations in huge datasets (Vapnik 1998).

In case of modeling of financial series one-way neural networks with one hidden layer are used mostly. The complexity of hidden layer is determined by number of hidden neurons which belongs to range $<2;2n>$, where n determines number of input variables (Azoff 1994). It is also recommended that number of network outputs should be as low as possible. Usually neural models with one output give the best results. Thus selection of output variable as well as input variables is important stage of determining neural model architecture.

According to Gatley's (1999) suggestions neural model output may represent the stock price after τ periods ($yT+\tau$), where T is a period, in which the last observation about the training sequence was made; the last change in price, the direction coefficients of the quotes trends function, indices of stock purchase and sale signals, indices showing whether current data represent market maximum or minimum, indices showing whether the date of the forecast represents global maximum or minimum, indices specifying transaction 'quality' in a given period of time or indices describing susceptibility to sudden fluctuations.

Models which generate stock purchase or sale at the output enjoy a considerable popularity due to the fact that defining the time to enter and leave the market floor is the most significant element of a successful investment. Proper definition of changes in trends has a particular meaning as regards short-time stock exchange speculations. The main tool for their assessment is technical analysis, which is based on finding similarities between the graphs for the previous periods and current ones. The same task may be carried out by neural networks.

The main aim of this paper was to create a computer system forecasting trends decisive in stocks prices changes and their value assessment. Thanks to this system a market participant would be able to gain information concerning changes in stocks prices in a short time, which would result in increased efficiency of investments on the capital market.

2 Material and methods

Data came from stocks' quotations of twenty one companies in the food-processing sector on the Warsaw Stock Exchange from 1998-01-01 to 2006-12-31. Each case in the data set corresponded to a three-month period of the company's operation and was described by attributes assigned to one of three categories: indices, quotations and volumes, self-assessment.

Quarter reports including balance sheet, profit and loss account (income statement), cash flow statement, statement of changes in equity, off-balance sheet

liabilities and additional information were used for the purpose of index analysis. According to Bednarski's (2007), Brigham's and Houston's (2005), Info Consulting's (2007) and New Trader's (2007) suggestions, symbols 'v0' and 'v1' were assigned to negative and positive values of indices respectively. Indices were organized into five groups: fluency indices, indebtedness indices, activity indices, profitability indices and market value indices.

Self-assessment was conducted for every group of indices in three variants. Whole group got positive note if 100% (the first variant), 75% (the second variant) or 50% (the third variant) of indices within had positive value 'v1'.

Each case in the data set was described by current share price (kurs_m0) and volume (wol0) representing share price and volume at the end of the stock exchange trading hour on the last working day of the quarter's first month, share prices and volumes from one, two, three and four months before (kurs_m-1, wol-1, kurs_m-2, wol-2, kurs_m-3, wol-3, kurs_m-4, wol-4, respectively), share price from seven months before (kurs_m-7). On the basis of listed variables few additional were calculated which represented proportional changes of share prices between months, quarters and half-year and proportional changes of volumes between months.

The analysis of collected data proceeded in few steps. Figure 1 presents the successive stages of data analysis methodology.

2.1 Preliminary data processing

K-means clustering method implemented in STATISTICA 7.1 (StatSoft Inc. 2005) was used which resulted in partition of the whole data set into smaller subsets. In the first stage, variable %kurs_m0, representing proportional changes between current share price and share price from one month before, was partitioned into two classes corresponding with the declining or rising trend in the price. Then, cluster analysis was conducted using *k*-means algorithm and cross-validation, separately for cases characterized by decreasing and increasing tendency. Cluster centers were matched with the observations which gave the maximum distance. An Euclidean metric was used as a distance measure.

Cross-validation was performed in order to select the optimal number of clusters. The idea of cross-validation consists on the partition of the whole sample into v subsets or random disjoint sub-samples. The same analysis is then used for observations out of v-1 sets (training samples), and the obtained model is used for subset v (sample which was not used for determining the clusters – a testing sample). Calculation of prediction accuracy is based on inspection how well the observations out of sample v are matched with homogenous clusters using the current solution calculated out of v-1 training samples. Results for successive v repetitions are aggregated (averaged) and give assessment of the model stability, which means the accuracy of matching observations with clusters (Stanisz 2006). The parameters of cross-validation were as follows: the number of iterations – 100, the maximum number of clusters – 10, convergence criterion – 0.05.

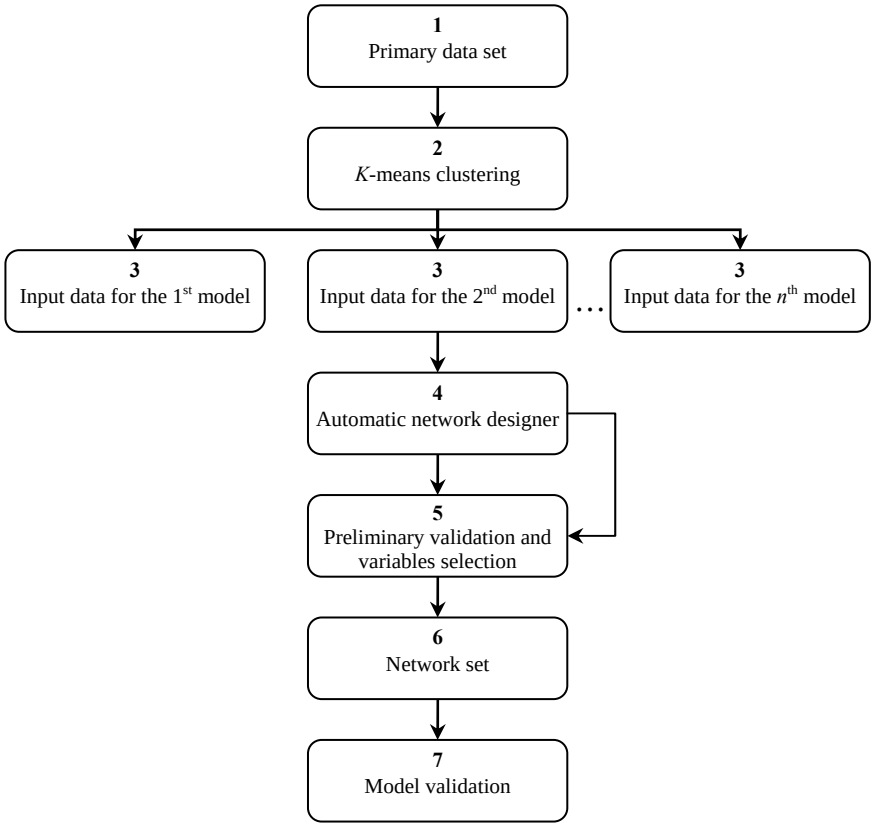

Figure 1. Data analysis methodology

2.2 Decisions support system

Input data set contained cases selected on the basis of current share price which belonged to range (0;6.6>. Neural model output represented share price (kurs_m1) that was predicted at the end of the last working day of the quarter's second month. The first stage of the experiment was to create and test models with all examined variables. Input data set was randomly divided into a training sample (80%), validating sample (10%) and testing sample (10%). An automatic network designer, implemented in STATISTICA 7.1 tested three-layer neural networks and selected models as regards maintaining balance between the error rate and network diversity. The number of networks to be retained was set to 100. An automatic designer tested models which had the number of hidden neurons in range <2;57>. After training the best thirty models with standard deviation ratio lower than 0.4, correlation higher than 0.91 and mean error lower than 0.5 were selected. Sensitivity analysis was then conducted. Variables were evaluated with regard to usefulness in forecasting variable kurs_m1.

The next step was building models using the inputs selected on the basis of previous networks analysis that have the same output. Actions presented at the above scheme were repeated a few times until the model met the assumptions.

The last stage involved building a set of networks using the group of selected models. Either all networks or selected networks may be included in the set. This depends on the final result – an error that will bias the forecast. The system of forecasts in the form of a set of networks was verified using the testing set, defining forecasts errors.

3 Research results

3.1 Preliminary data processing

Cluster analysis conducted for declining trend in share price pointed that optimal number of clusters was 8. Figure 2 presents error rates (costs) depending on number of clusters. The cross-validation algorithm reached the convergence criterion if the number of clusters was equal or more than four. Having analyzed the error function an expected number of clusters in actual analysis was set to five.

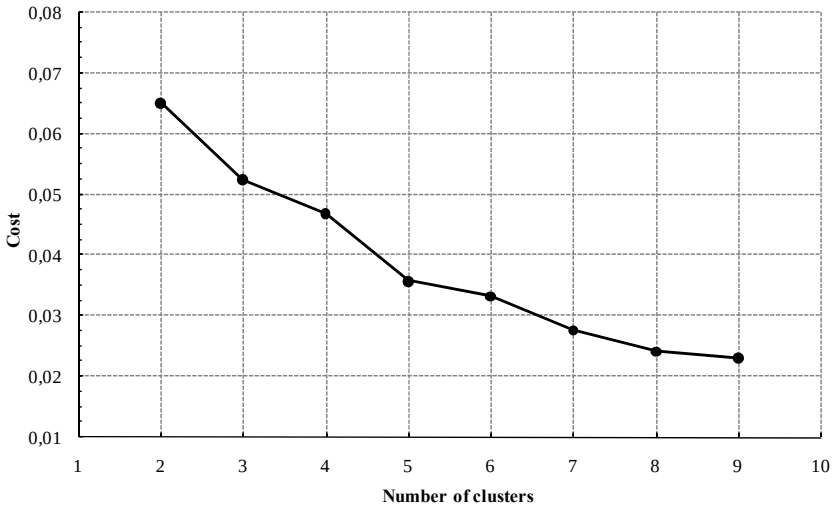

Figure 2. Error rates depending on number of clusters for declining part of course

Table 1 shows that there was only one case included in the first cluster whereas the second cluster contained only two cases. First three clusters were then merged so finally three clusters were made and ranges of the variable %kurs_m0 were (Tab. 2): <-61;-14>, <-13;-6>, <-5,0>.

Table 1. Number of cases in clusters – decreasing tendency

Cluster	1	2	3	4	5
Number of cases	1	2	26	54	93
Percentage	0.57	1.14	14.77	30.68	52.84

Cluster analysis conducted for increasing tendency in share price pointed that optimal number of clusters was 6. The cross-validation algorithm reached the convergence criterion if the number of clusters was equal or more than four (Fig. 3).

Having analyzed the error function an expected number of clusters in actual analysis was set to five.

Number of cases in particular clusters is presented in table 3. The fifth cluster contained only ten cases so it was joined with the fourth cluster. Finally four clusters were made and ranges of the variable %kurs_m0 were (Tab. 4): <1;6>, <7;16>, <17, 37>, <38;60>.

Table 2. Clusters characteristics – decreasing tendency

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	General
Minimum	- 61.00	- 37.00	- 26.00	- 13.00	- 5.00	- 61.00
Maximum	- 61.00	- 33.00	- 14.00	- 6.00	0.00	0.00
Mean	- 61.00	- 35.00	- 18.00	- 8.74	- 1.97	- 7.12
Std	0.00	2.83	3.68	3.26	1.67	62.97

Having analyzed the data set the following observations were made:

The data set for prediction included a wide range of values. For example, variable kurs_m0 had values from 0.4 PLN to 509 PLN, with the mean of 56 PLN and the median of 16.6 PLN. It may be inferred that for majority of cases value of this variable was found in range <0.4;30> (Fig. 4).

The same change in the share price may have different influence on shares with value of 100 PLN, 10 PLN or 2 PLN. If the share has a value of 2 PLN the decrease of 1 PLN reduces its value by 50 per cent. In case of share with value of 10 PLN, decrease of 1 PLN reduces its value by 10 per cent, whereas share that cost 100 PLN gains only 1 per cent in value. If the investor bought 10 000 stocks with price 2 PLN each, and the price would decrease by 1 PLN, then the investor would lose 10 000 PLN. However, if the investor bought 200 stocks with price 20 000 PLN, and the price would decrease by 1 PLN, the loss would by only 200 PLN. Therefore, the forecast error equal 1 PLN has a different significance for stocks with different prices.

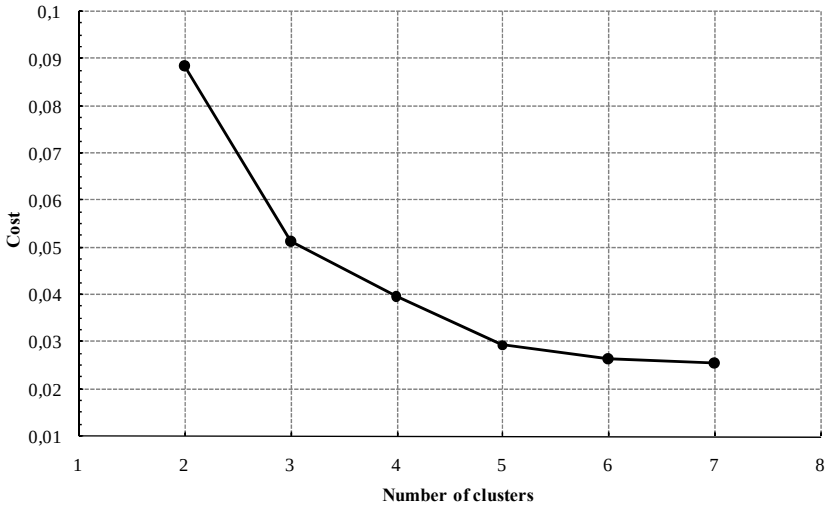

Figure 3. Error rates depending on number of clusters for rising part of course

Table 3. Number of cases in clusters – increasing tendency

Cluster	1	2	3	4	5
Number of cases	67	64	50	24	10
Percentage	31.16	29.77	23.25	11.16	4.65

Table 4. Clusters characteristics – increasing tendency

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	General
Minimum	1.00	6.00	12.00	21.00	52.00	1.00
Maximum	5.00	11.00	20.00	37.00	71.00	71.00
Mean	2.85	7.92	15.46	26.46	60.10	12.59
Std	1.32	1.50	2.51	5.12	5.97	172.14

Additional research was conducted as regards the change of the error rate depending on the data range. For this experiment three-layer neural networks were created for different data ranges ((0;501>, (0;35>, (0;12>, (0;7>).

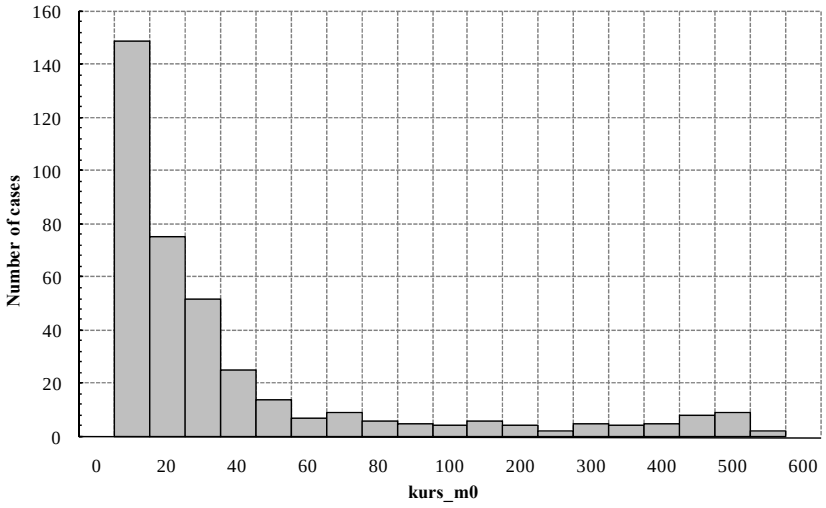

Figure 4. Bar chart for variable kurs_m0

The variables at the input (kurs_m0, kurs_m-1, kurs_m-2, kurs_m-3, kurs_m-4 and kurs_m-7) and at the output (kurs_m1) of all neural models were similar. Data set was randomly divided into a training set (80%), validating set (10%) and testing set (10%). The complexity of the hidden layer was defined from 2 to 13 (2n+1) neurons. An automatic network designer, implemented in STATISTICA 7.1, tested all available MLP learning algorithms and selected models as regards maintaining balance between the error rate and network diversity. The number of models to be retained was specified as 100. The next stage of this experiment consisted in selection models with standard deviation ratio near 0 and correlation near 1. On the basis of error rates generated by selected models mean absolute errors apart for learning, validating and testing set were computed (Fig. 5, Fig. 6).

Error rates for wider sets were much larger. In case of the first three specified ranges an error rate exceeded or approached 1 (values of error function for range (0;501> was not shown). Predicting share prices for the whole set of data, where the data range was between 0.4 and 501 would not give the desirable result. The error rate might exceed the initial stock value.

Further, the whole set was divided to subsets according to the variable kurs_m0 so that each subset was not smaller than 100 cases. For data set partition k-means clustering algorithm with cross-validation was used. The parameters of cross-validation were the same as those used for variable %kurs_m0.

The optimal number of clusters pointed by algorithm was 9. Figure 7 shows that error function begins flatten if the number of clusters is equal or more than 7. Therefore an expected number of clusters in actual analysis were set to seven.

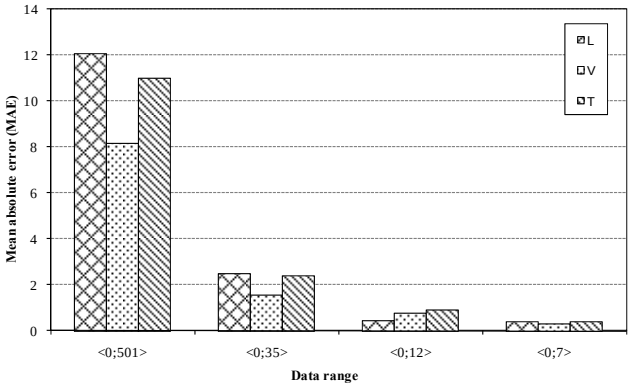

Figure 5. Error rate differences depending on range of variable kurs_m0 values

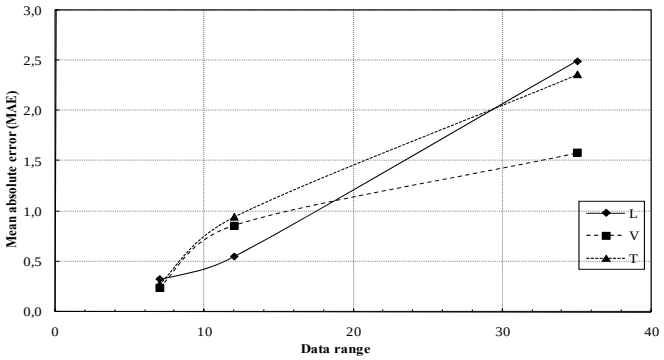

Figure 6. Error function depending on the data range

Only the first and the seventh cluster contained over 100 cases (Tab. 5) but data ranges in calculated clusters were wider than expected (Tab. 6).

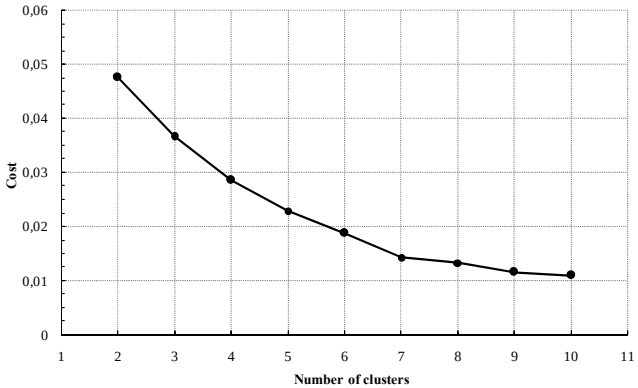

Figure 7. Error rates depending on the number of clusters

Table 5. Number of cases in clusters

Cluster	1	2	3	4	5	6	7
Number of cases	107	22	51	20	7	13	171
Percentage	27.36	5.63	13.04	5.11	1.79	3.32	43.73

Table 6. Clusters characteristics

Cluster	1	2	3	4	5	6	7
Minimum	12.65	65.00	31.90	400.00	138.00	262.50	0.52
Maximum	30.60	121.00	61.30	508.50	220.00	368.00	12.45
Mean	20.45	87.55	42.43	458.17	186.00	316.38	4.80
Std	4.83	15.68	9.09	32.49	28.12	36.28	3.45

Cluster analysis was then executed again with expected number of clusters set to nine. Table 7 shows that only the seventh cluster contained more than 100 cases and range of variable kurs_m0 values (Tab. 8) assured mean absolute error value (MAE) lower than 0.5 (see Fig. 6). Cases included in this cluster were used for further analysis and neural model creation.

Table 7. Number of cases in clusters

Cluster	1	2	3	4	5	6	7	8	9
Number of cases	72	19	50	13	16	6	125	70	20
Percentage	18.4	4.9	12.8	3.3	4.1	1.5	32.0	17.9	5.1

Table 8. Clusters characteristics

Cluster	1	2	3	4	5	6	7	8	9
Minimum	16.3	50.9	28.9	262.5	81.0	173.0	0.4	7.3	400.0
Maximum	27.8	78.0	48.0	368.0	138.0	220.0	6.6	16.1	508.5
Mean	21.0	61.8	36.3	316.4	97.9	194.0	2.9	11.5	458.2
Std	3.1	8.3	5.7	36.3	16.3	20.3	1.5	2.7	32.5

3.2 Decisions support system

The set of networks was composed of 21 models. Regression statistics for this set were as follows: the correlation coefficient 0.997, the standard deviation ratio 0.216 and mean absolute error 0.216. Testing sample, which did not take part in the training process, was used for the purpose of verification. The obtained predicted values were presented in table 9 and 10 with the actual stocks prices.

The second, fourth and seventh cell in table 9 as well as the third cell in table 10 presents incorrect forecasts of the future trend. Signals with correctly defined direction of changes in the stocks value are significantly predominant there. Attention should also be paid to the fact that the price in two out of four incorrect forecasts changes only slightly about a few hundredths of the value (see the second and the seventh cell in Tab. 9).

During elimination of variables, it turned out that the group of features connected with the stock price and the volume has a greater influence on the forecast result. These were past stocks and volumes prices, changes between respective dates of measurements expressed as percentage and qualitatively described jumps in prices and volumes. The error increase ratio was used as a significance measure of variables included in the neural model. It was one of the factors that allowed for the elimination of variables of little usefulness. During the first phase of the variables elimination its value was considerably lower for the variables describing the economic situation of a given company. Having analysed all the stages of variables elimination process, it was found out that variables describing the situation of a company on the stock exchange were enough to create a model for forecasting prices on the capital market. The explanation of this solution may be found in the rules of technical analysis ('history repeats itself').

Table 9. Verification of declining trend changes

	Beef-San 2000.Q2	**Beef-San 2001.Q4**	**Mieszko 2005.Q3**	**Mieszko 2006.Q4**	**Rolimpex 2003.Q1**	**Advadis 2006.Q1**	**Advadis 2006.Q2**
Current price	2.40	1.04	2.61	3.51	5.65	1.26	1.33
Actual price	2.20	1.00	2.82	3.24	6.45	1.33	1.31
Forecast	2.34	1.19	2.72	3.84	6.52	1.38	1.56

Table 10. Verification of rising trend changes

	Wilbo 2000.Q4	**Wilbo 2003.Q4**	**Beef-San 2001.Q3**	**Beef-San 2006.Q4**	**Mieszko 2004.Q2**	**Advadis 2000.Q4**	**Advadis 2006.Q4**
Current price	2.52	3.39	0.98	3.15	3.95	2.56	1.72
Actual price	2.35	3.90	1.19	2.94	3.11	2.51	1.78
Forecast	2.40	3.48	0.56	2.61	3.65	2.51	1.73

The forecasting process is based on past data concerning shares prices. It is also known that neural networks are suitable for seeking after hidden dependences in past data. It is then sensible that only variables representing price and volume were selected during variables elimination process.

4 Conclusions

K-means clustering enabled partition of the data set into smaller subsets that will be used to build forecast models. Neural modeling results point that large variance of input variables may negatively affect the prediction accuracy. Cases partition according to the variable representing current share price should favorably influence the quality of forecast models.

Using data mining for the analysis of the stocks prices issued by companies in the food-processing sector confirms large potential of these methods in supporting decisions about purchase or sale stocks on the capital market. Application of neural networks set as a tool for forecasting shares prices facilitated the forecasts quality in comparison to single models, which forecast effectively only for a certain range of data. The computer system supporting decisions about purchase or sale of stocks issued by the food-processing sector (in the form of a set of neural networks) on the Warsaw Stock Exchange was favourably verified which is confirmed both by good results of prediction of the changes in trend direction and low values of errors of the forecasted shares prices.

References

[1] Azoff E.,M., (1994). *Neural network time series forecasting of financial markets*, Chichester, John Wiley & Sons.
[2] Bednarski L., (2006). *Financial analysis of enterprise* (in polish), Warsaw, Polish Economic Publishers PWE.
[3] Brigham E. F., Houston J. F. 2005. *Fundamentals of financial management* (in polish), Warsaw, Polish Economic Publishers PWE.
[4] Gately E., (1999). *Neural networks for financial forecasting and transactional systems designing* (in polish), Warsaw, Financial Publishers WIG-PRESS.
[5] Hand D., Mannila H., Smyth P., (2005). *Principles of data mining* (in polish), Warsaw, Scientific-Technical Publishers WNT.
[6] Info Consulting, (2007). www.info-consulting.pl.
[7] New Trader, (2007). www.newtrader.pl.
[8] Stanisz A., (2006). *Statistical methods applied in medicine* (in polish), Vol. 3, Krakow, StatSoft Polska.
[9] StatSoft Inc., (2005). STATISTICA (Data Analysis Software System), version 7.1, Tulsa, OK., StatSoft.
[10] Vapnik V., (1998). *Statistical learning theory*, New York, John Wiley and Sons.