Research on Application of Data Mining Methods to Diagnosing Gastric Cancer

Arnis Kirshners[1], Serge Parshutin[1], and Marcis Leja[2]

[1] Riga Technical University, Institute of Information Technology, 1 Kalku str., LV-1658, Riga, Latvia
[2] University of Latvia, Faculty of Medicine; Riga Eastern Clinical University Hospital, Digestive Diseases Centre GASTRO, 6 Linezera str., LV-1006, Riga, Latvia
{arnis.kirsners,serge.parshutin}@rtu.lv,
cei@latnet.lv

Abstract. Constantly evolving technologies bring new possibilities for supporting decision making in different areas - finance, marketing, production, social area, healthcare and others. Decision support systems are widely used in medicine in developed countries and show positive results. This research reveals several possibilities of application of data mining methods to diagnosing gastric cancer, which is the fourth leading cancer type in incidence after the breast, lung and colorectal cancers. A simple decision support system model was introduced and tested using gastric cancer inquiry form statistical data. The obtained results reveal both the benefits and potential of application of DSS aimed to support a medical expert decision, and some shortcomings mainly connected with performing an appropriate data preprocessing before mining knowledge and building the model. The paper presents the technologies behind the DSS and shows the detailed evaluation process with discussions.

Keywords: gastric cancer, decision support system, data mining.

1 Introduction

Cancer is the worldwide problem in social health and one of the leading causes of death. Nevertheless it is known that the most of a cancer types are treatable. Referencing the World Health Organization data, at least 40% of all local cancer types are treatable and can be prevented, avoiding the risk factors, common not only for cancer, but also for the most chronic diseases. These risk factors are known and the most important of them are smoking, alcohol and other pernicious habits, activity shortage, adiposis (excessive weight) and different infectious agents. New medical technologies, new medicaments, vaccines, screening systems are continuously developed and introduced, all aimed at identification and treatment of cancer at initial stages, at improvement of life quality and life length for patients with cancer.

Most of the patients recourse to the experts having symptoms of the last stages of a disease, which significantly limits the list of possible treatments, thus having a negative impact on life length of a patient. People are too timid to

P. Perner (Ed.): ICDM 2012, LNAI 7377, pp. 24–37, 2012.

discuss their problems and recourse to experts having the disease symptoms with pain, fluxes, etc. In order to reveal a possible morbidity, a set of actions should be taken, which would contribute to early diagnosis of disease.

Even though globally the gastric cancer incidence is declining and in many Western countries the disease is not considered among the major health issues any more, globally the cancer of the stomach is still continuing to be an important healthcare problem. Gastric cancer is remaining the second leading cause of mortality worldwide within the group of malignant diseases after the lung cancer, and is accounting for almost 10% of cancer related deaths. Among men gastric cancer is the second (after lung cancer), but among women - the third leading (after breast and lung) cause of cancer-related deaths [12].

Today gastric cancer is the fourth leading cancer type in incidence (after the breast, lung and colorectal cancers). Close to a million new gastric cancer cases are diagnosed annually (989600 cases as reported by International Agency for the Research on Cancer (IARC) in 2008) [6]. The overall prognosis of the disease is remaining poor. The survival is closely related to the extent of the disease. If the disease is diagnosed at advanced stage, the survival is in general low. If an early cancer is diagnosed confined to the inner lining of the stomach wall, 95% 5-year survival could be reached [3].

Gastric cancer is well diagnosed using the upper endoscopy, however this is not a cheap type of analysis, thus there is a need for a decision support system for an earlier diagnosis, which would supply an expert with additional information for choosing whether the endoscopy should be performed in a specific case. The present work is a pilot research and discusses a possibility of using data mining methods for separating patients, who do need en endoscopy to be made from those, whom endoscopy is not obligate. Section 2 presents a model of such decision support system, showing its structure and describing inner processes. The experimental results are described and analysed in Section 3, followed by conclusions.

2 Model of the Decision Support System

The main objective of the proposed decision support system is to support a medial expert with additional information, helping him/her to make a decision whether a patient needs an endoscopy. It should be noted that a sphere of possible applications of such decision support systems is not limited to only diagnosing a gastric cancer. In most of developed countries decision support systems are widely used in medicine and other areas.

The decision support system contains two main modules - Data Mining module and Decision Support module (see Figure 1). The data preprocessing block is placed outside the DSS. The data preparation is an obligate process, but not necessary as part of a decision support system - the data preprocessing can be made outside DSS with any other tool available, however this does not decline the inclusion of data preprocessing module as part of DSS. Speaking about the medical data preprocessing, it should be noted that in the most cases classes in the dataset will be highly imbalanced, causing a high increase in a false negative rate. One of the solutions to this problem is data sampling - creating a subset(s) of data, where classes are more balanced, as compared to the initial dataset. Different sampling models exist - static, dynamic, active sampling [1], proposing different ways to choose examples. Another option for taking on the imbalanced classes is the distributed data mining, aggregating several models in order to gain a more precise result [1]. Besides sampling, the data preprocessing should include feature selection and transformation, as in most cases exclusion of less informative attributes increases an efficiency of the system [5, 9].

Data mining module contains tools for mining relationships in data and building the knowledge base for further application; it receives preprocessed statistical data and builds a relationship model, which is then saved in the knowledge base. In our specific case, the classification methods were chosen among other knowledge mining techniques. Classification model may contain a single classifier or a

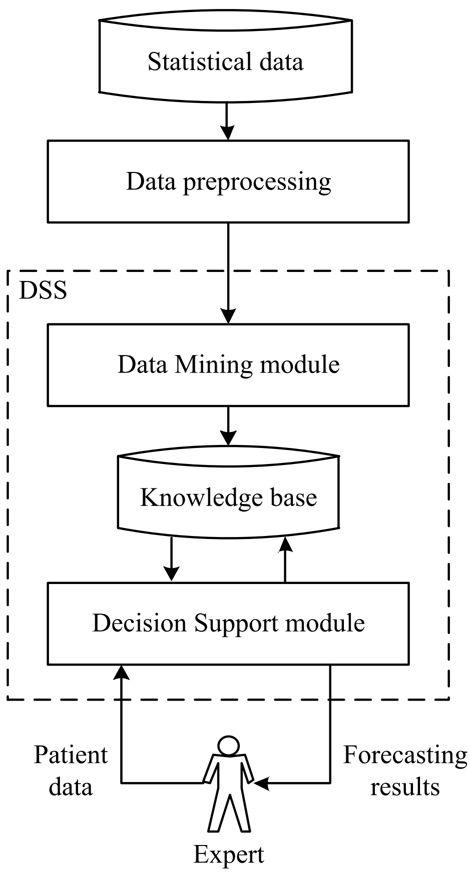

Fig. 1. Model of the decision support system

set of classifiers acting as a single one. The crisp or fuzzy classification may be applied, depending on a defined objective that should be gained.

Decision support module is the one that interacts with the user. Figure 1 shows an interaction process - an expert sends the patient data obtained from an inquiry form or via a direct contact with a patient, and receives an estimation of possible outcomes for a specific case, and then makes a final decision. The estimation of the outcomes is obtained using the knowledge base containing not only the rule sets and probability estimations (Naive Bayes), but also the efficiency coefficients for each classifier used, including classification accuracy, sensitivity (true positive rate), specificity (true negative rate) and a false-negatives rate. This coefficients can be used to show an expert the confidence of the forecast made by the decision support system.

The model of the decision system is simple and client oriented - it does not need a medical expert to have an advanced knowledge in statistics or data analysis, making it simple in application.

3 Experimental Results

In the previous section it was mentioned that in the current work the DSS is using the classification methods to mine knowledge from data. Three classification algorithms were chosen for evaluation - the Bayesian classification algorithm (Naive Bayes) [4, 5, 11], the decision tree classifier C4.5 [4, 5, 10, 11] and the classification rule induction algorithm CN2 [2, 7, 8]. All of the classifiers are known and the supplied references have a description of algorithms behind each classifier. Those three classifiers were chosen for the pilot research oriented to define whether or not simple classification algorithms can be used to process with a small dataset with class dominance. Other methods like SVMs or Nearest Neighbour classifiers were not used as the dataset contained mostly discrete attributes and was relatively small for application of SVMs.The experiments were performed using the medical data of patients who filled the gastric cancer inquiry form. The dataset contained 819 examples where 24 examples (3%) with positive diagnosis and 795 negative examples (97%) where described by 31 attribute - ID, target attribute and 29 descriptive attributes.

The diagnosis for each patient was assigned using the endoscopy, and it can be seen that in most cases the endoscopy was not necessary, as the final result was negative. The main objective of the proposed system is to lessen the false positive rate, simultaneously maintaining high sensitivity. Returning to the initial dataset the classes are highly imbalanced, which may lead to incorrectly interpreted results. All three classifiers were trained using all 819 examples with full feature set and tested using the 10-fold cross-validation; the results of experiments are given in Table 1.

All classifiers show a classification accuracy greater than 95%, but the sensitivity (true positive rate) of a target class - positive diagnosis, remains 0 or close to it, pointing out that classifiers were not able to correctly classify examples with a positive diagnosis. Such results are common for datasets with highly imbalanced classes, as classifiers perceive class significance equally weighted, thus

Table 1. Classification results with full dataset

Algorithm	CA	Sensitivity	Specificity
Naive Bayes	0.968	0.000	0.997
C4.5	0.957	0.042	0.985
CN2 Rules	0.951	0.000	0.980

in most classes a single rule is made - classify each record as one with dominatig class. Different options are available for improving the classification efficiency - application of cost matrix in training stage, using negative selection (anomaly detection) instead of common classification, data synthesis, feature selection, data sampling and others. In this research the feature selection and the data sampling options were applied. First, the correlation analysis [5, 9, 11] of 29 available descriptive attributes was made and 10 attributes were selected. Table 2 summarizes the results of an attribute correlation analysis.

Table 2. Attribute correlation analysis

No.	Attribute	F-value	p-value
1	Weight loss (T/F)	3.7813	0.0521
2	Age (years)	3.1399	0.0009
3	Weight loss in last 6 months (kg)	2.7813	0.0168
4	Vomiting (T/F)	1.7794	0.1825
5	Relatives have other tumours (T/F)	1.7006	0.1825
6	Constipations (T/F)	1.5049	0.2202
7	Heartburn with proximal spreading (T/F)	1.4589	0.2274
8	First-degree relatives have gastric cancer (T/F)	1.3195	0.2510
9	Cigarettes per day	1.2667	0.2762
10	Flatulence (T/F)	1.2174	0.2701

The list of selected attributes was reviewed by our medical expert and it was stated that at least two attributes - "Weight loss" and "Weight loss in last 6 months", can be removed from the list. It has been pointed out that if a patient has an unplanned weight loss, the additional laboratory analysis (endoscopy) will always be performed. Thus the two mentioned attributes were removed from our list, leaving eight attributes for further analysis. It was decided to perform data sampling in two different proportions, shown in Table 3. Five different datasets were randomly generated using each proportion, no duplication was applied.

All generated datasets contain all positive examples, available in an initial dataset, and randomly selected negative examples, the number of which is set using the defined proportion. The number of examples is relatively small and can

Table 3. Description of generated subsets

No.	Nr. of datasets	Proportion of classes	Positive ex.	Negative ex.	Total ex.
1	5 sets	1 x 2	24	48	72
2	5 sets	1 x 4	24	96	120

decrease the confidence of results in case if the cross-validation is applied, thus each of ten generated datasets was randomly split into training and testing sets using the 70% for training and 30% for testing [5,11]. The proportions of classes in train and test sets remained the same as in the subset before splitting (see Table 3). The experimental results are presented in the next two subsections - Subsection 3.1 summarizes the evaluation results, obtained using the No.1 datasets; the results using the No.2 datasets are given in Subsection 3.2.

3.1 Experimental Results Using No.1 Datasets

Each of the three classifiers was trained and tested using each of prepared datasets. In order to confirm that feature selection can increase an efficiency of the system, some experiments were performed with three different feature sets:

- First feature set with all 29 descriptive attributes;
- Second feature set with eight attributes from Table 2 excluding attributes 1 and 3;
- Third feature with six attributes, obtained by excluding from Second feature set attributes "Cigarettes per day" and "Flatulence" (see Table 2).

Table 4 shows the result obtained using the First feature set and training and testing each classifier with all five subsets. It can be seen that the average sensitivity of classifiers increased, as compared to the data in Table 1, but still remains less than 50%. The increase in sensitivity shows that changing the proportions of classes the classifier s were forced to create relationship model, containing both classes, however the results are highly dependent on the subset and the variation in sensitivity confirms it.

Table 5 shows the classification accuracy, sensitivity and specificity of classifiers, obtained using the Second feature set. The average sensitivity increased and the false-negatives rate decreased, comparing to the results in Table 4. The average specificity of each classifier remains high, as also the average sensitivity is greater than 50%, but still highly varies from set to set.

Table 6 provides evaluation results using the Third feature set with six attributes. The average results decreased, comparing to data in Table 5, showing that attributes 9 and 10 - the number of cigarettes smoked per day and the flatulence, should not be excluded from feature set.

Figure 2 depicts the average values of classification accuracy, sensitivity and specificity for all classifiers separately for each feature set. It can be seen that all

Table 4. Evaluation results using the First feature set

CA	Set 1	Set 2	Set 3	Set 4	Set 5	Average	St.dev.
Naive Bayes	0.64	0.73	0.73	0.68	0.50	0.655	0.084
C4.5	0.64	0.73	0.59	0.55	0.64	0.627	0.060
CN2 Rules	0.59	0.82	0.59	0.59	0.77	0.673	0.101
Sensitivity							
Naive Bayes	0.43	0.71	0.29	0.43	0.27	0.429	0.156
C4.5	0.71	0.71	0.14	0.27	0.57	0.486	0.232
CN2 Rules	0.27	0.86	0.00	0.43	0.43	0.400	0.277
Specificity							
Naive Bayes	0.73	0.73	0.93	0.80	0.60	0.760	0.108
C4.5	0.60	0.73	0.80	0.67	0.67	0.639	0.068
CN2 Rules	0.73	0.80	0.87	0.67	0.93	0.800	0.094
False-Negatives rate							
Naive Bayes	0.27	0.15	0.26	0.25	0.36	0.258	0.065
C4.5	0.18	0.15	0.33	0.33	0.23	0.247	0.075
CN2 Rules	0.31	0.07	0.35	0.29	0.22	0.249	0.096

Table 5. Evaluation results using the Second feature set

CA	Set 1	Set 2	Set 3	Set 4	Set 5	Average	St.dev.
Naive Bayes	0.64	0.86	0.68	0.68	0.64	0.700	0.084
C4.5	0.73	0.68	0.55	0.55	0.64	0.627	0.073
CN2 Rules	0.77	0.86	0.59	0.59	0.77	0.718	0.109
Sensitivity							
Naive Bayes	0.57	1.00	0.43	0.43	0.57	0.600	0.210
C4.5	0.71	0.86	0.29	0.29	0.57	0.543	0.229
CN2 Rules	0.43	0.86	0.43	0.43	0.43	0.514	0.171
Specificity							
Naive Bayes	0.67	0.80	0.80	0.80	0.67	0.747	0.065
C4.5	0.73	0.60	0.67	0.67	0.67	0.667	0.042
CN2 Rules	0.93	0.87	0.67	0.67	0.93	0.813	0.122
False-Negatives rate							
Naive Bayes	0.23	0.00	0.25	0.25	0.23	0.192	0.097
C4.5	0.15	0.10	0.33	0.33	0.23	0.23	0.094
CN2 Rules	0.22	0.07	0.29	0.29	0.22	0.217	0.078

Table 6. Evaluation results using the Third feature set

CA	Set 1	Set 2	Set 3	Set 4	Set 5	Average	St.dev.
Naive Bayes	0.68	0.73	0.73	0.50	0.55	0.636	0.095
C4.5	0.73	0.73	0.82	0.68	0.73	0.736	0.045
CN2 Rules	0.68	0.82	0.73	0.68	0.77	0.736	0.053
Sensitivity							
Naive Bayes	0.71	0.86	0.029	0.00	0.57	0.486	0.308
C4.5	0.43	0.86	0.43	0.00	0.43	0.429	0.270
CN2 Rules	0.14	0.71	0.14	0.00	0.43	0.268	0.256
Specificity							
Naive Bayes	0.67	0.67	0.93	0.73	0.53	0.707	0.131
C4.5	0.87	0.67	1.00	1.00	0.87	0.880	0.122
CN2 Rules	0.93	0.87	1.00	1.00	0.93	0.947	0.050
False-Negatives rate							
Naive Bayes	0.17	0.09	0.26	0.39	0.27	0.236	0.101
C4.5	0.24	0.09	0.21	0.32	0.24	0.218	0.073
CN2 Rules	0.30	0.13	0.29	0.32	0.22	0.252	0.068

classifiers have shown an increase in average sensitivity using the Second feature set with eight attributes and the value remains greater than 50%, however the average false negative rate remains above the 20% level.

The results of experiments with No.1 datasets (see Table 3) showed that sampling and feature selection can increase an efficiency of classifier, however the results show a high variance in estimations, especially in sensitivity. The main reason of that is the small number of examples in each subset, comparing to the initial dataset. Nevertheless individual results with sensitivity higher than 70% were reached.

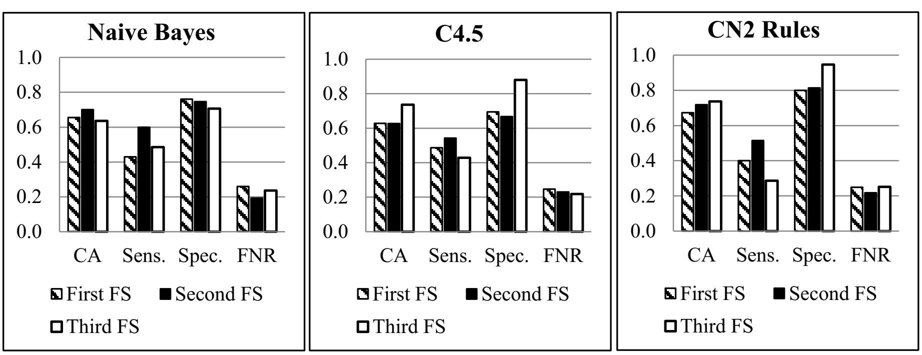

Fig. 2. Average values for classifiers in experiments with No.1 datasets

3.2 Experimental Results Using No.2 Datasets

This subsection shows the experimental results obtained using the No.2 datasets (see Table 3). The same feature sets were used as in previous subsection. Table 7 shows the evaluation results using the First feature set. As compared to the results for the same feature set, obtained using the No.1 datasets, the present results are significantly lower - the increase in negative class examples resulted in efficiency recession for all classifiers.

Table 7. Evaluation results using the First feature set

CA	Set 1	Set 2	Set 3	Set 4	Set 5	Average	St.dev.
Naive Bayes	0.75	0.81	0.78	0.72	0.69	0.750	0.039
C4.5	0.58	0.78	0.69	0.69	0.61	0.672	0.069
CN2 Rules	0.64	0.83	0.78	0.75	0.69	0.739	0.067
Sensitivity							
Naive Bayes	0.00	0.14	0.00	0.14	0.14	0.086	0.070
C4.5	0.14	0.29	0.14	0.14	0.29	0.200	0.070
CN2 Rules	0.14	0.57	0.14	0.14	0.00	0.200	0.194
Specificity							
Naive Bayes	0.93	0.96	0.96	0.86	0.83	0.910	0.056
C4.5	0.69	0.89	0.83	0.83	0.69	0.786	0.083
CN2 Rules	0.76	0.89	0.93	0.89	0.86	0.869	0.059
False-Negatives rate							
Naive Bayes	0.21	0.18	0.20	0.19	0.20	0.195	0.010
C4.5	0.23	0.16	0.20	0.20	0.20	0.198	0.022
CN2 Rules	0.21	0.10	0.18	0.19	0.22	0.181	0.041

The results obtained using the Second feature set are given in Table 8. The efficiency of classifiers is better than in the case of using the First feature set (see Table 7), but still lower comparing to experimental results with the No.1 datasets.

The results of the final set of experiments with No.2 datasets using the Third feature set with six attributes, are shown in Table 9, the efficiency recession remains, showing that sampling process is highly target specific and, if used improperly, can decrease efficiency of a classifier.

Figure 3 depicts the average values of classification accuracy, sensitivity and specificity for all classifiers separately for each feature set, training and testing classifiers on Nr.2 subsets. Comparing to the results in Figure 2, the only measure that improved is the false-negatives rate.

3.3 Evaluation of the Best Classification Model Obtained

Analysing results obtained in Subsection 3.1 and 3.2, it was decided to evaluate on the initial dataset the best classification models that were obtained using

Table 8. Evaluation results using eight attributes

CA	Set 1	Set 2	Set 3	Set 4	Set 5	Average	St.dev.
Naive Bayes	0.61	0.89	0.83	0.72	0.75	0.761	0.096
C4.5	0.61	0.86	0.69	0.81	0.78	0.750	0.088
CN2 Rules	0.69	0.64	0.81	0.81	0.81	0.750	0.070
Sensitivity							
Naive Bayes	0.29	0.43	0.14	0.14	0.29	0.257	0.107
C4.5	0.14	0.29	0.14	0.57	0.29	0.286	0.090
CN2 Rules	0.43	0.14	0.14	0.00	0.29	0.200	0.146
Specificity							
Naive Bayes	0.69	1.00	1.00	0.86	0.86	0.883	0.115
C4.5	0.72	1.00	0.83	0.86	0.90	0.862	0.090
CN2 Rules	0.76	0.76	0.97	1.00	0.93	0.883	0.104
False-Negatives rate							
Naive Bayes	0.20	0.12	0.17	0.19	0.17	0.171	0.028
C4.5	0.22	0.15	0.20	0.11	0.16	0.168	0.040
CN2 Rules	0.15	0.21	0.18	0.19	0.16	0.179	0.023

Table 9. Evaluation results using six attributes

CA	Set 1	Set 2	Set 3	Set 4	Set 5	Average	St.dev.
Naive Bayes	0.61	0.86	0.83	0.78	0.81	0.778	0.088
C4.5	0.72	0.86	0.83	0.81	0.81	0.806	0.046
CN2 Rules	0.72	0.81	0.78	0.81	0.81	0.783	0.032
Sensitivity							
Naive Bayes	0.43	0.43	0.29	0.43	0.29	0.371	0.070
C4.5	0.14	0.29	0.29	0.00	0.29	0.200	0.114
CN2 Rules	0.29	0.29	0.14	0.00	0.29	0.200	0.114
Specificity							
Naive Bayes	0.66	0.97	0.97	0.86	0.93	0.876	0.117
C4.5	0.86	1.00	0.97	1.00	0.93	0.952	0.052
CN2 Rules	0.83	0.93	0.93	1.00	0.93	0.924	0.055
False-Negatives rate							
Naive Bayes	0.17	0.13	0.15	0.14	0.16	0.149	0.017
C4.5	0.19	0.15	0.15	0.19	0.16	0.169	0.021
CN2 Rules	0.17	0.16	0.18	0.19	0.16	0.172	0.015

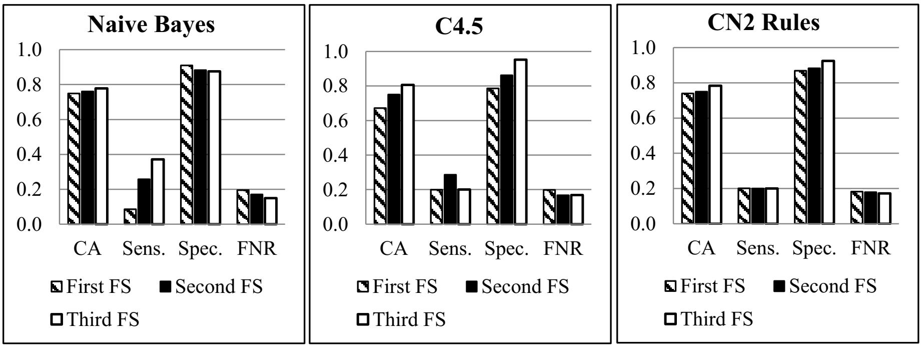

Fig. 3. Average values for No.2 datasets

sampled datasets. The higher sensitivity and the lowest false-negatives rate were reached training and testing classifiers on the second subset of the No.1 datasets, while using the Second feature set (see Table 5). The training records of the second subset were removed from the initial dataset, leaving 769 of 819 records for testing - 762 negative examples and 7 positive examples. Table 10 shows the obtained results.

Table 10. Experimental results using the best classification model

Classifier	CA	Sensitivity	Specificity	False-Negatives rate
Naive Bayes	0.86	1.00	0.80	0.00
C4.5	0.68	0.86	0.60	0.10
CN2 Rules	0.86	0.86	0.87	0.07

The obtained sensitivity of all three classifiers remained on the same level as in the results, described in Subsection 3.1 (see Table 5), however the specificity and classification accuracy decreased by 10% in general. Analysing obtained results it can be stated that classifiers show high true positive rate - sensitivity, resulting in correct diagnosis for the most of patients with gastric cancer, which is good. From the other hand, the specificity of classifiers remains on the level of 65-70%, which means that for about 30% of patients with negative diagnosis the decision support system suggested to make an endoscopy. This is a good result, comparing to the initial case, when an endoscopy was performed for each patient. Looking at the results from the other side - the false-negatives rate still remains above zero level. This means that some patiens with positive diagnosis will remain unthreated, meaning that the decision support system should not be used as a primary source for decision making. Figure 4 shows the classification tree built by the C4.5 algorithm. The tree returns good classification results, however contains some conflicting rules, like *IF Age* $\leq$ *66 AND Relatives does*

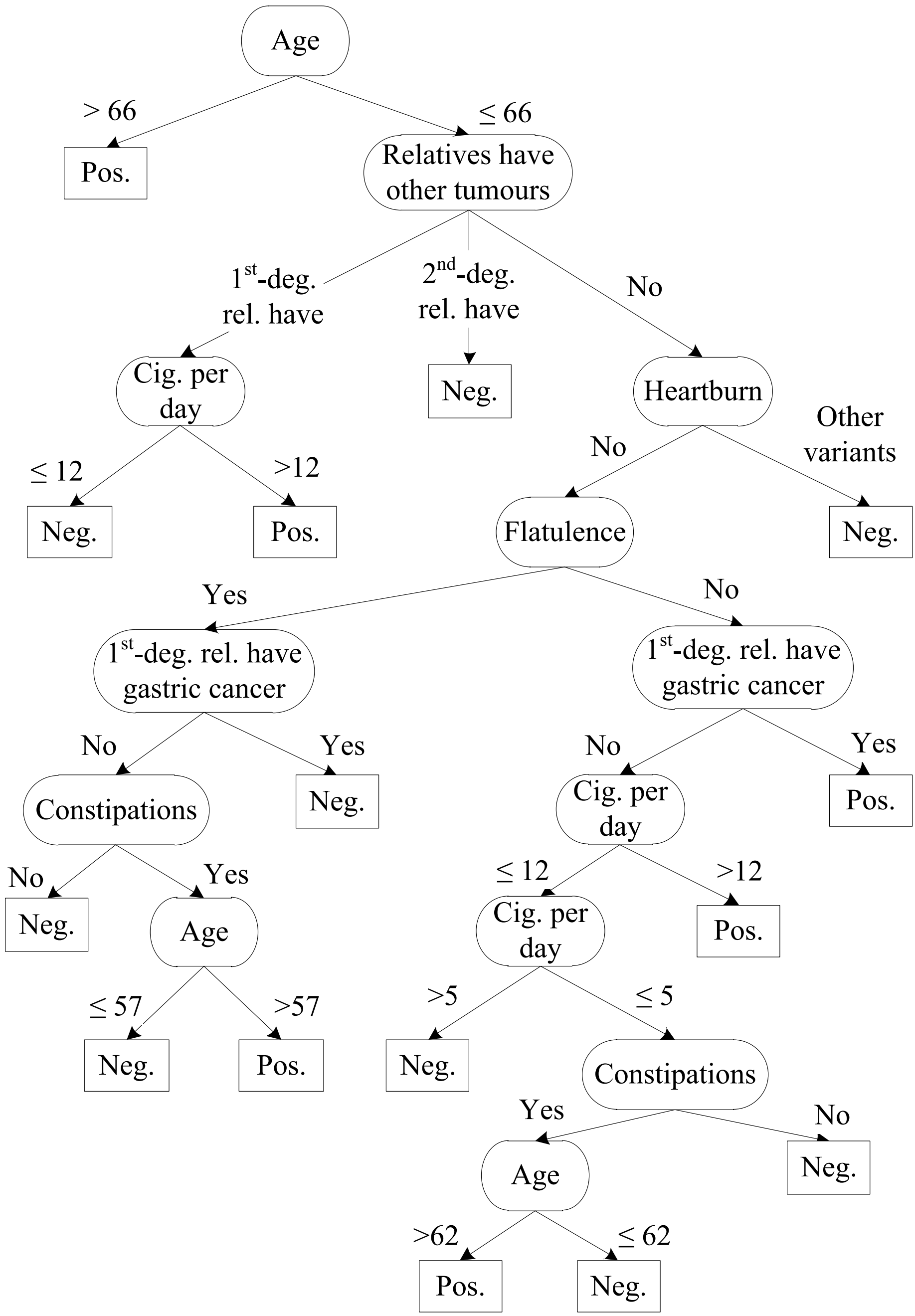

Fig. 4. Classification tree

not have other tumours AND Have no heartburn AND 1st-degree relatives have gastric cancer THEN result is Negative. This points out that classifiers are data specific and may contain rules that are normal for a training set, but do not concur with an opinion of a medical expert. That is one of the reasons for system to be a decision support not a decision system.

4 Conclusions

Gastric cancer is not the only disease the presented system can be used for. As it was mentioned earlier, the decision support systems are widely used in the healthcare. The present research showed one of possible applications of data mining methods in diagnosing the gastric cancer. The obtained results gave answers to different questions, connected with data preprocessing and especially feature selection and sampling, and defined directions for future research. The proposed decision support system is able to mine knowledge in medical data and apply it to evaluation of alternatives for each specific case. The experimental results have shown the average sensitivity greater than 50% and 86-100% at most, at the same time having classification accuracy and specificity close to 65-70% and false-negatives rate on the level of 20% on average. In comparison with an initial state when an endoscopy was performed for each patient, the application of the proposed DSS would lessen it by 70%, leaving 30% as false positives. The research in application of DSS in healthcare will be continued and for the future tasks it is planned to enlarge the initial dataset and recheck the results experimentally obtained and presented in the paper. Other option that will be considered is the application of association analysis and other anomaly detection techniques to mine knowledge in medical data.

Acknowledgements. This work has been partly supported by the European Social Fund within the project "Support for the implementation of doctoral studies at Riga Technical University".

This study was supported in part from the project of European Social Fund No. 009/0220/1DP/1.1.1.2.0/09/APIA/VIAA/016 "Multidisciplinary research group for early cancer detection and cancer prevention".

References

1. Aounallah, M., Quirion, S., Mineau, G.W.: Distributed Data Mining vs. Sampling Techniques: A Comparison. In: Tawfik, A.Y., Goodwin, S.D. (eds.) Canadian AI 2004. LNCS (LNAI), vol. 3060, pp. 454–460. Springer, Heidelberg (2004)
2. Clark, P., Niblett, T.: The CN2 induction algorithm. Machine Learning 3, 261–283 (1989)
3. Crew, A., Neugut, K.: Epidemiology of gastric cancer. World Journal of Gastroenterol 12, 354–362 (2006)
4. Dunham, M.: Data Mining Introductory and Advanced Topics. Prentice Hall (2003)

5. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann (2006)
6. Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., Forman, D.: Global cancer statistics. CA: A Cancer Journal for Clinicians 61, 69–91 (2011)
7. Lavrac, N., Flach, P., Kasek, B., Todorovski, L.: Rule induction for subgroup discovery with CN2-SD. In: 2nd International Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning, pp. 77–81. University of Helsinki (2002)
8. Lavrac, N., Flach, P., Kasek, B., Todorovski, L.: Subgroup discovery with CN2-SD. Journal of Machine Learning Research 5, 153–188 (2004)
9. Pyle, D.: Data Preparation for Data Mining. Morgan Kaufmann (1999)
10. Ruggieri, S.: Efficient C4.5. IEEE Transactions on Knowledge and Data Engineering 14, 438–444 (2002)
11. Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Education (2006)
12. World Health Organization (2011), http://www.who.int/en/