COMPUTATIONAL SCIENCE AND TECHNIQUES
Volume 1, Number 2, 2013, 155-170

Publisher: Klaipėda University
http://journals.ku.lt/index.php/CST
Online ISSN: 2029-9966

DATA MINING APPROACH TO PREDICT BRCA1 GENES MUTATION

Olegas Niakšu[1], Jurgita Gedminaitė[2], Olga Kurasova[1]
[1]Vilnius University, Institute of Mathematics and Informatics, Akademijos str. 4,
LT-08663 Vilnius, Lithuania
[2]Lithuanian University of Health Sciences, Oncology Institute, Eivenių str. 2,
LT-50009 Kaunas, Lithuania
olegas.niaksu@mii.vu.lt, jurgita.gedminaite@lmu.lt, olga.kurasova@mii.vu.lt

Abstract. Breast cancer is the most frequent women cancer form and one of the leading mortality causes among women around the world. Patients with pathological mutation of a BRCA gene have 65% lifelong breast cancer probability. It is known that such patients have different cause of illness. In this study, we have proposed a new approach for the prediction of BRCA mutation carriers by methodically applying knowledge discovery steps and utilizing data mining methods. An alternative BRCA risk assessment model has been created utilizing decision tree classifier model. The biggest challenge was a very small size and imbalanced nature of the initial dataset, which have been collected by clinicians during 4 years of clinical trial. Iterative optimization of initial dataset, optimal algorithms selection and their parameterization has resulted in higher classifier model performance, with acceptable prediction accuracy for the clinical usage. In this study, three data mining problems have been analyzed using eleven data mining algorithms.

Keywords: data mining applications, BRCA mutation, breast cancer, cancer reoccurrence prognosis, BRCA risk model.

Introduction

Nowadays, cancer is the second main cause of death in the developed countries and one of the main causes in the world. According to the statistics from World Health Organization, there are more than 10 million people, diagnosed with oncologic disease and about 6 million people will die for it in each year.

Breast cancer (BC) is the most common cancer in women worldwide. It is also the major cancer mortality reason among women. According to Parkin, about 89% of women diagnosed with BC are still alive 5 years after their diagnosis in Western countries, which is due to advances in detection and treatment (Parkin et al. 1999). Survivability is a major concern and is highly related with early diagnosis and optimal treatment plan.

BC diagnosis is a medical domain, which has a recognizable footprint in data mining applications. A number of articles (Bellaachia; Erhan 2006, Choi et al. 2009, Delen et al. 2005) investigates the utilization of different data mining (DM) methods: support vector machines, artificial neural networks, genetic algorithms, regression, etc.

The most popular DM models in the BC domain is diagnostic models which aim to distinguish benign from malignant tumor, or prognosis models, which predict BC patients survivability. However, less attention is paid to the more specific topics in the BC domain.

This paper deals with the issue of cancer suppression genes BRCA1 mutations. In this paper, we methodically apply a set of classical and emerging statistical and data mining

tools having a goal to answer questions formulated by clinicians, i.e. what are BRCA1 mutation prognostic factors, what are BC tumor reoccurrence factors, and if-and-what BRCA1 mutations influence to the course of decease.

The rest of the paper is organized as follows. Section 1 provides background information about BC disease, BRCA1 genes and DM applications in oncology. Section 2 deals with a real world example of a systematic DM methods application to the research data recently collected by oncology clinicians. The results of the research are outlined in Section 3. Finally the conclusions are provided in the last section.

1. Brief introduction to breast cancer decease, BRCA genes importance and data mining research in the domain

1.1. Breast cancer

Breast cancer is one of the leading causes of death among the women in many parts of the world. In 2008, approximately 1.90 million women across the world have been diagnosed with a form of invasive breast cancer. That results in a 23% of all cancers diagnosed in women and 11% of the total in men and women (Ferlay et al. 2008).

The process of breast cancer treatment typically starts from the identification of the malignant tumor. Hence, the information about the tumor from examinations and laboratory and radiology diagnostic tests are gathered. The stage of a cancer is one of the most important factors to define an optimum treatment plan. In general, the staging defines the spread of the cancer and its metastasis in the body. For the breast cancer TNM staging system is typically used, where T – Tumor, N – Nodes and M –Metastasis. First of all, patient's T, N, and M category values have been determined using gathered examinations and laboratory test data then this information is combined to determine a disease stage ranging from stage I to stage IV. The stage 0 also called carcinoma in situ, is an initial cancer stage, indicating high probability to develop invasive form of a cancer in a short period of time.

Despite significant efforts, scientists still do not know the exact causes of breast cancer; however some of the risk factors are known, i.e. genetic risk factors, family history, ageing, alcohol abuse, obesity. Thereof the current state of oncology research is highly dependent on genetic, clinical and treatment data collection and its analysis. The growing amount of heterogeneous data being collected in clinical settings highlights the importance of proper data mining techniques and application methodology.

1.2. BRCA genes

The gene named BRCA stands for breast cancer susceptibility gene. BRCA are human genes that belong to a class of genes known as tumor suppressors.

In normal cells, BRCA genes help ensure the stability of the cell's genetic material (DNA) and help prevent uncontrolled cell growth. Mutation of these genes has been linked to the development of hereditary breast and ovarian cancer. However not all mutations has proven breast cancer prognostic effect. A woman's risk of developing (pathogenic) breast

and/or ovarian cancer is greatly increased only if she inherits a deleterious (pathogenic) BRCA gene mutation. Men with these mutations also have an increased risk of a breast cancer. Both men and women who have harmful BRCA mutations may be at increased risk of other cancers.

It is known that BRCA gene mutations have regional and ethnical character. According to Janavicius, 86% of the mutation carriers from Lithuania and Latvia have BRCA c.4035delA and c.5266dupC mutations (Janavicius 2010).

The identification of patients having BRCA mutations is of great importance. Different risk models are used to calculate the likelihood of carrying a BRCA mutation. The BRCAPRO, Penn II, Myriad II, FHAT and BOADICEA models calculates risk on the basis of the inclusion of different cancer diagnoses within a family (Panchal et al. 2008). All models incorporate a family history of breast and ovarian cancer as a main prediction factor.

Penn II model, provided by Abramson Cancer Center of the University of Pennsylvania, has the best Sensitivity 0.93 among all mentioned risk models (Panchal et al. 2008).

In general, individuals with at least a 5–10% chance of having a mutation in either gene are considered good candidates for genetic testing. Identifying patients with BRCA mutation, allows applying risk-reducing preventive medical interventions, which is proven to be life-saving (National Cancer Institute 2013)

In this paper, we pay more attention to the analysis of other than family history predictors for carriers identification, and identification of any patterns, discriminating mutation carriers from non-carriers.

1.3. Medical data mining

Healthcare domain is known for its ontological complexity and variety of medical data standards and variable data quality (Chen et al. 2005, Cios et. al. 2002, Wilson et al. 2003). Adding to this privacy consideration makes medical knowledge discovery an open subject for the research during last decades. According to Fayyad et al. (1996), the knowledge discovery process comprises of a few steps: data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge representation. In this paper, we mostly address data mining step, covering data preparation steps.

Typically, the available medical datasets are fragmented and distributed; thereby the process of data cleaning and integration is a challenging task. Other important issues related to the use of personal healthcare data have origins in legal, ethical and social aspects. Various DM methods are used for the clinical decision support. Classification, clustering and association rules are among the most popular.

The literature analysis of DM application in oncology domain and particularly within breast cancer has shown constant academic interest in the field. We have found 9 articles published within 5 years and more than 30 articles have been published since year 2000 contributing to the topics of using DM to diagnose BC or to predict its development (DM techniques applied for automated image analysis were not counted). After carrying out a

literature screening, we came to the conclusion, that the typical topics of DM application in cancer research can be summarized as follows:

1) diagnosing various cancer forms (diagnosing breast cancer),
2) predicting patient's survivability,
3) predicting recurrence,
4) finding dependencies, patterns among clinical, pathological attributes of a patient.

Below we list examples of the latest research publications where data mining techniques have been used within BC research. There are currently no publications known to the authors in the field of DM application for BRCA mutations research.

Comparative analysis of different DM methods, i.e. Naïve Bayes, Artificial Neural Network (ANN) and C4.5 algorithm has been used by Bellachia A. et al. (2006) for the prediction of breast cancer patients' survivability rate over 7 variables. The authors have introduced the pre-classification step into knowledge discovery process. Their conclusion is that C4.5 algorithm outperformed its rivals.

Delen D. et al. (2005) compared ANN, decision trees, SVM and logistic regression techniques for BC survival analysis on twenty variables. Support vector machines method was the most accurate predictor with a test dataset accuracy of 92.85%.

Shukla A. et al. (2009) used ANN and neuro-fuzzy systems to create an optimal model for early BC diagnosis. The best results were achieved by feed-forward back-propagation ANN (accuracy 99.5%, no published information on the dataset).

Choi J.P. et al. (2009) also have compared the performance of an ANN to other methods, i.e. Bayesian Network and Hybrid Network to predict breast cancer prognosis over 9 variables. Finally they combined all three methods together. The best results were achieved by ANN with accuracy of 88.8%, following by Hybrid Network and Bayesian Network.

2. Predicting BRCA1 mutation and analyzing its influence applying data mining techniques

2.1. Research Data

The original medical research had been carried out in Oncology institute of Lithuanian University of Health Sciences from 2010 till 2013. The study group consisted of 83 women, who have been diagnosed with I-II stage breast cancer with the following tumor morphology: T1 N0, T2 N0, T3 N0, T1 N1, T2 N1. The list of observed clinical, morphological features (attributes), as well as interventions and therapies applied is provided in Table 1, together with attribute types and the number of distinct values of each nominal attribute quantities of nominal attributes.

Research duration was determined considering amount of patients and not less than 2 years period of disease progress monitoring. As the cancer stage is a strong predictive factor, only the early (I - II stage) breast cancer have been chosen to reduce the factors influencing the variation.

After laboratory confirmation of pathologic BRCA mutation all patients have been divided into 2 groups: (1) carriers – patients with pathologic BRCA gene mutation, and (2) non-carriers – patients without BRCA mutation.

The patients have been monitored by clinical monitoring protocol till 31st December, 2012. Monitoring of the breast cancer progress was initiated starting from morphological breast cancer diagnosis till the end of the trial. Both the local breast cancer recurrence in the same breast, and distant metastases of visceral organs, skeleton, skin, or central nervous system structures have been counted as breast cancer reoccurrence. However, malignant tumor development in another breast was not considered as cancer reoccurrence.

Table 1. The full list of attributes of initial dataset.

#	**Attribute**	**Attribute type***		#	**Attribute**	**Attribute type***
1	Age	Continuous		18	Triple neg. BC	Nominal (2)
2	Histology type	Nominal (5)		19	Family history type	Nominal (3)
3	cT	Nominal (5)		20	Prostate cancer fam. Hist.	Nominal (2)
3	pT	Nominal (6)		21	Pancreatic cancer fam. hist.	Nominal (2)
4	Multifocality	Nominal (2)		22	Colorectal cancer fam. Hist.	Nominal (2)
5	cN	Nominal (3)		23	Surgery type	Nominal (4)
6	pN	Nominal (2)		24	Chemotherapy type	Nominal (3)
7	G	Nominal (3)		25	Herceptin	Nominal (2)
8	L	Nominal (2)		26	Cht. complications	Nominal (4)
9	V	Nominal (2)		27	Reoccurrence	Nominal (2)
10	ER	Nominal (4)		28	Metastases	Nominal (2)
11	PR	Nominal (4)		29	Time to diseased	Continuous
12	HER2	Nominal (2)		30	Is Diseased	Nominal (2)
13	BRCA mutation	Nominal (6)		31	Monitoring period	Continuous
14	Bilateral BC	Nominal (2)		32	Time to reoccurrence	Continuous
15	Tumor size	Continuous		33	Adjuv. ST	Nominal (2)
16	CHEK2 mutation	Nominal (4)		34	Adjuv. HT	Nominal (5)
17	Affected l_m number	Continuous				

* For nominal attributes, a number of distinct values is given in brackets.

2.2. Knowledge discovery and data mining application

Firstly, we have applied statistical analysis methods, which are typically applied in medical research. The correlation of gathered attributes in the dataset were tested using χ^2 criterion with α=0.95. The cancer reoccurrence survival analysis was performed with Cox regression model and Kaplan–Meier. The statistical analysis revealed a few statistically significant attribute dependencies. BRCA1 mutation has statistically significant dependency on family history (p=0.001), age (p=0.001), tumor grade degree (p=0.004), progesterone receptors (p=0.03), triple negative BC (p=0.001).

The results of the statistical analysis have been already described in the article "A research of breast cancer patients with BRCA1/2 mutation and its influence to tumor

biological characteristics and disease progress", which has been submitted to the journal "Lietuvos bendrosios praktikos gydytojas" (Lithuanian General Practitioner).

Further analysis of the collected research data was performed applying a set of data mining techniques. In the scope of our study, we aimed to fine-grain a few questions of a clinical interest which have been not proved by classical statistical methods:

- Do patients with *BRCA1* pathogenic mutation have any specific clinical, morphological manifestations?
- What other patient features or feature groups can serve as predictors of pathogenic *BRCA1* mutation?
- If there are any predictive factors of breast cancer reoccurrence?
- If there is an impact of *BRCA1* mutation to the time of tumor reoccurrence?

According to Speckauskaite (2011) data mining in a clinical domain needs an iterative approach of data pre-processing, selection of optimal algorithm and its optimal parameterization. Following the proposed method and recommendations (Speckauskaite 2011, Bellazzi; Zupan 2008) we have applied an iterative procedure to define the Optimal Dataset (ODS) for each task, afterwards we found the best performing classification algorithm, then we optimized its parameters and finally validated the results with the domain experts (clinicians).

The collected research data, which formed the initial dataset had a very imbalanced structure. As showed in Table 2, the carriers counted 14%, and non-carriers – 86% of the whole patient group.

Table 2. The distribution of prediction class attributes.

Attribute	Positive attribute value		Negative attribute value	
	Number of patients	Percentage of the whole group	Number of patients	Percentage of the whole group
BRCA1 mutation	12	14%	71	86%
BC reoccurrence	22	27%	61	73%
Diseased patients	2	2%	81	98%

The following preprocessing steps were applied. Continuous attributes Age, Tumor size, Time to reoccurrence were discretized accordingly to D_age group, D_tumor size, and D_time to reoccurrence. Dataset was checked for outliers and missing values.

The set of 19 nominal attributes was used for BRCA1 mutation classification task. The attributes values frequency tables are visually shown by BRCA class colored histograms in Fig. 1. Blue color marks items with no BRCA1 mutation, and red color – items with a BRCA1 mutation.

Nominal attributes value distribution, as it can be visually seen in Fig. 1 does not indicate trivial single nominal attribute value dependency on dependent variable (BRCA BIT attribute).

For the dimensionality reduction, feature subset selection algorithms were used. More details are provided in Section 2.2.1.

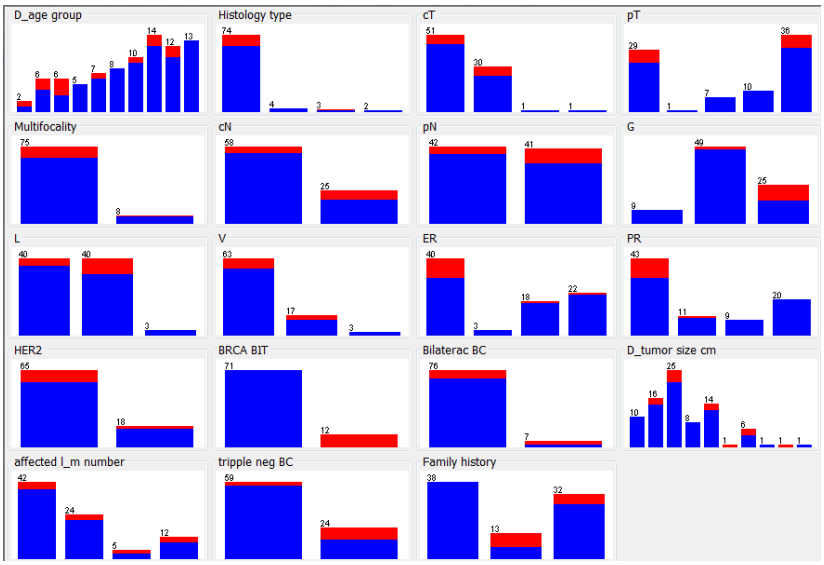

Figure 1. Histograms of nominal attributes values for BRCA1 classification task.

The research questions raised have been formulated as classification problems. Classification models for the prediction of BRCA1 carrier with dependent variable BRCA mutation, and for the prediction of BC reoccurrence with dependent variable Reoccurrence have been created. Also, association rules were used to identify hidden dependencies between dependent and independent variables. K-fold cross-validation technique was used for classification models evaluation. Time series analysis was carried out to evaluate if BRCA1 mutation influences time to reoccurrence or time to death.

A comparative analysis of the classification techniques was performed using algorithms implemented in WEKA (Hall et al. 2009), Orange (Curk et al. 2005) and Tibco Spotfire Mining (Tibco Software Inc. 2010):

- Classification trees – J48[1], Random Forest, Random tree, tree ensemble.
- Classification rules – ZeroR[2], OneR[3], and FURIA[4].
- Artificial neural networks – Multi-layer Perceptron, SOM[5].
- Regression – logistic regression
- Bayes – Naïve Bayes
- Meta – Ada Boost[6], Bagging

In our study, we can symbolically divide knowledge discovery process into two major iterations. Within the first iteration the main data preprocessing activities were carried out, and then different DM algorithms and their parameterizations were used to achieve the best performing result. In the second iteration, we have changed optimal dataset by

[1] J48 – WEKA implementation of C4.5 algorithm
[2] ZeroR – WEKA implementation of classification algorithm, using 0-R classifier
[3] OneR – WEKA implementation of classification algorithm, using 1-R classifier
[4] FURIA - fuzzy unordered rule induction classification algorithm
[5] SOM – self organizing map, a data visualization algorithm
[6] Ada Boost – boosting classification algorithm using Ada Boost M1 method

equaling the proportion of the records with all distinct values of the dependent attribute, and then applied the same set of DM algorithms to find the winning classifier. Afterwards the best model was tested on the initial dataset applying 10 fold cross-validation technique.

2.2.1. The results of the first iteration

Our first objective was to evaluate different classification algorithm types, and to find the most appropriate for our task. We had no noisy or missing data constraints, also we had a flexibility of discretizing continues data to nominal for the algorithms handling nominal attributes better and for getting the results, which are more meaningful for clinical interpretation.

Secondly, we have improved the classification results by changing default algorithm parameters. The algorithm parameterization had the results as follows:

Fuzzy Unordered Rule Induction Algorithm (FURIA) showed overall performance improvement after changing uncovered rules handling parameter to "vote for the most frequent class". Main algorithm parameters have been set as follows: T-Norm equals to Product T-norm, error rate $> \frac{1}{2}$ as stopping criterion, 2 optimization runs, 3 folds for pruning, random seed equals to 1, minimal weight of the instances in a rule equals to 2. See the details for BRCA1 classification problem in Table 3.

Table 3. FURIA algorithm optimization results.

Algorithm	Accuracy	Sensitivity	Specificity	ROC AUC
Furia initial	0.916	0.667	0.958	0.80
Furia optimized	0.940	0.667	0.986	0.81

Adaptive boosting meta-algorithm *AdaBoost* is known for good results with weak classifiers and is more resistant to overfitting. *AdaBoostM1* WEKA implementation was used. We have achieved Sensitivity improvement from 0.500 to 0.667 by using *DecisionStump* as a basis classifier and increasing the iteration number from 10 to 30. Main algorithm parameters have been set as follows: reweighting resampling was not used, weight threshold for weight pruning equals to 100, random seed equals to 1. However Specificity and ROC area values have reduced. See the details for *BRCA1* classification problem in Table 4.

Table 4. AdaBoost algorithm optimization results.

Algorithm	Accuracy	Sensitivity	Specificity	ROC AUC
AdaBoostM1 initial	0.891	0.5	0.958	0.802
AdaBoostM1 optimized	0.892	0.667	0.930	0.790

Another meta-algorithm bootstrap aggregating (bagging) results were improved by choosing J48(C4.5) as a base classification algorithm. Main algorithm parameters have been set as follows: 1 execution slot, 10 iterations, random seed equals to one, out-of-bag error was not calculated, bag size percentage equals to 100%. See the details for BRCA1 classification problem in Table 5.

Table 5. Bagging algorithm optimization results.

Algorithm	Accuracy	Sensitivity	Specificity	ROC AUC
Bagging with RepTree	0.855	0	1	0.705
Bagging with J48	0.880	0.417	0.958	0.853

The overall result of the first iteration is shown in Tables 6 and 7.

Table 6. *BRCA1* classifier models performance.

Algorithm	Accuracy	Sensitivity	Specificity	ROC AUC
J48 (C4.5)	0.880	0.667	0.915	0.825
Random Forest	0.855	0.167	0.972	0.774
Random tree	0.819	0.333	0.901	0.696
ZeroR	0.854	0.000	1.000	0.428
OneR	0.807	0.000	0.944	0.472
Furia	0.940	0.667	0.986	0.81
Multilayer perceptron	0.819	0.667	0.845	0.805
Multilayer perceptronCS	0.916	0.667	0.958	0.865
Logistic regression	0.795	0.500	0.845	0.738
AdaBoostM1	0.892	0.667	0.930	0.790
Bagging with J48	0.880	0.417	0.958	0.853

Table 7. Breast cancer reoccurrence classifier models performance.

Algorithm	Accuracy	Sensitivity	Specificity	ROC AUC
J48 (C4.5)	0.734	0.000	1.000	0.457
Random Forest	0.71	0.091	0.934	0.516
Random tree	0.639	0.227	0.787	0.484
ZeroR	0.735	0.000	1.000	0.457
OneR	0.675	0.000	0.918	0.459
Furia	0.747	0.091	0.984	0.633
Multilayer perceptron	0.687	0.455	0.770	0.576
Multilayer perceptronCS	0.687	0.455	0.770	0.596
NaïveBayes	0.639	0.136	0.820	0.508
Logistic regression	0.663	0.591	0.689	0.675
AdaBoostM1	0.651	0.000	0.885	0.319
Bagging with J48	0.687	0.045	0.918	0.546

Dimension reduction techniques including Principal Component Analysis, Particle Swarm Optimization based attribute search, Chi Squared attribute evaluation and Correlation Attribute evaluation have been used. The methods resulted in different

attribute sets. In our experiments, Particle Swarm Optimization algorithm for the attribute search has shown the best results.

However, the most of them had significantly worse classification accuracy comparing to the dataset with the full set of attributes. See Fig. 3 and Fig. 4 for different classifier models performance comparison. The only possible advantage of the dimension reduction is a shorter classification model building time, which was not applicable due to the small research dataset.

2.2.2. Association rules discovery

Additionally association rules discover algorithms were applied to test for non trivial dependencies. Apriori, PredictiveApriori, and HotSpot algorithms were used. Generic and class specific rules with a minimum support in the range of [0.01; 0.2] with confidence greater than 0.75 were searched using WEKA and Tibco Spotfire Miner.

Three set of rules were discovered iteratively: class independent, and class dependent with the BRCA mutation and Reoccurrence as class attributes. The search space was incrementally increased by decreasing minimum support and confidence values and by increasing maximum number of antecedents from 2 to 5, and by increasing the associated attribute set from 5 (attributes found in 1st iteration by dimension reduction techniques) to the full set of 35 attributes. In the largest search space within our experiments, association rules search has found from 46 thousand to 78 thousand rules. Such amount of the rules is due to the selected lower support and confidence value. The generated rule items were filtered and then analyzed by the clinician.

2.2.3. The second iteration results

In the second iteration, we have changed ODS by incrementally equaling the proportion of dependent binary (class) attribute values till it reached 50% to 50% distribution. The balancing of ODS gave significant results to the most of the classification algorithms. The results of the best performing algorithms are provided in Table 8 and Table 9. The initial dataset equal parts stratification was used to produce a balanced training dataset for the classifier. The classifiers based on the balanced ODS showed 0.90 accuracy, 0.95 Sensitivity, 0.85 Specificity and 0.96 ROC area value with meta algorithm Bagging, and 0.88 accuracy, 0.93 Sensitivity, 0.83 Specificity and 0.85 ROC area value with J48 tree algorithm.

Then classifier model was exported to PMML (Predictive Model Markup Language) format in Tibco Spotfire Miner and to WEKA model format in WEKA environment. Finally, the initial unbalanced dataset was used as a test dataset for the validation of the model.

The comparison of the classifier models performances was done using ROC and Gain charts. The best performing classification models tested on unbalanced dataset are presented in Table 8 and Table 9.

Table 8. *BRCA1* prediction classifier.

Algorithm	Accuracy	Sensitivity	Specificity	ROC AUC
Bagging	0.867	0.833	0.873	0.81

Table 9. Breast Cancer reoccurrence prediction classifier.

Algorithm	Accuracy	Sensitivity	Specificity	ROC AUC
Bagging	0.711	0.955	0.623	0.65

Comparing to the first iteration results, higher Sensitivity was achieved, but in turn Specificity has decreased, resulting in lower overall performance with ROC AUC value 0.81 for BRCA1 classifier which is weaker comparing to the performance of the first iteration classifier where ROC AUC was 0.85.

Visual comparison of the best performing Bagging algorithm classifiers is provided in Fig. 2 for BRCA1 class model and in Fig. 3 for Reoccurrence class model. ROC charts are often used to compare the performance of different models. The charts display Sensitivity of the models versus their Specificity. Cumulative Gain charts display the percentage of positive responses predicted by the models versus the percentage of the population. The best model for the data is the one with the highest curve above the straight diagonal line.

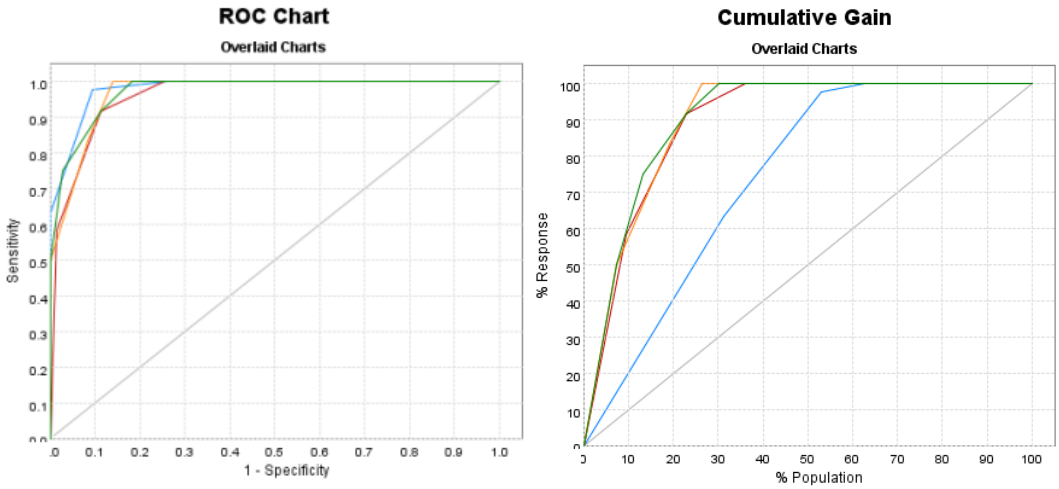

Figure 2. BRCA mutation prediction models performance charts.

In BRCA1 mutation classification ROC chart (Fig. 2), the best performance (blue line) is achieved by "stratified data" classifier trained and tested solely on balanced dataset, then the performance gradually decreases as follows: ODS with all 29 attributes (green line), ODS after dimension reduction with 5 attributes (orange line), "stratified data" classifier tested on initial dataset (red line). The same color notation is used for cumulative gain chart (Fig.2). The Gain ranking of the models is as follows: orange, green and red lines have similar Gain values, and then the blue line which represents "stratified data" classifier trained and tested solely on the balanced dataset has a lower Gain value.

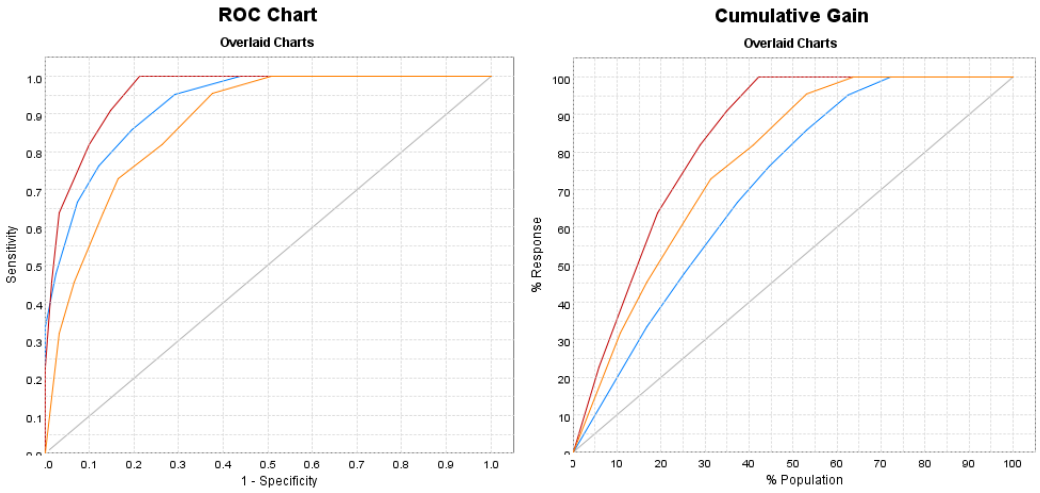

Figure 3. BC reoccurrence prediction models performance charts.

In BC reoccurrence classification ROC chart (Fig. 3), the best performance (red line) is achieved by ODS with all 34 attributes. The performance gradually decreases as follows: "stratified data" classifier trained and tested solely on the balanced dataset (blue line), and finally "stratified data" classifier tested on initial dataset (orange line). The same color notation is used for cumulative gain chart (Fig. 3). The Gain ranking of the models is as follows: ODS with all 34 attributes (red line), "stratified data" classifier tested on initial dataset (orange line), and finally "stratified data" classifier trained and tested solely on the balanced dataset (blue line).

2.2.4. Time series and survival analysis

Additionally survival and time-series analysis was performed to find any impact of BRCA1 mutations to the time of BC reoccurrence or to the time of death. In the cases of systematic reoccurring BC, we have researched possibilities to predict localization of metastases.

However, neither statistical linear regression or Cox regression, nor DM methods provided satisfactory results. The received results were statistically insignificant or without reasonable accuracy.

More than hundred association rules describing cancer metastases in lungs have been discovered by Apriori algorithm. However, all of them were rejected by the clinician expert as being trivial or being possibly resulted by algorithm over-fitting. Therefore we state that there were no prove found to support the hypothesis, that BRCA1 mutations have an impact to BC reoccurrence time and to time to death. Also, there is no proof of the impact of metastases localization.

3. Generalization of data analysis results

A set of different class data mining techniques has been applied for the advanced analysis of breast cancer patients' data. Two main classification problems have been formulated:

BRCA1 gene mutation prediction and BC reoccurrence prediction. The experimental work has been carried out in data mining platforms Weka, Orange and Tibco Spotfire Miner. The initial dataset consisting of 83 items was iteratively optimized to achieve better performance of DM algorithms. 10-fold cross validation was used for classifier models training and validation.

The BRCA1 classifier model with the best ROC AUC value was created using Multilayer Perceptron algorithm modification (MultilayerPerceptronCS in WEKA) with overall accuracy 0.92, Sensitivity 0.67, Specificity 0.96 and ROC AUC 0.87. However higher classifier Sensitivity and explicit interpretability of a model was required by the clinicians. Therefore decision tree J48 and decision rules Furia classifiers were used for the interpretation by the domain experts. Their performance is as follows: overall accuracy 0.88 and 0.94, Sensitivity 0.67 in both cases, Specificity 0.92 and 0.99, and ROC AUC 0.83 and 0.87.

To increase Sensitivity value, ODS was balanced and the best classifier results were achieved with Bagging algorithm. Its performance on the test dataset (10-fold cross-validated initial dataset): overall accuracy 0.87, Sensitivity 0.83, Specificity 0.87 and ROC AUC 0.81. The clinical interpretation of the decision trees and decision rules is presented in Section 4.

The optimal breast cancer reoccurrence classifier models were created in the second iteration, when the initial dataset was balanced, which significantly improved Sensitivity with remaining similar levels of Specificity and ROC AUC. The achieved performance of Bagging algorithm classifier: overall accuracy 0.71, Sensitivity 0.96, Specificity 0.62 and ROC AUC 0.65. The highest Specificity was achieved applying Furia decision rules algorithm: overall accuracy 0.75, Sensitivity 0.09, Specificity 0.98 and ROC AUC 0.63. The clinical interpretation of the decision trees and decision rules is presented in Section 4.

Survival and time-series analysis which was performed to find an impact of BRCA1 mutations to the time of disease reoccurrence or to the time of death has not shown any results with predictive value or allowing to reject Null hypothesis.

Discussion and conclusion

By analyzing breast cancer patient data, we have realized the importance of systematic approach in knowledge discovery process. The study has showed high importance of an optimal dataset forming for the classification accuracy. A dataset with balanced class attribute values was of key importance. Our experiments have not shown the positive impact of dimension reduction for the model accuracy.

Artificial neural networks have showed the best performance for BRCA1 gene mutation carrier prediction, but due to the lack of its expressivity, decision tree and decision rules methods were preferred by the clinicians. A few rules have been identified as potentially interesting for further analysis when larger patient dataset is available. These classification rules have been analyzed by the clinicians, compared to the current medical knowledge, and are discussed below.

Clinical conclusions

Remarkably, currently used publicly available clinical BRCA risk evaluation models are based purely on patient's family history, whilst our classifier models provide similar and in some cases better accuracy by including clinical and morphological patient features.

Our created breast cancer reoccurrence models have not included BRCA mutation as a possible predictor for a patient group with a recurrent tumor. This finding supports the researches (Robson et al. 2004), implying the importance of tumor's clinical-morphological features and diminishing the impact of BRCA mutation to the breast cancer reoccurrence. Though, other researches have reported on the worse survival rate for BRCA carriers (Brekelmans et al. 2009). Our prediction models have reconfirmed already known and used in clinical practice criteria, which indicate possible BRCA1 mutation. The family history attribute has high predictive value, especially when combined with clinical and morphological features such as bilateral BC, high grade tumor, medullary carcinoma, triple negative BC. Interestingly, classification trees models highlighted negative expression of progesterone receptors as possible BRCA1 mutation predictor, which is significantly narrower discrimination condition comparing to triple negative BC, which additionally includes estrogen R(-) and HER2(-) features.

Another finding is higher BRCA1 mutation probability for the patients with tumor size greater than 1 cm or when more than one axillary lymph node is affected. This can be explained by higher grade of BRCA1 associated tumors and higher proliferation.

By analyzing BC reoccurrence classifiers, contradictory rule was inducted. The rule depicts higher reoccurrence rate for the patients with higher estrogen receptors expression with average or higher grade and larger than 2 cm tumor. This finding is in contradiction with the known clinical practice. This rule most probably is explained by the limited dataset and shall be revalidated in the future research.

BC reoccurrence classifier reconfirmed the prognostic features approved in previous clinical researches: higher tumor grade, primary tumor size, negative progesterone receptors, young patient age, and type of chemotherapy used.

After additional validation on a larger dataset the created prediction models can be used as clinical decision support systems.

References

Bellaachia, a.; Erhan, g. (2006). Predicting Breast Cancer Survivability using Data Mining Techniques. 9th Workshop on Mining Scientific and Engineering Datasets. 6th SIAM International Conference on Data Mining.

Bellazzi, R.; Zupan, B. (2008).Predictive data mining in clinical medicine: Current issues and guidelines. International journal of medical informatics, 77, p. 81–97.

Brekelmans, C.T.; Tilanus-Linthorst, M.M.; Seynaeve, C.; et al. 2007. Tumor characteristics, survival and prognostic factors of hereditary breast cancer from BRCA2-, BRCA1- and non-BRCA1/2 families as compared to sporadic breast cancer cases. European Journal of Cancer, 43(5):867-76

Chen, H.; Fuller, S.; Friedman, C.; Hersh, W. (2005). Medical Informatics. Knowledge Management and Data Mining in Biomedicine. Springer Science

Choi, J.P.; Han, T.H.; Park, R.W. (2009). A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis. Journal of Korean Society of Medical Informatics, p. 49-57.

Cios, K. J.; Moore, G. W. (2002). Uniqueness of medical data mining. Artificial Intelligence in Medicine, 26. p. 1–24.

Curk, T.; Demšar, J.; Xu, Q.; Leban, G.; Petrovič, U.; Bratko, I. et al. (2005). Microarray data mining with visual programming. Bioinformatics, vol. 21(3), p. 396-398.

Delen, D.;Walker, G.;Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine, vol. 34, p. 113-127.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

Ferlay, J.; Shin, H. R.; Bray, F.; et. al. (2008). Cancer Incidence and Mortality Worldwide. International Agency for Research on Cancer. [internet] [Accessed: May 2013]. Available from: http://globocan.iarc.fr/

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, vol. 11(1), p. 10-18.

Janavičius, R. (2010). Founder BRCA1/2 mutations in the Europe: implications for hereditary breast-ovarian cancer prevention and control. EPMA Journal, 1(3):397-412.

National Cancer Institute, USA. BRCA1 and BRCA2: Cancer Risk and Genetic Testing. [internet] [Accessed: May 2013]. Available from: http://www.cancer.gov/cancertopics/factsheet/Risk/BRCA

Panchal, S. M.; Ennis, M.; Canon, S.; Bordeleau, L. J. (2008). Selecting a BRCA risk assessment model for use in a familial cancer clinic. BMC Medical Genetics, 9:116.

Parkin, D. M.; Pisani, P.; Ferlay, J. (1999). Global cancer statistics. A Cancer Journal for Clinicians, Vol. 49, Issue 1, p. 33–64.

Robson, M.E.; Chappuis, P.O.; Satagopan, J; et al. (2004). A combined analysis of outcome following breast cancer: differences in survival based on BRCA1/BRCA2 mutation status and administration of adjuvant treatment. Breast Cancer Research, 6(1):R8-R17.

Shukla, A.; Tiwari, R; Kaur, P. (2009). Knowledge based approach for Diagnosis of Breast Cancer. IEEE International Advance Computing Conference (IACC 2009). Patiala, India.

TIBCO Software Inc. (2010). TIBCO Spotfire Miner™ 8.2 User's Guide. [internet] [Accessed: May 2013]. Available from: http://stn.spotfire.com/stn/Site/NewsSPlus81.aspx

Wilson, A.; Thabane, L.; Holbrook, A. (2003). Application of DM techniques in pharmacovigilance, British Journal of Clinical Pharmacology (57) 2, p. 127-134.

Špečkauskienė, V. (2011). Development and analysis of informational clinical decision support method. Kaunas: Technologija.

Olegas Niaksu is a PhD student of the Institute of Mathematics and Informatics at Vilnius University. He conducts a research in the field of data mining applications within healthcare domain, combining scientific approach with his international professional experience in e-Health and medical informatics. ACM member, certified project manager, Microsoft certified professional.

Jurgita Gedminaitė graduated from the Faculty of Medicine of Kaunas University of Medicine in 2002. From 2007 has the position of medical oncologist in Hospital of Lithuanian University of Health Sciences Kaunas Clinics. Member of the Lithuanian Society for Medical Oncology and the European Society for Medical Oncology. Research interest – neuro-oncology, breast cancer.

Olga Kurasova received the doctoral degree in computer science (PhD) from Institute of Mathematics and Informatics jointly with Vytautas Magnus University in 2005. Recent employment is at the System Analysis Department of the Vilnius University, Institute of Mathematics and Informatics as senior researcher, and at the Informatics Department of Lithuanian University of Educational Sciences as associate professor. Research interests include data mining methods, optimization theory and applications, artificial intelligence, neural networks, visualization of multidimensional data, multiple criteria decision support, parallel computing. She is the author of more than 40 scientific publications.

DUOMENŲ TYRYBA BRCA1 GENŲ MUTACIJOS PROGNOZEI

Olegas Niakšu, Jurgita Gedminaitė, Olga Kurasova

Santrauka

Krūties vėžys yra dažniausiai moterims diagnozuojama vėžio rūšis ir viena iš dažniausiai pasitaikančių moterų mirties priežasčių visame pasaulyje. Pacientėms, turinčioms mutavusį BRCA geną, tikimybė susirgti krūties vėžiu siekia net 65 proc. Taip pat žinoma, kad turinčioms mutavusį BRCA geną pacientėms ligos eigą lemia skirtingos priežastys. Straipsnyje siūlome naują mutavusio BRCA geno nešiotojų identifikavimo būdą, metodiškai taikant žinių išgavimo žingsnius ir naudojant duomenų tyrybos metodus. Alternatyvus BRCA rizikos nustatymo modelis sukurtas, naudojant sprendimų medžių klasifikavimo modelį. Šio tyrimo ypatumas buvo labai mažos apimties ir nevienalyčiai pradiniai duomenys, kuriuos medikai surinko per ketverius klinikinių tyrimų metus. Iteracinis pradinių duomenų optimizavimas, tinkamų algoritmų parinkimas ir jų parametrų nustatymas lėmė didesnį pasirinkto klasifikavimo modelio efektyvumą ir priimtiną klinikiniam naudojimui prognozavimo tikslumą. Straipsnyje analizuojamos trys duomenų tyrybos uždaviniai, naudojant vienuolika duomenų tyrybos algoritmų.

Pagrindiniai žodžiai: duomenų tyrybos taikymai, krūties vėžys, vėžio remisijos prognozavimas, BRCA mutacijos, BRCA rizikos modelis