https://doi.org/10.31891/csit-2023-4-2
UDC 004.62: 616.006

Gennady CHUIKO, Olga YAREMCHUK
Petro Mohyla Black Sea National University

HANDLING THE BREAST CANCER RECURRENCE DATA FOR A MORE RELIABLE FORECAST

Breast cancer in women is a global problem that affects the gene pool. This sickness has become a prevalent cancer threat for Ukrainian women, while early detection and prophylactics notably raise survival chances, dropping the cost of treatment. Recurrence event control and forecasting are vital field areas of this problem.

This article deals with data that permits via machine-learning breast cancer recurrences in patients undergoing the therapy. The renewed data set presented in this paper contains 252 cases, of which 206 did not have recurrent events, but 46 did. This data set is an improved version of the well-known Ljubljana breast cancer data set from 1988.

The aim is a lift in the reliability of clinical prognoses of breast cancer recurrence using the updated and improved LBCD. The list of tasks accompanying this goal is as follows: Estimating relevance ranks for LBCD attributes; Evaluations of noise levels for attributes, mainly for the class attribute; Reduction of the dataset by removing irrelevant and noisy data; Imputing (restoring) the missed values for the class attribute; The simile of the performance for the initial and upgraded dataset.

Our updated dataset has fewer instances (252 instead of 286) and fewer attributes (six instead of ten), aside from the class attribute being noise-cleaned and its missed values being restored. As a result, the performance of the upgraded data set is much better than the original one, especially concerning cases of recurrence cancer. It allows clinicians a more reliable machine-learning diagnosis of breast cancer recurrence using the most known classifiers.

The used dataset is helpful in machine learning models' devising, which shall classify, detect, and forecast probabilities of recurrence events of breast cancer in clinics. The elaborated dataset ensures a much higher performance for machine learning algorithms than the initial prototype. Compared to the prototype, the dataset is more compact, comprising 252 instances instead of 286 and 6 attributes instead of 10. This dataset's class (category) attribute is entirely free of noise.

Keywords: machine learning, breast cancer dataset, recurrence events, noise cleaning, performance improving.

Геннадій ЧУЙКО, Ольга ЯРЕМЧУК
Чорноморський національний університет імені Петра Могили

ОБРОБКА ДАНИХ ПРО РЕЦИДИВИ РАКУ МОЛОЧНОЇ ЗАЛОЗИ ДЛЯ БІЛЬШ НАДІЙНОГО ПРОГНОЗУ

Рак молочної залози у жінок – глобальна проблема, яка впливає на генофонд. Ця хвороба стала основною онкологічною загрозою для українських жінок, а її раннє виявлення та профілактика значно підвищують шанси на виживання, знижуючи вартість лікування. Контроль рецмдивів та їх прогнозування є життєво важливими ділянками цієї проблеми.

Ця стаття стосується даних, які дозволяють за допомогою машинного навчання виявляти рецидиви раку молочної залози у пацієнтів, які проходять терапію. Оновлений набір даних, представлений у цій статті, містить 252 випадки, з яких 206 не мали рециливів, але 46 мали їх. Цей набір даних є вдосконаленою версією відомого набору про рак молочної залози створеного в Любляні 1988 року.

Метою є підвищення надійності клінічних прогнозів рецидиву раку молочної залози за допомогою оновленого та вдосконаленого LBCD. Перелік завдань, що супроводжують досягнення цієї мети, є наступним: Оцінка рангів релевантності для атрибутів LBCD; Оцінка рівнів шуму для атрибутів, головним чином для атрибуту класу; Скорочення набору даних шляхом видалення нерелевантних і зашумлених даних; Обчислення (відновлення) пропущених значень для атрибуту класу; Порівняння продуктивності для початкового та оновленого набору даних.

Наш оновлений набір даних має менше екземплярів (252 замість 286) і менше атрибутів (шість замість десяти), окрім мого атрибут класу очищено від шуму, і його пропущені значення відновлено. У результаті продуктивність оновленого набору даних набагато краща, ніж у прототипу, особливо щодо випадків рецидиву раку. Це дозволяє клініцистам проводити більш надійну діагностику рецидиву раку молочної залози за допомогою машинного навчання та найвідоміших класифікаторів.

Використаний набір даних є корисним для розробки моделей машинного навчання, які повинні класифікувати, виявляти та прогнозувати ймовірність рецидивів раку молочної залози в клініках. Розроблений набір даних забезпечує значно вищу продуктивність алгоритмів машинного навчання, ніж початковий прототип. Порівняно з прототипом, набір даних є більш компактним: 252 екземпляри замість 286 та 6 атрибутів замість 10. Атрибут класу (категорії) цього набору даних повністю очищений від шуму.

Ключові слова: машинне навчання, набір даних про рак молочної залози, рецидиви, очищення від шуму, підвищення продуктивності.

Introduction

Breast cancer is a highly prevalent form of cancer among Ukrainian women [1]. Women are the primary carriers of the national gene pool, which makes early diagnosis of breast cancer with the help of artificial intelligence incredibly important. Machine learning and deep learning techniques are gradually integrated into evidence-based medicine, and oncology and breast cancer diagnostics are no exception [2]. These computer-assisted techniques can improve clinical decision-making and patient outcomes[2, 3].

Our attention will be focused on the oldest among several well-known oncology breast cancer datasets, the Ljubljana Breast Cancer Dataset (LBCD). This dataset has been in use since 1988 [4], and it illustrates the cases where

a node cap can occur in a female's body, which can lead to the recurrence of breast cancer. Attributes of LBCD (Meta Data) are the following:

- *Age* – Age of the patient at the time of diagnosis.
- *Menopause* – 12 months after a woman's final period.
- *Tumor size* – Tumor size represents the size of the cancer tumor at the time of diagnosis.
- *Inv-nodes*- Number of lymph nodes in the armpit that contain the spread of breast cancer visible.
- *Node caps* – Though the outside of the tumor seems to be contained, cancer may expose the risk of metastasis to the lymph node.
- *Degree of malignancy* – Grade of cancer that is visible under a microscope.
- *Breast* – Which side of the breast does breast cancer occur.
- *Breast quadrant*- Regions from the nipple area where breast cancer occurred.
- *Irradiation*: Treatment that destroys cancer cells.

The class attribute has two possible values: {no-recurrence, recurrence}. There are a total of 286 cases (instances) in the dataset. Of these, 201 belong to the first class without repeats, while 85 belong to the second class with recurrences [4]. One can see that the dataset is pretty imbalanced regarding the possible classes.

Related works

Between 2001 and 2019, the UCI machine learning repository listed 147 papers that cited the dataset [4]. This means that the dataset was referenced in approximately eight articles per year. Recently, a few dozen papers have been added to this list. As a result, it is virtually impossible to analyze each of these works individually.

However, it is noteworthy that most of these works attempted to achieve better results by improving the machine-learning algorithms while keeping the dataset unchanged. Only a few authors have taken the opposite approach, focusing on improving the dataset through optimal feature selection [2, 5, 6], denoising [7,8], or restoring missing values [9]. It is rare for authors to use a combination of these three methods.

A balanced dataset contains an equal or almost equal number of samples from the positive and negative classes. The medical datasets often are out of this rule. One can find the consequences and solutions in the review [10]. For example, AU PRC (area under the Precision-Recall Curve) is better as the integral evaluation of the performance than traditional AU ROC (area under the Receiver Operator Characteristic) in this case [11]. We are going to take these recommendations further.

The Waikato Environment for Knowledge Analysis (Weka) is a Java-based software developed at the University of Waikato, New Zealand. It is free and licensed under the GNU General Public License [12]. Its purpose is to mine data, especially from vast datasets. The latest versions of Weka contain a modern collection of various algorithms and means of machine learning with powerful visual support. Weka was used in the research displayed in this paper.

Main goal and tasks of the research

Let us formulate the primary goal of this study. The aim is a lift in the reliability of clinical prognoses of breast cancer recurrence using the updated and improved LBCD.

The list of tasks accompanying this goal is as follows:

- Estimating relevance ranks for LBCD attributes
- Evaluations of noise levels for attributes, mainly for the class attribute.
- Reduction of the dataset by removing irrelevant and noisy data.
- Imputing (restoring) the missed values for the class attribute.
- The simile of the performance for the initial and upgraded dataset.

Experimental design, datasets, and methods

The raw initial dataset (LBCD) was borrowed from [4]. This dataset has 286 instances (201 without Breast Cancer recurrence events and 85 having ones). Each instance was described by ten, including the class attributes. The code table for nine attributes of the raw dataset, excluding the class, is hosted in Table 1.

Table 1.

Coding table for nine attributes of the raw dataset

Attribute	deg–malign	irradiated	node-caps	tumor-size	inv-nodes	age	breast-quad	breast	Meno pause
Code (IDs)	1	2	3	4	5	6	7	8	9

Feature ranking and selection

As a rule, authors who work with machine learning and have performed attribute ranking exploit for this purpose one, rarer two, or three algorithms. There will be seven ranking algorithms in use. Therefore, they also need a code table, Table 2.

Table 2.

Coding table for seven algorithms (evaluators) used for attribute ranking

Evaluator	Symmetrical_ Uncert_Attribu t	Pairwise_Corr elation_Attrib	Info_Gain_Att ribut	Gain_Ratio_At trib	Classifier_Attr ibut	Correlation_At trib	Cfs_Subset
Code (IDs)	a	b	c	d	e	f	g

Thus, one can frame a rank matrix for nine attributes obtained by seven evaluating algorithms. Hence, the ranks matrix has nine columns and seven rows. The highest rank is 1, and the lowest is 9. This matrix has such a form:

$$Rm = \begin{pmatrix} 2 & 3 & 1 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1 & 3 & 2 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1 & 2 & 4 & 5 & 3 & 6 & 7 & 8 & 9 \\ 3 & 2 & 1 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1 & 2 & 4 & 3 & 5 & 6 & 7 & 8 & 9 \\ 3 & 2 & 1 & 4 & 5 & 9 & 8 & 6 & 7 \\ 4 & 2 & 3 & 5 & 1 & 6 & 7 & 8 & 9 \end{pmatrix} \quad (1)$$

Thence, the columns of the matrix (1) are labeled by codes from Table 1, while the rows are by codes from Table 2. Note that most evaluators rank the last four attributes by lower ranks (6-9) almost in unison, being less united for the first five.

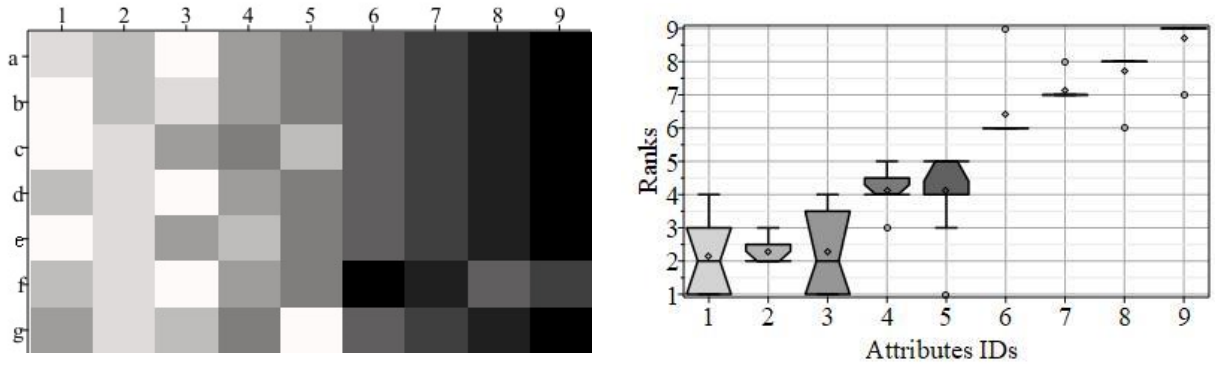

Fig 1 Heat Map of the Ranking matrice (left-hand side) and box plot (Tukey's chart) of its columns (right-hand side); the height of the boxes shows the corresponding interquartile ranges, while the horizontal segment at the boxes' "waist" is a median.

Figure 1 shows the Heat Map of the matrix (1) and the box-and-whiskers plot (Tukey's chart) for its columns. The structure of the rank's matrix, its heat map, and the box plot of its columns all testify about the presence of two subsets of attributes:

– The first five have between 1-th and 5-th ranks and manifest relatively high variability of ranks if the interquartile ranges evaluate that as in the box plot;

– The last four attributes have no visible variabilities (zero interquartile ranges), lower rank from 6 to 9, and occupy the dark side of the Heat Map.

So, the order of attributes in Table 1 corresponds to the ascending order of the rank matrix (1) columns, their medians, and the "darkening" of the heat map columns. After that, one can consider the last four attributes of Table 1 as less relevant in simile to the first five.

Noise cleaning and dataset reduction

First, we reduced the number of attributes from ten to six by removing the last four attributes with lower relevance ranks (from 6 to 9) in Table 1 from the dataset. Thus, the intermediate dataset had 286 instances and six attributes, including class.

Then, this dataset was filtered using CAIRAD (Invalid Record Analysis and Attribute Value Discovery [7]). This filter allows one to mark all questionable (incorrect) attributes as missing values. Next, all instances with three or more incorrect attributes, meaning half of them or more, were deleted. That reduced the dataset to 252 instances.

It is well-established that incorrect values in class attributes are the most harmful among all the noises present [8]. Even after applying filters, the dataset still contained 35 incorrect values, which accounted for 14% of the total values in the class attribute (see Fig.2).

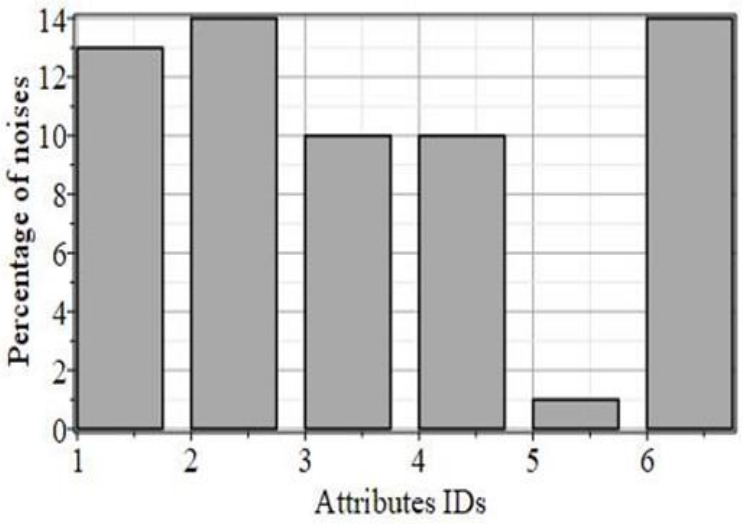

Fig. 2 Column graph for the noise levels for the first five attributes of Table 1 (1-5) and class attribute (6)

That initiated the imputation procedure (restoring correct values) for detected via filtering, denoted as missed ones. There exist many ways of such an imputation, and the algorithm [5] has been used here. The imputation was used only for the class attribute; other noises (1% to 14% dependent on the attribute) were considered missed values. As a result, the dataset was upgraded to 252 instances with six attributes each. Two hundred six of them are without recurrence events, whereas 46 have recurrence. Thus, the upgraded dataset is surely imbalanced even more than the raw one. Each instance has five nominal attributes with moderate noise levels. However, the class attribute, the sixth one, is free of noise.

The authors would like to underline two limitations of the upgraded dataset that is accessible in [13]:

First, the dataset still holds 16 instances from 252, with two missed attributes from six; perhaps these instances should also be removed.

Second, five attributes still have missing values, on a level of 1% to 14%, which may be restored.

Performance simile of datasets

To compare the performance of the two datasets, we analyzed the main performance indexes of the original [4] and upgraded [13] datasets using a version of the J48 classifier. Specifically, we used "weka.classifiers.trees.J48Consolidated -A -C 0.25 -M 2 -Q 1 -RM-C -RM-N 99.0 -RM-B -2 -RM-D 50.0".

Let us start with two confusion matrices, which serve as an origin for calculations of most performance indicators [14]. The structure of a confusion matrix is as follows :

– *True Positive (TP):* Observations are positive and are predicted to be positive. The value of TP is located in the left upper cell of the 2x2 matrix

– *False Negative (FN):* Observations are positive but are predicted to be negative. Right upper cell.

– *True Negative (TN):* Observations are negative and are predicted to be negative. Right lower cell

– *False Positive (FP):* Observations are negative but are predicted to be positive. Left lower cell.

So, the diagonal elements of the confusion matrix show the numbers of correctly classified instances for each binary class. In contrast, quantities of incorrectly classified instances are shown by non-diagonal cells. Table 3 displays both confusion matrices.

Table 3.

Confusion matrices for raw and upgraded datasets

Dataset	Raw	Upgraded
Confusion matrices	$\begin{pmatrix} 153 & 48 \\ 46 & 39 \end{pmatrix}$	$\begin{pmatrix} 190 & 16 \\ 5 & 41 \end{pmatrix}$

Note that the upgraded dataset has 252 instances instead of 286 in the raw one. Nevertheless, there is evidence that the number of correct predictions increased while the number of incorrect ones dropped. A more detailed comparison of the datasets is provided in Table 4.

Table 4

Performance indicators for raw and upgraded datasets

Datasets	Classes	Precision	Recall		F-measure	MCC	AU PRC
Raw	**no-recurrence**	0.769	0.761		0.765	0.219	0.758
	recurrence	0.448	0.459		0.453	0.219	0.434
Upgraded	**no-recurrence**	0.971	0.922		0.948	0.751	0.979
	recurrence	0.719	0.891		0.796	0.751	0.776

In evaluating prediction models, MCC (Matthew correlation coefficient) and AU PRC (area under the Precision-Recall Curve) are two critical measures. MCC considers all four elements of the confusion matrix and produces higher scores closer to 1 only if the prediction ensures reasonable rates for all four categories. In other words, MCC comprehensively evaluates the model's performance independent of a class. For the upgraded dataset, the MCC score tripled and achieved a value of 0.751, indicating a significant improvement in the model's accuracy.

Another integral but class-dependent performance indicator (AU PRC) increases, especially for the second class (cancer recurrence). Tables 3 and 4 show the higher classification performance concerning the updated data set. It means more reliable diagnostics, declared the goal in section 1.2.

Comclusions

The dataset [13] is helpful in machine learning models' devising, which shall classify, detect, and forecast probabilities of recurrence events of breast cancer in clinics.

The elaborated dataset ensures a much higher performance for machine learning algorithms than the initial prototype [4].

Compared to the prototype, the dataset is more compact, comprising 252 instances instead of 286 and 6 attributes instead of 10.

This dataset's class (category) attribute is entirely free of noise.

References

[1] World Bank. 2018. Breast Cancer in Ukraine: The Continuum of Care and Implications for Action. © World Bank, Washington, DC. License: CC BY 4.0." URI http://hdl.handle.net/10986/30144

[2] G. Chuiko, O. Dvornik, Y. Darnapuk, D. Honcharov, Y. Krainyk, O. Yaremchuk, "Attribute Selection, Outliers Impact Study and Visualization within Breast Cancer Detection." International Conference on Electronics and Information Technologies (ELIT), Lviv, Ukraine, 2023, pp. 1-5, https;//doi.org//10.1109/ELIT61488.2023.10310922

[3] M. Radak, H.Y. Lafta, and H. Fallahi, Machine learning and deep learning techniques for breast cancer diagnosis and classification: a comprehensive review of medical imaging studies. J Cancer Res Clin Oncol 149, 10473–10491 (2023). https://doi.org/10.1007/s00432-023-04956-z

[4]] M. Zwitter and M. Soklic. Breast Cancer, UCI Machine Learning Repository, 1988, https://doi.org/10.24432/C51P4M

[5] O. Goldstein, M. Kachuee, K. Kärkkäinen, and M. Sarrafzadeh,. (2019). Target-Focused Feature Selection Using a Bayesian Approach. ArXiv, abs/1909.06772. URL: https://www.semanticscholar.org/paper/Target-Focused-Feature-Selection-Using-a-Bayesian-Goldstein-Kachuee/841cf62aed427a1bdfe21333dcc82792fe7c593f

[6] T.A Mohammed, S. Alhayali, O. Bayat, and O.N. Ucan, (2018). Feature Reduction Based on Hybrid Efficient Weighted Gene Genetic Algorithms with Artificial Neural Network for Machine Learning Problems in the Big Data. Sci. Program., 2018, 2691759:1-2691759:10. URL: https://www.semanticscholar.org/paper/Feature-Reduction-Based-on-Hybrid-Efficient-Gene-in-Mohammed-Alhayali/3c4fe0598ab524c8c0f7e96691c8cc51c43a1b34

[7] M.G. Rahman, M.Z. Islam, T. Bossomaier, J. Gao. CAIRAD: A co-appearance-based analysis for incorrect records and attribute-values detection. The 2012 International Joint Conference on Neural Networks (IJCNN), Brisbane, QLD, Australia, 2012, pp. 1-10, https://doi.org/10.1109/IJCNN.2012.6252669

[8] S. Gupta, A. Gupta, Dealing with Noise Problem in Machine Learning Datasets: A Systematic Review. Procedia Computer Science,161, 2019, pp. 466-474, https://doi.org/10.1016/j.procs.2019.11.146

[9] B. Mathura Bai, Nimmala Mangathayaru, B. Padmaja Rani, An Approach to Find Missing Values in Medical Datasets, ICEMIS 15: Proceedings of The International Conference on Engineering & MIS 2015. September 2015, Article No.:70, pp. 1–7, http://dx.doi.org/10.1145/2832987.2833083

[10] A. Somasundaram and U. Srinivasulu Reddy, "Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data," Proc. of 1st International Conference on Research in Engineering, Computers and Technology (ICRECT), 2016, 28-34. URL: https://www.researchgate.net/publication/320895020_Data_Imbalance_Effects_and_Solutions_for_Classification_of_Large_and_Highly_Imbalanced_Data

[11] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," Proceedings of the 23-rd International Conference on Machine Learning, Pittsburgh, PA, 2006, 233–240, https://doi.org/10.1145/1143844.1143874

[12] R. R. Bouckaert, E. Frank,M. Hall, R. Kirkby, P. Reutemann, A. Seewald, D. Scuse. WEKA Manual for Version 3-9-5. 2020, University of Waikato. URI: https://osdn.net/projects/sfnet_Weka/downloads/documentation/3.9.x/WekaManual-3-9-5.pdf

[13] G. Chuiko, O. Yaremchuk, and D. Honcharov. Updated Ljubljana Breast Cancer Data Set: reduced and cleaned version, Mendeley Data, V2, 2023, https://doi.org/10.17632/fgs9pyfv2z.2

[14] I. Düntsch, G. Gediga, Confusion matrices and rough set data analysis, arXiv:1902.01487 [cs.LG] (or arXiv:1902.01487v1 [cs.LG] for this version), https://doi.org/10.48550/arXiv.1902.01487

Gennady Chuiko **Геннадій Чуйко**	DrSci, Professor of the Department of Computer Engineering Petro Mohyla Black Sea National University, Mykolaiv, Ukraine. e-mail: genchuiko@gmail.com ; https://orcid.org/0000-0001-5590-9404	професор кафедри комп'ютерної інженерії , Чорноморський національний університет імені Петра Могили, Миколаїв, Україна
Olga Yaremchuk **Ольга Яремчук**	PhD student of the Department of Computer Engineering Petro Mohyla Black Sea National University, Mykolaiv, Ukraine. e-mail: olga.yaremchuk.77@ukr.net; https://orcid.org/0000-0002-0891-4216	аспірант кафедри комп'ютерної інженерії , Чорноморський національний університет імені Петра Могили, Миколаїв, Україна