International Journal of Advances in Science Engineering and Technology, ISSN: 2321-9009 Volume-5, Issue-3, Jul.-2017
http://iraj.in

CHOOSING BEST MACHINE LEARNING ALGORITHM FOR BREAST CANCER PREDICTION

[1]MERARYSLAN MERALIYEV, [2]MEIRAMBEK ZHAPAROV, [3]KAMALKHAN ARTYKBAYEV

[1,2,3]Computer Science Department SDU Kazakhstan
E-mail: [1]meraryslan.meraliyev@sdu.edu.kz, [2]zhaparov.meirambek@sdu.edu.kz, [3]kamalkhan.artykbayev@is.sdu.edu.kz

Abstract- Throughout the 20th century, views about breast cancer have drastically changed. Breast cancer is the most common cancer in women worldwide, with nearly 1.7 million new cases diagnosed in 2012. This type of cancer is the second most common cancer overall. There is lot of information and data, which give opportunity for analyzing some processes, make some researches in classification and in data mining fields, test some tools of machine learning and make experiments for tuning main methods of supervised learning. Main part of project is creating useful tool for predicting breast cancer with high accuracy before getting ill or in initial stage of disease. This work is fascinating because the goal is to implement a lot of tools for creating web system, which can make effective prediction analysis. In other word, we can anticipate the future for women diseases.

Keywords- Breast Cancer, Diseases prediction, machine learning methods, scikit, Wisconsin Breast Cancer dataset.

I. INTRODUCTION

Breast cancer takes second place for the most cancer diagnoses among women, second to only skin cancer.

[1] And it's currently a widely discussed issue. If we look for this problem from clinical view, detecting early stage of breast cancer very difficult, but has some possible variants.

In USA, breast cancer is the most frequently diagnosed malignancy, and is the greatest cause of cancer death in women. About 1 in 8 U.S. women (about 12%) will develop invasive breast cancer over the course of her lifetime [2].

In 2017, an estimated 252,710 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S., along with 63,410 new cases of non-invasive (in situ) breast cancer. About 2,470 new cases of invasive breast cancer are expected to be diagnosed in men in 2017.

A man's lifetime risk of breast cancer is about 1 in 1,000. By analyzing design thinking process, we found some real problem, which related automatization of predicting process.

[3] As with most cancers, early detection of the disease can be crucial in increasing survivorship. Nowadays machine learning algorithms can be applied to this task.

These algorithms give highly accurate precise diagnoses and also decrease time of getting diagnoses. There are a lot of work, which related with applying machine learning algorithms for predicting breast cancer.
One of them - «Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis» Hiba Asri ,Hajar Mousannif, Hassan Al Moatassime, Thomas Noel. [16] In this paper discussed about applying Support Vector Machine, Decision tree, Naive Bayes and KNearest neighbors methods for predicting breast cancer for Wisconsin Breast Cancer dataset. The main objective is to assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity and specificity. Experimental results show that SVM gives the highest accuracy (97.13%) with lowest error rate. All experiments are executed within a simulation environment and conducted in WEKA data mining tool. In conclusion,

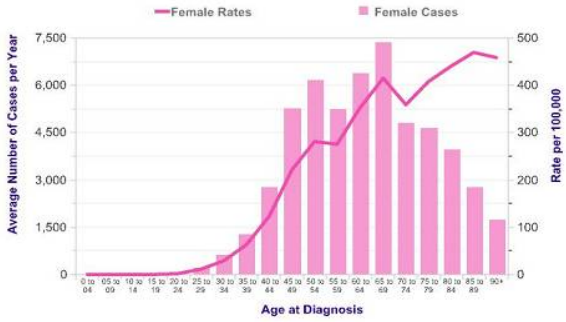

Fig. 1 Average number of cases per Year on Age of diagnosis [4]

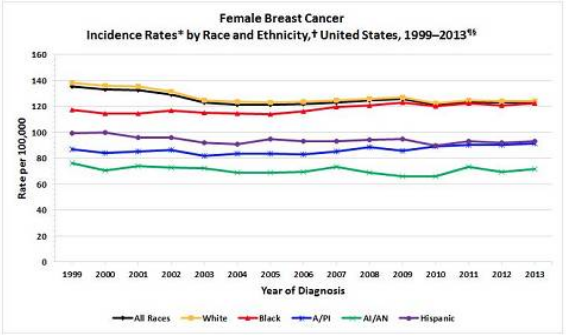

Fig. 2 Number of rates per year on age of woman by their races [5]

this paper show us that SVM has proven its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate. [2] In our work we try different machine learning algorithms to determine best ones for our dataset. We will examine 5 algorithms: K-nearest neighbors, Decision Tree Classification, Logistic Regression, SVM and Artificial Neural Network. To work with this algorithms we have chosen Wisconsin Diagnostic dataset from which the author extracts 32 features using Xcyt[6]. These features include: "area, radius, perimeter, symmetry, number and size of concavities, fractal dimension (of the boundary), compactness, smoothness (local variation of radial segments), and texture (variance of gray levels inside the boundary)".[7] We should apply some preprocessing techniques to prepare out data for our models. In next part we describe used preprocessing techniques. And in third part of this paper we explain each algorithm, how we used them and show results. The last part of this paper is conclusion, where we compare results of our research.

II. PRE-PROCESSING PART

Data pre-processing is an important part in any data analysis task. Because during the data gathering process there are many factors that may lead to irrelevant and out-of-range data, missing values, etc. Analyzing data which has not been carefully screened for such problems can produce misleading results. Thus, the quality of data and its representation is first and foremost before running any analysis. [8]

If in data there are many irrelevant information or noisy and unreliable data present, then the knowledge discovery and models creation will be more difficult. So, to solve this problem data should

be prepared and filtered. And this processes take considerable amount of processing time. Data pre-processing includes binarization, categorization, standardization, cleaning, feature extraction, etc. In our work we used only categorization technique, because we used dataset which is already good for classification algorithms. In our dataset there are two possible outputs: M which states for "Malignant" and B for "Benign". We transformed our dataset, to have only numeric values. So instead of M we took 1 and for B we took 0.

To check results of this pre-processing steps we should make correlation analysis. We will use Pearson correlation coefficient, which is the most familiar measure of dependence between two quantities. If we have a series of n measurements of X and Y written as xi and yi for i = 1, 2, ..., n, then the sample correlation coefficient can be used to estimate the population Pearson correlation r between X and Y. The sample correlation coefficient is written:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where x and y are the sample means of X and Y, and sx and sy are the sample standard deviations of X and Y. The results of this correlation analysis are shown in Figure 3.

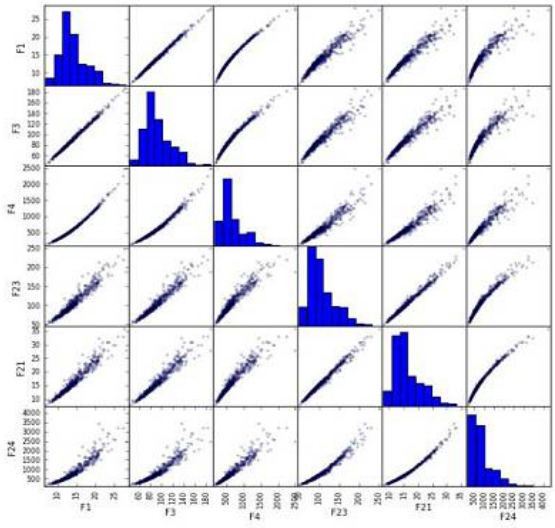

Fig. 3. Scatter plots and histograms showing correlation between features

III. MACHINE LEARNING ALGORITHMS ANALYSIS

3.1 KNN

In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression.[9] In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors. During the creation of best model of KNN algorithm we used Greedy Search to fit algorithm

with best parameters. And for KNN we tried to test number of neighbours from 1 to 41. The best results of our work on KNN algorithm are shown in Table 1.

	K-Fold 2	K-Fold 3	K-Fold 4	K-Fold 5	K-Fold 6	K-Fold 7	K-Fold 8	K-Fold 9	K-Fold 10	K-Fold 11
ACCURACY	1.94	0.95	0.96	0.97	0.95	0.99	0.97	0.98	0.98	0.98
SENSITIVITY	1.93	0.88	0.91	0.93	0.92	0.96	0.91	0.95	0.96	1
SPECIFICITY	1.94	0.99	0.97	1	0.98	1	1	1	1	0.97
PPV	1.83	0.88	0.9	1	0.98	1	1	1	1	0.92
NPV	1.98	0.93	0.97	0.96	0.92	0.98	0.95	0.98	0.97	1
F-SCORES	1.87	0.93	0.9	0.96	0.95	0.98	0.97	0.98	0.98	0.96
G-SCORES	1.87	0.93	0.9	0.96	0.95	0.98	0.97	0.98	0.98	0.96

Table 1: Results of KNN model

3.2 SVM

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The clustering algorithm which provides an improvement to the support vector machines is called support vector clustering.[10]

We tried to find best parameters for SVM model. And tried to change kernel and C parameters. For kernel we used linear and rbf types of kernel and for C parameter we tried values from 1 to 10. Results of best SVM model shown in Table 2.

	K-Fold 2	K-Fold 3	K-Fold 4	K-Fold 5	K-Fold 6	K-Fold 7	K-Fold 8	K-Fold 9	K-Fold 10	K-Fold 11
ACCURACY	0.94	0.96	0.96	0.97	0.98	0.99	0.98	1	0.98	1
SENSITIVITY	0.97	0.93	0.97	0.95	0.96	0.96	0.95	1	0.94	1
SPECIFICITY	0.93	0.98	0.96	0.97	0.99	1	0.98	1	1	1
PPV	0.81	0.97	0.89	0.95	0.96	1	0.95	1	1	1
NPV	0.99	0.96	0.99	0.97	0.99	0.98	0.98	1	0.98	1
F-SCORES	0.88	0.95	0.93	0.95	0.96	0.98	0.95	1	0.97	1
G-SCORES	0.89	0.95	0.93	0.95	0.96	0.98	0.95	1	0.97	1

Table 2: Results of SVM Model

3.3 Logistic Regression

In statistics, logistic regression, or logit regression, or logit model is a regression model where the dependent variable (DV) is categorical. This article covers the case of binary dependent variables—that is, where it can take only two values, such as pass/fail, win/lose, alive/dead or healthy/sick. Cases with more than two categories are referred to as multinomial logistic regression, or, if the multiple categories are ordered, as ordinal logistic regression.[11] In Logistic Regression algorithm we tried to change solver parameter. And as values for this parameter we used 'newton-cg ', 'lbfgs', 'liblinear' and 'sag'. Results of the best model are shown in Table 3.

	K-Fold 2	K-Fold 3	K-Fold 4	K-Fold 5	K-Fold 6	K-Fold 7	K-Fold 8	K-Fold 9	K-Fold 10	K-Fold 11
ACCURACY	0.95	0.97	0.95	0.97	0.97	0.98	0.97	0.98	0.98	0.98
SENSITIVITY	0.97	0.94	0.92	0.95	0.94	0.96	0.95	1	0.96	1
SPECIFICITY	0.94	0.98	0.97	0.97	0.98	0.98	0.98	0.98	1	0.97
PPV	0.83	0.97	0.97	0.95	0.97	0.96	0.95	0.95	1	0.93
NPV	0.99	0.97	0.94	0.97	0.97	0.95	0.98	1	0.97	1
F-SCORES	0.9	0.96	0.94	0.95	0.96	0.96	0.95	0.98	0.98	0.96
G-SCORES	0.9	0.96	0.94	0.95	0.96	0.96	0.95	0.98	0.98	0.96

Table 3: Results of Logistic Regression Model

3.4 Decision Tree Classification

Decision tree learning is a method commonly used in data mining.[12] The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown in the diagram at right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf. In Decision Tree Classification algorithm we used min_samples_split and min_samples_leaf, both of this parameters we tried to change from 1 to 5. And the results of our best test are shown in Table 4.

	K-Fold 2	K-Fold 3	K-Fold 4	K-Fold 5	K-Fold 6	K-Fold 7	K-Fold 8	K-Fold 9	K-Fold 10	K-Fold 11
ACCURACY	0.92	0.95	0.96	0.96	0.96	0.98	0.97	0.97	0.98	0.98
SENSITIVITY	0.9	0.94	0.88	0.93	0.96	0.95	1	1	1	1
SPECIFICITY	0.95	0.96	0.98	0.98	0.96	1	0.96	0.96	0.98	0.97
PPV	0.95	0.93	0.93	0.93	0.96	1	0.89	0.88	0.93	0.92
NPV	0.9	0.97	0.96	0.98	0.96	0.96	1	1	1	1
F-SCORES	0.92	0.94	0.9	0.93	0.96	0.97	0.94	0.94	0.96	0.96
G-SCORES	0.92	0.94	0.9	0.93	0.96	0.97	0.94	0.94	0.96	0.96

Table 4: Results of Decision Tree Model

3.5 ANN

Neural network models in artificial intelligence are usually referred to as artificial neural networks (ANNs); these are essentially simple mathematical models defining a function f : X → Y or a distribution over X or both X and Y, but sometimes models are also intimately associated with a particular learning algorithm or learning rule. A common use of the phrase "ANN model" is really the definition of a class of such functions (where members of the class are obtained by varying parameters, connection weights, or specifics of the architecture such as the number of neurons or their connectivity). For ANN algorithm we tried to work with activation and learning_rate parameters. For activation we used 'identity','logistic', 'tanh', 'relu' values and for learning_rate we tried 'constant', 'invscaling' and 'adaptive' values. The result of our best model shown in Table 5.

	K-Fold 2	K-Fold 3	K-Fold 4	K-Fold 5	K-Fold 6	K-Fold 7	K-Fold 8	K-Fold 9	K-Fold 10	K-Fold 11
ACCURACY	0.9	0.94	0.94	0.96	0.95	0.98	0.94	0.97	0.95	0.98
SENSITIVITY	0.91	0.86	0.78	0.9	0.88	0.96	0.9	0.85	0.89	0.92
SPECIFICITY	0.9	0.95	0.98	1	0.98	0.98	0.97	1	1	1
PPV	0.73	0.84	0.93	1	0.97	0.96	0.97	1	1	1
NPV	0.97	0.96	0.94	0.95	0.94	0.98	0.93	0.96	0.91	1
F-SCORES	0.81	0.85	0.85	0.95	0.92	0.96	0.94	0.92	0.94	0.96
G-SCORES	0.81	0.85	0.85	0.95	0.92	0.96	0.94	0.92	0.94	0.96

Table 5: Results of ANN Model

3.6 K-fold Cross-validation

Cross-validation, sometimes called rotation estimation[13], is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset).[14] The goal of cross validation is to define a dataset to "test" the model in the training phase (i.e., the validation dataset), in order to limit problems like overfitting, give an insight on how the model will generalize to

an independent dataset (i.e., an unknown dataset, for instance from a real problem), etc. In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 subsamples are used as training data. The cross-validation process is
then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged to produce a single estimation.

The advantage of this method over repeated random sub-sampling (see below) is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used,[15] but in general k remains an unfixed parameter.

3.7 Greedy Search

A greedy search algorithm is an algorithm that uses a heuristic for making locally optimal choices at each stage with the hope of finding a global optimum. In our work we used Greedy search to find best parameters for our models.

Conclusion In this work we the problem of breast cancer prediction and examined ways of its solution using
machine learning algorithms. We considered 5 modeling algorithms with Greedy Search and K-fold Cross-validation. Algorithms that were considered are Neural Networks, Decision Tree Classifier, Logistic Regression, K-nearest Neighbour and Support Vector Machines. During the process of creation this models, we also used Greedy Search algorithms to test different parameters for our models, to fit and find the best model for our dataset.

According to Table 6, the obtained results of modeling show that the algorithms SVM and KNN are the best ones for breast cancer prediction. In our future work we are going to request from Kazakhstan Ministry of Healthcare data
about breast cancer patients. And we will optimize our models for this data. Then we are going to create web based interface for clinics and patients to provide support for the medical community in clinical decision making by the web system given results.

	ANN	DTC	KNN	LOGIT	SVM
ACCURACY	0.98	0.98	0.99	0.98	1
SENSITIVITY	0.96	0.95	0.96	1	1
SPECIFICITY	0.98	1	1	0.98	1
PPV	0.96	1	1	0.95	1
NPV	0.98	0.96	0.98	1	1
F-SCORES	0.96	0.97	0.98	0.98	1
G-SCORES	0.96	0.97	0.98	0.98	1

Table 6: Best results of modeling for each algorithm

REFERENCES

[1] 1 Parker S. L. et al, “Cancer Statistics”, 1997 CA-A Cancer Journal for Clinicians, 47:5-27, 1997.
[2] 2 S.Sasikala, Dr.S.Appavu alias Balamurugan, Dr.S.Geetha, «A Novel Feature Selection Technique for Improved Survivability Diagnosis of Breast Cancer», Procedia Computer Science, pp. 16-23, 8 May 2015. [Online]. Available: http://www.sciencedirect.com
[3] 3 «U.S. Breast Cancer Statistics»,http://www.breastcancer.org, March 10,2017. [Online]. Available: http://www.breastcancer.org/symptoms/understand_bc/statistic s
[4] 4 Cancer Research UK, «Breast cancer: worldwide and UK trends». Available: www.cancerresearchuk.org/sites/default/files/cstream-node/cas
[5] es_crude_f_breast_I14.png
[6] 5 Breast Cancer Rates by Race and Ethnicity, 2013. Available: https://www.cdc.gov/cancer/ breast/images/2013-e-incidence-breast.gif
[7] 6 W. N. Street, Cancer Diagnosis and Prognosis via Linear-Programming-Based Machine Learning, Ph.D. dissertation, University of Wisconsin-Madison, 1994.
[8] 7 O.L. Mangasarian et al, “Breast cancer diagnosis and prognosis via linear programming”, Operations Research, 43(4), pages 570-577, July-August 1995.
[9] 8 Pyle, D., 1999. Data Preparation for Data Mining. Morgan Kaufmann Publishers, Los Altos, California.
[10] 9 Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175–185.

[11] 10 Ben-Hur, Asa, Horn, David, Siegelmann, Hava, and Vapnik, Vladimir; "Support vector clustering" (2001) Journal of Machine Learning Research, 2: 125–137.
[12] 11 Walker, SH; Duncan, DB (1967). "Estimation of the probability of an event as a function of several independent variables". Biometrika. 54: 167–178.
[13] 12 Rokach, Lior; Maimon, O. (2008). Data mining with decision trees: theory and applications. World Scientific Pub Co Inc.
[14] 13 McCulloch, Warren; Walter Pitts (1943). "A Logical Calculus of Ideas Immanent in Nervous Activity". Bulletin of Mathematical Biophysics. 5 (4): 115–133.
[15] 14 "Newbie question: Confused about train, validation and test data!". Retrieved 2013-11-14.
[16] 15 Werbos, P.J. (1975). Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.
[17] 16 Hiba Asri ,Hajar Mousannif, Hassan Al Moatassime, Thomas Noel, «Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis» Procedia Computer Science
[18] Volume 83 , 2016, Pages 1064-1069 [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050916 302575

★★★