3 Fainzilberg, L.S. Bayesian scheme of collective decision-making in conditions of contradictions," (in Russian) // Management and Informatics Problems. – 2002. – no. 3. – P. 112–122

4 Aidarkhanov, M.B. On the stability of the group classification algorithms // The 9th National Conference: Mathematical methods of pattern recognition, Moscow, 1991. – Moscow: «ALEV-V», 1991. – P. 3-4

5 Zhuravlev, Y.I., Biryukov, A.S. Some practical algorithms of recognition by precedents and methods of their correction // The 9th National Conference, presented at the Mathematical methods of pattern recognition, Moscow, 1991. – Moscow: «ALEV-V», 1991. – P. 190-191

6 Kuzmitsky, N.N. Topical issues of using of convolutional neural networks and their committees in pattern recognition // Vestnik of Brest State Technical University. – 2012. – no. 5. – P. 6–10

7 Rastrigin, L.A., Ehrenstein, R.K. Collective Recognition Method. – Moscow: Energoizdat, 1981. – P. 78

8 Zagoruiko, N.G. Applied methods of data and knowledge analysis. – Novosibirsk: Sobolev Institute of Mathematics of RAS, 1999. – ch. 8. – par. 6

IRSTI 50.09

M.M. Meraliyev[1], K.Ye. Orynbekova[1], D. Hasanov[1], M.K. Zhaparov[1]
[1]Suleyman Demirel University, Almaty, Kazakhstan

PARAMETERS OPTIMIZATION OF DECISION TREE AND KNN ALGORITHMS FOR BREAST CANCER PREDICTION

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.

Key words: breast cancer, diseases prediction, machine learning methods, scikit, Wisconsin Breast Cancer dataset.

***

Аңдатпа. 20-шы ғасырдың бойы, сүт безінің қатерлі ісігі туралы көріністер түбегейлі өзгерді. Сүт безінің рагы Бұл ісігінің түрі жалпы екінші ең көп таралған қатерлі ісік 2012 жылы диагнозы жуық 1,7 млн жаңа жағдайларда, бүкіл әлемде әйелдердің ең көп таралған қатерлі ісік болып табылады. Онда кейбір процестерді талдау үшін мүмкіндік береді ақпарат және деректер көп, болып табылады, жіктеу және деректер тау-кен салаларында кейбір зерттеулер жасауға, машина оқыту кейбір құралдарын тексеру және жетекшілік ететін оқыту негізгі әдістерін баптау үшін эксперименттер жасауға. Жобаның негізгі бөлігі ауырып алғанға дейін немесе аурудың бастапқы сатысында жоғары дәлдікпен сүт безі қатерлі ісігінің болжау үшін пайдалы құралы құру болып табылады. мақсаты тиімді болжау талдау жасауға болады веб жүйесін, құру үшін құралдар көп жүзеге асыру болып табылады, өйткені бұл жұмыс қызықты болып табылады. басқа сөзбен айтқанда, біз әйелдер ауруларының болашағын болжап болады.

Кілт сөздер: сүт безінің қатерлі ісігімен, ауруларды болжау, машина оқыту методтар, scikit, WisconsinBreastCancerdataset.

***

Аннотация. На протяжении 20-го века взгляды на рак молочной железы кардинально изменились. Рак молочной железы является наиболее распространенным раком среди женщин во всем мире, и в 2012 году диагностировано около 1,7 миллиона новых случаев. Этот тип рака является вторым по распространенности раком в целом. Существует много информации и данных, которые дают возможность анализировать некоторые процессы, проводить исследования в области классификации и в области интеллектуального анализа данных, тестировать некоторые инструменты машинного обучения и экспериментировать с настройкой основных методов контролируемого обучения. Основная часть проекта — это полезный инструмент для прогнозирования рака молочной железы с высокой точностью, до заболевания или на начальной стадии заболевания. Цель этой работы состоит в том, чтобы реализовать множество инструментов для создания веб-системы, которая может сделать эффективный анализ прогноза. Другими словами, мы можем предвидеть будущие болезни женщин.

Ключевые слова: рак молочной железы, предсказание заболеваний, методы машинного обучения, scikit, Wisconsin Breast Cancerdataset.

Introduction

In Kazakhstan and all over the world, people are suffering from limited medical resources and long waiting times to receive medical services. The

increasing population of Kazakhstan, the ageing population, the modern lifestyle, the climate change, and the new diseases that come into view have presented challenges for the Australian health organisations and state governments to set procedures and plans to manage and cope with the available medical resources, infrastructure, and to deliver a decent healthcare services for residents despite the shortages in medical personnel and equipment. In addition, medical services are essential for all individuals and it is the nation's responsibility to develop and sustain the medical infrastructures and services for all residents and citizens. In addition to the shortages in medical personnel and technology, incidents of prescription errors have been increasingly causing minor to major problems for patients. For example, serious health problems may occur because of Adverse Drug Effects (ADE). ADE caused by mistaken prescription, errors in dosage, miscommunication between physicians and pharmacy, dispensing and administering of drugs, and inappropriate number of drug intake [1]. For example, a study [2] shows that ADE may rank as the sixth leading cause of death in the United States after heart diseases, cancer, stroke, pulmonary diseases, and roads accidents. Those problems may be avoided by a systematic information transfer between different healthcare providers (hospitals, medical centres, pharmacies, pathologies, etc.).

Breast cancer has become a common disease around the world. Yearly, millions of women suffer from this debilitating life threatening disease, making it the second common non-skin cancer after lung cancer, and the fifth cause of death among cancer diseases in the world [3]. Discovering the disease in its early stages may reduce the breast cancer tragedy. Computing technologies and machine learning tools can be used to assist physicians in diagnosing and predicting the disease so they can provide the necessary treatment and prevent the impact, including the possibility of death. More specifically, breast cancer cause about 22.9% of all cancers in women excluding skin cancers [4]. For example, breast cancer caused 458,503 deaths worldwide in 2008 [4]. Breast cancer is targeting women 100 times more than men, although men tend to have poorer outcomes due to delays in diagnosis [5]. Survival rates for breast cancer vary greatly depending on the cancer type, stage, treatment, and geographical location of the patient. For instance, survival rates in the Western world are high. However, in developing countries survival rates are much poorer. Therefore, this work provides a hope, that this research and the related future work makes some contributions that can help in a better diagnosis of breast cancer for men and women worldwide, especially for countries with poor health services.

Preprocessing Part

Data collection phase may produce dataset that contains incomplete, inaccurate, and inconsistent data. Inaccurate data is having incorrect attribute values; this may due to data entry errors, faulty in data collection tools, errors in data transmission, and users may submit incorrect values just to fill

mandatory fields during surveying [6]. Incomplete data can occur for many reasons. For example, some attributes values were not important during data entry and some attributes values were not always available. Inconsistency occurs when there is a record that is in conflict with other records on the dataset [6]. Completeness, accuracy, and consistent data are the elements that define data quality. Data pre-processing is an important step in data mining process to satisfy data quality elements. Therefore, the current research is to utilize data pre-processing tasks to ensure the dataset is ready for mining process in order to produce accurate results as possible. The study has proposed a new approach for constructing missing feature values to satisfy the completeness element; also a comparison has been made between feature selection methods to find the best method that suites datasets, and performed some techniques to eliminate noise and outliers. At the end of the current phase, data should be ready for mining process.

Machine learning algoithms analysis KNN

The k Nearest Neighbour algorithm (k-NN) is an instance based machine learning algorithm. k-NN is very simple to understand but works amazingly well [7]. The idea behind k-NN method for classifying objects is based on the closest training cases in the feature space. The k-NN finds the k closest instances to a predefined instance and decides its class label by identifying the most frequent class label among the training data that have the minimum distance between the query instance and training instances. The distance is determined by the distance metric. Preferably, the distance metric minimise the distance between similar instances and maximise the distance between different instances. The following pseudo-code shows an Chapter 2: Background Study 22 illustration for k-NN implementation [8]. Examples of approaches to define the distance are the Euclidean and Manhattan methods.

Figure 5 shows an example of kNN.

```
procedure K-NN-Learner(TestingDataSet)
for each testing instance {
find the k most nearest instances of the training set
according to a distance metric (Euclidean distance or Manhattan
distance )
Resulting Class= most frequent class label of the k nearest instances
}
```

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.

Table 1

Results of KNN model

	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	0.95	0.96	0.97	0.95	0.99	0.97	0.98	0.98	0.98
SENSITIVITY	0.93	0.88	0.91	0.93	0.92	0.96	0.94	0.95	0.96	1
SPECIFICITY	0.94	0.99	0.97	1	0.98	1	1	1	1	0.97
PPV	0.83	0.98	0.9	1	0.98	1	1	1	1	0.92
NPV	0.98	0.93	0.97	0.96	0.92	0.98	0.95	0.98	0.97	1
F-SCORES	0.87	0.93	0.9	0.96	0.95	0.98	0.97	0.98	0.98	0.96
G-SCORES	0.87	0.93	0.9	0.96	0.95	0.98	0.97	0.98	0.98	0.96

Decision Tree Classification

Decision tree is a classification method which contains nodes, branches, and leafs. The first node on the tree or the top node is called the root node. Each node in the tree is connected with one or more nodes using branches, the last node in the tree that contains no outgoing branches is called leaf node. The leaf node indicate to termination or the outcome value [9] [10]. Figure 9 shows an example of a simple decision tree. Figure 9 shows how to solve a real time problem based on making questions and answers about attributes in the testing records. The terminology of such classification method is to keep asking question until conclusion is reached. The set of questions and answers could form a decision tree with set of nodes.

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 2.

Table 2

Results of Decision Tree Model

	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

K-fold Cross-validation

Cross-validation, sometimes called rotation estimation[11], 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).[12] 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,[13] but in general k remains an unfixed parameter.

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 2 modeling algorithms with Greedy Search and K-fold Cross-validation. Algorithms that were considered are Decision Tree Classifier and K-nearest Neighbour. 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 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.

Table 3

Best results of modeling for each algorithm

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

References:

1 Sorwar G. and S. Murugesan, Electronic medical prescription: an overview of current status and issues, in Biomedical knowledge management: infrastructures and processes for e-health systems / M. Cooper and M. Gururajan (Editors), 2010. – IGI Global Hershey, PA. – P. 61-81

2 Lazarou J., B. Pomeranz and P. Corey. Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies // Journal of the American Medical Association, 1998. – №279(15).

3 Mammography Screening Can Reduce Deaths from Breast Cancer. – 2002. – [sighted 2011 20/05/ 2011]; Available from: http: //www.iarc.fr /en /mediacentre/pr/2002/pr139.html.

4 Most Frequent Cancers in Men and Women. 2008. – [sighted 2012 20/01/2012]; Available from: http://globocan. iarc.fr/ factsheets/ populations /factsheet.asp?uno=900.

5 General Information About Male Breast Cancer. 2012 – [sighted 2012 30/12/2012]; Available from: http://www.cancer.gov/ cancertopics/pdq/ treat ment/malebreast/Patient/page1.

6 Han J. and K. M, Data Mining Concepts and Techniques. – San Franscisco: Morgan Kaufmann. – 2011. – Vol. 3.

7 Mei, Z., Q. Shen and B. Ye. Hybridized k-NN and SVM for gene expression data classification. Life Science Journal, 2009. – №6(1).

8 Parvin H., H. Alizadeh and B. B. MKNN: Modified k-Nearest Neighbor // in Proceeding of the World Congress on Engineering and Computer Science. – 2008: USA.

9 Tan P.-N., M. Steinbach and V. Kumar, Introduction to Data Mining. – 2006: Addison-Wesley.

10 Larose D., Discovering Knowledge in Data: An Introduction to Data Mining. 2005. – New Jersey: John Wiley & Sons, Inc.

11 Weisstein E.W. Euclidean Metric. 2011. – [sighted 2011 19/08/2011]; Available from: http://mathworld.wolfram.com/EuclideanMetric.html.

12 Young M., et al. Distance Metrics Overview. 2004 [sighted 2011 03/08/2011]; Available from:http://www.improvedoutcomes.com/ docs/ Web SiteDocs/Clustering/Clustering_Parameters/Distance_Metrics_Overview.htm

13 Grabusts P. The Choice of Metrics for Clustering Algorithms // Proceedings of the 8th International Scientific and Practical Conference. – Volume I1. – 2011. – P. 70-76.