Evaluation of diagnostic classifiers using artificial clinical cases

Karol Antczak[1], *Andrzej* Walczak[1,*] and *Michał* Paczkowski[1]

[1]Institute of Computer and Information Systems, Military University of Technology, Warsaw, Poland

Abstract. Evaluation of classifiers in diagnosis support systems is a non-trivial task. It can be done in a form of controlled and blinded clinical trial, which is often difficult and costly. We propose a new method for generating artificial medical cases from a knowledge base, utilizing the concept of so-called medical diamonds. Cases generated using this method have features analogous to that of double-blinded trial and, thus, can be used for measuring sensitivity and specificity of diagnostic classifiers. This is easy and low-cost method of evaluation and comparison of classifiers in diagnosis support systems. We demonstrate that this method is able to produce valuable results when used for evaluation of similarity-based classifiers as well as shallow and deep neural networks.

1 Introduction

1.1 Evaluation of diagnostic tests

Diagnostic and screening tests are important tools of medical diagnosis, especially in evidence-based approach. They serve two main purposes. The first one is to establish presence of some medical condition, while the second purpose is to determine absence of it. No diagnostic test is 100% accurate. Instead, there is always a possibility that given test can yield a wrong result. It is a significant risk factor and, as such, has to be minimized or at least identified. This is why it is so important to evaluate diagnostic tests.

Evaluation of diagnostic test should be done in a form of a *controlled trial* [1]. Such trial involves a group of people, typically divided into two subgroups: test and control. People from test group are diagnosed using diagnostic test being evaluated, while control group is diagnosed using the best currently available test serving the same purpose. This test is also called *gold standard test*. It is important to remember that, as it was already mentioned, even gold standard test is not fully accurate. Because of this, test is always evaluated in comparison to current gold standard.

Controlled trial can be randomized, meaning that patients are assigned to control and test groups randomly. Randomization of trial provides several benefits [2], the most important one being prevention of bias. Another benefit is increase of statistical power of evaluation. There are many randomization techniques, developed for maximizing specific benefits at the cost of others. The simplest one is just “coin-tossing” and assigning patients to either group with equal probability. More sophisticated approaches involve restricting allocations of patients or adaptive methods.

Another important technique for minimization of observer bias is blinding the trial. *Blinded trial* means that every patient from test and control groups does not know his assignment. If tester does not know the patient’s assignment, too, then the trial is double-blinded.

Evaluation of diagnostic test in the controlled trial is mostly performed by comparison of quality measures of test under evaluation and gold standard. There are various quality measures for diagnostic tests. Most of them are based on cardinalities of four subsets: True Positives (cases where test properly identified presence of disease), False Positives (where test incorrectly identified presence of disease), True Negatives (where test properly excluded disease) and False Negatives (where test incorrectly excluded disease). Two basic quality measures are sensitivity and specificity. They are defined as:

$$sensitivity = \frac{|TP|}{|TP|+|FN|} \tag{1}$$

$$specificity = \frac{|TN|}{|TN|+|FP|} \tag{2}$$

Sensitivity serves as an indicator how good the test is in identifying existing diseases. Specificity, on the other hand, measures ability of the test to identify patients without disease. In practice, these two measures are exclusive: increasing the sensitivity results in lower specificity and vice versa. Typically, diagnostic tests should have high sensitivity, while screening tests should be characterized by high specificity. Dependency between sensitivity and specificity can be visualized in a form of ROC curve.

1.2 Diagnosis Support Systems

[*] Corresponding author: andrzej.walczak@wat.edu.pl

Modern diagnosis relies not only on diagnostic tests and physician's knowledge, but also computer-based support systems. They can aid physician in retrieval and processing of medical data as well as support the very diagnosis process. Such systems can be viewed as a diagnostic test – given set of medical data, it can provide proposed diagnosis.

Typical *diagnosis support system* relies on two components: knowledge base and inference engine (classifier). *Knowledge base* represents information about the domain the system was created for support. In the domain of medical diagnosis, it can be information about symptoms, diseases, risk factors, etc. Knowledge can be represented in various ways. Some of the techniques used are ontologies, semantic networks, rules, frames and scripts. The input for the knowledge base can be achieved manually or semi-automatically by retrieving information from physicians, medical sources or by performing controlled trials.

Another component of diagnosis support system is inference engine, usually in a form of a *classifier*. Mathematically, classifier can be viewed as function v mapping a power set of input features X (e.g. symptoms) to a set Y of output classes (e.g. diseases):

$$f: P(X) \to Y \tag{3}$$

The parameters of function f are adjusted basing on information from knowledge base in a process of training.

The number of output classes can vary for different classifiers. For $|Y| = 2$, we deal with binomial (binary) classifier. For diagnostic classifiers, this usually means that classifier can diagnose only one disease, and provides one of two possible outputs:

- 1 - if patient have a disease
- 0 - if patient does not have a disease

Of course, there are also classifiers able to return a set of classes, i.e. $|Y| > 2$. They are known as multinomial classifiers.

Classification can be also achieved using regression models returning continuous values rather than binary ones. In such case, it is necessary to perform a quantization, for example, by choosing class with maximum value. Another way is to apply threshold function on values (decision threshold) or class ranks (rank threshold).

1.3 Evaluation of diagnostic classifiers

Diagnostic classifiers can be viewed as a specific kind of diagnostic tests, and, therefore, can be evaluated using the same methods. For binomial classifier, we can calculate its sensitivity and specificity using formulas (1) and (2). In case of multinomial classifiers, these measures are calculated per each of possible classes. There are, however, several problems with practical evaluation of diagnostic classifiers.

The first one is a need for real clinical cases for test and control groups. Organization of controlled clinical trial is a difficult task, involving many groups of stakeholders, and resources such as time, money and personnel. Moreover, clinical trials are not necessary for creation of knowledge base, and therefore, training of the classifier. As a result, many diagnosis support systems don't have properly measured quality metrics. This is a serious risk factor for the patient. What is worse, these systems are often publicly available. Safety concerns of such systems are described further in [3].

The second problem of classifier evaluation is uneven distribution of cases, leading to bias. It is important factor in case of rare diseases, but can also occur when there are specific requirements for patients. As a result, performed trial can have low statistical power.

2 Proposed approach

2.1 General assumptions

Each of problems mentioned in previous section occurs not only in a field of classifiers but is also relevant to evaluation of all kinds of medical tests. However, using knowledge base it is possible to generate clinical cases "artificially". It is relatively easy to generate artificial case by choosing random subset of symptoms. However, it is not trivial to determine whether it is a positive or negative case – this is caused by lack of "gold standard" test. To overcome this, we propose a method based on concept of "medical diamonds" introduced by Walczak and Paczkowski [4].

The general idea is very simple. Let us consider knowledge base with graph-based structure, containing two types of nodes: conditions and symptoms. Each disease is connected with one or more symptoms. Such connection means that given symptom can occur in associated disease. Symptoms are not exclusive - they can be connected with multiple diseases. Connections are also unweighted, meaning that each symptom has the same "strength". Example of such knowledge base is shown in Figure 1.

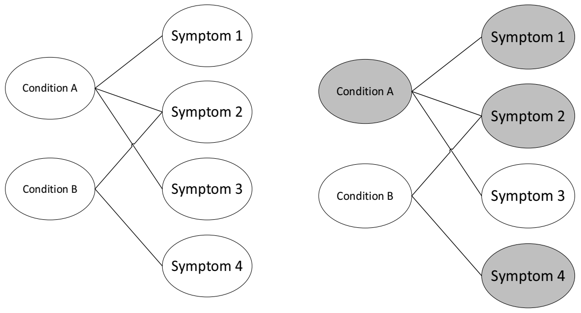

Fig. 1. Example of graph knowledge base (left) and sample case (right)

Note, that in order to ensure a proper quality of knowledge base, it should be verified by experts (diagnosticians) in order to exclude all inconsistencies and fill gaps. It should also be compatible with standards like International Classification of Diseases (ICD) 9 or 10.

2.2 Generating positive and negative cases

Now consider what a typical clinical case looks like in knowledge base of given structure. It will be a random subset of all available symptoms. In general, we are not able to certainly determine whether the patient has one disease or another. Thus, we are not able to identify Positive (P) and Negative (N) sets. However, it is possible that case will have one or more so-called medical diamonds. There are three kinds of medical diamonds, proposed by Walczak and Paczkowski [4]:

- 1st kind: symptom which identifies uniquely disease
- 2nd kind: symptom which identifies disease, but is also present in few others
- 3rd kind: group of symptoms always determining presence of disease

For knowledge base with graph-like structure, we are able to identify diamonds of the 1st kind. Presence of diamond in case determines, by definition, its disease. Hence, we are able to generate positive case for given disease D just be ensuring it will have at least one diamond of the 1st kind for D and some other symptoms not being diamonds for other diseases.

Using this method we are able to generate negative cases as well. It will be any case not having diamond (of the 1st kind) for D and having, among other symptoms, diamond for any other disease.

With positive and negative cases, we can now use them to test classifiers. Knowing which cases belong to each group, we can calculate sets TP, FP, TN and FN, and then sensitivity and specificity. This method has several advantages over clinical trial. We can generate set of any size, limited only by computational power. Additionally, set of such cases will have the features of double-blinded trial, thus reducing the bias. Thanks to this, it can be used to generate test and control sets.

3 Tests results

3.1 Test methodology

We have performed a series of experiments, measuring sensitivity and specificity of various classifiers for detecting psoriasis. We used artificial cases generated with proposed method. Three classifiers were tested: Jaccard index, shallow neural network and deep neural network. We analysed how sensitivity and specificity changes for different sizes of patient's symptoms.

3.2 Datasets

For case generation, we used knowledge base from SWD system. It is a system developed on Military University of Technology in Warsaw for diagnosis support for skin and respiratory diseases [5]. The knowledge base consists of 91 diseases, 766 symptoms and 112 possible values of symptoms. Data is stored in a form of triplets <disease, symptom, symptom value>, for example <psoriasis, lesions in the vicinity of the elbow, present>. There are 14086 such triplets in total. Data was entered by medical experts and reviewed multiple times in order to ensure its consistency. Moreover, all data is fully compliant with ICD-9 standard.

Positive set consisted of 500 cases identifying psoriasis. Each of them was generated with following algorithm:

- choose random diamond of the 1st kind associated with psoriasis
- choose remaining symptoms from other symptoms not being diamonds

Negative set consisted of 500 cases associated with scabies, generated with algorithm:

- choose random diamond of the 1st kind associated with scabies
- choose remaining symptoms from other symptoms not being diamonds

The number of symptoms in each case was a parameter N with values 1 to 300. For each value of N, sets were re-generated. Hence, for all tests, 350 000 random cases were generated. As one can see, this is much more than used in typical clinical trial.

3.3 Jaccard Index

First evaluated classifier was *Jaccard index*. It is a set similarity measure defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (3)$$

In formula 3, sets A and B are compared. As a classifier, Jaccard index chooses a disease for which set of associated symptoms is the most similar to symptoms in the case.

The measured sensitivity and specificity for this classifier are presented in Figure 2.

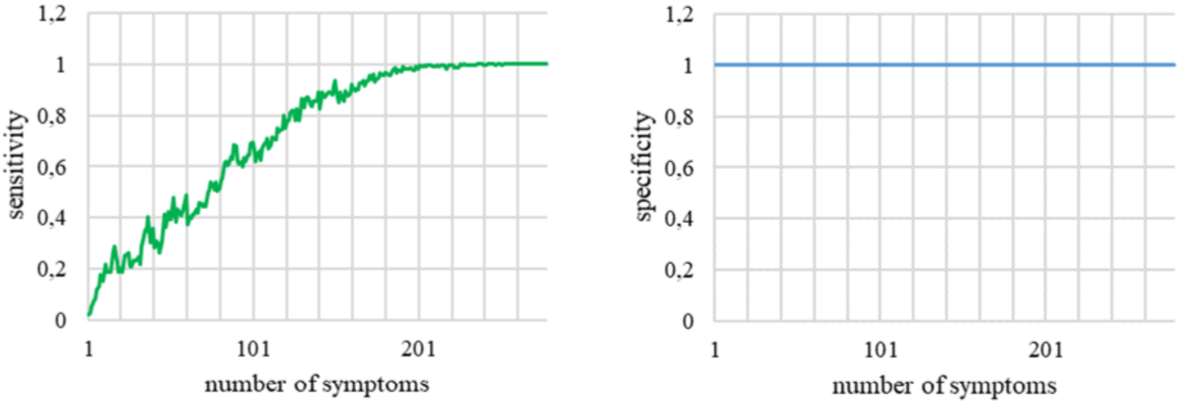

Fig. 2. Sensitivity and specificity of Jaccard Index classifier

One can observe that the sensitivity of Jaccard index increases with the number of symptoms. This is consistent with nature of this classifier. For large number of symptoms, many of them will be associated with correct disease. As a result, positive patient's symptoms will be the most similar to set of symptoms characteristic for psoriasis. This dependency is roughly monotonic, however there are several slopes. They can be identified as moments of appearance of diamonds of 2nd and 3rd kinds. They are more important than regular symptoms, but for Jaccard index, all symptoms have the same "importance". Thus, we revealed hidden feature of our dataset, which our classifier, due to its simplicity, was not able to discover.

Jaccard Index had constant specificity of 1, meaning that there were not any false positive cases. At the first sight, this can be quite surprising, given the fact that in tested model there are 275 common symptoms for psoriasis and scabies. However, there are more than 2 diseases in our knowledge base. Some of them, for example mycosis fungoides, have the same symptoms as common set for psoriasis and scabies as well as additional common symptoms associated with scabies (see figure 5). Generated case is always at least as similar (in terms of Jaccard index) to these diseases, as it is similar to psoriasis. As a result, there were no cases from negative set that were classified as the most similar to psoriasis by Jaccard index.

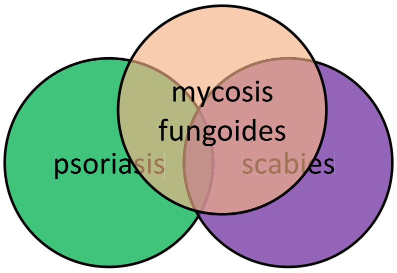

Fig. 5. Covering common symptoms of psoriasis and scabies by mycosis fungoides

3.4 Neural network

The second tested classifier was a shallow *neural network*. It consisted of 880 input neurons (all symptoms + symptoms values), 256 hidden neurons (with sigmoid activation function) and 91 output neurons (one for each disease). The network was trained with generated dataset (500 positive and 500 negative cases of various lengths) using 10000 iterations of gradient descent with learning rate 0.1. The loss function used during training was

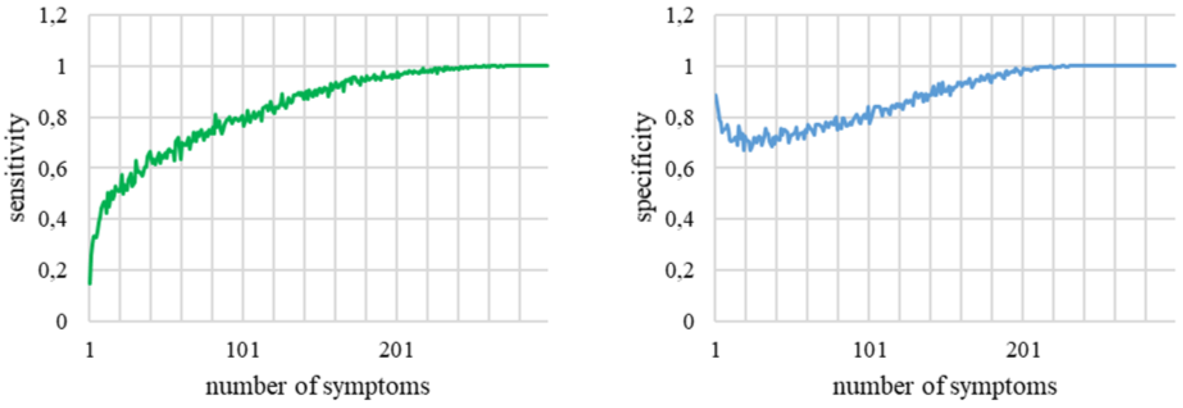

Fig. 3. Sensitivity and specificity of Neural Network classifier

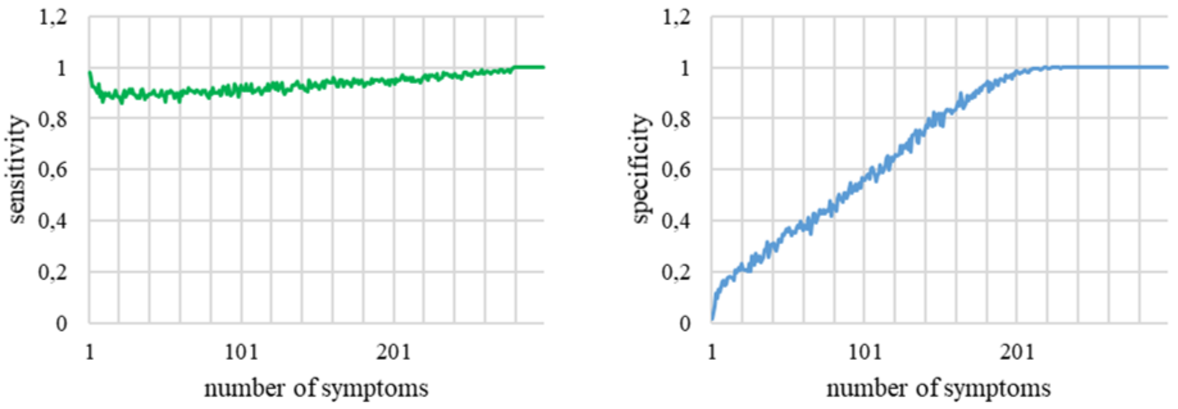

Fig. 4. Sensitivity and specificity of Deep Neural Network classifier

cross-entropy. We also added weight decay penalty (L2) to loss function in order to prevent overfitting.

Trained neural network performed better than Jaccard index in terms of sensitivity, especially for small case size. This indicates that network was able to identify “diamonds” and use them to correctly classify cases even for small number of other symptoms. Jaccard index could not utilize this property and required higher number of symptoms to perform as good. However Jaccard index had higher specificity than neural network for number of symptoms < 200. It is caused by the fact that Jaccard index is good representation of natural way of thinking of diagnosticians [3].

3.5 Deep neural network

The last model was a *deep neural network*. It represents an increasingly popular branch of machine learning, known as deep learning. Deep models attempt to extract and utilize high level features from data. We tested this classifier in order to determine whether it could extract diamonds, which can be viewed as multilevel features in medical data. Deep networks are typically multi-layered. Our deep network consisted of 3 hidden layers, 256 neurons each, with sigmoid activation function. It was trained using the same method as previous network.

Test results (Figure 4) showed that deep neural network had significantly better sensitivity than both of previous models. Additionally, sensitivity had stable values regardless of number of symptoms, however, at the cost of low specificity. Nonetheless, it seems that deep neural network was able to extract diamonds and use them even more efficiently than shallow network.

4 Discussion

4.1 Classifier evaluation

As we demonstrated, our method of measuring performance of classifiers can be successfully used for comparison of various models. Our results confirmed the prediction that the more “sophisticated” model, the more is it able to utilize hidden features of data. Additionally, we showed that there is no “silver bullet” in diagnostic classification. The higher sensitivity is achieved at the cost of lower specificity and vice versa. In this perspective it seems reasonable to use multi-classifier systems, for example in a form of classifiers ensembles [6].

4.2 Conclusions

We have presented a method for generating artificial patient cases from knowledge base. Such cases will be doubly blinded, since they will not be biased by patient (which is artificial) or classifier (which does not have a priori knowledge of the case). We have shown that it is possible to generate cases "medically plausible", meaning that they have one and only one, clearly determined, correct diagnosis. This is done by extracting hidden features of the knowledge base, in the form of medical diamonds. Because we can generate cases in any number, it is possible to use them for measuring sensitivity and specificity of classifier. Then, evaluation of classifier can be viewed as measuring its ability to uncover hidden features from dataset. Test results show that it is indeed possible to evaluate and compare various classifiers with our method.

We have not performed a meta-evaluation of proposed method. It seems reasonable to perform such meta-evaluation in a form of clinical trial, measuring sensitivity and specificity of given classifier on generated cases and comparing it to values obtained from performing classification on real cases. One can also think of performing "Turing Test"-like evaluation [7]. In this approach, artificial and real cases would be presented to expert, who should determine whether the case is a real one. This can be used as a quality measurement for knowledge base.

4.3 Other works

Generating artificial patient cases is a known task, existing nearly as long as knowledge-based diagnosis support systems. One of the first attempts was described in 1987 by Parker and Miller [7]. They used INTERNIST-1/QMR knowledge base to create simulated patient cases. The case was generated by choosing random disease, and then selecting random subset of all illnesses (other diseases, subdivisions and facets) linked to the selected disease. Next, a list of findings was generated, using knowledge of frequencies of finding given symptoms. The generated cases were then used for training of medical students. They also proposed using them for medical "Turing Test" described in previous section.

As for the newer attempts, in 2010 Buczak and others proposed a data-driven approach for generating artificial electronic medical records (EMRs) [8]. They were, similarly to us, motivated by difficulties in obtaining real-life medical data as well as privacy concerns regarding EMRs. The EMRs were generated for use in development of bio-surveillance algorithms. Other data-driven approaches included generating fictional clinical timeline [9] or entire population for modelling spreading of diseases [10].

The general problem of using synthetic data in supervised classification task was researched by Nonnemaker and Baird [12]. They concluded that using such data for classifiers training is for the most part safe and will not result in worsening of classifier's accuracy. A practical method for learning image classifiers from synthetic data was presented by Zhang and others [13]. They identified a problem of *distribution gap* between real and synthetic data. As a solution, they proposed using a deep learning model, namely Multichannel Autoencoder.

References

1. M. Rodger, T. Ramsay, D. Fergusson, Diagnostic

randomized controlled trials: the final frontier, *Trials,* **137**, 13 (2012)

2. G. D. Schulz KF, Generation of allocation sequences in randomised trials: chance, not choice, *Lancet,* **359**, 9305 (2002)
3. A. Walczak. K. Antczak, Patient Safety versus Computer Diagnosis, *MATEC Web of Conferences,* **76** (2016)
4. A. Walczak, M. Paczkowski, Medical data preprocessing for increased selectivity of diagnosis, *Bio-Algorithms and Med-Systems,* **12**, 1 (2016)
5. System Wspomagania Diagnostyki do informatycznego wspomagania badań naukowych w medycynie, *Wojskowa Akademia Techniczna*, [Online] http://www.isi.wat.edu.pl/sites/default/files/zalaczniki/swd.pdf.
6. K. Antczak, Rank thresholds in classifier ensembles in medical diagnosis, *Computer Science and Mathematical Modelling,* **3** (2016)
7. R. C. Parker, R. A. Miller, Using Causal Knowledge to Create Simulated Patient Cases: The CPCS Project as an Extension of INTERNIST-1, *Proceedings of the Annual Symposium on Computer Application in Medical Care* (1987)
8. A. Buczak, S. Babin, L. Moniz, Data-driven approach for creating synthetic electronic medical records, *BMC Medical Informatics and Decision Making,* **10**, 1, (2010)
9. D. M. Eddy, L. Schlessinger, Archimedes: An Analytical Tool for Improving the Quality and Efficiency of Health Care, *Building a Better Delivery System: A New Engineering/Health Care Partnership.*, Washington, National Academies Press (2005)

10 V. Ravichandran, D. Janes, Models of Infectious Disease Agent Study (MIDAS), National Institute of General Medical Sciences (2016). [Online] https://www.nigms.nih.gov/Research/SpecificAreas/MIDAS/Pages/default.aspx.

11 M. Kiedrowicz, T. Nowicki, R. Waszkowski, Business process data flow between automated and human tasks, *3rd International Conference on Social Science (ICSS 2016) December 9–11 2016*, pp. 471-477, (2016).

12 J. Nonnemaker, H. S. Baird, Using Synthetic Data Safely in Classification, *Proc. IS&T/SPIE Conf. on Document Recognition and Retrieval (DRR 2009)* (2009)

13 X. Zhang, Y. Fu, S. Jiang, L. Sigal, and G. Agam. Learning from synthetic data using a stacked multichannel Autoencoder, *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)* (2015)