Patient Safety versus Computer Diagnosis

Andrzej Walczak[1,a] and Karol Antczak[1]

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

Abstract. The development of the Internet technology has caused telemedicine diagnosis systems to be commonly available. The most crucial aspect of the patient's safety in such systems is a problem of safe diagnosis. There are two factors at the stake here. The first one is a human factor in the form of user; it is especially severe when the user is not a physician. The second one is the accuracy of the diagnosis process. The best way to handle possible diagnosis errors is to apply measures of sensitivity, specificity and ROC (Receiver Operating Characteristic) curve. We examine these measures on a sample diagnosis system against a real-life data. It turns out that values of these measures are strictly associated with another measure: an indication threshold. Therefore, the accuracy of diagnosis may be to a large extent determined by the chosen threshold. We propose several methods for minimizing the impact of this factor.

1 Introduction and Motivation

We are all witnesses to the software boom in telemedicine technology. A significant part of this technology is used for diagnostic data transmission. Such transmission allows proliferation of the measured data and creates wide audience for proper final diagnosis. However, there is also another aspect of telemedicine. Plenty of diseases are widely described on the Internet, which is quite often used for self-diagnosis. What is worse, there are multiple telemedicine services used for the same purpose. A study by Pew Research Center says that 35 % of American adults have used the Internet to figure out their or someone else's medical condition - they are the so-called "online diagnosers" [1]. About half of the online diagnosers talked with a clinician to confirm or deny their findings. Only 41% of them had their diagnosis confirmed.

Abuse of the self-diagnosis seems to be the source of potential danger for patients who are trying to diagnose themselves. This danger comes from two sources. The first one are the patients themselves. The activity of such online diagnosers can cause unfounded escalation of concerns about common symptomatology. The media coined a special term known as "cyberchondria" to name such condition; it is suggested that it can affect medical decisions and anxiety level of online diagnosers or their loved ones [2].

The second source of potential danger comes from the very process of self-diagnosis. One of the problems is its accuracy, or - more precisely – the issue of how to measure its reliability. A standard approach to medical diagnosis normally involves a combination of the research evidence with a physician's personal opinions and experience. Though nonetheless empirical to some degree, this process is highly dependent on subjective factors. As a result, the final decision may vary from clinician to clinician. Moreover, there is no objective measure of impact of the evidence.

With the advent of the Evidence Based Medicine (EBM), the emphasis was put on further objectivization of the decision-making process by applying a scientific method. The EBP uses tools such as statistical measures for objective evidence evaluation. It allows to rate diagnostic tests using sensitivity and specificity metrics as well as the Receiver Operating Characteristic curves.

A very important feature of medical diagnosis (both traditional and evidence-based) is that, in fact, each diagnosis is a kind of measurement - in quite clear, scientific sense. Yet, sorry to say, medicine does not apply proper measurement descriptions. What is measured are the diagnostic tests, but not the diagnosis itself. The history of medicine did not teach the tradition to describe a diagnosis with proper accuracy. A common use of the "diagnosis" is not in the form of measurement of symptoms, but rather a physician's decision. Such decision is made as a result of a single or multiple measurements of the disease symptoms using different tools or methods. The basic grounds for this decision are always similarities between the observed symptoms of the registered and gathered disease descriptions.

2 Background

Determination of diagnose accuracy is not easy. The exact definitions of sensitivity and specificity of a

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

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

diagnosis are sometimes determined for diagnostic methods, e.g. diagnostic tests. To obtain these values, the test is evaluated by comparison with the so-called "gold standard" test, which is the best test available under reasonable conditions.

The test sensitivity (also known as the "true positive ratio") is defined as a proportion of patients correctly diagnosed with a disease (TP) to all patients having the disease (TP + FP). The above may be noted down in using the following formula:

$$sensitivity = \frac{|TP|}{|TP|+|FN|} \qquad (1)$$

Specificity is defined as ratio of patients correctly ruled out by test (TN) to all patients not having disease (TN + FP). This can be expressed as:

$$specificity = \frac{|TN|}{|TN|+|FP|} \qquad (2)$$

Sensitivity is used to express the test's ability to confirm a disease, while specificity relates to its ability to rule out certain cases. Both of these measures can be combined to create the Receiver-Operating Characteristic (ROC) curves. Such curves are an excellent way to compare diagnostic tests. The British initially developed such tests for radar receiver testing during World War II. The sensitivity describes the "true positive rate" while (1-specificity) describes the "false positive rate" obtained during patients' population examination.

The above metrics are the result of the clinical examination. We suggest applying the sensitivity and specificity to describe accuracy of the diagnosis. In case of computer diagnostic systems, "nameplates" are generated, with the parameters providing descriptions analogous to those of diagnostic tests. As a consequence, the systems will be more "evidence-based" and hence, more secure for the patients.

Note, though, that the sensitivity and specificity depend on where we make the cut-point (a threshold for the assumed level of the rejected results). Let us explain the complicated links between the sensitivity, specificity and cut-point. In table 1, we have placed the values of similarities measured by means of the Jaccard coefficient:

$$d^J_{AB} = \frac{|A\cap B|}{|A\cup B|} = \frac{|A\cup B|-|A\cap B|}{|A\cup B|} \qquad (3)$$

Table 1. Jaccard coefficient between sample illnesses chosen from a medical database.

Illness code	B36	B85	C84.0	H26.8	L13.0	L20.0
B36	1,00	0,91	0,78	0,72	0,12	0,13
B85	0,91	1,00	0,85	0,77	0,14	0,14
C84.0	0,78	0,85	1,00	0,90	0,12	0,13
H26.8	0,72	0,77	0,90	1,00	0,12	0,12
L13.0	0,12	0,14	0,12	0,12	1,00	0,61
L20.0	0,13	0,14	0,13	0,12	0,61	1,00

Similarities exposed in table 1 illustrate that each patient's diagnosis should result in a set of indications containing different diseases. Obviously, the level of similarity between the disease and the patient's examination will be different for each indication. The example of such similarity may be as in fig.1, with the help of two different classifiers.

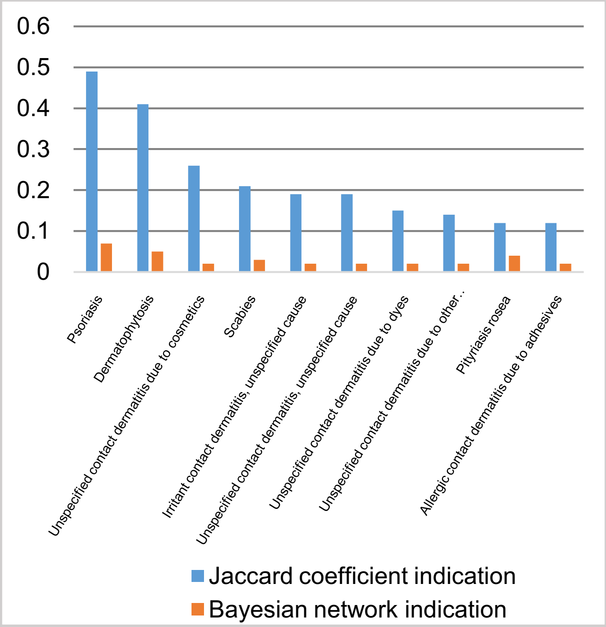

Figure 1. Similarities between the disease and the patient's examination.

A cut-point set in the classifiers' indications in figure 1 results in different choices for each classifier, whereas the threshold displacement causes different sensitivity and specificity. For the purpose of a safer diagnosis, proper values of the threshold must be used to obtain the expected sensitivity and specificity. In such sense, safer diagnosis seems to be a part of the Evidence Based Medicine. Speaking of the EBM, one of its important traits is that it involves the search for the most specific evidence - that is, the most unambiguous proof of a given disease. However, one should note that only in some cases we are able to perform the screening tests to evaluate the evidence in a form as shown in table 1. In many situations, it is necessary to depend only on a collection of clinical cases, which are not usually identical in terms of clinical research. This leads to the reduction of quality of the acquired data. Moreover, what can be considered a rule is the patient's state described by a set of several distinct proofs. It is generally impossible to assess sensitivity and specificity for each of them.

Another important factor is the approach presented in fig. 1, which contains some hidden information resulting from the inter-individual response to the cause of the disease. It means that even the diagnostic test, which confirms the existence of the evidence does not necessarily mean that a given patient is sick and vice versa: the test excluding the evidence does not entail certainty that the patient will not have a disease. Even such factor as the time difference between performing the tests, symptom incubation period and the patient's reaction strongly influences the result of the observation.

3 Tasks, methods and models

Let us examine how different cut-point values result in different quality of the classification. We have used a neural network to classify a set of 81 clinical cases from medical database of respiratory diseases. We evaluated neural network classifier by measuring its sensitivity and specificity for different cut-point values. The results are shown on fig 2. As one can see, there is a link between the sensitivity, specificity and cut-point. The higher is the cut-point value, the more increases the sensitivity, but at the expense of lower specificity. The more indications does the classifier propose, the more chance that a correct disease will be found among them.

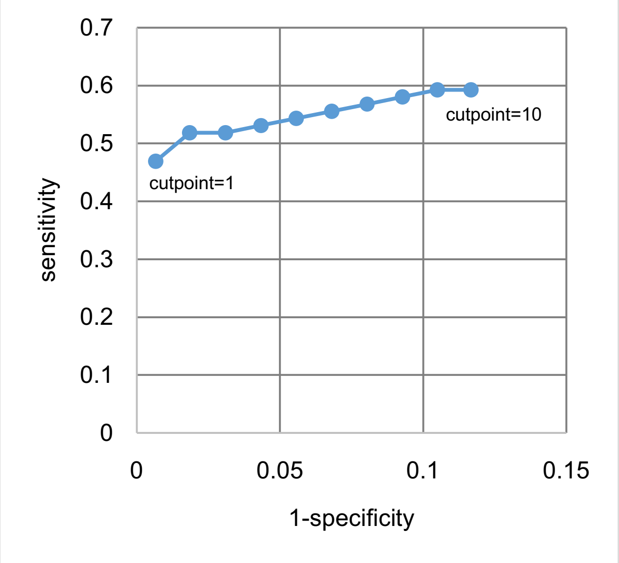

Figure 2. ROC curve for neural network classifier with various cut-point values.

How to make the classification more resistant to different cut-point values? We are unable to completely eliminate this dependency, but we may weaken it. The basic method is to increase the selectivity of indications. The selectivity can be measured as a standard deviation between the indication values. We can modify it so that one of the two indications has considerably higher values than the other. Thus, the changing of the cut-point to some extent will not change the set cut-point and the final classification will remain the same.

The selectivity varies between various classification methods. As we can observe on fig 1, the indications of the Bayesian Network have higher selectivity than the ones of the Jaccard coefficient. This is due to the fact that the Jaccard coefficient takes in consideration only amount of the symptoms, ignoring their significance. The Bayesian Network is able to extract the hidden information concerning the symptoms. In a way, this classifier works similarly to the EBM. We can improve the selectivity of the classifier by explicitly using the symptom weights during the classification process. The extraction of the hidden information to improve the selectivity may also be achieved by way of data pre-processing, as demonstrated by Walczak and Paczkowski [3]. They make the weights of the symptoms by introducing the concept of “medical diamonds”. Using such terminology, a “diamond of the first kind” is the symptom that identifies only one disease. On the other hand, the “diamond of the second kind” uniquely identifies one disease, but is also present in a few other ones. The “diamond of the third kind” defines a group of symptoms, which is always associated with a specific illness. This categorization allows weighing the symptoms and hence extracting some of their hidden properties.

Creating an ensemble of classifiers can also increase the selectivity. The ensemble of classifiers (also known as a “conference of classifiers”) is a set of classifiers connected in such a manner so as to produce single output resulting from the combined outputs from the base classifiers. Typically, this combination involves some kind of voting. As a result, the ensemble of classifiers may have better sensitivity and specificity than any of its base classifiers. The exact explanation of this feature was given by Dietterich [4]. It turns out, that ensemble can also improve the selectivity of indications. This phenomenon is based on the Pareto principle: the majority of good indications occur in the first few places. Fig. 3 shows the occurrence of this principle for the Jaccard coefficient and neural network. As shown by Antczak, in the ensembles of classifiers based on voting methods (namely Borda Count and Highest Rank), common indications of base classifiers are promoted, while the differences between them are leveled [5].

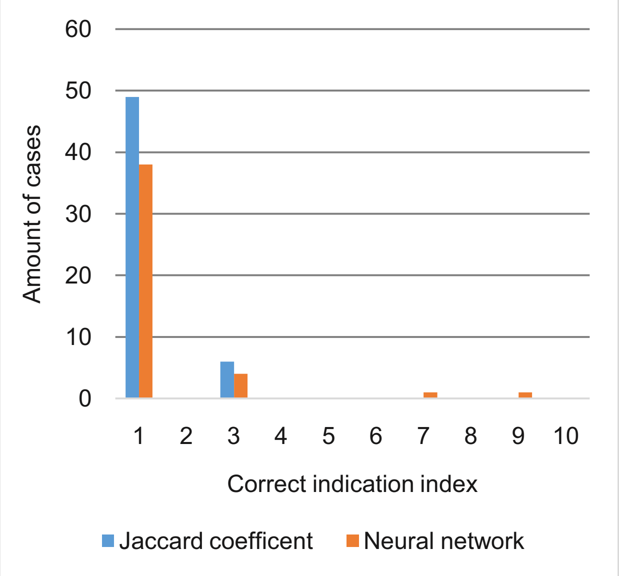

Figure 3. Amounts of correct indication indices for the Jaccard coefficient and neural network

Other methods developed in the recent years and extensively researched, are the Deep Learning techniques. It is a group of biologically inspired methods for pattern recognition and classification of complicated, high-level models. They work by "feature extraction" - in the medical diagnosis, these features can constitute the hidden information about symptoms. DL algorithms are able to extract it in a hierarchical fashion. The algorithms are useful for modeling the aforementioned medical diamonds and other "hidden" data features.

The problem of the cut-point influence can also be approached in a completely different way: instead of dealing with the selectivity, one can create a "solver" which will adjust the cut-point to ensure the sensitivity and specificity as required by a physician. This way the classifier would work as a diagnostic test, with strictly measured accuracy. The explicit formula binding the sensitivity, specificity and cut-point can be hard to obtain. It is, however, possible to use the probabilistic optimization methods (e.g. simulated annealing) to find a cut-point for the desired sensitivity and specificity.

4 Conclusions

The term "patient safety" means not only the proper protection of the patient's data, but also, what seems to be even more important, safe diagnosis. It is a vital aspect of the diagnosis as such and even more important in case of the Computer Diagnosis Support Systems. Since the diagnosis may not always be precise, the best way to control the risk is a clear determination of the properly described diagnosis errors. A part of such approach has been described above as the necessity to determine the "sensitivity-specificity-threshold" triplet. However, these measures may not be considered independently, as they are interrelated. Furthermore, a level of this association, interpreted as the selectivity of indications, may affect overall quality of the diagnosis. To minimize the impact of this factor, we proposed several methods, which mainly rely on the extraction of the hidden information about the symptoms. This approach would allow controlling the values of the sensitivity, specificity and threshold, thus, improving safety of the diagnosis.

It seems reasonable to further investigate the issue of maximizing the diagnosis selectivity. A more detailed comparative analysis of the various classification methods in terms of their selectivity can produce a lot of interesting findings. Additionally, the Deep Learning methods seem to be a promising direction of research. They already have some impressive applications, achieving human-level control [6]. It would be an interesting task to build a DL-based classifier for medical diagnosis and to test its ability as regards extracting the hidden features.

References

1. S. Fox and M. Duggan, *Health Online 2013*, (Pew Research Center 2013).
2. R. W. White and E. Horvitz, *AMIA Annu. Symp. Proc,* **2009**, pp. 696-700 (2009).
3. A. Walczak and M. Paczkowski, *BioAlgorithms and MedSystems,* **16**, 1, (2016)
4. T. G. Dietterich, "Ensemble Methods in Machine Learning," (Oregon State University 2000)
5. K. Antczak, "Rank Threshold in Classifier Ensmbles in Medical Diagnosis," *Computer Science & Mathematical Modelling,* **III**, no. 1 (2016, to be published)
6. V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature,* 518, p. 529–533, (2015)