Clinical Examples as Non-uniform Learning and Testing Sets

Piotr Augustyniak

AGH University of Science and Technology, 30 Mickiewicza Ave.
30-059 Krakow, Poland
august@agh.edu.pl

Abstract. Clinical examples are widely used as learning and testing sets for newly proposed artificial intelligence-based classifiers of signals and images in medicine. The results obtained from testing are usually taken as an estimate of the behavior of automatic recognition system in presence of unknown input in the future. This paper investigates and discusses the consequences of the non-uniform representation of the medical knowledge in such clinically-derived experimental sets. Additional challenges come from the nonlinear representation of the patient status in particular parameters' domain and from the uncertainty of the reference provided usually by human experts. The presented solution consists of representation of all available cases in multidimensional diagnostic parameters or patient status spaces. This provides the option for independent linearization of selected dimensions. The recruitment to the learning set is then based on the case-to-case distance as selection criterion. In result, the classifier may be trained and tested in a more suitable way to cope with unpredicted patterns.

1 Introduction

In medical research, the size of study cases population is an important, and usually the only considered factor of the result's reliability. However, even an intuitive approach says that two similar cases don't significantly enrich the learning or testing sets. Despite only a little influence the clinical researcher has on the available examples, each paper on clinical data-based research specifies only the cases count and neglects the features distribution therefore silently assuming it is gaussian [7] [3].

This observation and lack of justified guidelines for learning sets composition motivated us to the research on the intelligent recruitment. The presented method is proposed as an alternative for the random choice in the recruitment of cases to the learning set from all available medical examples. It is noteworthy that the human education, particularly in medicine, is also based on purposely preselected examples. Unlike the learning set, that determines the volume of competence of the AI classifier, the most natural method of recruitment test set members is the random choice.

The paper is organized as follows: In section 2 two alternative representations of medical cases are presented. The transformations between the representations

L. Rutkowski et al. (Eds.): ICAISC 2010, Part I, LNAI 6113, pp. 581–588, 2010.

and linearization of selected domains are also concerned in that section. Section 3 introduces the definition of the case-to-case distance and two methods of recruitment the cases as learning set members. Section 4 is dedicated to the description of conditions and results of tests of the basic QRS complex types recognition in the electrocardiogram with use of backpropagation neural network and both proposed learning set recruitment methods.

2 Management of Cases Representation

2.1 Parameter-Domain Representation of Cases

Detailed description of the patient on the cell level is rarely practical in the clinic for the reasons of huge amount of data and lack of the organ's physiology description. Health records usually provide several organ-specific descriptors and principal global parameters describing the whole organism in aspects representing it as organ's environment influencing its functionality. Each subject S can be described by the set of diagnostic parameters $S_p = \{p_1 \cdots p_N\}$, considered as the projection of his physiological state on the modality-dependent N-dimensional state space $\mathbf{S}^N$. The projection is limited due to restrictions on the count N of values available for measurement and inaccurate due to measurement errors ϵ_N and additive interferences δ_N [8].

2.2 Disease-Domain Representation of Cases

The description of the disease D usually involves a set of symptoms $D_s = \{s_1 \cdots s_M\}$, being characteristic patterns of M selected parameters [9]. Their coincidence is defined as a set of conditions $C_D\{M\}$, also called *disease templates* allowing the doctor to make evidence of certain pathology. Comparing to

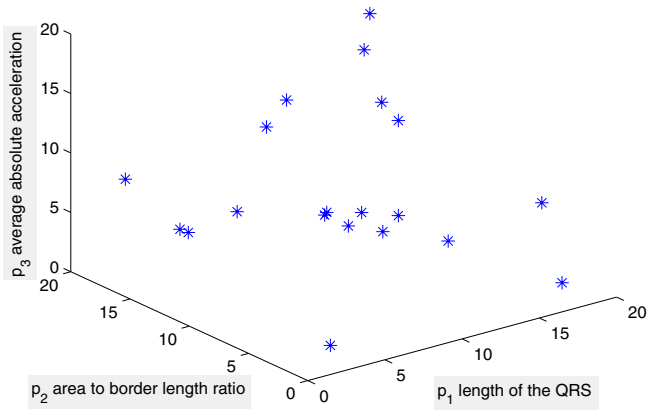

Fig. 1. Parameter-domain representation of cases. Example dimensions are: length of the QRS, area to border length ratio and average absolute acceleration.

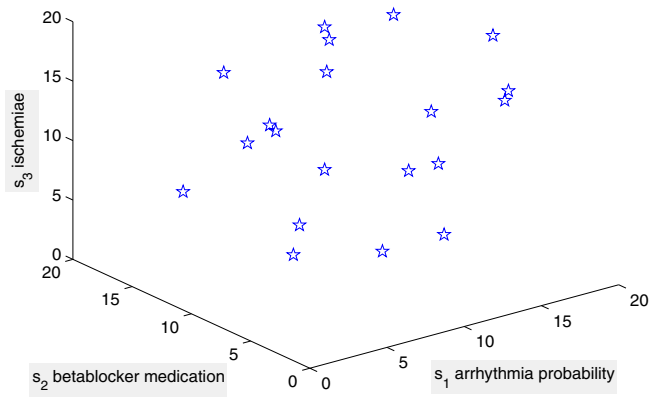

Fig. 2. Disease-domain representation of cases. Example dimensions are: arrhythmia probability, betablocker medication and ischemiae.

the *parameter-domain representation* being an initial quantitative description of the subject, the *disease-domain representation* is a result of the interpretation process and the final outcome of the diagnostics determining the subject treatment.

2.3 Transformation of Cases Representation

The medical diagnosis may be considered as matching of the parameter-domain and the disease-domain case descriptions. The subject $S \in \mathbf{S}$ is qualified to a certain category defined in the disease-domain space $\mathbf{D}^M$ and described as having a disease $D \in \mathbf{D}$ by means of his diagnostic parameters S_p best matching to recommended and fulfilling the essential criteria of the disease pattern D_s (eqn. 1). Although the patient may meet the criteria $C_D\{M\}$ for several diseases $D_1, D_2, D_3 \cdots$, only up to two, three of them, considered as most important are diagnosed and treated in medical practice.

$$p\,(D) = |\langle S_p, D_s \rangle| \quad where\ \forall m \in M\ D_s \subset C_D\,\{M\} \tag{1}$$

The patient's status available in his health record in the parameter-domain representation may be transformed to the normalized diseases space in which the probability of each disease is represented independently (2).

$$\mathbf{S}^N \rightarrow \mathbf{D}^M : 0 \leq p(D) = C_D\{M\} \leq 1 \tag{2}$$

This transformation is based on the quantitative measure of correlation between the subject's record and the *disease template*. The transformation is only partially mutually unambiguous, since - due to the data reduction - guessing the diagnostic parameters of a subject from the disease he or she has is hazardous. The transformation is performed by the human medic during the diagnosis process. If the available N parameters are not sufficient to separate two pathologies,

the representation of the subject may be completed by the complementary diagnosis yielding $N1$ new parameters and interpreted in an iterative way.

2.4 Domain Linearization of Cases Representation

Representation of the subject's state in the multidimensional parameter space assumes the independence of any two particular variables. In spite this is not always fulfilled, such representation opens the opportunity for linearization of the dimensions [1], where the nonlinearity considerably influences the distance calculations. The parameter p_n whose values have to be piecewise expanded or compressed in order to correctly represent the differences between particular diseases is transformed accordingly to (3):

$$p'_n = f\left(p_n\right) \tag{3}$$

where f is a nonlinear projection of the dimension p to p'. The projection f is a piecewise continuous function defined in all the domain of the parameter p_n in a heuristic way in order to provide equal separation of the normal and abnormal cases regardless the parameter's value.

3 Distance of the Case-Representation Space

3.1 Definition of the Case-to-Case Distance

In order to consider distance-based similarity of case representations, the notion of distance has to be defined in parameter-domain and diseases-domain spaces. Provided all dimensions of the space are linear, the following Cartesian definition of the distance is applicable (4):

$$d\left(p_N, p_K\right) = \sqrt{w_1\left(d_1^N - d_1^K\right)^2 + \cdots + w_m\left(d_m^N - d_m^K\right)^2} \tag{4}$$

where w_m is the m^{-th} weighting coefficient for each dimension in the disease state space.

3.2 Distance-Based Case Recruitment

Aiming at the generation of possible rich and representative learning set from a given set of available cases, the selection of appropriate cases may follow one of the following paradigm:

- maximum hiperspace volume, or
- equidistant hiperspace support.

First approach tends to *maximize the volume* of the competence hiperspace by selecting the cases most distant in the space as learning set members. This can be achieved by calculating first the gravity center of the cloud of all available

medical cases. First candidate is recruited as the most peripheral case. Next candidates are cases most distant from all of the previously selected, therefore the learning set consists of most atypical cases. The procedure ends after having recruited a given number of cases, or when cases exceeding the given distance are no longer available.

Second approach assumes the *equal distance* between the learning set samples. The procedure starts with the calculation of the distance between any two cases in the space. If the distance histogram is unimodal, the distance range is determined around its maximum population bin. First candidate is then recruited at random from cases belonging to the most typical distance bin. The recruitment of next candidates is based on how their distance to the already recruited case matches with the mid-range value. The procedure ends when the only remaining cases show out-of-range distance.

4 Test Conditions and Results

Both presented recruitment methods were implemented in the Matlab environment. Based on objects description consisting of up to 25 parameters each, they are capable to select from the initial population a subset complying with given distance or population criteria. Despite the method is designed to provide purposely selected learning sets for a wide class of artificial intelligence algorithms, we tested it on a three-layers backpropagation neural network [6], [10], [5],[12] applied to heartbeat types (QRS) classification in the electrocardiogram (ECG). The ECG signal originated from the MIT-BIH Arrhythmia Database [4] and was sampled with the frequency of 360 Hz.

Each QRS section was represented in normalized parameter- and disease-spaces. The versors of the parameter space were:

- length of the QRS,
- area to border length ratio,
- average absolute acceleration.

The versors of the disease space were:

- arrhythmia probability,
- betablocker medication,
- ischemiae (insufficient oxygenation, ST segment changes).

Each dimension was quantized to 20 levels, therefore the input layer contains 60 neurons. The middle layer is composed of 16 neurons, and the output layer contains 4 neurons, accordingly to the recognition of four basic QRS morphologies: normal, supraventricular, ventricular and undetermined.

From 1500 cases of heartbeats available from the database with medical annotations, 150 cases were randomly selected as the test set, and other 150 cases were recruited accordingly to the presented methods as the learning set. For the purpose of reference, the random recruitment was also used as an option.

The results classification accuracy for the parameter-space heartbeat representation are displayed in tab. 1

The results classification accuracy for the disease-space heartbeat representation are displayed in tab. 2.

Table 1. Parameter-space heartbeat representation. Percentage of correct heartbeat classification for the same test set and different recruitment method for learning set cases.

learning setrecruitment method	random	maximum volume	equidistant support
normal	93	97	98
supraventricular	63	93	95
ventricular	87	98	98
undetermined	51	78	71

Table 2. Disease-space heartbeat representation. Percentage of correct heartbeat classification for the same test set and different recruitment method for learning set cases.

learning setrecruitment method	random	maximum volume	equidistant support
normal	78	85	88
supraventricular	60	90	91
ventricular	81	85	91
undetermined	53	77	74

5 Discussion

The artificially prepared learning set (in case of both recruitment methods and independently in both representation domains) led to a considerably better result of the network learning, expressed by a better recognition result achieved in the test phase with use of the same randomly selected test set.

When the method *maximizing the volume* of competence space was used, the recognition of undetermined beats was slightly better comparing to the competitors. This was caused by the representation of more atypical beats within the learning set. The features of these beats lie far from the gravity center of the cases cloud in the parameter space.

On the contrary, the use of *equidistant hiperspace support* refines the allotment of beats into basic categories (normal/supraventricular/ventricular) at a price of few more atypical beats erroneously falling into the 'undetermined' category. Such behavior is most expected in real electrocardiogram interpreters and the proposed regularization of the non-uniform representation of the medical knowledge has been proven as the efficient method for ameliorating the learning set adequacy.

When the disease-space heartbeat representation is used, the percentage of correct heartbeat classification drops dramatically for two reasons:

– disease-domain representation is determined based on the electrocardiogram context wider than a single heartbeat,

– transformation of cases representation simplifies the cases description thus for the determination of the heartbeat types, the disease-domain is less representative than the parameter-domain.

In fact the determination of heartbeat types based on the disease-domain representation is a reversal of the typical diagnostic order. The regular interpretation first calculates the parameters, then based on parameter values determines the beat types, which in turn and in the context of neighboring results are used for assignment the disease-domain representation [11].

The presented investigation supported by the example of application to the problem of heartbeat classification, well known in electrocardiology, demonstrates that the recruitment of learning set members determines the quality of AI-based recognition systems. The selection method should consider:

– possibly diverse examples spanning the hiperspace of competence which volume is maximal,
– possibly regular distribution of the samples of medical knowledge along the particular dimensions of that hiperspace.

Acknowledgment

Scientific work supported by the Polish State Committee for Scientific Research resources in years 2009-2012 as a research project No. N N518 426736.

References

1. Aldroubi, A., Feichtinger, H.: Exact Iterative Reconstruction Algorithm for Multivariate Irregularly Sampled Functions in Spline-like Spaces: the L^p Theory. Proc. Amer. Math. Soc. 126(9), 2677–2686 (1998)
2. Augustyniak, P.: Automatic Understanding of ECG Signal. In: Kopotek, A., Wierzchon, S.T., Trojanowski, K. (eds.) Intelligent Information Processing and Web Mining, pp. 591–597. Springer, Heidelberg (2005)
3. Haussler, D.: Quantifying Inductive Bias: AI Learning Algorithms and Valiant's Learning Framework. Artificial Intelligence 36, 177–221 (1988)
4. Moody, G.B., Mark, R.G.: The MIT-BIH Arrhythmia Database on CD-ROM and Software for Use with it. In: Computers in Cardiology 1990, pp. 185–188 (1990)
5. Osowski, S.: Neural Networks for Information Processing. WUT Publishing House, Warsaw (2000) (in Polish)
6. Rutkowski, L., Tadeusiewicz, R. (eds.): Neural Networks and Soft Computing. Polish Neural Network Society (2000)
7. Stanisz, A.: Accessible Course of the Statistics with STATISTICA PL and Examples from Medicine. StatSoft Poland, Krakow (2006) (in Polish)
8. Straszecka, E., Straszecka, J.: Distance Based Classifiers and their Use to Analysis of Data Concerned Acute Coronary Syndromes. Image Processing & Communications 9(3-4), 53–69 (2003)
9. Straszecka, E., Straszecka, J.: Interpretation of Medical Symptoms Using Fuzzy Focal Element. In: Kurzynski, M., et al. (eds.) Computer Recognition Systems. Springer, Heidelberg (2005)

10. Tadeusiewicz, R.: Neural Networks. RM Academic Publishing House, Warsaw (1993) (in Polish)
11. Tadeusiewicz, R., Augustyniak, P.: Information Flow and Data Reduction in the ECG Interpretation Process. In: IEEE 27 Annual EMBS Conf., paper 88 (2005)
12. Tadeusiewicz, R., Ogiela, L.: Selected cognitive categorization systems. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2008. LNCS (LNAI), vol. 5097, pp. 1127–1136. Springer, Heidelberg (2008)