ICT for Health Science Research
A. Shabo (Shvo) et al. (Eds.)

doi:10.3233/978-1-61499-959-1-51

On the Trustworthiness of Soft Computing in Medicine

Dominik WOLFF [a, 1], Michael MARSCHOLLEK [a] and Thomas KUPKA [a]
[a] *Peter L. Reichertz Institute for Medical Informatics, University of Braunschweig - Institute of Technology and Hannover Medical School, Hannover Medical School, Hannover, Germany*

1. Introduction

By means of the huge developments in the field of applied artificial intelligence (AI) in recent years more and more typically human performed tasks are automated. Automated diagnostic tools like ECG diagnosis or the automated analysis of medical imagery are finding their way into hospitals' everyday life. This brings up the question of how dependable the results obtained by artificial intelligence are. Knowledge-driven symbolic AI uses an explicit representation of the expert's knowledge. If implemented precisely and validated against physicians, the resulting system could become as trustworthy as medical experts. In contrast, data-driven soft computing AI investigates given data, but normally behaves like a black box. This paper presents an analysis of the factors of the trustworthiness of soft computing models.

2. Methods

Soft Computing approaches can be regarded as empirical models due to their methodology. In the field of statistics and empiricism three main quality criteria exist: *objectivity*, *reliability* and *validity*. *Objectivity* describes the results' independence from boundary conditions. Therefore, a high level of intersubjective comparability must be available. *Reliability* is a measure for stability and is normally measured by reproducibility. A completely *reliable* measure exhibits no random error. The *validity* makes a statement about the resilience of the results and separates in *internal* and *external validity*. A high *internal validity* predicates, that observed changes in the dependent variable, for example a disease, are actually caused by the independent variables, for example risk factors, and not by a systematic error. This implies that a causal connection exists. On the other hand *external validity* describes the transmissibility of results from the small study cohort to the whole population. The *objectivity* is a necessary, but nor sufficient condition for *reliability*. The same applies for *reliability* and *validity*.

[1] Corresponding Author: Dominik Wolff, Peter L. Reichertz Institute for Medical Informatics, University of Braunschweig - Institute of Technology and Hannover Medical School, Hannover Medical School, Hannover, Germany, E-mail: Dominik.Wolff@PLRI.de

3. Results

Regarding soft computing methodology, *objectivity* as by definition is more a property of the data collection than of the model itself. For example labels have to be universally valid by general consensus. The models themselves should always be regarded as *reliable*, because they are deterministic systems. The training data can be not *reliable,* which applies if one input exhibits multiple different outputs. A high *internal validity* could be observed in a decreasing validation error, which determines the model's generalization capability. For *external validity* the validation dataset has to be a representative random sample. The representativeness of a random sample is mostly defined by the sample size. Using the central limit theorem from statistics formula (1) could be used to determine the needed sample size n

$$n \geq z_{1-\frac{\alpha}{2}}^2 \frac{S^2}{e^2} \tag{1}$$

with the $1-\frac{\alpha}{2}$ quartile of the normal distribution $N(\mu,\sigma)$ $z_{1-\frac{\alpha}{2}}$, standard deviation S and with the difference from the real value e. A common preprocessing technique in soft computing is to standardize the collected data, resulting in $\bar{X} = 0$ and $S = 1$. Table 1 shows minimum sample sizes for typical α and e values.

Table 1. Minimum sample sizes for typical α and e values rounded to the second decimal digit.

α \ e	0.05	0.01	0.005	0.001
0.05	1536.64	38416.00	153664.00	3841600.00
0.01	2653.90	66347.46	265389.80	6634746.00
0.005	3151.70	78792.49	315170.00	7879249.00
0.001	4330.96	108273.90	433095.60	10827390.00

4. Discussion

Objectivity and *reliability* must be ensured during the data collection process. Main focus should be on the data's unambiguity. The model itself should be regarded as *reliable*. The most important criterion, *validity*, is a property of the model itself. A high *internal validity* could be observed in a decreasing validation error. The standard measures F-score and mean squared error are good measurements for the *internal validity*, but most papers lack a meaningful threshold, leaving the reader just with a feeling about *internal validity* of the model. A threshold is problem dependent and not universally definable. For the *external validity* a suitable method of measurement was found. Table 1 can be used by other scientists to simply determine the validation set size needed. Many current research papers lack the validation set size or are smaller than the critical end validation set size of 1537.

5. Conclusion

The three trustworthiness main criteria from statistics were successfully transferred to soft computing, where the validity is most important. The results should be taken into account when designing AI. Current research often misses the external validity.