SciForum
Mol2Net

A Proposal about Normalization of Experimental Designs in Computational Intelligence

Carlos Fernandez-Lozano[1,*], Julián Dorado[1], Marcos Gestal [1]

[1] RNASA-IMEDIR Group, Faculty of Computer Science, University of A Coruna, 15071 A Coruña, Spain, Emails: julian@udc.es, mgestal@udc.es

* Corresponding author: Carlos Fernandez-Lozano, Information and Communication Technologies Department, Faculty of Computer Science, University of A Coruna, 15071 A Coruña, Spain; E-Mail: carlos.fernandez@udc.es; Tel.: +34-881-01-1302; Fax: +34-981-167-160.

Published: 4 December 2015

Abstract: Experimental analysis starts with very similar premises: given a specific problem, we need to either collect or generate a dataset and to choose the best model according to the performance. A set of techniques can be evaluated (i.e. statistical or metaheuristic approaches) as well as results from previous works that should be taken into account. Thus, it is necessary to analyse the behaviour of a method with respect to the others in equality of conditions. Therefore it is necessary to formalize an experimental design to solve as effectively as possible the problem with different approaches and to estimate the error rate; so different results from different methods can be compared. In this work we propose four phases for any experimental design: data extraction, data pre-processing, model learning and the selection of the best model. These generic phases encapsulate the main operations and steps that should be performed during an experimental analysis (some of them mandatory and other optional), independently of the kind of data or method used and are not mandatory and can be adapted to a new specific domain. The proposed experimental design has proven to be a vital contribution to compare different techniques under the same conditions in different scopes.

Keywords: Experimental Design; Statistical Analysis; Computational Intelligence

1. Introduction

Experimental Design in Computational Intelligence is one of the most important aspects on every research so it is crucial to correctly define all the steps that should be address to ensure that we achieve good results. A correct experimental design should also ensure that the results are reproducible for other researchers and that are comparable among different techniques or methods over the same dataset.

This work proposes a generic framework about Normalization of the Experimental Design to address these concerns. Of course, the framework is not a fixed workflow of different phases as it can be adapted to different fields, each of them with its particularities.

Our proposal encapsulates the operations or steps that any researcher should follow to get reproducible and comparable results on their investigations with state-of-the-art approaches or other researcher's results.

2. Materials and methods

This paper normalizes and formalizes experimental design in computational intelligence and proposes and defines four phases: extraction of data, pre-processing of data, learning and selection of the best model, see Figure 1. The following paragraphs describes more in depth each of them:

Data Extraction

Firstly, we should generate the dataset defining its particular characteristics. The definition must contain the variables involved in the study and a brief description of each of them to ensure the reproducibility of the tests by external researchers.

In order to ensure that the data is enough representative of the particular studied problem, the help of experts is needed to define the cases (i.e. regions of interests for medical imaging or case-control patients).

Data Pre-Processing

After the generation of the dataset, data is in a *raw* or *pure* state. Raw data is often difficult to analyze, so it usually requires a preliminary study or pre-processing stage. This study will check that there is no data with incomplete information, outliers or noise. In case that some of the aforementioned appears in the dataset, different approaches should be applied to avoid them. Only once this process finished, it is considered that data is ready to begin the analysis itself. To check the importance of this step, it is often said that 80% of the effort of a data analysis, is spent compiling data correctly for analysis [1].

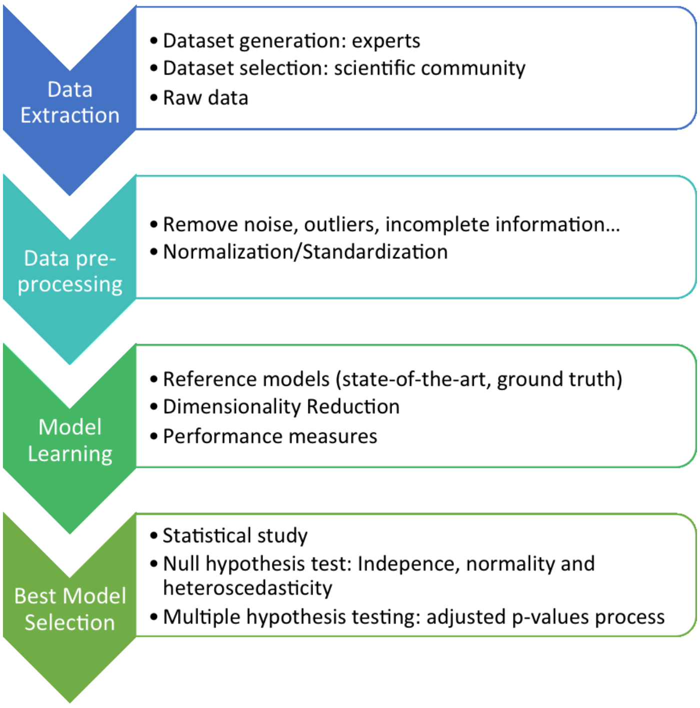
Furthermore, the variables typically present different scales or sizes, making them difficult to compare in equality of conditions. Thus normalization or standardization techniques are required to made data comparable. Of course, both techniques have their drawbacks and no one in better than the other. Furthermore, it is necessary to study the dataset for each particular problem before applying them. For example, if we try to apply a normalization step and there are outliers in the data (not removed previously), this step will scale useful data to a small interval. This is a non-desirable behavior. After a normalization step, data is scaled in the range [0,1] in case of numeric values. In case we performed a standardization process, data presents an average equal to zero and a standard deviation equal to one so they are independent of the unit of measure. Of course, depending of the kind of data, there are other well-known approaches for minimize the influence of the values.

Model Learning

Maybe the most important step within the process in computational intelligence. First of all a reference model is needed to check the results achieved for a proposed model or technique. This reference model can be extracted from a bibliography study of the field (state-of-the-art model) or constructed from a set of standard data (*gold standards* or *ground truth*) for example. In both cases the results from this reference model will be the ground truth along the following experimental design.

Once the reference model is established, it is time to build and test the model that is intended to develop in order to provide better solutions. The range of techniques available to solve any problem is usually very high. Some of these techniques are dependent on the field of study, so the researcher should review the state-of-the-art in its research field in order to choose the most suitable for his interests.

Some key points arise at this time such as the



- Dataset generation: experts
- Dataset selection: scientific community
- Raw data

- Remove noise, outliers, incomplete information...
- Normalization/Standardization

- Reference models (state-of-the-art, ground truth)
- Dimensionality Reduction
- Performance measures

- Statistical study
- Null hypothesis test: Indepence, normality and heteroscedasticity
- Multiple hypothesis testing: adjusted p-values process

Figure 1. Phases of the proposed Experimental Design

need for some measure of performance that clearly indicates how well the techniques have done this training phase. There are different well-known performance measures such as AUROC, accuracy or F-measure in classification problems or MSE, RMSE o R2 in regression problems. Sometimes it is necessary to evaluate the performance of the model using an ad-hoc measure.

Furthermore, it is desirable to avoid the overtraining of the techniques to the dataset to ensure that the model offers good results with unknown data. Techniques like *cross-validation* [2] can be useful for this point.

Finally the dimensionality of data should be taken into account. The bigger the dimensionality of the input data, the higher the number of examples necessaries for learning. Moreover, techniques for dimensionality reduction are usually interesting [3] for providing the best possible model [4] with the lower dimensionality. Thus, these techniques allow for a complexity reduction of the generated model. Furthermore, it also implies a reduction of the time and improves the overall capacity of the system.

Best Model Selection

In the previous phase, we state that there are several different measures accepted and well known as a good measure of performance for a classifier. This does not means that a researcher is able to compare different classifiers used in the same conditions and with the same dataset with just one run and this measure. At this point, it is needed to run several times each technique in order to ensure that our results are not biased because of the data. With these results per technique and in order to determine whether or not the performance of a particular technique is statistically better than the others, a null hypothesis test is needed. Furthermore, in order to use a parametric or a non-parametric test some required conditions must be checked: independence, normality and heteroscedasticity [5]. Note that these assumptions are not referring to the dataset used as input to the techniques but to the distribution of the performance of the techniques.

As part of a good experimental design for techniques comparisons, it is necessary to apply the proper test, according to the shape of the performance measure distribution. Most of the computational intelligence comparisons in the literature just apply a t-test between the performance measures to check if a technique is significantly better than the others. In some cases, this distribution does not fill the requirements of this parametric test, so a non-parametric test is required. Although the parametric test is perfectly fine to use a non-parametric test when the non-parametric test when the distribution does not fulfil the independency, normality and homoscedasticity assumptions.

Finally, after the null hypothesis test (parametric or non-parametric) is rejected, a post hoc procedure had to be used in order to address the multiple hypothesis testing and to correct the *p-values* with and adjusted *p-values* process (APV).

3. Results and Discussion

Normalization of experimental designs in computational intelligence is demonstrated using x datasets from different scientific fields. We validate this new methodology in cheminformatics and QSAR modeling with three different works. Several different Machine

Learning approaches were tested for finding the first classification model to predict cell death-related proteins [6]. In drug development it is of increased importance to find new molecular targets involved in specific diseases. Therefore, using protein star graphs for the peptide sequence information we find that the final model, reducing from 42 to 11 descriptors the original dataset [7] achieved the better results. Finally, we find for a more accurate ways of predicting residues for complex binding that can be used to model protein structure, dynamics and function [8]. We applied our experimental design as well in other fields such as bioinformatics, for example in image texture analysis problems for classification in a biomedical image texture dataset [9]. Aforementioned work used the four phases of the normalized experimental design, applying different feature selection approaches [3] for dimensionality reduction. Our results show that for all the generated datasets, our methodology reports results that are reproducible, comparable and achieved in equality of conditions. Thus, we are able to state, in each case, that we found the best model for each particular problem.

4. Conclusions

Normalization of experimental design in Computational Intelligence, as well as in other research fields is crucial. In this short communication paper we state that it is crucial to ensure that research is: reproducible, comparable and that our conclusions are based on results achieved in equality of conditions. Furthermore, for the very beginning of a research, authors should be involved in all the process that starts with the generation of the dataset, pre-processing of the data, dimensionality reduction and finally, statistical analysis. We proposed a general framework that could be used and adapted for different scenarios. Four phases could be adapted (crucial phases are mandatory but some steps are optional) for different research fields.

Acknowledgments

The authors acknowledge the support provided by the Galician Network of Drugs R+D REGID (Xunta de Galicia R2014/025) and by the "Collaborative Project on Medical Informatics (CIMED)" PI13/00280 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013-2016 and the European Regional Development Fund. This work was partially supported by the Galician Network for Colorectal Cancer Research (Red Gallega de Cáncer Colorrectal - REGICC, Ref.: CN 2014/039). This work is supported by the General Directorate of Culture, Education and University Management of Xunta de Galicia (Ref. GRC2014/049) and the European Regional Development Fund.

Conflicts of Interest

The authors declare no conflict of interest.

References and Notes

1. Tamraparni, D.; Theodore, J. *Exploratory data mining and data cleaning*. John Wiley & Sons, Inc.: 2003; p 203.
2. McLachlan, G.J.; Do, K.-A.; Ambroise, C. *Analyzing microarray gene expression data*. Wiley: 2004.
3. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507-2517.
4. Donoho, D.L. In *High-dimensional data analysis: The curses and blessings of dimensionality*, AMS Conference on Math Challenges of the 21st Century, 2000; pp 1-33.
5. García, S.; Fernández, A.; Luengo, J.; Herrera, F. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences* **2010**, *180*, 2044-2064.
6. Fernandez-Lozano, C.; Gestal, M.; González-Díaz, H.; Dorado, J.; Pazos, A.; Munteanu, C.R. Markov mean properties for cell death-related protein classification. *Journal of theoretical biology* **2014**, *349*, 12-21.
7. Fernandez-Lozano, C.; Cuiñas, R.F.; Seoane, J.A.; Fernández-Blanco, E.; Dorado, J.; Munteanu, C.R. Classification of signaling proteins based on molecular star graph descriptors using machine learning models. *Journal of theoretical biology* **2015**, *384*, 50-58.
8. Munteanu, C.R.; Pimenta, A.C.; Fernandez-Lozano, C.; Melo, A.; Cordeiro, M.N.D.S.; Moreira, I.S. Solvent accessible surface area-based hot-spot detection methods for protein–protein and protein–nucleic acid interfaces. *Journal of Chemical Information and Modeling* **2015**, *55*, 1077-1086.
9. Fernandez-Lozano, C.; Seoane, J.; Gestal, M.; Gaunt, T.; Dorado, J.; Campbell, C. Texture classification using feature selection and kernel-based techniques. *Soft Computing* **2015**, 1-12.