32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

Automatic Decision using Dirty Databases: Application to Prostate Cancer Diagnosis.

Oscar R. Marin, Daniel Ruiz, Antonio Soriano and Francisco J. Delgado.

Abstract— **Currently, the best way to reduce the mortality of cancer is to detect and treat it in its early stages. Automatic decision support systems, such as automatic diagnosis systems, are very helpful in this task but their performance is constrained by the integrity of the clinical input data. This could be a problem since clinical databases, in which these systems are based on, are commonly built up containing dirty data (empty fields, non-standard or normalized values, etc). This article presents a study of the performance of a clinical decision support system, based on an artificial neural networks, using sets of clean and dirty prostate cancer data. The study shows that is possible to obtain an implementation that allow us to avoid the problems associated to the database's lack of integrity and reach a similar performance using either clean or dirty data.**

I. INTRODUCTION

CANCER is a major public health concern in the developed countries. A total of 1,479,350 new cancer cases and 562,340 deaths from cancer were expected to occur in the United States in 2009 [1]. From those, approximately 192,000 men were diagnosed with prostate cancer, and 27,000 men were expected to die from this disease what makes prostate cancer the second most common cause of cancer death among men aged 80 years and older [2].

As it is the case of many other kinds of cancer, early detection of prostate cancer symptoms is the best way to treat the disease at its first stages reducing the morbidity and mortality [3]. The survival rate of prostate cancer soars from 34% when the cancer is detected at the advanced stage to nearly 100% at the early stage [4].

In prostate cancer case early detection is mainly based on a single biomarker, a protein called prostate-specific antigen (PSA) [2]. The PSA level in blood is generally low in healthy people, but it tends to rise in many patients with malignancies. Despite being the main marker, the specificity of using the concentration of PSA as an indicator of prostate cancer ranges from only 18% to 50% with a sensitivity of 70–90% [4]. Thus, the diagnosis process should be based primarily on PSA, but should take into account multiple factors including Digital Rectal Examination (DRE) results, free and total PSA, patient age, PSA velocity, PSA density, family history, ethnicity etc.

A clinical decision support system can be useful to help specialists in the difficult task of diagnosis [5]. A second expert opinion, even if it is from an artificial entity or software acting as a human expert, can support the decision of the doctor. In other cases, when the decision made by the artificial entity does not agree with the doctor's opinion, the clinical decision support system can suggest alternative tests to increase the degree of certainty in a specific diagnosis. In the medical problem we are working with, the prostate cancer, a clinical decision support system can help the specialist to improve the certainty in the diagnosis and, for example, to avoid useless biopsies.

One of the problems to use a clinical decision support system in the regular clinical activity is its lack of flexibility using incomplete or erroneous sets of data. In this paper we will analyze how can affect this kind of data to an automated decision system, taking as example the automated decision in prostate cancer diagnosis.

II. ARTIFICIAL NEURAL NETWORKS AND AUTOMATIC DIAGNOSIS

From its rising, Machine Learning theory has had as a main goal its applying, with several purposes, to health and clinical field [5]. This paradigm describes algorithms to solve automatic learning and classification problems which are the basis to implement systems for clinical decision support (CDSS) or computer-aided diagnosis (CAD). CDSS and CAD systems are nowadays common usage techniques in healthcare programs, including cancer screening and diagnosis [6] [7].

A. *ANNs as Classifiers.*

Within Machine Learning, ANNS are not the only, but an extensively used tool to perform automatic classification tasks [8]. In general, ANNs are able to model complex biological systems by revealing relationships among the input data that cannot always be recognized by conventional analyses [9].

To have a set of examples representing previous

Manuscript received April 1, 2010. This work was supported by the Ministerio de Ciencia e Innovacion of the Spanish government under grant TIN2009-10885 and co-financed by FEDER.

O. R. Marin is with the Bioinspired Engineering and Health Computing Research Group, University of Alicante, Alicante, P.O. 99 E-03080 Spain (e-mail: omarin@ibisrg.com).

D. Ruiz is with the Bioinspired Engineering and Health Computing Research Group, University of Alicante, Alicante, P.O. 99 E-03080 Spain (e-mail: druiz@ibisrg.com).

A. Soriano is with the Bioinspired Engineering and Health Computing Research Group, University of Alicante, Alicante, P.O. 99 E-03080 Spain (e-mail: soriano@ibisrg.com).

F. J. Delgado is with the University General Hospital of Elche, Elche, Spain. (e-mail: frandelgol@hotmail.com)

978-1-4244-4124-2/10/$25.00 ©2010 IEEE

experience is essential to construct an ANN-based classifier that assures good learning and generalization rates. These examples would be the inputs to the designed ANN. After applying to these values a collection of mathematical functions in different stages, an output is obtained. The output value has a useful meaning to classify the input or express the probability of being a member of a target class.

A classification of artificial neural networks divides them in two types, supervised and unsupervised neural networks. The former implies a supervision of the training where it is specified explicitly what output corresponds to an input. In the latter, it is expected that the neural network classify the inputs in different groups according to the output (but, in this case, it is not specified explicitly the relation between input and output). In this paper, we will work with a supervised artificial neural network

B. *Input Data Integrity: Clinical Data Vs Laboratory Data.*

As it is easy to see, the larger and more representative the set of available examples used for training the automatic classifier is, the better the future classification of query cases will be. However, there are several challenges and practical limitations associated with medical and clinical data compilation. Acquiring large volumes of data representing certain diseases is often challenging due to the low prevalence of the disease. In addition, collecting data from patients is time consuming and this came into conflict with the immediate way in which this data are obtained in clinical practice. Finally, bureaucratic limitations related to clinical data privacy protection legislation could make that only a few portion of the whole set of samples could be used for experimentation.

For all this reasons, having an input database with a lack of integrity ("Clinical Data") is a common problem at the time of training an automatic classifier and data must be treated and filtered ("Laboratory Data").

The literature considers input data preprocessing as one of the different stages in a machine learning performance. The database's non-consistency (Dirty Database) raises the need of these preprocessing methods to "clean up" the data so that machine learning algorithms will be able to tease out key information and correctly classify new samples [10].

We can find several researches about cleaning-data and preserving-integrity techniques. For instance, it has been proved that target-class imbalance, i.e. its underrepresentation in the set of collected examples, could lead to a poor performance of ANN-based CAD systems [11].

We focus this subject in a different way, trying to prove that sometimes a properly designed and trained ANN could reach the same rates of accuracy, sensitivity and specificity using either an incomplete, or dirty, data set or a clean one at the training stage.

III. Experimentation

Since a database of characteristics from prostate cancer diagnostic tests is at our disposal, we have built two different sets from this data. The first one contains the original samples with the least treatment needed to use it as a knowledge base. The second one is obtained after applying preprocessing and cleaning techniques like filling missing values, normalization, or pruning to original raw data [12].

Next we have designed and implemented two models of ANN based on each data set. Finally, we have applied to them training, validation and testing processes.

Each of this experimentation stages is widely explained in the following subsections.

A. *Prostate Cancer Database.*

Our clinical database contains 950 samples from prostate cancer diagnostic tests performed by an expert urologist. The samples could be divided into two classes depending on the test results: "healthy patient" and "patient who suffers from prostate cancer". Besides the diagnostic results for all of the samples, the tests also include values for 14 characteristics more for each patient. These characteristics are commonly used by urology experts for prostate cancer diagnosis: age, PSA in blood level, PSA density, prostate volume, rectal examination results, transitional zone flow, peripheral zone transitional, intralesional IR, intraprostatic IR, periprostatic IR, state of the prostate capsule, state of the seminal vesicles, quotient, and prostateseminal angle.

Not all the fields are numerical, 5 of them are filled using a subset of medical terms. In order to use these text fields, we have related each term with a number (e.g. adenoma, LD nodule, LI nodule and bilateral nodule, which are values of "rectal test results" fields, are translated to 1, 2, 3 and 4 respectively). On the other hand, the final diagnosis has two possible values: 'yes' or 'no' that we have identified with 1 and 0 respectively (see Table I).

Only a subset of 44 samples has values for all the fields. The remaining samples have a number of empty fields that ranges from 1 (34%), to a half of the whole number of fields (23%).

Even, in few cases the stored values seem to be clearly out of range for being much higher or lower than the expected values. This could be due to typing mistakes.

B. *Dirty and Clean Data Sets.*

We use the mentioned database to build two different sets of data: the dirty and clean sets.

The Dirty Data Set (DDS) is obtained applying to the original samples, the minimum number of transformations needed that allow us to use it as an input of an ANN. We decide to replace missing values with zeros. Since every chosen value would introduce noise into the whole data set, it seems to be the most suitable value in order to fill the empty fields avoiding the use of preprocessing techniques like calculating the average value of each characteristic.

The Clean Data Set (CDS) consists on the data obtained after applying the following preprocessing strategies to the original raw data.

1) Filling Empty Fields: As in the DDS, we have filled the empty fields, but in this case using the average value of the characteristic to the empty field belongs.

2) *Normalization:* The purpose of this operation is to normalize the spectra from different samples such that they are comparable. First we obtain a frequency histogram of each characteristic and then we divide the total range of values into different little ranges including each one a similar amount of samples. Finally, under the urological expert supervision, a number, ranging from 1 to the total number of ranges created, is assigned to each sample depending on which range it is included its original value.

3) Codifying Text Fields: It is the case of fields that in the original database had a textual value. We associate a numerical value to each category that appears in the textual characteristics fields in order to replace them and make it computable by ANN mathematical procedures.

4) Pruning: After examining the entire set of data, those samples that contained highly discriminant and clearly out of range values were deleted (12 samples).

5) Features Selection: We try to filter non-relevant features. There are 3 features that only contain values for 150 of the 950 samples from the database. The amount of empty fields is too much high, so, it seems convenient to exclude this features from the diagnosis set.

C. Designed Neural Network.

The ANN designed to analyze the data sets is a multilayer perceptron (MLP) that is based on the supervised learning model [13]. MLPs have been used with high rates of success in researches that implies the use of automated methods to support the cancer diagnosis [14] [15] [16] [17].

A MLP consists of two or more layers of simple processing units called perceptrons. The net has as much inputs as different features are in the input data set and a variable number of outputs. In our case, the designed MLPs have 14 inputs, the one that uses the dirty data set (MLP1), and 11 inputs the one based on the clean data set (MLP2). The output will be binary, 1 or 0, corresponding to positive or negative value of suffering from prostate cancer; therefore only one neuron will be included on the last layer.

The number of neurons on the intermediate, or hidden, layer has been determined by empirical test.

D. Training, Validation and Testing processes.

Ideally separate data sets should be used for each of these processes. However, in practice, some form of data partitioning of a single data set, such as cross-validation or bootstrap sampling, is employed due to the difficulties of obtaining a large number of samples [18] [19]. In our case we have divided the input data in three non-overlapping sets to carry out the training (60% of the input samples), validation (20%), and testing (20%) processes.

There are a set of customizable parameters that has to be obtained by testing. For this reason, we wrote an executable script to test in a batch way several parameters for each MLP and compiling metrics after the execution of each one.

This process allows us to choose the proper parameters configuration. Firstly, we try to find the number of hidden layers and the suitable size of each net's layer. We have tested designs that contain 1 and 2 hidden layers with a range from 5 to 20 neurons in each layer. Secondly, we look for the transfer function that will control the input data through the net. We tested different combinations of tan-sigmoidal, log-sigmoidal, and lineal transfer functions applied to the hidden layer transfer function and the output function. Besides this, we need to know the ideal training algorithm, which should be used during the training process. There is a wide variety using each one of them a different mathematical model to do this task. In this case the tested training algorithms have been five variations of the backpropagation general algorithm.

After the previous task we know the parameters' configuration that produces the best results for each of the implemented MLPs. Finally, we have repeated the test 15 times for each MLP using this best-configuration. The result of the experimentation for the dirty data set is shown in table II with the name of MLP1; the MLP with best results using the clean data set is the named MLP2 in Table II.

TABLE II
FEATURES OF THE DESIGNED MLPS

Net	Hidden Layer Neurons	Inputs	Outputs	Transfer Function	Trainig Algorithm
MLP1	10	14	1	Log Sigmoid	BFGS
MLP2	13	11	1	Tan Sigmoid	Resilient Backpropagation

E. Obtained Results.

After all the batch tests performed we compare the obtained results for each MLP (see Table III). As we can observe, the results obtained from the clean data set are quite similar to the obtained from the dirty data, in accuracy rate and also in sensitivity and specificity.

TABLE III
EXPERIMENTATION RESULTS

Net	Data Set	Design	Epoc.	Accuracy	Sensitivit.	Specificit.
MLP1	Dirty	[10 1]	5000	79%	89%	59%
MLP2	Clean	[13 1]	5000	76%	88%	60%

IV. CONCLUSION

We have built two data sets obtained from a clinical database that contains samples from patients who have been diagnosed with prostate cancer by an expert urologist. One of the data sets contains raw data without preprocessing (Dirty Data). The other's data (Clean Data) is the result of applying different preprocessing techniques to the original data.

Next we have designed two implementations of a multilayer perceptron, which is an ANN based on the supervised learning paradigm. Each of the MLP implementations has been trained, validated and tested using one of the data sets.

After this process, both systems are evaluated and compared in terms of clinically relevant metrics such as diagnosis accuracy, sensitivity and specificity, in order to check the influence of the data set used on the MLPs performance.

Looking at the results we can say that, the values for each of the measured metrics are quite similar in both MLPs performing. We should point out the similarity of the performance results, more than how good or not are these results.

We can conclude that if a classifier's design is robust enough, it is more unlikely to be affected in its performance by the lack of integrity and quality of the input data. This may make us questioning ourselves about the need, or not, of spending a lot of time and efforts on preprocessing input data sets instead of designing and implementing more robust classifiers.

As future works we will try to repeat this job procedure with other automatic classifiers based on different learning paradigms like unsupervised learning, competitive learning and competitive supervised learning. We might use too other automatic classifiers like genetic algorithms or support vector machines.

REFERENCES

[1] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu. (2009, May). USA Cancer Statistics, 2009. *CA Cancer Journal for Clinicians.* [Online]. *59.* pp. 225—249.

[2] A.M.D. Wolf, R. C. Wender, R. B. Etzioni, I. M. Thompson, A. D. D'Amico. (2010, March). American Cancer Society Guideline for Early Detection of Prostate Cancer: Update 2010. *CA Cancer Journal for Clinicians.* [Online].

[3] O. W. Brawley, D. P. Ankerst, I. M. Thompson. (2009, June). Screening for Prostate Cancer. *CA Cancer Journal for Clinicians.* [Online]. *59(4).* pp. 264—273.

[4] M.K. Brawer. (1999, September). Prostate-specific antigen: Current status, *CA Cancer Journal for Clinicians* [Online]. *49*. pp. 264–281

[5] D. Ruiz, A. Soriano, "A distributed Approach of a Clinical Support System Based on Cooperation" in *Mobile Health Solutions for Biomedical Applications*, P. Olla, J. Tan , Ed. IGI Global, 2009.

[6] P. J. Lisboa, A.F.G Taktak, "The use or artificial neural networks in decision support in cancer: A systematic review" *Neural Networks*, vol. 19, no. 4, pp. 408—415, May. 2006.

[7] M. F. Abbod, J. W. F. Catto, D. A. Linkens, F. C. Hamdy, "Application of Artificial Intelligence to the Management of Urological Cancer" *The Journal of Urology*, vol. 178, no.4, pp. 1150—1156, Oct. 2007.

[8] D. Andina, *Computational Intelligence for Engineering and Manufacturing*. Springer, 2007. pp 39—109.

[9] H. Shin, M. K. Markey, "A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples" *Journal of Biomedical Informatics,* vol. 39, pp. 227—248, Apr. 2006.

[10] E. Alpaydin, *Introduction to Machine Learning.* MIT Press, 2004.

[11] M. A. Mazurowski, P. A. Habas, J. M. Zurada. J. Y. Lo, "Training neural network classifiers for medical decision making: The effects of imbalanced data sets on classification performance" *Neural Networks*, vol. 21, no. 2-3, pp. 427—436, Mar-Apr. 2008.

[12] J. A Drakopoulos, A. Abdulkhader, "Training neural networks with heterogeneous data" *Neural Networks*, vol. 18, no. 5-6, pp. 596—601, Jul-Aug. 2005.

[13] S. Haykin, *Neural Networks and Learning Machines,* 3rd ed. Prentice Hall, 2008.

[14] T. Anagnostou, M. Remzi, M. Lykourinas, B. Djavan, "Artificial Neural Networks for Decision-Making in Urologic Oncology" *European Urology*, vol. 43, no. 6, pp. 596—603, June 2003.

[15] M. Çinar, M. Engin, E. Z. Egin, Y. Z. Atesçi, "Early prostate cancer diagnosis by using artificial neural networks and support vector machines" *Expert Systems with Applications,* vol. 36, no. 3, pp. 6357—6361, Apr. 2009.

[16] C. Stephan, N. Büker, H. Cammann, C. Xu "PSA-Velocity and artificial neural networks (ANN)-Velocity to differentiate Prostate Cancer from Benign Prostatic Disease" in *Proc. 22nd Annual Congress of the European Association of Urology*, Barcelona, Spain, 2007, pp. 224.

[17] E. Coiera, *Guide to Health Informatics,* 2nd ed. London: Hodder Arnold, 2003.

[18] H. Wang, H. Wong, H. Zhu, T. T. C. Yip. "A neural network-based biomarker association information extraction approach for cancer classification" *Journal of Biomedical Informatics,* vol. 42, no. 4, pp. 654—666, Aug. 2009.

[19] C. Stepahn, H. Cammann, H. Meyer, M. Lein, "PSA and new biomarkers within multivariate models to improve early detection of prostate cancer" Cancer Letters, vol. 249, no. 1, pp. 18—29, Apr. 2007.