* For the reference, please see the last page. *

Monotonicity and Logical Analysis of Data: A Mechanism for Evaluation of Mammographic and Clinical Data[1]

Boris Kovalerchuk and Evangelos Triantaphyllou
Department of Industrial Engineering, Louisiana State University, Baton Rouge, Louisiana

James F. Ruiz
Department of Radiology, Woman's Hospital, Baton Rouge, Louisiana

1. INTRODUCTION

Computer assisted diagnosis has become one of the promising methods for improving the accuracy and early detection of breast cancer. Standard back-propagation neural networks are presently very popular diagnostic tools, but this approach does not inform the physician user on how a conclusion was reached. Some promising results have obtained with rule based techniques[3]. This method has the advantage of providing the physician with a tool that promotes consistency and accuracy. However, these (and most other) models suffer from relatively small training sample sets which in turn limit statistical significance.

An approach called logical analysis of data (LAD), and which is based on inferring discriminant Boolean functions, has the potential to overcome these weaknesses. By discovering logical relationships in existing classes of disjoint observations, the method can improve the understanding of the diagnostic process The statistical significance problem is eliminated by exploiting the property of monotonicity that exists within mammographic evaluation and interpretation.

The resultant discriminant functions could be used many ways. In the evaluation of a new *"problem case"* the radiologist could use these functionsfor that case to draw a diagnostic conclusion. Alternatively, with a set of *"gold standard"* test cases, the functions could be used as a reproducible testing mechanism. Each radiologist could determine his/her own function, compare it to the gold standard and thereby identify areas of strength or weakness. Progress or improvement over time could then be objectively measured.

2. BACKGROUND

Currently, methods of computer-aided diagnostics include neural networks, nearest neighbor methods, descriminant analysis, cluster analysis, and linear programming based methods. [3,11] These techniques attempt to generalize from collections of available training data. Therefore, they rely on the fact that the more representative the data are, the more accurate the performance of these methods will

1 *The first two authors gratefully acknowledge the support from the Office of Naval Research (ONR) grant N0014-95-1-0639*

be. However, there **are some basic weaknesses** in using these techniques. For example, according to Jonson[6] the use of Baysian models in medical diagnosis is controversial, if not unethical, because the fundamental requirement of **strict randomness** rarely occurs in reality. Standard back-propagation neural networks techniques are problematic because they do not provide **bio-medical explanations**[4]. While newly developed intelligent hybrid systems, and in particular knowledge based neural networks[4], appear to be more promising, these systems have the potential to create too many (i.e., an exponential number) of rules[12]. In summary, there are two closely related difficulties with these methods as they may be applied to breast cancer diagnosis:

(i) **The available training data are often insufficient to guaranty statistically significant results by using ordinary mathematical methods.**

(ii) **The way these computerized diagnostic tools work and produce recommendations, is not appealing to medical doctors**.

The diagnostic problem considered in this research is a nested one. That is, it is composed of two interrelated sub-problems. The **first sub-problem** is related to the clinical question of whether a biopsy or short term follow-up is necessary or not. The **second sub-problem** is related to the question whether the radiologist believes that the current case is highly suspicious for malignancy or not. It is assumed that if the radiologist believes that the case is malignant, then he/she will also definitely recommend a biopsy. Formally, these two sub-problems are defined as follows:

The Clinical Management Sub-Problem: One and only one of two disjoint outcomes is possible. That is: 1) *"biopsy/short term followup is necessary"*; or 2) *"biopsy/short term followup is not necessary"*.

The Diagnosis Sub-Problem: Similarly as above, one and only one of two disjoint outcomes is possible. That is, a given case is: 1) *"highly suspicious for malignancy"*; or, 2) *"not highly suspicious for malignancy"*.

The above nested sub-problems provide a natural area for the application of a recent development in artificial intelligence. The technique aims at to discovering **logical relationships** which may be present in classes of disjoint observations. Each observation is described in terms of a number of characteristics or features. Then, the challenge is to find patterns which can be used to explain which features, or combinations of features, are present in each class. This development is called *Logical Analysis of Data (or LAD)* and consists of a group of methods.[2,7,8] The proposed logical analysis of data approach has several unique advantages related to: (1) the understanding/comprehension of the diagnostic process and recognition of sources of uncertainty in existing knowledge, (2) overcoming the statistical significance problem mentioned earlier by utilizing the ***property of monotonicity***, (3) resolving differences arising in conclusions that are based upon the same premises, (4) developing AI-based expert systems to assist in the diagnosis of breast cancer.

The validity and accuracy (reliability) of a classifier should be reasonably high for clinical applications. Many, if not the majority, of the pattern recognition studies in radiology operate with approximately 80 training cancer cases. To explore the issue of **reliability**, assume that $f(x_1,x_2,...,x_{11})$ is some discriminant function in the space of 11 binary (i.e., with 0 or 1 value) indirect diagnostic features (vectors,) see section 5. Also assume that this discriminant function was constructed by using a sample of 80 cases (many studies consider sample sizes around 80). Next, suppose that the function f discriminates the entire state space (which is of size: $2^{11} = 2{,}048$) by describing 2,048 distinct binary vectors (combinations of 11 features). The vectors are as follows: 1,600 (i.e., about 80%) of them are features suggestive of cancer and 448 (i.e., about 20%) are negative for cancer. A description of suspicious/benign features is given in section 5. Thus, in this illustrative scenario, at most 80 different vectors (5% of 1,600) were used to learn to discriminate 1,600 cancer vectors. At this point one may wish to ask the question: "Is this function (which was inferred by using a training sample of 5% of the combinations of features suggesting malignancy) sufficiently reliable to require a patient to undergo a surgical procedure ?" While this function can be useful, its statistical significance is very questionable for the reliable diagnosis of cancer. Observe that the concept of **state space** is different than the concept of **population**. The state space expresses all possible combinations of features. However, the actual population may not exhibit all these combinations of features (i.e., some combinations may be impossible) and these can be eliminated from consideration. We consider the ratio *S/N* (where *S* is the sample size and *N* is the size of the state space) as the index of the potential reliability.

In the light of the above considerations, consider the published data[11]. We estimated the sample/space ratio (i.e., the *S/N* value) for these data. That is, for 43 features $S/N = 133/10^{43} = 1.33\times10^{-41}$. This means that the available sample is 1.33×10^{-39} % of the total possible number of different vectors in the state space. Therefore, the question which is naturally raised is: *"Can this small number of training cases be considered as reliable in order to assist in accurately diagnosing new (and thus unknown) cases?"* According to some estimations[5] the recommended value of the *S/C* ratio (where *C* is the number of connections in a neural network) must be no less that 10:1 for potentially reliable learning of a neural network. (Note that this ratio is similar but not identical to the *S/N* ratio).

The 10:1 ratio is debated by Boon[1]. He compares the neural network with biological networks (e.g. radiologists), showing that for them the ratio of sample/connections is much worse, over 10^{10} times less, than the ones presented in studies which were criticized by Gurney[5]. Thus, Boone asks: "Is there any reason that we should hold a computer to higher standards than a human?" Maybe not, but we should ask of both systems: "Is learning based on small subsets sufficiently reliable to distinguish suspicious from not suspicious given the vast diversity of all potential mammographic images?" The Machine Learning Theory[10] shows that there are relatively simple concepts that no algorithm is capable of learning in a reasonable amount of time (polynomial time). **Therefore, the question about reliability is among the most fundamental questions of scientific rigor of mammographic diagnosis.** Thus, in this study we explore the question: "Are accessible, relatively small samples sufficiently representative for learning how to evaluate the broad range of mammographic appearances?".

3. THE REPRESENTATION HYPOTHESIS

One of the most common fundamental hypotheses supporting the usage of small samples in pattern recognition is the hypothesis that a small sample is representative for the whole population (the *representation hypothesis*): *"The training data must still form a representative sample of the set of all possible inputs if the network is to perform correctly"*[9]. Therefore, without confirmation of the representation hypothesis as it applies to mammography, all pattern recognition work with small samples will be extremely heuristic and of questionable reliability. Specifically, we study the following version of the representation hypothesis: the *hypothesis of narrow vicinity*. According to the hypothesis of narrow vicinity (the NV hypothesis): ***All real possible cases are in the narrow vicinity of an accessible small training sample.*** The NV hypothesis means that we can generalize a training sample with 80 vectors, for instance, for an additional 40-80 vectors, but not for an additional 1,600 vectors. It also means that we may exclude those 1,520-1,560 vectors from 1,600 that do not represent possible cases.

4. TASKS

Let us consider how one can confirm the narrow vicinity (NV) hypothesis when a small sample is available. This is a difficult methodological question. If one has a large sample available, then he/she does not need the NV hypothesis. On the other hand, without a large sample, one does not have direct data to confirm the hypothesis.

We develop a new methodology to avoid these difficulties. The main idea is to **extend insufficient clinical cases with information from an experienced radiologist**. Another approach is mentioned by Miller et al. [9]: "One obvious solution to the problem of restricted training and testing data is to create simulated data using either a **computer based or physical model.**" Unfortunately this way is still a theoretical possibility. There is no such model. We will use experienced experts as a **"biological"/expert model to generate new potential cases**. One can ask a radiologist to evaluate a particular case when a number of features take on a set of specific values. A typical query will have the following format: "If feature *1* has value V_1, feature 2 has value V_2, ..., feature n has value V_n, then should biopsy/short term follow-up be recommended or not? Or, does the above setting of values correspond to a highly suspicious case or not?"

The above queries can be defined with artificially constructed vectors or with artificially generated new mammograms by modifying existing ones. In this way one may increase a sample size but not as much as may be necessary. The technical weakness now is, roughly speaking, the same as before. It is practically impossible to ask a radiologist to generate many thousands of artificial, new mammographic appearances.

A logical analysis approach can avoid these difficulties in two possible ways. First, if the features can be organized in a **hierarchical manner,** then a proper exploitation of this structure can lead to a significant reduction of the needed queries. Second, if the property of monotonicity is applicable, then the available data can be generalized to cover a larger training sample[7,8].

5. METHODS AND RESULTS

We construct a hierarchy of medically interpretable features from a very generalized level to a less generalized level. For example, we consider the generalized binary feature x --*"Shape and density of calcification"* with grades (0-*"contra cancer"* and 1-*"pro cancer"*). On the second level we consider the feature x to be some function of five other features: $y_1, y_2, .., y_5$. That is, $x = \psi(y_1, y_2, .., y_5)$, where: y_1-- *"Irregularity in shape of individual calcifications"*; y_2-- *"Variation in shape of calcifications"*; y_3-- *"Variation in size of calcifications"*; y_4-- *"Variation in density of calcifications"* and y_5-- *"Density of calcifications"*.

Monotonicity in Boolean functions is the core of a powerful mathematical theory, which we applied in the breast cancer problem. For instance, consider the evaluation of calcifications in a mammogram. For simplicity and illustrative purposes assume that x_1 is the number and the volume occupied by calcifications, in a binary setting, as follows: (0-*"contra cancer"*, 1-*"pro cancer"*). Similarly with the same values we used features: x_2--{shape and density of calcifications}, x_3--{ductal orientation}, x_4--{comparison w. previous examination}, and x_5--{associated findings}.

Given the above definitions we can represent clinical cases in terms of binary vectors with these five features as: $(x_1, x_2, x_3, x_4, x_5)$. Next consider the two clinical cases which are represented by the two binary vectors: (00111) and (10100). If one is given that a radiologist correctly diagnosed (10100) as a malignancy, then, by utilizing the monotonicity property, we can also conclude that the second clinical case (10110) should also be malignancy.

A detailed description of the method and the specific steps can be found in [7,8] we show that the descriminant monotone Boolean functions for features on the upper level of the hierarchy are as follows. For the **Biopsy/Short** term followup Sub-Problem: $f_1(x) = x_2x_4 \vee x_1x_2 \vee x_1x_4 \vee x_3 \vee x_5$. Similarly, for the second sub-problem (**highly suspicious for cancer**) the function which we found was: $f_2(x) = x_1x_2 \vee x_3 \vee (x_2 \vee x_1 \vee x_4)x_5$ Regarding the second level of the hierarchy (which recall has 11 binary features) we interactively constructed the following functions (interpretation of the features is presented below): $x_1 = \varphi(w_1, w_2, w_3) = w_2 \vee w_1w_3$ and $x_2 = \psi(y_1, y_2, y_3, y_4, y_5) = y_1 \vee y_2 \vee y_3y_4y_5$. By combining the previous functions we obtained theformulas of all the 11 features for **biopsy**/short term followup:

$$f_1(x) = (y_2 \vee y_1 \vee y_3y_4y_5)x_4 \vee (w_2 \vee w_1w_3)(y_2 \vee y_1 \vee y_3y_4y_5) \vee (w_2 \vee w_1w_3)x_4 \vee x_3 \vee x_5$$

and for **highly suspicious for cancer**: $f_2(x) = x_1x_2 \vee x_3 \vee (x_2 \vee x_1 \vee x_4)x_5 =$

$$= (w_2 \vee w_1w_3)(y_1 \vee y_2 \vee y_3y_4y_5) \vee x_3 \vee (y_1 \vee y_2 \vee y_3y_4y_5) \vee (w_2 \vee w_1w_3 \vee x_4)x_5$$

To construct these functions with usual methods one needs more than **4000 training cases**[8] or to ask an experienced radiologist to diagnose these 4000 cases, presented as binary vectors. We constructed these discriminant functions for one hour in dialog with a radiologist. Our approach allowed to ask only **42 questions, i.e., about 100 times less diagnostic questions than the number of all possible cases.**

An analysis of these functions and real statistics has shown that that, in general, the hypothesis of narrow vicinity (NVH) is not valid for mammographic evaluation. **Recall, that this is exactly the hypothesis implicitly used by all traditional pattern recognition methods in breast cancer diagnosis!**

We have done this study for binarized the features presented below: x_1 - amount & volume of calcifications , w_1 - number of calcifications/sm^2; w_2 - volume, cm^3 w_3 -total number of calcifications; x_2 -shape & density of calcifications.

Note, we consider x_2 as a function (y_1,y_2,y_3,y_4,y_5) of y_1,y_2,y_3,y_4,y_5,. y_1 - irregularity in shape of individual calcifications; y_2 - variation in shape of calcifications; y_3 -variation in size of calcification; y_4 - variation in density of calcification; y_5 - density of calcification; x_3 - ductal orientation; x_4 - comparison with previous exam ; x_5 - associated findings.

We assume that radiologists implicitly use monotone regularities when they learn to diagnose breast cancer. This gives them a chance to be more successful than computer programs learning just by positive and negative examples. Incorporation of these verified regularities in computer learning programs opens a new direction of improvement in mammographic evaluation.

REFERENCES

1. J.Boon "Sidetracked at the Crossroads", Radiology, v. 193, n. 1, 28-30, 1994.

2. E.Boros, P. Hammer, T. Ibaraki, "Predicting cause-effective relations from incomplete discrete observations", SIAM J. on Discrete Mathematics, v. 7, n.4, 1994, 531-534.

3. C.D'orsi,D.Getty,J.Swets,R.Picket,S.Seltzer,B.McNeil, "Reading and decision aids for improved accuracy and standardization of mammographic diagnosis", Radiology, v. 184, 619-622, 1992.

4. L.M.Fu, "Knowledge-based connectionism for revising domain theories". IEEE Transactions on Systems, Man, and Cybernetics, v. 23, n. 1, 173-182, 1993.

5 J.Gurney "Neural Networks at the crossroads: caution ahead", Radiology, v. 193., n. 1, 27-28, 1994.

6 N. Jonson, "Everyday diagnostics-- a critique of the Bayessian model". Med. Hypotheses, v. 34, n. 4, 289-96, 1991.

7. B.Kovalerchuk, E.Triantaphyllou, E.Vityaev, "Monotone Boolean functions learning technique integrated with user interaction". In: Proc. of Workshop "Learning from examples", 12-th International Conference on Machine Learning, Tahoe City, CA, 41-48, 1995.

8. B.Kovalerchuk., E.Triantaphyllou, A. Deshpande, and E. Vityaev, "Interractive learning of a monotone Boolean function". Information Sciences, 1996 (in print).

9.A.Miller, B.Blott, T.Hames, "Review of neural network applications in medical imaging and signal processing", Medical & Biological Engineering & Computing, v. 30, 449-464, 1992.

10.R.Schapire, The design and analysis of efficient learning algorithms. MIT Press, 1992.

11.Y.Wu,M.Giger, K.Doi.C.Vyborny, R.Schmidt, C.Metz, "Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer", Radiology, v. 187, n. 1, 81-87, 1993.

12 J.Shavlik, "Combining symbolic and neural learning", Macine Learning, v.14, 1994, 321-331.

SCAR 96

COMPUTER APPLICATIONS TO ASSIST RADIOLOGY

Editors

Ray F. Kilcoyne, M.D., F.A.C.R.
Professor of Radiology and Orthopaedics
Vice-Chairman, Department of Radiology
University of Colorado Health Sciences Center,
Denver, Colorado

James L. Lear, M.D.
Professor of Radiology
University of Colorado Health Sciences Center,
Denver, Colorado

Alan H. Rowberg, M.D.
Associate Professor of Radiology
University of Washington,
Seattle, Washington

Published by

Symposia Foundation

A Public Non-Profit Organization Dedicated to the Health Care of the Public

Founded 1979