Quality of Life through Quality of Information
J. Mantas et al. (Eds.)
IOS Press, 2012

© 2012 European Federation for Medical Informatics and IOS Press. All rights reserved.
doi:10.3233/978-1-61499-101-4-781

Information-Based Sequential Selection of Clinical Tests in Risk Assessment

Naama PARUSH [a,1], Tal EL-HAY [a], Michal OZERY-FLATO [a], Ligita RYLISKYTE[b,c], Zydrune VISOCKIENE [b,c], Aleksandras LAUCEVICIUS [b,c]

[a] *IBM Research Haifa Labs, 165 Aba Hushi St., Haifa 31905, Israel*
[b] *Vilnius University Hospital Santariskiu Klinikos, Vilnius, Lithuania*
[c] *Faculty of Medicine, Vilnius University, Vilnius, Lithuania*

Abstract. We present a new framework for supporting decisions in sequential clinical risk assessment examinations. In this framework, the decision whether to perform a test depends on its expected contribution to risk assessment, given results of previous tests, and the contribution is quantified using information theory. In many cases adding an additional examination clearly improves the predictive model. However, there are cases in which the improvement is not constant for all values of previous tests, and quantification of possible improvement can support decision on further examinations. Using this approach can prevent many expensive, unpleasant or risky examinations. We demonstrate the use of this method on an example of type 2 diabetes onset study. The results show that reducing a considerable percent of the blood tests does not decrease the model's prediction power.

Keywords. Information gain, predictive models, sequential tests.

Introduction

Predictive models use information driven from clinical tests, lifestyle and other personal and medical history details to predict a patient's future condition. The biomarkers used by these models, are chosen based on their contributions to the prediction. In general, these risk assessment models require all chosen biomarkers.

Models that can perform without all biomarkers usually assume that the values are randomly missing or that the probability of the values being absent known already at the time of model generation [1]. Another model that can possibly perform without all biomarkers is the ordinary decision tree [2, 3]. An examination can be avoided when there is a path in the tree that does not involve (at least) one of the biomarkers and the other biomarkers lead to that path. However, even if such a path exists, it is predefined and cannot be dynamically adjusted at evaluation time to updated costs or personalized preferences.

We suggest a new framework, where the biomarkers are chosen sequentially at assessment time based on the amount of information they could contribute given the results of the previous examinations. In this framework physicians receive a quantitative measure that can help decide whether or not to use the additional biomarker, according to the physician's, patient's, or HMO's preferences.

The conditional information gain (cInfoGain) is a measure that quantifies the possible additional information based on information theory [4]. The next biomarker in

[1] Corresponding Author, e-mail: naamap@il.ibm.com

the sequence is examined only if the cInfoGain is high enough. The threshold can be set by the physician/patient/HMO while evaluating the cost vs. additional predictive power evaluated by the cInfoGain. Figure 1 illustrates the decision process using the cInfoGain measure. In this example, in order to predict Y (a future condition of the patient), biomarker A is examined. Thereafter, depending on the cInfoGain of biomarker B given the value of biomarker A, biomarker B is examined. When only biomarker A is examined, Y is predicted using model 1. When both biomarkers are examined Y is predicted using model 2.

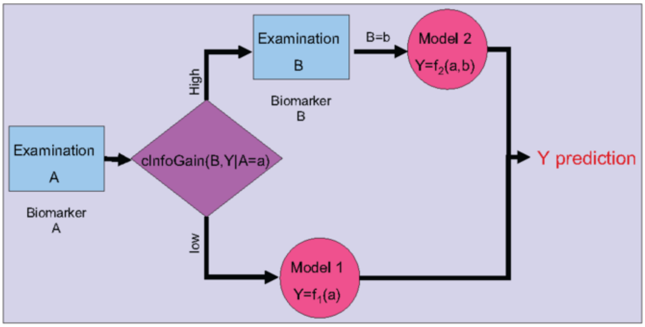

Figure 1. Schematic illustration of the decision process using the cInfoGain measure.

Using this approach can reduce many costly, unpleasant or risky examinations such as radiology scans or invasive pregnancy screening tests, while controlling the decrease in prediction power. The method is quantative and can be easily updated and adjusted to different populations.

In this paper we first present the conditional information gain measure, and thereafter demonstrate how it can be used to support decisions in sequential tests. Throughout this paper our demonstrations focus on data from a type 2 diabetes onset prediction study.

1. Methods

This method is based on information theory and in particular on the mutual information (MI) and entropy (H) measures [4]. The entropy measure quantifies in bits the distribution's “uncertainty” or “randomness”, and it ranges from 0 to $\log_2(|X|)$. $H(X) = 0$ when the identity of x is known with full certainty. $H(X) = \log_2(|X|)$ when x is totally random (i.e., uniformly distributed): $p(x) = 1/|X|$ for all values of x. Intermediate values correspond to intermediate levels of uncertainty. $H(X|Y)$ denotes the conditional entropy of variable X given variable Y.

The mutual information between two variables is defined as the “uncertainty” of one of the variables (in bits) reduced on average by knowledge of the other variable: $MI(X, Y) = H(X) - H(X|Y)$. Another definition of MI is the Kullback–Leibler divergence (D_{KL}) between the actual joint probability of X, Y [$p(x, y)$] and the expected independent probability [$p(x)*p(y)$].

In the proposed method, we assume there are n+1 biomarkers denoted by $A_1 \ldots A_n, A_{n+1}$ and a future outcome denoted by Y. We assume biomarkers $A_1 \ldots A_n$ are already measured and the physician needs to decide whether to perform examination A_{n+1}. The cInfoGain of A_{n+1} given A_1=a_1, …, A_n=a_n is defined as the conditional mutual information between biomarker A_{n+1} and Y given the values of the previous biomarkers:

$$cInfoGain(Y, A_{n+1} \mid A_1 = a_1,\ldots,A_n = a_n) = MI(Y, A_{n+1} \mid A_1 = a_1,\ldots,A_n = a_n) =$$

$$\int_{a_n}\sum_{y} p(a_{n+1}, y \mid a_1,\ldots,a_n)\log_2\left(\frac{p(A_{n+1}, y \mid a_1,\ldots,a_n)}{p(A_{n+1} \mid a_1,\ldots,a_n)p(y \mid a_1,\ldots,a_n)}\right)da_n$$

Calculating MI requires an estimation of the joint probability of biomarkers A_1,… A_n, A_{n+1} and Y. However, joint distributions are difficult to estimate especially when the data set is relatively small. Therefore, in the following examples, the distributions are estimated using a simple normal distribution parametric model and the cInfoGain is computed using numerical integration.

To demonstrate the method, we analyzed data from a prospective study of 366 non-diabetic middle-aged Lithuanian men and women (120 men, 246 women) having metabolic syndrome but without overt cardiovascular diseases. The patients were recruited from the Lithuanian High Cardiovascular Risk (LitHiR) primary prevention program [5] and were referred to the Vilnius University Hospital “Santariskiu Klinikos" tertiary care center for further assessment. During a follow-up period of three to four years, there were 31 cases of incident type 2 diabetes onsets. At baseline, all participants underwent several examinations and measurements including BMI (body mass index - weight in kilograms divided by the height in meters squared) and a fasting plasma glucose (FPG) test. We concentrated on calculating the cInfoGain of FPG on a prediction of diabetes onset given different BMI values and estimating the performance of a model that does not use FPG values corresponding to low cInfoGain. In this example, A_1=BMI, A_2=FPG and Y=diabetes onset. Logistic regression was used to develop predictive models for incident cases. Leave-one-out cross validation and receiver operating characteristic (ROC) analyses were used to assess the discriminatory power of the prediction models.

2. Results

Figure 2 depicts different cInfoGain curves for diabetes onset given various BMI values. The expected random cInfoGain was calculated using the mean and standard deviation of a random shuffling of the FPG biomarker. FPG contributes different cInfoGains for different BMI values (ranges from 0 to 0.3 bits) while the age curve is not significantly different than random.

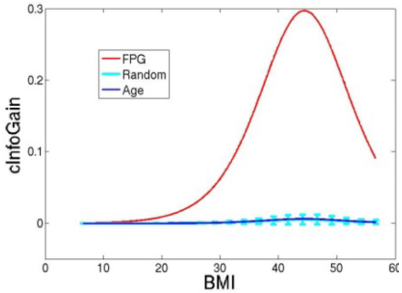

Figure 2. cInfoGain between diabetes onset and FPG, age and random FPG for different BMI values. The random FPG curve represents the average of 300 random curves and the error bar represents 1 std.

The cInfoGain of FPG is determined by the uncertainty of developing diabetes given the BMI value alone, and the information the FPG can provide at that point. We focused on three values of BMI: 15, 44.5 and 55, with respective cInfoGains of 0.003, 0.3 and 0.12. Figure 3 illustrates the probability of developing diabetes at these BMI values as a function of FPG and the corresponding probability of diabetes onset independent of FPG. At BMI=15, the probability of diabetes onset is close to 0, i.e. there is very little uncertainty. In addition, the probability of diabetes onset as a function of FPG is almost uniform and therefore not very informative. The probability of diabetes onset at BMI=44.5 and 55 ranges from 0 to 1. However at BMI=44.5 there is high uncertainty (there is almost equal probability of developing diabetes independent of FPG) whereas at BMI=55 there is little uncertainty (the probability of diabetes onset is close to 1), and therefore the cInfoGain at this point is lower than the cInfoGain at BMI=44.5.

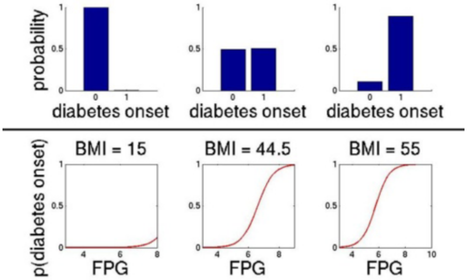

Figure 3. Probability of diabetes onset independent of FPG for selected values of BMI (15, 44.5, 55) and the corresponding probability of diabetes onset as a function of FPG

To test the effectiveness of this framework, the following analysis forgoes FPG tests that correspond to low cInfoGain values. First, we generated two logistic regression predictive models: BMI and BMI-FPG based models. The corresponding performance was evaluated by using leave-one-out cross-validation, and the estimated area under ROC curves (AUC) were 0.74 and 0.92 respectively (these ROC curves were significantly different [6], p-value <0.0001). Second, for each cInfoGain threshold, we estimated the AUC of a predictive model that uses the BMI model on records with smaller cInfoGain and uses the BMI-FPG model otherwise. We evaluated the performance of this model using the leave-one-out cross validation, and the cInfoGain of each record was calculated by estimating the joint distributions on all the data without the evaluated record. This analysis showed that the model performance did not decrease even when reducing up to 23% of FPG tests (corresponding to measuring FPG only for patients with BMI over 28, cInfoGain threshold 0.05). Figure 4 depicts for each cInfoGain threshold, the rate of change in AUC as a function of the percentage of records that did not pass that specific cInfoGain threshold. The percent of AUC change is defined as the distance of the AUC from the minimal AUC divided by the distance of the maximal AUC from the minimal AUC.

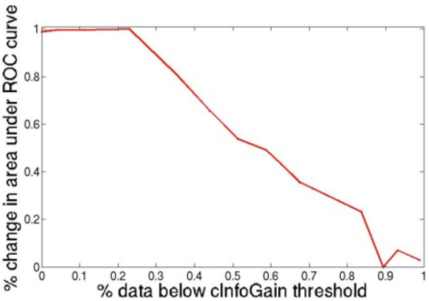

Figure 4. Area under ROC curve for different cInfoGain thresholds, and the corresponding percent of records that did not pass that cInfoGain threshold

3. Discussion

In the type 2 diabetes study data, it is clear that the predictive model using both BMI and FPG is significantly better than the model using only BMI (this is also evident in previous diabetes studies [7]). However, according to the new framework, a considerable number (23%) of the FPG tests can be avoided while maintaining the high predictive power of the full model. We demonstrated the power of our framework on a toy example that involves rather simple tests and suffers from small dataset limitations. Nevertheless the new framework is particularly valuable when the additional test requires invasive or risky procedures or when there are financial limitations and patients must be prioritized. Moreover, avoiding unnecessary examinations can decrease the probability of false positive results and further unnecessary examinations and treatment [8] and enable assessment at point of care. Future work will focus on the order of examinations and on combining a cost function. In addition richer non-Gaussian distributions can be considered for estimating the joint distributions, as well as corrections for finite sample effects [9].

References

[1] Saria S, Rajani AK, Gould J, Koller D, Penn AA. Integration of early physiological responses predicts later illness severity in preterm infants. *Science Translational Medicine* 2010;2(48):48-65

[2] Quinlan JR. Induction of decision trees. *Machine Learning* 1986 Mar;1(1):81-106

[3] Fonarow GC, Adams KF, Abraham WT, Yancy CW, Boscardin WJ. for the ADHERE Scientific Advisory Committee SG. Risk stratification for in-hospital mortality in acutely decompensated heart failure. *JAMA: The Journal of the American Medical Association* 2005 Feb;293(5):572 -580

[4] Cover TM, Thomas JA. Elements of information theory. New York: Wiley; 1991.

[5] Laucevicius A, Kasiulevicius V, Jatuzis D, Petrulioniene Z, Ryliskyte L, Rinkuniene E, Badariene J, Cypiene A, Gustiene O, Slapikas R. Lithuanian High Cardiovascular Risk (LitHiR) primary prevention programme – rationale and design. *Seminars in Cardiology*, 2012;18(3): In Press.

[6] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44(3):837-845

[7] Schwarz PEH, Bornstein SR, Hanefeld M. Elevated fasting glucose levels predicts IGT and diabetes also in middle-age subjects. *Diabetes Research and Clinical Practice* 2007 Jul;77(1):148-150

[8] Brewer NT, Salz T, Lillie SE. Systematic review: the long-term effects of false-positive mammograms. *Ann. Intern. Med.* 2007 Apr;146(7):502-510

[9] Slonim N, Atwal GS, Tkacik G, and Bialek W. Estimating mutual information and multi-information in large networks, http://arxiv.org/abs/cs.IT/0502017, 2005.