Data-Efficient Information-Theoretic Test Selection

Marianne Mueller[1], Rómer Rosales[2], Harald Steck[2], Sriram Krishnan[2], Bharat Rao[2], and Stefan Kramer[1]

[1] Technische Universität München, Institut für Informatik, 85748 Garching, Germany
[2] IKM CAD and Knowledge Solutions, Siemens Healthcare, Malvern PA 19335, USA

Abstract. We use the concept of conditional mutual information (MI) to approach problems involving the selection of variables in the area of medical diagnosis. Computing MI requires estimates of joint distributions over collections of variables. However, in general computing accurate joint distributions conditioned on a large set of variables is expensive in terms of data and computing power. Therefore, one must seek alternative ways to calculate the relevant quantities and still use all the available observations. We describe and compare a basic approach consisting of averaging MI estimates conditioned on individual observations and another approach where it is possible to condition on all observations at once by making some conditional independence assumptions. This yields a data-efficient variant of information maximization for test selection. We present experimental results on public heart disease data and data from a controlled study in the area of breast cancer diagnosis.

1 Information Maximization for Medical Test Selection

Consider a collection of patient records containing data generally indicative of various clinical aspects associated to the patient, e.g., patient demographics, reported symptoms, results from laboratory or other tests, and patient disease/condition. Let us define each single element (source) of patient data as a random variable V_m, e.g., patient age (demographics), presence of headache (symptom), blood pressure (test result), or occurrence of diabetes (disease). In this paper, we consider the set of variables $V = \{V_1, \ldots, V_M\}$ and we assume each variable to be discrete with a finite domain. For a given patient, a few of these variables may have been observed while others may have not. In some cases, we know such values of the variables with some probability.

Let us now consider the time when a diagnosis needs to be made for a specific patient, for which some of these variables have been observed, denoted as background attributes $Z_i \in V$, for $i \in \{1, ..., k\}$; e.g., demographic information and patient symptoms. The remaining $M - k$ variables denoted X_j have not been observed yet; e.g., the result of a lab test. Finally, for the patient, let Y denote a variable in which we are ultimately interested, but that we have not observed, such as the occurrence of cancer. This variable may not be observable directly

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 410–415, 2009.

or it may be observable at a high cost. Thus, we prefer to rely on other sources of information and try to infer the value of this variable.

Formally, we want to optimize $X^* = \texttt{arg}\max_j I(X_j, Y|Z_1 = z_1, \ldots, Z_k = z_k)$, where the quantity I is the mutual information (MI) between our variable of interest Y and the (test) variables whose values we could potentially obtain X_j, conditioned on the fact that we already know $Z_1 = z_1, \ldots, Z_k = z_k$. Using the definition of MI [1] and the shorthand $\mathbf{z} = (z_1, \ldots, z_k)$ to denote the assignment of multiple variables, the function of interest is:

$$I(X_j, Y|\mathbf{z}) = \sum_{x_j} \sum_{y} P(x_j, y|\mathbf{z}) \cdot \log \frac{P(x_j, y|\mathbf{z})}{P(x_j|\mathbf{z}) \cdot P(y|\mathbf{z})} \tag{1}$$

Assuming the variables follow a multinomial joint probability distribution that can be reliably estimated, this problem can be solved by testing each of the potential candidate variables individually to see which provides the most information.

In order to arrive at a diagnosis with high certainty, one observation may not be enough. In some cases, we may be allowed to observe more than one variable. Thus, this process can be repeated iteratively until a user-defined precision or confidence level is achieved. For example, this limit can be defined in terms of the amount of information that is left in the variable of interest (entropy). Once a variable X_i has been tested and observed, it can be incorporated as part of the background knowledge and the maximization problem is updated: $\texttt{arg}\max_{j \neq i} I(X_j, Y|x_i, \mathbf{z})$.

Ideally, the quantity to optimize is MI conditioned on all available data or observed variables. However, in practice this may be difficult because in general the conditional joint probabilities $P(x_j, y|z_1, z_2, \ldots, z_k)$ cannot be properly estimated from limited data for large k. The data required to properly estimate the conditional joints may grow exponentially with k. In addition, the number of unobserved variables plays an important role in terms of computational complexity (which can also grow exponentially with the number of unobserved variables).

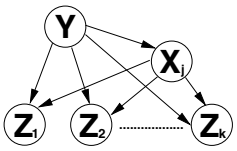
2 Combining All Available Background Information

Since with limited data our estimate of $P(x_j, y|z_1, z_2, \ldots, z_k)$ will not be accurate, we seek a *data-efficient* method – one whose data requirements do not grow exponentially with k – to compute or approximate P and thus be able to use all available background information to decide what test should be chosen.

2.1 Information Averaging

A simple heuristic consists of averaging the information conditioned on each background variable Z_i separately, $I(X_j, Y|\mathbf{z}) \approx \frac{1}{k} \cdot \sum_{i=1}^{k} \sum_{z_i} \alpha(z_i) I(X_j, Y|z_i)$, where $\alpha(z_i)$ is equal to the prior $P(z_i)$ if z_i is not observed and $\alpha(z_i) = 1(z_i)$ if observed. $1(z_i)$ is equal to one if z_i is the observed value for Z_i and zero otherwise. We will use this as our baseline method.

Table 1. Bayes net and approximations of probability distribution and entropy given $\mathbf{z}$



	Averaging	Joint Model (Q)
$P(y\|x_j,\mathbf{z})$	$\frac{1}{k}\cdot\sum_{i=1}^{k} P(y\|x_j,z_i)$	$\frac{P(x,y)\cdot\prod_{i=1}^{k} P(z_i\|x,y)}{\sum_{y\in Y} P(x,y)\cdot\prod_{i=1}^{k} P(z_i\|x,y)}$
$H(Y\|x_j,\mathbf{z})$	$\frac{1}{k}\sum_{i=1}^{k} H_P(Y\|x_j,z_i)$	$H_Q(Y\|x_j,\mathbf{z})$

One way to think about this heuristic is to consider k independent models involving Y, X_j, Z_i of the form $P(y, x_j, z_i) = P(z_i|x_j, y)P(x_j, y)$. MI can be computed for each model separately given the observed data. Similar to model averaging, the MI of X_j about Y can be found by averaging the individual information values, giving rise to the above equation.

2.2 Exploiting Conditional Independence Assumptions

We have so far assumed a very general model of the data, specifically a full multinomial distribution with a large number of degrees of freedom. However, if we relax this assumption, we can find a family of models for which the computation of the mutual information of interest is computationally and data efficient. These models make stronger conditional independence assumptions (simpler joint models), so that when the z_i's are observed, the computation of $I(X_j, Y|\mathbf{z})$ does not have large data requirements.

We consider the Bayes net in Table 1, where the background information Z_i depends on the test result X_j and the actual disease status Y. This network should not be viewed as reflecting causality, but simply as a statistical model. We obtain a new probability distribution Q:

$$Q(x_j, y, \mathbf{z}) = P(x_j, y) \cdot \prod_{i=1}^{k} P(Z_i = z_i|x_j, y). \tag{2}$$

The above Bayes network defines another multinomial model with the particular advantage that for the MI computations above, it only requires computing joint distributions of at most three variables, even if we do not know what variables Z_i will be observed beforehand. This is beneficial since we indeed do not know what variables will be observed for a particular case (patient).

For learning the model, we are interested in finding $P(z_i|x_j, y)\forall i \in \{1, ..., k\}$ and $P(x_j, y)$ since these distributions fully define the model. Using the maximum likelihood criterion we can see that: $P(z_i|x_j, y) = \frac{count(Z_i=z_i, X_j=x_j, Y=y)}{count(X_j=x_j, Y=y)}$ and $P(x_j, y) = \frac{count(X_j=x_j, Y=y)}{count(X_j=any, Y=any)}$. Inference can be performed efficiently also [2].

In summary, we have found a method that allows us to use all the available background information for deciding what test to perform. For this we have relaxed the model assumptions and use a class of models whose data requirements

are more practical. Inference and learning are computationally efficient in these models as well.

3 Validation and Results

First, we test our approaches on public heart disease data (HD) consisting of 920 patients and 14 attributes (two background attributes, 11 observable test attributes, and one binary class attribute).[1] On this dataset each patient is an instance. The variable of interest Y expresses if the patient has heart disease (411 patients) or not (509 patients). In a second step, we perform experiments on the breast cancer diagnosis data (BC) described in detail in the long version of this paper [2] (16 background attributes, 4 observable test attributes, and one binary class attribute). Here, we use lesions as instances. From 132 patients we analyze 216 biopsied lesions, 134 malignant and 82 benign. We use the biopsy result as the variable of interest Y. Note that the number of data points is small given the dimensionality of the data, which is quite common in controlled medical studies where the cost of data gathering is usually high.

We validate the two approaches considered above for each instance in a leave-one-out validation.[2] First, we determine for each instance which test X^* maximizes the information given the patient specific attributes $\mathbf{z}$. After observing the selected test $X^* = x_j$, we decide for the most likely diagnosis $y^* = arg\max_{y \in Y} P(y|x_j, \mathbf{z})$. We say the instance is *correctly assessed*, if y^* equals the actual state of disease of the instance, and *incorrectly assessed*, if the diagnosis was different. This accuracy measure could be easily improved by training a classifier on the features $\{X_j, Z_1, \cdots Z_k\}$ and predicting Y for each test instance.

The two approaches provide two different estimations of $I(X, Y|\mathbf{z})$. Our validation criteria (correctness, certainty, and information gain) additionally require the computation of $P(y|x_j, \mathbf{z})$ and the conditional entropy $H(Y|x_j, \mathbf{z})$. The discussion in Section 2 showed that it is not possible to compute this for large k. Therefore, we use the approximations shown in Table 1. H_P (H_Q) denotes the entropy of the probability distribution P (Q). To get a better grading of the results of our approaches, we compare it to two different ways of test selection. We evaluate: selecting a test *proposed* by information maximization, selecting one of the possible tests at *random* and selecting the *best*[3] test. Because of the different approximations (see Table 1), the performance of the *random* and *best* methods differs for each approach.

Correctness: Table 2 compares the ratio of instances that were assessed correctly: On both data sets it holds that if we select an examination at random,

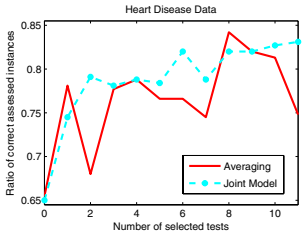
[1] `http://archive.ics.uci.edu/ml/datasets/Heart+Disease`

[2] We only consider instances with complete information on the test attributes as test instances (278 for HD and 138 for BC) because it is not possible to evaluate the cases where the result value of the selected test is missing.

[3] We determine the performances of all tests beforehand and select the test with the best performance (note that this is only possible in this controlled experiment).

Table 2. Ratio of correctly assessed instances. On the left: when iterating the test selection

Ratio of correctly assessed instances						
	Heart Disease			Breast Cancer		
	proposed	*best*	*random*	*proposed*	*best*	*random*
Av.	0.781	0.996	0.681	0.657	0.846	0.696
J.M.	0.745	0.989	0.665	0.748	0.944	0.691



in two thirds of the cases the decision about the status of disease is correct. On the HD data, selecting a test by information maximization leads to an improvement of about one tenth. On the BC data, selecting a modality by information maximization leads to an improvement in the second approach where we assume a simple joint model. Hence, the latter appears to give a more useful estimate of $I(X, Y|\mathbf{z})$. The difference of the two approaches on the HD data only becomes apparent when iterating the test selection process and adding the previously obtained test results as background knowledge, otherwise we have only two background attributes (Table 2). The second approach performs more stable and mostly outperforms the first approach.

Certainty of Decision: Results [2] show that on both data sets the certainty for the correctly classified lesions is higher (lower entropy) than for the incorrectly classified lesions. On the BC data, the difference is large for the joint model approach, but for model averaging these quantities are very similar. This indicates that the proposed joint model is better suited at modeling the information content in the test/decision variables.

Actual Information Gain: We compare the entropy of $P(Y|\mathbf{z})$ before performing a test with the entropy of $P(Y|x_j, \mathbf{z})$ after observing the test result $X_j = x_j$ (see [2]). Comparing both approaches to random, the joint model approach achieves a better result than the baseline approach for both data sets.

4 Conclusion

We have employed the concept of MI to address the problem of choosing tests efficiently[4]. We applied this to a problem in medical diagnosis. While MI is a well-understood concept, it is hard to calculate accurately for general probability models in practice due to small datasets and the underlying computational complexity. We have experimentally shown how certain model assumptions can help circumventing these problems. Making these assumptions, we obtain a

[4] Due to lack of space, the reader is referred to [2] for a discussion of related work.

comparatively data-efficient variant of test selection based on information maximization. Results indicate that the proposed joint model outperforms the information averaging approach by comparing the performance of each approach relative to a *random* and a *best* selection.

References

1. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley Interscience, Hoboken (1991)
2. Mueller, M., Rosales, R., Steck, H., Krishnan, S., Rao, B., Kramer, S.: Data-Efficient Information-Theoretic Test Selection, Technical Report TUM-I0910, Institut für Informatik, TU München (2009)