Integrating Non-Animal Test Information into an Adaptive Testing Strategy – Skin Sensitization Proof of Concept Case

Supplementary File C

BN construction details

Transforming training set into discrete variables
Strong and extreme sensitizers were pooled into one class that resulted in a balanced distribution of chemicals in each class. There were 26% non-sensitizers (NS), 22% weak (W), 30% moderate (M) and 22% strong including extreme (S). The continuous input variables were discretized by first applying a decision tree algorithm (Yuan and Shaw, 1995) with LLNA results as the target. Next, after a visual inspection, one additional state was added at the inflection point of the variable distribution function if it was not done automatically.

Learning structure of the latent variables
The data that were significantly dependent according to the t-test were clustered. The clustered manifest variables were used to construct latent variables local networks with EQ algorithm (Munteanu and Bendou, 2001). This learning step generated structure and probabilistic relationships between the manifest variables and respective latent variable.

Missing data imputation
For BN there is no direct way to perform imputation and one must decide on the sampling process outside the BN. Given that the patterns of missing data were not evenly distributed in the training set used in this study, a local imputation was performed per latent variable. First, local datasets associated with a latent variable were constructed from records for which there were no missing data. Imputation was performed separately in each cluster by replacing missing data with the values which were randomly sampled from the corresponding manifest variable marginal distributions of latent variable probability distribution (Gelman et al., 2003). Imputation transformed the data set with missing data into a complete dataset.

Learning the final structure of the network
Final structure of the network learning required connecting latent variables with the target and revealing potential direct connections between manifest variables and target that would provide information to explain the target in addition to one already contained in the latent variable. The final network's structure was learned from the complete dataset including the target variable as well as latent variables by EQ algorithm (Munteanu and Bendou, 2001).

Elucidation of CPTs – parameter learning
In order to parameterize the Bayesian network it is necessary to specify for each arc its conditional probability tables (CPTs). CPT is a multinomial distribution of a variable representing a node for each combination of parents' values. Knowing the structure of the graph and data attached to the nodes of the graph we learned CPTs by recursively applying Bayes' rule.

Methodology to guide testing

The framework can guide adaptive testing strategy based on information gain calculations. It uses information – theoretic concept of Value of Information and One step look – ahead hypothesis mutual information driven approach to identify the test that has the highest potential to refine the hypothesis variable. Once the evidence in this test is calculated the mutual information indices are recalculated for all the possible tests for which we do not have results, and again a test with the highest value is chosen as optimal to conduct.

Mutual Information (MI) is a measure of the dependence between two variables It quantifies the stored information in one variable about another variable. Mutual Information measures the general dependence while the correlation function measures the linear dependence, and Mutual Information is a more generalized quantity than the correlation function to measure the dependence. Entropy and Mutual Information are information – theoretic metrics measures used in VoI. Entropy is a measure of randomness and can be used as a measure of uncertainty in the distribution of a hypothesis variable Y, H(Y). Analyzing entropy changes as the new evidence is provided allows identifying the most informative information. Change in the entropy is measured as Mutual Information MI equal to MI(X, Y) = H(Y)-H(Y|X) where H (Y|X) is a conditional entropy of Y given an observation X. The Mutual Information MI(X, Y) is a measure of the information shared by X and Y, equivalent in reduction of entropy of Y from observing X and if Y is the hypothesis variable a measure of the value of observing X. Formally, the Mutual Information of two discrete random variables X and Y is defined as:

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p_1(x)\,p_2(y)}\right)$$

where p(x,y) is the joint probability distribution function of X and Y, and p1(x) and p2(y) are the marginal probability distribution functions of X and Y, respectively.

One step look – ahead hypothesis driven VoI in BNs amounts to computing the mutual information MI(X, Y) for all possible observations X and choosing the one that has the highest MI with the target variable Y. In the paper we use relative to the hypothesis MI, specifically MI(X, Y)/H(Y) that is expressed in %.