

Learning Causal Models for Noisy Biological Data Mining: An Application to Ovarian Cancer Detection

Ghim-Eng Yap and Ah-Hwee Tan

School of Computer Engineering
Nanyang Technological University
Nanyang Avenue, Singapore 639798
{yapg0001, asahtan}@ntu.edu.sg

Hwee-Hwa Pang

School of Information Systems
Singapore Management University
80 Stamford Rd, Singapore 178902
hhpang@smu.edu.sg

Abstract

Undetected errors in the expression measurements from high-throughput DNA microarrays and protein spectroscopy could seriously affect the diagnostic reliability in disease detection. In addition to a high resilience against such errors, diagnostic models need to be more comprehensible so that a deeper understanding of the causal interactions among biological entities like genes and proteins may be possible. In this paper, we introduce a robust knowledge discovery approach that addresses these challenges. First, the causal interactions among the genes and proteins in the noisy expression data are discovered automatically through Bayesian network learning. Then, the diagnosis of a disease based on the network is performed using a novel error-handling procedure, which automatically identifies the noisy measurements and accounts for their uncertainties during diagnosis. An application to the problem of ovarian cancer detection shows that the approach effectively discovers causal interactions among cancer-specific proteins. With the proposed error-handling procedure, the network perfectly distinguishes between the cancer and normal patients.

Introduction

In the year 2007, 22,430 American women are expected to be diagnosed with ovarian cancer, and 15,280 women could die from it (American Cancer Society 2007). As symptoms surface only in the advanced stages, the 5-year survival rate is just 45%. The chances of survival rise to 93% if the cancer is diagnosed early, but just 19% of cases are detected at early stages. There is hence a strong association between the high morbidity and the lack of a reliable early screening method.

Based upon the hypothesis that disease-specific proteins, or *proteomic biomarkers*, might be secreted into the blood stream from pathological changes in affected organs, recent discovery of differentially expressed proteins in ovarian cancer patients looks promising. Using a surface enhanced laser desorption and ionization time-of-flight (SELDI-TOF) mass spectroscopy to profile the proteins in the patients' sera, Petricoin *et al.* (2002) describe a technique that could identify ovarian cancer patients with perfect sensitivity (cancer-class recall) and a high specificity (normal-class recall) of 95%. Their approach has since been applied for the diagnosis of prostate (Wellmann *et al.* 2002) and other types of cancers.

Copyright © 2007, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Indeed, being able to provide an almost perfect sensitivity and specificity is a necessary requirement for cancer detection methods. As the prevalence of each cancer in the population is relatively low, a high predictive value is necessary to minimize incorrect diagnoses. This has been the focus of prior researches (e.g. Petricoin *et al.* 2002, Zhu *et al.* 2003).

However, being able to diagnose accurately on particular data sets does not guarantee that a learned classifier performs well on the general population. In fact, it is found that classifiers that are learned from different ovarian cancer data sets differ greatly, such that each classifier predicts accurately on its own data set, but performs poorly on another (Baggerly, Morris, & Coombes 2004). In addition to having high accuracy on representative samples, a model that describes biologically plausible causal interactions among genes and proteins is far more likely to be generalizable. Hence, a second important requirement of a cancer detection method is that it should be able to encode the causal feature interactions, in a way that we can interpret, comprehend, and further verify.

There are a number of challenges in satisfying the two requirements. Existing protein analysis techniques depend on mass spectral data streams that are infested with substantial electronic noise and chemical noise due to contaminants and matrix effects (Petricoin *et al.* 2002). Likewise, the gene expression data from high-throughput methods like DNA microarrays are extremely noisy (Friedman *et al.* 2000). For robust diagnoses, there is a need for an error-handling procedure that reliably discovers the unknown error rates underlying the observations and accounts for these when predicting.

To fulfil the second requirement of causal interpretability, a thorough knowledge discovery method for disease diagnosis has to examine all the candidate features so as not to overlook the handful of potential markers. In addition, there is a need to reliably discover the causal interactions among the features, so that the intrinsic mechanism of diagnosis can become clear. Ideally, the method must be able to learn the causal interactions from data without external information, but it should also allow for leveraging of prior knowledge in learning, as well as refinement by experts after construction.

In this paper, we introduce a knowledge discovery approach that can address these challenges. After applying a dimensionality reduction method, an automatic Bayesian network learning algorithm discovers the underlying causal interactions among features. The learned Bayesian network

causal model captures these feature interactions in a graph that can be readily inspected and refined by human experts. For robustness against noise, we introduce a novel procedure that automatically identifies the erroneous features and accounts for their uncertainties during predictions. Although we describe the proposed approach's usefulness within the noisy biological domain in this paper, the same approach can be effectively applied to analyze any noisy data in general.

We evaluate the mechanism by applying it to ovarian cancer detection. By utilizing our proposed error-handling procedure, the network achieves a perfect specificity and sensitivity on unseen data. At the same time, in contrast to the incomprehensible "black-box" nature of other prediction models, the learned network offers an interpretable model of the causal interactions among the proteins related to the cancer.

The rest of this paper is organized as follows. We start off by describing the ovarian cancer data set we have analyzed, and the related literature and prior results. Next, we describe the key phases of our knowledge discovery approach. Following that, we present the major evaluation results. Finally, we conclude this paper with a discussion of future work.

The Ovarian Cancer Data Set

We analyze our proposed mechanism based on the raw mass spectra data set that is publicly-available online at the web site of the United States National Cancer Institute's Center for Cancer Research's *Clinical Proteomics Program Data-bank* (<http://home.ccr.cancer.gov/ncifdaproteomics>) (Ovarian Data Set 8-7-02). The experiment aims to find proteomic patterns in blood serum that distinguish patients with ovarian cancer from those without. This is especially important for early detection in women who are at high risk of ovarian cancer due to a family or personal history of cancer, and for women with genetic predisposition to ovarian cancer due to abnormal *BRCA1* and *BRCA2* genes (Liede & Narod 2002).

The data set comprises of 162 serum profiles from ovarian cancer patients, and 91 profiles from normal subjects who do not have cancer. The ovarian cancer patients consist of 28 stage I patients, 20 stage II patients, 99 stage III patients, 12 stage IV patients, and 3 patients who were at an unspecified stage. The data set had been collected from blood samples of the human subjects using the Ciphergen WCX2 ProteinChip array. The samples were prepared with a robotic instrument, and the raw data without any baseline subtraction was posted for download. The profile for the subjects comprises 15,154 distinct mass-to-charge ratios (M/Z values) of intensities that range from 0.0000786 to 19995.513 (Sorace & Zhan 2003).

Related Work

Lilien, Farid, & Donald (2003) develop the Q5 algorithm for classifying the SELDI-TOF serum spectra. They use Principle Component Analysis (Duda & Hart 1973) to reduce the dimensionality before classifying using Linear Discriminant Analysis (Fisher 1936). The method's prediction confidence is defined in a normal Gaussian distribution centered at each of the class means, where a higher threshold allows for more confidence at the expense of fewer classifiable samples. For the same ovarian data set, they report that a particular thresh-

old exists for which Q5 classifies perfectly, but it is not clear how this crucial threshold value could be determined in general. More importantly, principle component analysis generates new features that cannot be related back to the proteins.

Sorace & Zhan (2003) have used the Wilcoxon nonparametric test to find M/Z values having the largest differences between cancer and normal sera, and stepwise discriminant analysis to develop rules for diagnosis. For this ovarian cancer data, they have presented two rules that can classify perfectly. However, their rules involve proteins with M/Z values as low as only 2.792 and 2.823. As the authors and later Baggerly, Morris, & Coombes (2004) have clearly highlighted, this can be a major point of concern since it is very difficult to offer any biological explanation for the observed difference in such a low M/Z region. As they have not learned the protein interactions, their method cannot be used to examine these features' relations to other proteins and ovarian cancer.

Li & Wong (2003) have compared the performance of two rule-induction algorithms, the C4.5 (Quinlan 1993) and their own PCL classifier, based on this same set of ovarian cancer data. They have reported that the decision tree that is learned by C4.5 misclassifies ten out of about 25 test samples, while their PCL classifier, which uses multiple significant rules as a committee, misclassifies only four samples. Their induced rules are also more interpretable compared to the functions that are defined within kernel-based methods. However, like the C4.5, their rule-based classifier model is susceptible to the noises in the protein values during diagnoses, and it cannot encode the causal interactions among important proteins.

Knowledge Discovery from Noisy Data Set

An important purpose in the analysis of biomedical data is to discover interactions or causal relationships among features. Not only can the discovered protein interactions be biologically informative, their dependencies can be effectively exploited for robust prediction of diseases like ovarian cancer.

Phase 1 - Data Preprocessing

Normalization To ensure comparability across the spectra, the intensity values are normalized according to the procedure outlined by the Clinical Proteomics Program Data-bank and described in Sorace & Zhan (2003). The normalization is done over all the 253 examples for all the 15,154 mass-to-charge M/Z identities using the following formula:

$$NI = \frac{RI - Min}{Max - Min} \quad (1)$$

where NI represents the normalized intensity, RI represents the raw intensity, and Min and Max refer to the minimum and maximum intensity of the pooled examples. After normalization, the intensity values will be between 0 and 1.

Dimensionality Reduction and Discretization We adopt entropy-based discretization (Fayyad & Irani 1993) in this work as it is known to be effective for: (i) ignoring trivial variations in intensity due to noise, and (ii) sifting through high-dimensional data to select important features that can distinguish samples from different classes. The method has

been successfully used for preprocessing high-dimensional biomedical data (Li, Liu, & Wong 2003; Tan & Pan 2005).

Entropy-based discretization combines the entropy-based splitting criterion of the C4.5 decision tree (Quinlan 1993) with a minimum description length stopping criterion. It recursively determines an optimal cutting point for each feature dimension that maximizes the separation of the classes. Features with no cutting points are deemed not as important and thus can be discarded. In this way, the method effectively reduces the feature dimensions and converts the continuous mass-to-charge (M/Z) markers into discrete features.

Phase 2 - Learning Bayesian Network from Data

The causal interactions among a set of variables can be modelled in the directed acyclic graphical model of a Bayesian network (Pearl 1988). The variable dependencies are captured qualitatively by the network's structure, i.e., the arcs linking the variables (*nodes*), and quantitatively by the table of conditional probabilities associated with each node. In this work, we adopt the *CaMML* program (Wallace & Korb 1999) for supervised Bayesian network learning from data.

CaMML stochastically searches over the entire space of causal models to find the best model that maximizes a Minimum Message Length (MML) posterior metric (Korb & Nicholson 2004). For each real model it visits, it computes a representative model and counts only on these representatives to overlook the trivial variations. The MML posterior of each representative is computed as the total MML posterior of its members. This total posterior approximates the probability that the true model lies within the MML equivalence class of the representative. The best model is hence the representative model with the highest MML posterior.

The learned Bayesian network is promising for analyzing interacting quantities such as expression data (Friedman *et al.* 2000). Firstly, Bayesian networks effectively represent causal dependencies among multiple interacting proteins. Secondly, they describe local interactions well, as the value of each node directly depends upon the values of a relatively small number of neighboring proteins. Finally, Bayesian networks are probabilistic in nature, and hence are capable of handling the noise by taking account of the uncertainty in the evidence presented for different network nodes.

Phase 3 - Discovering Erroneous Markers

A set of data is erroneous if it involves markers that carry some probability of being incorrect in their observed values. An error rate of e for a marker m implies that observations on m are wrong $e*100\%$ of the time. In practice, we do not know how many and which of the markers are erroneous, so our procedure must be flexible enough to discover multiple erroneous markers, each suffering an unknown rate of error.

As summarized in Algorithm 1, the discovery procedure takes in an erroneous data set, and a Bayesian network that concisely represents the data's joint probability distribution while ignoring most of the noises in the markers. Predicting with this model, we identify the top-most erroneous marker by using the proportion of misclassified training examples as an estimate for each marker's probability of error. Accounting for the estimated errors on this top marker by entering it

Algorithm 1 Error Discovery Procedure

Input: Training data (D) containing erroneous records, Bayesian network (BN) learned on D , and error threshold (t).
Output: Erroneous markers (M), and their est. error rates (R).

Step 1: Identify the top erroneous marker m_{top} .

Set the first marker, m_1 , as m_{top} .

for each record in D **do**

 Cover-up m_1 and predict its value using BN .

 Compute $P_{err}(m_1)$ as fraction of D that m_1 is misclassified.

 Set top to $P_{err}(m_1)$.

for each remaining marker m_i **do**

for each record in D **do**

 Cover-up m_i and predict its value using BN .

 Compute $P_{err}(m_i)$ as fraction of D that m_i is misclassified.

if $P_{err}(m_i) > top$ **then**

 Set m_{top} to m_i .

 Set top to $P_{err}(m_i)$.

if $top \geq t$ **then**

 Add m_{top} to M and add top to R .

else

return M and R as empty sets.

Step 2: Identify the remainder of sets M and R .

while \exists marker $\notin M$ **do**

 Estimate likelihoods of values in M using R .

 Identify the next-most erroneous marker m_{next} .

if $P_{err}(m_{next}) \geq t$ **then**

 Add m_{next} to M and update R .

else

 Update R ; Break.

return the non-empty sets M and R .

as an uncertain (*likelihood*) evidence, we look for the next-most erroneous marker, and so on until the estimated error rate falls below a predefined threshold. In this way, the procedure discovers the error markers and estimates their error rates, based on the intuition that the noisiest markers should also be the markers that are least consistent with the learned network's joint distribution. The error threshold serves to filter out the small degrees of natural randomness in the protein behaviors. In our experience, a suitable threshold value for protein spectra analysis is less than 0.1, such that as many of the possibly erroneous markers can be identified as possible.

Phase 4 - Predicting under Uncertainties

Predicting using Learned Bayesian Network We predict the disease category (e.g. cancer, normal) with the learned Bayesian network by feeding in the collected protein expression levels, and then allowing the beliefs within the network to be updated. The category with the highest posterior probability is the network's inference, or diagnosis. This process of prediction using Bayesian network is summarized below:

Step 1: For an unseen sample, present the corresponding serum expression profile to the learned Bayesian network.

Step 2: Let the network update the posterior probabilities of all its nodes based on the evidence. We use the fastest known algorithm for exact general probabilistic inference in a compiled Bayesian network, called message passing

in a join (or “junction”) tree of cliques (Neapolitan 1990).

Step 3: Predict, for the given sample, the disease category (cancer, non-cancer) with the largest posterior probability.

Empirically, we observe that this Bayesian network inference mechanism is highly efficient. In the experiments, each prediction took less than 0.1 seconds on a PC.

Likelihoods Estimation We would expect the new, or the *unseen*, samples to be noisy as well, with noise characteristics similar to samples in the training data. In the normal prediction scenario, where we are not aware of possible errors in the marker values, we enter each reading given in the test sample as a *specific evidence*. However, when there are errors in the readings of the marker, entering these erroneous readings as specific findings would very likely result in wrong predictions. Fortunately, the Bayesian network allows us to specify such potentially erroneous evidence as *likelihoods* to reflect our uncertainty regarding the evidence.

We should take into consideration the *prior probability* for each possible value of an erroneous marker in estimating its likelihoods (Korb & Nicholson 2004). The prior probability for the value v_m of a marker m is the proportion of examples within the training data for which v_m is present. For instance, the prior probability for each of the possible values of marker “MZ261.88643” (hereafter referred as MZ261) can easily be computed from the training data, giving us an array of priors $P(v_m)$, where $v_m \in \{0, 1, 2, 3\}$. Now, let the observation on MZ261 be O . The likelihoods for O are then given by $\{\text{prob}(O | \text{MZ261}=\text{“0”}), \text{prob}(O | \text{MZ261}=\text{“1”}), \text{prob}(O | \text{MZ261}=\text{“2”}), \text{prob}(O | \text{MZ261}=\text{“3”})\}$.

For the example where we observe the MZ261 as “3”, suppose we are aware that this observation carries an estimated error rate of e . This means that each observation for MZ261 might be wrong $e * 100\%$ of the time. This error rate for the erroneous marker could be estimated using the training error rate discovered by our error discovery procedure. The likelihoods for $O:\{\text{MZ261}=\text{“3”}\}$ would then be computed as $\{P(0) * e * P(3), P(1) * e * P(3), P(2) * e * P(3), P(3) * (1.0 - e)\}$, where $P(v)$ denotes the prior of v . This is because the probability of observing the MZ261 as “3” when it is actually “0”, “1” or “2” is simply the probability of one of these three values being present in the sample but we wrongly record the marker as “3”. We can generalize the estimation of likelihoods from the error rate e as follows:

Likelihood of the observed $v = P(v) * (1.0 - e)$, and
Likelihood of any other value $v' = P(v') * e * P(v)$.

Entering an error rate of zero for a marker is equivalent to entering a specific finding, so the above likelihood formulation is appropriate even when the reading is made with a full certainty. By properly accounting for these confidence in the erroneous marker values, our automatic error discovery and likelihoods estimation procedure enables the learnt network to overcome noise and predict accurately on unseen data.

Experimental Validation

The ovarian data set contains 162 cancer and 91 normal samples. Similar to Sorace & Zhan (2003), we randomly select 81 cancer patients and 45 controls for training, leaving the

Table 1: The top 10 proteins selected in decreasing order of information gain.

Protein (M/Z)	Information Gain	Cutting Points
MZ261.88643	0.8190	0.3771,0.4011,0.4388
MZ262.18857	0.7627	0.4317, 0.5107
MZ245.24466	0.7381	0.4508
MZ244.95245	0.7381	0.4350
MZ245.8296	0.7219	0.3908
MZ245.53704	0.7219	0.5052
MZ244.66041	0.7202	0.2098, 0.3608
MZ246.12233	0.6903	0.3478
MZ435.07512	0.6693	0.3647
MZ434.29682	0.6650	0.4461, 0.5249

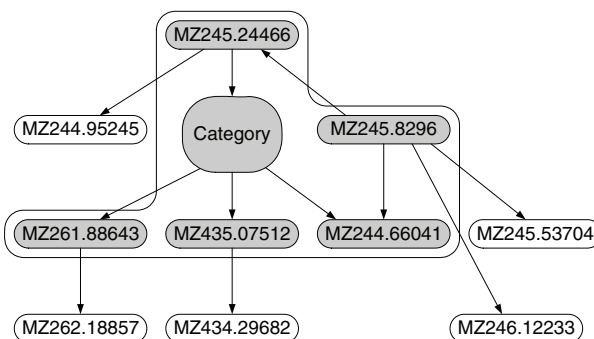


Figure 1: Bayesian network learned from the training data.

rest for testing. In this section, we first present the discretization results on which proteins are selected for further analysis. We then present our important results, on the discovered causal proteins interactions in the learned Bayesian network, and the approach’s efficacy at diagnosing for ovarian cancer.

Using the entropy-based feature selection and discretization method, only 4796 of the 15,154 M/Z values in the training data are partitioned into two to four intervals each, while there are no cutting points for the other attributes. This indicates that only $4796/15,154 = 31.6\%$ of the features can be considered as discriminatory and the rest are negligible. Deriving this much smaller set of informative proteins helps us to more efficiently identify the important causal interactions.

Next, we examine all the 4796 selected features and sort them in decreasing order of information gain, a formal measure that reflects the reduction in entropy that is obtained by splitting the training data on individual proteins. Perfect prediction on the training data is readily achieved using the top 10, top 15 and top 20 proteins, but based on just the top five proteins, two of the 45 control samples in the training set are misclassified. Therefore, we proceed to learn the causal interactions among all the top 10 proteins as listed in Table 1.

Figure 1 presents the Bayesian network that has been automatically learned from the training data based on the top 10 proteins with the highest information gains. The *Markov blanket* for the target node of *Category* has been demarcated, and composes of the *Category*’s parents, its children, and the parents of its children. In any Bayesian network, the *Markov*

Table 2: The discovered error information on the proteins. *Proteins are listed in decreasing order of est. error rate.*

Name of the Protein (<i>M/Z</i>)	The Estimated Error Rate
MZ244.66041	0.2063
MZ434.29682	0.1429
MZ261.88643	0.0952
MZ262.18857	0.0952
MZ435.07512	0.0476
MZ245.24466	0.0079
MZ244.95245	0.0079
MZ246.12233	0.0079
MZ245.53704	0.0
MZ245.8296	0.0

blanket of a node shields off the rest of the network from the node, and is all that is needed to predict the behavior of that node. Since each *M/Z* feature describes a serum protein, it should interest biologists that the disease’s category can be fully predicted by just the highlighted set of proteins.

It is interesting to note that this small set of proteins within the *Markov blanket* of our learned causal model in fact overlaps with the set of proteins that Sorace & Zhan (2003) have identified using stepwise discriminant analysis. Specifically, both “MZ261.88643” and “MZ435.07512” are found to possess an exceptionally strong diagnostic power in both experiments. Based on this finding, there is further reason to believe that these proteins are closely related to ovarian cancer.

We make use of this network for investigating whether our proposed error discovery and likelihoods estimation procedure can indeed contribute to more robust predictions. First, we employ this learned model to predict on the set of masked test data without applying our error-handling procedure. It is found that this learned model is reasonably accurate, misclassifying only one of the 81 cancer and one of the 46 normal samples. This gives a sensitivity (cancer-class recall) of 98.77%, a specificity (normal-class recall) of 97.83%, and a positive predictive value (cancer-class precision) of 98.77%.

We investigate whether the misclassifications can be corrected using our proposed error-handling procedure. First, we apply our proposed error discovery procedure of Algorithm 1 to identify the erroneous proteins and also obtain estimates of their error rates. Using a minimum error threshold of zero, the algorithm automatically discovers that eight out of the ten proteins have some evidence of errors (Table 2), although only the first few most-erroneous proteins have an estimated error rate that is above 0.10. This information on the individual protein’s error rate encompasses the important discovered knowledge about each measurement’s reliability.

Next, based on these estimated error rates, we compute the likelihoods estimates for each protein when entering its observed values into the Bayesian network during testing. Using the learned model presented in Figure 1, we apply our procedure to the same test samples. This produces a perfect classification for all 127 unseen samples. Details of the corrections in the two samples that are misclassified earlier are presented in Table 3. We can see that the beliefs of proteins

Table 3: Samples that are correctly classified only after applying our error discovery and likelihoods estimation procedure. *The proteins are in the same order as shown in Table 1.*

Protein values before error compensation	Class Label	Prediction (Prob.)
2, 0, 1, 1, 0, 0, 2, 0, 1, 2	Cancer	Normal (0.600)
0, 0, 1, 1, 1, 0, 1, 1, 0, 0	Normal	Cancer (0.564)
Protein values after error compensation	Class Label	Prediction (Prob.)
2, 0, 1, 1, 0, 0, 0, 0, 1, 2	Cancer	Cancer (0.832)
0, 0, 1, 1, 1, 0, 1, 1, 0, 0	Normal	Normal (0.619)

within the Bayesian network have been updated, to the extent that the most-probable protein values may differ from their findings. This results in the correct predictions, each of which has a higher probability, or a higher level of certainty.

We repeat the experiment for five rounds of ten-fold stratified validations. In each round, the samples are randomly divided into ten equal portions. For each fold, one portion is left out for testing, while we train on the remaining samples. By coupling entropy-based discretization with Bayesian network learning as we have proposed, the predictive accuracy, i.e., the average percentage of test samples that are correctly classified within the fifty splits, is 96.46%. We note that this already corresponds to just a single misclassification on average. With error discovery and likelihoods estimation, the accuracy improves further to 96.70%, with more of the splits producing perfect diagnostic sensitivity and specificity.

Apart from predictive accuracy, the direct interpretability of the discovered model of causal protein interactions is an important advantage of our proposed approach. For example, besides the information on the *Markov blanket*, we can also begin to understand the effects of errors in different proteins upon the diagnosis. With reference to Figure 1, the protein “MZ244.66041” is directly connected to the “Category” node that predicts the disease category, as well as to another protein “MZ245.8296”. From Table 2, we observe that these are respectively the most and least erroneous proteins. As such, our likelihoods estimation procedure would have entered the value of “MZ244.66041” with much lower certainty than “MZ245.8296”, allowing the latter to correct the former. True enough, we see from the first sample of Table 3 that a key reason for the correct prediction after error compensation is the correction in value of “MZ244.66041”.

Another important advantage of learning causal models is that explanations can be generated automatically from the learned Bayesian networks to aid in interpretation. In a companion paper (Yap, Tan, & Pang 2007), we have presented a novel approach that explains the Bayesian network’s diagnosis using a minimal set of features, by exploiting the fact that the conclusions of the diagnostic node can be completely explained with just the nodes in its *Markov blanket*. Empirical evaluations based on multiple real-world data sets show that, by focusing on the nodes within the Markov blankets, we are

able to generate high-quality explanations for the probabilistic inferences in learned Bayesian networks. The following example shows an explanation for the first sample in Table 3. This automatically-generated explanation clearly highlights the compensation for “MZ244.66041” during the diagnosis.

*BN predicts Cancer with probability 0.832 because
 MZ245.24466 is 1 with probability 0.968,
 MZ245.8296 is 0 with probability 1.0,
 MZ244.66041 is 0 with probability 0.600,
 MZ435.07512 is 1 with probability 0.993, and
 MZ261.88643 is 2 with probability 0.616.
 BN corrects MZ244.66041 from 2 to 0 because
 Given MZ245.8296 is 0,
 MZ244.66041 is 2 with probability 0.266, and
 MZ244.66041 is 0 with probability 0.600.*

Conclusion

We have presented a systematic mechanism for learning, extracting, and exploiting the knowledge of the causal interactions contained in noisy biological data for cancer detection. Based upon the strong statistical foundation of the Bayesian network, experimental results on a high-dimensional ovarian cancer data from the Clinical Proteomics Program Databank show that the proposed knowledge discovery approach contributes significantly to more robust predictive performance. Even with just ten proteins, the empirical results show that the normal and cancer sera could be perfectly distinguished.

In addition to predictive accuracy, the ability to generate interpretable knowledge is the key strength of our approach. In assigning each unseen sample to its most-probable category, the Bayesian network uses its probabilities as the basis for its confidence. Most importantly, the Bayesian network that is learned from data can be readily verified by, and integrated with, the prior knowledge from medical practitioners, biologists, and the other experts from different related fields.

Having a systematic approach to discover and to exploit causal interactions, our next step would be to work with the experts to interpret and validate the knowledge discovered by the system. Ultimately, we aim to find a diagnostic model for diseases like the ovarian cancer that not only is extremely sensitive and specific, but also describes biologically plausible interactions, as this is the only way for knowledge found from small data sets to generalize. Hopefully, this could lead to a better screening tool for women who are at high risk of ovarian cancer. This would form the core of our future work.

Acknowledgments

Ghim-Eng Yap is a graduate scholar of the Agency for Science, Technology & Research (A*Star). Hwee-Hwa Pang is partially supported by a grant from the Office of Research, Singapore Management University. This work is supported in part by the I²R-SCE Joint Lab on Intelligent Media.

References

American Cancer Society. 2007. *Cancer Facts & Figures 2007*. Atlanta: American Cancer Society.

Baggerly, K. A.; Morris, J. S.; and Coombes, K. R. 2004. Reproducibility of SELDI-TOF protein patterns in serum:

Comparing data sets from different experiments. *Bioinformatics* 20(5):777–785.

Duda, R., and Hart, P. 1973. *Pattern Classification and Scene Analysis*. New York, USA: Wiley.

Fayyad, U., and Irani, K. 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In *Procs. of ICML*, 1022–1029. Morgan Kaufmann.

Fisher, R. 1936. The use of multiple measures in taxonomic problems. *Annals of Eugenics* 7:179–188.

Friedman, N.; Linial, M.; Nachman, I.; and Pe’er, D. 2000. Using Bayesian networks to analyze expression data. In *Procs of RECOMB*, 127–135.

Korb, K. B., and Nicholson, A. E. 2004. *Bayesian Artificial Intelligence*. CRC Press.

Li, J., and Wong, L. 2003. Using rules to analyse biomedical data: A comparison between C4.5 and PCL. In *Proc. of WAIM*, 254–265. Chengdu, China: Springer.

Li, J.; Liu, H.; and Wong, L. 2003. Mean-entropy discretized features are effective for classifying high dimensional bio-medical data. In *Procs of SIGKDD Workshop on Data Mining in Bioinformatics*, 17–24. ACM Press.

Liede, A., and Narod, S. A. 2002. Hereditary breast and ovarian cancer in Asia: Genetic epidemiology of *BRCA1* and *BRCA2*. *Human Mutation* 20:413–424.

Lilien, R. H.; Farid, H.; and Donald, B. R. 2003. Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. *Journal of Computational Biology* 10(6):925–946.

Neapolitan, R. E. 1990. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. NY: John Wiley & Sons.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Petricoin, E. F.; Ardekani, A. M.; Hitt, B. A.; Levine, P. J.; Fusaro, V. A.; and et al. 2002. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 359:572–577.

Quinlan, J. R. 1993. *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Sorace, J. M., and Zhan, M. 2003. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* 4(24).

Tan, A.-H., and Pan, H. 2005. Predictive neural networks for gene expression data analysis. *Neural N.* 18:297–306.

Wallace, C. S., and Korb, K. B. 1999. Learning linear causal models by MML sampling. In Gamberman, A., ed., *Causal Models and Intell. Data Management*. Springer.

Wellmann, A.; Wollscheid, V.; Lu, H.; Ma, Z. L.; Albers, P.; and et al. 2002. Analysis of microdissected prostate tissue with ProteinChip arrays - a way to new insights into carcinogenesis and to diagnostic tools. *IJMM* 9:341–347.

Yap, G.-E.; Tan, A.-H.; and Pang, H.-H. 2007. Explaining inferences in Bayesian networks. *Forthcoming*.

Zhu, W.; Wang, X.; Ma, Y.; Rao, M.; Glimm, J.; and Kovach, J. S. 2003. Detection of cancer-specific markers amid massive mass spectral data. *U.S. PNAS* 100(25):14666–71.