Immune System Modeling with Infer.NET

Vincent Y. F. Tan, John Winn, Angela Simpson, Adnan Custovic*

Abstract

Graphical models allow scientific prior knowledge to be incorporated into the statistical analysis of data, whilst also providing a vivid way to represent and communicate this knowledge. In this paper we develop a graphical model of the immune system as a means of analyzing immunological data from the Manchester Asthma and Allergy Study (MAAS). The analysis is achieved using the Infer.NET [1] tool which allows Bayesian inference to be applied automatically to a specified graphical model.

Our immune system model consists firstly of a Hidden Markov Model representing how allergen-specific skin prick tests (SPTs) and serum-specific IgE tests (SITs) change over time. By introducing a latent multinomial variable, we also cluster the children in an unsupervised manner into different sensitization classes. For 2 sensitization classes, the children who are vulnerable to allergies and have a high probability of having asthma (22%) are identified. For 5 sensitization classes, children in the first cluster, those who are vulnerable to allergies, have an even higher probability of having asthma (42%). The second part of the model involves using the inferred sensitization class as a label and 8 exposure variables in a Bayes Point Machine. Using multiple permutation tests, we conclude that the level of endotoxins and gender have a significant effect on a child's vulnerability to allergies.

1 Introduction

There are a variety of factors that lead to childhood asthma, including environmental, physiological and genetic factors. The complex interplay between these factors affects the probability that a child is deemed to have asthma. In this paper, we focus on the immune system and explore this interaction of factors by using a *graphical model* to incorporate prior clinical and other structured domain knowledge into the analysis of data. Our data is obtained from the Manchester Asthma and Allergy Study (MAAS) [1], a birth-cohort study in the Manchester and Stockport area. The main aim of this study is to identify risk factors for the development of asthma and allergies in children. We will observe that *Infer.NET* can indeed incorporate prior knowledge into the graphical model such that the statistical analyses produce meaningful and interpretable results.

There are two types of variables that are of interest to us, the physiological variables and the exposure variables. These are explored in two parts of our graphical model.

2 The Model and Inference

The first submodel is the *allergic sensitization* model, a *Hidden Markov Model* (HMM)[2]. At the core of the model are four binary variables for each allergen representing whether the child is sensitized to that particular allergen at different ages. They are linked together in a Markov chain and we aim to infer the transition probabilities, i.e. the probabilities of gaining and losing sensitization at various ages. Our measurements are the SPTs and SITs, which are also treated as binary. The Infer.NET tool is able to infer transition probabilities in the light of missing measurements and can compute posterior distributions over such missing values. We also infer the true positive and true negative rates of both the SPT and SIT.

The model includes a latent variable for each child representing the 'Sensitization Class' of the child. This class is a multinomial variable with between two and six states, and is used to select between different sets of transition probabilities. Hence, the variable serves as a means to cluster the children, in a completely unsupervised manner, into classes of children with different sensitization properties. For instance, if the variable has just two states, it is plausible that the children will be clustered into those who are vulnerable to allergies and those who are not. Finally, we link this sensitization class variable to an observed binary variable,

*Vincent Tan (vtan@mit.edu) is with the Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology and he performed this work at Microsoft Research Cambridge, UK (MSRC). John Winn (jwinn@microsoft.com) is with MSRC. Angela Simpson and Adnan Custovic ({angela.simpson, adnan.custovic}@manchester.ac.uk) are with the School of Medicine, University of Manchester.

[1] http://research.microsoft.com/mlp/ml/infer/infer.htm

[2] We refer to the reader to http://web.mit.edu/vtan/www/escience for detailed figures of the allergic sensitization and exposure models.

namely asthma. We would like to infer the probability that a child has asthma given that he/she is in a particular cluster. This submodel is a concrete example showing how *clinical knowledge* can be incorporated into a *graphical model* and how Infer.NET can be applied to the model to produce scientifically meaningful results.

The second submodel is the *exposure model*, a *Bayes Point Machine* (BPM) [2]. The eight features used as inputs to the model are daycare, cat and dog ownership, endotoxin, glucan, elder siblings, mean body mass index and gender. The target variable is the inferred binary sensitization class.

We used *Variational Message Passing* [3] and *Expectation Propagation* [4] as the inference algorithms for the allergic sensitization and exposure models respectively. The graphical models were constructed in Infer.NET by selecting appropriate conjugate prior distributions for the latent variables, instantiating an inference engine depending in the algorithm, and finally running inference to obtain the posteriors of the latent variables.

3 Results

The Infer.NET tool provides posterior distributions over all parameters and latent variables. For example, the true positive rates of the SPT and SIT were found to be 74.5 $\pm$ 0.64% and 89.3 $\pm$ 0.45% respectively, showing that the IgE test is more accurate than the skin test, which makes sense because the IgE test is based on blood samples. We also noted a decreasing trend in the probability of gaining sensitization for the inhaled allergens (mite, cat, dog and pollen). This is again plausible because more children gain sensitization to these allergens as they grow up.

For 2 clusters, Cluster 1 (resp. Cluster 2) contains children who have many (resp. few) allergies. Also, if a child is in Cluster 1 (resp. Cluster 2), the probability of having asthma is $26/118 = 22\%$ (resp. $9/357 = 2.5\%$). Thus, the children are effectively partitioned into those who are vulnerable to allergies and asthma and those who are not. We show the structure of the 5 cluster case in Fig 1. As can be seen, if we increase the cardinality of the sensitization class variable to 5, children in Cluster 1, who are sensitized to many allergens, have an even higher probability of having asthma ($19/45 = 42\%$). More interestingly, we can discover more latent structure in the 5 cluster case: The children are clustered into general sensitization, inhaled allergen sensitization and no sensitization, with mite-only and pollen-only sensitization being somewhat smaller subclusters. Because the gain and lose sensitization probabilities were time-varying, the vulnerable children can be further clustered into those who are sensitized early (before 3) and those who are sensitized later in life (after 3).

We inferred the BPM weights for the exposure model. We computed the mean of the weights divided by the standard deviation m_w/s_w. The two (out of eight) weights that had the highest $|m_w/s_w|$ were endotoxin ($|m_w/s_w| = 2.6$) and gender ($|m_w/s_w| = 3.4$). To validate the significance of these weights, we performed *permutation testing* on the exposure model by randomly permuting the labels 500 times. The false positive rate was $195/(8 \times 500) = 5\%$ and hence the false discovery rate was $5\% \times 8/2 = 20\%$. Thus, we can be 80% certain that increased endotoxin exposure and being a girl reduce one's vulnerability to allergies.

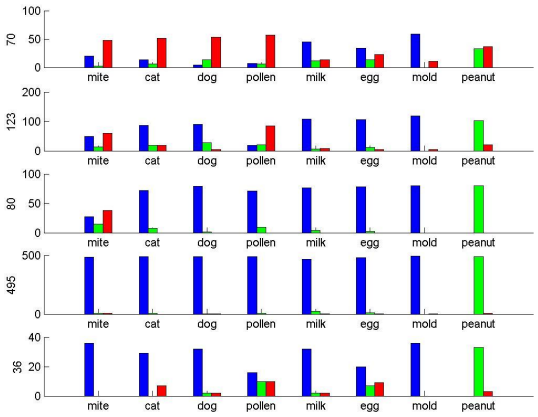

Figure 1: The 5 clusters. The length of the red (blue) bars denote the number of children who have many (few) sensitizations. Clus. 1 (top): General sensitization and early onset, Clus. 2: Inhaled allergen sensitization and late onset, Clus. 3: Mite-only sensitization, Clus. 4: No sensitizations, Clus. 5: Pollen-only sensitization.

4 Conclusion

We have presented a graphical model of the immune system and have shown how Infer.NET can be used to apply the model to analyze a large-scale immunological dataset. Such automatic graphical modeling tools have great potential in terms of speeding up the analyses of large datasets whilst allowing rigorous incorporation of domain knowledge.

References

[1] A. Custovic, B. M. Simpson, C. S. Murray, L. Lowe, and A. Woodcock, "The National Asthma Campaign Manchester Asthma and Allergy Study," *Pediatr Allergy Immunol*, vol. 13 (Suppl. 15), pp. 32 – 37, 2002.

[2] R. Herbrich, T. Graepel, and C. Campbell, "Bayes Point Machines," *JMLR*, vol. 1, pp. 245 – 279, 2001.

[3] J. Winn and C. M. Bishop, "Variational Message Passing," *JMLR*, vol. 6, pp. 661 – 694, 2005.

[4] T. Minka, "Expectation Propagation for approximate Bayesian inference," in *UAI*, 2001.