Bayesian Neural Networks used to Find Adverse Drug Combinations and Drug Related Syndromes

Roland Orre
Mathematical Statistics, University of Stockholm,
Stockholm, Sweden

Andrew Bate and Marie Lindquist
WHO Coll. Centre for Internl. Drug Monitoring, Uppsala Monitoring Centre,
Uppsala, Sweden

Abstract

The data mining task we are interrested in is to find associations between variables in a large database. The method we have earlier proposed to find outstanding associations is to compare estimated frequencies of combinations of variables with the frequencies that would be predicted assuming there were no dependencies. The method we now propose use the same strategy as an efficient way of finding complex dependencies, i.e. certain combinations of explanatory variables, mainly medical drugs, which may be highly associated with certain outcome events or combinations of adverse drug reactions (ADRs). Such combinations of ADRs may also be recognized as syndromes.

The method we use for data mining is an artificial neural network architecture denoted Bayesian Confidence Propagation Neural Network (BCPNN). To decide whether the joint probabilities of events are different from what would follow from the independence assumption, the "*information component*" $\log(P_{ij}/(P_iP_j))$, which is a weight in the BCPNN, and its variance plays a crucial role. We also suggest how this method might be used in combination with stochastic EM to analyse conditioned dependencies also between real valued variables, e.g. to consider the amount of each drug taken.

1 Introduction

We are studying an international data base of case reports, each one describing a possible case of adverse drug reactions (ADRs), which is maintained by the Uppsala Monitoring Centre (UMC), for the WHO international program on drug safety monitoring. Each case report, which can be seen as a row in a data matrix, consists of a number of variables, like drugs used, which amounts of each drug and how the drugs were taken, observed ADRs and other patient data like sex, age and resulting outcome events for the patient [1]. The fundamental problem is to find significant dependencies which might be signals of potentially important ADRs, to be investigated by clinical experts. The estimates of significance are obtained with a Bayesian approach via the variance of posterior probability distributions. The posterior is obtained by fusing a prior Dirichlet distribution with a batch of data, which is also the prior used when the next batch of data arrives.

2 Method

The Bayesian Confidence Propagation Neural Network (BCPNN) [2],[3], can be seen as one way of rewriting Bayes theorem into a form which is reminiscent of other feed forward artificial neural network architectures. It works by propagating probabilistic belief values for a set of inputs or explanatory variables into a set of outputs which are the beliefs that the given input represents one of a set of mutually exclusive classes, which are the response variables. In the work presented here the inputs constitute the drugs suspected of causing the adverse reactions and the outputs are the observed set of adverse reactions or the outcome event suspected to be caused by the reported drug or drug combination.

In the following, let a_j denote the j:th compononent of composite output A of a set of mutually exclusive outcome events. That is, output A could represent a certain adverse reaction [*true, false*] or it could be a set of events like [*alive, coma, death*] where a_j would represent one of these outcomes. In a similar way input D may represent the presence of a suspected drug on a report. More generally, let D, denote a multiple variable input event, where d_i is the input variable i of a set of conditionally independent ($P(D|A) = P(d_1|A) \cdot P(d_i|A) \cdots P(d_n|A)$) variables, and d_{ik} is the k:th mutually exclusive component of one of these input variables. Then $\pi_{d_{ik}}$ is the current "belief" on input event k of variable i. If we only have discrete input belief values ($\pi_{d_{ik}} \in \{0,1\}$) and none of the input states overlap, then the feed forward neural network like expression to produce posterior beliefs for a_j given the input belief values of $\pi_{d_{ik}}$ can be written

$$P(a_j|D) \propto \exp\left[\log P(a_j) + \sum_i \sum_k \log\left[\frac{P(d_{ik},a_j)}{P(d_{ik})P(a_j)}\right]\pi_{d_{ik}}\right]. \tag{1}$$

The weight expression

$$IC_{ijk} = \log\frac{P(d_{ik},a_j)}{P(d_{ik})P(a_j)} = \log\frac{P(d_{ik}|a_j)}{P(d_{ik})} \tag{2}$$

in (1) we denote the "information component" between state j and state k of variable d_i. Often we leave the k index out and just write IC_{ij} because the explanatory variables are usually binary and we are most often only interrested in the positive occurances, *i.e.* the *true* states of the variables. The motivation for the notation "information component" is that mutual information [4] in its discrete form can be regarded as a weighted sum of *information components*:

$$I(Y;X) = \sum_j \sum_k P(x_k,y_j)\log\frac{P(x_k,y_j)}{P(x_k)P(y_j)}, \tag{3}$$

which defines the amount of information passed on from one variable to another. The IC_{ijk} in particular is a measure of the information migrating from state k of variable i to state j of the other variable. Due to the properties of the logarithmic function the

expectation and variance for the IC_{ij} kan be expressed as

$$\begin{aligned} E(IC_{ij}) &= E(\log \frac{p_{ij}}{p_i p_j}) = E(\log p_{ij}) - E(\log p_i) - E(\log p_j), \quad (4) \\ V(IC_{ij}) &= V(\log p_{ij}) + V(\log p_{ij}) + V(\log p_{ij}) \\ &\quad -2cov(\log p_{ij}, \log p_i) - 2cov(\log p_{ij}, \log p_j) + 2cov(\log p_i, \log p_j). \end{aligned}$$

A reasonable model distribution for $P(d_{ik}, a_j)$ is Dirichlet [3]. However, here will the p_{ij}, p_i and p_j all become a special case of the Dirichlet, the Beta. When p being Beta(a,b) distributed there is an exact form [5] of the expectation and variance

$$E(\log p) = \frac{b}{a(a+b)} - b \cdot \sum_{n=1}^{\infty} \frac{1}{(a+n)\cdot(a+b+n)}, \quad (5)$$

$$V(\log p) = \sum_{n=0}^{\infty} \frac{b^2 + 2ab + 2bn}{(a+n)\cdot(a+b+n)^2}. \quad (6)$$

From the expectation and variance values (4) a probability interval for the IC_{ij} can be calculated, which is used as one signalling criterion when searching for unexpected associations between drugs and adverse reactions.

2.1 Combinations of Variables

As described above, the IC_{ij} is a useful measure to find new unexpected single drug ADR associations. The focus of interest for this paper is, however, extending this to analyse also combinations of variables. We want to find variables which interact conditionally, *i.e.* given a certain outcome a set of drugs may show an unexpected interaction, alternatively when a certain drug or combination of drugs is given as input we may find that a set of adverse reactions interact. The latter form of interaction between adverse reactions may lead to detection of *syndromes*. Earlier [3] we have indicated a way of finding such syndrome interactions by looking at conditioned probabilites for combinations of adverse reactions. By looking at, *e.g.* the quotient

$$\log \frac{P(A_1, A_2, A_3 | D_1)}{P(A_1, A_2, A_3)} = IC(A_1, A_2, A_3; D_1), \quad (7)$$

where A_j stand for an adverse drug reaction and D_i for a medical drug, we may find conditionally interacting triplets of adverse reactions. We demonstrated this earlier [3] by looking for a specific syndrome and sorting the results on the $IC(A_1, A_2, A_3; D_1)$ according to (7) and found that we got very high rankings on the combinations of adverse reactions known to appear within the syndrome picture. There are, however, certain limitations, this method will highlight strong dependencies between three states, but they may be due to strong lower order dependencies. The purpose of our search for interacting combinations is to find those where the interaction may not be explained by lower order interactions. Assume that we are looking for pairs where $P(A_1, A_2|D) >> P(A_1, A_2)$, further assume that $P(A_1|D)$ and $P(A_2|D)$ are independent as well as $P(A_1)$ and $P(A_2)$ being independent.

$$\frac{P(A_1, A_2|D)}{P(A_1, A_2)} = \frac{P(A_1, A_2, D)}{P(A_1, A_2)P(D)} = \frac{P(A_1, A_2|D)P(D)}{P(A_1, A_2)P(D)} = \frac{P(A_1|D)P(A_2|D)}{P(A_1)P(A_2)} \quad (8)$$

Under this assumption would the joint probability increase when a marginal probability (κ) increases. We were then looking for a measure which would make the combinations stand out despite lower order interactions. An idea for extension of the *IC*-measure was to use the *IC* between a set of ADRs conditioned on a drug

$$IC(A_1;A_2|D) = \log \frac{P(A_1,A_2|D)}{P(A_1|D)P(A_2|D)}, \tag{9}$$

which when compared with an unconditioned *IC* measuring the general interaction between the same set of adverse reactions

$$IC(A_1;A_2) = \log \frac{P(A_1,A_2)}{P(A_1)P(A_2)} \tag{10}$$

could tell us if the presence of a drug increases or decreases the interaction between the adverse reactions. For a drug related syndrome it could then be expected that when

$$IC(A_1,A_2;D) = \log \frac{P(A_1,A_2|D)}{P(A_1,A_2)} >> 0 \tag{11}$$

would $IC(A_1;A_2) << IC(A_1;A_2|D)$, but the investigation we have done so far, has, however, not given us indications about this. We intend to investigate these measures (9,10) more in the future, but the results we present in this paper are based on the measure in eq. (11) only, which when combined with the variance measure, eq. (5), gives us the ability to sort the obtained results on credibility levels for the *IC* values.

3 Results

The aim with our experiment was to verify that the algorithm could extract a well known syndrome which is considered drug related. The layer specification for the BCPNN was to consider the two classes "haloperidol" and "other drug" as inputs and let the output layer represent a subset of the power set of all adverse reactions (ADRs) occuring on every report.

The drug "haloperidol" is considered to be the main cause of the Neuroleptic Malignant Syndrome (NMS) and included in the symptom picture of NMS are the following four ADRs *creatine phosphokinase increased, fever, death and hypertonia*.

In the setup of this experiment we generated up to the fourth power of ADR combinations in the output layer. The criterion used to associate the ADR combination with the drug haloperiod was the one given by eq. (11).

The weights inside the BCPNN would then implement the following measures $IC(A_1;D)$, $IC(A_1,A_2;D)$, $IC(A_1,A_2,A_3;D)$ and $IC(A_1,A_2,A_3,A_4;D)$, which are referred to as $IC(A*;D)$ in the table below. In the analysis step we generated lists of these different *IC* values which were then filtered on two different criteria, either *IC* being greater than zero or $IC-2\sigma$ being greater than zero. The latter criterion, $IC-2\sigma > 0$, gives us an approximate credibility level of 97% for the *IC* to be positive.

ADR-combinations found where drug haloperidol is suspected					
ADR	#A*	#D-A*	$\#IC(A*;D) > 0$	$\#IC(A*;D) - 2\sigma > 0$	all within
single	2329	623	298	117	91
pair	125791	4952	4458	651	162
triple	496007	6290	6245	256	61
quad	433993	3315	3315	23	0

As could be expected the single term NMS, representing the syndrome, was on the top of all these lists. In the pairs, triplets and quadruples list NMS was also found to be strongly associated with some of the other symptoms which are included in the symptom picture of the selected ADR terms. The reason for this is that the syndrome is not so strictly defined as being composed of these symptoms only, it is enough that the patient has a couple of these symptoms to be diagnosed as NMS. We also found that the selected ADRs were high on all three lists. For the single ADR list all four reactions were among the highest 91 IC values. For the list with ADR pairs combinations of these four ADRs plus the syndrom itself were also found among the highest 162 IC values, and three of these were in the top eight. For the list with triple ADRs, complete combinations were found within the 61 highest. Of the quadruples none of them included a complete symptom picture, on the other hand if we look for the number of terms included in each combination it looks like this:
(3 2 2 2 2 2 3 1 3 2 1 2 3 1 1 3 2 2 1 0 1 1 0), *i.e.* the part with the highest $IC - 2\sigma$ values also contains most of the terms. For the triplets we obtain a similar picture:
(2 2 3 1 2 1 1 1 2 1 2 1 2 1 2 1 1 2 1 0 2 2 2 1 2 1 2 2 1 0 1 0 1 1 2 0 2 2 1 3 2 0 2 1 0 1
1 1 1 1 0 1 2 2 1 1 0 0 1 1 3 1 0 1 0 1 2 0 1 2 1 0 2 1 2 0 1 2 1 1 1 0 1 0 1 1 1 1 2 0 1 0
1 0 2 1 1 1 1 1 1 1 1 2 0 0 1 0 1 1 1 2 0 1 1 1 1 0 1 1 2 0 2 1 0 1 1 2 2 1 1 0 0 1 1 1 1 1
1 1 0 1 0 1 0 2 0 1 1 0 0 0 0 0 1 1 1 1 1 2 1 1 1 0 2 2 0 1 0 2 0 0 2 0 1 1 1 0 0 0 0 1 0 0
1 1 0 2 2 1 0 0 0 0 0 0 1 1 1 1 1 1 0 1 1 0 0 0 1 1 1 0 0 1 0 0 1 0 0 0 0 0 0 0 1 0 1 2 0 1 1
1 0 0 0 1 0 1 0 0 1 0 0 1 1 1 1 1 1 0 0 1 1 2 1 1 1)

4 Discussion on Real Valued Dependencies in R^{12000}

This way of serching for discrete value combinations also allows us to further investigate such variables which have real valued attributes. For each drug reported there is also an associated amount, which may be used to give further information about covariances and ranks of the input space. There are about 12000 drugs being reported and there are about 2 million reports where the amount of the drug taken is filled in for about half of the drugs being reported. To search for dependencies in a real space of dimensionality 12000 would be quite a demanding task. However, the dimensionality of the space that need to be considered in this problem is smaller as the there is a limited number of drugs and ADRs which can occur on one single report.

The approach is to adapt a set of Radial Basis Functions (multivaritate gaussians) to the real valued space using a stochastic EM-algorithm [6],[7] by creating a large set of *low dimensional gaussians* representing the density of the subspaces found on the reports as discrete combinations. Further analysis may then be done on these gaussians $N(\mu_i, \Sigma_i)$ which are parameterized by μ_i which is the center and Σ_i is the covariance matrix for the gaussian density function. The *rank* of the inverse covariance

matrix gives the dimensionality of the dependency, the *non diagonal elements* in the covariance matrix tell about partial covariates or colinearities. Linear and also non linear dependencies may be found by doing regression analysis on center values μ_i.

5 Summary

The BCPNN *(Bayesian Confidence Propagation Neural Network)* has shown to be a useful tool for data mining large data bases and is used on a regular basis for signalling of adverse drug reactions. The *IC (information component)*, which is the weight in a BCPNN, and its variance is an efficient and intutive measure of the strength and significance of a dependecy relation, which relates to information theory. Our approach to use the *IC* to find interactive discrete varible combinations seems promising but is being investigated further. The approach to use this kind of discrete feature detection in combination with multivariate analysis to find dependencies in sparse high dimensional real valued space will also integrate well with the BCPNN methodology.

References

[1] A. Bate, M. Lindquist, I. R. Edwards, S. Olsson, R. Orre, A. Lansner, and R. M. D. Freitas, "A bayesian neural network method for adverse drug reaction signal generation," *European Journal of Clinical Pharmacology*, vol. 54, pp. 315–321, 1998.

[2] A. Holst and A. Lansner, "A higher order bayesian neural network for classification and diagnosis," in *Computational Learning and Probabilistic Reasoning* (A. Gammerman, ed.), ch. 12, John Wiley & Sons, Ltd, 1996. Proc. of ADT, London, England, April 3-5, 1995.

[3] R. Orre, A. Lansner, A. Bate, and M. Lindquist, "Bayesian neural networks with confidence estimations applied to data mining," *Computational Statistics and Data Analysis*, pp. –, 1999. in press.

[4] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, 1988.

[5] T. Koski and R. Orre, "Statistics of the information component in bayesian neural networks," Tech. Rep. TRITA-NA-9806, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden, 1998.

[6] H. G. Tråvén, "A neural network approach to statistical pattern classification by "semiparametric" estimation of probability density," *IEEE Transactions on Neural Networks*, vol. 2, no. 3, pp. 366–118, 1991.

[7] R. Orre and A. Lansner, "Pulp quality modelling using bayesian mixture density neural networks," *Journal of Systems Engineering*, vol. 6, pp. 128–136, 1996.