The relevance vector machine and an in silico study of the oxytocine receptor gene (summary of PhD work in 2011)

Peter MARX
Advisor: Peter ANTAL

I. Introduction

Understanding connections in huge amount of data is one of the greatest challenges in bioinformatics. People search databases, and try to find connections between variables. In this short report, I will describe my research in two parts. In the first part, I will describe my work related to Bayesian feature selection in case of continuous target variables and the relevance vector machine (RVM)[1]. In the second part, I will briefly introduce my oxytocine receptor gene related work and my future work.

II. The Relevance Vector Machine

Last year I started to look for existing methods for Bayesian variable selection in continuous cases. The first method I have analyzed was stochastic search variable selection (SSVS)[2]. After I studied the pros and cons of SSVS (it has its limitations in nonlinear cases) I tried to find a method which also works in nonlinear cases. RVM was developed by Tipping. He wanted to extend the famous support vector machine with the following properties:

- get probabilistic solution;
- use non-Mercel kernels;
- get sparser result in case of bigger training sets.

The probabilistic solution is just in that case important, if you want to know the goodness of the result. For example, if you use SVM for classification you get the class of an input element. If you use RVM you get a probability for all classes. RVM gives sparser result but with the newest modification of SVMs we have the possibility to get comparable result. The upgraded SVM method has a little higher error than the first SVM but gives sparser result so we can also find an optimal ratio between sparsity and error using SVMs.

RVM takes the SVM model:

$$y(\mathbf{x};\mathbf{w}) = \sum_{i=1}^{N} w_i K(\mathbf{x}, \mathbf{x_i}) + w_0 \tag{1}$$

which we can write in a probabilistic model with added noise $t_n = y(x_n, \mathbf{w}) + \varepsilon_n$:

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}), \sigma^2) \tag{2}$$

where the standard deviation of the noise is σ^2. To avoid over-fitting we introduce some hyperparameters, to penalize the model complexity. $\boldsymbol{\alpha}$ gives to every weight a hyperparameter and β to variance.

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i|0, \alpha_i^{-1}) \tag{3}$$

$$p(\boldsymbol{\alpha}) = \prod_{i=0}^{N} Gamma(\alpha_i|a, b) \tag{4}$$

$$p(\beta) = Gamma(\beta|c, d) \tag{5}$$

where $\beta \equiv \sigma^{-2}$. In the above equations the value of a, b, c, d can be chosen by the user. You can find detailed information in Tipping (2001).

Learning the hyperparameters goes iteratively and learning the model goes by Bayesian inference.

$$p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{w}, \boldsymbol{\alpha}, \sigma^2)p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2)}{p(\mathbf{t})} \tag{6}$$

where

$$p(t) = \int p(t|\mathbf{w}, \boldsymbol{\alpha}, \sigma^2)p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2|t)d\mathbf{w}d\boldsymbol{\alpha}d\sigma^2 \tag{7}$$

In equation (6) the normalizing integral cannot be computed analytically so we have to decompose the posterior. The posterior distribution over the weights is the following:

$$p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2) = \frac{p(\mathbf{t}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\boldsymbol{\alpha})}{p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2)} \tag{8}$$

The solution is those relevance vectors where $\alpha_i < \delta$ and δ is a given threshold to filter those vectors where the α_i hyperparameter converges to infinity.

III. Social behavior and the oxytocine receptor

Beside the RVM research I am involved in a cooperative research project with the Semmelweis University (SE) and the Eötvös Loránd University (ELTE). We examine genes related to social behavior. First, we analyzed oxytocine receptor gene in three species: humans, wolves and dogs. We tried to describe the connection between OXTR and social behavior. Initially, I searched for available knowledge in the NCBI and Ensemble database where the human and the dog gene sequence can be found, but there is no public genome for wolves. Using this information I carried out an in silico study to compare the human and dog oxytocine receptor gene and the protein. The biggest difference between the dog and the human OXTR is a 5 amino acid long deletion in the dog OXTR. I also built a phylogenetic tree as you can see in Figure 1. The oxytocine receptor gene is a quite good indicator for clustering species. For detailed information read [3].

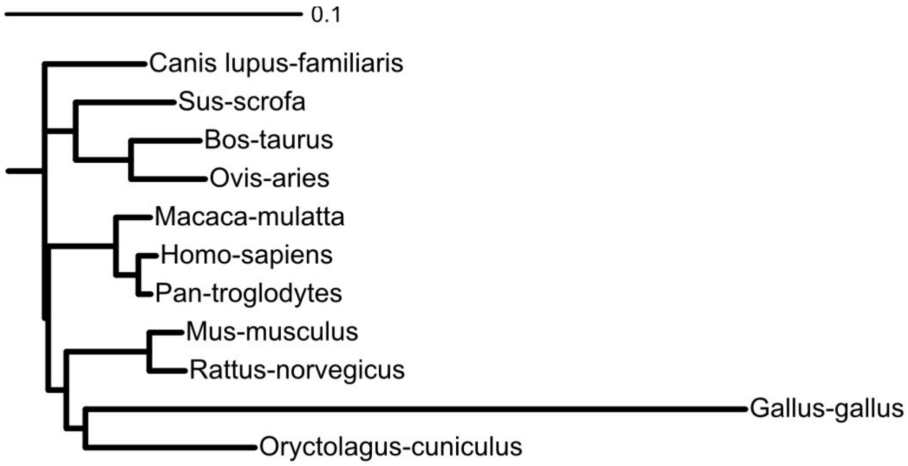

Figure 1: Phylogenetic tree [3]

The second task is to find other genes related to social behavior using Endavour [4]. I used Endavour for two training sets. In the first set, I used the genes which are connected to social behavior in Gene

Ontology [5]. This training set includes 27 genes. In the second case, I used a training set for all genes that were used by the research group at SE in former studies including 32 genes. The training set is based on experts' knowledge about social behavior. There is an overlap between the two training sets, so I am expecting similar results. As we do not have candidate genes, I ran the Endavour on the full human genome in both cases. As you can see on the left side in Table 1 after the fusion of results for each database but Gene Ontology in Endavour, we get the expected outcome. In the first case MAPK8IP1, NLGN1 and DLG2 while in the second case DRD5 are the only ones which were not included in the training set. Because of running on the full genome the first 20 highest ranked genes contains almost exclusively the training genes. We have to analyze the results for all genes with experts to validate the influence of these genes on social behavior. The analysis helps to find candidates. With these candidate genes together with some control genes, I can validate and complete the set of genes.

Table 1: Endavour results for the two different training sets. The results for the training set from Gene Ontology are on the left and the results from the second training set (SE gene set) are shown on the right hand site

symbol	Global prioritization score	rank	rank ratio	symbol	Global prioritization score	rank	rank ratio
HRAS	1	2.70e-16	0.000044	DRD2	1	1.39e-17	0.000044
KRAS	2	1.31e-14	0.0000879	DRD1	2	1.16e-16	0.0000879
DLG4	3	5.40e-13	0.000132	ADRA1A	3	5.37e-15	0.000132
MAPK8IP2	4	2.00e-12	0.000176	DRD3	4	5.89e-14	0.000176
DRD3	5	3.24e-12	0.00022	HTR2C	5	7.56e-14	0.00022
IL1B	6	1.22e-11	0.000264	ADRA2A	6	1.59e-13	0.000264
DRD4	7	4.08e-11	0.000308	HTR1B	7	1.65e-13	0.000308
NLGN3	8	1.28e-10	0.000352	BDNF	8	3.33e-13	0.000352
OXTR	9	1.98e-10	0.000396	NR3C1	9	4.65e-13	0.000396
MKKS	10	3.15e-10	0.00044	SLC6A3	10	5.14e-13	0.00044
TH	11	4.02e-10	0.000484	HTR2A	11	1.11e-12	0.000484
AVPR1A	12	6.62e-10	0.000528	CSNK1E	12	1.14e-12	0.000528
CHRNB2	13	1.14e-9	0.000572	DRD4	13	1.25e-12	0.000572
MAPK8IP1	14	5.50e-9	0.000616	DRD5	14	2.16e-12	0.000616
NLGN4X	15	8.45e-9	0.00066	EGLN2	15	3.88e-12	0.00066
NLGN1	16	9.98e-9	0.000704	OPRM1	16	1.27e-11	0.000704
DBH	17	1.49e-8	0.000748	SLC6A4	17	3.14e-11	0.000748
DLG2	18	3.42e-8	0.000792	COMT	18	4.35e-11	0.000792
AVP	19	8.97e-8	0.000835	MAOA	19	5.74e-11	0.000835
OXT	20	1.30e-7	0.000879	OXTR	20	1.19e-10	0.000879

Next I will run Endavour for the fused training sets and I will look for other training genes to get more general results. Another option is to perform a thematic search e.g. using a training set for the serotonergic genes. Another task is to use parts of these training sets for validation with Endavour. If I left out some genes, will the algorithm give these genes high rank? At the end I would like to have a set containing about 50 genes, which can be used in further studies. Last we would like to do a gene expression study to find SNPs and connect them to specific phenotype.

IV. Future Work

The next steps are the following. As I am traveling to the USA for a one-year research project in the next year I will concentrate on the oxytocine and the social behavior project. I will annotate the wolf OXTR gene and after finding new candidates for social behavior I will search for single nucleotide polymorphisms (SNP) in these genes. To carry out the research I will test SNP validation methods for next generation sequencers (NGS). In the RVM project the next step is to find a way to trace back the solution from the kernel space to the input space to make RVM for variable selection. The goal of this research is to extend the BayesEye so it can also handle continuous cases.

V. Conclusion

During last year I conducted two main researches. Both have some open questions. The relevance vector machine is a promising method which can be applied for Bayesian variable selection. Annotating and publishing the wolf genome is an interesting and challenging study. Getting deeper understanding of the social behavior of dogs and wolves can also lead to describe in more details fearful and aggressive behavior in humans. Last but not least, during this research project at the UCLA my task is to build long time cooperation between the Hungarian and the American research groups.

References

[1] M. E. Tipping, “Sparse bayesian learning and the relevance vector machine,” *Journal of Machine Learning Research*, 1(3):211–244, 2001, Cited By (since 1996): 1017.

[2] E. I. George and R. E. McCulloch, “Variable selection via gibbs sampling,” *Journal of the American Statistical Association*, 88(423):881–889, Sept. 1993.

[3] P. Marx, A. Arany, Z. Ronai, P. Antal, and M. Sasvari-Szekely, “Genetic variability of the oxytocin receptor: an in silico sudy,” *Neuropsychopharmacologia Hungarica*, 13(3):139–144, 2011.

[4] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau, “Gene prioritization through genomic data fusion,” *Nat Biotech*, 24(5):537–544, May 2006.

[5] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S., Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, “Gene ontology: tool for the unification of biology,” *Nature Genetics*, 25(1):25–29, May 2000.