A joint mixture model for clustering genes from Gaussian and beta distributed data

Xiaofeng Dai*, Timo Erkkilä, Olli Yli-Harja and Harri Lähdesmäki*

Department of Signal Processing, Tampere University of Technology, Tampere, Finland

Email: Xiaofeng Dai*- xiaofeng.dai@tut.fi; Harri Lähdesmäki*- harri.lahdesmaki@tut.fi;

*Corresponding author

Background

Cluster analysis has become a standard computational method for gene function discovery as well as for more general explanatory data analysis. A number of different approaches have been proposed for that purpose out of which different mixture models provide a principled probabilistic framework. Cluster analysis is increasingly often supplemented with multiple data sources nowadays and these heterogeneous information sources should be made as efficient use of as possible.

Data and Methods

Beta-Gaussian mixture model (BGMM) is built from beta mixture model (BMM) and Gaussian mixture model (GMM) with the assumption that for each component the beta distributed and Gaussian distributed data are independent. Three types of EM algorithms are developed to estimate the parameters of BGMM, which are the standard EM, an approximated EM and a hybrid EM. We propose to tackle the model selection problem by well-known model selection criteria, such as AIC (Akaike information criterion), AIC3(modified AIC), the BIC (Bayesian information criterion), and ICL-BIC (integrated classification likelihood-BIC).

Results and Discussion

The three EM algorithms that we have developed for BGMM perform similarly according to our simulation results. Performance tests with simulated data show that combining two different data sources into a single joint mixture model greatly improves the clustering accuracy compared to either of its two extreme cases, GMM or BMM. Applications with real mouse gene expression data (Gaussian distribution) and protein-DNA binding probabilities (beta distribution) [1] also demonstrate that BGMM can yield more biologically reasonable results compared to either of its two extreme cases. One of our applications has found three groups of genes that are likely to be involved in Myd88-dependent Toll-like receptor 3/4 (TLR-3/4) signaling cascades, which might be useful to better understand the TLR-3/4 signal transduction.

Conclusions

We have developed a novel mixture model (BGMM) for clustering genes based on Gaussian distributed and beta distributed data. The proposed BGMM can be viewed as a natural extension of BMM and GMM. The proposed BGMM method differs from other mixture model based methods in its integration of two different data types into a single and unified probabilistic modeling framework, which provides a more efficient use of multiple data sources than methods that analyze different data sources separately. Moreover, BGMM provides an exceedingly flexible modeling framework since many data sources can be modeled as Gaussian or beta distributed random variables, and it can be easily extended to integrate data of other parametric distributions which adds even more flexibility to this model-based clustering framework.

Acknowledgements

We would like to thank the Tampere Graduate School in Information Science and Engineering (TISE) for its financial support in this project.

References

1. Lähdesmäki H, Rust AG, Shmulevich I: **Probabilistic Inference of Transcription Factor Binding from Multiple Data Sources**. *Journal of the American Statistical Association* 2008, **3**:e1820.