

---

# Using Bayesian Networks for Model-Based Multiple Clusterings: An Example of Exploratory Analysis on NBA Data

---

Leonard K.M. Poon<sup>†</sup>

Nevin L. Zhang<sup>†</sup>

<sup>†</sup>Department of Computer Science and Engineering  
The Hong Kong University of Science and Technology  
Clear Water Bay, Hong Kong, China

Tao Chen<sup>‡</sup>

Tengfei Liu<sup>†</sup>

Yi Wang<sup>‡</sup>

<sup>‡</sup>Shenzhen Institutes of Advanced Technology  
Chinese Academy of Sciences  
1068 Xueyuan Avenue, Shenzhen, China

## 1 Abstract

Bayesian networks, while usually used for classification, have also been applied in an unsupervised learning setting for data clustering. A mixture of Bayesian networks is often used for the latter case [2]. In unsupervised training, the class variables in the models are considered as latent variables, and the parameters are estimated by the EM algorithm. One limitation of these models is that they have only single latent variables and are restricted to producing single clusterings, which is insufficient in many situations.

Recently, there is a growing interest in multiple clusterings, due to an awareness that data are often multifaceted and can be clustered in multiple meaningful ways. However, most of the multiple-clustering methods are distance-based [e.g. 1, and references therein]. They lack of a strong statistical basis and ease of interpretation. Model-based methods, in contrast, provide these advantages.

In this paper, we demonstrate how to use a family of Bayesian networks with multiple latent variables for model-based multiple clusterings. This family of Bayesian networks is called Pouch Latent Tree Models (PLTMs). Due to computational efficiency, they are restricted to having only tree structures, in which the leaf nodes represent manifest variables while the internal nodes represent latent variables. A hill-climbing search for structure, using the BIC score for model selection, is employed during training to automatically determine the number of latent variables and their cardinalities. The resulting latent variables can then be used to represent multiple clusterings on data.

In our previous work, we introduce PLTMs and empirically show that using PLTMs as a multiple-clustering method is more reasonable than clustering with variable selection [3]. In the present work, we apply PLTMs in an exploratory analysis on some real-world data. We use PLTM analysis to produce multiple clusterings on seasonal statistics of players from the Na-

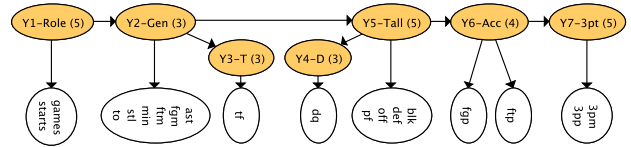


Figure 1: A PLTM obtained from NBA data. Shaded nodes denote discrete latent variables, while others denote continuous manifest variables. The leaf nodes are a shorthand notation for complete networks of the contained variables. Latent variables were renamed based on our interpretation.

tional Basketball Association (NBA). We show that the PLTM obtained (Figure 1) provides several meaningful clusterings on data. For example, we obtained a clustering  $Y_1$  related to role of a player based on the numbers of games played and started, and a clustering  $Y_7$  related to three-pointers based on number of three-pointers made and three-pointer percentage. Thanks to the capability of Bayesian networks, this method also allows interpretation of clusters through the conditional means of manifest variables and modeling of the relationships between clusterings with the conditional probability tables. For example, the CPT between  $Y_5$  and  $Y_6$  reveals that taller players are usually poorer in shooting free throws. We show that other multiple-clustering methods we compared with cannot provide such interesting findings on this data.

## References

- [1] D. Niu, J. G. Dy, and M. I. Jordan. Multiple non-redundant spectral clustering views. In *ICML*, 2010.
- [2] D. T. Pham and G. A. Ruz. Unsupervised training of Bayesian networks for data clustering. *Proc. R. Soc. A*, 465:2927–2948, 2009.
- [3] L. K. M. Poon, N. L. Zhang, T. Chen, and Y. Wang. Variable selection in model-based clustering: To do or to facilitate. In *ICML*, 2010.