Statistical Models for Discovering Knowledge from Relational Data

Iku Ohama, Panasonic Corporation

Relational data encoding pairwise relationships appears in many fields. For example, point-of-sale (POS) data of an e-commerce (EC) site is relational data between customers and items. One of the major motivations behind analyzing such relational data is to discover hidden interaction patterns underlying the given data. In this talk, I will focus on statistical models for discovering underlying structure from relational data.

Biclustering is one of the most popular techniques to extract useful insights from relational data. It abstracts the given data matrix to a low-dimensional block structure by simultaneously clustering both the row and column objects. Among the existing biclustering models, the infinite relational model (IRM) is one of the most widely used models, and it can automatically estimate the optimal number of clusters. Furthermore, the IRM can be inferred efficiently because the model parameters of the IRM can be analytically integrated out.

Although the IRM is very well designed, there are several limitations that are not acceptable in analyzing real-world relational data. First, the IRM assumes that each block in bicluster structure has a uniform density. In other words, the IRM assumes that all the objects within relational data are equally relevant to the underlying structure. However, in many real-world situations, there is heterogeneity in objects' relevance to the underlying structure. In such situation, the IRM produces many unexpected and non-informative clusters. Second, the IRM is a hard clustering model, which assumes each object is assigned to one of the estimated clusters. However, in real-world situations, it is natural to consider that each object is relevant to multiple clusters.

To overcome the above mentioned problems, we will discuss the statistical models for discovering advanced structure from relational data.

First, we develop relevance-dependent infinite biclustering (R-IB), which is an extension of the IRM. In the R-IB, we assume that each object has an additional hidden variable indicating a relevance value that determines how strongly the object relates to the cluster. A large relevance value means that the corresponding object strongly relates to the cluster, whereas a small relevance value indicates that the corresponding object is non-informative. Therefore, the relevance-dependent bicluster structure obtained by the R-IB facilitates understanding what each cluster means. Furthermore, similar to the

IRM, the inference for the R-IB can be performed efficiently because its model parameters can be integrated out.

Second, we develop gamma process edge partition model (GP-EPM), which is a multiple-membership extension of the R-IB. The GP-EPM allows each object to relate to arbitrary number of clusters with arbitrary relevance values, allowing deep insights to be gained from real-world relational data. The GP-EPM also has an efficient inference algorithm similar to the IRM and R-IB.

Finally, we will offer suggestions for future work and conclude the talk.

Keywords: relational data, discovering hidden structure, statistical model