Infinite mixtures for multi-relational categorical data

Janne Sinkkonen JANNE.SINKKONEN@TKK.FI
Helsinki University of Technology, and Xtract Ltd.

Janne Aukia JANNE.AUKIA@XTRACT.COM
Xtract Ltd., Hitsaajankatu 22, 00810 Helsinki, Finland

Samuel Kaski SAMUEL.KASKI@TKK.FI
Helsinki University of Technology, Department of Information and Computer Science, P.O. Box 5400, FI-02015 TKK, Finland

Keywords: nominal co-occurrence data, infinite latent topic models, rich network or graph data, nonparametric Bayesian, relational data

Abstract

Large relational datasets are prevalent in many fields. We propose an unsupervised component model for relational data, i.e., for heterogeneous collections of categorical co-occurrences. The co-occurrences can be dyadic or n-adic, and over the same or different categorical variables. Graphs are a special case, as collections of dyadic co-occurrences (edges) over a set of vertices. The model is simple, with only one latent variable. This allows wide applicability as long as a global latent component solution is preferred, and the generative process fits the application. Estimation with a collapsed Gibbs sampler is straightforward. We demostrate the model with graphs enriched with multinomial vertex properties, or more conceretely, with two sets of scientific papers, with both content and citation information available.

1. Introduction

Many types of data collections can be represented as graphs. These include social networks, metabolic networks in biology, and computer networks (Newman, 2003). Many methods for finding structure in graphs have been devised (Newman & Girvan, 2004; Handcock et al., 2007; Airodi et al., 2007), but the methods do not provide a framework for incorporating other rich data on network elements, such as vertex types.

We present a generalized model for inferring component structure in *enriched graphs* based on a component model for graphs (Sinkkonen et al., 2008). The enriched graphs may contain other types of data beyond simple edges, such as classes for edges, nominal data associated to the vertices, or both—or even something more complex.

From another viewpoint, the model is for general *multi-relational* data, and below we present it in this more broad sense. Here multi-relational data are heterogeneous categorical co-occurrences: The samples are tuples over discrete variables, and heterogeneous in that all samples need not be tuples of similar length or over the same variable. Tuples may also be internally heterogeneous. From this viewpoint, graphs are co-occurrences (edges) within a single categorical variable (vertices), while graphs with associated vertex data have additional co-occurrence type, between vertices and a nominal variable describing the vertices.

The co-occurrences given implicit knowledge about statistical relations between the variables, and these are modeled by a global latent component structure. The relational model is implicit in that occurrences between variables are independent within a component. Dependencies become modeled by the aggregate component structure.

Compared to other multi-relational proposals (Xu et al., 2006; Kemp et al., 2006), here (1) the generative process is simple, (2) the model has only one

Appearing in the 6^{th} *International Workshop on Mining and Learning with Graphs*, Helsinki, Finland, 2008.

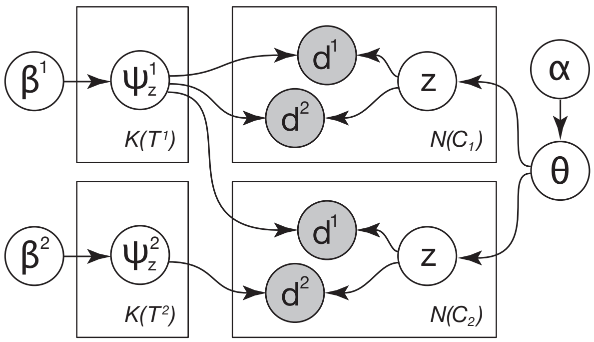

Figure 1. Plate representation for the graph model with nominal vertex data (Section 4).

latent variable, and it therefore produces global latent components, (3) we do not need to explicitly handle probabilities for co-occurrence combinations that do not occur in the data, making the model scalable. Estimation with collapsed Gibbs sampling is easy. From the viewpoint of complex relational models, our model is similar to LDA (Blei et al., 2003), although it allows rich data and has no hierarchy level of "documents".

2. Heterogeneous co-occurrence models

Let the data $\mathcal{D}$ consist of independent *co-occurrences* $\mathcal{D}_i$, $i = 1, \ldots, N$, that can (within a single data set) fall into several object classes described by C_i, $i = 1, \ldots, n(C)$. The structure of the co-occurrences is heterogeneous, but fixed within a class C. A co-occurrence of class C_k is a tuple of nominal values $(d^{(1)}, d^{(2)}, \ldots, d^{(h_k)})$, of size $h_k > 0$. If $h_k = 2$, the co-occurrences are dyadic and presentable as a co-occurrence matrix or a graph. To each variable d we associate a nominal variable type T; the types differ in their domains. Then a tuple $(T_a, T_b, \ldots, T_{h_i})$ of length h_i becomes associated to each C_i. The same variable type T may be shared by several nominal variables d, even within one C.

Note that althought the co-occurrences may often be dyadic, the model class includes triplets and higher-order co-occurrences. It also includes independent events, although they are likely to be of less use.

We assume the data are generated from latent components. A latent component z is drawn, from a multinomial with parameters θ, for each co-occurrence $\mathcal{D}_i$. Given the component z for the datum $\mathcal{D}_i$, and its class C, the nominal values $d^{(t)}$ are generated independently from the associated multinomials, having the types T_t. (Figure 3 and Section 4 offer an example with two co-occurrence classes and two nominal variable types.) Denote the parameters of the multinomials by $\psi_z^{(t)}$. Note that all multinomials of type T_t generated by the same component z share the same parameters $\psi_z^{(t)}$; this is the assumption that ties the co-occurrences together.

We have conjugate priors in the model, a Dirichlet or Dirichlet process (DP) prior for the latent components z_i, and Dirichlet priors for $\psi_z^{(t)}$. With the DP prior, the data are generated by

1. $\theta \sim \mathrm{DP}(\alpha)$; $\psi_z^{(t)} \sim \mathrm{Dir}(\beta^{(t)})$, $t = 1, \ldots, n(T)$;

2. For each $i \in 1, \ldots, N$:

   - $z_i \sim \mathrm{Mn}(\theta)$;
   - $d_i^{(j)} \sim \mathrm{Mn}(\psi_{z_i}^{(t_{ij})})$, $j = 1, \ldots, h_{k(C(\mathcal{D}_i))}$;

with the hyperparameter α controlling the component diversity, and the hyperparameters $\beta^{(t)}$ the evenness of the specific data type distributions. The index t_{ij} simply indicates that each d should be generated from the multinomial T to which it is associated via the description of the co-occurrence class $C(\mathcal{D}_i)$. The occurrence of classes C within $\mathcal{D}$ is not modelled—we have a model only for the contents of an occurrence $\mathcal{D}_i$ given its class C_i. That is, the amounts of data from various types are not modelled either.

All models of the class can be easily estimated by collapsed Gibbs sampling (Neal, 2000), and the rules for sampling the latent classes of the various co-occurrence types are simple enough that they can be derived automatically. Such a sampler gives only posterior samples of the latent memberships $\mathcal{Z}_i$ of the co-occurrences; The parameters ψ and θ are marginalized out. The sampler proceeds by removing one co-occurrence from the sampling "urn" at a time, then drawing a new assignment z for the sample, given assignments of other co-occurrences. An example is presented below in Section 4.

3. Model for graph topology

A trivial case of an undirected graph with one object type, $\{C_1 = (T_1, T_1)\}$, is described by Sinkkonen et al. (2008). The co-occurrences are edges of an undirected graph, with values of T_1 being vertices. Implementation details of that paper are directly applicable in the models of this paper.

4. Model for graphs with nominal vertex data

Another example is a model for two co-occurrence types, $\{C_1 = (T_1, T_1), C_2 = (T_1, T_2)\}$, where $n(C) = 2$. An interpretation is a graph with undirected edges (C_1), and a categorical variable T_2 generating vertex-specific properties (C_2). The corresponding plate model is presented in Figure 1.

The sampling formulas for the two object types are[1]

$$p(z|\mathcal{D}_i) \propto \frac{\{n_z, \alpha\}}{N + \alpha} \times \begin{cases} g^{(1)}_{z,l_1} g^{(1)}_{z,l_2} / (g^{(1)}_{z,\cdot}(g^{(1)}_{z,\cdot} + 1)) & \text{for } \mathcal{D}_i \in C_1 \ , \\ g^{(1)}_{z,l_1} g^{(2)}_{z,l_2} / (g^{(1)}_{z,\cdot}(g^{(2)}_{z,\cdot})) & \text{for } \mathcal{D}_i \in C_2 \ . \end{cases}$$

All counts, g, n, and N, in the samping formulas are *with the object removed* for which we are drawing the latent component. The total number of objects is denoted by N, while n_z is the number of objects (co-occurrences) associated to the latent component z. The first factor arises from the DP prior, with the case $n_z = 0$ corresponding to a new component, and we define $\{n_z, \alpha\} = \alpha$ for $n_z = 0$ otherwise n_z.

A matrix of counts $g^{(t)}_{z,l}$ exists for each type T_t, counting atomic events d assigned to a latent z. The index l is over the bins of the multinomial $\psi^{(t)}_z$. In the sampling formula associated to a co-occurrence class C_k, the indices $l_1, l_2, \ldots, l_{h_k}$ refer to the atomic events d within that type of co-occurrence. Priors β are included in the counts g as virtual data. The dot notation is used for summation.

In the general case of multiple object types, there is one sampling formula similar to those above for each co-occurrence class, and the structure with the g counters closely follows the structure of the object type.

5. Experiments

We tested the model of Section 4 on enriched versions of the Cora and Citeseer data sets (Sen & Getoor, 2007), and compared the model to the simple model of Section 3, which uses only the graph topology, and another simple model which uses only the vertex attributes. The slowest model for Figure 2 ran in 1.25 hours, with a conservative number of iterations (50,000) to assure convergence.

The sizes of the Cora and Citeseer sets are 2708 and 3312 vertices, 5429 and 4732 edges, and 1433 and 3703 indicators for the existence of unique words, respectively. At the time of writing this, the data sets, with more detailed descriptions, are available at `http://www.cs.umd.edu/~sen/lbc-proj/LBC.html` .

[1]We have assumed no self-links in the citation network. If papers were citing themselves, $g^{(1)}_{z,l_2}$ in the numerator of first formula would need to be $g^{(1)}_{z,l_2} + \delta_{l_1,l_2}$.

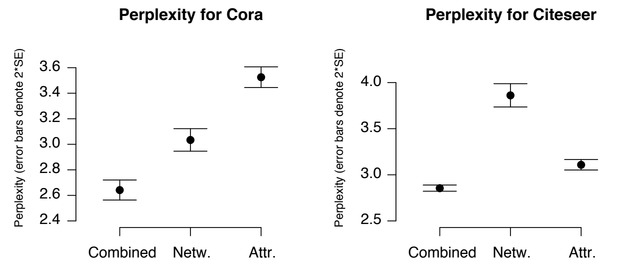

Figure 2. In terms of perplexity, the subject categories of Cora and Citeseer citation sets are best recovered with the model of Section 4, that is able to combine citation and content information ("Combined"; lower perplexity is better). The other candidates are the model of Section 3 ("Netw."), and a similar model for the article content only ("Attr."). The 2SE error bars are over ten runs. The models are with Dirichlet priors that work well in cases with a known number of categories. Note that the models are unsupervised—perplexities are not assumed to beat those from supervised models.

Figure 2 demonstrates how the components found with the models correspond to the correct Citeseer and Cora subject categories in terms of perplexity. Perplexities were computed from average cluster assignments z. Figure 3 shows in further detail how the components found align to the correct subject categories.

6. Discussion

We present an infinite mixture model for multi-relational data and demonstrate it with two enriched citation graphs. Although the original motivation for the model is to find communities (global components) from enriched large social networks, the model is likely to be more widely applicable to relational data.

If the counters g of the Gibss sampler are represented sparsely, the model is highly scalable, regardless of the number of co-occurrence types, that can be very high, even on the order of the data set size.

Comparisons to other approaches are missing from this work. Possible future enhancements to the model would be a hierarchy, and considering more complex latent structures what would still allow sparse representations for efficient estimation for large data sets. Estimation by mean-field approximations or stick-breaking samplers should also be evaluated.

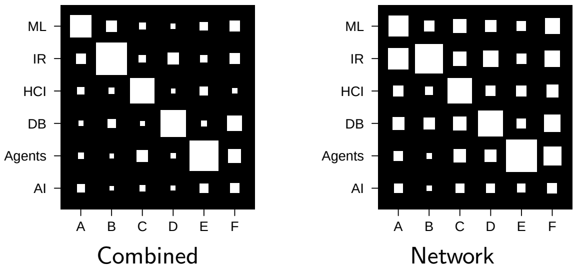

Figure 3. The average confusion matrices between real and computed clusters of Citeseer. *Left:* model for combined citation and content. *Right:* model for the citation information only. The model for combined data recovers the original subject categories except for Artificial Intelligence (AI) that is mixed with Machine Learning (ML) and Agents. Content information is helpful overall, but especially in separating Information Retrieval (IR) from ML.

Acknowledgments

This work was supported by Academy of Finland, grant number 119342. SK belongs to the Finnish CoE on Adaptive Informatics Research of the Academy of Finland and Helsinki Institute for Information Technology, and was partially supported by EU Network of Excellence PASCAL.

References

Airodi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2007). Mixed membership stochastic blockmodels. *ArXiv e-prints*.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Handcock, M. S., Raftery, A. E., & Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal Of The Royal Statistical Society Series A*, *170*, 301–354.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*. Menlo Park, USA: AAAI Press.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*, 249–265.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167–256.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*, 026113.

Sen, P., & Getoor, L. (2007). *Link-based classification* (Technical Report CS-TR-4858). University of Maryland, College Park, USA.

Sinkkonen, J., Aukia, J., & Kaski, S. (2008). Component models for large networks. *ArXiv e-prints*. arXiv:0803.1628.

Xu, Z., Tresp, V., Yu, K., & Kriegel, H.-P. (2006). Infinite hidden relational models. *Proceedings of the 22nd International Conference on Uncertainty in Artificial Intelligence (UAI 2006)*.