Joint Topic Modeling and Factor Analysis of Textual Information and Graded Response Data

Andrew S. Lan, Christoph Studer, Andrew E. Waters, Richard G. Baraniuk
Rice University, USA
{mr.lan, studer, waters, richb}@sparfa.com

ABSTRACT

Modern machine learning methods are critical to the development of large-scale personalized learning systems that cater directly to the needs of individual learners. The recently developed SPARse Factor Analysis (SPARFA) framework jointly estimates learner's knowledge of the latent concepts underlying a domain and the relationships among a collection of questions and the latent concepts, solely from the graded responses to a collection of questions. To better interpret the estimated latent concepts, SPARFA relies on a post-processing step that utilizes user-defined tags (e.g., topics or keywords) available for each question. In this paper, we relax the need for user-defined tags by extending SPARFA to jointly process both graded learner responses and the text of each question and its associated answer(s) or other feedback. Our purely data-driven approach (i) enhances the interpretability of the estimated latent concepts without the need of explicitly generating a set of tags or performing a post-processing step, (ii) improves the prediction performance of SPARFA, and (iii) scales to large test/assessments where human annotation would prove burdensome. We demonstrate the efficacy of the proposed approach on two real educational datasets.

1. INTRODUCTION

Traditional education typically provides a "one-size-fits-all" learning experience, regardless of the different backgrounds, abilities, and interests of individual learners. Recent advances in machine learning enable the design of computer-based systems that analyze learning data and provide feedback to the individual learner. Such an approach has the potential to revolutionize today's education by offering a high-quality, personalized learning experience to learners.

Several efforts have been devoted into building statistical models and algorithms for learner data analysis. In [4], we proposed a personalized learning system (PLS) architecture based on the SPARse Factor Analysis (SPARFA) framework for learning and content analytics, which decomposes assessments into different knowledge components that we call *concepts*. SPARFA automatically extracts (i) a question–concept association graph, (ii) learner concept knowledge profiles, and (iii) the intrinsic difficulty of each question, solely from graded binary learner responses to a set of questions. This framework enables a PLS to provide personalized feedback to learners on their concept knowledge, while also estimating the question–concept relationships that reveal the structure of a course.

The original SPARFA framework [4] relies on a post-processing step to associate instructor-provided question tags to each estimated concept. Inspired by the recent success of modern text processing algorithms, such as latent Dirichlet allocation (LDA) [2], we posit that the text associated with each question can potentially reveal the meaning of the estimated latent concepts without the need of instructor-provided question tags. Such an data-driven approach is advantageous as it easily scales to domains with thousands of questions.

In this paper, we propose *SPARFA-Top*, which extends the original SPARFA framework [4] to jointly analyze graded learner responses to questions and the text of the questions, responses, or feedback. To this end, we augment SPARFA by a Poisson model for the word occurrences associated with each question. We develop a computationally efficient block-coordinate descent algorithm that, given only binary-valued graded response data and associated text, estimates (i) the question–concept associations, (ii) learner concept knowledge profiles, (iii) the intrinsic difficulty of each question, and (iv) a list of most important keywords associated with each estimated concept. We show that SPARFA-Top is able to automatically generate a human readable interpretation for each estimated concept in a purely data-driven fashion. This capability enables a PLS to automatically recommend remedial or enrichment material to learners that have low/high knowledge level on a given concept.

2. THE SPARFA-TOP MODEL

SPARFA [4] assumes that graded learner response data consist of N learners answering a subset of Q questions that involve $K \ll Q, N$ underlying (latent) concepts. Let the column vector $\mathbf{c}_j \in \mathbb{R}^K$, $j \in \{1, \ldots, N\}$, represent the latent *concept knowledge* of the j^{th} learner, let $\mathbf{w}_i \in \mathbb{R}^K$, $i \in \{1, \ldots, Q\}$, represent the *associations* of question i to each concept, and let the scalar $\mu_i \in \mathbb{R}$ represent the *intrinsic difficulty* of question i. The student–response relationship is modeled as

$$\begin{aligned} Z_{i,j} &= \mathbf{w}_i^T \mathbf{c}_j + \mu_i, \quad \forall i, j, \quad \text{and} \\ Y_{i,j} &\sim \mathit{Ber}(\Phi(\tau Z_{i,j})), \quad (i,j) \in \Omega_{\text{obs}}, \end{aligned} \tag{1}$$

where $Y_{i,j} \in \{0, 1\}$ corresponds to the observed binary-valued graded response variable of the j^{th} learner to the i^{th} question, where 1 and 0 indicate correct and incorrect responses, respectively. $\mathit{Ber}(z)$ designates a Bernoulli distribution with success probability z, and $\Phi(x) = \frac{1}{1+e^{-x}}$ denotes the inverse logit link function. The set Ω_{obs} contains the indices of the observed entries (i.e., the observed data

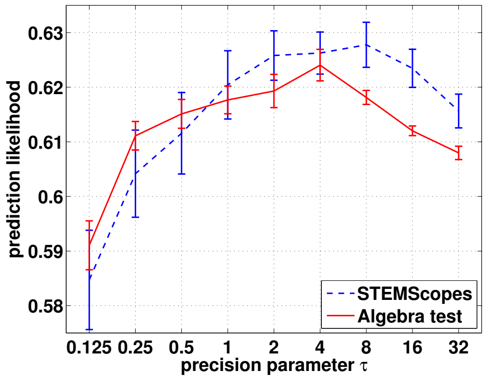


Figure 1: Average predicted likelihood using SPARFA-Top with different precision parameters τ.

may be incomplete). The precision parameter τ models the *reliability* of the observed binary graded response $Y_{i,j}$. In order to account for real-world educational scenarios, $\mathbf{w}_i$ is assumed to be sparse and non-negative [4].

We now introduce a novel approach to *jointly* consider graded learner response and associated textual information, to directly associate keywords with the estimated concepts. Assume that we observe the word–question occurrence matrix $\mathbf{B} \in \mathbb{N}^{Q\times V}$, where V corresponds to the size of the vocabulary, i.e., the number of *unique* words that have occurred among the Q questions. Each entry $B_{i,v}$ represents how many times the v^{th} word occurs in the associated text of the i^{th} question. Inspired by the topic model proposed in [6], the entries of the word-occurrence matrix $B_{i,v}$ in (2) are assumed to be *Poisson* distributed as follows:

$$A_{i,v} = \mathbf{w}_i^T \mathbf{t}_v \quad \text{and} \quad B_{i,v} \sim \mathit{Pois}(A_{i,v}), \quad \forall i, v, \qquad (2)$$

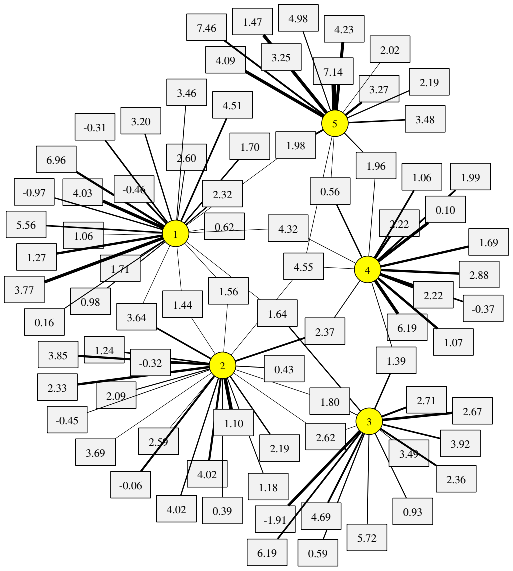
where $\mathbf{t}_v \in \mathbb{R}_+^K$ is a non-negative column vector that characterizes the expression of the v^{th} word in every concept. The latent factors $\mathbf{w}_i$, $\mathbf{c}_j$, $\mathbf{t}_v$ and μ_i are estimated through a block coordinate descent algorithm, which is detailed in [3].

3. EXPERIMENTS

We now demonstrate the efficacy of SPARFA-Top on two real-world educational datasets: an 8th grade Earth science course dataset provided by STEMscopes [5] and a high-school algebra test dataset administered on Amazon's Mechanical Turk [1], a crowdsourcing marketplace.

In Figure 1, we show the prediction likelihood defined by $p(Y_{i,j}|\mathbf{w}_i^T\mathbf{c}_j + \mu_i, \tau), (i,j) \in \bar{\Omega}_{\text{obs}}$ for SPARFA-Top on 20% holdout entries in $\mathbf{Y}$ and for varying precision values τ. We see that textual information can slightly improve the prediction performance of SPARFA-Top over SPARFA (which corresponds to $\tau \to \infty$), for both datasets. The reason for (albeit slightly) improving the prediction performance is the fact that textual information actually reveals additional structure underlying a given test/assessment.

Figure 2 shows the question–concept association graph along with the recovered intrinsic difficulties, as well as the top three words characterizing each concept, for the STEMscopes dataset. Compared to SPARFA (cf. [4, Fig. 2]), we observe that SPARFA-Top is able to relate all questions to concepts, including those questions that were found to be unrelated to any concept. Furthermore, the table in Figure 2 demonstrates that SPARFA-Top is capable of automatically generating an interpretable meaning of each concept.



Concept 1	Concept 2	Concept 3	Concept 4	Concept 5
Energy	Water	Plants	Water	Water
Water	Percentage	Buffalo	Soil	Heat
Earth	Sand	Eat	Sample	Objects

Figure 2: Question–concept association graph and most important keywords recovered by SPARFA-Top for the STEMscopes dataset; boxes represent questions, circles represent concepts, and thick lines represent strong question–concept associations.

4. CONCLUSIONS

We have introduced the SPARFA-Top framework, which extends SPARFA by jointly analyzing both the binary-valued graded learner responses to a set of questions and the text associated with each question via a Poisson topic model. Our purely data-driven approach avoids the manual assignment of tags to each question and significantly improves the interpretability of the estimated concepts by automatically associating keywords extracted from question text to each estimated concept. For additional details, please refer to the full version of this paper on arXiv [3].

5. REFERENCES

[1] Amazon Mechanical Turk. http://www.mturk.com/mturk/welcome, Sep. 2012.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, Jan. 2003.

[3] A. S. Lan, C. Studer, A. E. Waters, and R. G. Baraniuk. Joint topic mmodeling and factor analysis of textual information and graded response data. *arXiv preprint: arxiv.org/abs/1305.1956*.

[4] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk. Sparse factor analysis for learning and content analytics. Oct. 2012, submitted.

[5] STEMscopes Science Education. http://stemscopes.com, Sep. 2012.

[6] J. Zhu and E. P. Xing. Sparse topical coding. In *Proc. 27th Conf. on Uncertainty in Artificial Intelligence*, Mar. 2011.