Summary of the First ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data (U'09)

Jian Pei
Simon Fraser University, Canada
jpei@cs.sfu.ca

Lise Getoor
University of Maryland College Park, USA
getoor@cs.umd.edu

Ander de Keijzer
University of Twente, Netherlands
a.dekeijzer@utwente.nl

The importance of uncertain data is growing quickly in many essential applications such as environmental monitoring, mobile object tracking and data integration. Recently, storing, collecting, processing, and analyzing uncertain data has attracted increasing attention from both academia and industry.

Analyzing and mining uncertain data needs collaboration and joint effort from multiple research communities. Based on this motivation, we ran the First ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data (U'09) in conjunction with the 2009 SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09) at Paris.

The focus of this workshop was to bring together and bridge research in reasoning under uncertainty, probabilistic databases and mining uncertain data. Work in statistics and probabilistic reasoning can provide support with models for representing uncertainty, work in the probabilistic database community can provide methods for storing and managing uncertain data, while work in the mining uncertain data can define data analysis tasks and methods. It is important to build connections among those communities to tackle the overall problem of analyzing and mining uncertain data.

There are many common challenges among the communities. One is understanding the different modeling assumptions made, and how they impact the methods, both in terms of accuracy and efficiency. Different researchers hold different assumptions about the semantics for probabilistic models and uncertainty. This is one of the major obstacles in the research of mining uncertain data. Another challenge is the scalability of proposed management and analysis methods. Finally, to make analysis and mining useful and practical, we need real data sets for testing. Unfortunately, uncertain data sets are often hard to get and hard to share.

The theme of this workshop was to make connections among the research areas of probabilistic databases, probabilistic reasoning, and data mining, as well as to build bridges among the aspects of models, data, applications, novel mining tasks and effective solutions. By making connections among different communities, we aim at understanding each other in terms of scientific foundation as well as commonality and differences in research methodology.

Although the workshop was allocated to only half day, we had a very dynamic and exciting program. The workshop was among one of the best attended ones in conjunction with the conference. There were about 40 attendees when the workshop started.

We were lucky to have two excellent invited talks in the workshop. Professor Christopher Jermaine at Rice University gave a talk on "Managing and Mining Uncertain Data: What Might We Do Better?". In this talk, he expressed a few of his strongly-held opinions on the management and mining of uncertain data. He argued that those who work in the field should listen very carefully to complaints from machine learning experts, who often say, "but all of our methods were already designed to work with uncertain data, so you are wasting your time!" Furthermore, he contended that too much work aimed at managing uncertainty is tightly coupled to first-order logic and related ideas. He also argued that Bayesian approaches and Monte Carlo methods should be much more widely employed in this area. Finally, he argued that too much work in this area neglects the application domains where uncertainty is most important: "what if" analysis, risk assessment, and predication.

In his invited talk titled "Querying and Mining Uncertain Data: Methods, Applications, and Challenges", Dr. Matthias Renz at Ludwig-Maximilians Universität (LMU) München summarized several very interesting projects in his group exploring various aspects of mining uncertain data, particularly from the point of view of efficiency. The efficiency concern is particularly important for modern databases since they allow users to incorporate uncertainty of data in the hope of increasing the quality of query results. Dr. Matthias Renz gave an overview of modeling uncertain data in feature spaces and illustrated diverse probabilistic similarity search methods which are important tools for many mining applications. In this context, he discussed some current methods as well as the challenges in clustering uncertain data and mining probabilistic rules.

The two invited talks were very successful — they led to interesting discussions among the audience and the invited speakers. The invited speeches helped to highlight the interdisciplinary nature of the workshop.

The program committee accepted eight papers — four of them were 15 minute presentations and the other four were 10 minute presentations.

In the paper titled "Efficient Algorithms for Mining Constrained Frequent Patterns from Uncertain Data", Leung and Brajczuk argue that constrained frequent pattern mining from uncertain data is important since constrained frequent pattern mining and mining frequent patterns from uncertain data often happen in some common applications such as analyzing medical laboratory data. They developed

interesting techniques, including an UF-tree data structure and a mining algorithm to push user constraints into the mining process.

Namata and Getoor, in their paper titled "Identifying Graphs from Noisy and Incomplete Data", introduced an interesting graph identification problem – how to model the inference of a "cleaned" output network from a noisy input graph. They used an illustrative example to analyze the types of inferences involved and the inherent challenges. They also gave a simple yet general approach to the problem.

The paper "Learning from Data with Uncertain Labels by Boosting Credal Classifiers" by Quost and Denœux tackles the supervised learning problem when training data is associated with uncertain labels. Their approach uses the theory of belief functions and boosting techniques. A variant of the AdaBoost method is developed.

Dudas and Boström presented an application of mining uncertain data in their paper "Using Uncertain Chemical and Thermal Data to Predict Product Quality in a Casting Process". The uncertainty comes from the fact that the measurements cannot be directly aligned since they are collected at different time points. The authors used random forests to handle uncertain numeric feature values represented by intervals.

In the paper titled "On Perturbation Theory and an Algorithm for Maximal Clique Enumeration in Uncertain and Noisy Graphs", Hendrix, Schmidt, Breimyer, and Samatova considered the problem of maximal clique enumeration in the context of networks based on noisy or uncertain data. They proposed an algorithm that solves the maximal clique enumeration problem on altered or perturbed graphs with correctness guarantee and remarkable performance improvement over the traditional enumeration techniques in the cases of adding and removing edges from protein interaction data.

The paper "Exploiting Contexts to Deal with Uncertainty in Classification" by Zadrozny, Pappa, Meira, Gonçalves, Rocha, and Salles discusses how to account for uncertainty in classification methods, particularly when data attributes may not be accurate for classifying a given sample. They proposed a lazy classification strategy which incorporates the uncertainty into both the training phase and the classification phase, and extended the traditional KNN approach using the strategy.

In the paper titled "Lazy Naive Credal Calssifier", Corani and Zaffalon proposed a local or lazy version of the naïve credal classifier. The classifier retains good reliability even on small amounts of data carrying highly uncertain information.

In the paper titled "Decision Support and Profit Prediction for Online Auction Sellers", Chang and Lin presented another interesting application where uncertainty plays an important role. In order to obtain genuine sold probability and end-price, they applied probability calibration and sample selection bias correction when building the prediction models.

The workshop is just impossible without the great contributions by the authors, the excellent program committee, the external reviewers, and the attendees. We sincerely thank the authors who submitted to the workshop and the attendees of the workshop. We are deeply grateful to the program committee members for their informative reviews in less than two weeks. The program committee members are, in last name alphabetical order, Lyublena Antova, Lei Chen, Reynold Cheng, Nilesh Dalvi, Ming Hua, Nick Koudas, Xuemin Lin, Raymond Ng, Sunil Prabhakar, Christopher Re, Prithviraj Sen, and Guy de Tre. We also thank the external reviewers, in last name alphabetical order, Jinchuan Chen, Xiang Lian, Stephen Sun, and Xike Xie. Our special thanks go to Ming Hua as the publicity chair who publicized the workshop effectively, ran the workshop web site, and designed the art work for the proceedings. Last but not least, we thank Carlos Soares, the workshop chair of KDD'09, for his support in the course.

Inspired by the success of the workshop, we are thinking of the next version of the workshop. Please kindly contact us if you have any comments, suggestions, ideas, or interests.