The PKDD Discovery Challenges on Thrombosis Data

Petr Berka

Laboratory for Intelligent Systems,
University of Economics,
W. Churchill Sq. 4, Prague 3
berka@vse.cz

Abstract. The aim of the Discovery Challenge workshops held during PKDD conferences is to encourage a collaborative research effort when analyzing real world data. For PKDD'99 and PKDD2000 two data sets were available; from the financial and from the medical domain, for PKDD2001 only the medical data are used. There are two basic types of contibutions to the Challenge; in "method oriented" papers the authors describe their own approach and use the data mainly for demonstration, in "problem oreinted" papers the authors tried to solve a problem that can be interesting for the end user.

1 Discovery Challenges at the PKDD Conferences

The aim of the Discovery Challenge workshops held during PKDD conferences is to encourage a collaborative research effort when analyzing real world data. The idea came from Jan Zytkow, who suggested to organize such an event during PKDD' 99 in Prague. In contrast to competitive nature of KDD Cups held within KDD Conferences, the Discovery Challenge emphasises the aspect of cooperation.

Two data sets were available for the PKDD'99 and PKDD2000 Discovery Challenges. In the financial domain, the dataset describes clients of a bank, their accounts, transactions, permanent orders, granted loans and issued credit cards. In the medical domain, the dataset describes patients with collagen diseases. The PKDD2001 Challenge is organized only around the medical data.

Each participant could use any KDD techniques and discover as much knowledge as possible. Ideally each contribution includes the proposed business objectives (goals that may be of interest to database users), a brief summary of datamining effort, presentation of the discovered knowledge, and an explanation for database users how they can apply the discovered knowledge.

2 Medical Domain - the Thrombosis Data

The Thrombosis Data for the PKDD1999 Discovery Challenge were organized into three tables, TSUM_A, TSUM_B, TSUM_C. The tables can be connected by the ID number unique for each patient. Table TSUM_A gives basic information about patients (input by doctors). This dataset includes all patients (about 1000 records). Table TSUM_B gives special laboratory examinations (input by doctors) (measured by the Laboratory on Collagen Diseases). This dataset does not include all the patients, but includes the patients with these special tests. The data in table TSUM_C are data about laboratory examinations stored in Hospital Information Systems (Stored from 1980 to March 1999); all the data include ordinary laboratory examinations and have temporal stamps. The tests are not necessarily connected to thrombosis.

For the PKDD2000 Discovery Challenge, the data was restructured into 7 tables to eliminate problems with multi-valued attributes in the original tables (for details see the data description by Zytkow and Gupta in this volume). The same data tables are used also for the challenge this year.

3 The PKDD experience

Altogether ten papers on the thrombosis data analysis have been presented at the PKDD'99, PKDD2000 and PKDD2001 conferences. Most of the contributions deal with the classification of thrombosis, but there have also been papers dealing with temporal aspects of the data. We can distinguish two basic types of the contributions. The "method/algorithm oriented" papers focus on describing a new approach or system and use the data more or less for demonstration of the features of the method. The "problem oriented" papers try to formulate (and solve) a problem which can be interesting for end users or domain experts. Tables 1-3 summarize all the papers in terms of solved problem (task), described KDD steps, used mining algorithms and used system. All the papers are available from web at http://lisp.vse.cz/challenge.

Table 1. PKDD'99 results

1st. author	KDD task	KDD steps	DM method	tool
Beilken	correlations between lab. Test + thrombosis	vizualization	Display correlations	InfoZoom (own)
Levin	predict yes/no thrombosis	description	association rules, ranking objects	WizWhy (own)
Taylor	predict thrombosis, diagnoses	preprocessing, classification	classification and regression trees	

Table 2. PKDD 2000 results

1st. Author	KDD task	KDD steps	DM method	tool
Meidan	predict yes/no thrombosis	description	association rules, ranking objects	WizWhy (own)
Tawfik	causal and temporal patterns	preprocessing, description, (classification)	statistical techniques (Bayesian networks)	Tetrad

Table 3. PKDD 2001 results

1st. Author	KDD task	KDD steps	DM method	tool
Boulicaut	Classify collagen disease	preprocessing, description	association rules, classification rules	ac-miner-12 (own)
Coursac	classify thrombosis	preprocessing, classification	decision trees and rules	C5.0
Jensen	classify thormbosis	CRISP-DM	neural networks, decision rules, sequnece analysis, association rules	Clementine
Werner	classify severity of disease	classification	genetic programming	LilGP (own)
Zytkow	classify severity of disease	description, classification, interpretation	SQL, contingency tables	

4 Thrombosis data at another challenges

Beside the PKDD conferences, another challenges used the Thrombosis data as well. In September 1999, Shusaku Tsumoto organized a special session in the 38th SIG-FAI and the 45th SIG/KBS of Japanese Society for Artificial Intelligence.

5 References

5.1 Data Descriptions

Tsumoto, S. 1999. Guide to the Medical Data Set. In: (Berka P. ed.) PKDD'99 Workshop Notes on Discovery Challenge, Prague, p.45-47.

Zytkow, J., Tsumoto, S. & Takabayashi, K. 2000. Medical (Thrombosis) Data Description. In: (Siebes A. & Berka P. eds.) PKDD2000 Discovery Challenge, Lyon.

Zytkow, J. & Gupta, S. 2001. Guide to Medical Data on Collagen Disease and Thrombosis. In: (Berka P. ed.) PKDD2001 Discovery Challenge on Thrombosis Data, Freiburg.

5.2 PKDD'99

Beilken, C. & Spenke, M. 1999. Visual, Interactive Data Mining with InfoZoom -- the Medical Data Set. In: (Berka P. ed.) PKDD'99 Workshop Notes on Discovery Challenge, Prague, p.49-54.

Levin, B., Meidan, A., Cheskis, A., Gefen, O. & Vorobyov, I. 1999. PKDD99 Discovery Challenge -- Medical Domain. In: : (Berka P. ed.) PKDD'99 Workshop Notes on Discovery Challenge, Prague, p.55-57.

Taylor, C. 1999. PKDD' 99 DiscoveryChallenge:Medical Data Set. In: (Berka P. ed.) PKDD'99 Workshop Notes on Discovery Challenge, Prague, p.59-64.

5.3 SIG-FAI/KBS-9902

Ichise, R. & Numao, M. 2000. Knowledge Discovery from Medical Database with Multistrategy Approach (in Japanese). In: Proceedings of SIG-FAI/KBS-9902, Japan Assoc. of Artificial Intelligence, p.1-4.

Nakamoto Kazuki, Yoshida Mieko & Suzuki Einoshin, 2000. Analysis of Collagen-Disease Data Set Based on KDD Process Model (in Japanese), In: Proceedings of SIG-FAI/KBS-9902, Japan Assoc. of Artificial Intelligence, p.9-153.

Negishi, N., Suyama, A. & Yamaguchi, T. 2000. Automatic Composition of Inductive Applications to Collagen Diseases Database Using Inductive Learning Method Ontologies. In: Proceedings of SIG-FAI/KBS-9902, Japan Assoc. of Artificial Intelligence, p.5-8.

Tsukada, M., Inokuchi, A., Washio, T., & Motoda H. 2000. Discretization of Numerical Attributes on Structured Data for Basket Analysis, (in Japanese) In: Proceedings of SIG-FAI/KBS-9902, Japan Assoc. of Artificial Intelligence, p.17-24.

5.4 PKDD2000

Meidan, A., Cheskis, A., Gefen, O., Levin, B. & Vorobyov, I. 2000. The WizWhy analysis of the PKDD 2000 Discovery Challenge Medical Domain. In: (Siebes A. & Berka P. eds.) PKDD2000 Discovery Challenge, Lyon.

Tawfik, A. & Strickland K. 2000. Mining Medical Data for Causal and Temporal Patterns. In : (Siebes A. & Berka P. eds.) PKDD2000 Discovery Challenge, Lyon.

5.5 PKDD2001

Boulicaut, J.F. & Crémilleux, B. 2001. δ-strong Classification Rules for Predicting Collagen Diseases. In: (Berka P. ed.) PKDD2001 Discovery Challenge on Thrombosis Data, Freiburg.

Coursac, I., Duteil, N. & Lucas, N. 2001. pKDD 2001 Discovery Challenge - Medical Domain. In: (Berka P. ed.) PKDD2001 Discovery Challenge on Thrombosis Data, Freiburg.

Jensen, S 2001. Mining Medical Data for Predictive and Sequential Patterns: PKDD 2001. In: (Berka P. ed.) PKDD2001 Discovery Challenge on Thrombosis Data, Freiburg.

Werner, J.C. & Fogarty, T.C. 2001. Genetic Programming Applied to Collagen Disease & Thrombosis. In: (Berka P. ed.) PKDD2001 Discovery Challenge on Thrombosis Data, Freiburg.

Zytkow, J. & Gupta, S. 2001. Mining Medical Data using SQL Queries and Contingency Tables. In: (Berka P. ed.) PKDD2001 Discovery Challenge on Thrombosis Data, Freiburg.