Computer Science
PGM: H. Arimura
Computer Challenges of Massive Data Sets

Speaker: Einoshin Suzuki, Yokohama National University, Yokohama

Leveraging Value of Information with Exception Discovery

The speaker personally considers the objective of data mining as building software which leverages value of information. Data mining has originally focused on discovery of strong patterns, each of which holds true for a relatively large number of examples accurately. Such a pattern is, however, often well-known and rarely contributes to the abovementioned objective. On the other hand, weak patterns are numerous in number and discovering interesting weak patterns has mainly relied on domain dependent methods.
Exceptions have long been ignored as noise in machine learning, which has mainly concerned with accurate prediction. However, exceptions are often related with unexpectedness and usefulness thus have attracted attention of data miners. Motivated to build a genuine method for discovering interesting patterns, we have formalized simultaneous discovery of a strong pattern and its exceptional pattern as rule-pair discovery from a data set. Issues such as distinction of exceptions from noise and inefficiency of search have been resolved with an analytical solution based on simultaneous estimation and a search algorithm based on sound pruning respectively. Subjective evaluation of a domain expert for discovered rule pairs from a medical data set was encouraging. We attribute the reason of success to the quality of the data set donated by the domain expert and the structure of our rule pair which we believe to capture interestingness.
In this talk, I will describe several attempts to leverage value of information with exception discovery. These endeavors will be highlighted from the viewpoints of pattern representations, evaluation criteria, search methods, and application fields.

Reference
[1] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth: "From Data Mining to Knowledge Discovery: An Overview", Advances in Knowledge Discovery and Data Mining, U. M. Fayyad et al. (eds.), AAAI/MIT Press, Menlo Park, Calif., pp.1-34, 1996.

用語集

data mining: 【データマイニング】大量データからの有用情報の発見を目的とする研究分野．データに内在する妥当で新規性があり，有用となる可能性がある理解可能なパターンの工夫した特定．

leverage value of information: 【情報価値の向上】スピーカーの造語であり直感的な意味で用いている．例えば，患者データを病気診断手続きに変換することは通常，情報価値を向上することになる．

pattern: 【パターン】自由変数を含む一般的な表記（statement)．

strong pattern: 【強いパターン】 比較的多くの例について正確に成立するパターン．

example: 【例】 値のベクトルであり１つの対象物を表す．事例，タプル，レコードともいう．

weak pattern:【弱いパターン】比較的少ない例について（不正確に）成立するパターン．

interestingness: 【興味深さ】人間が抱く興味深さに関する評価指標．

machine learning: 【機械学習】学習という行為を通してその性能を向上するソフトウェアに関する研究分野．

rule:【ルール】前提部が成立すれば結論部が成立するという形式のパターン．前提部と結論部は，例に関する命題で表される．論理学のルールとは異なり，当てはまらない例があってもよい．例えば，検査値Ａが 30 以上であれば診断Ｂは陽性など．

data set: 【データ集合】例集合として表されるデータ．例えば，10 個の検査値で表された患者 100 人分のデータ．

search: 【探索】複数個の候補を調べることによる問題解決法．この場合，可能なルールペアを調べ，有望そうなものを表示すること．

simultaneous estimation: 【同時推定】データに基づいて未知母集団に関する複数パラメタの値を定めること．

sound pruning: 【健全な枝刈り】探索において，計算結果を変えることなく候補の一部を調べずに済ますこと．

subjective evaluation: 【主観的評価】データ以外の情報を用いて評価すること．領域専門家が，その領域における知識も用いて評価する場合などが当てはまる．

evaluation criteria: 【評価規準】パターンの良さを評価する方法．