

Traitement de données numériques par analyse formelle de concepts et structures de patrons

Vendredi 22 avril, 14h

INRIA NGE / LORIA, salle A-008

Venez nombreux !

Rapporteurs : Jean-François Boulicaut (INSA/Liris, Lyon)

Bernard Ganter (TU Darmstadt)

Examineurs : Bernard Girau (Université H. Poincaré, Nancy)

Sergei O. Kuznetsov (HSE, Moscou)

Céline Rouveirol (LIPN, Paris)

Encadrants : Amedeo Napoli (CNRS, Nancy)

Sébastien Duplessis (INRA, Nancy)

Résumé : Le sujet principal de cette thèse porte sur la fouille de données numériques et plus particulièrement de données d'expression de gènes. Ces données caractérisent le comportement de gènes dans diverses situations biologiques (temps, cellule, etc.). Un problème important consiste à établir des groupes de gènes partageant un même comportement biologique. Cela permet d'identifier les gènes actifs lors d'un processus biologique, comme par exemple les gènes actifs lors de la défense d'un organisme face à une attaque. Le cadre de la thèse s'inscrit donc dans celui de l'extraction de connaissances à partir de données biologiques. Nous nous proposons d'étudier comment la méthode de classification conceptuelle qu'est l'analyse formelle de concepts (AFC) peut répondre au problème d'extraction de familles de gènes. Pour cela, nous avons développé et expérimenté diverses méthodes originales en nous appuyant sur une extension peu explorée de l'AFC : les structures de patrons. Plus précisément, nous montrons comment construire un treillis de concepts synthétisant des familles de gènes à comportement similaire. L'originalité de ce travail est (i) de construire un treillis de concepts sans discrétisation préalable des données de manière efficace, (ii)

d'introduire une relation de similarité entre les gènes et (iii) de proposer des ensembles minimaux de conditions nécessaires et suffisantes expliquant les regroupements formés. Les résultats de ces travaux nous amènent également à montrer comment les structures de patrons peuvent améliorer la prise de décision quant à la dangerosité de pratiques agricoles dans le vaste domaine de la fusion d'information.

Mots-clés : Découverte de connaissances, fouille de données, analyse formelle de concepts, extraction de motifs numériques, bi-clustering, fusion d'information