

Instance Selection and Construction for Data Mining

edited by **Huan Liu**, *National University of Singapore*, and **Hiroshi Motoda**, *Osaka University, Japan*

THE KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE 608
February 2001 Hardbound 448 pp. ISBN 0-7923-7209-3

The ability to analyze and understand massive data sets lags far behind the ability to gather and store the data. To meet this challenge, knowledge discovery and data mining (KDD) is growing rapidly as an emerging field. However, no matter how powerful computers are now or will be in the future, KDD researchers and practitioners must consider how to manage ever-growing data which is, ironically, due to the extensive use of computers and ease of data collection with computers. Many different approaches have been used to address the data explosion issue, such as algorithm scale-up and data reduction. Instance, example, or tuple selection pertains to methods or algorithms that select or search for a representative portion of data that can fulfill a KDD task as if the whole data is used. Instance selection is directly related to data reduction and becomes increasingly important in many KDD applications due to the need for processing efficiency and/or storage efficiency.

One of the major means of instance selection is sampling whereby a sample is selected for testing and analysis, and randomness is a key element in the process. Instance selection also covers methods that require search. Examples can be found in density estimation (finding the representative instances — data points — for a cluster); boundary hunting (finding the critical instances to form boundaries to differentiate data points of different classes); and data squashing (producing weighted new data with equivalent sufficient statistics). Other important issues related to instance selection extend to unwanted precision, focusing, concept drifts, noise/outlier removal, data smoothing, etc.

Instance Selection and Construction for Data Mining brings researchers and practitioners together to report new developments and applications, to share hard-learned experiences in order to avoid similar pitfalls, and to shed light on the future development of instance selection. This volume serves as a comprehensive reference for graduate students, practitioners and researchers in KDD.

Contents and Contributors: Foreword; *R.S. Michalski*. Preface. Acknowledgments. Contributing Authors. **Part I: Background and Foundation.** 1. Data Reduction via Instance Selection; *H. Liu, H. Motoda*. 2. Sampling: Knowing Whole from its Part; *B. Gu, et al.* 3. A Unifying View on Instance Selection; *T. Reinartz*. **Part II: Instance Selection Methods.** 4. Competence Guided Instance Selection for Case-Based Reasoning; *B. Smyth, E. McKenna*. 5. Identifying Competence-Critical Instances for Instance-Based Learners; *H. Brighton, C. Mellish*. 6. Genetic-Algorithm-Based Instance and Feature Selection; *H. Ishibuchi, et al.* 7. The Landmark Model: An Instance Selection Method for Time Series Data; *C.-S. Perng, et al.* **Part III: Use of Sampling Methods.** 8. Adaptive Sampling Methods for Scaling Up Knowledge Discovery Algorithms; *C. Domingo, et al.* 9. Progressive Sampling; *F. Provost, et al.* 10. Sampling Strategy for Building Decision Trees from Very Large Databases Comprising Many Continuous Attributes; *J.-H. Chauchat, R. Rakotomalala*. 11. Incremental Classification Using Tree-Based Sampling for Large Data; *H. Yoon, et al.* **Part IV: Unconventional Methods.** 12. Instance Construction via Likelihood-Based Data Squashing; *D. Madigan, et al.* 13. Learning via Prototype Generation and Filtering; *W. Lam, et al.* 14. Instance Selection Based on Hypertuples; *H. Wang*. 15. KBIS: Using Domain Knowledge to Guide Instance Selection; *P. Wright, J. Hodges*. **Part V: Instance**

 **kluwer academic publishers**

日本総代理店 **株式会社 ニュートリノ**

Selection in Model Combination. 16. Instance Sampling for Boosted and Standalone Nearest Neighbor Classifiers; *D.B. Skalak*. 17. Prototype Selection Using Boosted Nearest-Neighbors; *R. Nock, M. Sebban*. 18. DAGGER: Instance Selection for Combining Multiple Models Learnt from Disjoint Subsets; *W. Davies, P. Edwards*. **Part VI: Applications of Instance Selection.** 19. Using Genetic Algorithms for Training Data Selection in RBF Networks; *C.R. Reeves, D.R. Bush*. 20. An Active Learning Formulation for Instance Selection with Applications to Object Detection; *K.-K. Sung, P. Niyogi*. 21. Filtering Noisy Instances and Outliers; *D. Gamberger, N. Lavra_*. 22. Instance Selection Based on Support Vector Machine; *S. Sugaya, et al.* Index.



日本総代理店 株式会社 ニュートリノ