US006842751B1

(12) **United States Patent**
Vilalta et al.

(10) **Patent No.: US 6,842,751 B1**
(45) **Date of Patent: Jan. 11, 2005**

(54) **METHODS AND APPARATUS FOR SELECTING A DATA CLASSIFICATION MODEL USING META-LEARNING**

(75) Inventors: **Ricardo Vilalta**, Stamford, CT (US); **Irina Rish**, White Plains, NY (US)

(73) Assignee: **International Business Machines Corporation**, Armonk, NY (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 507 days.

(21) Appl. No.: **09/629,086**

(22) Filed: **Jul. 31, 2000**

(51) **Int. Cl.**[7] **G06F 17/30**

(52) **U.S. Cl.** **707/6**; 707/102.1; 707/104.1; 706/12; 706/14; 706/20

(58) **Field of Search** 707/1–10, 100, 707/101; 706/16, 15, 28, 30

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,465,321	A	* 11/1995	Smyth	706/20
5,742,738	A	* 4/1998	Koza et al.	706/13
5,835,901	A	* 11/1998	Duvoisin et al.	706/19
5,884,294	A	* 3/1999	Kadar et al.	706/10
5,970,482	A	* 10/1999	Pham et al.	706/16
6,058,385	A	* 5/2000	Koza et al.	706/13
6,301,579	B1	* 10/2001	Becker	707/102
6,356,884	B1	* 3/2002	Thaler	706/28
6,728,689	B1	* 4/2004	Drissi et al.	706/14

OTHER PUBLICATIONS

"Information Extraction as a Bisis for High–Precision Text Classification"—Ellen Riloff and Wendy Lehnert—ACM Tranaction on Information Systems, vol. 12, No. 3, Jul. 1994, (pps: 296–333).*

* cited by examiner

Primary Examiner—Jean M. Corrielus
Assistant Examiner—Anh Ly
(74) *Attorney, Agent, or Firm*—Ryan, Mason & Lewis, LLP; Louis J. Percello, Esq.

(57) **ABSTRACT**

A data classification method and apparatus are disclosed for labeling unknown objects. The disclosed data classification system employs a model selection technique that characterizes domains and identifies the degree of match between the domain meta-features and the learning bias of the algorithm under analysis. An improved concept variation meta-feature or an average weighted distance meta-feature, or both, are used to fully discriminate learning performance, as well as conventional meta-features. The "concept variation" meta-feature measures the amount of concept variation or the degree of lack of structure of a concept. The present invention extends conventional notions of concept variation to allow for numeric and categorical features, and estimates the variation of the whole example population through a training sample. The "average weighted distance" meta-feature of the present invention measures the density of the distribution in the training set. While the concept variation meta-feature is high for a training set comprised of only two examples having different class labels, the average weighted distance can distinguish between examples that are too far apart or too close to one other.

24 Claims, 9 Drawing Sheets