Title: An unsupervised approach to feature discretization and selection

Author(s): Ferreira, Artur J.[1,3]; Figueiredo, Mário A. T.[2,3]

Source: Pattern Recognition **Volume:** 45 **Issue:** 9 **Special Issue:** SI

Pages: 3048-3060 **DOI:** 10.1016/j.patcog.2011.12.008 **Published:** Sep 2012

Document Type: Article

Language: English

Abstract: Many learning problems require handling high dimensional datasets with a relatively small number of instances. Learning algorithms are thus confronted with the curse of dimensionality, and need to address it in order to be effective. Examples of these types of data include the bag-of-words representation in text classification problems and gene expression data for tumor detection/classification. Usually, among the high number of features characterizing the instances, many may be irrelevant (or even detrimental) for the learning tasks. It is thus clear that there is a need for adequate techniques for feature representation, reduction, and selection, to improve both the classification accuracy and the memory requirements. In this paper, we propose combined unsupervised feature discretization and feature selection techniques, suitable for medium and high-dimensional datasets. The experimental results on several standard datasets, with both sparse and dense features, show the efficiency of the proposed techniques as well as improvements over previous related techniques. (C) 2011 Elsevier Ltd. All rights reserved.

Author Keywords: Feature Discretization; Feature Quantization; Feature Selection; Linde-Buzo-Gray Algorithm; Sparse Data; Support Vector Machines; Naive Bayes; k-nearest Neighbor

KeyWords Plus: Random Subspace Method; Microarray Data; Gene Selection; Classification; Algorithm; Cancer; Information; Redundancy; Relevance

Reprint Address: Ferreira, AJ (reprint author), Polytech Inst Lisbon, Inst Super Engn Lisboa, ADEETC Gabinete 16,Rua Conselheiro Emídio Navarro, P-1959007 Lisbon, Portugal.

Addresses:
1. Polytech Inst Lisbon, Inst Super Engn Lisboa, P-1959007 Lisbon, Portugal
2. Univ Tecn Lisboa, Inst Super Tecn, Lisbon, Portugal
3. Inst Telecomunicacoes, Lisbon, Portugal

E-mail Address: arturj@isel.pt

Funding:

Funding Agency	Grant Number
Polytechnic Institute of Lisbon	SFRH/PROTEC/67605/2010
Instituto de Telecomunicacoes	Pest-OE/EEI/LA0008/2011

Publisher: Elsevier SCI LTD

Publisher Address: The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, Oxon, England

ISSN: 0031-3203

Citation: Ferreira A J, Figueiredo M A T. An unsupervised approach to feature discretization and selection. Pattern Recognition. 2012; 45 (9): 3048-3060.