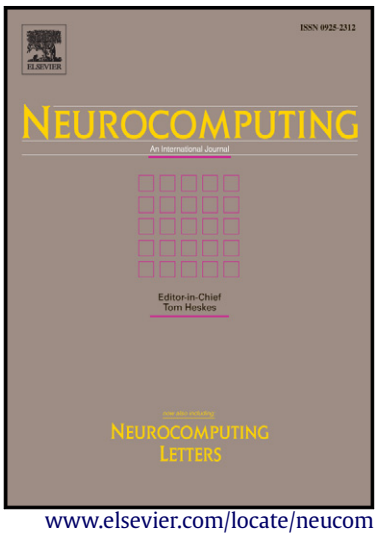
Author's Accepted Manuscript



www.elsevier.com/locate/neucom

Unsupervised Feature Selection through Gram-Schmidt Orthogonalization — A Word Co-occurrence Perspective

Deqing Wang, Hui Zhang, Rui Liu, Xianglong Liu, Jing Wang

PII: S0925-2312(15)01208-4
DOI: http://dx.doi.org/10.1016/j.neucom.2015.08.038
Reference: NEUCOM15968

To appear in: *Neurocomputing*

Received date: 22 July 2014
Revised date: 10 June 2015
Accepted date: 14 August 2015

Cite this article as: Deqing Wang, Hui Zhang, Rui Liu, Xianglong Liu, Jing Wang, Unsupervised Feature Selection through Gram-Schmidt Orthogonalization — A Word Co-occurrence Perspective, *Neurocomputing*, http://dx.doi.org/10.1016/j.neucom.2015.08.038

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Unsupervised Feature Selection through Gram-Schmidt Orthogonalization — A Word Co-occurrence Perspective

Deqing Wang[a,*], Hui Zhang[a], Rui Liu[a], Xianglong Liu[a], Jing Wang[b]

[a]*SKLSDE, School of Computer Science, Beihang University, Beijing, China 100191*
[b]*School of Economics and Management, Beihang University, Beijing, China 100191*

Abstract

Feature selection is a key step in many machine learning applications, such as categorization, clustering etc. Especially for text data, the original document-term matrix is high-dimensional and sparse, which affects the performance of feature selection algorithms. Meanwhile, labeling training instance is time-consuming and expensive. So unsupervised feature selection algorithms have attracted more attention. In this paper, we propose an unsupervised feature selection algorithm through **R**andom **P**rojection and **G**ram-**S**chmidt **O**rthogonalization (RP-GSO) from the word co-occurrence matrix. The RP-GSO algorithm has three advantages: (1) it takes as input dense word co-occurrence matrix, avoiding the sparseness of original document-term matrix; (2) it selects "basis features" by Gram-Schmidt process, guaranteeing the orthogonalization of feature space; and (3) it adopts random projection to speed up GS process. Extensive experimental results show our proposed RP-GSO approach achieves better performance comparing against supervised and unsupervised feature selection methods in text classification and clustering tasks.

Keywords: `Feature Selection, Random Projection, Gram-Schmidt Orthogonalization, Basis Features, Word Co-occurrence Matrix`

*Corresponding author
Email addresses: `dqwang@buaa.edu.cn Tel:8610-82338084 Fax:8610-82339924` (Deqing Wang), `hzhang@nlsde.buaa.edu.cn` (Hui Zhang), `liurui@nlsde.buaa.edu.cn` (Rui Liu), `xlong_liu@nlsde.buaa.edu.cn` (Xianglong Liu), `jim08@buaa.edu.cn` (Jing Wang)

Accepted manuscript

Download English Version:

https://daneshyari.com/en/article/10326425

Download Persian Version:

https://daneshyari.com/article/10326425

Daneshyari.com