View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Repositório do LNEG

Pattern Recognition Letters 45 (2014) 55–61

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Spectrometric differentiation of yeast strains using minimum volume increase and minimum direction change clustering criteria ☆

Nuno Fachada [a,*], Mário A.T. Figueiredo [b], Vitor V. Lopes [c], Rui C. Martins [d], Agostinho C. Rosa [a]

[a] ISR – Institute for Systems and Robotics, Instituto Superior Técnico, Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal
[b] IT – Instituto de Telecomunicações, Instituto Superior Técnico, Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal
[c] LNEG – Laboratório Nacional de Energia e Geologia, Estrada do Paço do Lumiar, 22, 1649-038 Lisboa, Portugal
[d] ICVS – Life and Health Sciences Research Institute, School of Health Sciences, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

ARTICLE INFO

Article history:
Received 6 July 2013
Available online 28 March 2014

Keywords:
Clustering
Minimum volume increase
Minimum direction change
Yeast
Spectroscopy

ABSTRACT

This paper proposes new clustering criteria for distinguishing *Saccharomyces cerevisiae* (yeast) strains using their spectrometric signature. These criteria are introduced in an agglomerative hierarchical clustering context, and consist of: (a) minimizing the total volume of clusters, as given by their respective convex hulls; and, (b) minimizing the global variance in cluster directionality. The method is deterministic and produces dendrograms, which are important features for microbiologists. A set of experiments, performed on yeast spectrometric data and on synthetic data, show the new approach outperforms several well-known clustering algorithms, including techniques commonly used for microorganism differentiation.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Spectroscopy, together with statistical analysis of spectra, is frequently used as a rapid microbiological identification method. Rapid, simple and low-cost identification of microorganisms opens several possibilities. For example, on pathogens in general, it has been shown that fast classification has a major impact on the morbidity, mortality, and duration of hospitalization [18]. For *Saccharomyces cerevisiae* (yeast), quick identification of different strains can yield significant economic advantages, as yeasts not only provide us with many distinctive types of aliment, but are also responsible for food spoilage and can be medically relevant [13]. Winemaking, a multibillion Euro industry, is a prime example, as it could prosper from rapid and comprehensive yeast identification and classification methods [9]. The international wine markets are constantly presenting new challenges, such as taste standardization or production of different and novel wine types with particular characteristics, which can in turn benefit from developing these techniques [6]. Additionally, new species of yeast are continually discovered and explored [15], which requires the classification of a high number of isolates, a task for which a rapid, simple, low-cost identification method is important [13].

Both supervised and unsupervised statistical techniques have been used on spectrometric data with varying degrees of success [19]. Principal Component Analysis (PCA) [12] is one of the latter methods, often employed as a dimensionality reduction step in a broader analysis [6,18]. The majority of methods used for strain differentiation are based on agglomerative hierarchical clustering (AHC) with typical off-the-shelf implementations and parameters [5,11,13,18–20,24,26].

This paper introduces two new clustering criteria for AHC, based on minimizing: (a) the total volume of clusters, as given by their respective convex hulls; and, (b) the global variance in cluster directionality. These are inspired by data produced when applying PCA to spectrometric data, although can be generically used in other problems. A set of experiments, performed on yeast spectrometric data and on synthetic data, show the new approach outperforms several well-known clustering algorithms, namely k-means [10], EM [7], Partitioning Around Medoids (PAM) [3] and AHC with common distance metrics and linkages [10].

The rest of the paper is organized as follows. First, in Section 2, previous work about spectroscopy as a fast identification method and clustering with volume-based metrics is discussed. Next, Section 3 describes the data sets and the dimensionality reduction methods used in this work. The novel clustering metrics, as well as their integration in AHC, are presented in Section 4. Results,

☆ This paper has been recommended for acceptance by Y. Liu.

* Corresponding author. Tel.: +351 21 8418273; fax: +351 21 8418291.

E-mail addresses: nfachada@isr.ist.utl.pt (N. Fachada), mtf@lx.it.pt (M.A.T. Figueiredo), vitor.lopes@lneg.pt (V.V. Lopes), rui.martins@ecsaude.uminho.pt (R.C. Martins), acrosa@laseeb.org (A.C. Rosa).

http://dx.doi.org/10.1016/j.patrec.2014.03.008
0167-8655/© 2014 Elsevier B.V. All rights reserved.