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.

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/

Section 5, show that using AHC with the novel volume and direction-based metrics offers better discrimination capabilities when compared with the remaining tested algorithms. Section 6 provides a global outline of what was accomplished in this work.

2. Related work

2.1. Spectroscopy as a rapid identification method

Fourier transform infrared spectroscopy (FTIR) was one of the first spectroscopy techniques to be able to distinguish and identify microorganisms. Helm et al. [11] used this method to successfully group bacteria according to species, metabolite production, presence of outer membrane and antigenic structure. Several spectral windows were preselected based on their specific information content and discrimination power; similarity between them was measured using Person's correlation coefficient. Different combinations of spectral windows and their weights, as well as the use of first or second derivative of spectra, were systematically tested until the resulting classification mostly agreed with the desired grouping criteria. Grouping was performed using AHC with average and Ward linkages. Kümmerle et al. [13] performed a similar cluster analysis using food-borne yeasts, and concluded that FTIR spectroscopy is limited for taxonomic purposes, because spectra of different species of the same genus generally did not cluster. However, they could successfully identify 97.5% of 722 independent yeast isolates by comparing spectrum similarities against the reference library used in the taxonomic analysis (comprised of 322 yeast strains), showing the potential of FTIR spectroscopy as an identification tool.

Maquelin et al. [18] aimed to identify *Candida* species with spectra obtained using confocal Raman microspectroscopy, by applying a mixture of both unsupervised and supervised techniques. PCA was first applied to the spectra to determine the respective principal components (PCs); a dendrogram was then generated with AHC using squared Euclidean distance and Ward's linkage on the most relevant PCs. Separate clusters were formed for the majority of species. The results of AHC were then used as a starting point for a supervised sequential species identification scheme based on Linear Discriminant Analysis (LDA). This process was used on two data sets which differed on how the *Candida* samples were prepared; when applied to a single data set, 100% correct identification was achieved, and when applied to both data sets combined, 97.0% of samples were correctly identified. Thus, the authors concluded that pretreatment or culturing of strains before spectrum measurement did not significantly influence the accuracy of the devised method.

The potential of visible (VIS) and near-infrared (NIR) spectroscopy to discriminate and identify yeast strains was demonstrated by Cozzolino et al. [6]. In this study, the 2nd derivative of yeast spectra was subjected to PCA. The resulting PCs were taken as independent variables of LDA, with the goal of classifying strains (with different deletion mutations) based on their metabolome. The observed differences were mostly consistent with the knowledge about specific yeast metabolic functions.

Silva et al. [24] used UV–VIS (Ultraviolet–Visible) and VIS–SWNIR (Visible–Short-wave NIR) diffusive reflectance spectra in order to discriminate different yeasts and bacteria. Spectrum data was normalized to account for various integration times, and the growth media spectrum was subtracted from the microorganisms spectra to increase spectral variance, because microorganisms were grown in distinct media. Finally, spectra were then subjected to light scattering correction, and underwent a modified PCA, with emphasis on statistical robustness. AHC, using Euclidean distance and average linkage, was performed on the PCs of UV–VIS and VIS–SWNIR data, leading to the conclusion that VIS–SWNIR produces higher discrimination ratios for all the studied microorganisms. In a posterior study from the same group [5], the authors experimented with yeast metabolic state identification under different growth conditions using a slightly broader spectral interval. Spectra were subjected to low-pass filtering to smooth the signal, followed by light scattering correction. The same modified PCA was applied to the 1st derivative of the spectra, and the resulting PCs underwent a similar clustering process. Results showed that spectroscopy has the potential for yeast metabolic state identification once the spectral signatures of colonies differ from one another, being possible to achieve 100% of classification.

2.2. Clustering with volume-based metrics

When using a minimum volume increase (MVI) criterion in AHC, the inter-cluster dissimilarity is equal to the increase of volume resulting from the merge of any given cluster pair. There are several forms of defining the volume of a cluster of points; when clusters are roughly shaped as convex sets, the volume of the corresponding convex hull or ellipsoid enclosure can be intuitively considered. To our knowledge, the convex hull volume has not been used before in a MVI clustering context, but some work exists regarding the use of minimum volume ellipsoids (MVE) for this purpose [2,14,17,23,27], comparing favorably with several off-the-shelf methods, such as *k*-means. MVE clustering presents several desirable features, such as scale-invariance because of the use of the Mahalanobis distance metric [27]. Further, data often exhibits a mixture of Gaussian distributions, which are shaped as ellipsoids, and thus suitable for this type of clustering. However, if the number of points in a cluster is insufficient or it lacks an ellipsoidal shape, the use of convex hull volume minimization may be advantageous.

MVI presents two main problems when compared with other clustering criteria. First, volume computations incur in higher computational costs, especially when used in a combinatorial context such as AHC. Second, during AHC initialization, in a m-dimensional problem, clusters with fewer than $m+1$ points do not have volume. Thus, MVI must begin with clusters containing at least $(m+1)/2$ points, so that all possible new clusters will have the minimum $m+1$ observations necessary for volume (however, initial clusters are not required to have volume, i.e. they can have zero contribution in Eq. (1). To address this issue, a divisive partitioning algorithm which keeps scale-invariance during MVE clustering is suggested by Kumar and Orlin [14]. However, it is not deterministic, and may lead to unbalanced clusters. In general, any clustering algorithm can be used for initial clustering, provided that each initial cluster contains at least $(m+1)/2$ observations, located close enough to ensure low volume.

3. The data

3.1. Spectrometric data from yeast

The yeast spectra was obtained in a study by Fonseca et al. [9] with VIS–NIR reflectance spectroscopy (450–1000 nm) using yeast strains from different locations and environments. Strains were grown in 96-well (8 rows × 12 columns) microplates at 30 °C in YPD[1] medium for 72 h, so colonies would occupy the entire well. Spectra were obtained inside a *biolog*, which is a box designed to

[1] Yeast Extract Peptone Dextrose, also often abbreviated as YPD, is a complete medium for yeast growth containing yeast extract, peptone, bidest, water, and glucose or dextrose.