Physica A 455 (2016) 52–64

Contents lists available at ScienceDirect

Physica A

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

Cancer detection based on Raman spectra super-paramagnetic clustering

José Luis González-Solís [a,*], Juan Ignacio Guizar-Ruiz [a], Juan Carlos Martínez-Espinosa [b], Brenda Esmeralda Martínez-Zerega [a], Héctor Alfonso Juárez-López [a], Héctor Vargas-Rodríguez [a], Luis Armando Gallegos-Infante [a], Ricardo Armando González-Silva [a], Pedro Basilio Espinoza-Padilla [a], Pascual Palomares-Anda [c]

[a] *Biophyics and Biological Science Laboratory, Centro Universitario de los Lagos, Universidad de Guadalajara, Enrique Díaz de León 1144, Paseo de la Montaña, Lagos de Moreno, Jalisco, Mexico*

[b] *Mathematics and Biotechnology Academy, Instituto Politécnico Nacional - UPIIG, Silao de la Victoria, 36275, Mexico*

[c] *Instituto Mexicano del Seguro Social, Av. P. los Insurgentes S/N, Col. Los Paraísos, León, CP 37320, Guanajuato, Mexico*

HIGHLIGHTS

- The clustering of Raman spectra of serum sample from cancer patients is analyzed.
- The super-paramagnetic clustering (SPC) method to classify Raman spectra is proposed.
- By identifying the most natural clusters, SPC method allows discriminating the control and cancer patients.
- In leukemia case, SPC detects a hierarchical structure allowing the identification of the patient leukemia type.
- The goal of this study is to apply a model of statistical physics to classify spectra allowing the design of a cancer detection method.

ARTICLE INFO

Article history:
Received 8 January 2015
Received in revised form 17 September 2015
Available online 2 March 2016

Keywords:
Super-paramagnetic clustering
Raman spectrum
Cervical and breast cancer
Leukemia
Monte Carlo simulation

ABSTRACT

The clustering of Raman spectra of serum sample is analyzed using the super-paramagnetic clustering technique based in the Potts spin model. We investigated the clustering of biochemical networks by using Raman data that define edge lengths in the network, and where the interactions are functions of the Raman spectra's individual band intensities. For this study, we used two groups of 58 and 102 control Raman spectra and the intensities of 160, 150 and 42 Raman spectra of serum samples from breast and cervical cancer and leukemia patients, respectively. The spectra were collected from patients from different hospitals from Mexico. By using super-paramagnetic clustering technique, we identified the most natural and compact clusters allowing us to discriminate the control and cancer patients. A special interest was the leukemia case where its nearly hierarchical observed structure allowed the identification of the patients's leukemia type. The goal of this study is to apply a model of statistical physics, as the super-paramagnetic, to find these natural clusters that allow us to design a cancer detection method. To the best of our knowledge, this is the first report of preliminary results evaluating the usefulness of super-paramagnetic clustering in the discipline of spectroscopy where it is used for classification of spectra.

* Corresponding author.
E-mail addresses: jluis@gmail.com, jluis0968@gmail.com (J.L. González-Solís).

http://dx.doi.org/10.1016/j.physa.2016.02.060
0378-4371/

1. Introduction

Cluster analysis is one of the most popular tool in exploratory data analysis. Clustering methods, that divide the data according to natural classes present in it, have been used in a large variety of engineering and scientific disciplines such as pattern recognition [1], learning [2], astrophysics [3].

The goal of clustering is to subdivide a set of items in such a way that points in a cluster are more similar to each other than to points in different clusters. Among these clustering techniques there is one that employs a methodology based in concepts of physics, which adopt a local criterion to build clusters by utilizing local structure of the data space, allowing us to identify high-density regions [4]. In Potts models [5], clusters appear naturally as regions of aligned spins. These procedures based on the Potts model introduce a finite temperature at which the division into clusters is stable and completely insensitive to the initial conditions. Furthermore, they complement other graph based algorithms [6] by providing a clustering criterion which is sensitive to collective features of the data set. This clustering technique that employs concepts of statistical physics and succeeds in correctly clustering most of the type of data is the known as the super-paramagnetic clustering (SPC) method [7].

The SPC method, that exploits the properties of phase transitions in disordered Potts ferromagnets, have been applied to the study of the expression levels of large numbers of genes [8,9] and protein sequences [10]. The research on hierarchical components in the correlation structure of an ensemble of stocks was pioneered by Mantegna [11] who used an algorithm known as minimum spanning tree to visualize the correlation structures between stocks in a connected tree. Alternative, methods for cluster identification have been proposed by Kullmann et al. [12,13] and Onnela [14]. Murua and Wicker [15] developed a Bayesian kernel-based clustering method that may be seen as a principled extension of the SPC and applied to the whole human proteome to uncover human genes that share common evolutionary history. In addition, data structures as the obtained in the analysis of the thermodynamic entropy of a cellular system [16–19], could be studied using the SPC model and whose results could play an important role in understanding of the biophysics of biological systems.

Agrawal et al. used the SPC method to identify genes that behave similarly across the samples and classify them as belonging to one group or cluster [8]. The quantities measured in an expression profiling experiment are light intensity ratio that reflect the expression levels of genes in the tissue sample that is used. The extraction of meaningful information from gene expression data is a complex task because of the large volume of the data and the expected complexity of its structure and organization. In Raman spectroscopy, a laser beam excites molecules in the samples and an inelastic scattering effect of the incident photons is observed resulting in a change in their energies (frequency shift or Raman shift). A spectrometer counts the scattered photons and measures the intensity and energy change or Raman shift of the resultant light in units of per centimeter. A column vector containing the information of the Raman shifts and their corresponding intensities is called Raman spectrum. Because each molecule has unique vibrations, the Raman spectrum of a biological sample will consist of a series of peaks or bands, characteristic of the biochemical composition of that sample. Clustering of Raman spectra allows an exploration of these kind of data, identifies spectra with similar characteristics and classifies them as belonging to one cluster. In our cancer study using Raman spectroscopy, we have a large volume of data where each spectrum has a complexity of thousands bands or peaks indicating the biochemical composition of the samples. In addition with a proper structure of data matrix, clustering of Raman spectra also tend to be significantly enriched for specific functional categories which may be used to infer a functional role for unknown molecules in the same cluster of molecules with a known functional role. In essence, in order to classify Raman spectra selecting those having high information content, we tested the ability of the SPC method to determine the physical number of clusters in a large volume of Raman spectra by exploring hierarchical structures which are also observed in some diseases.

Because Raman spectroscopy is a spectroscopic technique that provides information about the molecular structure in one sample, it is fast emerging as a promising alternative technique of application in biology and medicine, mainly in cancer diagnosis. Raman spectroscopy is a nondestructive analytical technique that provides fingerprint spectra with spatial resolution of an optical microscope with almost no sample preparation. This technique has showed to be an excellent technique to detect the breast cancer using biopsies [20] and serum samples [21], cervical cancer [22], leukemia [23], neuroblastoma and ganglioneuroma [24]. In these works, first principal component analysis (PCA) is applied to subsequently use the linear discriminant analysis (LDA) to classify the Raman spectra. PCA–LDA is a way of identifying patterns in data, and expressing the data in such a way as to highlight differences. When the principal component loading are plotted as a function of different variables, they reveal which variable accounts for the greatest difference. Nevertheless, PCA could be not recommended when hierarchical structures in the data must be observed.

We used the data points of our clustering problem as sites of an inhomogeneous Potts ferromagnet and presence of clusters in the data gives rise to magnetic grains. Working in the superparamagnetic phase of the model, we use the values of the pair correlation function of the Potts spins to decide whether a pair of spins does or does not belong to the same grain, and we identify these grains as the clusters of our data. This is the essence of our method.

SPC has been applied in several scientific disciplines as bioinformatics, neuroscience and visual scene analysis in all domains of data analysis and processing, nevertheless to the best of our knowledge, this is the first report of preliminary results evaluating the usefulness of SPC in the discipline of spectroscopy where it is used for classification of spectra.

This paper is organized as follows: In Section 2, we describe the SPC algorithm. In Section 3, we present the methodology to obtain the input data (distance matrix) for the study, and in Section 4, we show the results and full discussion. Section 5 contains the conclusions of this work.