Improving Regression Models by Dissimilarity Representation of Bio-chemical Data

Francisco Jose Silva-Mata, Catherine Jiménez, Gabriela Barcas,
David Estevez-Bresó, Niusvel Acosta-Mendoza(✉), Andres Gago-Alonso,
and Isneri Talavera-Bustamante

Advanced Technologies Application Center, 7th A Avenue # 21406 % 214 and 216,
Siboney, Playa, P.C. 12200, Havana, Cuba
{fjsilva,nacosta}@cenatav.co.cu

Abstract. The determination of characteristics by regression models using bio-chemical data from analytical techniques such as Near Infrared Spectrometry and Nuclear Magnetic Resonance is a common activity within the recognition of substances and their chemical-physical properties. The data obtained from the mentioned techniques are commonly represented as vectors, which ignore the continuous nature of data and the correlation between variables. This fact affects the regression modeling and calibration processes. For solving these problems, alternative representations of data have been previously used with good results, such as those ones based on functions and the others based on dissimilarity representation. By using the alternative based on dissimilarities, the obtained results improve the efficiency of the classification processes, but the experience in regression with this representation is scarce. For this reason, in this paper, in order to improve the quality of the regression models, we combine the dissimilarity representation with some adequate data pre-processing, in our case, we use the classical Partial Least Square regression as the modeling method. The evaluation of the results was carried out by using the coefficient of determination R^2 for each case and a statistical analysis of them is performed.

Keywords: Dissimilarity representation · Regression · Bio-chemical data

1 Introduction

Chemometrics, according to Massart [6], "is the chemical discipline that uses mathematical and statistical, and other methods that employ formal logic to design or select procedures and experiments for optimal measurement and to provide the relevant chemical information from the analysis of the chemical data". In this field, the development of classification methods, regression methods and analysis of substances from spectroscopic and chromatographic signals is one of

R. Vera-Rodriguez et al. (Eds.): CIARP 2018, LNCS 11401, pp. 64–71, 2019.
https://doi.org/10.1007/978-3-030-13469-3_8

the fundamental activities. The prediction of the concentration of a compound in a mixture spectrum or the properties of a material from its known structural parameters is a common task in chemistry [1].

Nowadays, instrumental techniques allow the analysis of pertinent data for different substances, such as: drugs, fuels, inks, medicines, etc. These data are usually represented as a sequence of independent values where each one can be: a measurement or observation made in time, response values for different wavelengths, etc. Usually, spectra are considered as vectors whose values correspond to the samples of the curve at a set of points, therefore, it ignores the continuous nature of data, the correlation between variables and, in many cases, it contains noisy and redundant information. In addition, the high dimension of data contrasts with the small number of samples that are usually counted in the laboratories. Each kind of spectroscopy and chromatography has its own characteristics, then the methods developed for a specific technique may not be applicable to the other ones. It is almost impossible to infer how a model will behave by simply making a reduction in the size of its variables. In order to lead with the mentioned problems, some alternative representations have reported, such as the known Dissimilarity Representation (DR) [7].

The DR allows that certain additional knowledge regarding data can be included making use of a measure of dissimilarity, which, in many cases, reduces the original dimension of data [13]. However, due to the physic-chemical nature of the measurement and the characteristics of each compound, not all dissimilarity measures achieve good results. The dissimilarity based approach has been used for the chemical substances classification, obtaining good results [8]; however, to the best of our knowledge, it has been purely applied in regression tasks [5,12]. Besides, it is not usually known which are the most appropriate dissimilarity measures for certain analytical techniques and what is their effect on the regression modeling.

Although the representation of data is a very important factor for the improvement of the regression model, it is not the only fact to take into account. The whole process to obtain a regression model is rather achieved through the combined use of an adequate pre-processing method and the selection of the most appropriated representation for each analytical technique. In case of using a representation based on dissimilarities, it is crucial the selection of the most appropriated dissimilarity measure.

The goal of this paper is to determine if regression models can be improved by using dissimilarities and which measures are more appropriated for each technique. We use the Partial Least Squares (PLS) as regression method, and Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) as preprocessing methods. For comparatively evaluating the quality of the regression models, the Coefficient of Determination (R^2) is used. It indicates the proportionate amount of variation in the response variable Y explained by the independent variables X in the linear regression model. The larger the R^2 is, the more variability is explained by the linear regression model.

This paper is organized in the following sections. In Sect. 2, we present a brief background. We detailed our proposal in Sect. 3. Section 4 is dedicated to the experiments description and the result analysis. In Sect. 5, the final conclusions of the work are presented.

2 Background

2.1 Pre-processing Methods

Data obtained from instrumental techniques often require that the differences in level and scale be eliminated, but must be known if is needed to pre-process the data and which is the adequate method to be used. MSC and SNV are probably the most widely used of pre-processing techniques for the Near Infrared technique (NIR) [9]. Imperfections, such as the effect of dispersions, are removed from the data matrix before modeling. MSC comprises two steps: (1) the estimation of the regression coefficients (additive and multiplicative contributions), and (2) the spectrum correction. The SNV concept is very similar to MSC except that the reference signal is not needed and each observation is processed separately.

2.2 Partial Least Squares

Partial Least-Squares Regression (PLS) is a widely used method in Chemometry for multivariate calibration. This method uses the information of the latent variables space (i.e. a combination of several variables to form a new one with a certain property). In PLS, the decomposition is carried out in such a way that the scores have the maximum covariance with the dependent variables. The covariance combines the large variances of the independent variables X and the high correlation with the response variable Y [3]. PLS is a linear model where the final latent variables predict the property and it is linear combinations of the original variables, supporting the collinearity between the variables. For modeling a single Y, the algorithm is known as PLS1 and for multivariate regression it is called PLS2.

2.3 Dissimilarity Representation

The Dissimilarity Representation (DR) is based on the role played by the proximity concept in classification problems, so the authors proposed to work in the space defined by the dissimilarity between objects [7]. When the objects to be classified are spectra or signals, the geometry and structure of them are used as discriminative characteristics between the different classes, depending on the used dissimilarity measure. Dissimilarity measures allow calculating and quantifying the differences between the samples; therefore, the selection of the appropriate measure for a given problem is a challenge and it depend on the characteristics of each particular problem. The use of the representation of data based on dissimilarities can be especially advantageous when the number of objects is very small or when the data are represented in high-dimensional spaces [8]. In DR, instead of having an $m \times n$ matrix where m represents the number of objects and n the measured variables, the data will be represented by an $m \times k$ matrix where k is the number of representative objects.

The commonly used dissimilarity measures are Chi-Square with a tolerance parameter epsilon of 0.1, Euclidean Distance (E), Kolmogorov-Smirnov Distance (KS), Cosine Distance (C), Shape Distance (Sh) with smoothing parameter $s = 2$, Spectral Angle Mapping (SAM), Pearson Correlation Coefficient (PCC), Bray-Curtis (BC), Correlation (Corr), Minkowski Distance (M) with $p = 5$ parameters, and Spearman Correlation (S).

3 Our Proposal

Our regression models can be explained by using the steps shown in Fig. 1. In general, we can obtain four possible regression model alternatives. The first one starts representing the original data by a vectorial representation which is modeled by using PLS and evaluated by using R^2 (see the sequence 1-4-5 of Fig. 1). The second alternative is similar to the first one but starting with a pre-processing step over the original data. The third alternative consists in representing the data in the dissimilarity space by using a selected measure over the vectorial representation of the original data (see the sequence 1-2-3-4-5 of Fig. 1). Finally, the fourth alternative is similar to the third one but starting with a pre-processing step over the original data. In this way, our four regression model are built, allowing us a comparison between the results of the regression model based on the space of the dissimilarities for different measures of similarity.

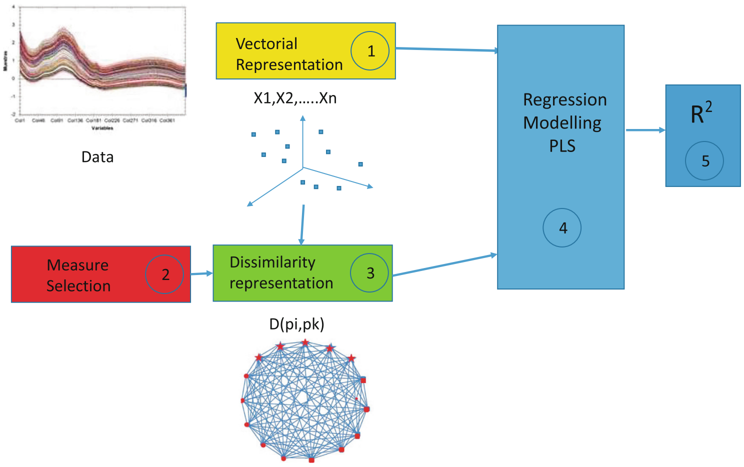

Fig. 1. Sequence steps to obtain the regression modelling.

4 Experiments

Following our four alternatives proposed in Sect. 3, we can evaluate the possible regression models. In our experiments, the PLS method was used with 10 components on each test. For comparing the regression models, the value of R^2 was calculated by each experiment. ANOVA is used to compare the experimental results because a higher value of R^2 does not guarantee a better model, for this fact, we want to verify that the difference between the values is not significant.

4.1 Used Datasets

The chosen dataset used in our experiments are Soil [10], Tecator [11], Cancer [2] (Metabolomic Cancer Diagnosis) and Wine [4] (see Fig. 2), where Soil and Tecator were obtained from the Near Infrared technique (NIR), while Cancer and Wine were obtained from the Nuclear Magnetic Resonance Technique (NMR). The Soil dataset is composed by soil samples originated in a field experiment in Abisko Sweden, where the samples comes from 36 parcels, with three subsamples of each parcel (in two different horizons, one of the inferior and two of the superior), giving a total of 108 samples, and the predicted values measured correspond to the ergosterol concentration. The Tecator dataset comes from the food industry and contains NIR spectra of meat samples, measured in a Tecator Infratec Food and Feed Analyzer. It has 215 samples consisting of 100 Absorbance values in the wavelength range 850 to 1050 nm and is associated with a description of the meat samples, obtained by a chemical that contains the percentage of fat, water and protein of each sample. The calibration problem addressed with this dataset consists in predicting the percentage of fat in the sample of meat from its NIR spectrum.

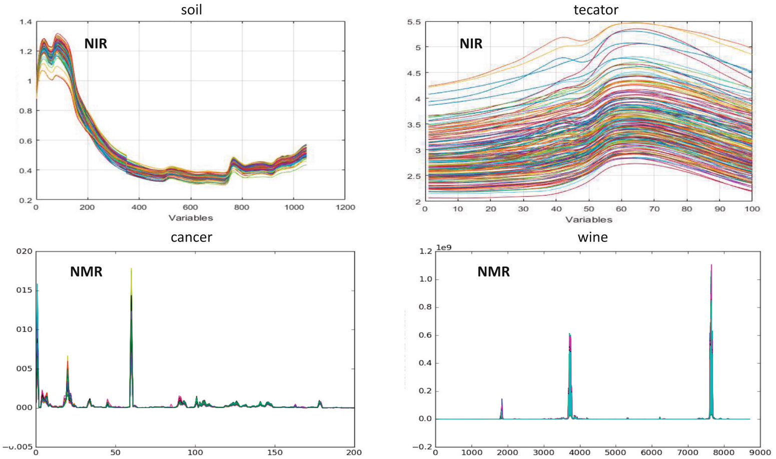

Fig. 2. NIR: Soil and Tecator (top) and NMR: Cancer and Wine (bottom) datasets.

The Cancer dataset was obtained by 1H-NMR spectroscopy (CPMG and NOESY-Presat) of human plasma samples (sodium citrate anticoagulant) and biomarker measurements (TIMP-1 and CEA). This study included patients undergoing to endoscopy of the large intestine due to symptoms that could be associated with ColoRectal Cancer (CRC), also known as bowel cancer and colon cancer. This dataset contains samples of control cases with one case of verified colorectal cancer and three controls for each case. The control group in this dataset correspond to subjects in which benign colorectal adenomas were found. The controls correspond to the age, sex and location of the tumors. The data of NMR are represented as PCA scores (first component) of the integrated peaks. Biomarker data is transformed into a record (base 2). It is known that TIMP-1 and CEA biomarkers change with the age and the genre. This has been corrected in the concentrations of biomarkers subtracting the concentration of a matched sample from another group of control (without findings). On the other hand, the Wine dataset is originated from the analysis by 1H-NMR of 40 table wines of different origin and color, where the value to predict is the pH of mentioned samples.

4.2 Experimental Results

The R^2 average results over each aforementioned dataset are shown in Fig. 3, which covers the four variant results. For each measure the results correspond to the use of the pre-processing step or not. From left to right (up to down) the first value is obtained without pre-processing, the second one is obtained by applying SNV and the third one is achieved by using MSC.

As it can be seen in the figures of Soil, the measures that had improvements with respect to the R^2 value using the vector representation without preprocessing were BC, E, SAM and Sh, measures that in turn have lower variance. Therefore, we can say that the model is more consistent when these measures are applied.

In the case of Tecator, except for the SAM, BC, E and M measures, the other seven measures reported improvements in the results of the dissimilarity, which is verified with the statistical analysis that having little variance in the results of R^2, the values obtained do not depend on the selected set.

For the Cancer dataset, R^2 values for BC, E, M, SAM and Sh measures increase by more than 30% and have the lowest variance values; while for the Wine dataset, the BC, E, M, SAM, Sh, S and X^2 measures present a high value of R^2, approximately 0.99 compared to 0.96 of the original data.

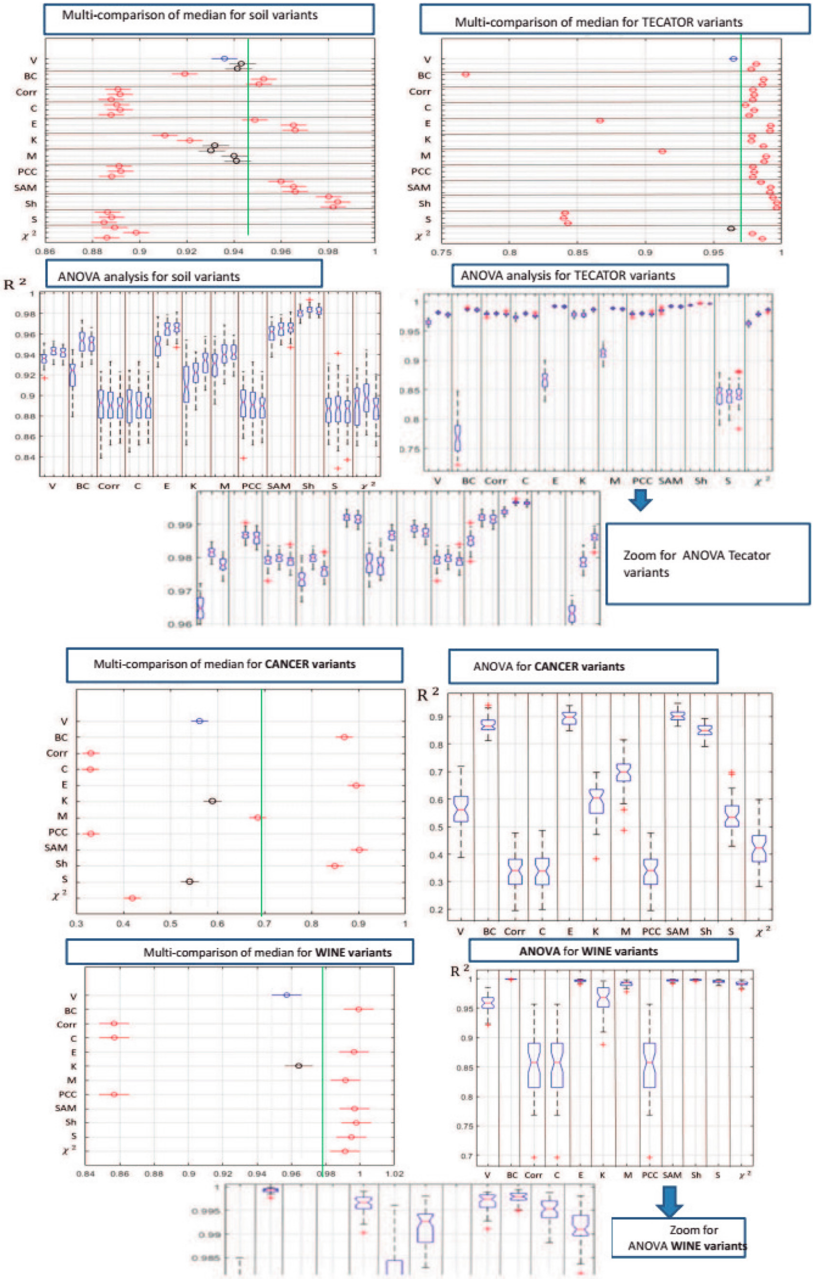

Fig. 3. Values of R^2, multicomparison and ANOVA analysis.

5 Conclusions and Future Work

The results indicate that the use of The Dissimilarity Representation with certain measures, in combination with a pre-processing (in cases where it is required), considerably improves the regression model in comparison with the model for the original feature based representation of data. The measures of dissimilarity that improved to a greater extent the models were: Euclidean, Shape and SAM, and those with the worst results were Kolmogorov-Smirnov and Cosine. The Dissimilarity Representation showed to be good alternative on obtaining the regression models in the selected chemical data for the analytical techniques NIR and NMR.

References

1. Brereton, R.G.: Chemometrics: Data Analysis for the Laboratory and Chemical Plant. Wiley, Chichester (2003)
2. Bro, R., et al.: Data fusion in metabolomic cancer diagnostics. Metabolomics **9**(1), 3–8 (2013)
3. Esbensen, K.H., Guyot, D., Westad, F., Houmoller, L.P.: Multivariate data analysis: in practice: an introduction to multivariate data analysis and experimental design. In: Multivariate Data Analysis (2002)
4. Larsen, F.H., van den Berg, F., Engelsen, S.B.: An exploratory chemometric study of 1H NMR spectra of table wines. J. Chemometr. J. Chemometr. Soc. **20**(5), 198–208 (2006)
5. Martin, Y.C., Lin, C.T., Hetti, C., DeLazzer, J.: PLS analysis of distance matrixes to detect nonlinear relationships between biological potency and molecular properties. J. Med. Chem. **38**(16), 3009–3015 (1995)
6. Massart, D.L.: Handbook of Chemometrics and Qualimetrics. Elsevier Science, Amsterdam (1997)
7. Pekalska, E., Duin, R.P.: Dissimilarity representations allow for building good classifiers. Pattern Recogn. Lett. **23**(8), 943–956 (2002)
8. Porro Munoz, D.: Classification of continuous multi-way data via dissimilarity representation (2013)
9. Rinnan, Å., van den Berg, F., Engelsen, S.B.: Review of the most common preprocessing techniques for near-infrared spectra. TrAC Trends Anal. Chem. **28**(10), 1201–1222 (2009)
10. Rinnan, R., Rinnan, Å.: Application of near infrared reflectance (NIR) and fluorescence spectroscopy to analysis of microbiological and chemical properties of arctic soil. Soil Biol. Biochem. **39**(7), 1664–1673 (2007)
11. Thodberg, H.: Statlib-datasets archive website (2018). http://lib.stat.cmu.edu/datasets/tecator
12. Zerzucha, P., Daszykowski, M., Walczak, B.: Dissimilarity partial least squares applied to non-linear modeling problems. Chemometr. Intell. Lab. Syst. **110**(1), 156–162 (2012)
13. Zerzucha, P., Walczak, B.: Concept of (dis)similarity in data analysis. TrAC Trends Anal. Chem. **38**, 116–128 (2012)