A Texture-based Classification Method for Proteins in Two-Dimensional Electrophoresis Gel Images

A Feature Selection Method using Support Vector Machines and Genetic Algorithms

Carlos Fernandez-Lozano, Jose A. Seoane, Marcos Gestal, Daniel Rivero, Julian Dorado and Alejandro Pazos

Information and Comunnications Technologies Department. Faculty of Computer Science, University of A Coruña, Campus Elviña s/n, 15071, A Coruña, Spain

Keywords: Texture Analysis, Feature Selection, Electrophoresis, Support Vector Machines, Genetic Algorithm.

Abstract: In this paper, the influence of textural information is studied in two-dimensional electrophoresis gel images. A Genetic Algorithm-based feature selection technique is used in order to select the most representative textural features and reduced the original set (296 feat.) to a more efficient subset. Such a method makes use of a Support Vector Machines classifier. Different experiments have been performed, the pattern set has been divided into two parts (training and validation) extracting a total of 30%, 20% and 0% of the training data, and a 10-fold cross validation is used for validation. In case of extracting 0% means that training set is used for validation. For each division 10 different trials have been done. Experiments have been carried out in order to measure the behaviour of the system and to achieve the most representative textural features for the classification of proteins in two-dimensional gel electrophoresis images. This information can be useful for a protein segmentation process.

1 INTRODUCTION

Proteomics is the study of protein properties in a cell or tissue aimed at obtaining a global integrated view of disease, physiological and biochemical processes of cells and regulatory networks. One of the most powerful techniques, widely used to analyze complex protein mixtures extracted from cells, tissues, or other biological samples, is two-dimensional polyacrylamide gel electrophoresis (2D-PAGE). In this method, proteins are classified by molecular weight (MWt) and iso-electric point (pI) using a controlled laboratory process and digital imaging equipment. Among others separation of proteins of a sample could also be done with several different techniques such as chromatography or mass spectrometry.

The main advantages of this approach are its robustness, its parallelism and its unique ability to analyze complete proteins at high resolution, keeping them intact and being able to isolate them entirely, however this method has also several drawbacks (Rabilloud, Chevallet et al., 2010).

In this work the most representative group of textural features are selected using Genetic Algorithms.

2 THEORETICAL BACKGROUND

The method proposed in this work intends to assist in 2D-PAGE image analysis by studying the textural information present within them. To do so, a novel combination of Genetic Algorithms (Holland, 1975) and Support Vector Machines (Vapnik, 1979) is presented. In this section, the main techniques used are briefly introduced and explained.

One of the most important characteristics used for identifying objects or regions of interest in an image is texture, related with the spatial (statistical) distribution of the grey levels within an image (Haralick, Shanmugam et al., 1973). Texture is a surface's property and can be regarded as the regular spatial organization of complex patterns, always

Fernandez-Lozano C., Seoane J., Gestal M., Rivero D., Dorado J. and Pazos A..
A Texture-based Classification Method for Proteins in Two-Dimensional Electrophoresis Gel Images - A Feature Selection Method using Support Vector Machines and Genetic Algorithms.
DOI: 10.5220/0004208704010404
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 401-404
ISBN: 978-989-8565-47-1

present even if they could exist as a non-dominant feature.

Genetic Algorithms (GAs) are search techniques inspired by Darwinian Evolution and developed by Holland in the 1970s (Holland, 1975). In a GA, an initial population of individuals, i.e. possible solutions defined within the domain of a fitness function to be optimized, is evolved by means of genetic operators: selection, crossover and mutation. The selection operator ensures the survival of the fittest, while the crossover represents the mating between individuals, and the mutation operator introduces random modifications. GAs possesses effective exploration and exploitation capabilities to explore the search space in parallel, exploiting the information about the quality of the individuals evaluated so far (Goldberg, 1989).

Vapnik introduces Support Vector Machines (SVMs) in the late 1970s on the foundation of statistical learning theory (Vapnik, 1979). The basic implementation deals with two-class problems in which data are separated by a hyperplane defined by a number of support vectors. This hyperplane separates the positive from the negative examples, to orient it such that the distance between the boundary and the nearest data point in each class is maximal; the nearest data points are used to define the margins, known as support vectors (Burges, 1998). These classifiers have also proven to be exceptionally efficient in classification problems of higher dimensionality (Chapelle, Haffner et al., 1999; Moulin, Alves Da Silva et al., 2004), because of their ability to generalize in high-dimensional spaces, such as the ones spanned by texture patterns.

3 MATERIALS

In order to generate the dataset, ten 2D-PAGE images of different types of tissues and different experimental conditions were used. These images are similar to the ones used by G.-Z. Yang (Imperial College of Science, Technology and Medicine, London). It is important to notice that Hunt et al. (Hunt, Thomas et al. 2005) determined that 7-8 is the minimum acceptable number of samples for a proteomic study.

For each image, 50 regions of interest (ROIs) representing proteins and 50 representing no-proteins (noise, black non-protein regions, and background) were selected to build a training set with 1000 samples in a double-blind process in the way that two clinicians select as many ROIs as they considered and after that, within the common ROIs clinicians selected proteins which are representatives (isolated, overlapped, big, small, darker, etc.).

4 PROPOSED METHOD

The first step in texture analysis is texture feature extraction from the ROIs. With a specialized software called Mazda (Szczypiski et al., 2009), 296 texture features are computed for each element in the training set. These features are based on the image histogram, co-ocurrence matrix, run-length matrix, image gradients, autoregressive models and wavelet analysis. Histogram-related measures conform the first-order statistics proposed by Haralick (Haralick, Shanmugam et al., 1973) but second-order statistics are those derived from the Spatial Distribution Grey-Level Matrices (SDGM).

All these feature sets were included in the dataset. The normalization method applied was the one set by default in Mazda: image intensities were normalized in the range from 1 to $Ng=2^k$, where k is the number of bits per pixel used to encode the image under analysis.

In this work, GA is aimed at finding the smallest feature subset able to yield a fitness value above a threshold. Besides optimizing the complexity of the classifier, feature selection may also improve the classifiers quality. In fact, classification accuracy could even improve if noisy or dependent features are removed.

GAs for feature selection were first proposed by Siedlecki and Skalansky (Siedlecki and Sklansky, 1989). Many studies have been done on GA for feature selection since then (Kudo and Sklansky 1998), concluding that GA is suitable for finding optimal solutions to large problems with more than 40 features to select from.

GA for feature selection could be used in combination with a classifier such SVM, KNN or ANN, optimizing it. In our method, based on both GA and SVM, there is no a fixed number of variables. As the GA continuously reduces the number of variables that characterize the samples, a pruned search is implemented. The fitness function (1) considers not only the classification results but also the number of variables used for such a classification, so it is defined as the sum of two factors, one related to the classification results and another to the number of variables selected. Regarding classification results, it apparently gives better results taking into account the F-measure than only using the accuracy obtained with image features (Müller, Demuth et al., 2008; Tamboli and

Shah, 2011). F-measure is a function made up of the recall (true positives rate or sensitivity: proportion of actual positives which are correctly identified as such) and precision (or positive predictive value: proportion of positive test results that are true positives) measurements.

$$Fitness = (1 - F) + \frac{numberActiveFeatures}{numberTotalFeatures} \quad (1)$$

Therefore individuals with less active genes are favored.

5 EXPERIMENTAL RESULTS

The method proposed in this work requires the division of the pattern set into two halves. To avoid overfitting, this work proposes to split the training dataset into training and validation sets to perform a validation of the obtained results. Once the GA finishes, the best individual found (the one with lowest fitness value) is tested, using a 10-fold cross validation (10-fold CV), to calculate the error of the proposed model with the validation set and using only the features in the best individual chromosome. This involves dividing the validation set into 10 complementary subsets, performing the analysis on 1 subset, retained as the validation data and the remaining k-1 subsamples are used as training data. This second partitioning provokes that either validation could be carried out with a very reduced number of data points. In this case, either training or validation sets, will surely not be representative of the search space that is being explored. Different experiments have been performed to verify this, in these experiments the pattern set has been divided into two parts (training and validation) extracting a total of 30 %, 20% and 0% of the training data to the validation set. In case of extracting 0%, which means that no validation is performed training set is used for validation using cross validation. For each division 10 different trials have been done; a different seed is used to divide randomly the elements of the dataset each time.

Parameters domains of the feature selection method were initially adjusted based on the literature in the way that are ranging in the case of the population size from 100 to 250 individuals, elitism from 0% to 2%, crossover probability from 80% to 98% and mutation probability from 1% to 5%. One-point, two-point, scattered, arithmetic and heuristic crossover functions were probed. Regarding with selection function, uniform, roulette and tournament functions were evaluated with uniform and Gaussian mutation functions.

Final combination set population size to 250 individuals, no elite, 95% crossover probability, 2% mutation probability, crossover scattered, tournament selection and mutation uniform.

SVM parameters domains are set for the kernel function as lineal, quadratic, polynomial (order ranging from 3 to 10) and Gaussian radial basis, with sigma parameter ranging from 0,1 to 10 and C parameter from 1 to 100. The RBF(2) kernel function is selected as the most accurate for solving this problem.

Experiments results shown in Table 1 separated in training error and the validation error calculated using 10-fold CV and the validation set for each division. Results are in mean of error of the 10 trials and standard deviation is in brackets. Best results in validation are achieved for the 0% division.

In these 10 trials, some features seem to be more relevant and appear recurrently as the solution of each trial. Skewness, S(0,5) InvDfMom, S(2,2) Correlat and S(0,4) InvDfMom appears at least in 5 solutions. Skewness is a measure of the degree of asymmetry of the image histogram distribution. Correlation analyzes the linear dependency of gray levels of neighboring pixels. When the scale of local texture is larger than the distance this measure is typically high. And inverse difference moment is the inverse of the contrast of the occurrence matrix so it is a measure of the amount of local uniformity present in the image.

The co-ocurrence matrix in Mazda (Szczypiski et al., 2009) is symmetric and the image is normalized. Co-ocurrence based parameters are computed up to 20 times, for (d,0), (0,d), (d,d) and (d,-d) where distance *d* is ranging from 1 to 5.

6 SUMMARY AND CONCLUSIONS

In this work we present a method for classification of proteins in two-dimensional electrophoresis gels using textural information. The proposed method is based on a feature selection process using GAs (Holland, 1975) and SVMs (Vapnik, 1979).

A dataset with 10 images, 100 ROIs for each one and 296 features per ROI is created. Two different clinicians have performed this manual protein detection. This is a high variability process; a refinement step based on the correlation of the results of the two clinicians was performed, in order

to select enough representative proteins.

Different experiments have been performed and in these experiments the pattern set has been divided into two parts (training and validation) extracting a total of 30 %, 20% and 0% of the training data to the validation set.

The proposed method has been successfully applied to different real images, including images with high complexity, which means larger number of proteins and larger deformation between images. Furthermore, the method presented have important implications for the analysis of two-dimensional electrophoresis gel images in the sense that this classification step can be very useful in order to discard over-segmented areas after a protein segmentation or identification process.

ACKNOWLEDGEMENTS

This work is supported by the General Directorate of Culture, Education and University Management of the Xunta de Galicia (Ref. 10SIN105004PR).

REFERENCES

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2(2): 121-167.

Chapelle, O., P. Haffner, et al. (1999). Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks* 10(5): 1055-1064.

Goldberg, D. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning, *Addison-Wesley Professional*.

Haralick, R. M., K. Shanmugam, et al. (1973). Textural features for image classification. *IEEE Transactions on Systems*, Man and Cybernetics smc 3(6): 610-621.

Holland, J. H. (1975). Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence, University of Michigan Press.

Hunt, S. M. N., M. R. Thomas, et al. (2005). Optimal Replication and the Importance of Experimental Design for Gel-Based Quantitative Proteomics. *Journal of Proteome Research* 4(3): 809-819.

Kudo, M. and J. Sklansky (1998). A comparative evaluation of medium- and large-scale feature selectors for pattern classifiers. *Kybernetika* 34(4): 429-434.

Moulin, L. S., A. P. Alves Da Silva, et al. (2004). Support vector machines for transient stability analysis of large-scale power systems. *IEEE Transactions on Power Systems* 19(2): 818-825.

Müller, M., B. Demuth, et al. (2008). An evolutionary approach for learning motion class patterns. 5096 LNCS: 365-374.

Rabilloud, T., M. Chevallet, et al. (2010). Two-dimensional gel electrophoresis in proteomics: Past, present and future. *Journal of Proteomics* 73(11): 2064-2077.

Siedlecki, W. and J. Sklansky (1989). A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters* 10(5): 335-347.

Szczypiski, P. M., M. Strzelecki, et al. (2009). MaZda-A software package for image texture analysis. *Computer Methods and Programs in Biomedicine* 94(1): 66-76.

Tamboli, A. S. and M. A. Shah (2011). A Generic Structure of Object Classification Using Genetic Programming. Communication Systems and Network Technologies (CSNT), *2011 International Conference* on.

Vapnik, V. N. (1979). Estimation of dependences based on empirical data [in Russian]. Nauka, English translation *Springer Verlang*, 1982.

APPENDIX

Table 1: Results separated in training and validation.

	% division	Accuracy	Sensitivity	Specificity	F
Training	70-30	0.9128(0.0103)	0.9450(0.0180)	0.8816(0.0131)	0.9141(0.0107)
	80-20	0.9183(0.0093)	0.9504(0.0118)	0.8866(0.0086)	0.9202(0.0097)
	100-0	0.9236(0.0035)	0.9593(0.0050)	0.8880(0.0072)	0.9262(0.0032)
Validation	70-30	0.9073(0.0058)	0.9442(0.0112)	0.8699(0.0118)	0.9110(0.0057)
	80-20	0.9165(0.0090)	0.9518(0.0107)	0.8812(0.0135)	0.9192(0.0085)
	100-0	0.9254(0.0040)	0.9580(0.0037)	0.8928(0.0068)	**0.9277(0.0037)**