Chapter 84
Semi-Supervised Clustering for the Identification of Different Cancer Types Using the Gene Expression Profiles

Manuel Martín-Merino
University Pontificia of Salamanca, Spain

ABSTRACT

DNA Microarrays allow for monitoring the expression level of thousands of genes simultaneously across a collection of related samples. Supervised learning algorithms such as k-NN or SVM (Support Vector Machines) have been applied to the classification of cancer samples with encouraging results. However, the classification algorithms are not able to discover new subtypes of diseases considering the gene expression profiles. In this chapter, the author reviews several supervised clustering algorithms suitable to discover new subtypes of cancer. Next, he introduces a semi-supervised clustering algorithm that learns a linear combination of dissimilarities from the a priory knowledge provided by human experts. A priori knowledge is formulated in the form of equivalence constraints. The minimization of the error function is based on a quadratic optimization algorithm. A L_2 norm regularizer is included that penalizes the complexity of the family of distances and avoids overfitting. The method proposed has been applied to several benchmark data sets and to human complex cancer problems using the gene expression profiles. The experimental results suggest that considering a linear combination of heterogeneous dissimilarities helps to improve both classification and clustering algorithms based on a single similarity.

INTRODUCTION

Classification techniques such as *k*-NN or Support Vector Machines (SVM) have been successfully applied to the identification of cancer samples using the gene expression profiles (Lanckriet, 2004; Martín-Merino, 2009b). This kind of supervised algorithms rely on a categorization of a subset of samples by human experts. Therefore, they are not able to discover new types of diseases which is usually more interesting for biologists. Besides, biological information provided by human experts is frequently very sparse and it is provided in the form of which pairs of objects are considered

DOI: 10.4018/978-1-4666-3604-0.ch084

similar. This kind of information cannot be incorporated into classification techniques.

Clustering algorithms group similar objects without any supervision by human experts. They are able to discover new types of diseases but classical algorithms cannot incorporate a priori knowledge provided by human experts.

Clustering algorithms depend critically on the choice of a good dissimilarity (Xing, 2003). A large variety of dissimilarities have been proposed in the literature (Cox, 2001). However, in real applications no dissimilarity outperforms the others because each dissimilarity reflects often different features of the data (Martín-Merino, 2009a). Instead of using a single dissimilarity it has been recommended in (Lanckriet, 2004; Martín-Merino, 2009b) to consider a linear combination of heterogeneous dissimilarities and data sources.

Therefore, new clustering algorithms should be developed that are able to adapt the metric to the problem at hand considering the a priori information provided by human experts in the form of similarity/dissimilarity constraints.

Several authors have proposed techniques to learn a linear combination of kernels (similarities) from the data (Lanckriet, 2004; Martín-Merino, 2009b; Soon Ong, 2005; Xiong, 2006). These methods are designed for classification tasks and assume that the class labels are available for the training set. However, for certain applications such as Bioinformatics, domain experts provide only incomplete knowledge in the form of which pairs of samples or genes are related (Huang, 2006). This a priory information can be incorporated via semi-supervised metric learning algorithms using equivalence constraints (Bar-Hillel, 2005). Thus, (Xing, 2003) proposed a distance metric learning algorithm that incorporates such similarity/dissimilarity information using a convex programming approach. The experimental results show a significant improvement in clustering results. However, the algorithm is based on an iterative procedure that is computationally intensive particularly, for high dimensional applications. To avoid this problem, (Bar-Hillel, 2005; Kwok, 2003; Wu, 2005) presented more efficient algorithms to learn a Mahalanobis metric. However, these algorithms are not able to incorporate heterogeneous dissimilarities and rely on the use of the Mahalanobis distance that may not be appropriate for certain kind of applications (Martín-Merino, 2009a; Martín-Merino 2009b).

The approach introduced in this chapter, considers that the integration of dissimilarities that reflect different features of the data should help to improve the performance of clustering and classification algorithms. To this aim, a linear combination of heterogeneous dissimilarities is learnt considering the relation between kernels and distances (Pekalska, 2001). A learning algorithm is proposed to estimate the optimal weights considering the similarity/dissimilarity constraints available. The method proposed is based on a convex quadratic optimization algorithm and incorporates a smoothing term that penalizes de complexity of the family of distances avoiding overfitting.

The metric learning algorithm proposed has been applied to a wide range of practical problems. The empirical results suggest that the method proposed helps to improve pattern recognition algorithms based on a single dissimilarity and a widely used metric learning algorithm proposed in the literature.

BACKGROUND

Supervised learning algorithms such as *k*-NN or SVM (Support Vector Machines) (Vapnik, 1998) have been applied to the classification of cancer samples with encouraging results (Martín-Merino, 2009b). However, classification algorithms are not able to discover new subtypes of diseases considering the gene expression profiles. This goal is often more interesting for human experts.

Clustering algorithms, allow us to discover new types of cancer that have not been previ-