Combination of Sparse and Wrapper Feature Selection from Multi-Source Data for Accurate Brain Tumor Typing

V. Metsis[1], O. C. Andronesi[2,3], H. Huang[1], M. N. Mindrinos[4], L. G. Rahme[5], F. Makedon[1], and A. A. Tzika[2,3]

[1]Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, United States, [2]NMR Surgical Laboratory, Dept. of Surgery, Harvard Medical School and Massachusetts General Hospital, Boston, MA, United States, [3]Athinoula A. Martinos Center of Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Boston, MA, United States, [4]Dept. of Biochemistry, Stanford University School of Medicine, Stanford, CA, United States, [5]Molecular Surgery Laboratory, Dept. of Surgery, Massachusetts General Hospital and Shriners Burn Institute, Harvard Medical School, Boston, MA, United States

Introduction

Previous research has shown the value of analyzing Gene Expression profiles and Magnetic Resonance Spectroscopy (MRS) data to type brain tumors [1, 2, 3]. Genomic and MRS data have been used separately and in combination to build robust classifiers to differentiate among different classes of brain tumors. In their recent work Metsis et al. [3] presented Machine Learning framework for feature selection and classification, which combines features from Affymetrix Microarray Gene Expression data and High Resolution Magic Angle Spinning (HRMAS) Proton (^{1}H) Magnetic Resonance Spectroscopy (MRS) data coming from the same subjects, to classify brain biopsy samples. It is shown that the combination of the two datasets significantly improves the classification accuracy compared to the accuracy achieved by each individual dataset. In this work we verify the advantage of combining the two above datasets and we introduce a new feature selection method based on Joint $\ell_{2,1}$-Norms Minimization which further improves classification accuracy in the multiclass problem.

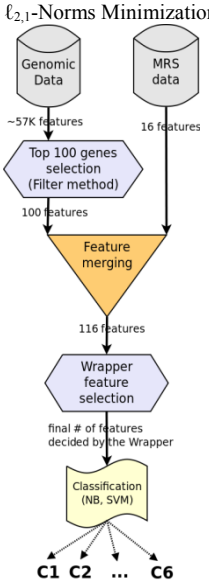

Fig. 1: Feature selection & classific. Framework.

Methods

The dataset used in this work consists of a total of 46 brain tumor and normal (control) biopsy samples coming from 16 different people. The HRMAS ^{1}H MRS and gene expression data were obtained ex vivo from the same samples, forming a set of 5 tumor classes and 1 control class. The tumor samples are analyzed as follows: 11 glioblastoma multiforme (GBM); 8 anaplastic astrocytoma (AA); 7 meningioma; 7 schwanoma; and 5 from adenocarcinoma. The 9 control samples come from epileptic surgeries. From the MRS data we extracted and used as features 15 significant metabolites: choline (Cho), phosphocholine (PC), glycerophosphocholine (GPC), phosphoethanolamine (PE), ethanolamine (Etn), γ-aminobutyric acid (GABA), n-acetyl aspartate (NAA), aspartate (Asp), alanine (Ala), polyunsaturated fatty acids (PUFA), glutamine (Gnl), glutamate (Glu), lactate (Lac), taurine (Tau) and lipids (Lip). From the gene expression profiles we used as features the full set of genes (~57K) provided by the Affymetrix human genome gene chip. We experimented with feature selection from both dataset types in order to reduce redundancy and noise before using them for classification. To achieve the highest possible classification accuracy we used a hybrid feature selection method which involved both filter [4] and wrapper [5] feature selection algorithms. We applied filter feature selection to select the top 100 genes in terms of their discriminative power between the different classes and then we applied the wrapper feature selection approach to further reduce the number of genes used for the final classification. This way we took utilized the advantages of each method and alleviated their disadvantages. The filter methods are much faster and can handle a big initial number of features, which is the case with the number of genes in the human genome, but they perform poorly when the final set of features to be selected is very small. Contrary the wrapper methods are much more accurate but their high computational cost prohibits their use in high dimensional features spaces. For the MRS data we only used wrapper feature selection since the small number of initial features allowed us to do so. Naïve Bayes [6] and SVM [7] classifiers were employed for classification. The framework used to combine the features from the two different datasets and perform the classification is illustrated in Fig.1. More details can be found in [3].

The main contribution of this paper lies in the application of a novel sparse filter feature selection method, namely $\ell_{2,1}$-Norms Minimization, which resulted in a significant increase in the classification accuracy compared to the previous results on the same datasets. This method reduces the feature dimensionality by performing sparsity regularization on the initial feature set which gives a high weight to the most discriminative features and small weight to the rest of them. The optimal weights (coefficients) are obtained by performing $\ell_{2,1}$-Norms Minimization on the linear regression objective function. The minimization problem to be solved is: $\min_W \frac{1}{\gamma}\|X^T W - Y\|_{2,1} + \|W\|_{2,1}$, where $X = [x_1, x_2, ..., x_n] \in R^{d \times n}$ is the data matrix, $Y = [y_1, y_2, ..., y_n] \in R^n$ is the vector of labels (classes) and $W \in R^{d \times c}$ is the matrix of coefficients to be computed. For more details on how to efficiently solve this optimization problem please refer to [8]. The other filter methods used were χ^2-statistic (χ^2), Information Gain (IG) and Relief-F (RF) [4].

Classifier / Dataset	NB	SVM
Mebolites only	Wrapper 72.34%	Wrapper 78.72%
Genes only	χ^2 + Wrapper 82.98%	$\ell_{2,1}$ + Wrapper 68.09%
Combined	$\ell_{2,1}$ + Wrapper **95.75%**	IG + Wrapper 80.85%

Table 1: Best results for each dataset and each classifier for the 6 class classification task. The feature selection method that achieved the higher accuracy along with the accuracy itself is shown in each table cell.

Results

To compare with the previous classification accuracy results reported on the given datasets [3] we followed the same experimentation process, but this time we replaced the previously used filter feature selection methods with our newly introduced one. Our findings showed, again, that the combination of data from two different sources yields higher accuracy compared to the accuracy that we obtain get by using each of the datasets separately. Also, same as before, our experiments reported perfect accuracy in the ability of the system to differentiate between tumor and non-tumor (normal) samples when the two datasets are combined. For the more difficult task of 6-class classification problem (5 tumor types + 1 normal) though, the use of the new feature selection method significantly increased the accuracy from the 87.23% that was the best previous performance to **95.75%**. This accuracy was achieved by performing a 10-Fold cross validation on the combined data using Naïve Bayes for wrapper feature selection and as classifier to do the final classification. The SVM achieved a relatively lower accuracy due to its inability to be successfully used as a wrapper feature section method because of the high computational complexity required for tuning its parameters. The final set of features that were selected by our system to achieve the above accuracy was a combination of 4 metabolites (Asp, Etn, GPC, PE) and 9 genes (ADM, CD24, ACTB, HSPA1B, CRYAB, MPZ, ABCA2, ID4, PTGDS). The discriminative power of this relatively small set of metabolites and genes may be suggesting that they can be used as possible Biomarkers related to the development of brain tumors and further investigation of their properties would be worthwhile.

Discussion

The results of this study support previous findings that the combination of heterogeneous datasets, in this case gene expression profiles and metabolite levels, boost the accuracy of brain tumor typing. In addition, the introduction of ℓ2,1-Norms Minimization feature selection method which here is used for gene selection, significantly increases the accuracy of the multiclass classification task. Our machine learning based feature selection and classification framework not only reduces the noise in the dataset and achieves high classification accuracy, but also, the final set of important features that it selects can be used as indicators of possible Biomarkers for the detection and typing of brain tumors.

References

1. Tzika et al., *International Journal of Molecular Medicine* 20: 199-208, 2007.
2. Andronesi et al., *International Journal of Oncology* 33: 1017-1025, 2008
3. Metsis et al, *Artificial Intelligence Applications and Innovations III*, pp. 233-240, 2009.
4. Guyon and Elisseeff, *The Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.
5. Kohavi and John, *Artificial intelligence*, vol. 97, no. 1, pp. 273-324, 1997.
6. Metsis et al., *Third Conference on Email and Anti-Spam (CEAS)*, 2006
7. Cristianini and Shawe-Taylor, *Support vector machines*, 2000
8. Nie et al, *Efficient and Rubust Feature selection via Joint l2,1-Norms Minimization*, NIPS 2010.