BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 5 2011, pages 736–738
doi:10.1093/bioinformatics/btq721

Data and text mining

Advance Access publication December 24, 2010

PUGSVM: a caBIG^{TM} analytical tool for multiclass gene selection and predictive classification

Guoqiang Yu[1], Huai Li[2], Sook Ha[1], Ie-Ming Shih[3,4,5], Robert Clarke[6], Eric P. Hoffman[7], Subha Madhavan[6], Jianhua Xuan[1] and Yue Wang[1,*]

[1]Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, [2]Bioinformatics Unit, RRB, National Institute on Aging, NIH, Baltimore, MD 21224, [3]Department of Gynecology and Obstetrics, [4]Department of Pathology and [5]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD 21231, [6]Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC 20057 and [7]Research Center for Genetic Medicine, Children's National Medical Center, Washington, DC 20010, USA

Associate Editor: John Quackenbush

ABSTRACT

Summary: Phenotypic Up-regulated Gene Support Vector Machine (PUGSVM) is a cancer Biomedical Informatics Grid (caBIG^{TM}) analytical tool for multiclass gene selection and classification. PUGSVM addresses the problem of imbalanced class separability, small sample size and high gene space dimensionality, where multiclass gene markers are defined by the union of one-versus-everyone phenotypic upregulated genes, and used by a well-matched one-versus-rest support vector machine. PUGSVM provides a simple yet more accurate strategy to identify statistically reproducible mechanistic marker genes for characterization of heterogeneous diseases.

Availability: http://www.cbil.ece.vt.edu/caBIG-PUGSVM.htm.

Contact: yuewang@vt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 2, 2010; revised on December 19, 2010; accepted on December 20, 2010

1 INTRODUCTION

Microarray gene expression studies provide new opportunities for the molecular characterization of heterogeneous diseases (Clarke *et al.*, 2008; Wang *et al.*, 2008). Multiclass gene selection is an imperative task for identifying phenotype-associated mechanistic genes and achieving accurate diagnostic classifications (Liu *et al.*, 2002). Most existing multiclass gene selection methods heavily rely on the direct extensions of two-class gene selection methods (Dudoit *et al.*, 2002; Golub *et al.*, 1999). However, simple extensions of binary discriminant analysis to multiclass gene selection are suboptimal and not well matched to the unique characteristics of a multicategory classification problem. Specifically, the existing methods will select gene subsets that preserve the distances of well-separated classes, while likely create error-prone large overlap between neighboring classes (Loog *et al.*, 2001).

In Phenotypic Up-regulated Gene Support Vector Machine (PUGSVM; see schematic flowchart and mathematic descriptions of the method in the Supplementary Material) (Yu *et al.*, 2010), we select genes that are highly expressed in one phenotype relative to each of the remaining phenotypes, namely One-Versus-Everyone Phenotypic Up-regulated Genes (OVEPUG). OVEPUG provides evenhanded gene resources for discriminating both neighboring and well-separated classes, and intends to assure the statistical reproducibility and biological plausibility of the selected genes. OVEPUG is well suited for small sample size problems (a concern with many current microarray datasets). We also implement other competing peer methods for multiclass gene selection, including signal-to-noise ratio (SNR) (Golub *et al.*, 1999), *t* statistic (*t*-stat) (Liu *et al.*, 2002), pooled ratio of between-groups to within-groups sum of squares (BW) (Dudoit *et al.*, 2002) and SVM-based recursive feature elimination (SVMRFE) (Li and Yang, 2005). In the classification component of PUGSVM, we implement the OVRSVM committee classifier, which has proved highly successful in multicategory classifications with finite or limited amounts of high-dimensional data in real-world applications (Rifkin and Klautau, 2002). We also include several other popular multicategory classifiers, including *k* nearest neighborhood (kNN) (Golub *et al.*, 1999), naïve Bayes classifier (NBC) (Liu *et al.*, 2002) and OVOSVM (Liu *et al.*, 2005).

PUGSVM was developed through the caBIG (cabig.nci.nih.gov) In Silico Research Centers of Excellence (ISRCE) effort, and offers users across the broader cancer research community a unique yet effective tool for identifying multiclass gene markers and predicting clinical outcomes in cancer treatment. PUGSVM is an open-source software package. The Java and Matlab codes and documents are freely available at the authors' web site, enabling users to easily modify the program and add new functions or extensions (http://www.cbil.ece.vt.edu/caBIG-PUGSVM.htm).

2 DESCRIPTION

2.1 Software

The components of PUGSVM and their input/output relationships are illustrated in Figure 1. We use caBIG existing tools to load, preprocess and normalize gene expression data from in-house (i.e. Georgetown Database of Cancer; GDOC) or public databases

[*] To whom correspondence should be addressed.

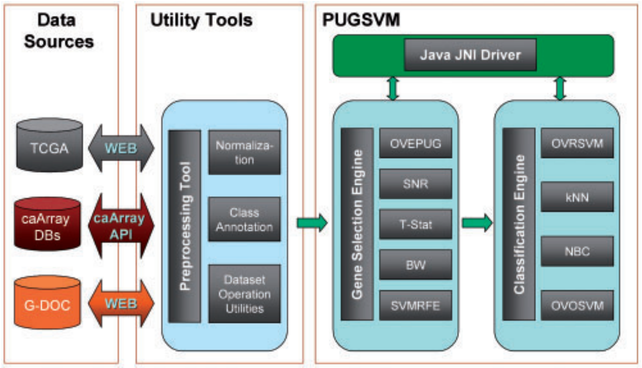


Fig. 1. The components and input/output of PUGSVM.

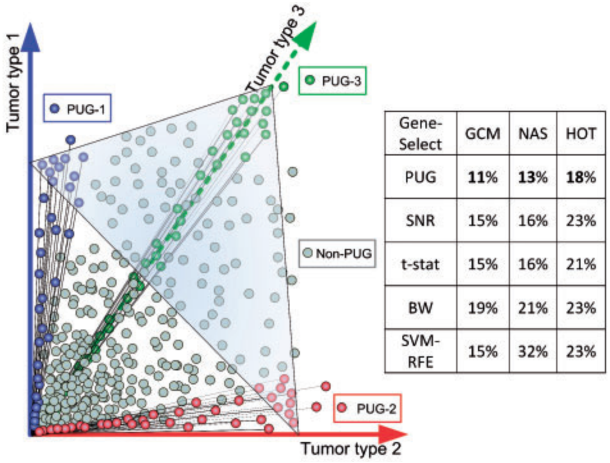
(e.g. caArray, TCGA). The processed data with class labeling are fed to the gene selection component. The selected PUGs are then used to train and test the classifiers for predictive classification. The output of PUGSVM is a set of gene markers with generalizable performance.

The OVEPUG and other algorithms in the gene selection component are implemented in Matlab. We use Matlab compiler to generate C++ shared function libraries. The OVRSVM algorithm and other classifiers are implemented in C++ with simple calling interfaces. The user interface is implemented in Java, and C++ shared libraries are called from Java using the Java Native Interface. PUGSVM has been tested on Microsoft Windows and Linux platforms. Users can run PUGSVM directly on a computer without an installed version of MATLAB.

2.2 Case study

We applied PUGSVM on the benchmark Global Cancer Map dataset (Ramaswamy *et al*., 2001) that is widely used for evaluating multicategory classification algorithms. Besides OVRSVM coupled with OVEPUGs, the combinations of competing gene selection methods (SNR, *t*-stat, BW and SVMRFE) with OVRSVM are also tested for comparisons. Figure 2 shows that OVEPUGs significantly improves the overall multicategory classification compared to all other combinations. Furthermore, gene markers selected by PUGSVM are confirmed to be important tumor-associated genes by literature survey and our domain experts. For example, the top 10 PUGs associated with prostate cancer include several genes strongly associated with prostate cancer including prostate-specific antigen and its alternatively spliced form 2 and prostatic secretory protein 57. Estrogen receptor α (ESR1), a PUG marker for uterine cancer, is known to be overexpressed in human uterine cancer, and the Hox7 gene, another PUG marker, is known to contribute to uterine function in cow and mouse models, especially at the onset of pregnancy.

Norway AScites (NAS) dataset is a unique dataset with samples taken from ascites. We applied PUGSVM on NAS data and obtained a significantly improved classification performance than competing methods as shown in Figure 2. Several top-ranking gene products identified by OVEPUG have been well established as tumor-type specific markers and many of them have been used in clinical diagnosis. For example, mucin 16, also known as CA125, is a food and drug administration (FDA)-approved serum marker to monitor disease progression and recurrence in ovarian cancer patients. Fatty acid synthase (FASN) is often upregulated in breast cancer and this enzyme is amenable for drug targeting using FASN inhibitors, suggesting that it can be used as a therapeutic target in breast cancer.



Gene-Select	GCM	NAS	HOT
PUG	**11%**	**13%**	**18%**
SNR	15%	16%	23%
t-stat	15%	16%	21%
BW	19%	21%	23%
SVM-RFE	15%	32%	23%

Fig. 2. Illustration of PUG scheme (left) and the classification error rates with different gene selection methods on three datasets (right).

Additional case studies on the Human Ovarian Tumors, National Cancer Institute 60 cancer cell lines dataset, University of Michigan cancer dataset, Central Nervous System tumors and Muscular Dystrophy dataset can be found in the Supplementary Material.

3 DISCUSSION

PUGSVM is a comprehensive open-source tool that consists of interconnected components for multiclass gene selection and predictive classification. The core algorithms in the package are OVEPUG and OVRSVM, which address several critical yet subtle issues in molecular characterization of heterogeneous diseases for both biological research and clinical applications. PUGSVM emphasizes the statistical reproducibility of the selected gene markers under small sample size, supported by their biologically plausible interpretations. Several competing gene selection and classification methods are also incorporated into the package to demonstrate the superior performance of PUGSVM via objective comparisons.

Through the caBIG ISRCE effort, we plan to adapt PUGSVM to identify gene markers as druggable targets and predict breast cancer resistance and recurrence after tamoxifen treatment. We have established several workflow pipelines for using PUGSVM. For example, we first download breast cancer gene expression datasets from TCGA and G-DOC, and then normalize and label the data using caBIG tools, such as GenePattern. We feed the processed data to PUGSVM. The gene candidates identified by PUGSVM will be sent to VIsual Statistical Data Analyzer for visualization, mapping onto known signaling pathways and further analysis. We are currently working to create Taverna workflow modules for all analyses and making them publicly available to the cancer community.

Funding: This work is supported in part by the National Institutes of Health, under Contract No. HHSN261200800001E and Grants CA109872, CA149147, CA139246, NS029525.

Conflict of Interest: none declared.

REFERENCES

Clarke,R. *et al*. (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat. Rev. Cancer*, **8**, 37–49.

Dudoit,S. *et al*. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science,* **286**, 531–537.

Li,F. and Yang,Y. (2005) Analysis of recursive gene selection approaches from microarray data. *Bioinformatics*, **21**, 3741–3747.

Liu,H. *et al*. (2002) A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Inform.*, **13**, 51–60.

Liu,J.J. *et al*. (2005) Multiclass cancer classification and biomarker discovery using GA-based algorithms. *Bioinformatics*, **21**, 2691–2697.

Loog,M. *et al*. (2001) Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *IEEE Trans. Pattern Anal. Mach. Intell.*, **23**, 762–766.

Ramaswamy,S. *et al*. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.

Rifkin,R. and Klautau,A. (2002) In defense of one-vs-all classification. *J. Mach. Learn. Res.*, **5**, 101–141.

Wang,Y. *et al*. (2008) Approaches to working in high-dimensional data spaces: gene expression microarrays. *Br. J. Cancer*, **98**, 1023–1028.

Yu,G. *et al*. (2010) Matched gene selection and committee classifier for molecular classification of heterogeneous diseases. *J. Mach. Learn. Res.*, **11**, 2141–2167.