Enliven Archive
Intensifying Thoughts

Review Article www.enlivenarchive.org **Enliven: Journal of Genetic, Molecular and Cellular Biology**

ISSN: 2379-5700

Using Feature Selection and Transductive SVM to Predict the Gene-Expression-Based Cancer Subtypes

Janet Titus

Department of Genetics, Stellenbosch University, JC Smuts Building, De Beer Rd, Stellenbosch Central, Stellenbosch, 7600, South Africa

***Corresponding author**: Janet Titus, Department of Genetics, Stellenbosch University, JC Smuts Building, De Beer Rd, Stellenbosch Central, Stellenbosch, 7600, South Africa, E-mail: titusnet@outlook.com

Received Date: 05th July 2018
Accepted Date: 21st August 2018
Published Date: 01st September 2018

Citation: Titus J (2018) Using Feature Selection and Transductive SVM to Predict the Gene-Expression-Based Cancer Subtypes. Enliven: J Genet Mol Cell Biol 5(2): 005.

The ability to conduct advance gene expression profiling owing to developments in microarray technology has greatly enabled scientists to perform cancer classification or to predict the diverse types of cancer and their treatment options [1]. The researcher exploited the microarray cancer data to classify tissue samples into their subtypes or malignant and benign [2]. Furthermore, the data is valuable since it can be relied to explore all the cancer subtypes to identify the likely gene markers which assist in achieving positive diagnosis of the exact type of cancer as shown below [3].

Nonetheless, designing an appropriate classifier has been hindered by the minor sizes of sample. In most cases, data obtained from a sample with medical follow-up (the labeled data) only works with conventional supervised classifiers [4]. Alternatively, the researchers chose to disregard the massive number of microarray data that failed to contain sufficient follow-up details also known as the labeled data [5]. Latest empirical studies focusing on cancer diagnosis indicate that the unlabeled data can be used to develop the semi-supervised learning technique which greatly assist in improving prediction accuracy [6]. Actually, the method is considered as very effective in providing solutions to diverse biological challenges for instance, the discovery cancer subtypes using gene expression, predicting transcription factor–gene interaction, as well as classification of protein [7]. Therefore, they incorporated a new method that combined transductive support vector machine (TSVM) and gene (feature) selection [8]. Additionally, the researchers prove that transductive support vector machine enhances the correctness of prediction in comparison to the typical inductive support vector machines (ISVM) as well as reveal that there is a possibility of identifying the gene markers [9].

The prospective gene markers were obtained using a forward greedy search algorithm. Additionally, they designed a transductive support vector machine by exploiting the microarray data's selected genes [10]. According to the article, such areas as identification of gene markers and classification of semi-supervised cancer have greatly benefited owing to the success of the suggested method in comparison to using the low-density separation technique and the inductive support vector machines [11]. Classifying the diverse kind of tumor is considered as a critical first step towards drug development and diagnosis of cancer disease [12]. Nonetheless, the clinical cancer research is negatively affected by problems experienced while predicting the prognosis while discovering the tumor [13]. Predicting correctly the varied types of tumor might assists in minimizing toxicity among patients as well as offering superior treatment of cancer [14, 15].

Microarray technology has enable researchers to conduct experiments across diverse conditions to investigate the gene expression profiles which has helped in predicting varied subtypes of cancer and the treatment options [16]. Nonetheless, the minor sample size used in microarray based-cancer inquiries is a big challenge with regards to getting comprehensive and correct prediction models [17]. The studies are usually restricted by the lack of sample besides being costly and time consuming. The researchers were successful in designing a model of predicting human cancers and drug option using gene expression profiles [18]. The TSVM was shown to be effective especially those which employ two feature selection methods as part of the transductive inference learning framework [19]. The selection of samples in the transductive support vector machines is based on geometric assessment of the feature space; besides, the training sets included support vector-like samples that had comprehensive details. Importantly, the transductive support vector machine being suggested by the researchers can be considered as an important process that helps to correctly describe the hyper-plane consistent with the transductive procedure that focuses on integrating the training and the unlabeled samples [20].

By using the unlabeled gene expression data in the transductive learning technique as recommended in the article, the researchers were able to attain positive empirical accomplishment [21]. Nonetheless, in case of varied distributions being followed by the unlabeled and labeled data, integration of the unlabeled data might have led to achievement of an undesirable performance [22, 23].

The study findings prove that the semi-supervised learning can greatly contribute towards tackling the current clinical challenges. Furthermore, the empirical findings demonstrate that the suggested method is more effective in comparison to the LDS and the ISVM [24]. In addition, as part of the expanded scope of future studies, the researchers propose the need to rely on the fuzzy rough set theory in order to establish significant gene makers [25]. Besides, they seek to improve the semi-supervised or transductive learning by introducing the theory to enhance the efficiency and effectiveness of the suggested method.

Table 1: Gene Markers found in the Cancer Datasets [1]

Gene Image ID	**Description**
Leukemia dataset:	
M27891_at	Cystatin C (amyloid angiopathy and cerebral hemorrhage), CST3
Y07604_at	Non-metastatic cells 4, protein expressed in
SRBCT dataset:	
784224	Fibroblast growth factor receptor 4
812105	Transmembrane protein
207274	Human DNA for insulin-like growth factor II (IGF-2); exon 7 and additional ORF
782811	High mobility group (nonhistone chromosomal) protein isoforms I and Y
344134	Immunoglobulin lambda-like polypeptide 3
MLL dataset:	
31375_at	-
31385_at	Ribosomal protein L28
31394_at	Serpin Peptidase inhibitor, clade I (pancpin), member 2
31441_at	ribonuclease, RNase A family, 2 (liver, eosinophil-derived neurotoxin) pseudogene
DLBCL dataset	
M59829_at	Heat shock 70kDa protein I -like
X53961_at	Transferrin, peptidase S60 transferrin lactoferrin
U46006_s_at	Cysteine and glycine-rich protein 2
X85785_mal_at	Duffy blood group, Chemokin receptor

References

1. Maulik U, Mukhopadhyay A, Chakraborty D (2013) Gene-Expression-Based Cancer Subtypes Prediction through Feature Selection and Transductive SVM. IEEE Trans Biomed Eng. 60: 1111-1117.
2. Bandyopadhyay S, Maulik U, Debadyuti Roy (2008) Gene Identification: Classical and Computational Intelligence Approaches. IEEE Transactions on Systems 38: 55-68.
3. Bandyopadhyay S, Mitra R, Maulik U, Zhang MQ (2010) Development of the Human Cancer Microrna Network. Silence 1: 6.
4. Sauravjoyti Sarmah, Dhruba K. Bhattacharyya (2010) An Effective Technique for Clustering Incremental Gene Expression Data. IJCSI International Journal of Computer Science 7: 31-41.
5. Xiaojin Zhu, Zoubin Ghahramani, John D. Lafferty (2003) Semi-supervised Learning Using Gaussian Fields and Harmonic Functions. In Proceedings of the 20th International conference on Machine learning (ICML-03): 912-919.
6. Mikhail Belkin, Partha Niyogi, Vikas Sindhwani (2006) Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. Journal of Machine Learning Research 7: 2399-2434.
7. Weston J, Leslie C, Ie E, Zhou D, Elisseeff A (2005) Semi-supervised Protein Classification Using Cluster Kernels. Bioinformatics 21: 3241-3247.
8. Swathi M (2017) Clustering Enhancement Using Similarity Indexing to Reduce Entropy. Enliven: Bioinform 4: 001.
9. Kristin P Bennett, Ayhan Demiriz (1999) Semi-Supervised Support Vector Machines. Advances in Neural Information Processing Systems 3: 368-374.
10. Vapnik Vladimir, A Sterin (1977) On Structural Risk Minimization or Overall Risk in a Problem of Pattern Recognition. Automation and Remote Control 10: 1495-1503.
11. Avrim L Blum, Pat Langley (1997) Selection of Relevant Features and Examples in Machine Learning. Artificial Intelligence 97: 245-271.
12. Steinfeld I, Navon R, Ardigò D, Zavaroni I, Yakhini Z (2008) Clinically Driven Semi-Supervised Class Discovery in Gene Expression Data. Bioinformatics 24: 90-97.
13. Yuhua Qian, Jiye Liang, Witold Pedrycz, Chuangyin Dang (2010) Positive Approximation: An Accelerator for Attribute Reduction in Rough Set Theory. Artificial Intelligence 174: 597-618.

14. Ujjwal Maulik, Anirban Mukhopadhyay (2010) Simulated Annealing Based Automatic Fuzzy Clustering Combined with ANN Classification for Analyzing Microarray Data. Computers & Operations Research 37: 1369-1380.
15. Burges Christopher JC (1998) A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2: 121-167.
16. Swathi M (2017) Drug Prediction of Cancer Genes Using SVM. Enliven: Pharmacovigilance and Drug Safety 4(2): 001.
17. Ujjwal Maulik, Anirban Mukhopadhyay (2010) Simulated Annealing Based Automatic Fuzzy Clustering Combined with ANN Classification for Analyzing Microarray Data. Computers & Operations Research 37: 1369-1380.
18. Swathi M (2018) Enhancement of K-Mean Clustering for Genomics of Drugs. Enliven: J Genet Mol Cell Biol 5: 001.
19. Olivier Chapelle, Alexander Zien (2005) Semi-Supervised Classification by Low Density Separation. AISTATS 2: 57-64.
20. Gentles AJ, Plevritis SK, Majeti R, Alizadeh AA (2010) Association of a leukemic stem cell gene expression signature with clinical outcomes in acutemyeloid leukemia. Jama 304: 2706-2715.
21. Yisong Chen, Guoping Wang, Shihai Dong (2003) Learning with Progressive Transductive Support Vector Machine. Pattern Recognition Letters 24: 1845-1855.
22. Manoranjan Dash, Huan Liu (2003) Consistency-Based Search in Feature Selection. Artificial Intelligence 151: 155-176.
23. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286: 531-537.
24. Dupuy A, Simon RM (2007) Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting. J Natl Cancer Inst. 99: 147-157.
25. Ernst J, Beg QK, Kay KA, Balázsi G, Oltvai ZN, Bar-Joseph Z (2008) A Semi-Supervised Method for Predicting Transcription Factor–Gene Interactions in Escherichia Coli. PLoS Comput Biol. 4: e1000044.
26. Ein-Dor L, Zuk O, Domany E (2006) Thousands of Samples Are Needed To Generate A Robust Gene List For Predicting Outcome In Cancer. Proc Natl Acad Sci U S A 103: 5923-5928.
27. Mukhopadhyay A, Bandyopadhyay S, Maulik U (2010) Multi-Class Clustering of Cancer Subtypes throughSvm Based Ensemble of Pareto-Optimal Solutions for Gene Marker Identification. PloS one 5: 13803.
28. Maulik Ujjwal (2011) Analysis of Gene Microarray Data in a Soft Computing Framework. Applied Soft Computing 11: 4152-4160.