RESEARCH ARTICLE

Evaluation of Machine Learning Algorithm Utilization for Lung Cancer Classification Based on Gene Expression Levels

Maxim D Podolsky[1]*, Anton A Barchuk[2], Vladimir I Kuznetcov[3], Natalia F Gusarova[1], Vadim S Gaidukov[1], Segrey A Tarakanov[1]

Abstract

Background: Lung cancer remains one of the most common cancers in the world, both in terms of new cases (about 13% of total per year) and deaths (nearly one cancer death in five), because of the high case fatality. Errors in lung cancer type or malignant growth determination lead to degraded treatment efficacy, because anticancer strategy depends on tumor morphology. Materials and Methods: We have made an attempt to evaluate effectiveness of machine learning algorithms in the task of lung cancer classification based on gene expression levels. We processed four publicly available data sets. The Dana-Farber Cancer Institute data set contains 203 samples and the task was to classify four cancer types and sound tissue samples. With the University of Michigan data set of 96 samples, the task was to execute a binary classification of adenocarcinoma and non-neoplastic tissues. The University of Toronto data set contains 39 samples and the task was to detect recurrence, while with the Brigham and Women's Hospital data set of 181 samples it was to make a binary classification of malignant pleural mesothelioma and adenocarcinoma. We used the k-nearest neighbor algorithm (k=1, k=5, k=10), naive Bayes classifier with assumption of both a normal distribution of attributes and a distribution through histograms, support vector machine and C4.5 decision tree. Effectiveness of machine learning algorithms was evaluated with the Matthews correlation coefficient. Results: The support vector machine method showed best results among data sets from the Dana-Farber Cancer Institute and Brigham and Women's Hospital. All algorithms with the exception of the C4.5 decision tree showed maximum potential effectiveness in the University of Michigan data set. However, the C4.5 decision tree showed best results for the University of Toronto data set. Conclusions: Machine learning algorithms can be used for lung cancer morphology classification and similar tasks based on gene expression level evaluation.

Keywords: Computer aided diagnosis - lung cancer - ROC curve - data set - classifiers - gene expression

Asian Pac J Cancer Prev, **17 (2)**, 835-838

Introduction

According to the International Agency for Research on Cancer there were about 13 % (1,825 thousand) of new lung cancer cases of the total number of new cancer cases and about 19.4 % (1,590 thousand) deaths of the total number of deaths owing to lung cancer in the world in 2012. In the structure of cancer pathology lung cancer takes first place for men and third place for women (Ferlay et al., 2015).

Key elements in reduction of mortality rate among lung cancer carriers are early detection, accurate determination of cancer histological type and adequate treatment. Errors in lung cancer type or, in general, malignant growth type determination lead to treatment efficiency degradation, because anticancer strategy depends on tumor morphology (morphogenesis). For example, early malignant pleural mesothelioma is optimally treated by extrapleural pneumonectomy followed by radiochemotherapy, whereas metastatic lung cancer is cured by chemotherapy (Pass, 2001). At that, lung cancer five-year survival rates remain low, for instance, in South Korea they reached 20.7 % in 2007-2011 (Jung et al., 2014).

At present it is optimal to use machine learning methods to ascertain a definite diagnosis. Their final aim is to obtain trained algorithms which compute type and developmental character of malignant growth by usage of one or several classification attributes. These algorithms can be used by clinicians as auxiliary tools to process huge amounts of patient data for establishing diagnosis (Sun et al., 2013; Yu et al., 2015).

In population screening machine learning methods are used to differentiate between benign and malignant lung nodules based on low-dose computed tomography

*[1]ITMO University, [2]NN Petrov Research Institute of Oncology of the USSR Ministry of Health, [3]KBST ITMO LLC, Saint Petersburg, Russia *For correspondence: max.d.podolsky@gmail.com*

Table 1. Comparing of Machine Learning Methods for Dana-Farber Cancer Institute Data Set Where Five Classes are Presented

Machine learning classifier	Class Adenocarcinoma		Class Squamous cell lung carcinoma		Class Pulmonary carcinoid		Class Small-cell lung carcinoma		Class Healthy lung samples	
	AUC	MCC	AUC	MCC	AUC	MCC	AUC	MCC	AUC	MCC
k-NN, k=1	0.87	0.74	0.84	0.74	0.99	0.91	0.95	0.57	0.78	0.73
k-NN, k=5	0.94	0.75	0.98	0.73	0.99	0.88	0.98	0.63	0.93	0.83
k-NN, k=10	0.95	0.73	0.97	0.80	0.99	0.88	0.99	0.63	0.93	0.83
NB_normal	0.84	0.64	0.89	0.59	0.96	0.85	0.67	0.57	0.90	0.75
NB_histogram	0.82	0.61	0.87	0.59	0.95	0.81	0.67	0.57	0.90	0.74
SVM	0.88	0.94	0.94	0.89	1.00	1.00	0.97	0.91	0.98	0.90
C4.5 Decision Tree	0.92	0.83	0.87	0.71	0.99	0.97	0.83	0.65	0.95	0.93

(Wang et al., 2013), which is considered as a widespread standard in detecting and analysis of lung diseases. In case of expected tumors sampled it is possible to use technics on the basis of gene activity pattern in affected cells (Han et al., 2013; Liu et al., 2013). Thus gene expression levels can be used as classification attributes which characterize production rate of protein in lung tumor cells compared with healthy cells (Cheng et al., 2012).

To accomplish the task of cancer type classification the following algorithms are typically applied, such as Support Vector Machines, Random Forests, Decision Tree, Boosting, K-Nearest Neighbor, LASSO, neural networks (Lei Win et al., 2014; Cai et al., 2015). At that, effectiveness of various algorithms differs depending on analyzed data sets. To evaluate effectiveness of the algorithms and compare them it is accepted to use Receiver Operating Characteristic curve (ROC curve) and Matthews Correlation Coefficient (MCC) as a measure of the quality of binary (two-class) as well as non-binary classifications (Baldi et al., 2000).

Materials and Methods

Materials

To evaluate effectiveness of several machine learning algorithms we have processed four publicly available data sets related to gene expression:

i) Dana-Farber Cancer Institute, Harvard Medical School (Bhattacharjee et al., 2001); Consists of 203 samples: 139 correspond with adenocarcinoma, 21 -squamous cell lung carcinoma, twenty -pulmonary carcinoids, six -small-cell lung carcinoma, seventeen -healthy lung samples. Each sample is described by 12600 gene expression levels. Research task for this data set was to classify cancer types.

ii) University of Michigan (Beer et al., 2002); Consists of 96 samples: 86 -primary adenocarcinoma (where 67 -stage I, nineteen -stage III), ten -non-neoplastic tissue. Each sample is presented by expression levels of 7,129 genes. The task was to detect adenocarcinoma.

Samples of primary tumor and adjacent non-neoplastic tissue were taken during surgical intervention from May 1994 to June 2000 in the University of Michigan Hospital. Peripheral portions of resected lung carcinomas were sectioned, evaluated by a study pathologist and compared with routine H&E sections of the same tumors, and utilized for mRNA isolation. Regions chosen for analysis contained a tumor cellularity greater than 70%, no mixed histology, potential metastatic origin, extensive lymphocytic infiltration or fibrosis.

iii) University of Toronto, Ontario, Canada (Wigle et al., 2002)

Consists of 39 samples of non-small cell lung cancer. Twenty four samples correspond to patients with lung cancer recurrence (stage I -eight patients, stage II -thirteen patients, stage III -three patients). The remaining fifteen patients are disease-free (stage I -ten patients, stage II -two patients, stage III -free patients). The two groups were broadly similar in distribution of age and sex. Each sample is presented by expression levels of 2,880 genes. The task was to detect recurrences.

The samples were taken during lobectomy or pneumectomy of patients examined in University of Toronto, then snap-frozen and placed to liquid nitrogen to preserve them. Adenocarcinoma was confirmed in nineteen patients, squamous cell carcinoma -in fourteen, the rest six patients had adenosquamous carcinoma, large cell undifferentiated carcinoma or carcinoid tumor.

Patients were under observation for more than a year, on the average -around 26 months for patients with recurrence and 24 months for the rest.

iv) Brigham and Women's Hospital, Harvard Medical School (Gordon et al., 2002)

Consists of 181 samples of malignant tissue, where 31 -malignant pleural mesothelioma and 150 -adenocarcinoma. Samples were divided in two sets: training (sixteen samples of each cancer type) and testing (the remaining 149 samples). Each sample is presented by expression levels of 12,533 genes. The task is to make a binary classification of malignant pleural mesothelioma and adenocarcinoma.

Table 2. Averaged AUC and MCC of Dana-Farber Cancer Institute Data Set

Machine learning classifier	Averaged AUC	Averaged MCC
k-NN, k=1	0.87	0.75
k-NN, k=5	0.96	0.77
k-NN, k=10	0.97	0.76
NB_normal	0.85	0.66
NB_histogram	0.84	0.63
SVM	0.91	0.93
C4.5 Decision Tree	0.92	0.83

AUC, area under receiver operating characteristic curve; MCC, Matthews correlation coefficient

Table 3. Comparing of Machine Learning Methods for Data Sets of University of Michigan, University of Toronto, and Brigham and Women's Hospital

Machine learning classifier	University of Michigan		University of Toronto		Brigham and Women's Hospital	
	AUC	MCC	AUC	MCC	AUC	MCC
k-NN, k=1	1.00	1.00	0.67	0.24	0.98	0.89
k-NN, k=5	1.00	1.00	0.58	0.08	1.00	0.96
k-NN, k=10	1.00	1.00	0.54	0.15	0.99	0.84
NB_normal	1.00	1.00	0.72	0.38	0.97	0.80
NB_histogram	1.00	1.00	0.63	0.19	0.96	0.89
SVM	1.00	1.00	0.70	0.41	0.99	0.97
C4.5 Decision Tree	0.99	0.94	0.83	0.67	0.78	0.40

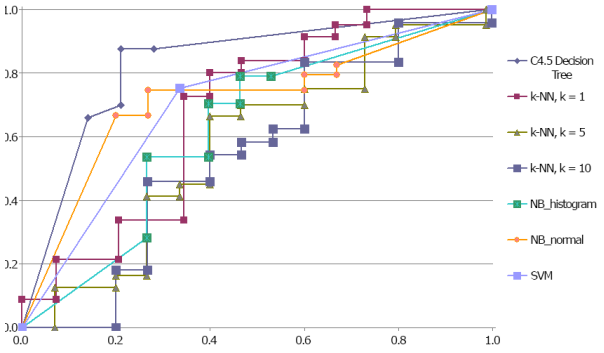

Figure 1. ROC Curves for University of Toronto Data Set. Abbreviations: ROC, receiver operating characteristic

Samples were taken and snap-frozen during surgical operations from 1993 to 2001 in Brigham and Women's Hospital, Boston, MA, USA.

All data sets can be downloaded using the reference: http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html

Methods

We have used seven machine learning algorithms or their versions to analyze the data sets:

i) k-nearest neighbors algorithm (k-NN, k=1, k=5, k=10); *ii*) Naive Bayes classifier both with assumption of the normal distribution of attributes (NB_normal) and distribution through histograms (NB_histogram); *iii*) Support vector machine (SVM); *iv*) C4.5 decision tree.

To train the algorithms using data sets of Dana-Farber Cancer Institute, University of Michigan and University of Toronto 10-fold cross validation was used. For Brigham and Women's Hospital data set we have used training and testing samples that have been already prepared. After ROC curves construction Area under ROC (AUC) and MCC were calculated.

Results

Dana-Farber Cancer Institute, Harvard Medical School data set: The data set contains the samples of five classes. Due to increasing of degrees of freedom k*(k-1) we constructed the curve and calculated MCC for each class while combining other classes and labeling them as "not considered class" (Table 1). In Table 2 related averaged results are shown.

Data sets from University of Michigan, University of Toronto, Brigham and Women's Hospital have binary classification and are summarized in table 3. ROC curves of University of Toronto data set are depicted on Figure 1.

Discussion

It is expected to have false-positive or false-negative results of differentially expressed genes due to the noisiness and scatter of processed data. To acquire accurate qualitative and quantitative data it is necessary to analyze experimental results carefully.

Support vector machine algorithm showed best results for Dana-Farber Cancer Institute (MCC 0.93) and Brigham and Women's Hospital data sets (MCC 0.97). At that k-nearest neighbors with k = 5 showed MCC 0.96 for the second data set. High values prove that SVM based on assessment of gene expression levels can be used to classify lung cancer by histological types, as well as classify adenocarcinoma and mesothelioma. Obtained data confirm results of the study Li et al. (2014) where SVM showed high accuracy in adenocarcinoma and squamous cell lung carcinoma classification. However, SVM showed second result after Bayes tree algorithm in identification and validation of the methylation biomarkers of non-small cell lung cancer (Guo et al., 2015). In addition it is effectively used to predict lung cancer type between small-cell one and non-small cell one, for example, in study Hosseinzadeh et al. (2013) SVM showed the best accuracy in analysis of protein attributes.

All algorithms except C4.5 decision tree (one classification error) were capable to accurately distinguish between adenocarcinoma and healthy lung in University of Michigan data set. However, C4.5 decision tree showed best result (MCC 0.67) in University of Toronto data set. The reason for lower effectiveness of other algorithms can be small quantity of the samples.

In conclusion among compared machine learning algorithms SVM tends to be the most appropriate auxiliary tool in lung cancer screening, while others showed sufficient effectiveness to be used in the tasks of gene expression levels assessment. It gives the opportunity to predict tumor growth and its metastasis with improved performance decreasing burden on clinicians determining the diagnosis. Machine learning algorithms can be used to substantially (15–25%) improve the accuracy of predicting cancer susceptibility, recurrence and mortality (Cruz and Wishart, 2006).

Acknowledgements

This paper is sponsored by the Ministry of Education and Science (Minobrnauka) of the Russian Federation within the project RFMEFI57814X0008.

References

Baldi P, Brunak S, Chauvin Y, Andersen CAF, and Nielsen H (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412-24.

Beer DG, Kardia SLR, Huang CC, et al (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*. **8**, 816-24.

Bhattacharjee A, Richards WG, Staunton J, et al (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, **98**, 13790-5.

Cai Z, Xu D, Zhang Q, et al (2015). Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Mol BioSyst*, **11**, 791-800.

Cheng P, Cheng Y, Li Y, et al (2012). Comparison of the gene expression profiles between smokers with and without lung cancer using RNA-Seq. *Asian Pac J Cancer Prev*, **13**, 3605-9.

Cruz JA, and Wishart DS (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, **2**, 59-77.

Ferlay J, Soerjomataram I, Dikshit R, et al (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Intern J Cancer*, **136**, 359-86.

Gordon GJ, Jensen RV, Hsiao LL, et al (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, **62**, 4963-7.

Guo S, Yan F, Xu J, et al (2015). Identification and validation of the methylation biomarkers of non-small cell lung cancer (NSCLC). *Clinical Epigenetics*, **7**, 3.

Han Y, Wang XB, Xiao N, and Liu ZD (2013). mRNA Expression and Clinical Significance of ERCC1, BRCA1, RRM1, TYMS and TUBB3 in postoperative patients with non-small cell lung cancer. *Asian Pac J Cancer Prev*, **14**, 2987-90.

Hosseinzadeh F, Kayvan Joo AH, Ebrahimi M, Goliaei B (2013). Prediction of lung tumor types based on protein attributes by machine learning algorithms. *Springer Plus*, **2**, 238.

Jung KW, Won YJ, Kong HJ, et al (2014). Cancer statistics in korea: incidence, mortality, survival, and prevalence in 2011. *Cancer Res Treat*, **46**, 109-23.

Lei Win S, Htike ZZ, Yusof F, Noorbatcha AI (2014). Gene expression mining for predicting survivability of patients in earlystages of lung cancer. *Int J Bioinformatics Biosciences*, **4**, 1-9.

Li J, Li D, Wei X, Su Y (2014). In silico comparative genomic analysis of two non-small cell lung cancer subtypes and their potentials for cancer classification. *Cancer Genomics Proteomics*, **11**, 303-10.

Liu M, Pan H, Zhang F, et al (2013). Screening of differentially expressed genes among various TNM stages of lung adenocarcinoma by genomewide gene expression profile analysis. *Asian Pac J Cancer Prev*, **14**, 6281-6.

Pass HI (2001). Malignant pleural mesothelioma: surgical roles and novel therapies. *Clinical Lung Cancer*, **3**, 102-7.

Sun T, Wang J, Li X, et al (2013). Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set. *Computer Methods Programs Biomedicine*, **111**, 519-24.

Wang JJ, Wu HF, Sun T, et al (2013). Prediction models for solitary pulmonary nodules based on curvelet textural features and clinical parameters. *Asian Pac J Cancer Prev*, **14**, 6019-23.

Wigle DA, Jurisica I, Radulovich N, et al (2002). Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Res*, **62**, 3005-8.

Yu Z, Lu H, Si H, et al (2015). A highly efficient gene expression programming (GEP) model for auxiliary diagnosis of small cell lung cancer. *PLoS ONE*, **10**, 125517.