Lung cancer classification and applications using genetic algorithm to optimize prediction models

Anshoo Singh department of computer science and engineering

Abstract- carcinoma is **one in every of the foremost** fatal **sorts of** cancer **round the** world. It is estimated by Cancer Research Fund International in 2012, there is 1.8 million new cases of this disease were diagnosed. Early diagnosis and classification of this condition prompts medical professionals on safer and **simpler** treatment of the patient. Availability of microarray technology has paved the **thanks to** exploring the genes and its association in various diseases like **carcinoma**. This study of genetic algorithm as **a way** of feature (genes) selection for the SVM artificial neural network to classify **carcinoma** status of a patient. Genetic algorithm identified genes that classify patient **carcinoma** status with notable predictive performance.

I.INTRODUCTION

Lung cancer has been identified as **a significant** health issue for both developed and developing countries. In 2000, over **a million** deaths **are** reported worldwide with 53% occurring in developed countries and 47% in less developed countries[22]. As of 2012, 1.8 million cases **are** diagnosed and estimates suggest that by 2030, **carcinoma** will reach around 10 million deaths **annually** [1][21].

Surgical removal of **carcinoma** still remains the gold standard in preventing **carcinoma**. Early diagnosis of **carcinoma** is therefore important **to stop** the spread of the cancer.

Treatment of **carcinoma** also varies **looking on** the type tumor present. Classification **of various** tumor types is thus important **to confirm** higher survival rates. However, classification of lung cancers is challenging [4][19].

Currently, cancer classification **is predicated** on subjective interpretation of histopathological and clinical data. Classification also depends on **the positioning** of origin of the tumor. Clinical information **could also be** incomplete **and therefore the**n|every now and then} and the wide classes of most tumors lack morphologic features which are essential in classification [23].

Cellular function **is set** by its gene expressions. Humans have approximately 20,000 to 25,000 genes, each of which **contains a** particular sequence [24]. Genes are first transcribed into **RNA** (mRNA) (transcription), mRNA is then **wont to** synthesize protein (translation). **the entire** process from transcription to translation **is named organic phenomenon**.

Cells which are **in a very** state of disease have different gene expressions from normal functioning cells. Genes which differ in expression are used as biological markers **to point** particular disease states.

With **the arrival of recent** technologies **like** the

DNA microarray, we are now **ready to** measure gene expression levels of thousands of genes **in a very** given cell or tissue. Microarray technology made it possible **to look** systematically for markers of cancer classification and outcome prediction **in a very form of** tumor types [14]. Microarray technology thus became **a very important** tool for studying the transcriptome of cancer cells.

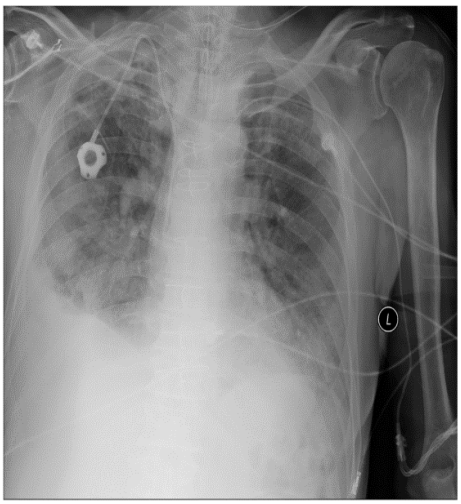
An application of microarrays is by classification analysis. Microarray data **is employed to see** if genes are active, hyperactive or inactive in various tissues. Then samples are classified into two or more groups[4]. However, classification analysis using microarray data becomes difficult **due to** small sample size, high dimensionality **of information** (gene expressions from all 20,000+ genes) and presence of fragments(noise and irrelevant information). Thus **we've** to implement strategies **to stop** misclassification and improve our analysis.

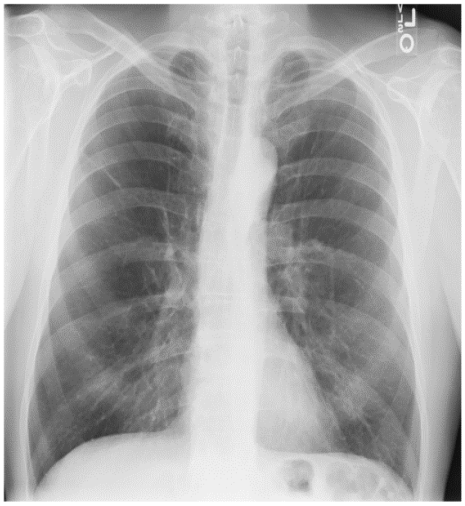
In this study, **we are going to** use the GALGO package developed for R for classification analysis of lung cancer data. GALGO allows **the event** and analysis of statistical models **employing a**

This NIH Chest X-ray Dataset is comprised of 112,120 X-ray images with disease labels from 30,805 unique patients. To create these labels, the authors used Natural Language Processing to text-mine disease classifications from the associated radiological reports. The labels are expected to be >90% accurate and suitable for weakly-supervised learning. The original radiology reports are not publicly available but you can find more details on the labeling process in this Open Access paper: "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases." (Wang et al.)

In this sample dataset (about 5% of the full dataset):

✗ Contains 5,606 images with size 1024 x 1024





✗ Class labels and patient data for the entire dataset:

- Image Index: File name
- Finding Labels: Disease type (Class label)
- Follow-up #
- Patient ID
- Patient Age
- Patient Gender
- View Position: X-ray orientation
- OriginalImageWidth
- OriginalImageHeight
- OriginalImagePixelSpacing_x

II. REVIEW OF RELATED LITERATURE

Microarray technology was used for tumor

classification and cancer diagnosis in works of Golb et. aI, Ben-Dor et al. and Alizadeh et al. These techniques, using two or three classes, returned test success rates of 90-100% **for many** binary class data. However, expansion of **the matter** to multiple tumor classes decreases perfonnance **of those** methods drastically because classification **for various** cancer types **isn't** yet clearly defined. This makes methods like Golub et al. and Slonin et al. **supported organic phenomenon**, starting with a feature selection **to require** possible correlation with an ideal gene marker particularly difficult. Also, complex relationships between genes affect the discriminant analysis in classification [26].

Tibshirani et al. and Ooi et al. used discriminant approaches which consider genetic interactions. Tibshirani et al. (2001) was successful **find** genes for classifying small round cell blue cell tumors and leukemias using **the straightforward** nearest prototype (centroid) classifier. Ooi et al (2002) used genetic algorithm maximum likelihood classification method (GAIMLHD) and found out that **the strategy** penn its substantial feature reduction in classifier genesets without compromising predictive accuracy[25].

Pan et al. (2003) used a hybrid genetic algorithm - based clustering (HGACLUS) schema and combined the advantages of simulated annealing for fmding an optimal/near optimal set of mediods **and located** that HGACLUS performed more accurately and more robustly than other methods in simulated data, embryonal CNS data and NC160 data [27].

Liu and Lin (2005) used the Genetic algorithm to identify **a collection** of key features **and mix** the silhouette statistic with a **kind of** linear discriminant analysis. They found that the GA/silhouette algorithm with the one-minus Pearson distance metric achieved **the simplest** performance and outperformed many previous methods.

Zhu et al. (2007) used a Markov Blanket-Embedded Genetic Algorithm (MBEGA) **for choosing** genes. The embedded Markov blanket based operators add or delete features (genes) from **an answer to enhance the answer** and increase accuracy. **the strategy** is effective and efficient in eliminating redundant and irrelevant features based on both Markov blanket and predictive power in classifier model [10].

Yang et al. (2009) used a hybrid filter/wrapper method called IG-GA for feature selection in microarray datasets. Information gain (lG) was **accustomed** select important feature subsets (genes) **and therefore the** genetic algorithm was used for actual feature selection. The method was used on eleven classification problems from

literature and has shown that the methods simplify the number of **organic phenomenon** levels effectively and either obtains higher classification accuracy or uses fewer features [28].

Cabrera (2014) developed a **computer virus** which can assess presence of **carcinoma** and further classify subtypes of **carcinoma** using normalization by decimal scaling, quantile normalization, min-max normalization and z-score transfonnation. The median absolute deviation (MAD) and **signal/noise ratio** (SNR) was **employed in** dimension reduction **for selecting** gene markers. Data was subdivided into test and training sets then **accustomed** classify patients by the support vector machine model [29].

This study utilized genetic algorithm (GA) as a method of feature (genes) selection to optimize performance of SVM and ANN in classifying lung cancer status of a patient.

III. METHODOLOGY

A. Dataset and Preprocessing

The dataset consisted of 203 patients subdivided into 17 normal lung patients, 6 small cell lung carcinoma (SM), 21 **epithelial cell** lung carcinoma (SQ), 20 pulmonary carcinoids (CO), and 139 adenocarcinoma (AD) patients. **there have been** 12600 genes (features) which were preprocessed using **the quality** nonnal score method. Selected features **accustomed** classify **carcinoma** were **supported the very best variance** trimming down the dataset to 3312 genes [19].

B. Prediction Models Optimization and model Validation

Artificial neural networks (ANN) and support vector machine (SVM) were used as prediction models to classify **carcinoma**. SVM and ANN are powerful tools in solving multiclass prediction problems as in the case of **carcinoma** classification. These models were optimized using genetic algorithm (GA) implemented in R Galgo package especially in terms of feature (genes) selection. Validation of the fmal models was done using cross validation method **which attracts** sample bootstraps **to handle the matter** of small test sample [30].

C. **fixing** the Genetic Algorithm

Two separate genetic algorithms each for ANN and SVM models were run using Galgo R Statistical package. Chromosome size was set to 50 genes (features) with a target accuracy rate (fitness) of **a minimum of** 97%.

IV. RESULTS AND DISCUSSION

A. Genetic Algorithm with SVM as Classifier

A total of 160 sets of calculations (chromosomes) with **at the most** 200 generations each set for the classification problem were perfonned. Seven sets of solution chromosomes (set of genes) satisfying the desired accuracy rate of **a minimum of** 97% correct cancer class prediction were obtained. For the majority of the calculations which yielded a lung cancer class prediction of less than 97%, the accuracy ranges from 86% to higher than 96% (See Figure 1).

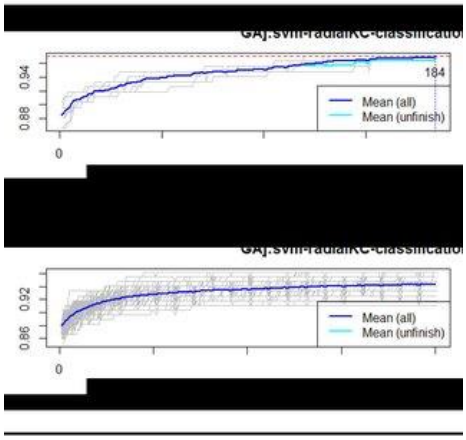


Fig. I. Classification Performance of Solution and No-solution Chromosomes with Genetic Algorithm Using SVM

Based on **the answer** chromosomes, a final model was derived with **regard to** parsimony and prediction accuracy rate. Figure 3 shows the candidate models obtained using forward selection method with accuracy rate plotted on the y-axis **and also the** selected features (genes) on the x-axis. **it absolutely was** revealed that of **the highest** 22 models presented, **the only** model that predicts lung cancer status of a patient at a high accuracy rate was model no. 15 consisting of 43 genes. **it's** a sensitivity rate **that's** between 79.5% and 98.9% with **a median** of 91.16. Moreover, model specificity is between 89.4% and 100% with **a median** of 97.8% (Figure 4). Adenocarcinoma and normal patients were predicted with **the best** accuracy at 97.92% and 97.14 % respectively. Also, good prediction accuracy rates were observe **for tiny** cell lung carcinoma (SM), squamous cell lung carcinoma (SQ) and pulmonary carcinoid (CO) patients with correct prediction accuracy rate of 83.33%, 81.03% and 91.17% respectively. **the** prediction accuracy rate was 95.87% **and also the** average accuracy rate was 91.16% (Table 1.)

B. Genetic Algorithm with ANN as Classifier

Same with **the strategy** using SVM, **a complete** of 160 sets of calculations (chromosomes) with **at the most** 200 generations each set for the classification problem were run. Majority (154 chromosomes) of the sets of analyses in this method satisfied **the specified** accuracy rate of at least 97% correct prediction. Only 6 chromosomes (no solution) obtained an accuracy rate of **but** 97% (Figure 2).

Figure 5 displays the candidate models derived using forward selection method. **the highest** 5 models were plotted with accuracy rate plotted on the y-axis **and therefore the** selected features (genes) on the x-axis. **the chosen** [mal model was model no. 1 with 45 features. Sensitivity rate is from 72.9% to 98.1 % with **a median** of 89.34%. The minimum specificity is 93.3% and the maximum is 99.3% with **a median** of 97.12% (Figure 6).

Fitness 1

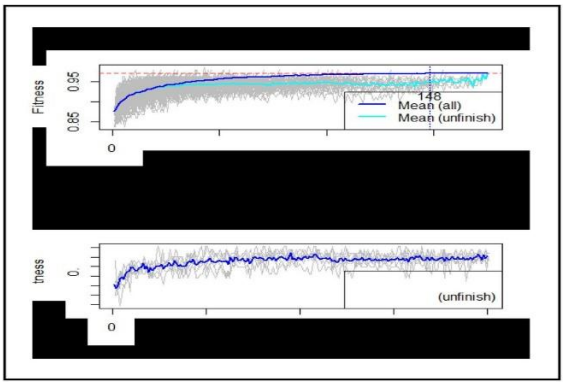


Fig. 2. Classification Peiformance oj Solution and No-solution Chromosomes with Genetic Algorithm Using ANN

Table 2 revealed that the [mal model was able to classify adenocarcinoma and normal patients with accuracy rates of 97.92% and 92.6 % respectively. Moreover, squamous cell lung carcinoma (SQ) and pulmonary carcinoid (CO) patients were classified at accuracy rates of 82.49% and 86.0% respectively. Lowest classification accuracy was observed on small cell lung carcinoma (SM) patients at an accuracy rate of 64.64%. Overall correct prediction rate was 93.66 % with an average of 84.75%

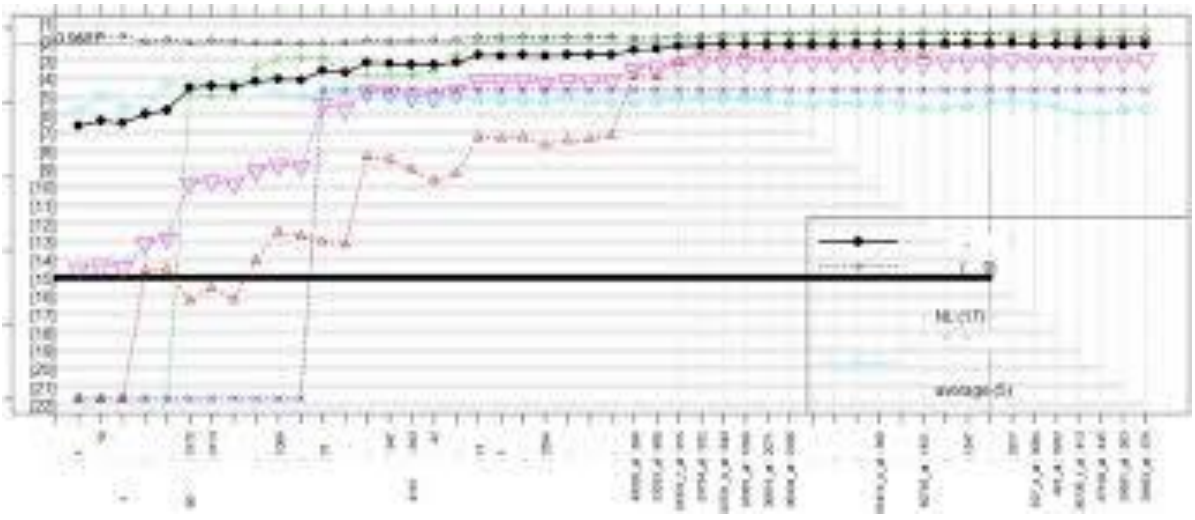


Fig. 3. Candidate Models Derived Using Forward Selection Method with Genetic Algorithm Using SVM3

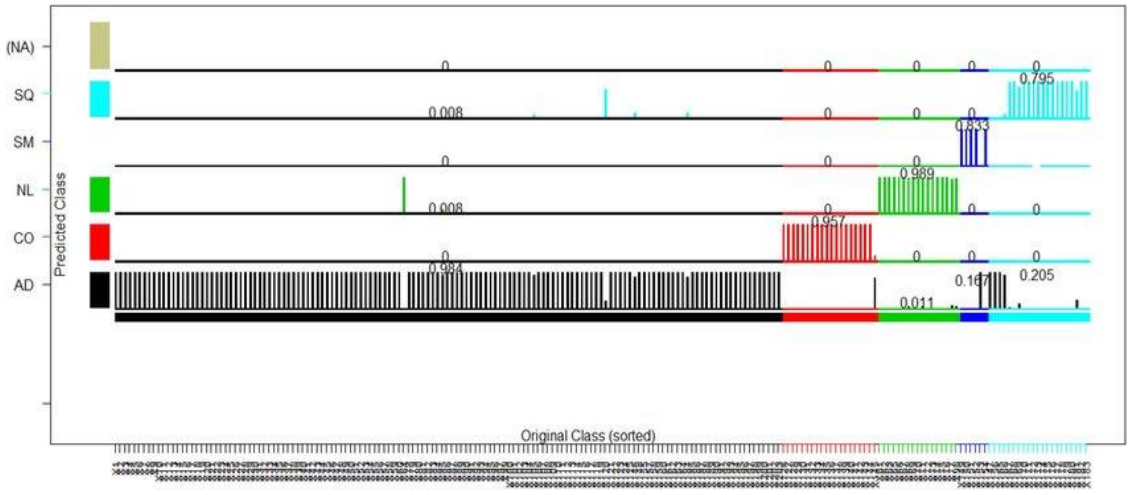
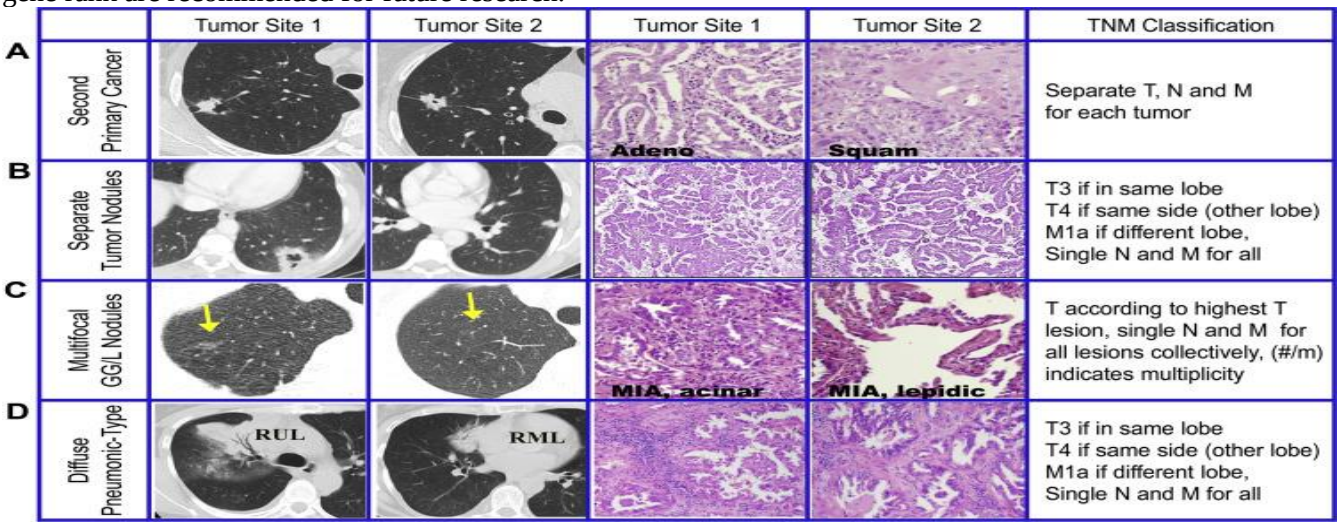


Fig. 4. Model Sensitivity and Specificity for each lung cancer patient class

V. CONCLUSION AND FUTURE WORKS

This study successfully classified **carcinoma** status of patients on **the premise** of **organic phenomenon** data using GA to optimize prediction models. The models were able to retain accurate prediction even with **a awfully** small number of features selected. GA selected quite similar numbers of genes for support vector machine and artificial neural networks and both yielded notable overall correct prediction accuracy rates of 95.87% and 93.66 % respectively. On **the opposite** hand, modelling of interrelationships among **the chosen** features and examining how classification performance behaves with gene rank are recommended for future research.



REFERENCES:

[1] World Cancer Research Fund International. "Lung Cancer". http://www.wcrCorgicancer statistics/data specific cancers/lung cancer statistics.php#BO TH. Accessed: March 2014

[2] T. Golub, D. Slomin, P. Tamayo, C. Huard, et aI., "Molecular Classification of Cancer: Class Discovery and Prediction by Gene Expression Monitoring" in Science, Vol. 286, October 1999, pp. 531-537.

[3] A. A1izadeh, M. Eisen, R. Davis, C. Ma, et aI., "Distinct Types

of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling" in Nature 403, February 2000, pp. 503-511.
[4] Y. Lu and J. Han, "Cancer Classification Using Gene Expression Data" in Information Systems (Special Issue on Data Management in Bioinformatics) 28, 2003, pp. 243-268.
[5] E. Tang, P. Suganthan and X. Yao, "Feature Selection for Microarray Data Using **statistical method** SVM and Particle Swarm Optimization," Computational Intelligence in Bioinformatics and Computational Biology 2005. [Proceedings of the 2005 IEEE Symposium, November 2005].
[6] D. Dasgupta, "Artificial Neural Networks and Artificial Immune Systems: Similarities and Differences", Proceedings of the IEEE Systems, Man, and Cybernetics, vol. I, 1997, pp. 873-878.
[7] V. Bevilacqua, G. Mastronardi, F. Menolascina, A. Paradiso and S. Tommasi, "Genetic Algorithms and Artificial Neural Networks in Microarray Data Analysis: a Distributed Approach," Engineering Letters, vol. 13, no. 3, 2006, pp.335-343.
[8] V. Trevino and F. Falciani, "GALGO: an R package for multivariate variable selection using genetic algorithms," in Bioinforrnatics Oxford Journals Vol. 22, no. 9, 2006, pp. 1154-1156.
[9] C. Ooi and P. Tan, "Genetic algorithms applied to multi-class prediction for the analysis of **organic phenomenon** data," in Bioinforrnatics Oxford Journals Vol. 19, no. 1, 2003, pp. 37-44.
[10] Z. Zhu, Y. Ong and M. Dash, "Markow Blanket-Embedded Genetic Algorithm for Gene Selection," in Pattern Recognition Vol. 40, issue I I, November 2007, pp. 3236-3248.
[ll] Y. Saeys, I. Inza and P. Larranaga, "A Review of Feature Selection Techniques in Bioinformatics," in Bioinformatics Oxford Journals Vol. 23, Issue 19, 2007, pp. 2507-2517.
[12] W. Daelemans, et aI., "Combined optimization of feature selection and algorithm parameter interaction in machine learning of language," in Proceedings of the 14th European Conference on Machine Learning (ECML-2003), Lecture Notes in **engineering** Vol. 2837, 2003, pp. 84-95.
[13] 1. Guyon and A. Elisseeff, "An introduction to variable and feature selection," **within the** Journal of Machine Learning Research Vol. 3, Issue 3, 2003, pp. 1157-1182.
[14] H. Liu and H. Motoda, "Feature Selection for Knowledge Discovery and **data processing**," **within the** Springer International Series in Engineering and **engineering** Vol. 454, 1998.
[15] 1. Yu, S. Ongarello, R. Fiedler, X. Chen, G. Toffolo, C. Cobelli , et aI., "Ovarian cancer identification **supported** dimensionality reduction for high-throughput mass spectrometry data," in Bioinformatics Oxford Journals, Vol. 21, Issue 10, pp. 2200-2209.
[16] T. Lal, O. Chapelle, J. Weston and A. Elisseeff, "Embedded Methods," in Feature Extraction Studies in Fuzziness and Soft Computing, Vol. 207, 2006, pp. 137-165.
[17] W. Krijnen, Chapter 8: Classification Methods in Applied Bioinformatics using R, 2009, pp. 161-163.
[18] B. Ripley, Pattern Recognition and Neural Networks,

Cambridge: **Cambridge** Press, 1996.
[19] A. Bhattacharjee, W. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, et aI., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses",PNAS, Vol. 98, no. 24, 2001, pp. 13790-13795.
[20] V. Trevino and F. Falciani, "Galgo: Genetic Algorithms for Multivariate Statistical Models from Large-scale Functional Genomics Data," R package version 1.2, 2014. http://bioinformatica. mty. itesm. mx/? q=node/82
[21] P. Jha, M. Kent Ranson, S. Nguyen, D. Yach, "Estimates of global and regional smoking prevalence in 1995 by age and sex". American Journal of Public Health, 92(6), pp.1 002-1 006. 2002.
[22] GLOBOCAN 2000. Cancer incidence, mortality and prevalence worldwide. IARC Cancer Base NO. 5. Version 1.0.1 [online database]. Lyon, International Agency for Research on Cancer, 2001.
[23] Ramaswamy, S., Tamayo, P., Ritkin, R., Mukherjee, S., Yeang, c., Angelo, M., et aI., "Multiclass cancer diagnosis using tumor gene expression signatures", Proc Natl Acad Sci USA, 98 (26), 15149-15154, 2001.
[24] Stein, L., D., "Human genome: End of the beginning", Nature, 431, 915-916, 2004.
[25] R. Tibshirani, T. Hastie, B. Narasimhan and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression". Proc. Natl. Acad. Sci. USA., 99: 6567-6572, 2002.
[26] T. Lin, R. Liu and S. Chen, "Genetic Algorithms and Silhouette Measures Applied to Microarray Data Classification", Proceedings of 3rd Asia-Pacific Bioinforrnatics Conference, Singapore, 17-21 January 2005.
[27] H. Pan, 1. Zhu and D. Han, "Genetic Algorithms Applied to Multi-Class Clustering for **organic phenomenon** Data", Genomics Proteomics Bioinforrnatics. 1(4):279-87. November 2003.
[28] C.H. Yang, L. Chuang and C.H. Yang. "IG-GA: A Hybrid Filter/Wrapper Method for Feature Selection of Microarray Data", Journal of Medical and Biological Engineering, 30(1): 23-28. August 2009.
[29] J. Cabrera, "Lung Cancer Classification System", Special