Artificial Neural Networks Methods for Identification of the Most Relevant Genes from Gene Expression Array Data

Conference Paper · August 2003

DOI: 10.1109/IJCNN.2003.1224066 · Source: IEEE Xplore

CITATIONS
9

READS
121

1 author:

Zvi Boger
Ben-Gurion University of the Negev
57 PUBLICATIONS **1,064** CITATIONS

SEE PROFILE

Artificial Neural Networks Methods for the Identification of the Most Relevant Genes from Gene Expression Array Data

Zvi Boger
OPTIMAL – Industrial Neural Systems Ltd.
Be'er Sheva, 84243, Israel
and
Rockville, MD 20852
zboger@bgumail.bgu.ac.il

Abstract **- Gene array studies can assess the global expression patterns of thousands of genes under multiple conditions. This paper demonstrates the application of large-scale artificial neural networks (ANN) for gene array analysis and cancer cell identification, by training ANN model from data generated by gene arrays experiments of four different small, round blue-cell tumors. Recursive input pruning of the ANN model and re-training techniques were used for the identification of the more relevant genes. Out of the original list of 2308 genes, an ANN model with 9 gene inputs and 5 neurons in the hidden layer correctly classified the four cancer cell types in the training set with only one non-recognition in the validation set. Doubtful mis-classifications can be identified as such by pattern analysis of the outputs of the ANN hidden neurons. Causal Index calculation shows the influence of each of the identified most relevant genes on the cancer cell classification, and thus constitutes an important new knowledge extracted from the ANN model.**

I. INTRODUCTION

Artificial neural networks (ANN) modeling are used for analyzing data when no mathematical relationships between the inputs and the outputs of a system are known. Recently there is a growing interest in the application of ANN in medical and biological sciences [1]. A description of ANN technique for bio-modeling is given in [2], and its application for prognostic and diagnosis in oncology is critically reviewed in [3]. The interest in ANN techniques for gene expression modeling and analysis was evident in the last 2002 World Congress Computational Intelligence, where several types of ANN architectures were reported for these applications [4a-4f].

"Global gene expression profiling using microarrays is emerging as a key technology for understanding fundamental biology of gene function, development, and for discovering new classes of diseases such as cancer and for understanding their molecular pharmacology", write Liang and Kachalo [5]. In their paper they describe the microchip gene arrays and the computational techniques that are currently used for analyzing the thousands of responses from each experiment. However, there remain many challenges that face the bio-informatic practitioners in extraction the knowledge hidden in the mass of information.

Other researchers have used ANN modeling for analysis of gene expression arrays of cancer cells, [6-10], usually identifying several tens of genes that can classify correctly the cancer types.

Many researchers use the popular Principal Component Analysis (PCA) technique for input dimensionality reduction. One of the reasons is the prevailing perception of large-scale ANN models as very hard to train, and that the resulting ANN model will be suspect of "over-training", as the number of the ANN model parameters will be much larger than the number of training examples. However, there are published algorithms that can successfully train large scale ANN models [11], and progressive dimensional reduction algorithms that will discard the less relevant inputs [12].

Another reason that researchers are not using ANN modeling is its "black box" image, without an explanation facility that would reveal the reasons for the decisions or predictions made by it. Again, there is a simple ANN model analysis method that can extract useful knowledge from it, in a form of Causal Indices (CI) that relates the influence of the inputs on the model outputs [13].

This paper will describe the application of these algorithms to the gene array data analysis and knowledge extraction. The paper structure is as follows – a brief exposition of the ANN training and knowledge extraction algorithms used here; the gene array data used in this

analysis will be described; the classification results reviewed; the new knowledge discussed and the conclusion will present some of the open issues resulting from this paper.

II. Brief Introduction To Large-Scale Artificial Neural Networks Modeling

ANN model is trained by learning from known examples. A network of simple mathematical "neurons" is connected by weights. Adjusting the values of the weights between the "neurons" during the training of the ANN is done by "back-propagation" of the errors between the ANN outputs and the known data outputs.

Once the ANN is trained, and verified by presenting examples not used in the training, the ANN is used to generate the model outputs from the new examples presented to it. The reader is referred to the many books and journal papers published on these subjects, and to a periodically updated review in the comp.ai.neural-nets discussion group [14].

There are several obstacles in applying ANN to large systems containing large number of inputs and outputs. Most ANN training algorithms need thousands of repeated presentations ("epochs") of the training examples to finally achieve small modeling errors. Large ANN tends to get stuck in local minima during the training. As most ANN training start from initial random connection weights sets, and the number of neurons in the hidden layer are usually determined by heuristic rules, many re-training trials are needed to achieve good models.

The PCA-CG training algorithm [11] can easily train large scale ANN models, as it starts from non-random initial connection weights, obtained by the mathematical processing of the training data set. A proprietary algorithm avoids, and escapes, local minima. These algorithms were successfully used to train ANN models of industrial plants with hundreds of inputs and outputs [15-16]. A comprehensive list of these and other applications is available in [17].

A knowledge extraction and dimensionality reduction technique is the ranking of the inputs according to their relevance to the ANN prediction accuracy [12]. It is based on the observation that the less relevant inputs contribute small proportion of the variance of the inputs of hidden layer neurons. This is the result of either the small relative variance of the input values or the small connection weights assigned to them during the training of the ANN. The least relevant inputs may be discarded and the ANN re-trained with the reduced input set giving often better prediction accuracy. The explanations for this possible improvement are:

a) The elimination of noise or conflicting data in the non-relevant inputs;

b) Reduction of the number of connection weights in the ANN that improves the ratio of the number of examples to the number of connection weights, thus reducing the chance of over-fitting.

Additional technique for knowledge extraction is the calculation of CI that results in quasi-quantitative relationships between the inputs and outputs of the ANN model.

The CI is calculated as the sum of the products of all "pathways" between each input to each output,

$$CI = \sum_{j=1}^{h} W_{kj} * W_{ji} \qquad (1)$$

where there are h hidden neurons, W_{kj} are the connection weights from hidden neuron j to output k, W_{ji} are the connection weights between input i to hidden neuron j.

The CI reveals the direction (positive or negative) and the relative magnitude of the relationship of the inputs on the particular output. Although somewhat heuristic, it is more reliable than the local sensitivity checks, as it is based on the whole ANN trained on all the available states. It has to be remembered, however, that the interactions between inputs are not readily visible.

Another knowledge extraction technique is the analysis of the hidden neurons' outputs. It was already observed [18] that in a well-trained ANN these outputs tend to be close to 0 or 1, and their information content calculation is useful [19]. The "binary" patterns can be used for clustering of examples, and the calculation of the mean attributes of each group reveal what makes each cluster different [20-21].

III. The NCI Gene Expression Experiments

Khan and co-workers in the National Cancer Institute described an ANN modeling technique to classify a gene array data, and published the data they used [7]. Their study rational and experimental techniques are quoted from that paper: "The small, round blue cell tumors (SRBCTs) of childhood, which include neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS), are so named because of their similar appearance on routine histology. However, accurate diagnosis of SRBCTs is essential because the treatment options responses to therapy and prognoses vary widely depending on the diagnosis. As their name implies, these cancers are difficult to distinguish by light microscopy, and currently no single test can precisely distinguish these cancers. In clinical practice, several techniques are used for diagnosis, including immunohistochemistry, cytogenetics, interphase fluorescence *in situ* hybridization and reverse transcription (RT)-PCR. Immunohistochemistry allows the detection of protein expression, but it can only examine one protein at a time. … However, molecular markers do not always provide a definitive diagnosis, as on occasion there is failure to detect the classical translocations, due to either technical difficulties or the presence of variant translocations."

"Gene-expression profiling using cDNA microarrays permits a simultaneous analysis of multiple markers, and has been used to categorize cancers into subgroup. However, despite the many statistical techniques to analyze gene-expression data, none so far has been rigorously tested for their ability to accurately distinguish cancers belonging to several diagnostic categories. … Here we applied ANNs to decipher gene-expression signatures of SRBCTs and used them for diagnostic classification."

"To calibrate ANN models to recognize cancers in each of the four SRBCT categories, we used gene-expression data from cDNA microarrays treated as separate samples. Filtering for a minimal level of expression reduced the number of genes to 2308."

Full details of the NCI gene expression experiments and the resulting data were published. 63 samples of four SRBCT tissues were tested with cDNA array of 6567 genes, filtering for a minimal level of expression reduced the number of genes to 2308. The data preprocessing used is PCA that reduced the dimensionality of the inputs to 10 linear combinations of the data, explaining 63% of the variability of the data. 3750 ANN models were trained with these inputs. The ANN modeling results were used to identify 96 genes that could be used as inputs to ANN classification model. This model classified correctly all the 63 training examples, and then used to diagnose cancer and non-cancer cells, with good results.

IV. Data Preprocessing And ANN Model Training

Data pre-processing is an important part of a good ANN modeling. In order to avoid numerical problems, all inputs are scaled to a uniform range. In this work the input preprocessing technique used was zero centering by subtraction of each column mean, and division of the result by the column standard deviation. The four classification outputs were set as [0.1, 0.9] values, if absent or present, believed to give faster ANN training convergence than the [0, 1]. The division of the 88 examples to training and validation set followed the original division save for the moving of three non-cancer examples to the training set, leaving two non-cancer examples in the validation set.

The PCA-CG ANN algorithm used in the training recommended 5 hidden neurons, and thus 2308-5-4 architecture was used for the classification model. It usually took two to three hundred presentations "epochs" of the training examples to get small training error, with no early stopping. The trained ANN was analyzed to discard the less relevant inputs and a smaller ANN re-trained. This procedure was repeated until the smallest number of relevant inputs was found that still gives adequate classification.

V. Analysis Of The ANN Modeling Results

The ANN training and validation RMS errors as a function of the (decreasing) number of inputs are given in Fig. 1. It can be seen that even if the validation errors may increase during the dimensionality reduction, eventually the ANN performance with a quasi-optimal small number of inputs is superior to the initial large ANN. The reason for these interim increases in

the validation errors may be a local minimum that the algorithms could not avoid, or get out of. However, even in these cases the correct identification of the less-relevant inputs continued.

The ANN outputs of the training set were clearly classified, with the correct cancer class output in the 0.88 – 1.00 range, and the other outputs, including the non-cancer examples, in the range of 0.00 – 0.09. The trained ANN outputs when presented with the validation set can be seen in Table I. The correct classification values are marked **bold**, and incorrect values are underlined**.** Table I also includes the values of the five hidden neurons, and their rounded "binary" patterns. The last column of this table shows the correct pattern from the training examples.

It can be seen that only one validation example was not correctly classified, with all other having correct ANN outputs in the range 0.57 – 1.00. Some examples also have outputs higher than 0.4 in the wrong classes, which may cause some doubts, but these examples generate abnormal binary patterns in their hidden neurons, which may be used to avoid wrong classification.

As seen in Fig. 1, the minimal good set of inputs is at nine genes, with both the training and validation errors rising sharply with further reduction of the gene list. The CI values of the nine genes for each classification output are presented in Table II, as well as the names of the genes. It is interesting to observe that each gene has one large positive CI value (marked **bold**), which may indicate that it is a "marker" the particular cancer type inserted in the last column. Indeed, some of these genes may be already known to be involved in cancer [22].

VI. Discussion

The ANN modeling results given here are encouraging, demonstrating the applicability of the large-scale training algorithms to the rapidly evolving, and potentially important, gene expression array data analysis. It also show that the danger of "over-training" is somewhat exaggerated, as already observed in [23].

However, some caveats have to be made. The successful training of a model with large number of noisy inputs is dependent on the number of hidden neurons, which may lead to several "quasi-optimal" sets of inputs [24]. Also necessary is to agree on an efficient way of reducing the number of genes from the typical 32 to 64 thousand genes, to the more manageable number of several thousand inputs to the ANN model.

References

[1] P.J.G. Lisboa, "A review of evidence of health benefit from artificial neural networks in medical intervention," *Neural Networks*, vol. 15, pp. 11-39, 2002.

[2] J.E. Dayhoff and J.M. Deleo, "Artificial neural networks – opening the black box," *CANCER Supplement*, vol. 91(8) pp. 1615-1635, 2001.

[3] G. Schwarzer, W. Vach and M. Schumacher, "On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology," *Statistics in Medicine*, vol. 19, pp. 541-561, 2000.

[4] *Proc. World Conference of Computational Intelligence, WCCI'02*, Honolulu, Hawaii, 2002.

a. E. Keedwell, A. Narayanan, and D. Savic, "Modeling gene regulatory data using artificial neural networks," paper IJCNN1204.

b. Y. Liang, E. O. George, and A. Kelemen, "Bayesian neural network for microarray data," paper IJCNN1359.

c. J. Ryu and S.-B. Cho, "Gene expression classification using optimal feature/classifier ensemble with negative correlation," paper IJCNN1419.

d. M. Su, M. Basu, and A. Toure, "Multi-domain gating network for classification of cancer cells using gene expression data," paper IJCNN1210.

e. C. Deng, P. Zhang, A. Wang, B. J. Trummer, and D. Wang, "Normalization of cDNA microarray data by using neural networks," paper IJCNN1072.

f. R. Xu, G. C. Anagnostopoulos, and D. C. Wunsch II, "Tissue classification through analysis of gene expression data using a new family of ART architectures," paper IJCNN1526.

[5] J. Liang and S. Kachalo, "Computational analysis of microarray gene expression profiles: clustering, classification, and beyond," *Chemometrics and Intelligent Laboratory Systems*, vol.62, 199-216, 2002.

[6] F.A. Azuaje, "Computational neural approach to support the discovery of gene function and classes of cancer," *IEEE Trans. Biomed. Eng.*, vol. 48(3), pp. 332-329, 2001.

[7] J. Khan, J. S. Wei1, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7(6), pp. 673-679, 2001.

[8] Y. Xu, F.M. Selaru, J. Yin, T.T. Zou, V. Shustova, Y. Mori, F. Sato, T.C. Liu, A. Olaru, S. Wang, M.C. Kimos, K. Perry, K. Desai, B.D. Greenwald, M.J. Krasna, D. Shibata, J.M. Abraham and S.J. Meltzer, "Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer," *Cancer Res*; vol. 62(12), pp. 3493-3497, 2002.

[9] M. Ellis, N. Davis, A. Coop, M. Liu, L. Schumaker, R.Y. Lee, R. Srikanchana, C.G. Russell, B. Singh, W.R. Miller, V. Stearns, M. Pennanen, T. Tsangaris, A. Gallagher, A. Liu, A. Zwart, D.F. Hayes, M.E. Lippman, Y. Wang and R. Clarke, "Development and validation of a method for using breast core needle biopsies for gene expression microarray analyses," *Clin. Cancer Res.* Vol. 8(5), pp.1155-66, 2002

[10] F.M. Selaru, Y. Xu, J. Yin, T. Zou, T.C. Iiu, Y. Mori, J.M. Abraham, F. Sato, S. Wang, C. Twigg, A. Olaru, V. Shustova, A. Leytin, P. Hytiroglou, D. Shibata, N. Harpaz

and S.J. Meltzer, "Artificial neural networks distinguish among subtypes of neoplastic colorectal lesions," *Gastroenterology*, vol. 122(3), pp. 606-613, 2002.
[11] H. Guterman, "Application of principal component analysis to the design of neural networks," *Neural, Parallel and Scientific Computing*, vol. 2, pp. 43-54, 1994.
[12] Z. Boger and H. Guterman, "Knowledge extraction from artificial neural networks models," *Proc. of the IEEE Int'l Conf. on Systems Man and Cybernetics, SMC'97*, pp. 3030-3035, Orlando, Florida, 1997.
[13] K. Baba, I. Enbutu and M. Yoda, "Explicit representation of knowledge acquired from plant historical data using neural network," *Proc. of the International Joint Conference on Neural Networks*, vol. 3, pp. 155-160, San Diego, 1990.
[14] W. Sarle, FAQ – Weekly reminder ftp://ftp.sas.com/pub/neural/FAQ.html
[15] Z. Boger, "Application of neural networks to water and wastewater treatment plants operation," *Trans. of the Instrument Society of America*, vol. 31 (1), pp. 25-33, 1992.
[16] Z. Boger, "Experience in industrial plant model development using large-scale artificial neural networks," *Information Sciences – Applications*, vol. 101(3/4), pp. 203-215, 1997.
[17] Z. Boger, "Who is afraid of the BIG bad ANN?" Paper IJCNN02 1215, *Proc. of the Int'l Joint Conf. on Neural Networks, World Conference of Computational Intelligence, WCCI'02*, pp. 2000-2005. Honolulu, Hawaii, 2002.
[18] L. Bochereau and P. Bourgine, "Extraction of semantic features logical rules from a multi-layer neural network," *Proc. of the International Joint Conference on Neural Networks*, vol. 2, pp. 579-583, Washington, DC, 1990.
[19] R. Kamimura and S. Nakanishi, "Hidden information maximization for feature detection and rule discovery," *Network Computation in Neural Systems*, vol. 6, pp. 577-602, 1995.
[20] Z. Boger, "Hidden neurons as classifiers in artificial neural networks models," *Proc. 5th Cconference on Artificial Neural Networks and Expert Systems*, ANNES'01, pp. 4-7, Dunedin, New Zealand, 2001.
[21] Z. Boger, "Finding patient cluster attributes using auto-associative ANN modeling," *These proceedings*, 2003.
[22] J. Khan, personal communication. 2002.
[23] S. Lawrence, C.L. Giles and A.C. Tsoi, "Lessons in neural network training: Overfitting may be harder than expected," *Proc. of the 14th National Conference on Artificial Intelligence, AAAI-97*, pp. 540-545, Menlo Park, 1997.
[24] Z. Boger, R.E. Cavicchi and S. Semancik, "Analysis of conductometric micro-sensor responses in a 36-sensor array by artificial neural networks modeling," *Proc. of the International Symposium on Olfaction and Electronic Nose, ISOEN '02*, Rome, September 2002, in press.

TABLE I
ANN CLASSIFICATION RESULTS OF THE VALIDATION SET

#	Name	Class	ANN Outputs					ANN Hidden Neuron Outputs					HN Pattern	Training HN
			EWS	RMS	NB	BL		H1	H2	H3	H4	H5		
67	TEST-8	NB	0.00	0.00	**0.94**	0.00		0.96	0.93	0.15	0.88	0.09	11010	11010
68	TEST-10	RMS	0.45	**1.00**	0.12	0.00		0.97	0.04	0.99	0.32	1.00	10101	10111
69	TEST-13	None	0.10	0.12	0.00	0.14		0.37	0.52	0.93	0.90	0.53	01111	01110
70	TEST-3	None	0.50	0.00	0.00	0.04		0.11	0.99	0.99	0.51	0.52	01110	01110
71	TEST-1	NB	0.00	0.00	**1.00**	0.00		1.00	0.99	0.00	0.78	0.01	11010	11010
72	TEST-2	EWS	**0.57**	0.03	0.28	0.00		0.49	1.00	0.92	0.26	0.49	01101	01101
73	TEST-4	RMS	0.01	**1.00**	0.00	0.00		1.00	0.01	0.97	0.99	0.96	10111	10111
74	TEST-7	BL	0.00	0.00	0.00	**0.93**		0.06	0.07	0.24	0.99	0.03	00010	00010
75	TEST-12	EWS	**0.71**	0.00	0.19	0.00		0.26	0.97	0.78	0.22	0.54	01101	01101
76	TEST-24	RMS	0.01	**1.00**	0.00	0.00		1.00	0.00	0.94	0.99	0.98	10111	10111
77	TEST-6	EWS	**0.98**	0.00	0.52	0.21		0.04	1.00	0.15	0.00	0.77	01001	01101
78	TEST-21	EWS	**1.00**	0.05	0.01	0.06		0.00	0.87	0.84	0.00	0.99	01101	01101
79	TEST-20	EWS	0.22	0.15	0.00	0.44		0.12	0.31	0.84	0.81	0.58	00111	01101
80	TEST-17	RMS	0.00	**0.57**	0.14	0.00		0.93	0.42	0.89	0.95	0.50	10111	10111
81	TEST-18	BL	0.01	0.00	0.00	**0.92**		0.02	0.08	0.37	0.98	0.04	00010	00010
82	TEST-22	RMS	0.01	**1.00**	0.00	0.00		0.99	0.00	1.00	1.00	0.99	10111	10111
83	TEST-16	NB	0.01	0.00	**1.00**	0.00		0.96	0.96	0.00	0.69	0.02	11010	11010
84	TEST-23	NB	0.02	0.02	**0.61**	0.12		0.68	0.70	0.25	0.81	0.19	11010	11010
85	TEST-14	NB	0.00	0.00	**0.82**	0.19		0.66	0.79	0.01	0.82	0.02	11010	11010
86	TEST-25	NB	0.00	0.00	**1.00**	0.01		0.94	0.93	0.01	0.80	0.03	11010	11010
87	TEST-15	BL	0.00	0.00	0.00	**1.00**		0.00	0.00	0.01	0.99	0.01	00010	00010
88	TEST-19	EWS	**1.00**	0.01	0.00	0.00		0.01	1.00	0.98	0.00	0.97	01101	01101

TABLE II
CAUSAL INDICES OF THE MINIMAL SET OF RELEVANT GENES

Gene #	EWS	RMS	NB	BL	Gene Name	Class Indication
856535	**7.3**	-0.7	-0.8	-0.8	methylenetetrahydrofolate dehydrogenase (NADP+ dependent), methenyltetrahydrofolate cyclohydrolase, formyltetrahydrofolate synthetase caveolin 1, caveolae protein, 22kD	EWS
855864	-1.1	-0.6	**12.6**	-2.9	PTK2 protein tyrosine kinase 2 cadherin 2, N-cadherin (neuronal)	BL
811956	-2.6	**10.6**	-0.7	-1.8	Homo sapiens GTP binding protein mRNA, complete cds Human DNA for insulin-like growth factor II (IGF-2); exon 7 and additional ORF	NB
810801	**5.6**	-3.7	0.2	-3.2	Homo sapiens agrin precursor mRNA, partial cds antigen identified by monoclonal antibodies 12E7, F21 and O14	EWS
767188	-1.2	-0.4	-5.0	**8.4**	UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 3 major histocompatibility complex, class II, DM alpha	BL
472186	-0.5	0.7	**4.6**	-2.7	RAB32, member RAS oncogene family transcriptional intermediary factor 2	NB
344272	**8.8**	-1.9	1.3	-1.5	epithelial membrane protein 3, Fc fragment of IgG, receptor, transporter, alpha	EWS
143661	1.3	-3.4	0.0	**5.7**	ESTs, protein kinase, cAMP-dependent, regulatory, type II, beta	BL
130884	-1.7	**11.2**	-5.2	-0.1	phosphate cytidylyltransferase 2, ethanolamine fibroblast growth factor receptor 5	RMS

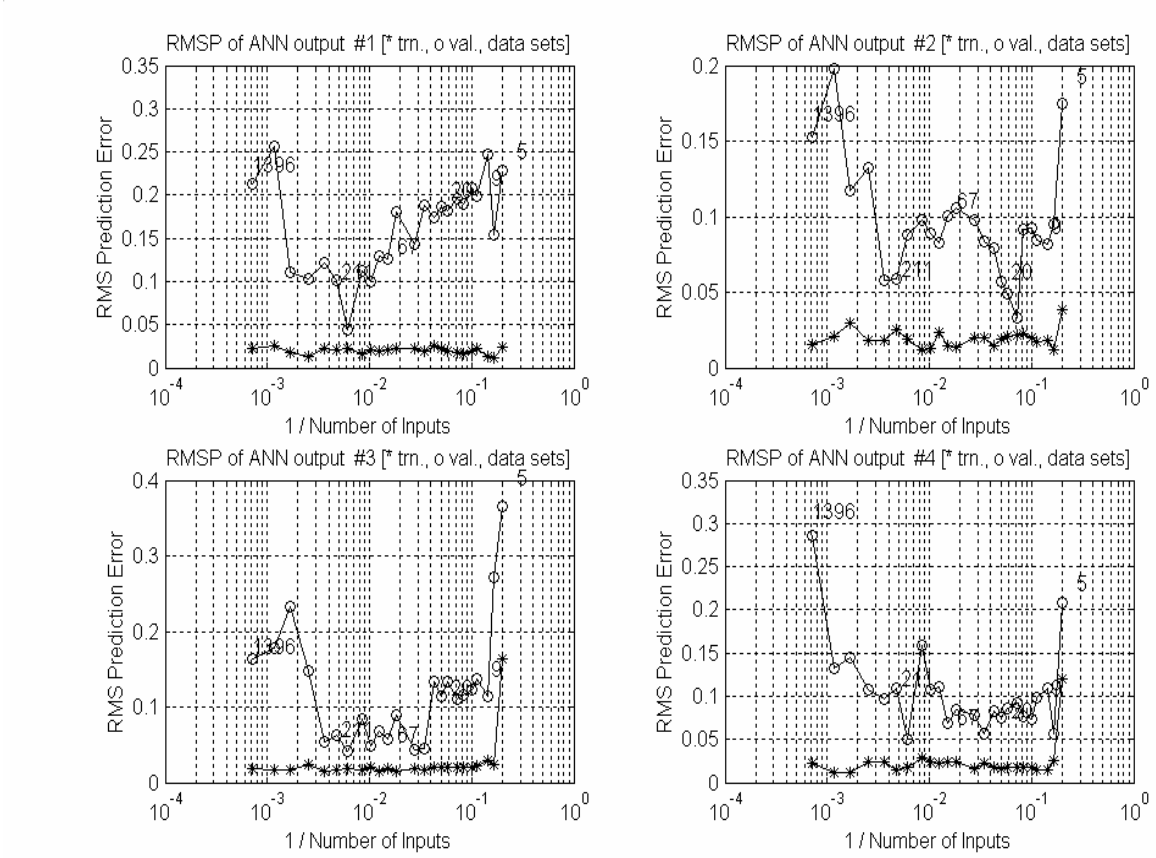

Fig. 1. ANN training (*) and validation (o) RMS errors as the function of the (decreasing) number of inputs. Each sub-figure is for a different cancer class output (#1 – EWS, #2 – RMS, #3 - NB, #4 – BL).