JACOBS PUBLISHERS

OPEN ACCESS

Jacobs Journal of Biostatistics

Research article

A Study of Statistical and Machine Learning Methods for Cancer Classification Using Cross-Species Genomic Data

Cuilan Gao[1*], Behrouz Shamsaei[2], Stan Pounds[3]

[1]The University of Tennessee at Chattanooga

[2]The University of Cincinnati

[3]St. Jude Children's Research Hospital

*Corresponding author: Dr. Cuilan Gao, The University of Tennessee at Chattanooga; Email: cuilan-gao@utc.edu

Received: 11-11-2016

Accepted: 12-02-2016

Published: 12-13-2016

Abstract

Use of gene expression profiling of animal model of a certain disease gives pre-clinical insights for the potential efficacy of novel treatments and drugs. Selection of an animal model, accurately resembling the human disease, profoundly reduces the research cost in resources and time. In this paper, we introduce and compare three different methods for classification of sub-types of cancer via cross-species genomic data. A statistical procedure based on analysis of variance (ANOVA) of similarity of gene expression between human and animal is used to select the animal model that most accurately mimics the human disease. Two other commonly used methods, logistic regression, and artificial neural networks are also examined and analyzed for the same data sets. The implementing procedure of each of these algorithms is discussed. Computational cost, advantage, and drawback of each algorithm are scrutinized for classification of simulated data and a real example of medulloblastoma (a type of brain cancer).

Keywords: ANOVA; Logistic Regression; Artificial Neural Networks; Classification; Gene Expression Data; Cancer Sub-type

Introduction

Current cancer classification includes more than 200 types of cancer. For the patient to receive appropriate therapy, the clinician must identify the cancer types as accurately as possible. Unlike many cancers in adults, childhood cancers are not strongly linked to lifestyle or environmental risk factors. In recent years, scientists have made great progress in understanding how certain changes in DNA can cause cells to become cancerous. Some genes (part of our DNA) help cells grow, divide or stay alive while others slow down cell division or cause cells to die at the right time. Cancers can be caused by DNA changes that turn on or turn off functions of cells. Different types of cancer in childhood may be caused by different types of genes changes. To appropriately classify cancer types, therefore, molecular diagnostic methods are needed. The classical molecular methods look for the DNA, RNA or protein of a defined marker that is correlated with a specific type of cancer and may or may not gives bio-logical information about cancer generation or progression. Gene expression data [1] by microarray or next-generation sequencing (NGS) has been emerged as an efficient technique for cancer classification, as well as for diagnosis, prognosis, and treatment purposes [2-6].

Gene expression profiling measures the expression levels of thousands of genes simultaneously. Most expression pro-filing studies focus on comparing the gene expression among biological conditions within the same species. For example, the experiments may compare the transcriptomes of tumors and normal tissue or between tumors arising in the same tissue. However, few work about comparing transcriptome data generated from different species had been done. Johnson et al. used a method called AGDEX to validate a novel mouse model of a human brain tumor [7]. A complete demonstration about AGDEX can be found in the paper by Pounds et al. [8]. Additional work about genomic expression patterns cross-species can be found in [9]. The rationale about the cross-species comparison is that animal and human shared the majority of patterns of regulation across orthologous genes. Therefore the animal gene expression profiling can be used to build a predictive model that may be used in the analysis of human diseases.

Statistical methods for cross-species gene expression analysis could be used for diagnosis of human disease based on the similarity in the genomic profile between human and animal data. We have proposed a method to find the most accurate animal model of a certain human disease [10]. This method is based on analysis of variance (ANOVA) of similarities between gene expression of human samples and gene expression of animal samples from a set of animal models to identify the animal model which is the best model for a certain type of cancer. This scheme, that will be analyzed in this paper, defines and computes a chosen metric of similarity between each human sample and animal sample from each animal model resulting in multiple groups of similarities. Then a random block ANOVA model is used to compare the group means of similarities among different animal models. Finally, post-hot multiple comparisons is applied to seek the best animal model of the human disease.

Logistic regression is a common tool in supervised classification problems both in statistics and machine learning areas. Mount et al. [11] applied this approach to identify gene expressions predictive of early death versus long survival in early-stage disease. Beane et al. [12] have used LR for lung cancer diagnosis that integrates genomic and clinical features. Stephenson et al. [13] have also integrated gene expression and clinical data using logistic regression modeling to predict prostate carcinoma reoccurrence after radial prostatectomy. One goal of many cancer genome projects is to discover cancer-related gene selection and sub-types. Zhou et al. [14] proposed a Bayesian approach to gene selection and classification using the logistic regression model. Logistic regression is often used for two-class classification and if input samples are classified in more than two categories, a multi-class classification is required [15]. Test samples are categorized in the most probable category. Algorithm of multi-class classification of human cancer type based on a trained logistic regression algorithm is illustrated in this paper.

Artificial Neural Networks (ANN) are computer-based algorithms which are modeled on the structure and behavior of neurons in the human brain and can be trained to recognize and categorize complex patterns. Pattern recognition is achieved by adjusting parameters of the ANN by a process of error minimization through learning from experience. They can be calibrated using any type of input data, such as gene-expression levels generated by cDNA microarrays or next-generation sequencing technology; and the output can be grouped into any given number of categories [6]. Implementation of the artificial neural networks algorithm in classification and diagnostic prediction of cancers using gene expression profiling is addressed in [2-16] and an example of this implementation in lung cancer prediction can be found in [17].

In this paper, gene expression profiles of animal models are used to predict the human cancer type. Mapping procedure between animal and human data is performed to match the orthologous genes between human and animal. Further-more, human gene expression data is deployed for cross-validation and test the trained algorithms. The advantages and disadvantages of the proposed methods, ANOVA test, logistic regression and artificial neural networks algorithm are observed and compared with artificial data and a real example of pediatric Medulloblastoma.

Description of Genomic Data

In microarray experiments, thousands of DNA sequences are aligned in probes and are exhibited in a high-density array positioned on a microscope slides. The relative expression of each probe(gene) is measured by comparing intensities of mRNA from tumor and reference tissue, see Figure 1. Thus gene expression levels of thousands of genes in each sample are simultaneously measured. Other sources of genomic data from sequencing technologies such as next-generation sequencing are obtained by using technologies to amplify and compare expression of target DNA sequences with reference (digital counts data).

We generally illustrate the microarray-based human and animal gene expression data as following. $ge_{i,j}$ represents the expression level the j^{th} gene of the i^{th} animal or human sample.

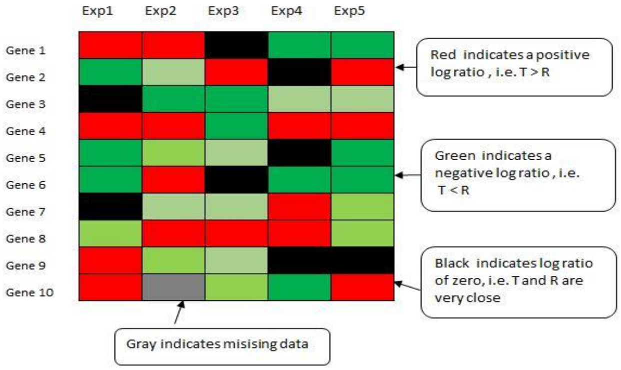


Figure 1. Gene expression matrix from multiple microarray experiments. The expression matrix is a representation of data from multiple microarray experiments. Each element is a log ratio of gene expression level of the testing sample (T) and gene expression level of the reference sample (R).

$$\overbrace{\left\{ \overbrace{\begin{pmatrix} gene-1 \\ gene-2 \\ gene-3 \\ \vdots \\ gene-n \end{pmatrix}}^{sample1} \overbrace{\begin{pmatrix} gene-1 \\ gene-2 \\ gene-3 \\ \vdots \\ gene-n \end{pmatrix}}^{sample2} \cdots \overbrace{\begin{pmatrix} gene-1 \\ gene-2 \\ gene-3 \\ \vdots \\ gene-n \end{pmatrix}}^{sample-h} \right\}}^{\text{Human model}} \quad (1)$$

where $gene_j$ represents the expression level of the j^{th} gene of the i^{th} animal or human sample. We assume i^{th} animal model has a samples, each including n genes.

$$\overbrace{\left\{ \overbrace{\begin{pmatrix} gene-1\\ gene-2\\ gene-3\\ \vdots\\ gene-n \end{pmatrix}}^{sample1} \overbrace{\begin{pmatrix} gene-1\\ gene-2\\ gene-3\\ \vdots\\ gene-n \end{pmatrix}}^{sample2} \cdots \overbrace{\begin{pmatrix} gene-1\\ gene-2\\ gene-3\\ \vdots\\ gene-n \end{pmatrix}}^{sample-a} \right\}}^{\text{Animal model}-\text{i}} \quad (2)$$

Each animal model corresponds to a particular subtype of disease or cancer. Several samples may belong to the same model.

ANOVA Models

Animal models play a pivotal role in translation biomedical research. The scientific value of an animal model depends on how accurately it mimics the human disease. In principle, microarrays collect the necessary data to evaluate the transcriptomic fidelity of an animal model in terms of the similarity of gene expression levels (by microarray gene expression file) with the human disease. We access this type of similarity by using different types of similarity metrics between each pair of a vector of gene expression of the human sample and a vector of gene expression of an animal sample from a certain animal model. Thus we end up with a set of similarity measurements. Then ANOVA method is used to analyze whether the similarity is associated with the types of animal models and the human sample labels.

In the ANOVA model described in the following equation, the type of animal model (model label) is considered as a fixed effect of similarity, and the human sample is considered as a random effect of similarity since the human samples are randomly selected from a large population of medulloblastoma patients.

$$\text{ANOVA } model: Y_{im} = \eta + \alpha_m + \varepsilon_{im}$$
$$\begin{cases} 1 < m < number\ of\ animal\ models \\ 1 < i < number\ of\ human\ samples \end{cases} \quad (3)$$

where η is the grand mean, α_m is the effect of m^{th} model which is the similarity of m^{th} animal model to the humam samples and ε_{ij} is the associated error. Null hypothesis or H_0 is: there is no statistically significant difference in similarities among the animal models, i.e $\alpha_1 = \alpha_2 = ... = \alpha_m$, where α_i is the effect of i^{th} the animal model. To test the hypothesis via ANOVA of similarity, we define four different similarity metrics as following.

A. Metrics of similarity

The similarity metric is a measurement of how similar of gene expression levels between a pair of human sample and animal sample. There are many different ways to measure the similarity. We will use four different metrics to measure this kind of similarity.

1) *Semi-Correlation:* One way to show the correlation between two matrices is to find the correlation between the columns of the two matrices. Experience shows that this method may result in non-normally distributed data. So to remedy this deficiency a new metric is defined based on the correlation coefficient between i^{th} human sample and the samples of m^{th} animal model. To define Y_{im} in (3), the column similarity is defined as

$$x_{i,j,m} = \sum(h_i - \bar{h}_i) \times (a_{m,j} - \bar{a}_{m,j})^T \quad (4)$$

where in (4) $a_{m,j}$ is the j^{th} sample of the m^{th} animal model, h_i is the *ith* human sample, $\bar{*}$ refers to the mean of variable $*$ and $*^T$ denotes the transpose of $*$. Y_{im} is the vector of all $(x_{i,j,m})$s in the m^{th} animal model.

2) *Cosine:* Second metric, like the first metric, defines a similarity between human and animal samples, which are columns of the human model and each animal model. This metric is also used in [11] and characterizes the *Cosine* between two vectors.

$$x_{i,j,m} = \frac{\sum(h_i \times a_{m,j}^T)}{\sqrt{\sum(h_i)^2} \times \sqrt{\sum(a_{m,j})^2}} \quad (5)$$

Definition of parameters in (5) is like the counterparts in (4).

3) Euclidean Distance: Euclidean distance is a simple yet powerful way to measure the similarity between two vectors. The Euclidean distance between vectors $a_{m,j}$ and $a_{m,j}$ is defined as follows:

$$Y_{im} = \sqrt{\sum \left(h_i - a_{m,j}\right)^2} \quad (6)$$

In other words, euclidean distance is the square root of the sum of squared differences between corresponding elements of the two vectors $h_i$and $a_{m,j}$.

4) *Pearson's Correlation Coefficient:* Pearson's correlation coefficient is calculated by dividing the semi-covariance (4) by the product of the standard deviations of gene expression of a human sample and gene expression of an animal sample. The formula for Pearson's correlation coefficient is omitted here as it is a simple and standard statistical concept. The advantage of the Pearson's correlation coefficient over the Euclidean Distance is that it is more robust against data that is not normalized.

B. Post hoc analysis

A simple way to look at the difference between animal models is to check the group means of similarities. Typically a larger mean value is indicative of similarity between the animal model to the human disease or cancer. The significant difference of animal models can be evaluated by the mixed ANOVA method. And finally, a multiple comparisons procedure test (Tukey's test or Hsu's Best) is conducted to identify the most similar animal model to human disease or cancer.

Logistic Regression Algorithm

Logistic regression is a standard method for building prediction models for a binary outcome and has been extended for disease classification with microarray data by many authors [19–23]. The idea of logistic regression is to use linear regression to predict probabilities of class labels. In mathematics, the goal is to minimize the cost function in regularized logistic regression C(Φ) defined in (7).

$$C(\Phi) = -\frac{1}{m}\left(\sum_{i=1}^{m}\left[y^{(i)}\log\left(h_\Phi\left(x^{(i)}\right)\right) + \left(1-y^{(i)}\right)\log\left(1-h_\Phi\left(x^{(i)}\right)\right)\right]\right) + \frac{\omega}{2m}\sum_{j=1}^{n}\Phi_j^2 \quad (7)$$

Where Φ is a matrix of weight vectors and x^i is the feature vector of each sample. y^i is a binary vector defined either 0 (for the case of non-existence) or 1 (for the existence of a criterion) for each sample, ω is the regularization factor used for preventing over-fitting the algorithm and $h_\Phi(x)$ is defined as the Sigmoid function defined in (8).

$$h_\Phi(x) = \frac{1}{1+e^{-(\Phi.x)}} \quad (8)$$

In equation (8), $\Phi.x$ denotes the dot product of weight vector Φ and the feature vector x and e is the exponential function. Gradient decent method is typically utilized to minimize the cost function (7). To update the weight vectors in each step, the derivative of the cost function with respect to the weight vectors should be computed.

$$\Phi_j = \Phi_j - \beta\frac{\partial(C(\Phi))}{\partial\Phi_j} \quad (9)$$

where β is the step size in gradient decent algorithm. It can be shown that derivative of the cost function with respect to weight vectors is

$$\frac{1}{m}\sum_{i=1}^{m}\left(h_\Phi\left(x^{(i)}\right)-y^{(i)}\right)x_j^{(i)} + \frac{\omega}{m}\Phi_j \quad (10)$$

Simultaneous updating the weight vectors will result in the reduction of the cost function.

Artificial Neural Networks Algorithm

Artificial neural networks is a method of training an algorithm for classification of a binary input data. For example, in the case of four-class classification problem, the binary classification vectors can be defined as

$$y^{(1)}=(\ 1\ 0\ 0\ 0\ )^{\mathrm{T}}, y^{(2)}=(\ 0\ 1\ 0\ 0\ )^{\mathrm{T}},\ldots.. \quad (11)$$

In this paper, animal samples are used to build an artificial neural network, connecting each input sample to a type of cancer. Artificial neural networks algorithm is typically a two-step perceptron, forward and back propagation steps. In practice, these networks consist of L layers of networks, that two of them are the input and output layers and the remaining layers are the hidden layers. Each layer has s_l a number of units that can be varied in different layers; And as defined in (11), each neural networks typically has K ≥ 3 output units. In forward propagation, perceptron is obtained with implementing the Sigmoid (logistic) function defined in (8) in different layers of the neural network. The back propagation step is used to calculate the error and consequently optimize the weight vectors. By having multiple nodes on each layer, K separable problems can be classified. Fig. 1 shows a neural network with three hidden layers; In each input and hidden layer one bias unit is added. Each unit is connected to every unit except the bias unit on the right layer.

A. Training the algorithm

In neural networks algorithm, activation vector $a^{(l)}$ for the l^{th} network layer is defined as

$$a^{(l)} = h_\Phi\left(a^{(l-1)}\right) = \frac{1}{1+e^{-(\Phi^{(l-1)}.a^{(l-1)})}} \quad (12)$$

where h_Φ is the Sigmoid function and $a^{(1)} = x$ or the input layer. Defining that, $a^{(K)}$ is the activation of the last layer or response of the algorithm for the output layer. The aim of the algorithm is to obtain weight vectors or $\Phi^{(l)}$ for each layer for better prediction of the output layer. The gradient decent method is used for solving this nonlinear problem.

So, training the neural networks algorithm consists of the following steps:

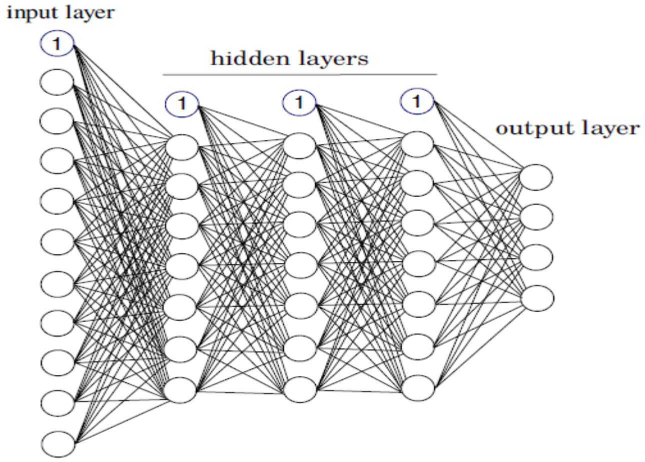


Figure 2. An example of a neural network with three hidden layers designed for four class classifications.

1) Random initialization of weight vectors for each network layer.

2) Minimizing the cost function by updating the weight vector of each layer until reaching the predefined criteria.

a) Forward propagation to compute the cost function C(Φ).

b) Back propagation to compute the derivative of the cost function with respect to weight vector in each layer $\frac{\partial(C(\Phi))}{\partial\Phi_{j,i}^{(l)}}$

c) Update weight vectors $\Phi_{j,i}^{(l)}$.

In the above algorithm, cost function shall be defined as

$$\begin{aligned} C(\Phi) = -\frac{1}{\bar{m}}\sum_{i=1}^{m}\sum_{k=1}^{k}\Big[y_k^{(i)}\log\Big(h_\Phi\big(x^{(i)}\big)\Big)_k \\ + \big(1-y_k^{(i)}\big)\log(1-h_\Phi(x^{(i)}))_k\Big] \\ + \frac{\omega}{2\bar{m}}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\big(\Phi_{j,i}^{(l)}\big)^2 \end{aligned} \tag{13}$$

where $\bar{m}$ is the number of samples, h_Φ is the sigmoid function and ω is the regularization factor. As discussed, the weight vectors should be updated by calculating the derivative of the cost function with respect to the weight vectors

$$\Phi_{j,i}^{(l)} = \Phi_{j,i}^{(l)} - \beta\frac{\partial(C(\Phi))}{\partial\Phi_{j,i}^{(l)}} \tag{14}$$

Back propagation process shall be summarized as

1) Set $\Delta_{i,j}^{(l)} = 0$

2) Calculate $\delta^{(l)}=a^{(l)}-y^{(i)}$ for each layer.

3) Compute $\delta^{(L-1)},\delta^{(L-2)},\ldots,\delta^{(2)}$

4) It can be shown that

$$\left|\begin{array}{l} \frac{\partial(C(\Phi))}{\partial\Phi_{i,j}^{(l)}} = \frac{1}{\bar{m}}\Delta_{i,j}^{(l)} + \omega\Phi_{i,j}^{(l)} \quad if\ j \neq 0 \\ \frac{\partial(C(\Phi))}{\partial\Phi_{i,j}^{(l)}} = \frac{1}{\bar{m}}\Delta_{i,j}^{(l)} \quad if\ j = 0, (for\ bias\ unit) \end{array}\right.$$

B. Cross-validation and testing the network

There are some factors in the neural networks algorithm that should be cited during network architecture selection. One of them is the number of hidden layers. The second one is the number of units in the hidden layers supposing that hidden layers have the same number of units, and the third is the regularization factor. Besides, the accuracy of the trained network should be examined. For these reasons, human samples are randomly divided into two groups of cross-validation and test sets. For the real data example discussed in section (VII-B), cross-validation and test groups include 40 and 66 samples respectively. The algorithm with a different number of hidden layers and units is examined with cross-validation samples. Cost efficient and accurate network is selected and tested with the test samples.

Feature Selection

Feature selection is the process of selecting a subset of features, or in this context, genes, in the training classification algorithms. The genomic data typically have a smaller number of examples than the number of features or genes that prevents the algorithms to over-fit the data or include the random noise in the training data. But to eliminate the under-fitting, the feature selection is essential. Also, feature selection often increases classification accuracy by reducing the noise pollution that might have infiltrated the data. Therefore, the t-test algorithm discussed in [24] is implemented for the sake of dominant feature selection.

Examples and Results

To compare the performance of the statistical and machine learning algorithms, two examples are examined here. In the first example, a set of simulated gene expression for human and animal are generated and the described processes in the previous section are applied on. In the second example, the algorithms are implemented on pediatric medulloblastoma data.

A. Simulated data

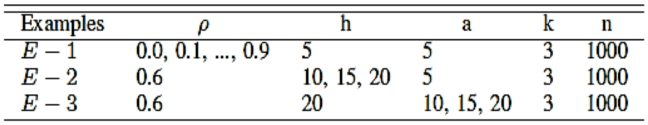
In the simulated data, we generated *h* human sample and *k* animal models each model with samples. Each sample contains *n* genes. These data, resembling the gene expression levels, are generated as random normal data *N* (0, 1). First, animal gene expression levels is generated as

$$A_1 = \frac{\rho}{\sqrt{1-\rho^2}} \times H + E \tag{15}$$

where A_1 is the gene expression matrix of the first animal model, *H* is gene expression matrix of the human, *ρ* is a scaling factor between 0 to 1, (0 < ρ < 1) and *E* is random error generated from *N* (0, 1). Arbitrary animal models can be generated with the same number of genes, however, for the illustration purpose, the number of models is restricted to three and specified in the table (I).

TABLE I

PARAMETER SETTINGS



Examples	ρ	h	a	k	n
$E-1$	0.0, 0.1, ..., 0.9	5	5	3	1000
$E-2$	0.6	10, 15, 20	5	3	1000
$E-3$	0.6	20	10, 15, 20	3	1000

1) *ANOVA results of the simulated data:* For all of these examples, the assumptions of ANOVA i.e. normality and homogeneity of variance, are tested by Shapiro test and Bartlett test, resulting in satisfaction of ANOVA assumptions. Accumulative means of the proposed examples are tabulated in tables (II), (III) and (IV). In these tables *S - C* and *Cos* represent the results of the semi-correlation and cosine similarity schemes. These tables show that means of similarity of the first model are significantly higher than the means of other two models.

TABLE II

ACCUMULATIVE MEAN IN GROUPS IN SETTING 1

Setting			Model1		Model2		Model3	
ρ	h	a	$S-C$	Cos	$S-C$	Cos	$S-C$	Cos
0	5	5	197.71	0.19	104.05	0.11	105.93	0.11
0.1	5	5	228.32	0.23	104.73	0.11	106.13	0.10
0.2	5	5	259.87	0.25	105.03	0.12	106.65	0.12
0.3	5	5	293.48	0.27	105.24	0.10	106.91	0.11
0.4	5	5	330.61	0.287	104.83	0.10	107.99	0.11
0.5	5	5	373.52	0.30	105.22	0.11	107.63	0.10
0.6	5	5	426.10	0.312	105.67	0.11	108.24	0.09
0.7	5	5	496.19	0.321	105.89	0.12	108.55	0.10
0.8	5	5	603.73	0.325	105.33	0.11	108.34	0.12
0.9	5	5	826.44	0.328	105.65	0.11	109.23	0.12

TABLE III

ACCUMULATIVE MEAN IN GROUPS IN SETTING 2

Setting			Model1		Model2		Model3	
ρ	h	a	$S-C$	Cos	$S-C$	Cos	$S-C$	Cos
0.6	10	5	267.88	0.196	109.49	0.109	111.97	0.109
0.6	15	5	209.76	0.153	106.30	0.105	94.33	0.093
0.6	20	5	182.66	0.133	94.69	0.093	95.11	0.097

TABLE IV

ACCUMULATIVE MEAN IN GROUPS IN SETTING 3

Setting			Model1		Model2		Model3	
ρ	h	a	$S-C$	Cos	$S-C$	Cos	$S-C$	Cos
0.6	20	10	181.87	0.136	96.65	0.097	102.07	0.103
0.6	20	15	179.99	0.136	100.78	0.101	101.24	0.103
0.6	20	20	180.58	0.138	100.21	0.102	105.82	0.105

F tests of ANOVA show that the similarities (by all metrics) of between these three models and human samples are significantly different (p < 0.0001). And finally results of the Turkey test for the sample in the table (V) shows that *model1* is significantly different from *model2* and *model3*, thus we conclude that *model1* is the most similar model to these human samples.

TABLE V

RESULTS OF THE TURKEY TEST FOR THE SIMULATED DATA OF THIRD SETTING (BY COSINE METRIC)

	diff	lwr	upr	p adj
$model2 - model1$	-0.03564	-0.05499	-0.01629	0.00004
$model3 - model1$	-0.03345	-0.05279	-0.01410	0.00015
$model3 - model2$	0.00219	-0.01715	0.02154	0.96171

2) *Logistic regression and artificial neural networks results of the artificial data:* After training the LR and ANN algorithms, each human sample is multiplied by the weight vector of each class and the most probable class is regarded as the type of cancer. For solving this problem with the neural networks scheme, regularization factor, and step size are defined as

$$\omega=0,\ \beta=0.1 \tag{16}$$

Table (VI) shows the percentage of accuracy in the result of the logistic regression and the artificial neural networks with one hidden layer and with 50, 100 and 150 units.

TABLE VI

ACCUMULATIVE MEAN IN GROUPS IN SETTING 1

Example	Setting			Accu. % of LR	Accu. % of ANN		
	ρ	h	a		s_l in hidden layers		
					50	100	150
E-1	0	5	5	0	20	20	20
	0.1	5	5	40	40	40	40
	0.2	5	5	60	80	80	80
	0.3	5	5	100	100	100	100
	0.4	5	5	100	100	100	100
	0.5	5	5	100	100	100	100
	0.6	5	5	100	100	100	100
	0.7	5	5	100	100	100	100
	0.8	5	5	100	100	100	100
	0.9	5	5	100	100	100	100
E-2	0.6	10	5	100	100	100	100
	0.6	15	5	100	100	100	100
	0.6	20	5	100	100	100	100
E-3	0.6	20	10	100	100	100	100
	0.6	20	15	100	100	100	100
	0.6	20	20	100	100	100	100

Accuracy in TABLE (VI) shows that these algorithms are capable of solving problems with small sample size and a large number of features, provided the feature selection is done.

B. Real data

We use a real example to examine the performance of the three methods described in the previous sections. The details of the data used for this analysis is described in sections (VII-C and VII-D).

C. Human and animal data

In our previously published paper in the journal Cancer Cell [18], four different mouse models were generated to mimic subtypes of medulloblastoma (a type of brain cancer). However, this paper cannot answer the question that which animal model is the most accurate model given a set of mouse models and what sub-type of medulloblastoma of each human sample belongs to. In this paper, we will use the same data to develop methods that classify the human cancer types using the cross-species genomic data. The animal data are mouse gene expression using 430V2 and HT430PM chips which can be found in NCBI database by accession numbers GSE33199 and GSE33200, respectively. The mouse data consist of four sub-types of medulloblastoma: 5 samples of *normal*, 5 samples of the *stem*, 5 samples of *prog* and 5 samples of *ptch* (one sample was damaged in the experiment), each sample contains 45101 probe sets. The human gene expression data are the same data as described in [18], which consist of 106 samples and 54675 probe sets.

D. Mapping cross-species genome data

A mapping procedure is induced to find the orthologous genes between human data and the mouse data. For this reason, Affymetrix best-match data set available at www.affymetrix.com is utilized to define around 75000 pairs of ortholog-matched genes (probe-sets). Furthermore, a filtering procedure is utilized to eliminate the repeated features (genes) in the human and animal samples.

1) *ANOVA results of the real data:* By ANOVA of the four different similarity metrics, all these metrics except Euclidean distance show that *ptch* medullablatoma of the mouse is significantly closer to the human samples ($p < 0.0001$, see Table (VII)). By Euclidean distance, ptch slightly larger than that of ptch, but still significantly ($p < 0.0001$) lower that of the norm and prog, which may be due to that Euclidean distance is less robust to non-normal data.

TABLE VII

COMPARISON OF FOUR SIMILARITY METRICS

Similarity	norm	prog	stem	ptch	p-value
Cosine	0.9059	0.9134	0.9001	0.9424	< 0.0001
Semi-Corr.	1806.04	2159.8	2102.9	2873.6	< 0.0001
Euclidean Dist.	176566	175943	171760	171829	< 0.0001
Pearson's Corr.	0.012	0.162	0.174	0.102	< 0.0001

2) *Logistic regression results of the real data:* Algorithm for implementing multi-class logistic regression is described in this paper. After training the algorithm, each human sample is multiplied by the weight vector of each class and the most probable class will be the type of cancer. By doing so, this algorithm is able to find the type of cancer correctly in all of the human cases.

3) *Artificial neural networks results of the real data:* For solving this problem with the artificial neural networks scheme, regularization factor, and step size are defined as

$$\omega=0,\ \beta=0.1 \tag{17}$$

Updating the weight vectors is repeated until reducing the cost function with three orders of magnitude. Table (VIII) shows the result of the neural networks with one and two hidden layers, each with 50, 100 and 150 units.

TABLE VIII

ACCURACY OF THE NEURAL NETWORK FOR DIFFERENT ARCHITECTURE SETTINGS

No hidden layers	1			2		
s_l in hidden layers	50	100	150	50	100	150
Cross validation error	0.12	0.00	0.00	0.02	0.00	0.00
Test error	0.05	0.01	0.00	0.00	0.00	0.00

Error in the table (VIII) is defined as the percentage of the accurate results in the cross-validation and test sets. For this problem, one hidden layer with 150 units suffices the efficiency and accuracy of the neural network.

Comparison of the Algorithms

In the proposed statistical scheme, post hoc analysis is inevitable. It means that the statistical analysis is not complete until a multi-step procedure is followed. While in the machine learning algorithms there is a need for parameter selection, i.e. choosing a number of hidden layers, feature selections, and the regularization factor to avoid over or under fitting the data. Since the genomic data typically has a small number of samples and a large number of features (genes), the algorithms are more prone to over-fit the data. For this reason, in the provided examples the regularization factor is chosen as zero. After these selections are done, the multiclass classification is complete after training the algorithm.

The ANOVA-based method requires the assumption of normality of the data and homogeneity of the variance, where the machine learning algorithms do not necessitate these requirements.

The ANOVA scheme only can specify one type of cancer for the human model, while in the LR and ANN algorithms, human samples can be categorized in different types of cancer.

In contrast, ANOVA scheme is a cost efficient scheme and does not require minimization of the cost functions, thus reduces the computational complexity.

In both of the LR and ANN, methods feature filtering and selecting are needed to avoid the data noise and find the global minimum. Finding the global minimum requires an optimization software, while the procedure in the ANOVA scheme is straightforward. Also, avoiding the local minimums and finding the global minimum in this process can be challenging when the ratio of a number of features to a number of training samples is high.

Even though logistic regression is easy to implement, multiclass classification is costly if there are more than two classes. ANN algorithm is the most expensive and the most difficult to implement an algorithm but because of the nature of this algorithm that allows increasing the number of hidden layers and hidden nodes, this algorithm tends to behave better than the other two schemes.

Conclusions

Three different methods, including the proposed ANOVA based analysis of similarity, logistic regression and artificial neural networks for classification of cancer types via analysis of cross-species gene expression data are investigated in this paper. Implementation procedure of each method is described. Benefits and drawbacks of these methods are examined and discussed with an artificial data and a real example of pediatric brain tumor. Among these three methods, the proposed ANOVA-based method yields a comparable result with less computation complexity. ANN is a powerful tool in the domain of data analysis. In practice, the samples of the human tumor are very limited as cancer are a rare disease, which will limit the power of parametric statistical methods. The power of ANN implementations has been enhanced by training the animal data to obtain the predictive model. As next-generation based genomic data becomes cheaper and more available than it is today, these classification methods via analytic of genomic data will continue to play an important role in diagnostic, prognostic and predictive software applications in the field of cancer genomics.

Competing Interests

The authors declare that they have no competing interests.

References

1. Zheng CH, Huang DS, Kong XZ, Zhao XM. Gene Expression Data Classification Using Consensus Independent Component Analysis. Genomics Proteomics Bioinformatics. 2008, 6(2): 74-82.

2. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine. 2001, 7: 673-679.

3. Khan J, Simon R, Bittner M, Chen Y, Leighton SB et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. Cancer Res. 1998, 58(22): 5009-5013.

4. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999, 286(5439): 531-537.

5. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000, 403(6769): 503-511.

6. Bishop CM. Neural Networks for Pattern Recognition. Clarendon Press, Oxford. 1995.

7. Ji W, Zhou W, Gregg K, Yu N, Davis S et al. A method for cross-species gene expression analysis with high- density oligonucleotide arrays. Nucleic Acids Research. 2004, 32(11): e93.

8. Johnson RA, Wright KD, Poppleton H, Mohankumar KM, Finkelstein D et al. Cross-species genomics matches driver mutations and cell compartments to model ependymoma. Nature. 2010, 466(7306): 632-636.

9. Pounds S, Gao CL, Johnson RA, Wright KD, Poppleton H et al. A procedure to statistically evaluate the agreement of differential expression for cross-species genomics. Bioinformatics. 2011, 27(15): 2098-2103.

10. Shamsaei B, Gao C. On the Evaluation of the Most Accurate Pediatric Medulloblastoma Animal Model. JSM Proceedings, Biometrics Section, Seattle,Washington, American Statistical Association. 2015, 3098-3106.

11. Mount DW, Putnam CW, Centouri SM, Manziello AM, Pandey R et al. Using logistic regression to improve the prognostic value of microarray gene expression data sets: application to early- stage squamous cell carcinoma of the lung and triple negative breast carcinoma. BMC Med Genomics. 2014, 7: 33-39.

12. Beane J, Sebastiani P, Steiling K, Dumas YM, Lenburg ME et al. A prediction model for lung cancer diagnosis that integrates genomic and clinical features. Cancer Prev Res. 2008, 1(1): 56-64.

13. Stephenson AJ, Smith A, Kattan MW, Satagopan J, Reuter VE et al. Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. Cancer. 2005, 104(2): 290-298.

14. Zhou X, Liu KY, Wong STC. Cancer classification and prediction using logistic regression with Bayesian gene selection. J Biomed Inform. 2004, 37(4): 249-259.

15. Kim Y, Kwon S, Song SH. Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data. Computational Statistics and Data Analysis. 2006, 51(3): 1643-1655.

16. Oustimov A, Vu V. Artificial neural networks in the cancer genomics frontier. Transl Cancer Research. 2014, 3(3): 191-201.

17. Adetiba E, Olugbara OO. Lung Cancer Prediction Using Neural Network Ensemble with Histogram of Oriented Gradient Genomic Features. The Scientific World Journal. 2015, 2015.

18. Kawauchi D, Robinson G, Uziel T, Gibson P, Rehg J et al. A mouse model of the most aggressive subgroup of human medulloblastoma. Cancer Cell. 2012, 21(2): 168-180.

19. Eilers PHC, Boer JM, Ommen GV, Houwelingen HCV. Classi-

fication of microarray data with penalized logistic regression. Proceedings of SPIE. 2001, 4266: 187-198.

20. Fort G, Lambert-Lacroix S. Classification using partial least squares with penalized logistic regression. Bioinformatics. 2005, 21(7): 1104-1111.

21. Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics. 2002, 18(1): 39-50.

22. Shen L, Tan EC. Dimension reduction-based penalized logistic regression for cancer classification using microarray data. IEEE/ACM Trans Comput Biol Bioinform. 2005, 2(2): 166-175.

23. Zhu J, Hastie T. Classification of gene microarrays by penalized logistic regression. Biostatistics. 2004, 5(3): 427-443.

24. Zhou N, Wang L. A Modified T-test Feature Selection Method and Its Application to the HapMap Genotype Data. Genomics Proteomics Bioinformatics. 2007, 5(3-4): 242-249.