Tumor Gene Characteristics Selection Method Based on Multi-Agent

Yang Li[*], Ye Mingquan and Zhang Hao

Department of Basic Public, WanNan Medical College, Wuhu Anhui 241002, China

Abstract: For the tumor gene expression profile data that aiming to high-dimension small samples, how to select the classification feature of samples among thousands genes effectively is the difficult problems for analysis on tumor gene expression profile. First to partition the data set into K average divisions, to use Lasso method performing feature selection on each respectively, and then merge each selected division of subset together to perform feather selection again, and get the final feature gene. This experiment adopts the Support Vector Machine (SVM) as classifier, to take the classification performance of feature gene set by Leave One Out Cross-Validation (LOOCV) method as evaluation standard, improve classification accuracy and with algorithm in good stability. Because of lowered dimensions in each time of calculation, it solves the problem of overhead computational-expensive, and also solves the problem of "over-fitting" in a certain grade. Thus it gets conclusion that the K-partitioning Lasso method shall be an effective method for tumor feature gene selection.

Keywords: Feature selection, K-partitioning lasso, support Vector Machine, tumor gene.

1. INTRODUCTION

DNA micro array (Gene Chip) technology is a major technological breakthrough in molecular biology field, it was applied widely in each field of biology and medical research, such as: large scale DNA sequencing, disease diagnosis, gene regulation and interaction relationship mining etc. Among the above application, the tumor diagnosing and typing is the most attractive point for researchers. The tumor gene expression profile is that to determine the level of gene expression value among the tissue samples by using DNA micro array technology, it provides a brand new means for phymatology research. Tumor gene expression profile data has such characteristics as small samples, high dimensions, high noise and high redundance, and shall easily cause the emerging of "dimensionality curse" and "over-fitting" phenomenon [1], how to make effective analysis on tumor gene expression profile data of high dimension small samples and mine the gene related to tumor from them is already the focus of the researchers. The target of feature selection is to reduce data noise and redundance, and improve the sample classification accuracy rate and model's generalization ability. At present, there're three methods available for gene selection of tumor gene expression profile data, they are: Filter methods, Wrapper methods and the Embedded methods [2, 3], in this article the Lasso method it adopted belongs to the Embedded methods. The Filter methods is independent of classifier, with excellence in rapid calculation but the defect that without consideration of correlation between genes, its classification accuracy rate is not higher enough, its typical algorithm includes the χ^2-statistic, t-statistic, ReliefF, Signal to Noise Ratio (SNR) [4] etc. The typical Wrapper feature selection methods is based on heuristic search method, it usually take the classifier to adjust the feature gene subset, with the purpose for selecting the optimized subset; and such method has advantages in less selected feature gene, high classification accuracy rate, and the defects with very high time complexity. For example, Li *et al.* [5] combined the Genetic Algorithm (GA) with KNN classifier to select the feature gene, Chen *et al.* [6] combined GA with SVM classifier and adopted the distance of support vector as sufficiency function, got the fine results. The Embedded methods includes the function of feature selection during the training process of classifier, its advantage is with higher classification accuracy rate and its time complexity is lower than the Wrapper methods, and its defects is that the results of feature selection only depends on classifier selection. For example, the Ramón *et al.* [7] took the Random Forest applied for gene selection and classification, Ma *et al.* [8] combined K-means with Lasso method to perform the feature selection and structure prediction model for gene expression profile data, got quite good results.

In this article, it takes tumor gene expression profile data set as the specific object of study, combing with Lasso method to propose the K-partitioning Lasso feature selection method. The experimental results shows that the feature gene selected by the K-partitioning Lasso method is with less gene quantity, it reduces the redundant characteristic accordingly with high classification accuracy, and also with lower time complexity. The algorithm has good stability, and because of lower dimensions of calculation in each time, it solves the problem of overhead computational-expensive, and it enables the equilibrium between the numbers of samples and the numbers of genes in a certain grade, solves the problem of "over-fitting"; meanwhile it takes analysis and comparison on the methods used in this article and those existing methods of feature selection, to make further explanation on availability of K-partitioning Lasso method. Thus it takes conclusion that the K-partitioning Lasso is such an effective tumor gene feature selection method.

1874-110X/15 2015 Bentham Open

2. THE K-PARTITIONING LASSO FEATURE SELECTION BASED ON LASSO METHOD

The problem of Least Absolute Shrinkage and Selection Operator (Lasso) is proposed initially by the scholar Tibshirani in 1996 [9], to be used for describing a class of optimization problems with the constraint. With the primary thoughts that under such constraining condition which the sum for these absolute values of regression coefficient less than a constant, to make the residual sum of squares minimized, so that it may produce some regression coefficients that equal firmly as zero, to get the explicable model. It supposed that the data (X,Y) contains the numbers of samples as *n*. and numbers of features as *;,* X= $(x^1,\ldots,x^j,...,x^m)$, thereinto $x^j=(x_{1j}, x_{2j},\ldots,x_{nj})^T$ is the independent variable, $Y=(y_1,\ldots,y_i,\ldots,y_n)^T$, thereinto y_i is the response variable, $i=1,2,...,n$, $j=1,2,...,m$ and x^j is standardized, y_i is centralized. To make linear regression on the independent variables towards the response variable, and set a limit to a certain norms of regression coefficient β= (β$_1$, β$_2$,…, βm) that it shall not be exceeding a certain threshold value *t.*

The standardization on x^j, and the centralization of y_i :

$$\sum_{i=1}^{n} y_i = 0,\ \sum_{i=1}^{n} x_{ij} = 0,\ \sum_{i=1}^{n} x_{ij}{}^2 = 1,\ j = 1,2,\cdots,m \tag{1}$$

Lasso: The minimization on residual sum of squares

$$\arg\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{m}x_{ij}\beta_j\right)^2\right\}\ \textit{subject to}\ \sum_{j=1}^{m}\left|\beta_j\right|\le t \tag{2}$$

LARS algorithm

Input: the data (X,Y), X= $(x^1,\ldots,x^j,...,x^m)$ includes n samples and m features, $Y=(y1,\ldots,yi,\ldots,yn)^T$ shall be matched with labels of the n samples

Output: F is the feature subset strongly related to class label.

1) To standardize all the independent variables of X, that the Mean value is zero, variance is 1; then to centralize all the response variables of Y, that the Mean value is zero, to be shown as formula (1). To record the residual as $r = y - \hat{y}$, supposes that both of regression coefficients β=(β$_1$, β$_2$,…, β$_m$) is all zero.
2) To find the variable x^j that with the highest correlation to residuals r.
3) Starting from zero and along with the direction of transvection between x^j and r to adjust β_j – which as the coefficient for x^j and calculate the residuals until finding another variable x^k that with the highest correlation to r.
4) Continuously coming along with the direction of transvection between (x^j, x^k) and current r to adjust β_j and β_k, until finding another variable x^p that with the second highest correlation to current residuals r.- in case of there's non-zero regression coefficient reduced as zero, then to delete its matching variable from current variable set, and arrange the calculation again.
5) Repeating all above steps, until all variables enter in model for solving, the algorithm comes to end.

Fig. (1). LARS algorithm.

Thereinto, *t≥0* is an adjustable parameter, while *t* has quite smaller value, the coefficient of those variable with lower correlation shall be compressed into zero, thus these variables shall be deleted accordingly to achieve the purpose of feature selection; while the value of *t* is bigger enough, the restraint shall not be valid any longer, under this situation all of these attributes shall be selected.

The scholars of Efron *et al.* [10] proposed the Least Angel Regression (LARS) in 2004 and it solved the calculation problem of Lasso well. The LARS algorithm is a process of residual fitting, it assures that while going through the solving path, those variables that been selected in regression model shall be the same with current residual's correlation coefficient, LARS algorithm shall find the optimized solution for Lasso effectively, its algorithm process is shown in Fig. (**1**).

At present, there're three methods to determine the parameter t – cross validation, expanded cross validation and unbiased estimation to prediction risks [11]. Among them the cross validation method has the widest application, thus it adopts 10-fold cross validation in the LARS algorithm to determine the parameter t, meanwhile set the max iterations as 1000.

Among the tumor gene expression profile analysis, Golub *et al.* [4] proposed the signal-to-noise ratio evaluation index shall be the most simple and with widest application of such kind a filtering feature gene selection method. Thus it adopts SNR as the comparison of algorithms.

The SNR method measures the importance of gene through calculating the ratio between the inter-class looseness and intra-class tightness of each gene on all of these samples, that is to measure how many sample classified information does the gene contain, the exact calculation formula is given below:

$$S(g) = \frac{\left|\mu_1 - \mu_2\right|}{\sigma_1 + \sigma_2} \tag{3}$$

Thereinto, μ_1 and μ_2 presents the Mean value respectively that the gene *g* expressed in the two classes, and σ_1 and σ_2 is their standard deviation. The higher signal- noise ratio of one gene, the closer correlation it may has to such classification.

3. TUMOR FEATURE GENE SELECTION METHOD BASED ON K-PARTITIONING

The tumor gene expression profile data is the small samples data, it usually contains dozens of samples even more than ten, however the dimensions number of gene may exceeding thousands in comparison. By using Lasso feature selection method to perform feature selection for such high dimension small samples data, it often emerge the problems such as overhead computational expensive and "overfitting". In order to solve such problems, in this article it proposes such an improved Lasso feature selection method – K-partitioning Lasso method. Its main idea is that to reduce the numbers of dimensions by partitioning feature set, then to treat the partitioned feature subset with LARS algorithm. The actual process of K-partitioning Lasso method is shown in Fig. (**2**).

Firstly to divide the feature set of data X into K feature subsets equally, to set X[i] as the *i* part of feature subset after the feature set divided into K parts, then to use LARS algorithm to make feature selection on each part of feature X[i], to merge all the K parts of selected feature subset together, and use LARS algorithm to make feature selection again, till get the feature gene subset that strongly related to class label.

For example, in our experimental data of prostatic carcinoma gene expression data, there're 102 samples in data set and feature dimensions as 12600; to directly us Lasso feature selection algorithm it would be huge computational overhead, and easy to lead to "over-fitting" problem. If it divides such gene set into 100 parts, there're only one hundred more dimensions in each data subset, thus it leads to an equilibrium between feature dimensions and numbers of samples, and then to solve "over-fitting" problem effectively meanwhile reduces the computational overhead.

K-partitioning Lasso method firstly to make feature selection on each feature subset X[i] after partition, eliminate those features not related to class label, and keep the feature in each subset that related to class label appropriately, among them there're also feature genes; later to make feature selection again on those selected K feature subsets, it shall delete part of redundant gene. Suppose that G_1, G_2 are all feature genes that strongly related with class label, and being redundant by each other, the two genes is divided into two different feature subsets; if directly using other feature selection method to handle with them, G_1, G_2 may has great possibility to be selected as feature gene, but through K-partitioning Lasso method, for G_1 andG_2 is not in same divided feature subset, it may cause that during first feature selection on each feature subset G_1 andG_2shall be both kept as feature gene, however during the second feature selection on those kept feature subsets again, because the existence of G_1 (G_2), G_2 (G_1)shall be eliminated accordingly. As a result, the K-partitioning Lasso method shall eliminate redundant gene.

From this it can be seen that, K-partitioning Lasso method is not only suitable for feature selection of tumor gene expression data and solving "over-fitting" problem, but also to eliminate redundant features, and with quite less numbers of selected features, and that reduce computational overhead, improve calculation speed of algorithm, it shall be an effective feature selection method.

```
K-partitioning Lasso Algorithm
Input: the data (X,Y), X= $(x^1,\ldots,x^j,\ldots,x^m)$ contains n samples and m features; Y=$(y_1,\ldots,y_i,\ldots,y_n)^T$ matches to the labels of n samples; the numbers of divided parts is K.
Output: the feature subset FS in strong relation to class label,
F=[ ]; // to be initialized as null.
FS=[ ]; // FS to be initialized as null.
for i=1:K
    A[i]= LARS (X[i]); // X[i] is the i part of feature subset after feature set divided into K parts, using LARS algorithm to make feature selection on each X[i], to store the result in A[i].
end
for i=1:K
    F=F$\cup$A[i]; // to merge all feature subsets A[i], put them into F.
end
FS=LARS(F); // using LARS to make feature selection on merged feature subset F.
```

Fig. (2). K-partitioning lasso algorithm.

4. THE EXPERIMENTAL RESULTS AND ANALYSIS

The genetic chip is such a method to measure gene expression level; it can access gene expression profile data of tissue samples rapidly. The gene expression profile data can be shown by matrix, as following Fig. (**3**).

$$\textit{Numbers of samples } N\left[\begin{array}{ccccc} \multicolumn{4}{c}{\overbrace{\hphantom{g_{1,1}\quad g_{2,1}\quad \cdots\quad g_{M,1}}}^{M0o}} & \overbrace{\hphantom{l_1}}^{fle} \\ g_{1,1} & g_{2,1} & \cdots & g_{M,1} & l_1 \\ g_{1,2} & g_{2,2} & \cdots & g_{M,2} & l_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ g_{1,N} & g_{2,N} & \cdots & g_{M,N} & l_N \end{array}\right]$$

Fig. (3). The matrix for gene expression profile data.

To set G={g_1,g_2,…,g_M}as such a gene set composed by all the genes in a sample, among them the g_i($1\le i\le M$)={ $g_{i,1}$,$g_{i,2}$,…,$g_{i,N}$} presents a gene, ∣G∣=M presents the numbers of all genes. To set S={s_1,s_2,…,s_N}as the composed sample set, among them ∣S∣=N presents the numbers of samples, each sample s_j($1\le j\le N$)={ $g_{1,j}$,$g_{2,j}$,…,$g_{M,j}$,l_j}presents all the gene expression values under a certain condition. Thereinto, $g_{i,j}$ presents the gene expression value of gene g_i in sample s_j.

In this article to adopt a public data set - the prostatic carcinoma gene expression profile data set (prostate) to verify the validity of K-partitioning Lasso feature selection method. The prostate data set has been released by Singh *et al.* in 2002 [12], the prostatic carcinoma gene expression profile data set consists of 102 samples totally, among them there're 50 prostatic carcinoma tissue samples and 50 normal tissue samples, each sample are composed by 12500 genes. Such prostatic carcinoma gene expression profile data set can be accessed from the download url as: http://linus.nci.nih.gov/~brb/DataArchive_New.html [13].

In this article, the hardware platform adopted PC configuration as: Intel Xeon 5110 dual-core process, 2GB RAM, 250GB hard disk; and software environment configuration: Operation System – Windows XP, JAVA development platform as JDK 1.6, Weka development environment is Weka 3.7.3, in this article the algorithm is accomplished under Matlab 7.0 environment. Weka is a public data mining working platform, it merges lot of machine learning algorithm, including the preprocessing, classification, regression and clustering of data etc. Thus, after select feature gene by using feature selection algorithm, to call the classification algorithm in Weka to compare the classification accuracy rate of feather genes, thereby to verify the efficacy of K-partitioning Lasso algorithm.

Firstly to eliminate the noise data from the prostate data set. For the gene expression level in some gene column of the prostate gene expression profile data set are all as zero value, so it eliminates those noise gene data with all zero value in 394 columns, and get the gene quantity of the prostate data set as 12206, it's the preparation for gene standardization in next step.

Table 1. The comparison of performance that K-partitioning lasso feature selection algorithm has under different K value.

K-partitioning Lasso method	T(s)	Genes	Acc
K=10	70	38	98.04%
K=20	74	41	99.02%
K=40	83	40	99.02%
K=60	119	20	97.06%
K=80	136	42	99.02%
K=100	152	66	100%
K=120	262	28	97.06%

After that, to standardize the prostate data set and make reflection to [-1,1] interval. After standardization, each gene expression profile has the average value as zero, standard deviation as 1, its purpose is make convenience for comparing and calculating correlation coefficient, meanwhile to keep the relationship with the original samples [14].

Passing through pre-processing to data, using the proposed K-partitioning Lasso algorithm and SNR filtering method respectively to make feature gene selection. The SNR method for comparison is also realized under the environment of Matlab 7.0.

SVM classifier has such advantage that it could process high dimensions data and with very high classification accuracy and strong anti-noise capacity, it doesn't need the user to adjust and input a lot of parameters, and the number of vectors it support after training is usually quite small, such advantages is very efficient for the gene expression profile data that with increased matrix dimensions gradually. Brown *et al.* [15] applied several common used classification methods into classification procedure of tumor gene expression profile and compared the classification results, his study found that using SVM classifier has the best effect. Thus it shall take SVM as experimental classifier. The common used kernel function of SVM includes that linear transvection kernel, polynomial transvection kernel, radial basis kernel and Sigmod transvection kernel. The Radial Basis Function Kernel (RBF) [16-22] applies to non-linear classification, in comparison with polynomial Kernel, Sigmod Kernel function, RBF kernel function need fewer parameters. Just because of this, in this experiment it takes classification on tumor samples with SVM classifier based on RBF. The form of RBF kernel function is given below:

$$K\left(x,y\right)=e^{-\|x-y\|^2/2\sigma^2} \tag{4}$$

In field of gene expression profile data classification, the cross validation method is more recommended. In this experiment, taking the LOOCV method as the classification performance evaluation index, its idea is that to keep one different sample as test sample from the sample set every time, in addition to use other samples training the classifier, while each part of samples has performed test set for one time, to make statistics on the ratio between the number of misclassified samples and original sample scale, and take it as the error in classification as *Err*, to record the classification accuracy as *Acc=1-Err*.

In Table **1** it explains the time performance, classification accuracy and numbers of feature gene in the prostate tumor gene expression profile data set by using K-partitioning Lasso feature selection algorithm while the *K* is taken various values. In experiment, to set the *K* value as 10, 20, 40, 60, 80, 100, 120 respectively, using K-partitioning Lasso feature selection algorithm to make feature selection, and record the implementing time of algorithm *T* and the numbers of feature gene *Genes* respectively, then to call the SVN classifier in Weka to classify the gene data set after selection respectively, record the classification accuracy of LOOCV as *Acc*.

Through Table **1** it can be known that K-partitioning Lasso algorithm reduces the redundant characteristic, the numbers of selected feature gene is more less, and with higher classification accuracy ranging in 97 ~ 100%, in case of *K* =100 to select 66 features, its classification performance can reach to 100%, the experimental result shows that the candidate feature subset consisted by such 66 feature genes has included enough sample classification information. With K-partitioning Lasso method, it can get faster implementation, and get the result of feature selection within shorter period, to make comparison of time performance; during experiment it directly used Lasso method to make feature selection on the prostate data, but just because of its too large amount of calculation, the lasting time is much longer than the K-partitioning Lasso method, and the feature gene selected by Lasso method has just 10 dimensions, with classification accuracy rate at 95.10%, it's lower than that by K-partitioning Lasso method.

Under the premise of classification accuracy, K-partitioning Lasso algorithm shall need even less time overhead, but finally it get basically the same classification accuracy, as per Fig. (**4**) shown. Therefore, the conclusion can be made as that the K-partitioning Lasso feature selection method shall has good stability.

From the theoretical analysis, Lasso method itself is just the feature selection method with high efficiency, it can screen out the variable which strongly related to class label, and K-partitioning Lasso algorithm is just with the basis of Lasso feature selection method to set parameter K to accomplish following task: to delete those feature being not related to class label in each part of feature subset, then make feature selection on those selected features to further eliminate redundant feature, meanwhile those features strongly related to class label shall be reserved from the beginning to the end, as a result, it can assure the validity and stability of K-partitioning Lasso algorithm.

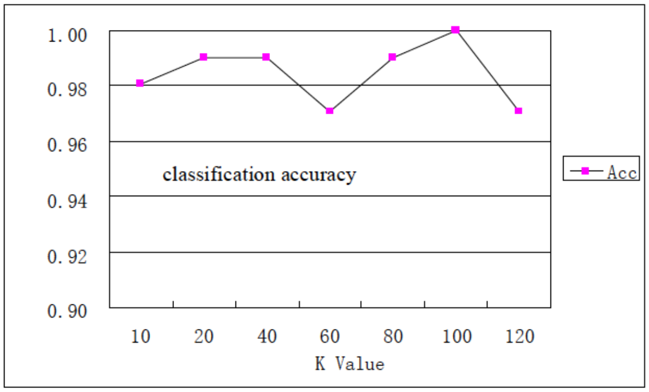

Fig. (4). The classification accuracy of K-partitioning lasso algorithm under different K value.

Table 2. The comparison on classification accuracy between K-partitioning lasso and SNR feature selection algorithm.

Genes	K-Partitioning Lasso Acc	SNR Acc
38	98.04%	92.16%
41	99.02%	92.16%
40	99.02%	92.16%
20	97.06%	94.12%
42	99.02%	92.16%
66	100%	92.16%
28	97.06%	91.18%

To further verify the validity of K-partitioning Lasso feature selection method, it takes comparison on the typical feature selection method – SNR with this algorithm, as SNR method is gene sequencing type method, to ensure the fairness of such comparison conditions, it compares respectively the classification accuracy in case of the numbers of feature genes is same as the selected gene numbers with K-partitioning Lasso algorithm, by using SVM classifier, adopting RBF as kernel function, the specific results is shown in the Table **2**. From Table **2** it can be seen that in case of different numbers of gene were selected, the classification accuracy of K-partitioning Lasso feature selection method is good than that of SNR method. So there shall be conclusion that K-partitioning Lasso method shall be an effective tumor gene feature selection method.

5. SUMMARY

The DNA microarray technology is the strong tool to analyze tumor gene, but the research on gene expression profile data analysis method is still under exploring period, it still face to many challenges, among these a very key task is to select the feature gene. Through improved Lasso method, this article proposed K-partitioning Lasso feature selection algorithm, by using prostatic carcinoma gene expression profile data set, combining SVM to classify the data set, and makes comparison between results of K-partitioning Lasso and SNR. The experiment results show that K-partitioning Lasso method reduced redundant feature, has higher classification accuracy, and such algorithm has good stability, and solves the problem of overhead computational expensive and “over-fitting”. Thus the conclusion shall be given as that K-partitioning Lasso method is an effective tumor feature gene selection method. Based on the study work of this article, the next study may hopefully abstract tumor classification rules with biomedical significance and tumor-related gene, in order to explore the mechanism for tumor emerging and development and gene regulatory network.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This paper is supported by Anhui Province provincial natural science research project No. KJ2014A266.

REFERENCES

[1] M. Zhang, Z. Lv, X. Zhang, G. Chen, and K. Zhang, “Research and Application of the 3D Virtual Community Based on WEBVR and RIA,” *Computer and Information Science,* vol. 2, no. 1, pp. 84, 2009.

[2] T. Su, Z. Lv, S. Gao, X. Li, and H. Lv, “3D seabed: 3D modeling and visualization platform for the seabed,” In*: Multimedia and Expo Workshops (ICMEW)*, IEEE International Conference on, pp. 1-6, 2014.

[3] X. Li, Z. Lv, B. Zhang, W. Wang, S. Feng, and J. Hu, “WebVRGIS Based City Bigdata 3D Visualization and Analysis,” In*: Pacific Visualization Symposium (PacificVis)*, 2015.

[4] Y. Geng, J. Chen, and K. Pahlavan, “Motion detection using RF signals for the first responder in emergency operations: A PHASER project,” In: *24nd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, London,Britain, 2013.

[5] Y. Geng, and K. Pahlavan, “On the Accuracy of RF and Image Processing Based Hybrid Localization for Wireless Capsule Endoscopy,” *IEEE Wireless Communications and Networking Conference (WCNC)*, 2015.

[6] J. He, Y. Geng and K. Pahlavan, “Toward Accurate Human Tracking: Modelling Time-of-Arrival for Wireless Wearable Sensors in Multipath Environment,” *IEEE Sensor Journal*, vol. 14, no. 11, 3996-4006, 2014.

[7] Z. Lv, L. Feng, H. Li, and S. Feng, “Hand-free motion interaction on Google Glass,” In *SIGGRAPH Asia Mobile Graphics and Interactive Applications*, pp. 21, ACM, 2014.

[8] Z. Chen, S. M. Arisona, X. Huang, M. Batty, and G. Schmitt, “Detecting the dynamics of urban structure through spatial network analysis,” *International Journal of Geographical Information Science,* vol. 28, no. 11, pp. 2178-2199, 2014.

[9] W. Li, J. Tordsson, and E. Elmroth, “An aspect-oriented approach to consistency-preserving caching and compression of web service response messages,” In*: Web Services (ICWS)*, 2010 IEEE International Conference on, pp. 526-533, 2010.

[10] S. Li, Y. Geng, J. He, and K. Pahlavan, “Analysis of Three-dimensional Maximum Likelihood Algorithm for Capsule Endoscopy Localization”, In: *5th International Conference on Biomedi-*

cal Engineering and Informatics (BMEI), Chongqing, China, pp. 721-725, 2012.

[11] Y. Geng, J. He, H. Deng and K. Pahlavan, "Modeling the Effect of Human Body on TOA Ranging for Indoor Human Tracking with Wrist Mounted Sensor," In: *16th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, Atlantic City, NJ, 2013.

[12] G. N. Brock, J. R. Shaffer, R. E. Blakesley, M. J Lotz, and G. C. Tseng, "Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes," *BMC Bioinformatics*, 2008.

[13] Y. Saeys, I. Lnza, and P. Larrañaga, "A review of feature selection technique in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507-2517, 2007.

[14] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531-537, 1999.

[15] L. Li, C. R. Weinberg, T. A. Darden, and L.G. Pedersen, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method," *Bioinformatics*, vol. 17, no. 12, pp. 1131-1142, 2001.

[16] X. W. Chen, "Margin-based wrapper methods for gene identification using microarray," *Neurocomputing*, vol. 69, no. 16-18, pp. 2236-2243, 2006.

[17] D. U. Ramón, and A. A. Sara, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, pp. 3, 2006.

[18] S. G. Ma, X. Song, and J. Huang, "Supervised group Lasso with applications to microarray data analysis," *BMC Bioinformatics*, vol. 8, pp. 60. 2007.

[19] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society Series B-Methodological*, vol. 58, no. 1, pp. 267-288, 1996.

[20] B. Efron, T. Hastie, and I. Johnstone, "Least Angle Regression," *Journal of the Institute of Mathematical Statistics*, vol. 32, no. 2, pp. 407-499, 2004.

[21] D. Singh, P. G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, and W.R. Sellers, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203-209, 2002.

[22] Y. D. Zhao, and R. Simon, "BRB-ArrayTools Data Archive for Human Cancer Gene Expression: A Unique and Efficient Data Sharing Resource," *Cancer Informatics*, vol. 6, pp. 9-15, 2008.

Received: June 10, 2015 Revised: July 29, 2015 Accepted: August 15, 2015