Asian Journal of Information Technology 16 (2-5): 206-211, 2017
ISSN: 1682-3915
© Medwell Journals, 2017

An Entropy based Mean Score Feature Selection Method for Identification of Biomarkers using Mirna Expression Profiles for Cancer Classification

[1]M. Anidha and [2]K. Premalatha
[1]Anna University, Chennai, India
[2]Deptartment of CSE, Bannari Amman Institute of Technology,
Sathyamangalam, India

Abstract: MicroRNAs are small non-coding RNA molecules which are important developments in the cancer biology. miRNA microarrays are useful tools to identify potential biomarkers for variety of cancers. Due to high dimensionality of microarrays, it is very hard to identify cancer oncogenes and classify tumor samples. Feature selection is very essential task in the process of classification and identification of biomarker genes by selecting relevant genes. In this research, Entropy Based Mean Score (EBMS) is employed to identify the biomarker genes in miRNA microarrays. This is based on Fisher score which has the benefits of information gain and achieves maximum classification accuracy. The proposed research is tested on benchmark datasets with SVM and ANN for classification. The experimental results show that the EBMS method outperforms the existing methods and it is suitable for effective feature selection.

Key words: Feature selection, fisher score, EBMS, classification, SVM, ANN, 10-fold cross validation

INTRODUCTION

The emergence of microRNAs has been one of the defining developments in cancer biology over the past decade and the explosion of knowledge in this area has brought forward new diagnostic and therapeutic opportunities (Hayes *et al*., 2014). Importantly, each tumor type has a distinct microRNA signature that distinguishes it from normal tissues and other cancer types. Most cancers can be further subclassified into prognostic groups based on these signatures (Lu *et al*., 2005). It is documented that microRNAs have roles in all of the cancer hallmarks defined by Hanahan and Weinberg (2011). Many of the miRNAs identified to date have been associated with cancer. The misregulation of these miRNAs is evident in a broad range of different cancer types, indicating that they can function as conventional tumor suppressors and oncogenes. Examples of oncogenic miRNAs are miR-10b, miR-155, miR-21 and miR-17-92, a cluster of miRNA genes that contains 7 miRNAs (Bader and Lammers, 2011). Microarray technology accelerates the analysis of thousands of miRNA expression profiles simultaneously. However, the most challenging issue in this high dimensional data analysis is huge number of features and small number of samples. According to cancer statistics, if the tumor is diagnosed at early stage then the survival rate is higher with the help of cancer therapeutics. It is very essential to select highly relevant and informative features which leads to classify the samples accurately and identifying biomarkers for cancer therapeutics. Feature selection algorithms can be categorized into supervised (Weston *et al*., 2003; Song *et al*., 2007). Unsupervised (Dy and Brodley, 2004; Mitra *et al*., 2002) and semi-supervised feature selection (Zhao and Liu, 2007; Xu *et al*., 2009, 2010). Supervised feature selection methods can further be broadly categorized into filter models, wrapper models and embedded models. Reducing the number of irrelevant/redundant features can drastically reduce the running time of the learning algorithms and yields a more general classfier. Feature selection for classfication initially performs feature selection to select a subset of significant genes. These significant genes are processed in learning algorithm to know the accuracy of the feature selection method. The feature selection phase might be independent of the learning algorithm like filter models or it may iteratively utilize the performance of the learning algorithms to evaluate the quality of the selected features, like wrapper models. Filter algorithms consist of two steps. Initially it ranks the features based on some criteria and secondly it chooses top ranked features as a subset which are relevant and non-redundant and the subset is suitable for the classifier. Previously, a number of performance criteria have been proposed for filter-based feature selection such as fisher score (Xie and Wang, 2011) methods based on

Corresponding Author: M. Anidha, Anna University, Chennai, India

mutual information (Koller and Sahami,1996; Yu and Liu, 2003) and relief and its variants (Kira and Rendell, 1992). Fisher score computes ranks for the features based on its classes. Instances within one class are assigned values which are different for the instances of different classes (Xie and Wang, 2011). Fisher score evaluates features individually; therefore, it cannot handle feature it cannot handle feature redundancy. Due to its computational efficiency and simple interpretation, information gain is one of the most popular feature selection methods. It is used to measure the dependence between features and labels and calculates the information gain between the feature and the class labels (Koller and Sahami, 1996; Yu and Liu, 2003). The proposed researche adapts the advantage of information gain to overcome the drawback of redundancy with the F-score method by combining it with mean score.

Literature review: Ulfenborg *et al.* (2013) stated that decision trunks have clear advantages over other classifiers due to their transparency, interpretability and their correspondence with human decision-making and clinical testing practices. They implemented k-TSP, a variant of TSP (Top Scoring Pairs) to identify set of k markers and they have implemented polarization score (p-score) as a feature selection technique (Ulfenborg *et al.*, 2013). Sokilde *et al.* (2014) revealed that t-tests were used for each histology and ranking of significant miRNAs. In addition to that the feature selection embedded in the Least Absolute Shrinkage and Selection Operator (LASSO) classification algorithm was applied (Sokilde *et al.*, 2014). Wach (2013) applied binary logistic regression models and a backward elimination method with the likelihood ratio as the determinant (inclusion $p<0.05$; exclusion $p>0.1$) (Wach *et al.*, 2013). Target genes of the best discriminative miRNAs were predicted by the miRanda algorithm were mapped to predefined Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Wach *et al.*, 2013). Lul *et al.* (2005) stated that multi class classification is performed with kNN and euclidean distance measure to reduce the feature space (Lu *et al.*, 2005). Xu *et al.* (2009) revealed that Particle Swarm Optimization (PSO) is used for selecting important miRNAs that contribute to the discrimination of different cancer types and the default ARTMAP is used to classify broad types of cancers based on their miRNA expression fingerprints.

MATERIALS AND METHODS

System description: The proposed feature selection method receives pre-processed high dimensionality microarray data set as an input. The first step is reducing the total number of features in the input data set to a smaller subset of relevant and non-redundant features using the entropy based mean score ranking technique. These significant miRNAs are used by the SVM and ANN for classification. At this point one can measure and record the test classification accuracy which is equal to the number of correctly classified test samples divided by the total number of introduced test samples.

Concept of F-score: Features which are highly discriminative will have smaller distances within the same class and different and larger values to instances from different classes with this idea the F-score is defined as:

$$F_i = \frac{(\mu_{i0} - \mu) + (\mu_{i1} - \mu)}{(\sigma_{i0} + \sigma_{i1})} \tag{1}$$

where, μ_{i0}, μ_{i1} and μ are mean of class 0, class 1 and whole set of ith feature respectively. σ_{i0} and σ_{i1} are respectively the variance of class 0 and class 1 of ith feature.

Concept of Shannon's entropy: Shannon entropy is one of the most important metrics in information theory. Entropy measures the uncertainty associated with a random variable, i.e., the expected value of the information in the message. In information gain, a feature is relevant if it has a high information gain:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log 2p(x_i) \tag{2}$$

Proposed methodology: The input data set is in the form of N×M matrix in which N represents features (miRNAs) and M represents samples. The classification system consists of three steps:

- Feature selection for choosing highly relevant and informative with the help of EBMS method
- Classification with SVM and ANN
- Accessing the performance of classifiers with 10-fold cross validation

Feature selection: To select the highly informative features from miRNAs, F-score based method known as EBMS is employed. Initially the F-score is computed according to Eq. 1. To calculate the mean score the following strategy is applied:

- Number of instances greater than mean of class 0
- Number of instances less than mean of class 0
- Number of instances greater than mean of class 1
- Number of instances less than mean of class 1

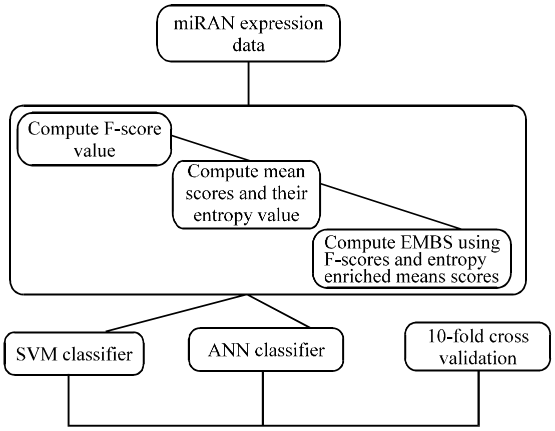
miRAN expression data
Compute F-score value
Compute mean scores and their entropy value
Compute EMBS using F-scores and entropy enriched means scores
SVM classifier
ANN classifier
10-fold cross validation

Fig. 1: Block diagram of the proposed method

Using the above values entropy is computed according to the Eq. 2 and the F-score values are modified according to their entropy values. miRNAs are ranked according to the EBMS values. Top ranked miRNAs are selected for classifiers input.

Pseudocode:

```
• Input the miRNA expression samples $S_{i=1...n}$
• Compute $F_i$ for all the features
• Compute the mean scores a,b,c and d
• Compute entropy for the above mean scores H(a), H(b), H(c), H(d) and the mean score is calculated as:

    $MS = (H(a)+H(d))-(H(b)+H(c))$

• $EBMS = F_i/MS$
• All the features are ranked according to their EBMS values
• Top ranked 100 and 50 miRNAs are selected as an input to the classifiers
```

Classification: Support Vector Machines (SVM) is a learning method based on statistical learning theory proposed by Vapnik and it is a powerful tool for data classification and estimation (Fig. 1). SVM can handle a nonlinear classification efficiently by mapping input samples from the input space into a high dimensional feature space with the nonlinear kernel function. In the feature space, SVM tries to maximize the generalization performance by solving a quadratic programming optimization problem and finds the optimal separating hyperplane. Artificial Neural Networks (ANN) are excellent computational models that have been implemented to solve different kind of problems. The pattern classification, forecasting and regression problems are areas where the ANN have demonstrated to be an efficient technique (Garrro *et al*., 2016). ANN have been widely applied in DNA microarrays (Garro *et al*., 2016). The 10-fold cross validation method is a scaled down version of the leave one out method where the dataset is divided into 10 partitions. Each of these partitions is used as a testing set while the classifier is trained on the remaining samples.

RESULTS AND DISCUSSION

Experimental design: The datasets used in this study are available at http://sourceforge.net/projects/trunk classifier/files. This datasets is part of the collection of database, created by Ulfenborg *et al.* (2014), Karin Klinga-Levan and Bjorn Olsson, Systems Biology research Centre, School of Life Sciences, University of Skovde, Sweden (Benjamin *et al*., 2013). Table 1 shows the description of the dataset. ST-stage; TR-treatment status; GR-grade; ER-estrogen receptor status; MY-MYC amplification status; HI-histological subtype; DI-differentiation status; SU-survival. In order to evaluate the performance of the proposed method, the measures like sensitivity, specificity and accuracy are considered. The measures are computed using the following equation (Table 2):

$$\text{Classification accuracy (\%)}: (TP+TN)/(TP+FP+FN+TN)$$

$$\text{Sensitivity (\%)} = TP/TP+FN\times 100$$

$$\text{Specificity (\%)} = TN/FP+TN\times 100$$

The proposed work is implemented using R version 3.2.4. The datasets used in this paper contain expression values of miRNAs from cancers of Lung, Breast and neuroblastomas. Most of the datasets were used for more than one classification tasks such as normal versus malignant and early versus late stage as in Sotiriou Breast cancer data. For breast cancer datasets, histologic grade 1 was considered as low grade and histologic grade >1 as high grade. For neuroblastoma datasets, the International Neuroblastoma Staging System (INSS) stage 1-2 was defined as early stage and INSS stage >2 as late stage (Ulfenborg *et al*., 2013). The Wang Y, Sotiriou datasets were log 2-transformed before classification (Ulfenborg *et al*., 2013). The evaluation of classification accuracy was performed using 10-fold cross validation. The average accuracy for the given dataset was defined as the proportion of correctly classified test samples (which is equivalent to the number of true positives plus the number of true negatives divided by the total number of

Table 1: miRNA datasets

Data set name	Cancer	Number of probes	Class	Total samples
Angulo_HI	Lung	20185	Adenocarc./Squam	66
Angulo_DI	Lung	20185	Well diff./ Poorly diff.	51
Takeuchi_HI	Lung	21619	Adenocarc./Squam.	125
Takeuchi_DI	Lung	21619	Well diff./poorly diff.	59
Takeuchi_SU	Lung	21619	Alive/dead	149
WangY	Breast	22283	ER+/ER-	286
VandeVijver_ER	Breast	13359	ER+/ER-	295
VandeVijver_SU	Breast	13359	Alive/dead	295
Sotiriou_TR	Breast	22283	Tamoxifen/untreated	189
Sotiriou_GR	Breast	22283	Low grade/hgrade	167
Sotiriou_ER	Breast	22283	ER+/ER-	183
WangQ_ST	Neurobl	12625	Early stage/late stage	101
WangQ_MY	Neurobl	12625	MYC-/MYC+	101

Table 2: Classification performance with SVM and ANN

	Sensitivity (%)		Specificity (%)		ClassificationAccuracy (%)	
Data set	SVM	ANN	SVM	ANN	SVM	ANN
Angulo_HI	98	98.0	100	95	98.80	95.40
Angulo_DI	85	75.0	84	99	84.00	88.00
Takeuchi_HI	97	98.0	94	88	96.45	95.20
Takeuchi_DI	91	88.0	92	91	91.03	90.80
Takeuchi_SU	59	66.0	68	67	61.89	66.80
WangY	77	79.6	97	88	91.22	85.40
VandeVijver_ER	94	87.0	95	96	94.83	93.80
VandeVijver_SU	79	81.0	51	44	73.30	71.00
Sotiriou_TR	99	100	100	74	99.57	90.60
Sotiriou_GR	81	78.0	68	64	76.10	72.28
Sotiriou_ER	95	43.0	88	95	88.78	84.98
WangQ_ST	92	84.0	78	71	87.71	81.14
WangQ_MY	98	99.0	93	90	97.20	97.60

Table 3: Biomarkers from the chosen miRNAs

Data set	ID Ref/probe set/entrez Id	Description
Angulo_HI	2335322869	SUN1-Sad1 and UNC84 domain containing 1ZNF510-zinc finger protein 510
Angulo_DI	67536754	SSTR3-somatostatin receptor 3SSTR4-somatostatin receptor 4
Takeuchi_HI	68149034	STXBP3-syntaxin binding protein 3CCRL2-chemokine (C-C motif) receptor-like 2
Takeuchi_DI	317322	ACACA-acetyl-CoA carboxylase alphaUBE2D2-ubiquitin-conjugating enzyme E2D 2
Takeuchi_SU	5272	SERPINB9-serpin peptidase inhibitor, clade B (ovalbumin), member 9
WangY	205225_at209603_at	ESR1-estrogen receptor 1GATA3-GATA binding protein 3
VandeVijver_ER	KIAA0575KIAA0882NAT1	GREB1TBC1D9N-acetyltransferase1(arylamine N-acetyltransferase)
VandeVijver_SU	KIAA0098KIAA0159TPI1	T-complex protein 1 subunit epsilon.Condensin complex subunit 1.Triosephosphate isomerase is an enzyme that in humans is encoded by the TPI1 gene.
Sotiriou_TR	216844_at	ZC3H7B-zinc finger CCCH-type containing 7B
Sotiriou_GR	211762_s_at214710_s_at202779_s_at	KPNA2-karyopherin alpha 2 CCNB1-Cyclin B1UBE2S- ubiquitin-conjugating enzyme E2S
Sotiriou_ER	208628_s_at208627_s_at204751_x_at	YBX1-Y box binding protein 1DSC2-desmocollin 2
WangQ_ST	36096_at37915_at	REEP1-receptor accessory protein 1BAZ2B-bromodomain adjacent to zinc finger domain, 2B
WangQ_MY	35158_at	MYCN v-myc myelocytomatosis viral related oncogene, neuroblastoma derived

samples). The EBMS technique had achieved a high accuracy value of 99.57% for Sotiriou-TR dataset and the average classification accuracy of 87.76% with SVM and 85.62% with ANN. The highest 95% CI value for Angulo dataset is 0.8424-0.9992, Takeuchi dataset is 0.865-0.9899, WangY data set is 0.7172-0.8837, Vande Vijver is 0.8785-0.9669, Sotiriou dataset is 0.8662-0.9762 and for WangQ dataset is 0.8935-0.9995. With the help of Empirical analysis the number features selected as an input to the classifier are 100 which attains maximum classification accuracy for all the datasets (Table 3).

The DSC2 identified from sotiriou-ER is one of the important marker gene of breast cancer Metastasis (Culhane and Quackenbush, 2009). The over expression of KPNA2 is directly related with Invasive Ductal Carcinoma (IDC) which will remain localized and it helps to identify the primary origin. The over expressions of cyclin B1 identified from sotiriou-GR play an important role in human breast carcinogenesis. ZGPAT is a gene which encodes zinc finger CCCH and is a tumor suppressor of breast carcinogenesis. GREB1 acts as a regulator of hormone-dependent cancer growth in breast

Table 4: Comparison of classification Accuracy with existing works

	Proposed method		
Data set	SVM	ANN	P-Score and DTC Ulfenborg*et al.* (2013)
Angulo_HI	98.80	95.40	89.39
Angulo_DI	84.00	88.00	74.50
Takeuchi_HI	96.45	95.20	94.40
Takeuchi_DI	91.03	90.80	86.44
Takeuchi_SU	61.89	66.80	68.45
WangY	91.22	85.40	89.86
VandeVijver_ER	94.83	93.80	100.00
VandeVijver_SU	73.30	71.00	71.19
Sotiriou_TR	99.57	90.60	98.94
Sotiriou_GR	76.10	72.28	76.04
Sotiriou_ER	88.78	84.98	76.50
WangQ_ST	87.71	81.14	83.16
WangQ_MY	97.20	97.60	100.00

Table 5: Comparison of performance with other techniques

References	Techniques	Averageclassification accuracy (%)
Ulfenborg *et al.* (2013)	P-Score and DTC	85.29
Sokilde *et al.* (2014)	LASSO	85.00
Wach *et al.* (2013)	Backward elimination method with logistic regression	86.00
This research	EBMSand SVM	87.76

and prostate cancers. GATA3 can be particularly useful as a marker for metastatic breast carcinoma, especially triple-negative and metaplastic carcinomas which lack specific markers of mammary origin. Finally, GATA3 labeling may help distinguish metaplastic carcinoma from malignant phyllodes tumors (Table 4 and 5).

Lysosomal-associated transmembrane protein 4B is a protein that in humans is encoded by the LAPTM4B gene. Over expression of LAPTM4B has been found in breast cancer and elevated LAPTM4B level contributes to chemotherapy resistance in breast cancer. It was found that over expression of LAPTM4B causes anthracyclines (doxorubicin, daunorubicin and epirubicin) resistance by retaining drug in the cytoplasm and decreasing nuclear localization of drug and drug induced DNA damage. Y box binding protein 1 also known as Y-box transcription factor or nuclease-sensitive element-binding protein 1 is a protein that in humans is encoded by the YBX1 gene (Eliseeva *et al*., 2011). YBX1 is a potential drug target in cancer therapy. YB-1 helps the replication of adenovirus type 5, a commonly used vector in gene therapy. Thus, YB-1 can cause an "oncolytic" effect in YB-1 positive cancer cells treated with adenoviruses (Eliseeva *et al*., 2011). Ubiquitin-conjugating enzyme E2S (UBE2S) plays important role in breast cancer treatment (Ayesha *et al*., 2016). Two new studies report the identification of activating ESR1 gene mutations in aromatase inhibitor-resistant metastatic breast cancers. This insight into therapeutic resistance suggests new approaches that may be useful in the management of endocrine-resistant breast cancer (Oesterreich and Davidson, 2013). Mutations in ESR1, the gene encoding the ER, found in approximately 20% of patients with metastatic ER-positive disease who received endocrine therapies such as tamoxifen and aromatase inhibitors. These mutations are clustered in a 'hotspot' within the Ligand-binding Domain (LBD) of the ER and lead to ligand-independent ER activity that promotes tumour growth, partial resistance to endocrine therapy and potentially enhanced metastatic capacity (Jeselsohn *et al*., 2015). SSTR3-somatostatin receptor 3, SSTR4-somatostatin receptor 4 are useful tool in the clinical management of neuroendocrine lung cancers (Muscarella *et al*., 2011; Landemaine *et al*., 2008).

CONCLUSION

In this study, it is proved that the proposed Entropy based mean score feature selection technique had achieved maximum classification accuracy when it is combined with SVM classifier compared to ANN classifier. The results shown above depicted that the better performance of the proposed technique for the datasets selected than the existing techniques. The algorithm is tested with different cancer types with different classes such as histological subtypes, stages, survival state, estrogen receptor positive and negative, etc., the classification accuracy and the 95% CI values are maximum with the help of highly informative and relevant features. The chosen signature miRNAs and genes are very useful to know about the cancer stage and more beneficial for the cancer therapeutics.

REFERENCES

Ayesha, A.K., T. Hyodo, E. Asano, N. Sato and M.A. Mansour et al., 2016. UBE2S is associated with malignant characteristics of breast cancer cells. Tumor Biol., 37: 763-772.

Bader, A.G. and P. Lammers, 2011. The therapeutic potential of microRNAs. Innovations Pharm. Technol., 1: 52-55.

Culhane, A.C. and J. Quackenbush, 2009. Confounding effects in: A six-gene signature predicting breast cancer lung metastasis. Cancer Res., 69: 7480-7485.

Dy, J.G. and C.E. Bradley, 2004. Feature selection for unsupervised learning. J. Mach. Learn. Res., 5: 845-889.

Eliseeva, I.A., E.R. Kim, S.G. Guryanov, L.P. Ovchinnikov and D.N. Lyabin, 2011. Y-box-binding protein 1 (YB-1) and its functions. Biochem., 76: 1402-1433.

Garro, B.A., K. Rodriguez and R.A. Vazquez, 2016. Classification of DNA microarrays using artificial neural networks and ABC algorithm. Appl. Soft Comput., 38: 548-560.

Hanahan, D. and R.A. Weinberg, 2011. Hallmarks of cancer: The next generation. Cell, 144: 646-674.

Hayes, J., P.P. Peruzzi and S. Lawler, 2014. MicroRNAs in cancer: Biomarkers, functions and therapy. Trends Mol. Med., 20: 460-469.

Jeselsohn, R., G. Buchwalter, C.D. Angelis, M. Brown and R. Schiff, 2015. ESR1 mutations a mechanism for acquired endocrine resistance in breast cancer. Nat. Rev. Clin. Oncol., 12: 573-583.

Kira, K. and L. Rendell, 1992. A practical approach to feature selection. Proceedings of the 9th International Workshop on Machine Learning, July 1-3, 1992, California: Morgan Kaufmann, pp: 249-256.

Koller, D. and M. Sahami, 1996. Toward optimal feature selection. Proceedings of the International Conference on Machine Learning, July 3-6, 1996, Bari, Italy, pp: 284-292.

Landemaine, T., A. Jackson, A. Bellahcene, N. Rucci and S. Sin *et al*., 2008. A six-gene signature predicting breast cancer lung metastasis. Cancer Res., 68: 6092-6099.

Mitra, P., C.A. Murthy and S.K. Pal, 2002. Unsupervised feature selection using feature similarity. IEEE Trans. Pattern. Anal. Mach. Intell., 24: 301-312.

Muscarella, L.A., V. D'Alessandro, A.L. Torre, M. Copetti and A.D. Cata *et al*., 2011. Gene expression of somatostatin receptor subtypes SSTR2a, SSTR3 and SSTR5 in peripheral blood of neuroendocrine lung cancer affected patients. Cell. Oncol., 34: 435-441.

Oesterreich, S. and N.E. Davidson, 2013. The search for ESR1 mutations in breast cancer. Nat. Genet., 45: 1415-1416.

Sokilde, R., M. Vincent, A.K. Moller, A. Hansen and P.E. Hoiby *et al*., 2014. Efficient identification of miRNAs for classification of tumor origin. J. Mol. Diagn., 16: 106-115.

Song, L., A. Smola, A. Gretton, K.M. Borgwardt and J. Bedo, 2007. Supervised feature selection via dependence estimation. Proceedings of the 24th International Conference on Machine Learning, June 20-24, 2007, Corvallis, OR., pp: 823-830.

Ulfenborg, B., K.K. Levan and B. Olsson, 2013. Classification of tumor samples from expression data using decision trunks. Cancer Inf., 12: 53-66.

Wach, S., E. Nolte, A. Theil, C. Stohr and T.T. Rau et al., 2013. MicroRNA profiles classify papillary renal cell carcinoma subtypes. Br. J. Cancer, 109: 714-722.

Weston, J., A. Elisseeff, B. Scholkopf and M. Tipping, 2003. Use of the zero norm with linear models and kernel methods. J. Mach. Learn. Res., 3: 1439-1461.

Xie, J. and C. Wang, 2011. Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. Ex. Syst. Appl., 38: 5809-5815.

Xu, R., J. Xu and D.C. Wunsch, 2009. MicroRNA expression profile based cancer classification using default ARTMAP. Neural Netw., 22: 774-780.

Xu, Z., R. Jin, J. Ye, M. Lyu and I. King, 2010. Discriminative semi-supervised feature selection via manifold regularization. Proceedings of the 21th International Joint Conference on Artificial Intelligence, June 21-July 8, 2010, Germany, pp: 1033-1047.

Yu, L. and H. Liu, 2003. Feature selection for high-dimensional data: A fast correlation-based ?lter solution. Proceedings of the International Conference on Machine Learning, August 21-24, 2003, ACM, Washington, DC., USA., ISBN:1577351894, pp: 856-863.

Zhao, Z. and H. Liu, 2007. Semi-Supervised Feature Selection Via Spectral Analysis. In: Proceedings of the SIAM International Conference on Data Mining, Apte, C.V. (Ed.). Society for Industrial and Applied Mathematics Publisher, Philadelphia, Pennsylvania, ISBN:9780898716306, pp: 641-646.