Comparative Study of ANFIS-Based Wrapper Model for Classification of Cancer and Normal Genes on Microarray Gene Expression Data

Sarita Chauhan, Aakashdeep Sharma, Abhishek Brahmabhatt, Namrata Singh, Puneet Sharma

Abstract- A novel way to enhance the performance of a model that combines genetic algorithms and neuro fuzzy logic for feature selection and classification is proposed. This research work involves designing a framework that incorporates genetic algorithm with neuro fuzzy for feature selection and classification on the training dataset. It aims for reducing several medical errors and provides better prediction of diseases. Medical diagnosis of diseases is an important and difficult task, and a proposed method performs feature selection and parameters setting in an evolutionary way. The wrapper approach to feature subset selection is used in this paper because of the accuracy. The performance of the ANFIS classifier was evaluated in terms of training performance and classification accuracy. The objective of this research is to simultaneously optimize the parameters and feature subset without degrading the ANFIS classification accuracy.ANFIS is compared with three other classifiers which are Support Vector Machine (SVM), K-Nearest Neighbour (KNN) and Classification And Regression Trees (CART). ANFIS gives the best results for original data of all the datasets and the predictions for noisy data are adequate in comparison with three others classifiers.

Keywords - ANFIS; Feature Selection; ; Cancer Classification

I. INTRODUCTION

Genetic Algorithm (GA) was first suggested by Holland, and has recently been used in a range of problems including pattern recognition, bioinformatics and text categorization. Early detection of medical problems such as prostate cancer and diabetese is important to increase the chance of successful treatment. Various soft computing methods have been used for the detection of a potential medical problem. Thus, a reliable method for both feature selection and classification is required. The feature selection is based on a new genetic algorithm and classification is based on Adaptive neuro fuzzy inference system (ANFIS). Feature selection is another factor that impacts classification accuracy. Many practical pattern classification tasks require learning an appropriate classification function that assigns a given input pattern, typically represented by a vector of attribute values to a finite set of classes.

Manuscript Received on March 2015.

Sarita Chauhan, Asst. Prof., M L V Textile and Engineering College Bhilwara, India.
Aakash Deep Sharma, Under Graduate, B.Tech Student, M L V Textile and Engineering College Bhilwara, India.
Abhishek Brahmabhatt, Under Graduate, B.Tech Student, M L V Textile and Engineering College Bhilwara, India.
Namrata Singh, Under Graduate, B.Tech Student, M L V Textile and Engineering College Bhilwara, India.
Puneet Sharma, Under Graduate, B.Tech Student, M L V Textile and Engineering College Bhilwara, India.

Feature selection is used to identify a powerfully predictive subset of fields within a database and reduce the number of fields presented to the mining process. By extracting as much information as possible from a given data set while using the smallest number of features, we can save significant computation time and build models that generalize better for unseen data points. These include the clustering and classification because it is essential to the developments of neuro fuzzy systems particularly in medical-related problems. After a training phase with train data, the classifiers such as Support Vector Machine(SVM), K-Nearest Neighbor(KNN), Classification And Regression Trees (CART), would map some of input features to one of the existing labels.

Microarrays usually have many biomarkers as same as features in datasets. Every biomarker shows an expression level of a gene. Microarrays have huge redundancy and high dimensionality, hence feature selection or in other words, gene selection is a main phase of microarray samples classiication for cancer prognosis. In general, feature selection methods can be divided into two categories. In the first category based on filtering[1], feature selection and classifica- tion method are proceeded distinctively. A single or multiple selection criteria must satisfy to identify a final subset of features. Because of independence between two phases, filter model is fast but it needs to select exacting methods and limitations for achieving high precision. In the second category is called wrapped model[2], feature selection process must be embedded for each classifier and precision is achieved overall. This approach needs more computation than the filtering method but it is possible to achieve better precision because of optimized feature selection for particular classifier. A classifier should be run with different groups of candidate features and after evaluating the result, those which are more efficient would be selected.

In this work we apply Adaptive Neuro-Fuzzy Inference System (ANFIS) as a classifier to select genes using wrapped approach of cancer microarrays. Then we add noise to our selected genes of datasets and compute precisions again. Finally we compare our result for original and noisy data with three other classifiers included SVM[7], KNN[8] and CART[9].

II. GENE SELECTION

Feature selection problems are NP-hard[11]. There are thousands of genes in our test bed in this paper which make our work difficult.

So, we used meta-heuristic methods for gene selection. Because of high dimensionality these choices are justifiable. Furthermore, due to existing of redundancy and several possible optimal solutions in our datasets, these methods often can find one of them for problem.

In next topics, it is described two meta-heuristic methods; genetic algorithm[12] and particle swarm optimization[13] which we used in this paper.

Genetic Algorithm

Genetic Algorithm (GA) is an optimization method which was introduced by John Holland in the 1970s for the first time. GA helps researchers to find single or multiple enough good solutions for optimization problems. To apply this method, a population of possible solutions for a real world problem must be encoded to some unique strings, called chromosome. Each chromosome is made up of some genomes and represents an individual. Genomes may be encoded to binary, real number, character or other formats. Besides, we would have one or several objective function(s). Through objective function(s), calculation of profit or cost for every individual is possible. At the beginning, a generation of individuals is initialized randomly. Subsequently, every individual is evaluated and ranked by the objective function(s). Better individuals based on their fitness would be selected as the parents for the next generation reproduction. Also, some of the best individuals will be copied to the next generation directly as elites. A small part of population would be mutated to generate new individuals which will have some changed blocks that we did not see them in the previous generation. New generations will be reproduced by some operators like crossover, mutation and replacement. These operators work by objective function(s) values and some stochastic variables in some specified intervals. This process would be repeated enough to satisfy one or multiple stop conditions. Steps of our implemented algorithm are expressed below:

0: Initialize the first generation. 1:
Evaluate the first generation.
2: Until stop condition is not satisfied, do:
3: Select best parents.
4: Reproduce new children from their parents with probability of Pr .
5: Mutate every chromosome with probability of Pm.
6: Evaluate and rank new generated population.
7: Replace Pr percent of the best of new population.
8: Go to 2.
9: Represent the best solutions based on their chromosomes.
10:End

In this work we assume Pr= 0.8, Pm =0.4, and Pr=0.2 when the maximum of iterations equals 500 and the number of population equals 50. Furthermore, the reproduction method is one point crossover.

III. ANFIS

Zadeh proposed the fuzzy set theory in the 1960s. Fuzzy set theory through its general approaches has covered classical sets theory[15]. In a fuzzy set, every element belongs to a membership function with a degree of membership. These degrees are real numbers between closed interval [0, 1]. A fuzzy set would be equivalent to a classical set (also known as crisp set in fuzzy set theory) when exacting degrees equal 0 or 1. Fuzzy logic handles changing of the variables and reasoning from them. Then Zadeh introduced possibility theory versus probability theory that every membership degree considers as a possibility measure. Appropriate operators in possibility theory obtain preliminary for different types of fuzzy reasoning. After these events, researchers developed new concepts based on fuzzy sets theory such as Fuzzy Inference System (FIS). Different FISes are based on some components like fuzzifiers for inputs, defuzzifiers for outputs, fuzzy knowledge base contained fuzzy if-then rules and inference engine. FISes generate crisp or fuzzy outputs by a wide variety of crisp or fuzzy input. Inference engine in FIS maps fuzzy inputs to a fuzzy output by fuzzy if-then rules. Takagi, Sugeno and Kang proposed a fuzzy inference method in the 1980s (known as TSK). Format of TSK type fuzzy if-then rules is like this:

If x is A_1 and y is B_1 , then $f_1 = p_1x + q_1 y + r_1$,

Where A_1 and B_1 are fuzzy sets or linguistic labels and x and y are corresponding values and subsequently f_1 is a crisp function.

Another concept has been interested by researchers in artificial intelligence fields is model learning based on an input-output pair. Therefore artificial Neural Networks (ANNs) which have the capability of the different kinds of learning methods were developed for the last several decades. These networks are composed from 3 layers (input layer, hidden layer and output layer) within some nodes. There are some weighted links with interconnection roles among nodes of layers. Every nod has a firing threshold to generate an output depends on sum of inputs of nods. Furthermore, there are some various architectures of ANN for similar and different functions.

Jang combined the capability of soft transitions between concepts and uncertain data in fuzzy logic with the available learning potentialities of neural networks in a new concept, called neuro-fuzzy system[16]. Adaptive Neuro-Fuzzy Inference System (ANFIS) is a type of neuro-fuzzy system which was introduced by Shing and Jang in 1990s. ANFIS has been a brilliant multifunction tool in many fields; however it widely used for nonlinear mapping function and prediction. In this paper we have used ANFIS as a predictor for cancer diagnosis. ANFIS composed from five layers. First layer compute membership degrees for real number input values. Every node i output is computed such as:

$O_1^1 = \mu_{Ai}(x)$, i = 1,2,3,.....,n.....(3)

Where O_1^1 represents degrees of x input that is applied to membership function of A_i . Every continuous and piecewise differentiable function which results between closed interval[0, 1] is acceptable in this layer as membership function. Thus In this paper we use double sigmoid function (dsigmf) whichcalculate from the difference between two sigmoidal functions, as membership function of our designed ANFIS classifier.dsigmf results more appropriate than Gaussian function which has used before in[17] for same application.

$\mu(x) = dsigmf(x,a_1 , c_1 , a_2 , c_2)$
$= \mu_1 (x , a_1 , c_1) - (x , a_2 , c_2)$

$= 1/ (1+ \exp (-a_1 * (x-c_1))) - 1/ (1+ \exp (-a_2 * (x-c_2)))$.....(4)

a_1 , c_1 , a_2 , c_2 in this layer called premise parameters. Number of membership functions is not a fix number. There is not any clear approach to find optimal number of nodes in hidden layers in the same way as neural networks. In this work, we assign two membership functions for each input.

Inputs of nodes in second layer are multiplied to each other. In other words, nodes of this layer capture their inputs from first layer and pass their products as firing strength.

$O_i^2 = w_i = {}^{Qm}_{j=1} \mu_{Ai}(x_j)$, i = 1,2,.....,n..... (5)

Where w_i represents firing strength for m input. Every T-Norm can be used in this equation as a multiple operators.

The third layer plays normalization role in ANFIS Architecture. Normalized weight would be computed from the firing strength of all nodes of second layer, relation is:

$O_i^3 = \bar{w}_i = w_i / {}^{Pn}_{j=1} w_j$, i = 1,2,...n..... (6)

Every node in the 4th layer has a function f_i which represents by some parameters called consequent parameters. The normalized weights from third layer would multiply by corresponding parametric functions. For linear function contained two inputs x and y, output is computed by following equation:

$O_i^4 = \bar{w}_i f_i = \bar{w}_i (p_i x + q_i y + r_i)$.....(7)

Furthermore, different types of functions in this layer are possible. Though constant functions delay convergence in nonlinear mapping, they improve prediction and ANFIS is needed as a predictor in cancer diagnosis. So we choose constant function in this work for each rule.

Final output in layer 5th is sum of total outputs of layer 4th:

$O_i^5 = {}^{P}_{i} \bar{w}_i f_i = {}^{P}w_i f_i / {}^{P}w_i$.....(8)

There are many possible learning methods for training ANFIS like any other artificial neural networks, however researchers often use classical back propagation or hybrid learning rule. Hybrid learning rule together with ANFIS were introduced by Shing and Jang[16]. In this article we have used Hybrid learning rule for training our ANFIS classifier Because of its advantage such as high speed and high performance in computing versus back propagation method.

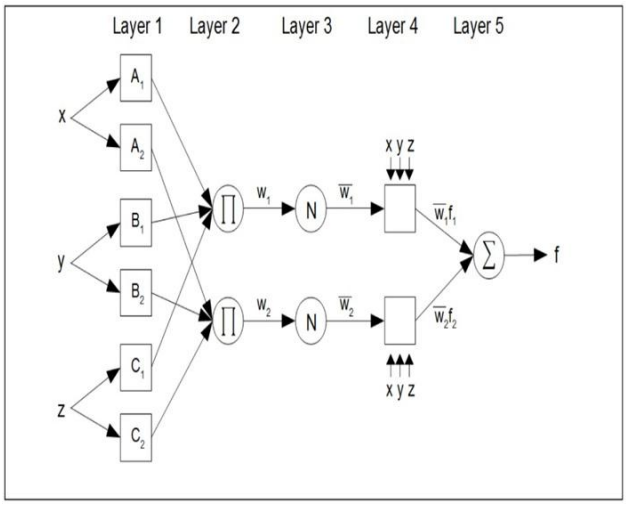

Figure 1. ANFIS layer model

IV. KNN & SVM

Support vector machine (SVM) is one of the most powerful supervised learning algorithms in gene expression analysis. The samples intermixed in another class or in the overlapped boundary region may cause the decision boundary too complex and may be harmful to improve the precise of SVM. In the present paper, hybridized k-nearest neighbor (KNN) classifiers and SVM (HKNNSVM) is proposed to deal with the problem of samples in the overlapped boundary region and to improve the performance of SVM.

V. CART

The CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). K-Nearest Neighbor method can create both classification and regression models as well.

VI. EXPERIMENTAL RESULTS

In this research we use GA as a gene selector for our ANFIS classifier in wrapped model, mentioned before. Therefore selected genes and their values pass to the ANFIS and neuro-fuzzy classifier was trained during iterations. Output of ANFIS using thresholds change into piecewise constant function that every constant represent a label[18]. Subsequently, test data are applied to trained classifier and classifier predicts a label value for each input vector. The precision would be computed based on ratio of number of correct classified samples to number of all test data.. Group of genes is chosen which are more robust after adding some noise than the other when precisions of GA and PSO are equal. We apply ANFIS, SVM, KNN and CART, as classifiers into six cancer datasets[19] and four normal data sets which include:

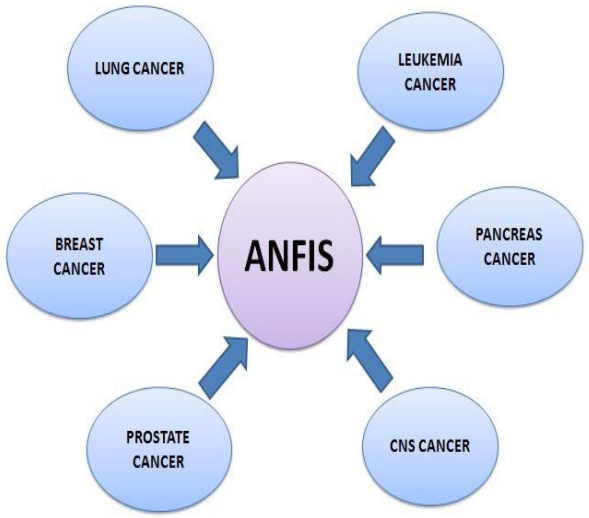

Figure 2: Anfis classifier with diseases

- Prostate Cancer:
  Prostate cancer dataset contained 260 training and 48 testing of relapse and non-relapse samples from patients in 15009 genes.
- Leukemia:
  Dataset has 260 training and 48 testing samples over 15009 probes.

- Lung:
  Colon tumour contained 308 instances which has been divided by us to 260 training and 48 testing samples genes from 15009.
- Pancrease Cancer:
  This dataset includes 260 training and 48 testing normal and tumour instances which every instance is described by 15009 genes.
- Breast Cancer:
  This dataset includes 260 training and 48 testing tissue samples with 15009 genes.
- CNS:
  Dataset has 260 training and 48 testing samples over 15009 probes.
- Normal Brain:
  Dataset has 260 training and 48 testing samples over 15009 probes.
- Normal Kidney:
  Dataset has 260 training and 48 testing samples over 15009 probes.
- Normal Lung:
  Dataset has 260 training and 48 testing samples over 15009 probes.

- Normal Breast:
  Dataset has 260 training and 48 testing samples over 15009 probes.

The selected genes and their precisions in binary and ternary groups with GA method for four classifiers are shown in TABLE II. All the algorithms run thorough the MATLAB 7.14 under Windows operating system.

VII. CONCLUSIONS

In this paper we compared four classifiers on ten datasets which used the selected genes by GA method, in wrapped model. It was predictable that our results using the ANFIS-Based wrapped model for fewer genes, was more appropriate than filter model of single ANFIS and was equivalent or better than filter model of ensemble ANFIS (contained several the different single ANFISes) invilved Information Gain (IG) as gene selection approach in[17].Though, the classic SVM had very good results, ANFIS-Based wrapped model results were better than the results of the SVM, in general. Altogether ANFIS classifier in comparison with the other ones gave good results for each of binary and ternary groups of the genes. Finally, obtained results were more interpretable than the results of the other classifiers, because of final acquired fuzzy rules which are learned by ANFIS.

CANCER TYPES	GROUPS	ANFIS	KNN	SVM	CART
PROSTATE	2G	100	98.20	100	100
	3G	100	100	100	97.40
LEUKEMIA	2G	100	100	100	100
	3G	100	100	100	100
LUNG	2G	100	97.38	100	100
	3G	100	100	100	100
PANCREAS	2G	100	100	100	98.20
	3G	100	100	100	100
BREAST	2G	100	100	100	100
	3G	100	100	100	100
CNS	2G	100	98.80	100	100
	3G	100	100	100	100
NORMAL BRAIN	2G	100	100	100	99.50
	3G	100	100	100	100
NORMAL KIDNEY	2G	100	100	100	100
	3G	100	100	100	100
NORMAL LUNG	2G	100	99.56	100	99.40
	3G	100	100	100	100
NORMAL BREAST	2G	100	100	100	100
	3G	100	100	100	100

TABLE III DIFFERENT PRECISIONS FOR THE ORIGINAL DATA.

CANCER TYPES	GROUPS	ANFIS	KNN	SVM	CART
PROSTATE	2G	3125,4832	10123,4321	3451,2750	0643,6792
	3G	7654,3648	6357,8996	6767,9898	4217,9231
LEUKEMIA	2G	97587,3845	12134,5674	3536,7777	0233,8341
	3G	8357, 8354	11234,7654	7652,7809	13082,9012
LUNG	2G	3689,8654	10987,9012	10387.10239	5741,8364
	3G	9465,2578	10243.1000	9835,3867	6262,6729
PANCREAS	2G	0976,2546	1111,1243	3565,2876	6390,1345
	3G	8765,9367	3343,8890	10001,12002	2901,835
BREAST	2G	9836,4839	0983,12345	1345,15009	10364,12083
	3G	4684.8536	14897,8643	14999,15006	0456,2371
CNS	2G	3568,7864	2453,2654	12456,12908	7354,8282
	3G	9734,2346	5462,7934	12956,2498	12548,9328
NORMAL BRAIN	2G	9089,2222	2537,10678	2099,9264	5674,1381
	3G	5463,9999	10678,4444	1022,7248	7817,9100
NORMAL KIDNEY	2G	2221,3232	2675,12999	12028,3729	0034,1582
	3G	3454,5656	13765,13456	7123,10372	9025.4012
NORMAL LUNG	2G	9876,1209	3456,6515	9801,9456	2019.9021
	3G	9709,1234	1313,4242	6489,4567	7820,3014
NORMAL BREAST	2G	3214,7624	5673,7585	12678,12098	8029,6940
	3G	5325,9357	2743,1443	2341,4563	4731,3065

TABLE I INDEXES OF THE SELECTED GENES IN BINARY, TERNARY AND QUATERNARY GROUPS, FOR EACH OF THE CLASSIFIERS IN THE DIFFERENT DATASETS.

CANCER TYPES	GROUPS	ANFIS	KNN	SVM	CART
PROSTATE	2G	100	95.31	99.32	100
	3G	100	100	100	95.32
LEUKEMIA	2G	100	95.42	100	100
	3G	100	100	100	100
LUNG	2G	100	96.33	99.24	100
	3G	100	100	100	100
PANCREAS	2G	100	96.44	100	100
	3G	100	100	100	100
BREAST	2G	100	97.49	100	100
	3G	100	100	100	96.45
CNS	2G	100	98.68	100	100
	3G	100	100	100	100
NORMAL BRAIN	2G	100	98.77	100	100
	3G	100	100	100	99.23
NORMAL KIDNEY	2G	100	98.26	100	100
	3G	100	100	100	100
NORMAL LUNG	2G	100	99.95	100	100
	3G	100	100	100	99.76
NORMAL BREAST	2G	100	99.36	`00	100
	3G	100	100	100	100

TABLE II MAXIMUM OBTAINED FITNESS FOR THE CLASSIFIERS AND DATA SETS IN BINARY AND TERNARY GROUPS.

REFERENCES

1. C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Now, A survey on filter techniques for feature selection in gene expression microarray analysis, Ieeeacm Trans. Comput. Biol. Bioinforma. Tcbb, vol. 9, no. 4, pp. 1106 1119, 2012.
2. A. El Akadi, A. Amine, A. El Ouardighi, and D. Aboutajdine, A twostage gene selection scheme utilizing MRMR filter and GA wrapper, Knowl. Inf. Syst., vol. 26, no. 3, pp. 487500, 2011.
3. T. Howlader and Y. P. Chaubey, Noise reduction of cDNA microarray images using complex wavelets, Image Process. Ieee Trans., vol. 19, no. 8, pp. 19531967, 2010.
4. N. Giannakeas, D. I. Fotiadis, and A. S. Politou, An automated method for gridding in microarray images, in Engineering in Medicine and Biology Society, 2006. EMBS06. 28th Annual International Conference of the IEEE, 2006, pp. 58765879.
5. L. Ying and C. Li, Based adaptive wavelet hidden Markov tree for microarray image enhancement, in Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008. The 2nd International Conference on, 2008, pp.314317.
6. A. Figueroa, P. S. Tsai, E. Bent, and R. Guo, Robust spots finding in microarray images with distortions, in Engineering in Medicine and Biology Society, 2008EMBS 2008. 30th Annual International Conference of the IEEE, 2008, pp. 13391342.
7. J. C. H. Hernandez, B. Duval, and J.-K. Hao, SVM-based local search for gene selection and classification of Microarray data, in Bioinformatics Research and Development, Springer, 2008, pp. 499 508.
8. C.-P. Lee, W.-S. Lin, Y.-M. Chen, and B.-J. Kuo, Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method, Expert Syst. Appl. Int. J., vol. 38, no. 5, pp. 46614667, 2011.
9. T. Jacobson, Bayesian Classification and Regression Tree Analysis (CART), 2010.
10. X. H. Wang, R. S. Istepanian, and Y. H. Song, Microarray image enhancement by denoising using stationary wavelet transform, NanobioscienceIeee Trans., vol. 2, no. 4, pp. 184189, 2003.
11. I. Guyon and A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res., vol. 3, pp. 11571182, 2003.
12. M. S. Mohamad, S. Omatu, S. Deris, and M. Yoshioka, An Iterative GASVM-Based Method: Gene Selection and Classification of Microarray Data, in Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living, Springer, 2009, pp. 187194.
13. E. Alba, J. Garcia-Nieto, L. Jourdan, and E. G. Talbi, Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms, in Evolutionary Computation, 2007. CEC 2007. IEEE Congress on, 2007, pp. 284290. M. S. Mohamad, S. Omatu, S. Deris, and M.
14. Yoshioka, Particle swarm optimization with a modified sigmoid function for gene selection from gene expression data, Artif. Life Robot., vol. 15, no. 1, pp. 2124, 2010.
15. L. X. Wang, A Course on Fuzzy Systems. Prentice-Hall press, USA, 1999.
16. J. S. R. Jang, ANFIS: Adaptive-network-based fuzzy inference system, Syst. Man Cybern. Ieee Trans., vol. 23, no. 3, pp. 665685, 1993.
17. Z. Wang, V. Palade, and Y. Xu, Neuro-fuzzy ensemble approach for microarray cancer gene expression data analysis, in Evolving Fuzzy Systems, 2006 International Symposium on, 2006, pp. 241246.
18. T. S. K. M. M. Hassan, Adaptive Neuro Fuzzy Inference System (ANFIS) For Fault Classification in the Transmission Lines, Online J. Electron. Electr. Eng. Ojeee Vol2no1 Vol, vol. 1.
19. J. Li and H. Liu, Kent Ridge Bio-medical Data Set Repository, 2002.
20. D. Bozdag, A. S. Kumar, and U. V. Catalyurek, Comparative analysis of biclustering algorithms, in Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology, 2010, pp. 265274.