International Journal of ChemTech Research
CODEN (USA): IJCRGG, ISSN: 0974-4290, ISSN(Online):2455-9555
Vol.10 No.4, pp 442-447, 2017

A Support Vector Machine Based Dynamic Clustering and Classification on Gene Expression Data

L.Sharmila[1]*, U.Sakthi[2] and Suresh Sagadevan[3]

[1]Department of Computer Science and Engineering, Sathyabama University, Chennai, India,
[2]Department of Computer Science and Engineering, St.Joseph's Institute of Technology, Chennai, India.
[3]Department of Physics, AMET University, Chennai, India.

Abstract : In this paper a novel method is proposed for classifying human genes using gene expression data taken from DNA microarray. This method functionally classifies the data and the data is taken from hybridization experiments. The novelty of the method is dynamically calls the feature vectors of the data and classify based on the theory of support vector machines. Since SVM is a regulated machine learning approach it take in the information and keep earlier learning about the quality capacities which thinks about the new approaching qualities utilizing similitude work. The problems faced by unsupervised learning methods like SOM (Self Organizing Map) and HCM (Hierarchical Clustering Method) and are overcome by SVM. Hence in this paper SVM method is used for training a small portion of the gene data and tests a major portion of the gene data whereas the test data is too dynamic. The experiment is carried out in MATLAB software and the results are verified.
Keywords : Gene Expression Data, Clustering and Classifying, Micro Array, Support Vector Machine, Machine Learning Approach.

Introduction

Presently to measure the translation levels of a creature's qualities at a specific moment of time, microarray quality expression studies are consistently utilized. These mRNA levels serve as a proxy for either the level of synthesis of proteins encoded by a gene or perhaps its involvement in a metabolic pathway. Differential term between a control organism and an experimental or diseased organism can thus emphasize genes whose function is related to the experimental challenge. Here, conservative diagnostic procedures involve morphological, clinical, and molecular studies of the tissue, which both are enormously subjective in their analysis and cause problem and humiliation to the patient Smaller scale cluster tests offer an option (or extra), target method for cell association through some encoded utilitarian of the quality expression levels for another tissue test of a new sort. By the "huge p, little n" issue; the factual heartiness of these strategies is still hampered while conceivably capable, a microarray slide can traditionally hold countless quality sections whose reactions here go about as the indicator variables(p), while the quantity of patient tissue tests (n)available in such studies is significantly less.

Related Approaches for Gene Clustering

In quality expression investigations, there are numerous cases of allegedly effective utilizations of both various leveled bunching and apportioning procedures. This area represents the differences of strategies which

have been utilized. Eisen et al[1] utilized agglomerative various leveled bunching with their un-focused relationship based contrast metric as portrayed above for development time-course microarray information from maturing yeast. This approach has since been followed in similar studies by Chu et al[2], Spellman et al[3], Iyer et al [4], Perou et al [5]and Nielsen et al[6]. Alternatively, Wen et al[7] used Euclidean various leveled gathering on vectors with the time arrangement of expression levels for each linked with the slants between them to consider balance yet comparative examples. SOMs have been taken care when swinging to non-show based apportioning techniques; Tamayo et al[8] utilized SOMs for bunching of various time arrangement of quality expression information. Comparable methodologies have likewise been utilized by Golub et al [9] for disease tissue class identification and expectation and Kasturi et al [10] for quality expression - time arrangement where the last first standardizes the information to permit the utilization of Kullback-Leibler dissimilarity as the separation metric. Tavazoie et al[11] represented expression time series in T dimensional space and used the k-means grouping algorithm.

For finding more delicate cluster structures, many model based variations have been developed beyond these generic methods. This has been particularly valuable with regards to time arrangement of quality expression tests. Ramoni et al [12] demonstrated quality expression time arrangement with autoregressive procedures, giving the going with free programming CAGED. Luan and Li [13] grouped quality expression time arrangement with blended impacts with B-splines; Bar-Joseph et al [14] utilized cubic splines for every quality with spline coefficients compelled to be comparable for qualities in a similar bunch. They likewise utilized a period distorting calculation to adjust time arrangement to comparable expression profiles in various stages. Utilizing a full MCMC Bayesian approach, Wakefield et al [15] performed grouping with a premise work representation for the expression time arrangement joining arbitrary impacts. Yeung et al[16] utilized the blend of normal appropriations programming MCLUST of Fraley and Raftery[17]for a scope of genuine and engineered quality expression information sets, at some point filed. Pan et al [18] utilized an indistinguishable model from MCLUST however on a two-specimen t-measurement of differential expression for every quality instead of the full quality expression information framework.

Medvedovic and Sivaganesan[19] utilized the Gibbs illustration strategies for Neal for Dirichlet prepare blend models to give a Bayesian rendition. Alon et al [20] utilized a divisive calculation iteratively appropriate two Gaussians at every phase with self-steady conditions. Heard et al [21] utilized a blend of Gaussian procedures with premise work representations for bunching of quality expression time arrangement, with a conjugate model expelling the requirement for MCMC. Graphical models have additionally been endeavored. Ben-Doret et al [22] gave two alternative graphical model-based clustering algorithms, clustering genes on a similarity matrix, PCC and CAST. Zhou et al [23] connected qualities with exceedingly related quality expression in a graphical model and grouped qualities through a briefest way investigation recognizing "transitive qualities." Dobra et al [24] endeavored to truly demonstrate the entire covariance structure of the qualities utilizing Gaussian graphical models.

Dynamic Clustering Approach

Considering SVM here due to it have manor mathematical features where it provides good analysis on GE, select a best similarity function, sparseness of solution for data scalability and can handle large set of feature space. SVM can also identify the outliers which provide high security. SVM identify the best set of genes where it is a common function using gene expression data. Finally SVM can predict the functional roles from GED. The test dataset is constantly new or has a place with the same dataset than the prepared dataset which are obscure to SVM.

For every quality X, the expression vector(X) $\vec{}$, the piece work K(X, Y) is utilized to gauge the likeness among qualities X and Y can be acquired utilizing the spot item from the information space

$$\mathbf{K(X, Y)} = \rightarrow \boldsymbol{Y} \rightarrow = \sum_{i=1}^{79} \boldsymbol{X_i Y_i}$$

In order to simplify technically 1 is added to the above kernel function as:

$$\mathbf{K(X, Y)} = \boldsymbol{X} \rightarrow \boldsymbol{Y} \rightarrow + \mathbf{1}$$

During the dot product on kernel, both feature space of X and are same and which is the input space of 79-dimension. Now SVM classifies after separating hyperplane in the space. To create a quadratic for separating the surface from the input space the above kernel function is squared as:

K(X, Y) = $(X \rightarrow Y \rightarrow +1)^2$

The separation includes the features of all pairs in mRNA gene expression interactions where $1 \leq i, j \leq 79$. If there should arise an occurrence of expanding the part as far as force of degree d, it can be characterized by

$(X \rightarrow Y \rightarrow +1)^d$.

In the component space of the SVM portion for all level of X a portion of the elements for all d-overlap collaborations between mRNA estimations, which are demonstrated by the terms type of X_i1,X_12,… ,X_id where 1≤i,j≤79. In out examination the level of the portions is characterized as d= 1, 2, and 3.

The radial basis kernel is experimented where it has a Gaussian form

$K(X, Y) = \exp\left(-\|X \rightarrow Y \rightarrow\|^2 / 2 \propto^2\right)$,

$\propto$ is the Gaussian width. To make the positive examples as closest to the nearest negative examples is set to equal to the median value of the Euclidean distance.

Experimental Design

In this paper the entire dataset is divided in to three groups. The classifiers are trained with 2/3 portion of the data and tested on the remaining data. Since more number of data is taken for training process the labeled and prior knowledge about the GED stored in SVM is high and leads to compare any gene expression data comes for testing dynamically. Also the learning (training) process is repeated three more times to handle different genes as test data. In the proposed SVM a spiral premise work portion is utilized to build the force of effectiveness contrasting and the other piece capacities exists in SVM. The SVM method and C4.5 methods are programmed in MATLAB software the results are taken in the form of numbers because this paper provides pure numerical analysis on the experimental results.

The data set is taken from MIPS Yeast Genome Database, where both training and testing data are included. The dataset has predefined classes of six such as:

- Tricarboxylic acid cycle (TCA) (targetmips~0),
- respiration (targetmips~1),
- cytoplasmic ribosomes (targetmips~2),
- proteasome (targetmips~3),
- histones (targetmips~4) and
- Helix-turn-helix proteins (targetmips~5)

The performance of the classifier is measured by computing the True positive (TP), True Negative (TN), False Positive (FP) and False Negatives (FN). The TP, TN, FP and FN are genuine class individuals, non-individuals; a part perceived as a non-part, non-part is delegated a part separately. As per the quantity of qualities what numbers of qualities are named the above classes utilizing SVM is said to be the execution of SVM. The proficiency of the SVM can likewise be figured as far as cost as:

Cost (SVM) = FP(SVM) + 2 * FN(SVM)

Where, FP(SVM) is the false positives obtained by SVM and FN(SVM) says the false negative of SVM. The result FN has more weight than FP due to the number of positive examples is very less comparing with the number of negatives. Finally the obtained cost values is compared with the Cost(N) which is the null learning procedure, and it classifies all the test data as negative. From this the cost can be saved using the learning procedure than comparison procedure SVM as:

S(SVM) = Cost(N) – Cost(SVM)

Results and Discussion

From the experiment it is noticed that the SVM learned and recognize the gene classes after training process in DNA microarray expression data. It can be compared with the non-SVM methods to notice the performance of the SVM. The SVM classifier is compared with the test results of the decision tree classifier C4.5 and it is given in Table-1. The performance of SVM is evaluated by comparing with the standard machine learning settings where all the methods must provide a positive or negative class label on the data after successful classification. Basically the class labels can be given only for trained data and it is compared with the test data for labeling process. In table-1 the performance of SVM with 1, 2 and 3 power based dot product of the radial function SVM, and Decision Tree learning (C4.5) methods. First column shows the class of gene expression data, second column says the methods and the other columns says the obtained false positive, false negative, true positive, and true negative rates with cost.

Table-1: Examination of error rates for different classification techniques

Class	Method	FP	FN	TP	TN	S(M)
TCA	D-p 1 SVM	18	5	12	2,432	6
	D-p 2 SVM	7	9	8	2,443	9
	D-p 3 SVM	4	9	8	2,446	12
	Radial SVM	5	9	8	2,445	11
	C4.5	7	17	0	2,443	−7
Resp	D-p 1 SVM	15	7	23	2,422	31
	D-p 2 SVM	*7*	*7*	*23*	*2,430*	*39*
	D-p 3 SVM	*6*	*8*	*22*	*2,431*	*38*
	Radial SVM	*5*	*11*	*19*	*2,432*	*33*
	C4.5	*18*	*17*	*13*	*2,419*	*8*

Table-2: Misclassification on Gene Expression Data

Methods	Real Data	Misclassified
SVM	20 genes	1 gene
C4.5	20 genes	2 gene

Functional Misclassification on Gene Data

In the experiment there are 20 data is taken from un-known and un-annotated data and fed into SVM and C4.5 classifiers. After some number experiment is carried out from the 20 genes 19 genes are classified correctly by SVM and 18 genes are classified by C4.5. This misclassification happens sometimes due to disagreement with MYGD reflection on various perspectives provided on the gene expression data. The misclassification results in shown in Table-2, and it is notices that SVM is better than C4.5 in terms of classification.

Functional Class Prediction on Gene Expression Data

SVM method is validated in terms of genes unknown functions to calculate the classification accuracy. To do this SVM is tested with un-annotated yeast genes in the experiment. ROF function is called by SVM dynamically to predict the class of the gene expression by overlapping or adjusting with the annotations of the adjacent class members. Since ORF function is utilized in the earlier research works on dsDNA and mRNA dataset for predicting the classes even ORF does not know the gene classes. The accuracy is not much more by

overlapping and adjusting annotated genes but combine with SVM classifier the accuracy in terms of prediction is improved because SVM has prior knowledge about the learned data set. Table-3 shows the un-annotated genes which are predicted as a class member after three or four times repeated the same process on different set of data. Finally the SVM agree that the gene expression data are near the indicated functional class members in the data space.

Table-3: Predicted functional classifications for previously un-annotated genes

Class	Gene	Locus	Comments
TCA	YHR188C		Conserved in worm, *Schizosaccharomycespombe*, human
	YKL039W	PTM1	Major transport facilitator family; likely integral membrane protein; similar YHL017w not co-regulated.
Resp	YKR016W		Not highly conserved, possible homolog in *S. pombe*
	YKR046C		No convincing homologs
	YPR020W	ATP20	Subsequently annotated: subunit of mitochondrial ATP synthase complex
	YLR248W	CLK1/RCK2	Cytoplasmic protein kinase of unknown function

Conclusion

In this paper, it is portrayed that the precision of SVM is exact as far as order on quality expression information. Here SVM is utilized to characterize the quality information as far as practical classes as indicated by the microarray structure. There are several experiments applied for predicting the classes in un-annotated data (yeast genes). Within the methods SVM utilizes higher dimensional kernel function for predicting the class and it is best for prediction. Taking in the information in higher dimensional way can give more data about the information since SVM can perform superior to anything different techniques like C4.5. Here SVM utilizes simple dot product functions on the features. From the experimental results given in Table-1, Table-2 and Table-3, it has been noticed that SVM is capable of classifying gene expression data even by utilizing other functions like ORF and other data where they can provide the feature information. In this paper the experiment is carried out only taking a sample data, small in size. In future SVM is evaluated by using number data in the experiment and compare it with the existing other traditional approaches.

References

1. Neal RM, “Markov chain sampling methods for Dirichlet process mixture models”, J Comput Graph Stat. 2000;9(2):249–265.
2. Chu S, DeRisi J, Eisen MB, et al. The transcriptional program of sporulation in budding yeast. Science. 1998;282(5389):699–705.
3. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression ProcNatlAcadSci USA. 1998;95(25):14863–14868.
4. Iyer VR, Eisen MB, Ross DT, et al. The transcriptional program in the response of human fibroblasts to serum. Science. 1999;283(5398):83–87.
5. Perou CM, Jeffrey SS, van de Rijn M, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. ProcNatlAcadSci USA. 1999;96(16):9212–9217
6. Nielsen TO, West RB, Linn SC, et al.Molecular characterization of soft tissue tumours: a gene expression study. Lancet. 2002;359(9314):1301–1307.
7. Wen X, Fuhrman S, Michaels GS, et al. Large-scale temporal gene expression mapping of central nervous system development. ProcNatlAcadSci USA. 1998;95(1):334–339.
8. Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. ProcNatlAcadSci USA. 1999;96(6):2907–2912.

9. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531–537.
10. Kasturi J, Acharya R, Ramanathan M. An information theoretic approach for analyzing temporal patterns of gene expression. Bioinformatics. 2003;19(4):449–458.
11. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999; 22(3):281– 285.
12. Ramoni MF, Sebastiani P, Kohane IS. Cluster analysis of gene expression dynamics. ProcNatlAcadSci USA. 2002;99(14):9121–9126.
13. Luan Y, Li H. Clustering of time-course gene expression data using a mixed-effects model with Bsplines. Bioinformatics. 2003;19(4):474–482.
14. Bar-Joseph Z, Gerber G, Gifford D, Jaakkola T, Simon I. A new approach to analyzing gene expression time series data. In: Proceedings of the 6th Annual International Conference on Research in Computational Molecular Biology (RECOMB ’02). Washington, DC; 2002:39–48.
15. Wakefield J, Zhou C, Self S. Modelling gene expression over time: curve clustering with informative prior distributions. In: Bernardo JM, Bayarri MJ, Berger JO, Heckerman D, Smith AFM, West M, eds. Proceedings of the 7th Valencia International Meeting. Vol 7 of Bayesian Statistics. New York, NY: The Clarendon Press, Oxford University Press; 2003:721– 732.
16. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. Bioinformatics. 2001; 17(10):977–987.
17. Fraley C, Raftery AE. Model-based clustering, discriminant analysis and density estimation. J AmStat Assoc. 2002;97(458):611–631.
18. Pan W, Lin J, Le CT. Model-based cluster analysis of microarray gene-expression data. Genome Biol. 2002;3(2):Research0009.
19. Medvedovic M, Sivaganesan S. Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics. 2002;18(9):1194–1206.
20. Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. ProcNatlAcadSci USA. 1999;96(12):6745–6750.
21. Heard NA, Holmes CC, Stephens DA. A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves. London, UK: Imperial College; 2004. Technical Report.
22. Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. J Comput Biol. 1999;6(3–):281– 297.
23. Zhou X, Kao MC, Wong WH. Transitive functional annotation by shortest-path analysis of gene expression data. ProcNatlAcadSci USA. 2002;99(20):12783–12788.
24. Dobra A, Hans C, JonesB, Nevins JR, YaoG, WestM. Sparse graphical models for exploring gene expression data. J Multivariate Anal. 2004;90(1):196–212.

*****