Hybrid Local Feature Selection In DNA Analysis Based Cancer Classification

Mrs.M.Akila
(student- CCE)
,Mr.S.Senthamarai kannan
M.E., Ph.D(Professor- CSE)
Sethu Institute of Technology, Virudhunagar, Tamil Nadu.
e-mail: akianush@gmail.com,stanfordssk@gmail.com

ABSTRACT

Feature selection, as a preprocessing step to machine learning, is effective in reducing dimensionality, removing irrelevant data and increasing learning accuracy. The development of microarray dataset technology has supplied a large volume of data to many fields. In particular, it has been applied to prediction and diagnosis of cancer, so that it helps us to exactly predict and diagnose cancer. To precisely classify cancer we have to select genes related to cancer. The challenging task in cancer diagnosis is how to identify salient expression genes from thousands of genes in microarray data because extracted genes from microarray dataset have many unwanted datas not related to cancer. In this project we attempt explore a novel hybrid wrapper and filter feature selection algorithm for classification problem using a memetic framework i.e., a combination of genetic algorithm (GA) and local search (LS) has been proposed.. The LS is performed using correlation based filter methods are discritize, ranking and redundancy elimination with symmetrical uncertainty (SU) measure .using this hybrid method we can able find cancer related gene, From the larger amount of gene datas .using that smaller dataset doctors can able to find the affected gene and provide better treatment. The efficiency and the effectiveness of the method are demonstrated through extensive comparisons with other methods using real-world datasets of high dimentionality

Keywords: Feature Selection, Memetic Algorithm, Filters ,wrappers, genetic algorithm, symmetrical uncertainty.

I. INTRODUCTION:

Cancer classification, which can help to improve health care of patients and the quality of life of individuals, is essential for cancer diagnosis and drug discovery. An accurate prediction of cancer has great value in providing better treatment and response to therapy varying from different aspects . However, traditional diagnostic methods are mainly based on the morphological and clinical appearance of cancer. They have limited contributions because cancers usually result from many environmental factors, and even the same tumor may have different symptoms under different conditions. Thus, it is necessary to inject systemic approaches into the problem of cancer diagnosis and prediction. From the bio-medical perspective, each kind of disease is associated with certain genes in tissues and the mutation of genes may give rise to the occurrence of certain diseases. Fortunately, the advent of DNA microarray technique, which allows simultaneously measure the expression levels of thousands of genes in a single experiment , makes the accurate prediction of cancer possible and easier. Since it is capable of comparing the gene expression levels in tissues under different conditions, the microarray technique may bring many advantages to cancer prediction and make the diagnosis result more objective, accurate and reliable. During past years, this method has drawn a great deal of attention from both biological and engineering fields

2.Discretization

The discretization is the process of dividing the continuous data into a discrete one, with help of information therory using **MDL algorithm** based on minimum description length principle.

MDL Algorithm for Discretization

MDLM(D)
(1) $MDL = \infty$
(2) For all feature subsets L
1.1 Compute $Length_L = \sum_{i=1}^{i=q} \frac{p_i}{2} \log \frac{|D_L(i)|}{|D_L|} + h_L$
where $h_L = \frac{1}{2}(N - M)(N + M + 3) \log P + \sum_{i=1}^{i=q} M(M + 3) \log P_i$,
N – total number of features,
M – number of features in the candidate subset,
P – total number of instances in D,
P_i – number of instances with class label i,
q – total number of class labels,
D_L – covariance matrix formed from all the useful feature vectors,
$D_L(i)$ – covariance matrix formed from the useful feature vectors,
of class i,
$|.|$ – denotes determinant.
1.2 If $Length_L < MDL$ then
$T = L, MDL = Length_L$
(3) Return T

3.RANKING

Applying ranking filter we can obtain a number of top ranked genes. The ranking is done with help of SU which is based on **Information theory.** information in the theory is known as **entropy**,

3.1 Entropy

The **entropy**, *H*, of a discrete random variable *X* is a measure of the amount of *uncertainty* associated with the value of *X*.

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i) \qquad (1)$$

3.2 Cross entropy

The **cross entropy** between two probability distributions measures the average number of bits needed to identify an event from a set of possibilities, if a coding scheme is used based on a given probability distribution *q*, rather than the "true" distribution *p*.

$$\mathrm{H}(p,q) = -\sum_{x} p(x) \log q(x).$$

3.3 Mutual information (transformation)

Mutual information measures the amount of information that can be obtained about one random variable by observing another. The mutual information of *X* relative to *Y* is given by:

$$I(X,Y) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)\,p(y)} \qquad (3)$$

When one variable becomes completely redundant with the knowledge of the other. Another symmetrical measure is the *symmetric uncertainty* [19], given by

$$SU(X,Y) = 2\left[\frac{I(X|Y)}{H(X)+H(Y)}\right] \qquad (4)$$

II. SYSTEM DESIGN

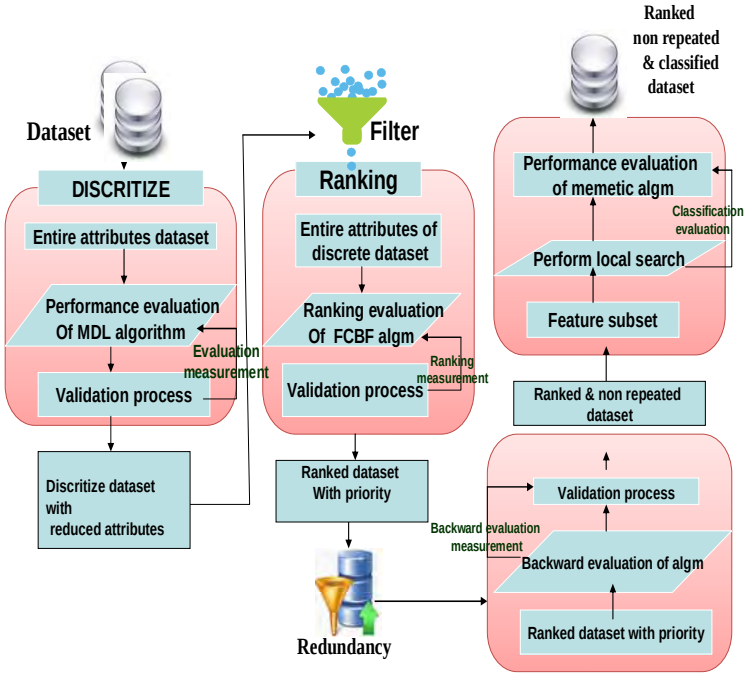

Fig1:system design

4. Redundancy Elimination

Applying redundancy filter to eliminate repetitive gene in the data set. FCBF is an efficient and fast algorithm which uses interdependence of features together with the dependence to the class. FCBF achieves this goal by giving every feature a temporary predominance in the elimination process and making them start eliminating features from the features which are least correlated with the class

FCBF Algorithm

Input: S ($F1$, F2,…, FN, C) // a training data set

δ // a predefined threshold

Output: S_{best} // an optimal subset

1 begin

2 Discretize using MDL method

3 for i = 1 to N do begin

4 calculate SUi, c for Fi;

5 if ($SUi,c \geq \delta$)

6 append Fi to S'_{list} ;

7 end;

8 order S'_{list} in descending *SUi, c* value;

9 $Fp = getFirstElement\ (S'_{list}\ )$;

10 do begin

11 $Fq = getNextElement\ (S'_{list}, Fp)$;

12 if (Fq <> NULL)

13 do begin

14 $F'\ q = Fq$;

15 if ($SUp, q \geq SUq, c$)

16 remove Fq from S'_{list};

17 $Fq = getNextElement\ (S'_{list,}\ F'\ q)$;

18 else $Fq = getNextElement\ (S'_{list,}\ Fq)$;

19 end until (Fq == NULL);

20 $Fp = getNextElement\ (S'_{list,}Fp)$;

21 end until (Fp == NULL)

22 $S_{best} = S'_{list}$;

23 end;

5. Classification:

the proposed Correlation based Memetic feature selection Algorithm (MA-C) for classification problems which is depicted in the GA population is randomly initialized with each chromosome encoding a candidate feature subset. Subsequently, a local search (LS) is performed. The LS is performed on all gene or portion of the gene, to reach a local optimal solution or to improve the feature subset.

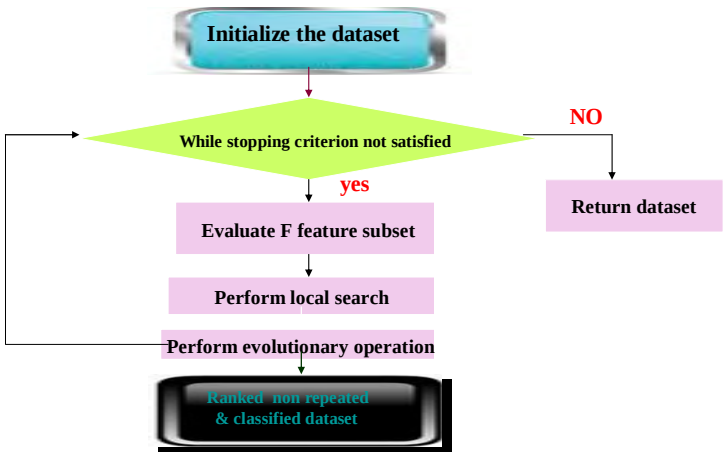

Fig2:memetic search

Table 1: microarray datasets used in the experiment

Dataset	No of genes	No of Samples	No of classes	Max no of genes
CNS	7129	60	2	100
Breast	24481	97	4	100
lung	7129	96	2	100
MLL	10289	102	2	100
LEUKEMIA	7219	72	2	75

Table 2: no of instances after global search and memetic search

Dataset	No of genes	After GA	After MA-C
CNS	7120	24	22
Breast	2309	24	23
lung	12601	27	25
MLL	12583	19	19
LEUKIMIA	7130	24	21

CONCLUSION

The goal is to improve classification performance and to accelerate the search to identify important feature subsets. In particular, the filter method fine-tunes the population of GA solutions by adding or deleting features based on SU measure. Hence, our focus here is on filter methods that are able to assess the goodness of the individual features. Empirical study of MA-C on several commonly used datasets from the UCI repository , indicates that it outperforms recent existing methods in the literature in terms of classification accuracy, selected feature size and efficiency. Further, we also investigate the balance between local and genetic search to maximize the search quality and efficiency find the correct gene which is affected. So Implementing this future Enhancement may take a considerable amount of time. Implementing Causes a Very effective system and an error free data pattern.

REFERENCES

[1] A. L. Blum and P. Langley(1997) “Selection of relevant features and example machine learning” , Artificial Intelligence, 97:245–271
[2] D. A. Bell and H. Wang. (2000), “A formalism for relevance and its application in feature subset selection”, Machine Learning, 41(2):175–195..
[3] D. D. Jensen and P. R. Cohen (2000), “Multiple comparisons in induction algorithms”, Machine Learning, 38(3):309–338.
[4] G. H. John, R. Kohavi, and K. Pfleger (1994),” Irrelevant feature and the subset selection problem” ,In Proceedings of the Eleventh International Conference on Machine Learning, pages 121–129
[5] Huawen Liu a, Lei Liu a,Ã, Huijie Zhang b (2010) “Ensemble gene selection for cancer classification” Pattern Recognition 43 2763–2772 on machine learning
[6] Hongxing He, Huidong Jin and Jie Chen,(2005) ,”Automatic Feature Selection for Classification of Health Data”, AI 2005: Advances in Artificial Intelligence – Springe

[7] I. Guyon and A. Elisseeff.(2003) “An introduction to variable and feature selection”, Journal of Machine Learning Research, 3:1157–1182
[8] K.-J. Kim, S.-B. Cho, (2008)” An evolutionary algorithm approach to optimal ensemble classifier for DNA micro array data analysis”, IEEE Transactions on Evolutionary Computation 12 (3) 377–388
[9] L-Y. Yeh, (2008) “ Applying data mining techniques for cancer classification on gene expression data”, Cybernetics and Systems: An International Journal 39
[10] Lei yu & Huan Liu,(2004), “Efficient Feature Selection via Analysis of Relevance and redundancy”, Journal of machine learning Research 5 , 1205-1224
[11] Lei Yu, Baris Senliol , Gokhan Gulgezen and Zehra Cataltepe,(2008),” Fast Correlation Based Filter(FCBF) with a Different Search Strategy”, ISCIS ’08.
[12] M. A. Hall (2000).,”Correlation-based feature selection for discrete and numeric class machine learning”, In Proceedings of the Seventeenth International Conference on Machine Learning, pages 359–366
[13] Muhammad Atif Tahir *, Jim Smith” (2010),”Creating diverse nearest-neighbour ensemble using simultaneous metaheuristic feature selection”, Science direct Pattern Recognition Letters 31 1470–1480
[14] R. Kohavi and G. H. John (1997)” Wrappers for feature subset selection**”**, Artificial Intelligence, 97(1-2): 273–324.
[15] Ranjit Abraham, Jay B. Simha and S. Sitharama Iyengar (2008), “Effective Discretization and Hybrid feature selection using Naïve Bayesian classifier for Medical data mining” , International Journal of Computational Intelligence Research.
[16] S.Senthamarai kannan,N.Ramaraj,G.Mainkandan 2009,”A novel hybrid feature selection via correlation based memetic search algorithm”a book on intelligent information management system and technologies
[17] S.Senthamarai Kannan, N.Ramaraj 2007, “Concept based Clustering for Descriptive Document Classification”, The CODATA Data Science Journal, volume 6, pages 91-98.
[18] U.M. Fayyad, K.B Irani,(1993), “Multi-interval discretization of continuous-valued attributes for classification learning” , In Proceedings of the 13th International Joint Conference on Artificial Intelligence, pp. 1022–1027.
[19] Witten, Ian H. & Frank, Eibe (2005), “Data Mining: Practical Machine Learning Tools and Techniques”,Morgan Kaufmann, Amster - dam
[20] Y. S. Ong and A. J. Keane 2004, "Meta-Lamarckian in Memetic Algorithm," IEEE Trans. Evolutionary Computation, vol. 8, no. 2, pp. 99-110.
[21] Zili Zhang,Pengyi yang(2006),”An ensemble of classifier with genetic algorithm based feature selection”,IEEE Transactions on Intelligent information

AUTHORS

M.AKILA completed B.Tech information technology in vickram college of engineering,India.now doing M.E computer and communication in sethu institute of technology India.her area of interest are datamining,networks and softwareengineering

DR.S.SENTHAMARAI KANNAN working as professor and head of the department of m.e computer science in sethu institute of technology.he has published more than twenty five papers in both nation and international journals.his area of interest are datamining e_learning networks software engineering.he is guiding many research scholars and presented papers in many conferences