

Enhanced Classification of Microarray Data Using Artificial Intelligence Searching Techniques

Sreehitha Anakala, Raja Sowmya Maganti , Brindha.G.R
School of Computing, SASTRA Deemed University, India

Abstract:

The objective of the study is to process the high dimensional dataset through feature selection techniques and in turn improve the classification accuracy using supervised classifier. The performance of classifying gene expression is to be enhanced using artificial intelligence searching algorithms like hill climbing and best first search. Hill climbing search is an optimization algorithm to find local optimum. Best first search is used to find the cheapest optimum path. The nodes in the path are considered as genes in micro array. The classifier which is to be used in this study is k-Nearest Neighbour. The performance of the classifier is proved using measurements such as accuracy, sensitivity, specificity and compared with existing studies; so that the robustness of the proposed method is proved.

Keywords: Feature selection, Hill Climbing Search, Best First Search, k-Nearest Neighbour, Microarray classification

I. Introduction:

The preprocessing of microarray data classification deals with extraction of significant genes through utilization of machine learning and data preprocessing techniques. Since many studies show that majority of genes don't contain relevant data for classification task, data preprocessing is an important effort to achieve accurate, reliable and efficient performance during classification. Gene Selection is one of the common techniques used for preprocessing of microarray data.

Gene selection methods are generally classified into four categories which include filter, wrapper, hybrid and embedded approaches. Filter approach ranks the features based on relevance of genes independent of learning algorithms. Wrappers utilize the learning method to evaluate the subset of genes repeatedly. Hybrid approach is a combination of both filter and wrapper methods. Firstly a subset of genes is selected through filter approach and the final gene set is selected through wrapper approach and vice-versa. In Embedded approach, initial gene set used to train and get model based on rank evaluation of genes.

Up to now, different strategies have been proposed to measure the relevance of genes which includes MGSACO(Microarray gene Selection based on ant colony optimization) which is based on unsupervised learning model. A subset of genes with maximum relevance and minimum redundancy is selected through MGSCAO. Though the probability distribution changes by

iteration and also sequences of decisions are not independent, it was able to select a subset of genes with minimum redundancy and maximum relevance using new fitness functions [1].

Greedy algorithms are often used to obtain optimal and quick approximate solutions to many complex optimization problems [2]. The solutions provided by this algorithm are not optimal always but are close to optimal. The main idea of greedy algorithm is to solve the problem through several steps [3]. The most attractive action at each step is taken without worrying too much about the effect of the step on the global picture. Using Particle swarm optimization, the classification performance is achieved in short computation time and less number of features, but it may get stuck in the local optimum and because of iterative process it may have low convergence rate [4]. High dimensionality is reduced using Ant Colony Optimization, which has a notable ability to generate a gene subset with the smallest size and salient features while yielding high classification accuracy [5]. Genetic algorithms are adaptive heuristic search algorithms that lead to significant feature reduction although randomized; they are not random [6]. They exploit the historical information and direct the search into a region of better performance. The main disadvantage of this algorithm is degeneracy, when many chromosomes represent the same solution.

II. Proposed Method

In this paper, feature selection techniques like hill climbing and best first search are used to improve the classification accuracy using supervised classifier k-NN (Fig 1).

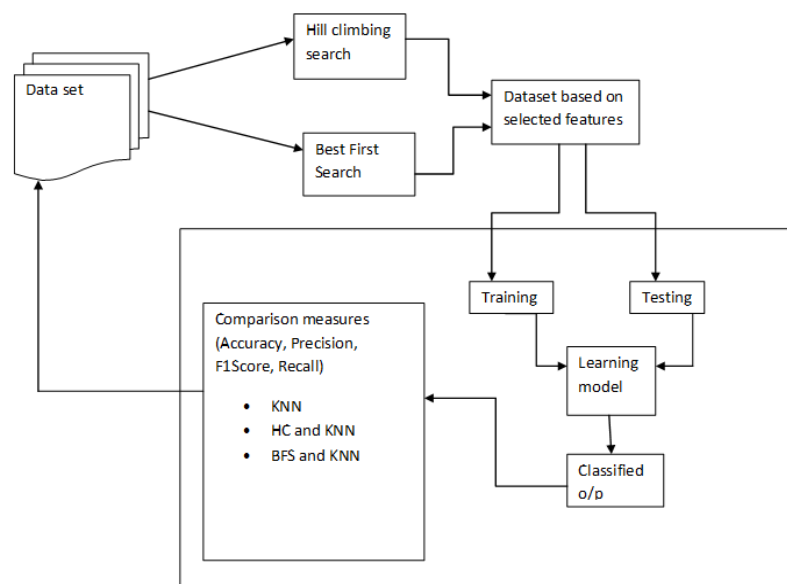


Fig 1 Proposed Work Flow

2.1 Hill Climbing Search

Hill climbing search is an optimization algorithm to find the local optimum. It is a heuristic algorithm that starts with a random solution, then it tries to find a better solution by changing a single element of the solution incrementally. If that change produces a better solution than the previous one, an incremental change is made to the new solution, repeating until no furthermore improvements can be found. It also helps in solving pure optimization problems where the objective is to find the best state according to the objective function.

Algorithm for hill climbing Search

function HILL-CLIMBING(*problem*) **return** a state that is a local maximum

input: *problem*, a problem

local variables: *current*, a node.

neighbor, a node.

current $\leftarrow$ MAKE-NODE(INITIAL-STATE[*problem*])

loop do

neighbor $\leftarrow$ a highest valued successor of *current*

if VALUE [*neighbor*] $\leq$ VALUE[*current*] **then return** STATE[*current*]

current $\leftarrow$ *neighbor*

2.2 Correlation Based Feature Selection

The main problem in machine learning is to find a set of features from which to construct a classification model for particular task. So, feature selection issue for machine learning is addressed through a correlation based approach. Good feature sets are those that contain features that are highly correlated with the class, yet unrelated with each other. CFS (Correlation based Feature Selection) is an algorithm that couples the evaluation formula with an appropriate correlation measure and a heuristic search strategy. The evaluation function biases towards subsets that contain functions that are highly correlated with the class and totally uncorrelated with each other. The feature acceptance depends upon the extent to which it predicts classes in areas of instance spaces that are not already predicted by other features.

2.3 Best First Search

BFS is an algorithm in which a feature (node) is selected for expansion based on evaluation function. The node with lowest evaluation is selected for expansion. First, the data is preprocessed by converting into discrete values and passed to CFS. Using symmetric uncertainty measures, CFS calculates the feature-class, correlation between features and then it searches the feature subset space using best first search. The subset with the highest rank found during the

search process is used to reduce the dimensionality of both the original training data and the testing data. These reduced datasets are then passed to machine learning scheme for training and testing. The training and testing data contain only the features selected by CFS. The dimensionally reduced data can then be passed to a KNN algorithm for classification and prediction.

The proposed BFS-CFS-KNN algorithm:

- 1: To start with OPEN list containing the start state, the CLOSED list empty and $BEST \leftarrow$ start state.
- 2: Let assign $s = \arg \max e(x)$ (get the state from OPEN with the highest evaluation).
- 3: Eliminate s from OPEN and add to CLOSED.
- 4: If $e(s) \geq e(BEST)$, then $BEST \leftarrow s$.
- 5: For every child t of s that is not in the OPEN or CLOSED list, evaluate and add to OPEN.
- 6: If BEST changed in the last set of expansions, go to 2
- 7: Return BEST.
- 8: Obtain the new data set
- 9: Construct both training and test data discrete.
- 10: Classify (X, Y, x) //X:training data, Y: class labels of X, x:unknown sample
- 11: for $i=1$ to m do

Compute distance $d(X_i, x)$
- 12: Compute set I containing indices for the k smallest distances $d(X_i, x)$

Return majority label for $\{ Y_i \text{ where } i \in I \}$

The first half of the algorithm from step one to eight is used to select the subset using Best First Search and then the second half of the algorithm from nine to twelve are for classification using KNN.

III. Results and Discussion

The proposed method has been tested by extracting the required features from three datasets golub, breast cancer and alonds (Fig 2). The classification is done to identify the class label and we evaluate the experimental results with precision and recall. These two classic measurements are utilized to measure performance of the classification. Accuracy is another criterion to measure the overall performance of the classification. These measures are calculated based on

four possible outcomes – true positive(TP), true negative(TN), false positive(FP), false negative (FN) of the confusion matrix.

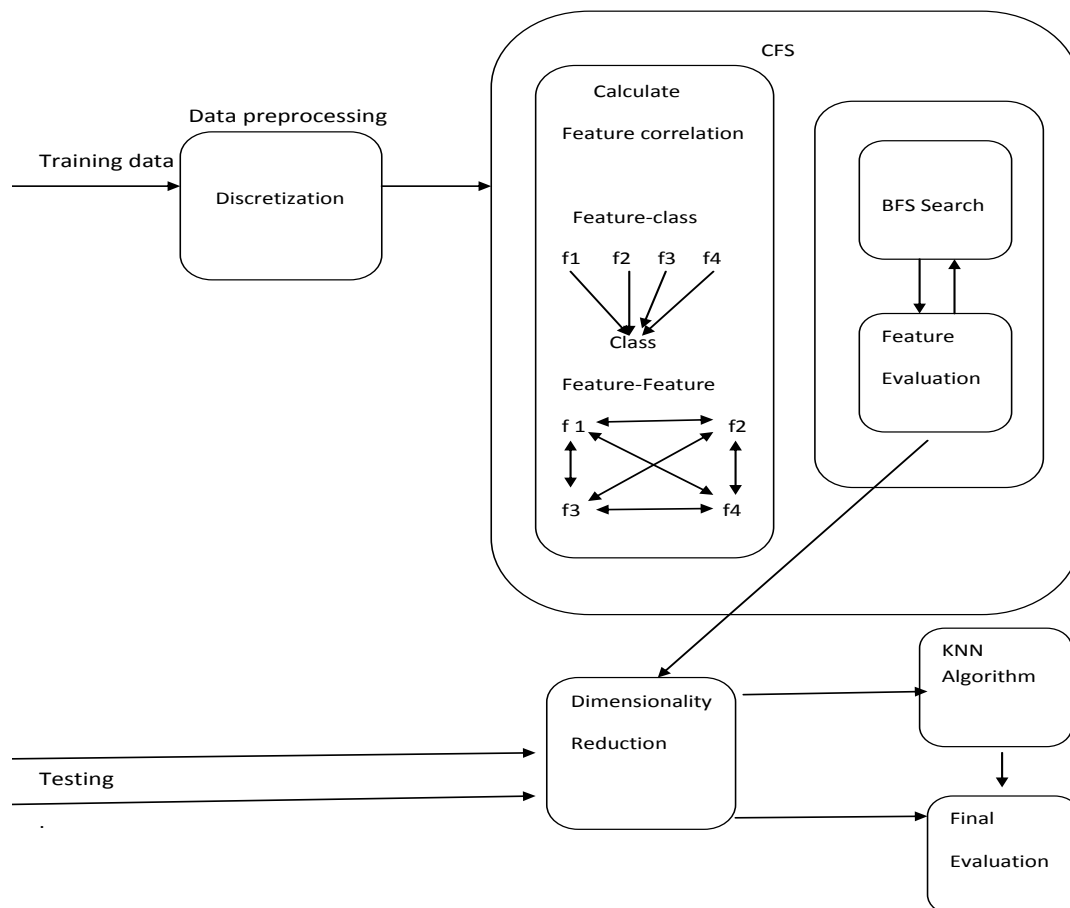


Fig 2. Proposed Model for BFS based CFS and KNN Algorithm

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

GOLUB:

Starting with 3051 genes ,when the proposed algorithm HCS(Hill climbing search) is applied ,the accuracy obtained is 92.8.When the proposed method is repeatedly applied, the accuracy is increased to 95.8 at 61 genes. When CFS (Correlation based feature selection) is applied at 723 the accuracy obtained is 100 which is the highest possible accuracy (Table 1.a).

Table 1.a Accuracy Comparison for GOLUB

Number of genes	Accuracy after applying HCS	No of genes obtained after CFS	Accuracy after applying CFS
3051	92.8	-	-
1530	93.2	-	-
723	93.2	11	100
346	93.0	8	99.3
177	94.4	10	90.1
99	95.2	3	89.1
61	95.8	6	95.8

ALONDS:**Table 1.b Accuracy Comparison for ALONDS**

Number of genes	Accuracy after applying HCS	No of genes obtained after CFS	Accuracy after applying CFS
2000	71.3	-	-
966	68.3	15	87.9
470	71.4	12	84.2
227	69.7	7	87.2
117	69.2	5	81.3
56	74.0	5	81.3

Here another dataset (ALONDS) with 2000 genes is considered (Table1.b). The proposed method CFS(Correlation based feature selection) is applied and highest accuracy(87.9) is obtained at 966 genes and when HCS (Hill Climbing Search) is repeatedly applied the highest accuracy(74.0) is obtained at 56 genes.

BREAST CANCER:**Table 1.c Accuracy Comparison for BREAST CANCER**

Number of genes	Accuracy after applying HCS	No of genes obtained after CFS	Accuracy after applying CFS
1213	88.7	-	-
580	89.8	17	92.3
287	89.5	32	84.6
136	85.5	28	86.0
68	83.7	15	82.7
31	87.9	5	77.9
13	81.4	5	70.4

The proposed methods are applied on another dataset (BREAST CANCER) with varying genes(1213). Here for this particular dataset the highest accuracies after applying the methods HCS(89.8) and CFS(92.3) are obtained at 580 genes (Table 1.c).

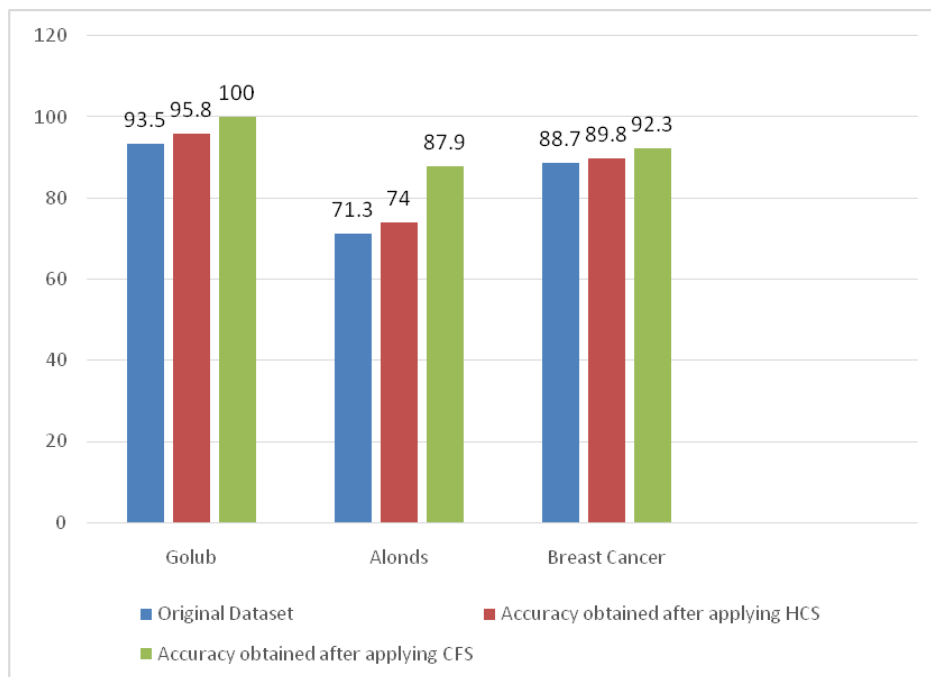


Fig 3. Accuracy Comparison of all data sets

Recall and Precision:

The fraction of relevant instances to retrieved instances is called precision where the fraction of relevant instances to total number of instances is called Recall. F1 Score is a measure of test's accuracy and is the simple harmonic mean of precision and recall. Here when HCS is applied, the performance measures precision, recall and f1 score are 100 for golub and 83.3 for alonds where as breast cancer gives different values, recall(100), precision(93.5) and f1 score(96.64).when CFS is applied recall is 100 for all datasets and precision is 100 for golub and 83.3 for alonds and 93.5 for breast cancer ,f1 score is highest for golub(100) followed by breast cancer (96.64) and alonds (90.88).

Table 2 a Performance measures comparison for GOLUB

GOLUB	RECALL	PRECISION	F1-Score
ORIGINAL	100	100	100
HCS	100	100	100
CFS	100	100	100

Table 2 b Performance measures comparison for ALONDS

ALONDS	RECALL	PRECISION	F1-Score
Original	80	100	88.88
HCS	83.3	83.3	83.3
CFS	100	83.3	90.88

Table 2 c Performance measures comparison for BREAST CANCER

BREAST CANCER	RECALL	PRECISION	F1-Score
Original	93.75	82.3	87.65
HCS	100	93.5	96.64
CFS	100	93.5	96.64

CONCLUSION:

In this paper, the artificial intelligence algorithms such as hill climbing search and best first search are used for feature selection. The computational efficiency of this methods and the goodness of KNN classifier were combined to improve the efficiency. The performance of the proposed methods was examined on three datasets using kNN. The experimental results show that the proposed methods HCS and BFS are able to select a subset of genes with minimum redundancy and maximum relevance. Through the intelligence searching methods significant genes are selected from 1213 to just 13 features, so that even the lazy classifier kNN can be used for classification.

REFERERNCES:

- [1] SinaTabakhi , Ali Najafi , RezaRanjbar , Parham Moradi , Gene selection for microarray data classification using a novel ant colony optimization, *Neurocomputing*, Volume 168 Issue C, November 2015,pp.1024-1036
- [2] J.Kruskal, “Greedy algorithm for the minimum spanning tree problem”, *Proceedings of the American Mathematical Society*, 1956, pp. 48-50.
- [3]K. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall, Englewood Cliffs, NJ, 1982.
- [4]Y.Saeys,I.Inza,P.Larrañaga,Areviewoffeatureselectiontechniquesinbioinformatics, *Bioinformatics*23(2007)2507–2517.
- [5] Y. Li, G. Wang , H. Chen, L. Shi, L. Qin, An ant colony optimization based dimension reduction method for high -dimensional datasets, *J. Bionic Eng.* 10 (2013)231–241.
- [6]J.J. Liu, G. Cutler, W. Li, Z. Pan, S. Peng, T. Hoey, L. Chen, X.B. Ling, Multiclasscancer classification and biomarker discovery using GA-based algorithms,*Bioinformatics* 21 (2005) 2691–2697.

