A Parallel Multi Objective Optimization Genetic Algorithm Gene Feature Selection on Microarray Based Cancer Classification Using Neuro-Fuzzy Inference System

A.Natarajan [1] , Dr.R.Balasubramanian [2]
[1] Asst.Prof, Department of Information Technology
Jayaraj Annapackiam CSI College of Engg, Nazareth, Tamilnadu
[2] Professor and Head, Department of Computer Science and Engineering,
M.S University, Tirunelveli, Tamilnadu.

Abstract: Feature selection has played a very important role in the field of data mining and machine learning. The high performance parallel and distributed computing is used for gene expression analysis and finding the thousands of genes simultaneously. The classification and validation of molecular biomarkers for cancer diagnosis is an important problem in cancer genomics. The microarray data analysis is used for extracting the biologically useful data from the huge amount of expression data to know the current state of the cell. Most cellular processes are regulated by changes in gene expression. This is a great challenge for computational biologists who see in this new technology the opportunity to discover interactions between genes. In this research we propose a Parallel Multi Objective Optimization Genetic Algorithm for Gene Feature Selection and the best features are evaluated by Adaptive Neuro Fuzzy Inference System classifier. More importantly, the method can exhibit the inherent classification difficulty with respect to different gene expression datasets, indicating the inherent biology of specific cancers.

Keywords:- Microarray, Gene Expression, Parallel Multi Objective Optimization GA, Gene Feature selection, ANFIS.

1.INTRODUCTION

1.1Microarray Data

DNA microarray experiments are used for cancer classification and prediction. The microarray technology has been used in many biomedical researches. The one major problem in applying gene expression summary to cancer classification is that number of features distinguishes the number of genes. Some important cancer informatics studies have exposed that the small collection of genes selected by feature selection methods can give the good classification results. Feature Selection [14] (FS) is a very important task in machine learning for cancer classification with the goal of identifying very important features subsets in a microarray data .It is one of the most key problems in the field of data mining and bioinformatics[13]. The classification and validation of molecular biomarkers for cancer diagnosis is an important problem in cancer genomics. The selection of applicant genes is very essential to identify accurately the origin of cancer and treatment. With the appearance and fast development of DNA microarray technologies, making gene expression profiles for different cancer types has already become a hopeful means for cancer classification.

Genetic algorithms (GAs) [5], a form of inductive learning strategies are adaptive search techniques initially introduced by Holland

(Holland, 1975). Genetic Algorithms are inspired from Darwin's theory of evolution. By simulating nature evolution and emulating biological selection and reproduction techniques, the GA can solve complex problems in a strong search domain. The algorithm starts with a set of randomly generated solutions called *population*. The population size remains constant throughout the genetic algorithm. At each iteration the populations are evaluated based on their fitness quality with respect to the given application domain to form new solutions called *offspring* which retains many features of their parents. Offsprings are formed by two main genetic algorithm operators such as crossover and mutation. Crossover operates by randomly selecting a point in the two selected parent gene structures and exchanging the remaining segments of the parents to create new offspring. Therefore, crossover combines the features of two individuals to create two similar offsprings. Mutation operates by randomly changing one or more components of a selected individual. It acts as a population perturbation operator and is a means for inserting new information into the population. This operator prevents any stagnation that might occur during the search process.

In this proposed work, Parallel Multi Objective Optimization Genetic Algorithm feature selection is implemented based on multi objective optimization genetic algorithm operators and function. This method uses two different algorithms such as contribution and entropy to find the pareto optimal solutions for ranking. The best solutions are evaluated based on Adaptive Neuro Fuzzy Inference System (ANFIS) classifier...The System model is shown below fig 1.

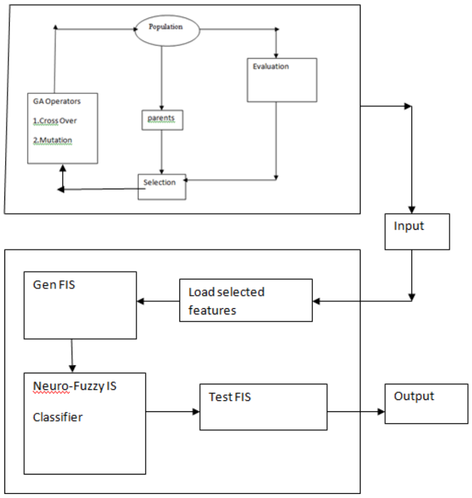


Figure 1.System Model

1.2 Related Work

Soumen Kumar Patel, at el [12] has explained a novel feature selection method which was based on Multiobjective Genetic Algorithm using rough set theory. This method proposed to choose important informative gene set, which classify the cancer dataset very efficiently. This method has used two fitness functions individually based on the concepts of strong mathematics such as rough set theory and probability theory. The lack of diversity of population is overcome by jumping gene mutation. The only drawback of this method is that the population size can be set within the range 100 to 1000 only.

Asha Gowda Karegowda [2] has proposed a wrapper approach with genetic algorithm for generation of subset of attributes with different classifiers such as Naïve Bayes, Bayes Networks,C4.5 and Radial basis functions. The above classifiers are experimented on the Diabetes datasets, Breast cancer datasets , Heart Statlog and Wisconsin Breast cancer. The main disadvantage of this approach is that the computing time is very high for the large datasets.

In 2009, QingzhongLiu, Andrew H. Sung [17] has proposed a new feature selection method called Recursive Feature Addition method on microarray based breast cancer data. The RFA gene feature selection method provides good classification accuracy than the other methods. In this method, serial programming is used for classification which slows down the computational speed.

2. PROPOSED METHOD

2.1 Feature Selection Using Parallel Multi Objective Optimization Genetic Algorithm

The multi-objective problem requires adaptation of optimization methods[10]. The main difference with mono-objective optimization problems is that in multi-objective problems, there is not a single solution for which all criteria are optimal but a set of solutions for which there are no other solutions better for all the criteria. These solutions are called Pareto-optimal. The notion of Pareto-optimality is defined in terms of dominance.

We propose to use a GA to find all the Pareto-optimal solutions which are all interesting potential rules. They are located on a boundary known as the Pareto-front. We would like the solutions to cover the Pareto-front as well as possible to obtain a good representation of this front. This "A priori" approach offers multiple solutions to the decision maker, which can select the solution that is best suited according to nonformal additional criteria, without requiring additional searches.

Multi-objective optimization, solutions quality can be assessed in different ways. Some approaches compare the obtained front with the optimal Pareto front .Others approaches evaluate a front with a reference point [18]. Some performance measures do not use any reference point or front to evaluate an algorithm [19, 20],especially when the optimal Pareto front is not known at all. Here, we have to compare different versions of the proposed model, without knowing the true Pareto front. We propose to use two complementary types of performance indicators that allow to compare two by two Pareto fronts obtained by different algorithms: the contribution and the entropy [21]. The contribution indicator quantifies the domination between two sets of nondominated solutions. The entropy indicator gives an idea about the diversity of the solutions found. This GA has been implemented by using parallelization tool like Open MPI [11].

2.2 Contribution

The contribution of PO_1 relative to PO_2 is roughly the ratio of non dominated solutions produced by PO_1. This way, let C be the set of solutions in $PO_1 \cap$ PO2. Let W_1 (resp. W_2) be the set of solutions in PO_1 (resp. PO_2) that dominate solutions of PO_2 (respectively in PO_1). Similarly, let L_1 (respectively L_2)be the set of solutions in PO_1 (resp. PO_2)that are dominated by solutions of PO_2 (resp PO_1). The set of solutions in PO_1 (respectively PO_2) that are comparable to solutions in PO_2 (respectively PO_1) is $N_1 = PO_1 \setminus (C \cup W_1 \cup L_1)$ (respectively $N_2 = PO_2 \setminus (C \cup W_2 \cup L_2)$)

Parallel Operator (PO): The search operators of the method are run in parallel.

This way the contribution is stated as

$$\mathrm{CONT}(PO_1 / PO_2) = \frac{|C|/2 + |W_1| + |W_2|}{|C| + |W_1| + |N_1| + |W_2| + |N_2|} \longrightarrow (1)$$

2.3 Entropy

Let PO_1 and PO_2 be the Pareto fronts of a given Multi Objective Optimizatio Problem respectively calculated with two algorithms **A** and B, and let be $PO = ND\ (PO_1 \cup PO_2)$, with $N\ D$ representing the nondominated set. Then, the N-dimensional space, where N is the number of objective functions to optimize, is clustered. For each space unit with at least one element of PO, the number of present solutions of PO_1 is calculated. This way, the relative entropy, $E(PO_1, PO_2)$ of a set of non dominated solutions PO_1regarding to the Pareto frontier PO is defined as

$$E\ (PO_1, PO_2) = -1/\log(\textstyle\sum_{i=1}^{c} \quad (\frac{ni}{c} \log \frac{ni}{c}) \rightarrow$$

(2)

where C is the cardinality of the non-empty space units of PO, and n_i is the number of solutions of set PO_1 inside the corresponding space unit. The more diversified the solution set PO_1 on the frontier the higher the entropy value $(0 \le E \le 1)$.

2.4 Steps of feature selection through Parallel MOOGA

1. First save the preprocessed dataset in Open MPI Tool.
2. Load the data.
3. Create an initial population.
4. Create a fitness function for the Genetic Algorithm
5. Set Genetic Algorithm operators
6. Run the Multi Objective Optimization GA and get the best solution

   Pareto front obtained by contribution and Entropy :Performance Metrics
7. Get the Best discriminative Features.

3. ANFIS:ADAPTIVE NEURO FUZZY INFERENCE SYSTEM CLASSIFICATION

3.1 Fuzzy Inference System:

The fuzzy inference system that we have considered is a model that maps

1. Input characteristics to input membership functions,
2. Input membership function to rules,
3. Rules to a set of output characteristics,
4. Output characteristics to output membership functions, and
5. The output membership function to a single-valued output, or
6. A decision associated with the output.

Architecture of ANFIS

The ANFIS is a framework of adaptive technique to assist learning and adaptation. This kind of framework formulates the ANFIS modeling highly organized and not as much of dependent on specialist involvement. To illustrate the ANFIS architecture, two fuzzy if-then rules according to first order Sugeno model are considered

$$Rule\ 1: If\,(x\ is\ A_1)and\ (y\ is\ B_1)then\ (f_1 = p_1x + q_1y + r_1)$$
$$Rule\ 2: If\,(x\ is\ A_2)and\ (y\ is\ B_2)then\ (f_2 = p_2x + q_2y + r_2)$$

where x and y are nothing but the inputs, Ai and Bi represents the fuzzy sets, fi represents the outputs inside the fuzzy region represented by the fuzzy rule, pi, qi and ri indicates the design parameters that are identified while performing training process.

The ANFIS architecture to execute these two rules is represented in figure 2, in which a circle represents a fixed node and a square represents an adaptive node.

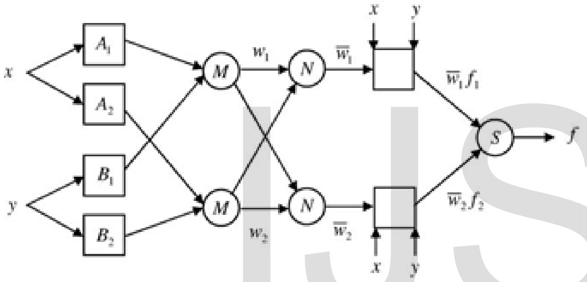


Figure 2.ANFIS Architecture

In the first layer, every node is adaptive node. The outputs of first layer are the fuzzy membership grade of the inputs that are represented by:

$$O_i^1 = \mu_{A_i}(x) \quad i = 1,2 \longrightarrow \quad (3)$$

$$O_i^1 = \mu_{B_{i-2}}(y) \quad i = 3,4 \longrightarrow \quad (4)$$

It can be noted that layer 1 and the layer 4 are adaptive layers. Layer1 contains three modifiable parameters such as ai, bi, ci that are associated with the input membership functions. These parameters are called as premise parameters. In layer 4, there exists three modifiable parameters as well such as {pi, qi, ri}, related to the first order polynomial. These parameters are called consequent parameters.

3.2 Learning algorithm of ANFIS

The intention of the learning algorithm is to adjust all the modifiable parameters such as{ai , bi, ci} and {pi, qi, ri}, for the purpose of matching the ANFIS output with the training data.

If the parameters such as ai, bi and ci of the membership function are unchanging, the outcome of the ANFIS model can be given by:

$$f = \frac{w_1}{w_1 + w_2} f_1 + \frac{w_2}{w_1 + w_2} f_2 \longrightarrow \quad (10)$$

Substituting Eq. (7) into Eq. (10) yields:

$$f = \bar{w}_1 f_1 + \bar{w}_2 f_2 \longrightarrow \quad (11)$$

Substituting the fuzzy if-then rules into Eq. (11), it becomes:

$$f = \bar{w}_1(p_1x + q_1y + r_1) + \bar{w}_2(p_2x + q_2y + r_2) \longrightarrow \quad (12)$$

After rearrangement, the output can be expressed as:

$$f = (\bar{w}_1x)p_1 + (\bar{w}_1y)q_1 + (\bar{w}_1)r_1 + (\bar{w}_2x)p_2 + (\bar{w}_2y)q_2 + (\bar{w}_2)r_2 \longrightarrow \quad (13)$$

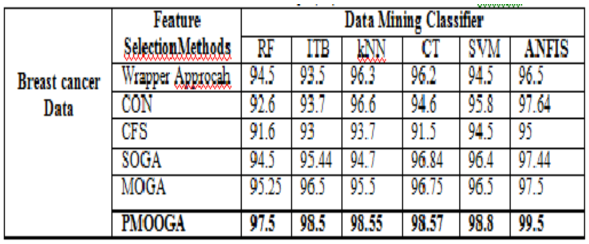
Which is a linear arrangement of the adjustable resulting parameters such as p1, q1, r1, p2, q2 and r2. The least squares technique can be utilized to detect the optimal values of these parameters without difficulty. If the basis parameters are not adjustable, the search space becomes larger and leads to considering more time for convergence. A hybrid algorithm merging the least squares technique and the gradient descent technique is utilized in order to solve this difficulty. The hybrid algorithm consists of a forward pass and a backward pass. The least squares technique which acts as a forward pass is utilized in order to determine the resulting parameters with the premise parameters not changed. Once the optimal consequent parameters are determined, the backward pass begins straight away. The gradient descent technique which acts as a backward pass is utilized to fine-tune the premise parameters equivalent to the fuzzy sets in the input domain. The outcome of the ANFIS is determined by using the resulting parameters identified in the forward pass. The output error is utilized to alter the premise parameters with the help of standard back propagation method. It has been confirmed that this hybrid technique is very proficient in training the ANFIS.

4.EXPERIMENTS AND RESULTS

Our proposed method has been implemented in Multi core processor environment [1],[4]. The best features have been taken from Parallel MOOGA gene feature selection method and the ANFIS classifier is used to evaluate the gene subsets using CUDA enabled MATLAB Tool.

We carried out this experiment on publicly available microarray breast cancer datasets available at Kent Ridge Bio-medical Data Set Repository (http://datam.i2r.a-star.edu.sg/datasets/krbd/).The existing feature selection methods are applied for this breast cancer data sets and classification accuracy are measured using Orange data mining and machine Learning tool[15].The feature selection methods[3],[8] like Wrapper approach, Consistency Subset Selection (CON) [9], Correlated Feature Selection (CFS) [6], Single Objective Genetic Algorithm (SOGA) [7], Multi-Objective GA (MOGA) and the proposed Parallel Multi Objective GA Optimization (PMOOGA) are applied on breast cancer data sets. It is then measured by various classification methods like Random Forest, Interactive Tree builder (ITB),K-NN, Classification tree, and SVM,in Orange tool. Our proposed PMOGA feature selection method with ANFIS provides better classification accuracy than the other feature selection algorithm and the computing time is very less than the other methods. Table 1 shows the Classification Accuracy of various methods in terms of time and graphically shown in Figure 3 and Table 2 shows the execution time of ANFIS classifier on different feature selection methods.

Table 1: Classification Accuracy (%) Of Breast Cancer Data .



	Feature SelectionMethods	Data Mining Classifier					
		RF	ITB	kNN	CT	SVM	ANFIS
Breast cancer Data	Wrapper Approcah	94.5	93.5	96.3	96.2	94.5	96.5
	CON	92.6	93.7	96.6	94.6	95.8	97.64
	CFS	91.6	93	93.7	91.5	94.5	95
	SOGA	94.5	95.44	94.7	96.84	96.4	97.44
	MOGA	95.25	96.5	95.5	96.75	96.5	97.5
	PMOOGA	**97.5**	**98.5**	**98.55**	**98.57**	**98.8**	**99.5**

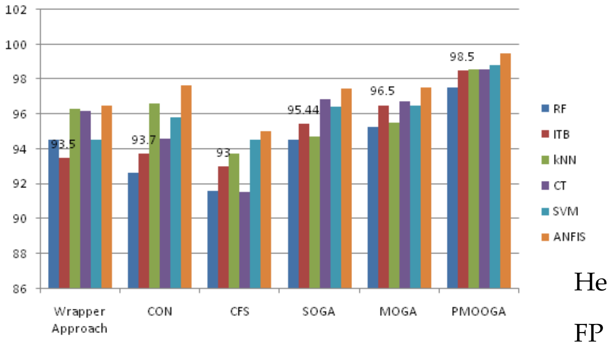


Figure 3.Classification Accuracy

Table 2. Execution time of ANFIS classifier on different feature selection methods.

Feature Selection Methods	Executing Machine	ANFIS Classifier	
		Accuracy (%)	Time in Hrs
Wrapper Approcah	Sequential Programming	96.5	18.5
CON		97.64	18
CFS		95	17.7
SOGA		97.44	18
MOGA		97.5	18.5
PMOOGA	**Parallel Computing**	**99.5**	4.5

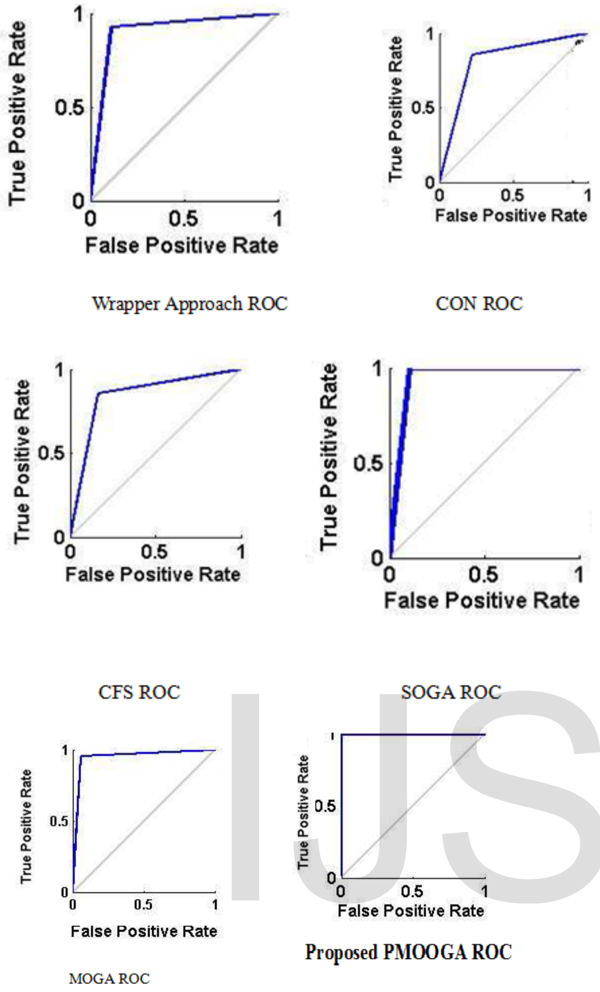
The Cross-fold validation is used for measuring the classification accuracy in terms of time which is available at Orange tool and PMOOGA is implemented by us using Parallel computing. The existing feature selection methods has implemented in Sequential programming. From Table 2 we observe that the proposed PMOOGA method is better than other methods in terms of time , feature selection and classification accuracy.

Some statistical measurements like Recall (True Positive Rate), Specificity(False Positive Rate) of the classifiers are calculated using equation (14), (15).

$$\begin{aligned} \text{TPR (Recall)} &= TP / P \\ &= TP / (TP+FN) \longrightarrow \quad (14) \end{aligned}$$

$$\begin{aligned} \text{FPR (Specificity)} &= FP / N \\ &= FP / (P+TN) \longrightarrow \quad (15) \end{aligned}$$

Here the TP is positive object classified as positive, FP is positive object classified as negative, TN is negative object classified as negative and FN is negative object classified as positive. The ROC curves of experimental dataset reduced by Wrapper approach, CON, CFS,SOGA, MOGA and proposed PMOOGA for ANFIS classifier are shown in Fig. 4.It is observed that, the ROC graph corresponding to the proposed method PMOOGA based on ANFIS classifier rises almost vertically from (0, 0) to (0, 1) and then horizontally to (1,1) whereas the graph for other methods are not so vertical and horizontal. This indicates perfect and truly significant classification performance on the dataset.



Wrapper Approach ROC

CON ROC

CFS ROC

SOGA ROC

MOGA ROC

Proposed PMOOGA ROC

Figure 4. ROC curve of Breast Cancer dataset for ANFIS classifier

Table 3. Breast cancer feature set .

Data Set	Sub Types of Breast Cancer	No of selected Attributes	Best Features
Breast Cancer	Luminal A, Luminal B, HER2+ Basal Like	20	F4186,F4831,F2475,F1540,F2592, F8463,F46,F3947,F1515,F4713, F71,F976,F83,F62,F1686, F54,F7601,F7511,F885,F7413.

5. CONCLUSION AND FUTURE WORK

In this research, we have applied a combination of Parallel MOOGA-ANFIS approach to the Breast cancer dataset diagnosis problem. The objective of the work is to find the presence of breast cancer sub types. The proposed work also helps to minimize the computing time, maximize the accuracy and better specificity and recall measures compare to other methods. Feature selection is an essential part of this research. With the help of feature selection method on microarray data, the computation cost reduces and also the classification accuracy increases. The principle of feature selection has been implemented using the Parallel MOOGA and the best solution is obtained by Multi objective optimization performance indicator contribution and entropy algorithms. It has been shown that for effective and efficient diagnosis the ANFIS with 20 features shows good accuracy. The Selected features are shown in Table 3. This technique is fast in execution, efficient in classification and easy in implementation. In this research work standard microarray data sets are taken from Kent Ridge Bio-medical Data Set Repository. In future, real time data from cancer patients has to be taken. The classification accuracy should also be clinically verified

6.REFERENCES

1 August A.D., Chiou K.P.D, Sendag R., Yi J.J., “Programming Multicores: Do Application Programmers Need to Write Explicitly Parallel Programs?”, Computer Architecture Debates in IEEE MICRO, pp. 19-32. 2010

2 Asha Gowda Karegowda, M.A.Jayaram, A.S. Manjunath, *Feature Subset Selection Problem using Wrapper Approach in Supervised Learning*, International Journal of Computer Applications 1(7):13–17, February 2010.

3 Aboul Ella Hassaneian, *Classification and feature selection of breast cancer data based on decision tree algorithm*, Studies and Informatics Control ,vol12, no1,March 2003.

4 Charles Severance,Kevin Dowd,: High Performance Computing,revised edition 2005.

5 Goldberg, D.E., Holland, J.H.: Genetic algorithms and machine learning. Machine Learning 3(2), 95–99 (1988).

6 Hall, M.A.: Correlation-based feature selection for machine learning. Diss., The University of Waikato (1999)

7 Hong T.P,Wang H, and ChenH, Simultaneously applying multiple mutation operators in genetic algorithms, *J. Heuristics*, 6, 439–455 (2000).

8 Kemal Polat, Seral Sahan, Halife Kodaz and Salih Günes, *A New Classification Method for Breast Cancer Diagnosis: Feature Selection Artificial Immune Recognition System (FS-AIRS),* In Proceedings of ICNC (2). pp.830~838. (2005)

9 Liu Yu, L.,H.: Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research 5, 1205–1224 (2004)

10 Mohammed Khabzaoui, Clarisse Dhaenens, A Cooperative Genetic Algorithm for Knowledge Discovery in Microarray Experiments,*Parallel Computing for Bioinformatics and computational Biology*,2006 JohnWiley & Sons, Inc.

11 Quinn M. J. *Parallel Programming in C with MPI and OpenMP*, McGraw-Hill, NewYork,2004.

12 Soumen Kumar Patel, Asit Kumar,: Gene Selection Using Multi-objective Genetic Algorithm Integrating Cellular Automata and Rough Set Theory, SEMCCO 2013, Part II, LNCS 8298, pp. 144–155, 2013.© Springer International Publishing Switzerland 2013.

13 Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques.
Morgan Kaufmann, USA (2005)

14 Xiaosheng Wang1 and Osamu gotoh" A Robust Gene selection Method for Microarray-based cancer Classification" Cancer Informatics 2010 15–30.

15 http://orange.biolab.si/docs/latest/tutorial/rst/

16 http://en.wikipedia.org/wiki/CUDA.

17 Qingzhong Liu,Andrew H. Sung,: Feature Selection and Classification of MAQC-II Breast Cancer and Multiple Myeloma Microarray Gene Expression Data

18 C. M. Fonseca and P. J. Fleming,An overview of evolutionary algorithms in multiobjective optimization, *Evol. Comput.*, 3 (1), 1–16 (1995).

19 D. A. vanVeldhuizen and G. B. Lamont, On Measuring Multiobjective Evolutionary Algorithm Performance, in In 2000 Congress on Evolutionary Computation, Piscataway, New Jersey, Vol. 1, July, 2000, pp. 204–211.

20 J. D. Knowles, D. W. Corne, and M. J. Oates, On the Assessment of Multiobjective Approaches to the Adaptive Distributed Database Management Problem, in *Proceedings of the Sixth International Conference on Parallel Problem Solving from Nature (PPSN VI),*September, 2000, pp. 869–878.

21 M. Basseur, F. Seynhaeve, and E.-G. Talbi, Design of multi-objective evolutionary algorithms:application to the flow-shop scheduling problem, in *Congress on Evolutionary Computation CEC'02*

22 A. Zibakhsh, M. Saniee Abadeh," Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness functions". Engineering Applications of Artificial Intelligence 26 (2013) 1274–1281.