Simultaneous Gene Selection and Cancer Classification using a Hybrid Intelligent Water Drop Approach

Manish Kumar[1], Shameek Ghosh[2], Jayaraman Valadi[2, 3*], Patrick Siarry[4*]

[1]Bioinformatics Center, University of Pune, Pune, India
[1]rishimanish123@gmail.com
[2]Centre for Development of Advanced Computing, Pune, India
[2]shameekg@cdac.in
[3]Shiv Nadar University, Uttar Pradesh – 203207, India
[3]jayaraman.valadi@snu.edu.in
[4]Université Paris-EstCréteil, Val-de-Marne, LiSSi (EA 3956),France
[4]siarry@u-pec.fr

Abstract. Computational Analysis of gene expression data is extremely difficult, due to the existence of a huge number of genes and less number of samples (limited number of patients). Thus,it is of significant importance to provide a subset of the most informative genesto a learning algorithm, for constructing robust prediction models. In this study, we propose a hybrid Intelligent Water Drop (IWD) - Support Vector Machines (SVM) algorithm, with weighted gene ranking as a heuristic, for simultaneous gene subset selection and cancer prediction. Our results, evaluated on three cancer datasets, demonstrate that the genes selected by the IWD technique yield classification accuracies comparable to previously reported algorithms.

Keywords: Gene Selection; Cancer Classification; Intelligent Water Drop based Optimization; Weighted Ranking

1 Introduction

The number of genes in a microarray gene expression dataset is normally much greater than the number of samples (instances), which makes the disease prediction problem difficult to solve since, out of thousands of genes, most genes do not correlate with the prediction process. To improve model accuracy, it is thus important to select a subset of relevant genes from the data. This is known as *Gene Selection or Feature Selection* and it helps in getting rid of irrelevant and noisy genes [1]. Two important categories of gene selection methods are: 1) *wrappers* and 2) *filters* [1]. Wrappers use a learning algorithm to score the quality of gene subsets based on their predictive power. On the other hand, filters select subsets of genes independently of the chosen predictor and evaluate the quality of genes considering their statistical properties. The

* Corresponding authors

problem of gene selection typically falls into the category of large-scale global optimization. Various nature-inspired optimization algorithms have been reported to solve such problems. Among these, swarm intelligence based methods have their own niche and sophisticated operators, which explore novel swarm based approaches to tackle optimization problems. According to the No-Free-Lunch Theorem [11], all metaheuristic based approaches report the same performance results when averaged over all possible objective functions. Thus, even though the spectrum of gene selection problems is quite huge, the numbers of reported swarm based metaheuristics are much less in comparison. Hence, in this study, we present a hybrid Intelligent Water Drop Optimization (IWD) based filter-wrapper approach for selecting a relevant subset of genes most predictive of a certain type of cancer.

2 Methodology

2.1 Intelligent Water Drop based optimization

The Intelligent Water Drop (IWD) algorithm has been inspired by the study of the real behavior of natural drops in a flowing water source from high altitude to low altitude regions. Shah-Hosseini extended this natural concept to introduce the Intelligent Water Drop (IWD) algorithm for the Travelling Salesman Problem (TSP) [2]. An IWD consists of two major properties - 1) the soil content of the IWD - *soil(IWD)* and 2) the velocity of the IWD - *vel(IWD)*. The IWD soil and velocity content dynamically change based on the path taken by the same, while flowing through the discrete problem landscape. Depending on the IWD movement, some soil is thus removed from the traversed path and the corresponding path soil is updated dynamically in the process. Such a flow results in the lowering of soil content in optimal routes based on the problem environment. One can thus say that the paths with lesser soil content may be the most relevant for the search of a near optimal solution. Based on the original formulation for a TSP problem, we may consider a graph G= (V, E) where V is the set of nodes and E, the set of edges. An IWD can thus be randomly placed at any node (say i). To select the next node (j), it follows the probability transition as given in equation (1).

$$P(i,j) = \frac{f(soil(i,j))}{\sum_{k \in unvisited} f(soil(i,k))} \quad (1)$$

$$f(soil(i,j)) = \frac{1}{\varepsilon + g(soil(i,j))} \quad (2)$$

$$g(soil(i,j)) = \begin{cases} soil(i,j) & \text{if } minsoil >= 0 \\ soil(i,j) - minsoil & \text{if } minsoil < 0 \end{cases} \quad (3)$$

Here P(i,j) indicates the transition probability associated with node j. k specifically denotes all the nodes that are still to be visited. ε is an algorithmic parameter. Thus the selection of a node depends probabilistically on the amount of soil present on the edges between adjacent nodes given by soil(i,j). Here *minsoil* indicates the least soil available on a path between any node i and j. As illustrated in equations (1) to (3), the state transition probability of an IWD is thus proportional to the soil content available in the edge between nodes i and j. While each IWD incrementally moves from one

node i to j while constructing a solution, the IWD soil content (soil(iwd)) and the velocity of the same(vel(iwd)) are also updated based on equations (4-5).

$$\Delta vel^{IWD}(t) = \frac{a_v}{b_v + c_v \times soil^{2\alpha}(i,j)} \quad (4)$$

$$\Delta soil(i,j) = \frac{a_s}{b_s + c_s \times time^{2\theta}(i,j)} \quad (5)$$

Here the IWD velocity is changed by a Δ vel component. a_v ,b_v ,c_v , and α are algorithm specific parameters. Similarly the soil content of an IWD is also increased by Δsoil which is the soil content removed by the IWD while moving from location i to j. $time^{2\theta}(i,j)$ is the time required for the IWD to move from i to j which is given as –

$$time(i,j) = \frac{HUD(i,j)}{vel(IWD)} \quad (6)$$

HUD is characterized as a heuristic which can be used to measure the desirability/undesirability of an IWD to select an edge between i and j. θ, in this case, is an algorithmic parameter. Thus, a larger IWD velocity contributes to minimizing the time taken by an IWD to move from i to j. The time factor in turn influences the amount of soil to be removed from a path (as shown in equation 5). Once the IWD properties are computed, the soil content of the complete solution path can be updated based on equation 7.

$$soil(i,j) = \rho_0 \times soil(i,j) - \rho_n \times \Delta soil(i,j) \quad (7)$$

whereρ_o and ρ_n are between 0 and 1.According to the original IWD algorithm for the TSP, $\rho_o = 1 - \rho_n$.

2.2 IWD based Feature Selection

For the feature selection problem, we consider each node (in the graph above) as a feature. Thus if a gene expression dataset consists of 1000 features, then a possible solution could be a feature subset composed of {11,23,391,510,999} with the subset size as 5. Here each element is a feature index. An initial set of IWDs are thus placed at random features from where they commence their flow. Each IWD moves to the next feature by following the probability transition given by equation (1). Once a feature has been visited, a local soil update between features i and j are performed by equation (7) as mentioned before. In the process, the IWD soil content and velocity are also updated by equation 4-5. This process, continues until a complete feature subset of the required size is constructed by the IWD. The feature subset is then used to generate a corresponding reduced dataset with the given features indices. The reduced dataset is thus fed as input to a classifier like SVM, which consequently returns a 10 fold classification cross validation accuracy (10 fold CVA).The 10 fold CVA is thus considered as the fitness measure for the corresponding feature subset(or the IWD solution). Subsequent IWDs also build up their solution vectors (feature subsets) similarly. After each iteration, the feature subset with the maximum 10 fold CVA gets selected as the

iteration best solution (TIB). A certain amount of soil is removed from the edges of the iteration-best solution based on the quality of the feature subset. Thus if TIB is given as (6,13,91,121,992), then the edges to be updated are 6-13,13-91,91-121 and 121-992.This is done according to equation (8).

$$soil(i,j) = \rho_s \times soil(i,j) - \rho^{IWD} \times \frac{1}{N_{IB} - 1} \times soil_{IB^{IWD}} \quad (8)$$

Here $soil_{IB^{IWD}}$ represents the soil content of the iteration-best IWD (which owns the iteration best feature subset). N_{IB} is the number of features in TIB. ρ^{IWD} is the global soil updating parameter selected from [0, 1]. ρ_s is set as $(1+ \rho^{IWD})$. Therefore an edge with lesser soil content turns out to have better prospects in the future in the constructing a good solution.

In addition we also maintain a global best feature subset which is given by the maximum of all the iteration best solutions. The above process is repeated till a termination criterion is reached. During this stage, the global best feature subset is reported as the most optimal solution to the feature selection problem. The IWD gene selection algorithm is thus stated as below.

2.3 Weighted Gene Ranking

A weighted gene ranking composed of three filters namely Information Gain (IG), Chi-square(CS) and Correlation based feature selection (CFS), are provided as input to the IWD algorithm [3].The heuristic information for each individual gene is obtained by calculating the weighted sum of the IG, CS and CFS scores which were obtained using the WEKA[4] data mining library. The computation of the weighted sum of a gene (WRg) is as shown in equation (9).

$$WR_g = w_1 * IG_g + w_2 * CS_g + w_3 * CFS_g \quad (9)$$

Here, w1,w2 and w3 are the weights provided for IG, CS and CFS rankings. The WRg is consequently provided as HUD(i,j) for the j-th feature, as shown in equation (6), in a modified form as given in equation (10).

$$time(i,j) = \frac{1}{WR_j + vel(IWD)} \quad (10)$$

The weighted gene value (WRj) is thus used to probabilistically guide the IWD search.

2.4 Support Vector Machine

Support Vector Machines (SVM) were introduced by Vapnik et al [5-6] and successively extended by a number of other researchers. SVM uses a maximum margin linear hyperplane for solving binary linear classification problems. For problems that are non-linearly separable, SVM transforms the data into higher dimensional features and then employs a linear hyperplane. To deal with intractability issues it also employs appropriate kernel functions allowing computations in the input space itself. In particular, SVM with recursive feature elimination (RFE) was used by Vapnik et al [7] for

gene selection and achieved notably high accuracy levels. For our purposes, we employ the libSVM [8] library for evaluation of our candidate solutions during each generation.

3 Results and discussion

Microarray gene expression datasets specify the expression levels of different genes, which are available publicly. Three such datasets were obtained from the Kent Ridge Biomedical datasets repository [8] and the libSVM repository [7].The Colon cancer dataset consists of 2000 genes and 62 samples. The Breast Cancer data constitues 7129 genes and 44 samples. The leukemia dataset consists of 7129 genes and 72 samples. Each of the three datasets constitutes a binary classification problem. Extensive simulations were carried out for each dataset with separate gene rankings as Information gain, Chi-square, CFS and the weighted heuristics as described earlier. Based on the simulations, one can say that comparable results for all three datasets were observed, while considering a maximum of 50 IWDs and 100 generations. Mostly towards the end of 100 generations, the fitness values of the feature subsets would converge and not show much improvement. Parameter tuning was also carried out extensively for weighted ranking to get the best results. Our simulations indicate that the SVM kernel and filter weighting parameters have a more profound influence and have thus tuned the same extensively for maximizing algorithm performance. The algorithm parameters for IWD are as shown in Table II.A comparison of the weighted IWD-SVM performance is provided along with the some recently reported best results for the same datasets. The results of the simulations are as given in Table III.

Table II. *IWD Parameters*

IWD Algorithm Parameters	*Values*
No. of IWDs	50
w_1,w_2,w_3	0.5,0.3,0.2
No. of Generations	100
a_v,b_v,c_v,α	1,0.01,1,1
a_s,b_s,c_s,θ	1,1,0.01,2
ρ_o , ε	0.1,0.5
cost,gamma(for radial basis function as SVM kernel),Folds	50,0.02,10

Table III: *Comparison of IWD-SVM with previously reported classification accuracies [1, 9, 10]. ACO-AM: Ant Colony Optimization with AntMiner, ACO-RF: Ant Colony Optimization with Random Forests.*

Colon	85.48% (SVM)	95.47%(ACO-AM)[1]	96.77%(ACO-RF) [1]	95.16%(IWD-SVM)
Breast	60.02% (SVM)	92.00% (Bagging)[9]	94.00%(Ensemble Predictors)[10]	97.72 %(IWD-SVM)
Leukmia	94.73% (SVM)	96.00%(ACO-AM)[1]	69.00% (nearest neighbor search)	97.22%(IWD-SVM)

According to results in Table III, IWD-SVM performs well in comparison to previously reported algorithms namely for all the three datasets [1,9,10]. The IWD based gene subset sizes selected were 15 for Colon, 15 for Breast and 19 for Leukemia. In addition, simulations with simple filters like Information Gain, Chi-square ranking and CFS were carried out separately with IWD for similar subset sizes. As per our results, the IWD-SVM with weighted ranking demonstrated superior performances than IWD-Infogain, IWD-Chi-Square and IWD-CFS.

4 Conclusion

The hybrid IWD-SVM has shown good results consistently on comparison with the highest accuracies for colon cancer, breast cancer and leukemia cancer datasets. In general, IWD is robust and flexible for discrete optimization owing to their typical swarm based emergent behavior.

Acknowledgment.VKJ gratefully acknowledges the Council of Scientific & Industrial Research (CSIR), New Delhi, India for financial support in the form of an Emeritus Scientist grant.

References

1. Sharma, Shimantika, et al. : Simultaneous informative gene extraction and cancer classification using aco-antminer and aco-random forests.In Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012, India, January 2012. pp. 755-761. Springer Berlin Heidelberg, 2012.
2. Shah-Hosseini H. :Problem solving by intelligent water drops. In IEEE Congress on Evolutionary Computation, CEC 2007. pp. 3226-3231 (2007)
3. J. Han and M. Kamber. Data mining: concepts and techniques. Morgan Kaufmann, 2006.
4. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., and Witten I.H. :The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 11, no. 1, 10-18(2009).
5. Boser B. E., Guyon I.M., and VapnikV. N.: A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory, ser. COLT '92. New York, NY, USA: ACM, pp. 144–152 (1992)
6. Guyon I., Weston J., Barnhill S., and Vapnik V.: Gene selection for cancer classification using support vector machines. Machine Learning, vol. 46, pp. 389–422, (2002).
7. Chang C.-C. and LinC.-J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, vol. 2, pp. 27:1–27:27, (2011)
8. Kent ridge bio-medical dataset, URL: http://datam.i2r.astar.edu.sg/datasets/krbd/
9. Martn-Merino M., BlancoA. and De Las RivasJ.: Combining dissimilarity based classifiers for cancer prediction using gene expression profiles. BMC Bioinformatics, vol. 8, (2008).
10. CongG., TanK.-L., TungA. K. H., and XuX.: Mining top-k covering rule groups for gene expression data. In Proceedings of the ACM SIGMOD international conference on Management of data, ser. SIGMOD '05. New York, NY, USA: ACM, pp. 670–681 (2005).
11. Wolpert DH, Macready WG.: No free lunch theorems for optimization. IEEE Trans Evol Comput 1(1):67–82, (1997)