Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Gene Selection in Microarray Datasets Using Progressively Refined PSO Scheme

Yamuna Prasad, K. K. Biswas
Indian Institute of Technology Delhi
New Delhi, India 110016
{yprasad, kkb}@cse.iitd.ac.in

Abstract

In this paper we propose a wrapper based PSO method for gene selection in microarray datasets, where we gradually refine the feature (gene) space from a very coarse level to a fine grained one, by reducing the gene set at each step of the algorithm. We use the linear support vector machine weight vector to serve as the initial gene pool selection. In addition, we also examine integration of other filter based ranking methods with our proposed approach. Experiments on publicly available datasets, Colon, Leukemia and T2D show that our approach selects only a very small subset of genes while yielding substantial improvements in accuracy over state-of-the-art evolutionary methods.

Introduction

DNA microarray technology produces expression levels of thousands of genes simultaneously in a sample. The data produced through microarray technology has expression levels of thousands of genes while the sample size is very small. This raises the issue of generalization in the classification process to determine which genes are responsible for cancer detection (Li et al. 2008). Gene selection plays a very important role in improving accuracy of classifiers. Amongst various approaches wrapper based methods have been shown to have higher accuracies.

In the literature, hybrid Particle Swarm Optimization (PSO/GA) method with Wilcoxon's rank test (Li et al. 2008), Novel Hybrid Framework (NHF) (Zhao et al. 2011), Genetic Swarm Algorithm (GSA) (Ganesh K. et al. 2012) and Binary Matrix Shuffling Filter (BMSF) (Zhang et al. 2012) have been shown to outperform many state-of-the-art gene selection methods.

In this paper, we propose a wrapper based PSO method which progressively refines the selected gene set. The initial top ranking gene set is obtained through linear SVM weight vector, as well through Wilcoxon's rank test (Li et al. 2008). Our experiments on five publicly available benchmark microarray datasets show that our proposed method selects minimal subsets of genes while improving the overall prediction accuracy.

Proposed Methods

In the first version, we develop a hybrid PSO using a linear SVM weight vector to generate top K genes and applying standard PSO on this gene set to carry out gene selection (named PSW method). The Progressively Refined PSO method starts by exploring the whole search space randomly in the first step and then in the successive steps the search space is refined. The PSO method is applied iteratively on this reduced search space maintaining a non-decreasing accuracy to reduce the gene set (steps 4 - 10). This process iterates as long as the gene set keeps on reducing but stops as soon as classification accuracy gets degraded (step 6 - 8, i.e stopping criteria). The algorithm returns the reduced set of informative genes and the best accuracy achieved (referred to as RPSW approach henceforth).

Algorithm 1: Progressively Refined PSO

Input: Dataset X, Number of features M, Number of particles n, Number of PSO iterations T, Maximum iterative depth D, Acceleration constants $C1$ and $C2$ and inertia weight ω.
Output: Selected genes S, Global best accuracy G.

1. Initialize $FLAG$ = False.
2. Assign all features to set S (i.e. $|S| = M$) and set $G = 0$.
3. **for** iter = 1 to D **do**
4. Execute PSO (Li et al. 2008) using X, M, n, T, $C1$, $C2$ and ω.
5. Compute the best subset of genes GB and global best accuracy G'.
6. **if** $G' \geq G$, **then** $S = GB$.
7. **if** $|S| = M$, **then** $FLAG$ = True.
8. **if** ($FLAG$ is True **or** $G' < G$), **then** *B*reak.
9. Set $M = |S|$ and $G = G'$.
10. Reduce the dataset X to X' with selected genes S.
11. **end for**
12. Return S and G

The complexity of algorithm increases as the iterative depth increases. We have used linear SVM (Fan et al. 2008) to evaluate the fitness of the particles. The total complexity

Table 1: Average testing accuracy (with number of genes selected in parentheses) in ten runs.

Methods	10CV			LOOCV		
	Colon	**Leukemia**	**T2D**	**Colon**	**Leukemia**	**T2D**
SV	78.7(all)	86.1(all)	55.9(all)	80.3(all)	84.7(all)	52.9(all)
P	87.4(517.9)	91.3(1864.2)	67.1(5264.3)	90.1(543.1)	90.0(2214.1)	68.8(5807.5)
PSO-GA(Li et al. 2008)	88.7(16.3)	95.1(21.0)	-	-	-	-
NHF (Zhao et al. 2011)	81.3(51.0)	92.4(89.0)	-	-	-	-
BMSF-SVM(Zhang et al. 2012)	94.4(-)	98.3(-)	-	95.2(7)	98.6(3)	-
GSA(Ganesh K. et al. 2012)	-	-	-	96.8(10)	94.4(10)	97.1(10)
RP	91.8(5.8)	96.5(10.1)	91.5(6.0)	93.4(8.0)	94.6(9.4)	92.9(7.2)
RPWL	93.0(5.5)	92.2(5.4)	100(4.9)	92.1(5.3)	93.2(6.4)	100(5.4)
RPSW	**97.7(8.5)**	**100(7.6)**	**100(5.4)**	**98.0(7.9)**	**100(8.9)**	**100(5.7)**

of the proposed Progressively Refined PSO (PRPSO) algorithm is $O(DTnT_{SVM})$, where n is the number of particles, T is the number of PSO iterations, D is the maximum iterative depth and T_{SVM} is the complexity of linear SVM classifier. In the second version, we form the initial set of top K genes using Wilcoxon's rank test (Li et al. 2008), and integrated it with progressively refined PSO. We refer to this approach as RPWL, henceforth.

Experiments

Dataset We experiment with five publicly available benchmark microarray datasets, namely Colon, Lymphoma, Leukemia, RAOA and T2D (Ganesh K. et al. 2012). For comparing classification accuracy, we have named various methods as follows:

SV: SVM with all the features
P: PSO with all the features
RP: PRPSO method
RPWL: PRPSO with Wilcoxon method
RPSW: PRPSO with Linear SVM weight vector

We followed training and test splitting of (Li et al. 2008) and (Ganesh K. et al. 2012) for 10CV and LOOCV strategies respectively. Gene selection is performed on training data only. For PSO implementation, we used $C1$ = 2, $C2$ = 2, ω =0.9, K=40 and T=100 (Li et al. 2008). The maximum number of iterative depth D is set to 20 in PRPSO method. We use Linear SVM (Fan et al. 2008) to compute the accuracy. The cost parameter C in SVM is tuned using 5-fold cross-validation on training dataset only for computing fitness of a particle in PSO. The experiments are repeated ten times and average accuracies are reported.

Results We have presented the average accuracies in 10 runs for 10CV and LOOCV in Table 1 for 3 datasets namely Colon, Leukemia and T2D. In the table, dash $(-)$ indicates that results for those datasets have not been reported in the literature. It is observed that the proposed RPSW method achieves highest accuracy while selecting a fewest number of features for 10CV as well as LOOCV. For T2D dataset, both RPSW and RPWL exhibits similar accuracy pattern. Further, all the proposed RP, RPWL and RPSW methods achieve higher accuracy with a small number of features than the all other methods. We have also computed Bonferroni correction P-values for RPSW method against the RPWL, RP and P methods. The p-values for Colon dataset at $\alpha = 0.05$ are 2.39e-13, 0 and 0 for RPWL, RP and P with RPSW respectively. This shows the significance of RPSW method over the RPWL, RP and P methods. Similar trends are observed for all the datasets.

Conclusion

In this paper, we have proposed a progressively refined PSO approach for gene selection. We have also integrated our method with Wilcoxon's rank test. Additionally, we propose to use the SVM weight vector for feature ranking and integrate this with progressively refined PSO approach (RPSW). The proposed RPSW method gives best set of accuracies and at the same time coming up with a very small sized gene set.

References

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.* 9:1871–1874.

Ganesh K., P.; Aruldoss A. V., T.; Renukadevi, P.; and Devaraj, D. 2012. Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm. *Expert Syst. Appl.* 39(2):1811–1821.

Li, S.; Wu, X.; Tan, M.; and . 2008. Gene selection using hybrid particle swarm optimization and genetic algorithm. *Soft Comput.* 12(11):1039–1048.

Zhang, H.; Wang, H.; Dai, Z.; Chen, M.-s.; and Yuan, Z. 2012. Improving accuracy for cancer classification with a new algorithm for genes selection. *BMC Bioinformatics* 13(1):1–20.

Zhao, W.; Gang, W.; Hong-bin, W.; Hui-ling, C.; Hao, D.; and Zheng-dong, Z. 2011. A Novel Framework for Gene Selection. *Int. J. of Advancements in Computing Technology* 3(3):184–191.