Chapter 3
Top Scoring Pair Decision Tree for Gene Expression Data Analysis

Marcin Czajkowski and Marek Krętowski

Abstract Classification problems of microarray data may be successfully performed with approaches by human experts which are easy to understand and interpret, like decision trees or Top Scoring Pairs algorithms. In this chapter, we propose a hybrid solution that combines the above-mentioned methods. An application of presented decision trees, which splits instances based on pairwise comparisons of the gene expression values, may have considerable potential for genomic research and scientific modeling of underlying processes. We have compared proposed solution with the TSP-family methods and decision trees on 11 public domain microarray datasets and the results are promising.

1 Introduction

A powerful tool for structural and functional analysis of genomes may be developed from DNA chips [6,21]. The entire set of genes of an organism can be microarrayed on an area not greater than 1 cm^2 and makes possible processing of thousands of expression levels simultaneously in a single experiment [12]. Nowadays, DNA chips may be used to assist diagnosis and to discriminate cancer samples from normal ones [2,10]. Extracting accurate and simple decision rules that contain marker genes is of great interest for biomedical applications. However, finding a meaningful and robust classification rule is a real challenge, since in different studies of the same cancer, diverse genes consider to be marked [25].

Typical statistical problems that often occur with microarray analysis are dimensionality and redundancy. In particular, we are faced with the *"small N, large P problem"* [27,28] of statistical learning because the number of samples (denoted by N) comparing to the number of genes (P) remains quite small as N usually does not exceeded one or two hundreds where P is usually several thousands. The high ratio

M. Czajkowski (✉)
Faculty of Computer Science, Bialystok University of Technology, Bialystok, Poland
e-mail: m.czajkowski@pb.edu.pl

H.R. Arabnia and Q.-N. Tran (eds.), *Software Tools and Algorithms for Biological Systems*, Advances in Experimental Medicine and Biology 696,
DOI 10.1007/978-1-4419-7046-6_3,

of features/observations may influence the model complexity [16] and can cause the classifier to overfit the training data. Furthermore, most of the genes are known to be irrelevant for an accurate classification, so gene selection prior the classification should be considered to simplify calculations, decrease model complexity, and often to improve accuracy of the following classification [22].

Recently, many new approaches including hierarchical clustering [1], machine learning [38], methods based on Support Vector Machine [23, 39], neural networks [7], and much more are applied in microarray classification. Usually they are "black-box" approaches, concentrated mostly on the improvement of accuracy. However, they generate very complex models that are difficult to interpret from medical point of view. Nonlinear models may achieve high accuracy but do not provide any significant and easily understood rules. Simple models like decision trees or rule extraction systems, however, may help in understanding underlying processes. Causal relationships between specific genes can be influenced by other causal relationships which can be identified in the model only if such a model is easy to understand. The restriction in considering the model interpretability in classification system often affects performance in classification. This may be called a trade-off between credibility and comprehensibility of the classifiers [30].

In this chapter, we propose a hybrid solution denoted as TSPDT that applies TSP-family algorithm and decision trees. The rest of the chapter is organized as follows. In the next section, TSP-family algorithms and decision tree classifiers for gene expression analysis are briefly recalled. At the end of this section, TSPDT hybrid solution is presented. In Sect. 3, the proposed approach is experimentally validated on 11 real microarray datasets. At the end, the chapter is concluded and possible future works are suggested.

2 Methods

2.1 A Family of TSP Algorithms

Top Scoring Pair (TSP)-family classifiers are one of the most promising techniques that classify gene expression data using a simple decision rules. These statistical methods are widely used to identify marker genes in microarray datasets [36]. Strong point of the TSP is its parameter-free, data-driven learning property and invariance to any simple transformation of data like normalization or standardization. TSP methods are also used in other algorithms as for example feature selection in SVM classifiers [37]. In this section, we will be focused on the original TSP method [14], its ensemble counterpart k-TSP [30], and our extension Weight k-TSP [8]. There is an additional restriction that k should not exceed 10 in the original article due to computational limitations. Computational cost was reduced by omitting cross-validation procedure for automatic determining parameter k (for all algorithms), and comparison was performed on all possible odd number of pairs manually like in [37].

TSP method was presented by Donald Geman [14] and is based on pairwise comparisons of gene expression values. Despite its simplicity to other methods, classification rates for TSP are comparable or even exceed other classifiers. Discrimination between two classes depends on finding pairs of genes that achieve the highest ranking value called "score".

A k-Top Scoring Pairs (k-TSP) classifier proposed by Aik Choon Tan [30] is a simple extension of the original TSP algorithm. The main feature that differs from those two methods is the number of TSPs included in final prediction. In the TSP method, there can be only one pair of genes, and in k-TSP classifier the upper bound denoted as k can be set up before the classification. The parameter k is determined by a cross-validation and in any prediction the k-TSP classifier uses not more than k top scoring disjoint gene pairs that have the highest score.

In classification Weight k-TSP proposed by us [8], selection of TSPs and prediction were changed comparing to TSP and k-TSP. Ranking modification results from limitation in finding optimal TSPs by TSP. Extended prediction was also proposed to improve accuracy for different types of datasets.

2.2 TSP Decision Tree

There have been several attempts to use decision trees for the classification analysis of the gene expression data. Dudoit et al. [11] compare some classification principles, among which there is the CART system [3]. Tan et al. [29] present the application of C4.5, bagged and boosted decision trees. Usage of the ensemble scheme for classification can be also found in Valentini et al. [34], where bagged ensembles of SVM are presented. The committee of individual classifiers is also presented in [18], where ensembles of cascading trees are applied to the classification of the gene expression data. Evolutionary algorithms with decision trees have also been applied in some classification algorithms for microarray datasets [15]. In [37], authors compare decision trees with SVMs on gene expression data and conclude that bagging and boosting decision trees perform as well as or close to SVM. However, ensemble methods alike decision trees with complex multivariate tests based on linear or nonlinear combination splits are more a "black-box" approach. They are difficult to understand or interpret by human experts and do not have potential for genomic research and scientific modeling of underlying processes. Thus, our goal is to improve classification accuracy of the decision trees in a way which will make them still easy to understand. Better decision trees of this kind imply more informative analysis of gene expression data. By combining the strength of the TSP-family algorithms with decision trees, we believe to receive simple decision rules and competitive classier with applications to microarray data.

Decision trees (also known as classification trees) [24] represent one of the main techniques for classification analysis in data mining and knowledge discovery. TSP Decision Tree like many popular decision trees is based on top-down greedy search [26]. First, the test attribute (and the threshold in the case of continuous attributes)

is decided for the root node. Then, the data are separated according to the splitting rule in the current node, and then each subset goes to the corresponding branch. In our research, we have tested splitting rules based on the TSP, k-TSP, and Weight k-TSP algorithms. We have adapted solutions presented in previous section to find the best pair (or pairs) of genes that will separate the data. Root node of the decision tree has identical splitting rule to analogous TSP method. Differences occur in lower parts of the tree where new rules are generated based on the instances in each node. The process is recursively repeated for each branch until leaf node is reached. C4.5-like pessimistic pruning was applied [31] to prevent data overfit. Illustration of the TSPDT is enclosed in the next section, in Fig. 1.

3 Results and Discussions

Performance of the above-mentioned classifiers was investigated on public available microarray datasets summarized in Table 1. Datasets came from Kent Ridge Bio-medical Dataset Repository [19] and are related to the studies of human cancer, including: leukemia, colon tumor, prostate cancer, lung cancer, breast cancer, ovarian cancer, etc. Typical tenfolds cross-validation was applied for datasets that were not arbitrarily divided into the training and the testing sets. To ensure stable results, for all datasets average score of 10 runs is shown.

We have implemented and analyzed TSP-family algorithms, that is TSP, k-TSP, Weight TSP, Weight k-TSP, and proposed hybrid solution, which use this method as splitting criterion, adequately: TSPDT, k-TSPDT, Weight TSPDT, and Weight k-TSPDT. Maximum number of gene pairs k used in all algorithms was default (equal 9) through all datasets. We also included classification results for the traditional decision trees such as J48 (pruned C4.5), ensembles: Bagging, Adaboost, Random Forest, and popular rule learner JRip (RIPPER). Data mining tool called Weka [33] was used for the performance experiments on the above-mentioned methods.

Table 1 Kent Ridge Biomedical gene expression datasets

	Datasets	Abbreviation	Attributes	Training set	Testing set
1	Breast cancer	BC	24,481	34/44	12/7
2	Central nervous system	CNS	7,129	21/39	–
3	Colon tumor	CT	6,500	40/22	–
4	DLBCL vs follicular lymphoma	DF	6,817	58/19	–
5	Leukemia ALL vs AML	LA	7,129	27/11	20/14
6	Lung cancer – Brigham	LCB	12,533	16/16	15/134
7	Lung cancer – University of Michigan	LCM	7,129	86/10	–
8	Lung cancer – Toronto, Ontario	LCT	2,880	24/15	–
9	Ovarian cancer	OC	15,154	91/162	–
10	Prostate cancer	PC	12,600	52/50	27/8
11	Prostate cancer outcome	PCO	12,600	8/21	–

Table 2 Comparison of TSP decision trees accuracy and model size with original methods. The highest classifiers accuracy for each dataset was bolded

	Classifiers accuracy and size							
Datasets	TSP	TSPDT	k-TSP	k-TSPDT	W. TSP	W. TSPDT	W. k-TSP	W. k-TSPDT
1. BC	52.63	73.68	68.42	**78.95**	63.16	57.89	47.37	57.89
	2.00	4.00[a]	18.00	3.00[a]	2.00	7.00[a]	18.00	6.00[a]
2. CNS	49.00	64.17	58.50	63.00	52.17	63.83	50.83	**65.50**
	2.00	6.62[a]	18.00	3.99[a]	2.00	7.09[a]	18.00	3.39[a]
3. CT	83.64	80.29	**88.93**	84.88	85.86	77.71	87.33	80.86
	2.00	4.68[a]	18.00	3.00[a]	2.00	4.79[a]	18.00	4.49[a]
4. DF	72.75	87.54	87.82	**95.25**	88.52	88.52	88.04	90.71
	2.00	6.02[a]	18.00	2.60[a]	2.00	3.17[a]	18.00	2.6[a]
5. LA	73.53	76.47	**91.18**	**91.18**	76.47	76.47	**91.18**	**91.18**
	2.00	2.00[a]	18.00	2.00[a]	2.00	2.00[a]	18.00	1.00[a]
6. LCB	76.51	78.52	83.89	83.89	**97.32**	**97.32**	96.64	94.63
	2.00	2.00[a]	18.00	2.00[a]	2.00	1.00[a]	18.00	1.00[a]
7. LCM	95.87	**98.94**	95.23	97.77	93.02	97.69	97.50	98.40
	2.00	2.90[a]	18.00	2.05[a]	2.00	1.94[a]	18.00	1.89[a]
8. LCT	50.92	58.50	58.42	55.33	55.17	62.42	**72.92**	71.25
	2.00	7.16[a]	18.00	3.38[a]	2.00	4.73[a]	18.00	4.17[a]
9. OC	99.77	**100.00**	**100.00**	**100.00**	97.24	96.45	98.62	98.03
	2.00	1.00[a]	18.00	1.00[a]	2.00	3.07[a]	18.00	3.41[a]
10. PC	76.47	85.29	91.18	94.12	91.18	76.47	**100.00**	**100.00**
	2.00	3.00[a]	18.00	3.00[a]	2.00	8.00[a]	18.00	3.00[a]
11. PCO	72.67	74.17	58.83	59.17	50.17	59.67	88.00	**97.00**
	2.00	1.01[a]	18.00	1.02[a]	2.00	1.38[a]	18.00	1.00[a]
Average	73.07	79.78	80.22	82.14	77.30	77.68	83.49	85.95
	2.00	3.67[a]	18.00	2.46[a]	2.00	4.02[a]	18.00	2.90[a]

[a] Value represents the number of execution TSP algorithm in the decision tree. To calculate the maximum pessimistic number of attributes used in decision tree model, multiply this value by adequate size of TSP (or k-TSP, W.k-TSP) method

Classification was performed with default parameters through all datasets for all algorithms. All classifications were preceded by a step known as feature selection where a subset of relevant features is identified. We decided to use popular for microarray analysis method Relief-F [20] with default number of neighbors (equal 10) and 1,000 features subset size. Experimental results on tested datasets are confronted in Table 2.

3.1 Comparison of TSP-Family Algorithms Methods

Table 2 summarizes classification performance for TSP decision trees and traditional TSP-family methods. First row for each dataset represents classification accuracy and the second row classification size. Results show that for over

two-thirds of datasets, classification accuracy increased (or did not change) when hybrid decision tree with TSP methods was applied. Average accuracy for all datasets also increased, from nonsignificant 0.38% for Weight TSPDT to 6.71% for TSPDT. For some datasets, however, like colon tumor, we can observe lower credibility of the decision tree solution. In this case, proposed method overfit to the training data.

However, it should be noted that one of the factor that stands for promising or (with respect to colon tumor dataset) worse results for the decision trees with TSP-family methods is the higher number of features in decision tree model. Maximum pessimistic number of attributes used in the whole decision tree model is from 2.46 to 4.02 times higher to TSP methods. We can observe that average model size of the proposed approach is larger to analogues TSP methods. In case of the decision trees, however, comprehensibility of the generated rules do not have to be always lower to the compared TSP methods.

Let us compare decision rules for TSPDT with k-TSP methods on one of our tested datasets – breast cancer (*LCB*) [35]. In the Fig. 1, we can observe generated rules and trees for analyzed methods and in Table 3 selected gene names. The k-TSP method is an ensemble TSP algorithm where prediction depends on simple majority voting. For every tested instance, there can be different combination of pairs of genes affecting the right decision which may cause difficulties in finding and

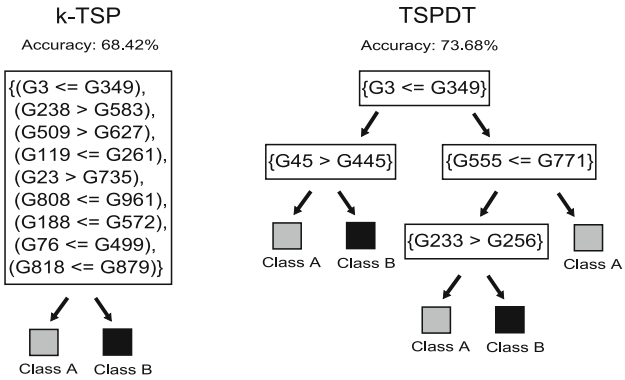

Fig. 1 Outcome for breast cancer dataset for k-TSP and TSPDT method

Table 3 Gene labels and names from Fig. 1

ID	Name	ID	Name	ID	Name	ID	Name
G3	Contig37376 RC	G23	NM 001661	G45	NM 003163	G76	Contig32185 RC
G119	NM 020120	G188	Contig54742 RC	G233	Contig42059 RC	G238	Contig31000 RC
G256	NM 005217	G261	NM 005243	G349	NM 004798	G445	Contig40635 RC
G499	U90911	G509	U82987f	G555	Contig51943 RC	G572	Contig65439
G583	Contig38288 RC	G627	Contig6089 RC	G735	Contig40557 RC	G771	NM 000017
G808	AF161553	G818	NM 019028	G879	NM 018457	G961	Contig32125 RC

extracting significant rules. In our opinion, the TSPDT, alike TSP method, generates much more simpler rules to k-TSP. In the longest path, there have to be tested three pairs of genes; however, no voting procedure is involved. Accuracy for this datasets is also higher for TSPDT method than to k-TSP. If we compare general credibility of the TSPDT and k-TSP algorithm, the results are similar; however, in our opinion proposed solution is much more easier to analyze and interpret by human experts.

3.2 Comparison of Traditional Decision Trees and Rule Classifiers

Performance of the proposed hybrid solution was also compared with popular decision trees and rule classifiers on datasets presented in Table 1. In our research, we have confronted the J48 Tree (pruned C4.5 decision tree) [32], Adaboost (boosting algorithm using Adaboost M1 method) [13], Bagging (reducing variance meta classifier) [4], Random Forest (algorithm constructing a forest of random trees) [5], and rule learner JRip (RIPPER: Repeated Incremental Pruning to Produce Error Reduction) [9]. Results enclosed in Table 4 confirm suggestions [17] that above-mentioned ensemble classification algorithms often outperforms a single classification methods (like C4.5), especially in classifying gene expression data. On the other side, ensemble methods may be rather "black-box" solutions that do not extract simple decision rules. Accuracy results for most of the classifiers are higher from the original TSP algorithm, but they are lower to their extensions. TSPDT solutions also have higher average credibility to those methods.

Table 4 Comparison of TSP decision trees accuracy with original methods. The highest classifiers accuracy for each dataset was bolded

	Classifiers accuracy				
Datasets	J48	Adaboost	Bagging	Random Forest	JRip
1. BC	52.63	57.89	63.15	68.42	73.68
2. CNS	56.66	75.00	71.66	73.33	65.00
3. CT	85.48	79.03	79.03	77.41	74.19
4. DF	79.22	90.90	85.71	88.31	77.92
5. LA	91.17	91.17	94.11	82.35	94.11
6. LCB	81.87	81.80	82.55	93.28	95.97
7. LCM	98.95	96.87	97.91	98.95	93.75
8. LCT	58.97	69.23	61.53	66.66	64.10
9. OC	97.23	99.20	97.62	98.41	98.81
10. PC	29.41	41.17	41.17	29.41	32.35
11. PCO	42.85	47.61	61.90	71.42	42.85
Average	70.40	75.44	76.03	77.09	73.88

4 Conclusion

Performed experiments suggest that proposed hybrid solution may successfully compete with decision trees and popular TSP algorithms for solving classification problems of microarray data. Results suggest that TSPDT may improve credibility and for some cases comprehensibility of the prediction model generated from TSP-family methods.

Furthermore, many possible improvements for decision trees with TSP algorithm still exist. One of the interesting field of endeavor is the tree pruning and the number of gene pairs tested in a single tree node. The problem of overfitting the tree model to the datasets (like in case colon tumor) is still not fully resolved, and we believe that the improvement on this field would reduce the tree size and increase classifier accuracy.

Acknowledgements This work was supported by the grant W/WI/5/08 from Białystok Technical University.

References

1. Alon, U., Barkai, N.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the USA, 96(12):6745–6750 (1999)
2. Bittner, M., Meltzer, P.: Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature, 406:536–540 (2000)
3. Breiman, L., Friedman, J.: Classification and Regression Trees, Wadsworth International Group, Belmont, CA, USA (1984)
4. Breiman, L.: Bagging predictors. Machine Learning, 24(2):123–140 (1996)
5. Breiman, L.: Random forests. Machine Learning, 45(1):5–32 (2001)
6. Brown, P.O., Botstein, D.: Exploring the new world of the genome with DNA microarrays. Nature Genetics, 21:33–37 (1999)
7. Cho, H.S., Kim, T.S.: cDNA microarray data based classification of cancers using neural networks and genetic algorithms. Nanotechnology, 1:28–31 (2003)
8. Czajkowski, M., Krętowski, M.: Novel extension of k-TSP algorithm for micro-array classification. Lecture Notes in Artificial Intelligence, 5027:456–465 (2008)
9. Cohen, W.W.: Fast Effective Rule Induction, Twelfth International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, USA, 115–123 (1995)
10. Dhanasekaran, S.M.: Delineation of prognostic biomarkers in prostate cancer. Nature, 412:822–826 (2001)
11. Dudoit, S.J., Fridlyand, J.: Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association, 97:77–87 (2002)
12. Duggan, D.J., Bittner, M.: Expression profiling using cDNA microarrays. Nature Genetics, 21(suppl 1):10–14 (1999)
13. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm, Thirteenth International Conference on Machine Learning, San Francisco, CA, USA, 148–156 (1996)
14. Geman, D., dAvignon, C.: Classifying gene expression profiles from pairwise mRNA comparisons. Statistical Applications in Genetics and Molecular Biology, 3(1):19 (2007)
15. Grześ, M., Krętowski, M.: Decision tree approach to microarray data analysis. Biocybernetics and Biomedical Engineering, 27(3):29–42 (2007)

16. Hastie, T., Tibshirani, R.: The Elements of Statistical Learning. Springer, New York (2001)
17. Hu, H., Li, J.: A Maximally Diversified Multiple Decision Tree Algorithm for Microarray Data Classification, Workshop on Intelligent Systems for Bioinformatics, Hobart, Australia (2006)
18. Jinyan. L., Huiqing, L.: Ensembles of cascading trees, Proceedings of the Third IEEE International Conference on Data Mining, 585–588 (2003)
19. Kent Ridge Bio-medical Dataset Repository: http://datam.i2r.a-star.edu.sg/datasets/index.html
20. Kononenko, I.: Estimating Attributes: Analysis and Extensions of RELIEF. In: European Conference on Machine Learning, Catania, Italy, 171–182 (1994)
21. Lockhart, D.J., Winzeler, E.A.: Genomics, gene expression and DNA arrays. Nature, 405:827–836 (2000)
22. Lu, Y., Han, J.: Cancer classification using gene expression data. Information Systems, 28(4):243–268 (2003)
23. Mao, Y., Zhou, X.: Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection. Journal of Biomedicine and Biotechnology, 2:160–171 (2005)
24. Murthy, S.: Automatic construction of decision trees from data: A multi-disciplinary survey. Data Mining and Knowledge Discovery, 2:345–389 (1998)
25. Nelson, P.S.: Predicting prostate cancer behavior using transcript profiles. Journal of Urology, 172:28–32 (2004)
26. Rokach, L., Maimon, O.: Top-down induction of decision trees classifiers - A survey. IEEE Transactions on Systems, Man, and Cybernetics - Part C, 35(4):476–487 (2005)
27. Sebastiani, P., Gussoni, E.: Statistical challenges in functional genomics. Statistical Science, 18(1):33–70 (2003)
28. Simon, R., Radmacher, M.D.: Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. Journal of the National Cancer Institute, 95:14–18 (2003)
29. Tan, A.C., Gilbert, D.: Ensemble machine learning on gene expression data for cancer classification. Applied Bioinformatics, 2:75–83 (2003)
30. Tan, A.C., Naiman, D.Q.: Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics, 21:3896–3904 (2005)
31. Quinlan, R.: Inductive knowledge acquisition: A case study. Addison-Wesley, Boston, MA, USA, chapt. 9, 157–173 (1987)
32. Quinlan, R.: C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, USA (1993)
33. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco, CA, USA (2005)
34. Valentini, G., Muselli, M.: Bagged Ensembles of SVMs for Gene Expression Data Analysis, International Joint Conference on Neural Networks 2003, Portland, OR, USA (2003)
35. Veer, L. J., Dai, H.: Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415:530–536 (2002)
36. Xu, L., Tan, A.C.: Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. Bioinformatics, 21(20):3905–3911 (2005)
37. Yoon, S., Kim, S.: k-Top scoring pair algorithm for feature selection in SVM with applications to microarray data classification. Soft Computing - A Fusion of Foundations, Methodologies and Applications, 14(2):151–159 (2009)
38. Zhang. H., Yu, C.Y.: Recursive partitioning for tumor classification with gene expression microarray data. Proceedings of the National Academy of Sciences of the USA, 98(12):6730–6735 (2001)
39. Zhang, C., Li, P.: Parallelization of multicategory support vector machines (PMC-SVM) from classifying microarray data. BMC Bioinformatics, 7(Suppl 4):S15 (2006)