Applying Decision Trees to Gene Expression Data from DNA Microarrays: A Leukemia Case Study*

Oscar Picchi Netto[1,2], **Sérgio Ricardo Nozawa**[3]
Rafael Andrés Rosales Mitrowsky[1], **Alessandra Alaniz Macedo**[1],
José Augusto Baranauskas[1]

[1]Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto
Universidade de São Paulo
Av. Bandeirantes, 3900 - Ribeirão Preto, SP - 14040-901 - Brazil

[2]Faculdade de Medicina de Ribeirão Preto
Universidade de São Paulo
Av. Bandeirantes, 3900 - Ribeirão Preto, SP - 14040-901 - Brazil

[3]Laboratório de Genética Química
Centro Universitário Nilton Lins
Av. Prof. Nilton Lins, 3259 - Manaus, AM - 69058-030 - Brazil

apnetto@aluno.ffclrp.usp.br, snozawa@niltonlins.br

rrosales@ffclrp.usp.br, ale.alaniz@usp.br, augusto@usp.br

Abstract. *Analyzing gene expression data is a challenging task since the large number of features against the shortage of available examples can be prone to overfitting. In order to avoid this pitfall and achieve high performance, some approaches construct complex classifiers, using new or well-established strategies. The main objective of this communication is to construct classifiers that can be human readable as well as robust in performance in microarray data using decision trees. Using one well-known leukemia dataset, a publicly available gene expression classification problem, we show the feasibility of decision trees on microarray data. Summarizing our results, we have obtained simple decision trees with performance comparable to related work.*

1. Introduction

The DNA microarray technology allows monitoring the expression of thousands of genes simultaneously [Schena et al. 1995]. Thus, it can lead to better understanding of many biological processes, improved diagnosis, and treatment of several diseases [Paik et al. 2006, van de Vijver et al. 2002]. However data collected by DNA microarrays are not suitable for direct human analysis, since a single experiment contains thousands of measured expression values. Several approaches have been suggested towards exploiting data mining from microarray data [Golub et al. 1999,

*This work was supported by CNPq/FAPEAM — INCT ADAPTA.

Gamberger et al. 2004, Molla et al. 2004, Dudoit et al. 2000], including supervised and unsupervised machine learning algorithms [Mitchell 1997].

In this work we have applied the supervised approach, learning from data with class labels from a dataset consisting of several gene expression values. Each example in the microarray dataset is composed by thousands gene expression measurements obtained from monitoring patients' tissue with some specific disease. Each example is also labeled accordingly the respective illness. The aim of machine learning algorithms is to start from such labelled dataset, building classifiers that can successfully classify new, previously unseen examples. Such classifiers are important because they can be used for diagnosis purposes in medicine and because they can help to understand the dependencies between classes (diseases) and features (gene expression values) [Rosenfeld et al. 2008b].

However, the problem of building classifiers from microarray experiments is dimensionality sparseness: in general, there are a large number of features (gene expression measurements) against few examples (patients monitored). In high dimensional domains like this it is well known that many induction algorithms degrade in performance — accuracy and run time. In fact many machine learning algorithms were not developed to deal with high dimensionality. Therefore it is not always straightforward using such algorithms directly in these datasets. One possible solution to this problem consists in reducing high dimensional datasets through feature selection [Blum and Langley 1997, Liu and Motoda 1998], which can provide the building of more accurate classifiers.

Besides the high dimensionality (large number of features), as mentioned before, the gene expression domain suffers from dimensionality sparseness: the high number of features contrasts with a very small number of examples. It is known from the literature that in such cases classifiers are prone to overfitting [Hastie et al. 2001] because actually weak/irrelevant features can appear to be relevant simply by chance to machine learning algorithms due the available data sample [Ein-Dor et al. 2005]. Overfitted classifiers are characterized by low specialization error but high generalization error, in other words, there is a significantly increased error on unseen examples when compared to the training set error [Domingos 1999]. One possible solution to avoid overfitting adopted by state-of-the-art researches construct complex classifiers through ensemble of classifiers. Weighted voting of informative genes is used by [Golub et al. 1999] whereas [Chow et al. 2001, Li and Wong 2002] employ support vector machines (SVM); [Ramaswamy et al. 2001] use scores of top ranked patterns. Such classifiers provide high predictive accuracy but since they include many features they are not useful for human expert interpretation.

An alternative way to see this problem consists in preserving the logical connections among features enabling the induction comprehensible classifiers by human experts, where classifiers are expressed as rules for the labels of gene expression data [Gamberger et al. 2004, Rosenfeld et al. 2008a]. In this case the aim is to build classifiers composed by rules with few conditions (typically 2-5 features). It is important to stress that these simple classifiers may have lower predictive accuracy than more complex classifiers but they explicitly emphasize the importance of the correlation among expressed and/or non-expressed genes [Baranauskas and Monard 2003]. In this study we follow this approach, aiming to induce simple classifiers (decision trees) describing important genes to foresee the target class label of unknown patients.

The remaining of this paper is organized as follows. In Section 2 basic tree induction concepts are provided, followed by the description of the microarray dataset we have used in our experiments as well as the experimental setup. Section 3 presents and discusses our experimental results, followed by a comparison with three previous works in the literature in Section 4. Finally, in Section 5 our conclusions are presented.

2. Experimental Methodology

2.1. Decision Trees

Induction of decision trees is a machine learning approach that has been applied on several tasks. Decision trees (DT) are well-suited for large real world tasks since they scale well and can represent complex concepts by constructing simple yet robust logic-based classifiers amenable to direct expert interpretation [Monard and Baranauskas 2003]. Top-down induction of decision trees algorithms [Quinlan 1993, Fayyad and Irani 1992] in general choose a feature that partitions the training data according to some evaluation function. Partitions are then recursively split until some stopping criterion is reached. After that, the decision tree is pruned in order to avoid overfitting. In our experiments we have used the algorithm J48 (with default parameters) from Weka[1] [Witten and Frank 2005], a library of several machine learning algorithms. J48 is a Java implementation of the well-known C4.5 algorithm [Quinlan 1993].

The key to the success of a decision tree learning algorithm depends on the evaluation function used to select the feature for splitting. If a feature is a strong indicator of an instance's class value, it should appear as early as possible in the tree — near the root. Most decision tree learning algorithms use a heuristic for estimating the best feature. J48 (as well as C4.5) uses a modified version of the entropy measure from information theory. For our purposes, it is sufficient to state that this measure yields a positive real number, where zero indicates an uniform set (only one class value is present for all instances of that feature) and larger values indicate a set where there is more likelihood of all class values being present (mixture of classes). In this case, the evaluation function searches for minimizing the entropy. Specifically, by default J48 uses the gain ratio metric G to choose the best feature to split.

The treatment of missing feature values — represented as '?' — in J48 follows information theory as well and receives a special treatment. Suppose that only a fraction F of some feature's value is known. In this case J48 takes this into account when computing the information gain: it considers only the fraction F of known values, whereas the gain ratio in this case G' is proportional to F, and therefore $G' < G$. Readers interested on technical details about the treatment of missing values should refer to [Quinlan 1993]. For our purposes, it is sufficient to state that features with missing values are treated differently from those without missing values by J48.

2.2. The Leukemia Dataset

In our study we have used a real world leukemia microarray experiment performed by [Golub et al. 1999]. Leukemia is a cancer of bone marrow or blood cells, i.e., a generalized neoplasic proliferation or an haematopoietic cells accumulation (cells involved

[1] www.cs.waikato.ac.nz/~ml/weka

in blood formation process) with or without peripheral blood involvement. In general, leukemias can be grouped into four categories. Myeloid and lymphoid leukemias can be acute or chronicle whereas myeloid and lymphoid both denote cell types involved. Thus, four main types of leukemias are: Acute Myeloid Leukemia (AML), Chronic Myeloid Leukemia (CML), Acute Lymphoblastic Leukemia (ALL) and Chronic Lymphoblastic Leukemia (CLL).

In the dataset provided by [Golub et al. 1999], each microarray experiment corresponds to a patient (example); each example consists 7219 genes expression values (features). Each patient has a specific disease (class label), corresponding to two kinds of leukemia (ALL and AML). There are 72 patients (47 ALL and 25 AML).

2.3. Experimental Setup

The original study of [Golub et al. 1999] split patients into two disjoint sets[2]: the training set contains 38 examples (27 ALL and 11 AML) and the test set contains 34 examples (20 ALL and 14 AML). Considering the shortage of examples it is a common technique in machine learning to use cross-validation or bootstrap [Kohavi 1995, Hastie et al. 2001] rather than isolating training and test sets. However, in order to be able to compare our results with those published by [Golub et al. 1999, Gamberger et al. 2004, Chow et al. 2001] we have also decided to use the same train/test dataset split approach used by all of them.

The original dataset provided by [Golub et al. 1999] contains nominal (discretized) and continuous values of gene expression data. Therefore we have performed experiments using both sort of values. In fact, we have defined three different datasets: two of them contain nominal (S) and continuous (S'') values both without any further data preprocessing; the third one (S') contains nominal values preprocessed as follows.

Nominal values in the original dataset consists of the discretization provided by Affymetrix, the microarray manufacturer, from continuous values. There are three nominal values: 'A': gene is absent or not expressed; 'P': gene is expressed or present and 'M': the level of the expression is marginal among 'A' and 'P'. Since 'M' is marginal, an important issue microarray experimenters should face when using decision tree induction concerns how to represent missing values: using 'M' values or replacing all of them by '?'. As explained before, using only known values (in this case, 'M') the gain ratio G is different than the gain ratio G' using missing values (in this case, '?'). This representational issue could lead to different classifiers (trees) and, in this case, to different diagnosis. Therefore in this study we have analyzed the impact of both representational forms for the leukemia data.

Let S_1 denote the original training set with 38 examples and s_1 the original test set with 34 examples, both of them containing 'A','P' and 'M' values. Dataset S'_1 is equivalent to S_1 but 'M' values were replaced by '?' and analogously for s'_1 from s_1. Datasets S''_1 and s''_1 correspond to training and test sets, respectively, both containing continuous values.

As mentioned before, analyzing gene expression data can be prone to overfitting

[2]Training and test sets can be downloaded from `http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi`.

due to the large number of features against the shortage of available examples; one way to avoid this pitfall concerns preprocessing data using feature selection. However previous work showed feature selection methods have a tendency for increasing the number of rules in the classifier when compared with the classifier generated on its own, *i.e.* without feature selection [Baranauskas et al. 1999]. For this reason we have adopted a different strategy as follows: we have also defined additional datasets $S_2, S_3, \ldots, S_{10}$ and $s_2, s_3, \ldots, s_{10}$ in the following way: using S_1 as training set, a decision tree T_1 was induced and evaluated on s_1; then, genes appearing on T_1 have been removed from both S_1 and s_1 defining a new training set S_2 and a new test set s_2, respectively. After that, using S_2 as training set, a decision tree T_2 was induced and evaluated on s_2; again, genes appearing on T_2 have been removed from both S_2 and s_2 defining a new training set S_3 and a new test set s_3, respectively and so on. Analogously for datasets $S'_2, S'_3, \ldots, S'_{10}$ and $s'_2, s'_3, \ldots, s'_{10}$ from S'_1 and s'_1, respectively. The same applies to datasets $S''_2, S''_3, \ldots, S''_{10}$ and $s''_2, s''_3, \ldots, s''_{10}$ from S''_1 and s''_1, respectively. This approach have generated 30 decision trees $T_1, T_2, \ldots, T_{10}$, $T'_1, T'_2, \ldots, T'_{10}$ and $T''_1, T''_2, \ldots, T''_{10}$. For simplicity, we denote datasets S refering to $S_1, S_2, \ldots, S_{10}$ and analogously for the other datasets and trees. Using the above strategy allows us to check the behavior of the leukemia dataset in subsequent induced trees for overfitting. Our results are presented in the next section.

3. Results & Discussion

The error rates, area under ROC curve (AUC) and rank used by Friedman-Nemenyi test [Demšar 2006] for all decision trees are presented in Table 1. On average, datasets S' (nominal values whereas 'M' has been replaced by '?') present the best performance both in terms of error rate 23.24% (rank 1.90) and AUC 0.75 (rank 1.85). Considering individual trees, the best performance was achieved by T''_3 with 5.88% error rate and AUC 0.94, followed by both T_1 and T'_1 with 8.82% error rate and AUC 0.90. It can also be noticed that performance (error rates and AUC) is not monotonically increasing/decreasing as genes were removed from training sets: for instance, trees T_7 and T_9 present better performance than T_3, T_4, T_5, T_6, T_8 and T_{10}. Similar behavior can be observed in trees T' and T''.

Even considering average rank for datasets S' is the best one, applying the non-parametric Friedman-Nemenyi test (95% confidence) we found no statistically significant differences among the three representational forms. However, comparing individually the best trees found for each representational form, *i.e.* comparing T_1 with $T_2, \ldots, T_{10}$ indicates T_1 is significantly better than all of them, except T_2. Analogously, T'_1 is significantly better than all others, except T'_2 and T'_7. Finally, T''_3 is significantly better than all T'', except T''_1.

Table 2 presents a brief description of genes appearing in decision trees T_1, T'_1 and T''_3 showed in Figure 1. Gene ID numbers have been obtained from NCBI[3]. We have found evidences in the research literature (last column of this table) that these genes are related not only to cancer data in general, but also in leukemia cases, especially concerning the AML-ALL problem. The gene appearing in the root of both trees T_1 and T'_1, M84526_at, was chosen by [Dobra 2008, Schachtner et al. 2007] as one of the most significant gene related to leukemia as well to the AML-ALL problem. The other gene, L05424-cds2_at,

[3] www.ncbi.nlm.nih.gov

Table 1. Error rates and area under ROC curve (AUC). The ranks in the parentheses are used in computation of the Friedman test

Tree	Error %	AUC	Tree	Error %	AUC	Tree	Error %	AUC
T_1	8.82 (2.0)	0.90 (2.5)	T_1'	8.82 (2.0)	0.90 (2.5)	T_1''	8.82 (2.0)	0.91 (1.0)
T_2	14.71 (1.5)	0.84 (2.0)	T_2'	14.71 (1.5)	0.86 (1.0)	T_2''	29.41 (3.0)	0.68 (3.0)
T_3	29.41 (3.0)	0.73 (3.0)	T_3'	26.47 (2.0)	0.74 (2.0)	T_3''	5.88 (1.0)	0.94 (1.0)
T_4	26.47 (2.5)	0.73 (2.5)	T_4'	26.47 (2.5)	0.73 (2.5)	T_4''	20.59 (1.0)	0.76 (1.0)
T_5	23.53 (1.5)	0.74 (1.5)	T_5'	23.53 (1.5)	0.74 (1.5)	T_5''	38.24 (3.0)	0.61 (3.0)
T_6	35.29 (1.5)	0.60 (1.0)	T_6'	35.29 (1.5)	0.59 (2.0)	T_6''	52.94 (3.0)	0.46 (3.0)
T_7	20.59 (3.0)	0.73 (3.0)	T_7'	17.65 (2.0)	0.80 (2.0)	T_7''	11.76 (1.0)	0.88 (1.0)
T_8	32.35 (1.5)	0.66 (1.5)	T_8'	32.35 (1.5)	0.66 (1.5)	T_8''	35.29 (3.0)	0.59 (3.0)
T_9	20.59 (1.5)	0.77 (1.5)	T_9'	20.59 (1.5)	0.77 (1.5)	T_9''	32.35 (3.0)	0.61 (3.0)
T_{10}	23.53 (2.0)	0.55 (3.0)	T_{10}'	26.47 (3.0)	0.71 (2.0)	T_{10}''	11.76 (1.0)	0.88 (1.0)
Average	23.53 (2.0)	0.73(2.15)		23.24 (1.90)	0.75 (1.85)		24.71 (2.10)	0.73 (2.0)

Table 2. Genes appearing on trees T_1, T_1' and T_3''

Gene ID	Description	References
M84526_at	Human adipsin/complement factor D	[Dobra 2008] [Schachtner et al. 2007]
L05424-cds2_at	Human cell surface glycoprotein CD44	[Screaton et al. 1992] [Krause et al. 2006]
CST3	Cystatin C (amyloid angiopathy and cerebral hemorrhage)	[Tang et al. 2009] [Sun et al. 2004]

does not appear directly in the literature, but the surface glycoprotein CD44 created by this gene appeared in many articles although the role of CD44 in neoplasia is less well defined; metastatic potential can be conferred on non-metastasis cell lines by transfection with a variant of CD44 and high levels of CD44 are associated with several types of malignant tumors. Finally the CST3 appears in [Tang et al. 2009] as one of the most significant genes in the leukemia data identified by their study; it is related with acute lymphoblastic leukemia B-cell (ALLB). CTS3 appears frequently in the leukemia associated literature, for instance, [Sun et al. 2004] report CST3 as one from 163 genes related with leukemia, being one from 13 genes identified in their results with respect to AML.

4. Comparison with Previous Works

Next we compare our results (T_1, T_1' and T_3'') with those from [Golub et al. 1999, Chow et al. 2001, Gamberger et al. 2004]. Since there are differences in the presentation of the performance in the mentioned literature, we have followed the same approach adopted by [Gamberger et al. 2004] in order to convert them into a unified error rate, defined as the proportion of incorrectly classified testing instances among all testing instances. Besides that, we also observe the number of genes (features) used in the individual classifiers. Results are reported in Table 3.

To compute the error rate of the voting approach [Golub et al. 1999], we consider that their predictor provided a class decision in 29 of the 34 testing examples and this decision was always correct (no error). For the five undecided examples, we consider the error rate of the majority vote on the test set ($14/34 = 41.18\%$) and therefore the overall

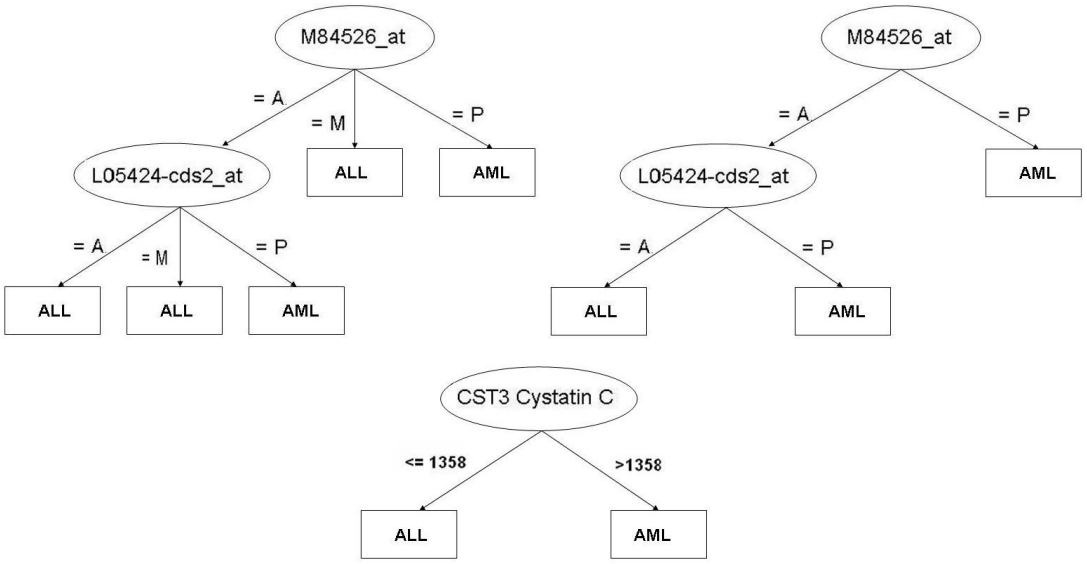

Figure 1. Decision Trees T_1 (upper left), T_1' (upper right) and T_3'' (center bottom)

error rate is $\frac{29}{34}0\% + \frac{5}{34}41.18\% = 6.06\%$.

Table 3. Comparing error rates

Leukemia	Classifier	Error %	No. of genes
AML	DT(nominal)	8.82	2
	DT(continuous)	5.88	1
	SD [Gamberger et al. 2004]	20.59	2
	SVM I [Chow et al. 2001]	11.76	50
	Voting [Golub et al. 1999]	6.06	50
ALL	DT(nominal)	8.82	2
	DT(continuous)	5.88	1
	SD [Gamberger et al. 2004]	5.88	2
	SVM II [Chow et al. 2001]	5.88	50
	SVM III [Chow et al. 2001]	2.94	50

The subgroup discovery approach [Gamberger et al. 2004] found two rules, each obtained by viewing one of the classes as the target class. To be able finding the error rate in this situation, we followed the same principle adopted by their authors: viewing each of the two rules as an individual binary classifier, interpreted under the closed-world assumption. Under this case, if the AML rule antecedent is not satisfied, then ALL is considered as the predicted class (the inverse is applied for the ALL rule). Thus, each of the two rules is then assigned its own error rate. Rule for class AML presents $\frac{7}{34} = 20.59\%$ error rate and rule for class ALL presents $\frac{2}{34} = 5.88\%$ error rate.

The SVMs approach [Chow et al. 2001] provides a binary decision for all examples in the test set. Thus, error rates are computed using the provided counts of incorrect classifications, yielding error rates of $\frac{4}{34} = 11.76\%$, $\frac{2}{34} = 5.88\%$ and $\frac{1}{34} = 2.94\%$ for SVM I, II and III, respectively.

Finally, since decision tree provide classifiers for both classes and there are no undecided examples the error rates for both T_1 and T_1' are $\frac{20}{34}\frac{1}{20} + \frac{14}{34}\frac{2}{14}\% = 8.82\%$. The error rate for T_3'' is $\frac{20}{34}\frac{1}{20} + \frac{14}{34}\frac{1}{14}\% = 5.88\%$.

As can be seen decision trees using nominal values for AML are better than SD and SVM I, but worst than voting. However, the voting scheme from [Golub et al. 1999] needs 50 genes whereas DT uses only 2. Considering ALL, DT using nominal values presents larger error rates than all the others but, again, DT uses only 2 genes as well as SD. In contrast, DT using continuous values presented better ou equal performance than all methods, except for SVM III. Due to differences on computing values in this table, it was not possible to perform a significance test in order to compare results statistically.

5. Conclusions

For human analysis it is important the induction of simple logic-based classifiers, i.e. in the form of small trees describing the target concept. However these simple decision trees may be of a lower predictive quality than more complex classifiers. In extended experiments (not shown here), induced decision trees include 2-4 gene expression features, in contrast to markers obtained from voting schemes or SVM using 50 genes. In this case, decision trees can explicitly stress the importance of the correlation of the expressed/non-expressed genes. Contrary to SD, DT provide disjoint rules which can be more interesting to represent the extracted knowledge from de expert point of view [Fugimoto et al. 2009]. Considering the three representational form for microarray data, we found no significant difference among them for decision trees on the leukemia dataset.

References

[Baranauskas and Monard 2003] Baranauskas, J. A. and Monard, M. C. (2003). Combining symbolic classifiers from multiple inducers. *Knowledge-Based Systems*, 16(3):129–136.

[Baranauskas et al. 1999] Baranauskas, J. A., Monard, M. C., and Horst, P. S. (1999). Evaluation of CN2 induced rules using feature selection. In *Proceedings of the Argentine Symposium on Artificial Intelligence (ASAI/JAIIO/SADIO)*, pages 141–154, Buenos Aires, Argentine.

[Blum and Langley 1997] Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1–2):245–271.

[Chow et al. 2001] Chow, M., Moler, E., and Mian, I. (2001). Identifying marker genes in transcription profile data using a mixture of feature relevance experts. *Physiol. Genomics*, 5:99–111.

[Demšar 2006] Demšar, J. (2006). Statistical comparison of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30.

[Dobra 2008] Dobra, A. (2008). Dependency networks for genome-wide data. Technical Report 547, Department of Statistics, University of Washington.

[Domingos 1999] Domingos, P. (1999). The role of occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3:409–425.

[Dudoit et al. 2000] Dudoit, S., Fridlyand, J., and Speed, T. (2000). Comparison of discrimination methods for the classification of tumors using gene expression data. Technical report, University of California, Berkeley.

[Ein-Dor et al. 2005] Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178.

[Fayyad and Irani 1992] Fayyad, U. M. and Irani, K. B. (1992). The attribute-selection problem in decision tree generation. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 104–110, Menlo Park, CA. American Association for Artificial Intelligence.

[Fugimoto et al. 2009] Fugimoto, P. M., Sales, L. D. F., Pereira Júnior, G. A., Passos, A. D. C., Alves, D., and Baranauskas, J. A. (2009). Análise comparativa entre Árvores de decisão e TRISS na predição de sobrevida de pacientes traumatizados. In *IV Congresso da Academia Trinacional de Ciências*, page 10 p., Foz do Iguaçu, PR.

[Gamberger et al. 2004] Gamberger, D., Lavrač, N., Zelezny, F., and Tolar, J. (2004). Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. *Journal of Biomedical Informatics*, 37:269–284.

[Golub et al. 1999] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.

[Hastie et al. 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning, data mining, inference and prediction.* Berlin: Springer.

[Kohavi 1995] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, pages 1137–1145.

[Krause et al. 2006] Krause, D. S., Lazarides, K., von Andrian, U. H., and Etten, R. (2006). Requirement for CD44 in homing and engraftment of BCR-ABL-expressing leukemic stem cells. *Nature Medicine*, 12(10):1175–1180.

[Li and Wong 2002] Li, J. and Wong, L. (2002). Geography of differences between two classes of data. In *PKDD '02: Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 325–337, London, UK. Springer-Verlag.

[Liu and Motoda 1998] Liu, H. and Motoda, H., editors (1998). *Feature Extraction, Construction and Selection: A Data Mining Perspective.* Kluwer Academic Publishers.

[Mitchell 1997] Mitchell, T. M. (1997). *Machine Learning.* McGraw–Hill.

[Molla et al. 2004] Molla, M., Waddell, M., Page, D., and Shavlik, J. (2004). Using machine learning to design and interpret gene-expression microarrays. *AI Mag.*, 25(1):23–44.

[Monard and Baranauskas 2003] Monard, M. C. and Baranauskas, J. A. (2003). *Indução de Regras e Árvores de Decisão*, chapter 5, pages 115–139. In [Rezende 2003].

[Paik et al. 2006] Paik, S., Tang, G., Shak, S., Kim, C., Baker, J., Kim, W., Cronin, M., Baehner, F. L., Watson, D., Bryant, J., Costantino, J. P., Geyer, Charles E., J., Wickerham, D. L., and Wolmark, N. (2006). Gene Expression and Benefit of Chemotherapy in

Women With Node-Negative, Estrogen Receptor-Positive Breast Cancer. *J Clin Oncol*, 24(23):3726–3734.

[Quinlan 1993] Quinlan, J. R. (1993). C4.5: *Programs for Machine Learning*. Morgan Kaufmann. San Francisco, CA.

[Ramaswamy et al. 2001] Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Gerald, W., Loda, M., Lander, E. S., and Golub, T. R. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*, 98(26):15149–15154.

[Rezende 2003] Rezende, S. O., editor (2003). *Sistemas Inteligentes - Fundamentos e Aplicações*. Manole.

[Rosenfeld et al. 2008a] Rosenfeld, N., Aharonov, R., Meiri, E., Rosenwald, S., Spector, Y., Zepeniuk, M., Benjamin, H., Shabes, N., Tabak, S., Levy, A., et al. (2008a). MicroRNAs accurately identify cancer tissue origin. *Nature biotechnology*, 26(4):462–469.

[Rosenfeld et al. 2008b] Rosenfeld, N., Aharonov, R., Meiri, E., Rosewalt, S., and Spector, Y. (2008b). MicroRNAs accurately identify cancer tissue origin. *Nature Biotechnology*, 26(4):462–469.

[Schachtner et al. 2007] Schachtner, R., Lutter, D., Theis, F., Lang, E., Tomé, A., Saez, J. G., and Puntonet, C. (2007). *Blind Matrix Decomposition Techniques to Identify Marker Genes from Microarrays*. Springer Berlin / Heidelberg.

[Schena et al. 1995] Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470.

[Screaton et al. 1992] Screaton, G., Bell, M., Jackson, D., Cornelis, F., Gerth, U., and Bell, J. (1992). Genomic structure of dna encoding the lymphocyte homing receptor cd44 reveals at least 12 alternatively spliced exons. *Proc. Natl. Acad. Sci.*, 89:12160–12164.

[Sun et al. 2004] Sun, Y., Dong, L.-J., Tian, F., Wang, S.-Q., Jia, Z.-L., and et al. (2004). Identification of acute leukemia-specific genes from leukemia recipient/sibling donor pairs by distinguishing study with oligonucleotide microarrays. *Journla Of Experimental Hematology*, 12:450–454.

[Tang et al. 2009] Tang, L.-J., Jiang, J.-H., Wu, H.-L., Shen, G.-L., and Yu, R.-Q. (2009). Variable selection using probability density function similarity for support vector machine classification of high-dimensional microarray data. *Talanta*, 79(2):260 – 267.

[van de Vijver et al. 2002] van de Vijver, M. J., He, Y. D., van 't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002). A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *N Engl J Med*, 347(25):1999–2009.

[Witten and Frank 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann.