Noname manuscript No.
(will be inserted by the editor)

Classification of Cancer Data: Analyzing Gene Expression Data using a Fuzzy Decision Tree Algorithm

Simone A. Ludwig · Stjepan Picek · Domagoj Jakobovic

Received: date / Accepted: date

Abstract Decision tree algorithms are very popular in the area of data mining since the algorithms have a simple inference mechanism and provide a comprehensible way to represent the model. Over the past years, fuzzy decision tree algorithms have been proposed in order to handle the uncertainty in the data. Fuzzy decision tree algorithms have shown to outperform classical decision tree algorithms. This chapter investigates a fuzzy decision tree algorithm applied to the classification of gene expression data. The fuzzy decision tree algorithm is compared to a classical decision tree algorithm as well as other well-known data mining algorithms commonly applied to classification tasks. Based on the five data sets analyzed, the fuzzy decision tree algorithm outperforms the classical decision tree algorithm. However, compared to other commonly used classification algorithms, both decision tree algorithms are competitive, but they do not reach the accuracy values of the best performing classifier. One of the advantages of decision tree models including the fuzzy decision tree algorithm is however the simplicity and comprehensibility of the model as demonstrated in the chapter.

Keywords medical data sets · fuzzy decision tree · soft discretization · cancer data

Simone A. Ludwig
Department of Computer Science
North Dakota State University
Fargo, ND, USA
E-mail: simone.ludwig@ndsu.edu

Stjepan Picek
KU Leuven, ESAT/COSIC and iMinds
Kasteelpark Arenberg 10, bus 2452, B-3001 Leuven-Heverlee, Belgium
E-mail: stjepan@computer.org

Domagoj Jakobovic
Faculty of Electrical Engineering and Computing
University of Zagreb, Croatia
E-mail: domagoj.jakobovic@fer.hr

1 Introduction

Data mining is the process of extracting useful information from the knowledge that is hidden in large volumes of data. The aim in data mining is to find patterns and relationships of data using data analysis tools and other techniques to build models. There are two distinct models in data mining: predictive models and descriptive models. The predictive models use data with known outcomes to develop a model that is then used to explicitly predict the different outcomes. The other model is the descriptive model, which is used to describe patterns in existing data. Both types of models provide an abstract representation of the data, which can then guide in the understanding of the data analyzed.

Data mining techniques have proved to be indispensable when working with large sets of data. The data mining community has been active in research of various techniques as well as new applications of data mining for more than 50 years. Naturally, during that time a plethora of techniques was designed to deal with various scenarios where one well known methodology is based on decision trees. We can trace the roots of its popularity to the fact that such methods can easily be interpreted by humans and the extracted knowledge can be clearly presented and visualized (Breiman, Friedman, Olshen, & Stone, 1984). However, often we encounter problems where decision trees need to have a strict division between feature values in data sets. In order to deal with that, Fuzzy Decision Tree (FDT) algorithms emerged (R. L. Chang & Pavlidis, 1977). This chapter investigates the improvements in classification accuracy that fuzzy decision trees may exhibit compared to classical decision tree algorithms.

When discussing the areas where data mining techniques play an important role, the biomedical domain is doubtless a prominent one. Here, the data can be various measurements taken from patients (e.g. heart rhythm or electrocardiogram) or the genes themselves. In order to query the expression of a multitude of genes, gene expression profiling is used. It presents the measurement of the activity of a large number of genes at once in order to be able to verify the cellular function. When the focus is on cancer data sets, gene expression profiling is used to more accurately classify tumors. Besides classifying tumors, with more powerful gene expression techniques it is also possible to classify tumor subclasses.

The objective of these methods is to discover not only a single association but several associations of genes. For this purpose, many features must be considered, with typically very few of them being significant for any given classification. Additionally, relatively few data points are available for learning.

Although very popular in practice, classical decision trees share some disadvantages that are revealed under these conditions. Specifically, their performance tends to deteriorate with the increase of features and emergence of complex interactions. Since most decision trees divide the search space into mutually exclusive regions, often the resulting tree must include several copies of the same subtree to accurately represent the data. Furthermore, their greedy

behavior is prone to over-fitting to the training set, as well as irrelevant features and noise.

In contrast to that, fuzzy decision trees do not need to assign a data instance with a single branch, but may simultaneously assign more branches to the same instance with a gradual certainty. In this way, fuzzy decision trees retain the symbolic tree structure, but are able to represent concepts by producing continuous classification outputs with gradual transitions between classes.

In this work, we experiment with a fuzzy decision tree algorithm with the goal of analyzing gene expression cancer data. Besides the comparison with a decision tree algorithm, we also compare the proposed algorithm with several other well known algorithms for classification. The results present the advantages of fuzzy decision trees over classical decision trees for multiple data sets in this domain.

This chapter is an extended version of the paper published in (Ludwig, Jakobovic, & Picek, 2015), and is arranged as follows: Section 2 describes the related work. The proposed approach is introduced in Section 3. The experimental setup and results are demonstrated in Section 4. In the final section (Section 5) the conclusions of this research are discussed.

2 Related Work

We divide the relevant research into two categories; the first is concerned with fuzzy decision tree development and applications, and the second with the applications of data mining techniques in the analysis of medical data. However, since this still encompasses a huge research area, we concentrate only on a subset of papers exploring cancer data research.

The development of fuzzy variants of decision tree induction has been around for quite a while (R. L. Chang & Pavlidis, 1977; Janikow, 1998), but they become a topic of interest in recent applications. These approaches provide examples for the application of "fuzzification" to standard machine learning methods.

There are many variations of fuzzy decision trees. Soft Decision Trees (SDT) are presented in (Olaru & Wehenkel, 2003), which combine tree-growing and pruning to determine the structure and refitting and backfitting to improve the generalization capability. The authors empirically show that SDTs are more accurate than standard decision trees. In (An & Hu, 2012), the authors propose fuzzy-rough classification trees with a new measure to quantify the functional dependency of decision attributes on condition attributes within fuzzy data. The experiments show that fuzzy-rough classification trees outperform existing decision tree induction algorithms on 16 real-world datasets.

Fuzzy decision trees have been applied to various domains; in (P.-C. Chang, Fan, & Dzan, 2010) they are integrated with genetic algorithms for data classification in database applications, and in (Lai, Fan, Huang, & Chang, 2009)

for developing a financial time series-forecasting model, where they were also combined with a genetic algorithm.

In (Biswal & Dash, 2013), the authors use a FDT-based classifier for the measurement, identification, and classification of various types of power quality disturbances and they report robust performance under different noise conditions. A fuzzy knowledge-based network is developed in (Mitra, Konwar, & Pal, 2002) based on the linguistic rules extracted from a fuzzy decision tree. The effectiveness of the system, in terms of recognition scores, structure of decision tree, performance of rules, and network size, is extensively demonstrated on three sets of real-life data.

For the biomedical applications, we first enumerate several surveys on the data mining techniques and cancer data. In the scope of cancer data analysis, a survey with a comprehensive study of various cancer classification methods is given in (Lu & Han, 2003). The authors conduct an analysis of the efficiency of methods based on their speed, accuracy and ability to reveal biologically meaningful gene information. Another survey on data mining techniques and breast cancer data is given in (Padmapriya & Velmurugan, 2014). In their work, the authors discuss the algorithms ID3 and C4.5. In (Palivela, Yogish, Vijaykumar, & Patil, 2013), the authors compare several data mining techniques on breast cancer data. A survey on decision tree classifiers in gene micro array data analysis is given in (Polaka, Tom, & Borisov, 2010). A general framework of sample weighting to improve the stability of feature selection methods is proposed in (Yu, Han, & Berens, 2012).

Experimentation with a multiclass classifier based on SVM (Support Vector Machine) algorithm is reported in (Ramaswamy et al., 2001). The authors use samples of 14 common tumor types and achieve an overall classification accuracy of 78%. A method of gene selection with reliability analysis is devised in order to help differentiate between histologically similar cancers (Li & Casey, 2004). In (Cuperlovic-Culf, Belacel, & Ouellette, 2005), the question is addressed on how to correctly select diagnostic marker genes from the gene expression profiles.

New astrocytic tumor micro-array gene expression data set is experimented with using an artificial neural network algorithm (Petalidis et al., 2008). With this algorithm the authors address grading of human astrocytic tumors, derive specific transcriptional signatures from histopathologic subtypes of astrocytic tumors, and assess whether these molecular signatures define survival prognostic subclasses. Another artificial neural networks approach for classifying cancers to specific diagnostic categories based on their gene expression signatures is provided in (J. Khan et al., 2001).

DNA micro-array analysis with supervised classification has shown to identify a gene expression signature to be strongly predictive of a short interval to distant metastases for breast cancer patients (van't Veer et al., 2002). With this strategy it is possible to select the patients who would benefit from chemotherapy or hormonal therapy. The problem how to select a small subset of genes from large patterns of data recorded on DNA micro-arrays is addressed

in (Guyon, Weston, Barnhill, & Vapnik, 2002). The authors experiment with SVM algorithms based on recursive feature elimination.

Another novel method called decision trunks that is based on decision trees to classify cancer using expression data is proposed in (Ulfenborg, Klinga-Levan, & Olsson, 2013). The results suggest that the new algorithm performs at least as good as the state of the art algorithms when considering accuracy.

The use fuzzy decision trees to predict breast cancer survivability is reported in (M. U. Khan, Choi, Shin, & Kim, 2008). The authors compare decision trees and fuzzy decision trees and find FDT to be more robust and balanced than DT. A logistic regression and decision trees for survivability prognosis in patients with breast cancer is given in (Wang, Makond, & Wang, 2013). The authors show that logical regression has better statistical power in predicting five-year survivability.

In (Hamdan & Garibaldi, 2010), an adaptive fuzzy inference system technique for the estimation of survival prediction in cancer patients is proposed. Three methods, namely, decision trees, artificial neural networks, and logistic regression to develop prediction models for breast cancer survivability is given in (Delen, Walker, & Kadam, 2005). The authors found decision trees to be the predictor with the best accuracy.

3 Fuzzy Decision Tree Classifier

Supervised classification is a very important and frequently used technique that is applied in the area of medical informatics. The most commonly used classification algorithms include logic-based algorithms, neural network algorithms, statistical learning algorithms, instance-based learning algorithms, and support vector machine algorithms.

In terms of learning-based models, there are two groups: decision trees and rule-based classifiers. Decision trees classify instances by sorting them based on feature values. A decision tree classifier builds a decision tree model that can be used for the classification of unseen data. The decision tree model consists of a series of observations (branch nodes) that lead to conclusions (leaf nodes). The main difference between classical decision tree modeling and fuzzy decision tree modeling is the use of crisp or soft discretization, respectively. Classical decision tree modeling uses crisp discretization, whereby the decision space is partitioned into a set of non-overlapping subspaces using the crisp discretization method. For soft discretization, the decision space is partitioned into a set of overlapping subspaces. For both classical and fuzzy decision trees, each path from the root node to a leaf node represents a classification rule.

The algorithm of the FDT classifier starts by sorting the continuous values of a feature. It then produces a possible candidate "cut-point", and "fuzzifies" the "cut-point" by using an entropy evaluation function. This checking of the best "cut-point" is done recursively and is applied to all attributes. Once all attributes have been soft discretized, the attribute with minimum value is selected to generate two child branches and nodes. This steps repeats until

one of the stopping criteria is met. A detailed description of the algorithm can be found in (Chen & Ludwig, 2013; Ludwig et al., 2015).

In order to show the decision trees that are generated by a DT and FDT classifier, a diabetes data set (obtained from the UCI repository (Frank & Asuncion, 2010)) has been analyzed. The diabetes data set consists of 8 features, 768 instances and 2 classes. The decision trees generated by a classical DT (J48) (WEKA's J48 algorithms was used (Witten, Frank, & Hall, 2011)) and our FDT (Java implementation) are shown in Figures 1 and 2, respectively. What we can see is that both decision trees are roughly of equal complexity, but different decision trees were generated in terms of the features used.

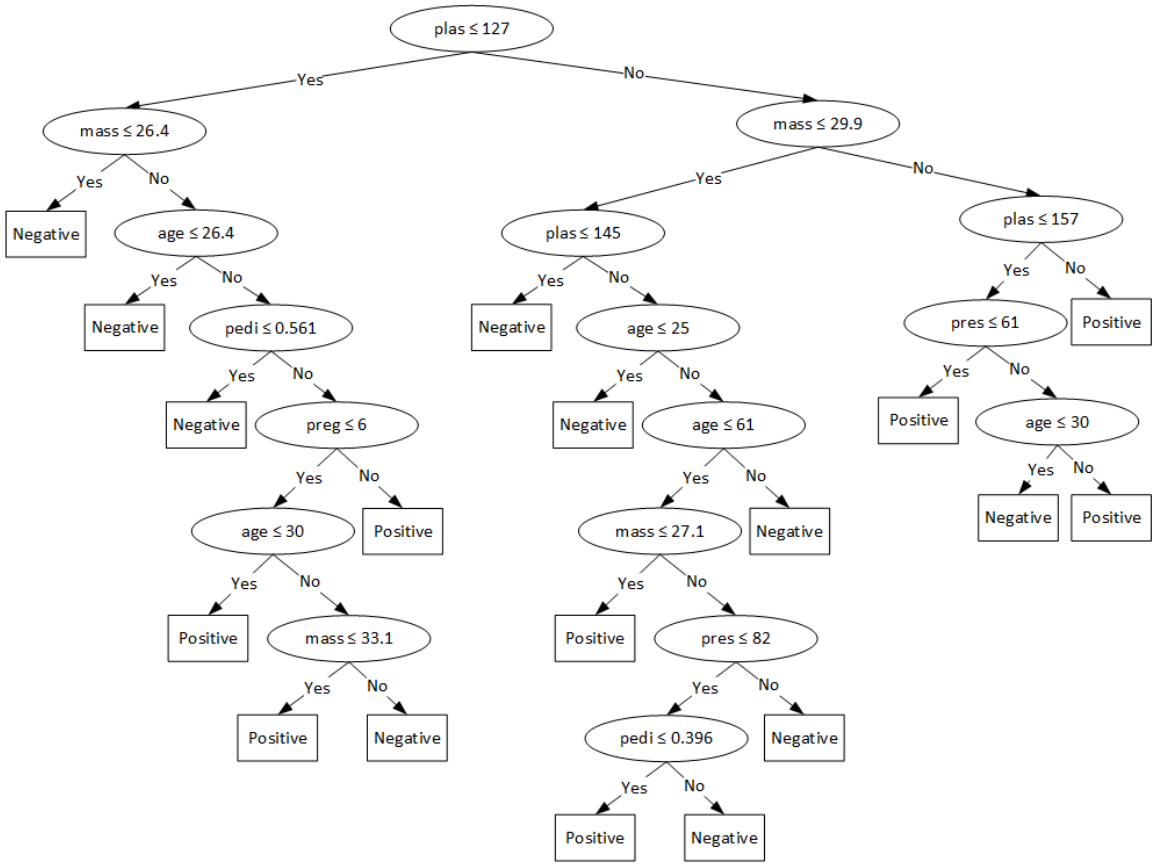

Fig. 1: Decision tree obtained from FDT classifier for the Ovarian cancer data set

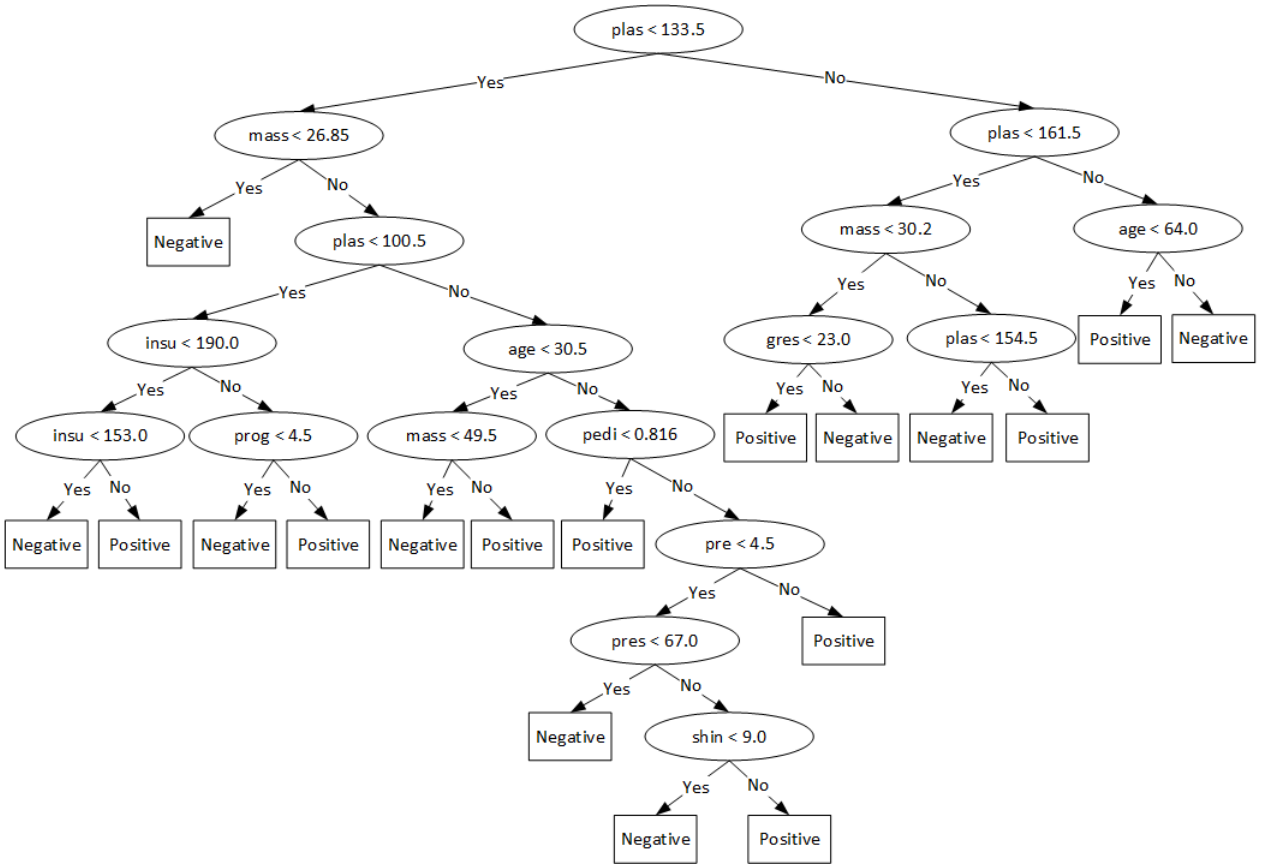

Fig. 2: Decision tree obtained from FDT classifier for the Prostate data set

Table 1: Details of binary data sets used for experiments

Data set name	# of features	# of instances	Class balance	Size	Short description	Ref.
Colon tumor	full: 2,000 reduced: 26	62	40/22	1.2 MB	Data collected from colon-cancer patients; tumor biopsies showing tumors ("negative"), and normal ("positive") biopsies are from healthy parts of colons of the same patients	(Alon et al., 1999)
Leukemia	full: 7,129 reduced: 81	72	47/25	2.2 MB	Data collected from bone marrow samples; distinction is between Acute Myeloid Leukemia ("AML"), and Acute Lymphoblastic Leukemia ("ALL") without previous knowledge of these classes	(Golub et al., 1999)
Lung cancer	full: 12,533 reduced: 160	181	150/31	12 MB	Data collected from tissue samples; classification between Malignant Pleural Mesothelioma ("MPM"), and ADenoCArcinoma ("ADCA") of the lung	(Gordon et al., 2002)
Ovarian cancer	full: 15,154 reduced: 35	253	162/91	34 MB	Data to identify proteomic patterns in serum that distinguish ovarian cancer ("cancer") from non-cancer ("normal")	(Petricoin et al., 2002)
Prostate cancer	full: 12,600 reduced: 75	136	77/59	5.5 MB	Data from prostate tumor samples, whereby the non-tumor ("normal") prostate samples, and tumor samples ("cancer") are identified using 12,600 genes	(Singh et al., 2002)

4 Experiments and Results

The FDT was implemented in Java as outlined in the previous section. The classical decision tree algorithm used for comparison is WEKA's J48 decision tree implementation (Witten et al., 2011). Other algorithms based on naive Bayes, Bayesian network, logistic regression, radial basis function neural network, and support vector machine are also used and compared with. All algorithms are further introduced in one of the following subsections.

In addition, since feature selection is a normal preprocessing step in data mining, WEKA's attribute selection method is used to filter out the relevant features. Results of both, FDT and J48, are given for the complete data set (all features) as well as the reduced feature set selected by the attribute selection method. 10-fold cross-validation was used for the training and testing of all experiments.

4.1 Data Sets

The data sets[1] that have been chosen for this investigation are listed in Table 1. All data sets contain gene data information for different types of cancer. The number of features (all numeric) for the original data set (full) as well as after feature selection is applied is also given (reduced) in the column. The number of instances and the class balance of the binary data sets are also listed. Furthermore, a short description is provided and more details can be found looking up the references listed in the last column.

4.2 Evaluation Measures

In order to evaluate the medical data sets, the following measures have been chosen based on the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN):

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}. \tag{1}$$

$$\text{Sensitivity} = \frac{TP}{TP + FP}. \tag{2}$$

$$\text{Specificity} = \frac{TN}{FP + TN}. \tag{3}$$

Another measure used to evaluate medical data sets is the Receiver Operating Characteristic (ROC) (Swets, 1996) curve, which is said to be a good indicator of the relationship between sensitivity and specificity. The AUC (Area Under the Curve) is calculated as follows:

$$\text{AUC} = \frac{1 - (1 - \text{Specificity}) + \text{Sensitivity}}{2}. \tag{4}$$

[1] http://datam.i2r.a-star.edu.sg/datasets/krbd/

4.3 Comparison Algorithms

The implemented FDT algorithm is compared with a classical decision algorithm known as J48 (Quinlan, 1993), which is implemented in WEKA. J48 is an extension of the C4.5 and the earlier ID3 algorithm (Quinlan, 1979).

The other comparison algorithms that are used for this investigation are:

- **NB**: is a Naive Bayes classifier implementation using estimator classes, whereby numeric estimator precision values are chosen based on the analysis of the training data.
- **BN:** implements a Bayes Network learning algorithm that uses various search algorithms and quality measures.
- **Log:** is a logistic regression model classifier. The classifier is based on a multinomial logistic regression model with a ridge estimator.
- **RBF:** is a radial basis function neural network model classifier. The classifier normalizes all attributes, and the initial centers for the Gaussian radial basis functions are identified using k-means.
- **SMO:** implements the sequential minimal optimization algorithm for training a support vector classifier. All missing values are replaced and nominal attributes are transformed into binary ones. In addition, all attributes are normalized by default.
- **BG:** implements the Bagging algorithm, which is an ensemble meta-algorithm that improves the accuracy and stability of learning algorithms that are used for classification and regression tasks.
- **RotF:** is the abbreviation for the Rotation forest algorithm that is a combination of decision trees with binary partitioning. Each decision tree is created based on the subset of training data with a bootstrap sample method.
- **RanF:** implements the Random forest algorithm. RanF uses a combination of decision trees with binary partitioning. Each tree is created based on training data with bootstrap sampling.

4.4 Experimental Results

Table 2 shows the accuracy, sensitivity and specificity values of the data sets using the complete feature set, i.e., using the complete data sets with all features. We can see that in terms of accuracy, the Ovarian cancer data sets achieves the highest values closely followed by the lung data set. However, comparing both data sets in terms of sensitivity and specificity reveals that the Ovarian cancer data set performs better scoring in the lower ninety percent.

Table 3 shows the same measures as Table 2, however, this time the feature set of the data sets are reduced after feature/attribute selection has been applied. We can see that the accuracy values are higher with the exception of the Lung cancer data set that scored the same accuracy. In terms of sensitivity and specificity, improved values can also be observed. Therefore, we can conclude that overall the feature reduction method improved the accuracy.

Table 2: Results of FDT measures with full feature set

Data set	Accuracy	Sensitivity	Specificity
Colon tumor	0.7746	0.8409	0.7200
Leukemia	0.8250	0.8475	0.6400
Lung cancer	0.9553	0.7879	0.9539
Ovarian cancer	0.9589	0.9175	0.9470
Prostate cancer	0.7985	0.8571	0.7885

Table 3: Results of FDT measures with reduced feature set

Data set	Accuracy	Sensitivity	Specificity
Colon tumor	0.8028	0.8864	0.7826
Leukemia	0.8750	0.8983	0.7391
Lung cancer	0.9553	0.7879	0.9540
Ovarian cancer	0.9711	0.9485	0.9662
Prostate cancer	0.8836	0.7662	0.7188

Table 4: Results of comparison of FDT and J48 with full and reduced feature set

Data set	Full feature set		Reduced feature set	
	FDT	J48	FDT	J48
Colon tumor	0.7746	**0.8226**	0.8028	**0.8710**
Leukemia	**0.8250**	0.7917	**0.8750**	0.8472
Lung cancer	**0.9553**	0.9503	0.9553	**0.9613**
Ovarian cancer	**0.9594**	0.9565	**0.9711**	0.9605
Prostate cancer	**0.7985**	0.7941	**0.8836**	0.8824

Table 4 shows the accuracy values comparing FDT with J48 as well as showing the effect of using the complete data set with all the features versus using the reduced data set. As can be seen by the values in bold, on the full data set FDT outperformed J48 four out of five times, and on the reduced data sets FDT outperformed J48 three out of five times.

Figure 3 shows the AUC values for the data set with and without feature selection. The AUC values are often used since it shows the interplay between sensitivity and specificity. As can be seen by the figure, the AUC is higher for the reduced feature data sets with the exception of the Prostate cancer data set.

Table 5 shows the comparison of FDT, J48, the naive Bayes classifier (NB), the Bayesian network algorithm (BN), the logistic regression (Log), radial basis function network (RBF), and the support vector machine algorithm (SMO).

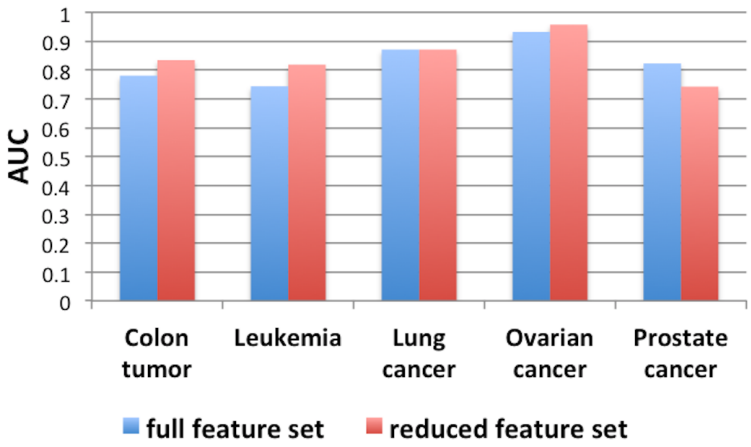

Fig. 3: Comparison of AUC values for different data sets with full and reduced feature set

Based on the five data sets, the SMO algorithm performs best out of all classifiers. It scores best 7 out of 10 times when applied to the full data sets as well as the reduced data sets. SMO is closely followed by NB and BN (both scoring best 4 times). In particular, SMO achieves 100% accuracy on the Lung cancer data set and the Ovarian cancer data set. The overall conclusions that can be drawn are that the SMO clearly outperforms all other classifiers including FDT and J48. FDT only achieves close results on the Lung and Ovarian data sets.

Table 5: Results of comparison of FDT with other WEKA algorithms in terms of accuracy

Data set		FDT	J48	NB	BN	Log	RBF	SMO	BG	RotF	RanF
Colon tumor	full	0.7746	0.8225	0.5323	0.7581	0.7097	0.7903	**0.8548**	0.7903	0.7742	0.7581
	reduced	0.8028	0.8710	0.8548	**0.9032**	0.7581	0.8710	0.8548	0.8710	0.8871	0.8226
Leukemia	full	0.8245	0.7917	**0.9861**	0.9722	0.9028	0.9306	**0.9861**	0.9028	0.9306	0.8750
	reduced	0.8750	0.8472	**1.0000**	**1.0000**	0.9583	**1.0000**	0.9861	0.8889	0.9583	0.9722
Lung cancer	full	0.9553	0.9503	0.9834	0.9834	0.9889	0.9779	**0.9945**	0.9779	0.9669	0.9834
	reduced	0.9553	0.9613	**1.0000**	**1.0000**	0.9945	0.9945	**1.0000**	0.9779	0.9890	**1.0000**
Ovarian cancer	full	0.9594	0.9565	0.9249	0.9210	0.9841	0.8340	**1.0000**	0.9723	0.9658	0.9605
	reduced	0.9711	0.9605	**1.0000**	0.9960	**1.0000**	**1.0000**	**1.0000**	0.9723	**1.0000**	0.9881
Prostate cancer	full	0.7985	0.7941	0.5588	0.6618	0.8456	0.6617	**0.9118**	0.8529	0.9044	0.7941
	reduced	0.8836	0.8824	0.6176	**0.9559**	0.7647	0.7647	0.8676	0.8676	0.9412	0.9412

Investigating the generated models of the FDT algorithm in the form of decision trees (as shown in Figures 5 - 9) as compared to the best-performing SVM (SMO) classifier reveals that only a fraction of the features are used for the model of FDT, whereas all features are used for the SMO model generation (the reduced feature set is used). This is true for all other non decision tree algorithms. Table 6 lists the number of features of the model created by SMO and other algorithms, and FDT, respectively. For example, for the Colon tumor data set only 6 as compared to 26 features are used for the model of FDT versus all others including SMO, and even a wider gap is observed for the Lung cancer data set on which FDT uses 3 features whereas SMO and others use 160 features. This demonstrates that the FDT models are much simpler in terms of complexity as well as comprehensibility. To show an example of the models created by FDT and SMO, the model of a decision tree generated by FDT is shown in Figure **??**, and the model generated by SMO on the lung cancer data set is as given in Figure 4 (output from WEKA console):

```
=== Classifier model (full training set) ===
SMO
Kernel used: Linear Kernel: K(x,y) = <x,y>
Classifier for classes: negative, positive
BinarySMO
Machine linear: showing attribute weights, not support vectors.
-0.2258 * (normalized) attribute143
+ 0.8376 * (normalized) attribute249
+ -0.237 * (normalized) attribute258
+ -0.4451 * (normalized) attribute279
+ 1.1883 * (normalized) attribute377
+ -0.1269 * (normalized) attribute467
+ -1.0661 * (normalized) attribute576
+ -0.5733 * (normalized) attribute625
+ -0.7617 * (normalized) attribute682
+ -0.5918 * (normalized) attribute763
+ 0.9659 * (normalized) attribute765
+ 0.2894 * (normalized) attribute897
+ -0.8163 * (normalized) attribute1042
+ -0.6559 * (normalized) attribute1153
+ -0.207 * (normalized) attribute1200
+ -0.1432 * (normalized) attribute1227
+ -0.5952 * (normalized) attribute1325
+ -0.0822 * (normalized) attribute1328
+ -0.529 * (normalized) attribute1412
+ 0.8739 * (normalized) attribute1423
+ 0.6139 * (normalized) attribute1560
+ -0.4861 * (normalized) attribute1562
+ 0.175 * (normalized) attribute1635
+ -0.1088 * (normalized) attribute1671
+ -0.8822 * (normalized) attribute1772
+ 0.2362 * (normalized) attribute1917
+ 0.0996
```

Fig. 4: WEKA's output of the model generated of the SMO classifier applied to the Lung cancer data set

What can be observed by the comparison of the model generated by SMO versus the decision tree model generated by FDT, is the simple and easy to visualize and understand model that is generated by the decision tree model. The mathematical formula involving all attributes as given as the SMO model is more difficult to describe and interpret. Besides SMO, the other machine learning algorithms used for comparison involve a mathematical model generation that is similar in outcome than the SMO model.

To further discuss and interpret the generated decision trees, let us look at the decision tree generated for the Lung cancer data set (see Figure **??**). The constructed decision tree is based on three decision node, namely *1394_at*, *34320_at*, and *37716_at*. Given this decision tree, a unseen example can then be routed down the tree to reach a decision node in order to present the output. For example, if a patient has the following values: *1394_at=420*, *34320_at=2100*, and *37716_at=1500*, then the output will be *Mesothelioma*. The decision tree model is very intuitive since the resulting model is easy to understand and assimilate by humans. That is the reason for its popularity in particular in the medical domain.

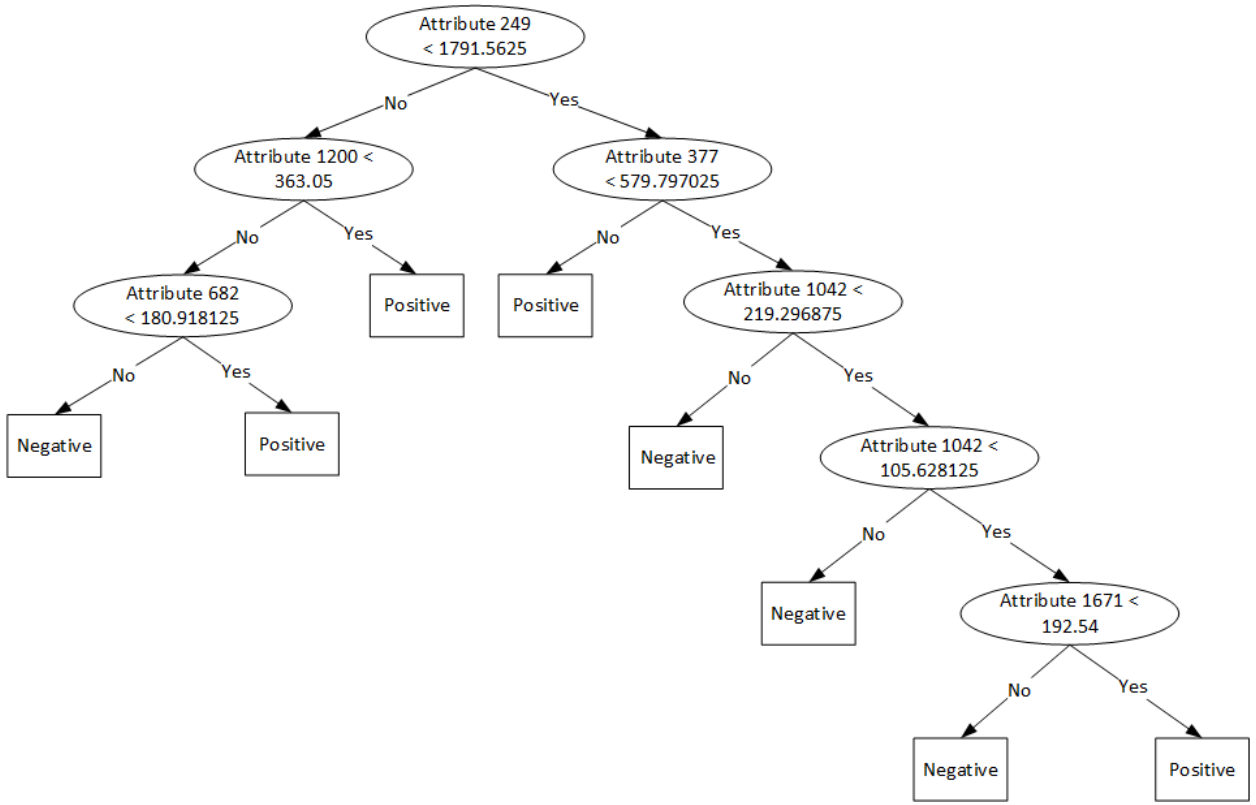

Fig. 5: Decision tree obtained from FDT classifier for the Colon tumor data set

Table 6: Comparison of features used for the generation of the model

	SMO and other algorithms	**FDT**
Colon tumor	26	6
Leukemia	81	4
Lung cancer	160	3
Ovarian cancer	35	6
Prostate cancer	75	8

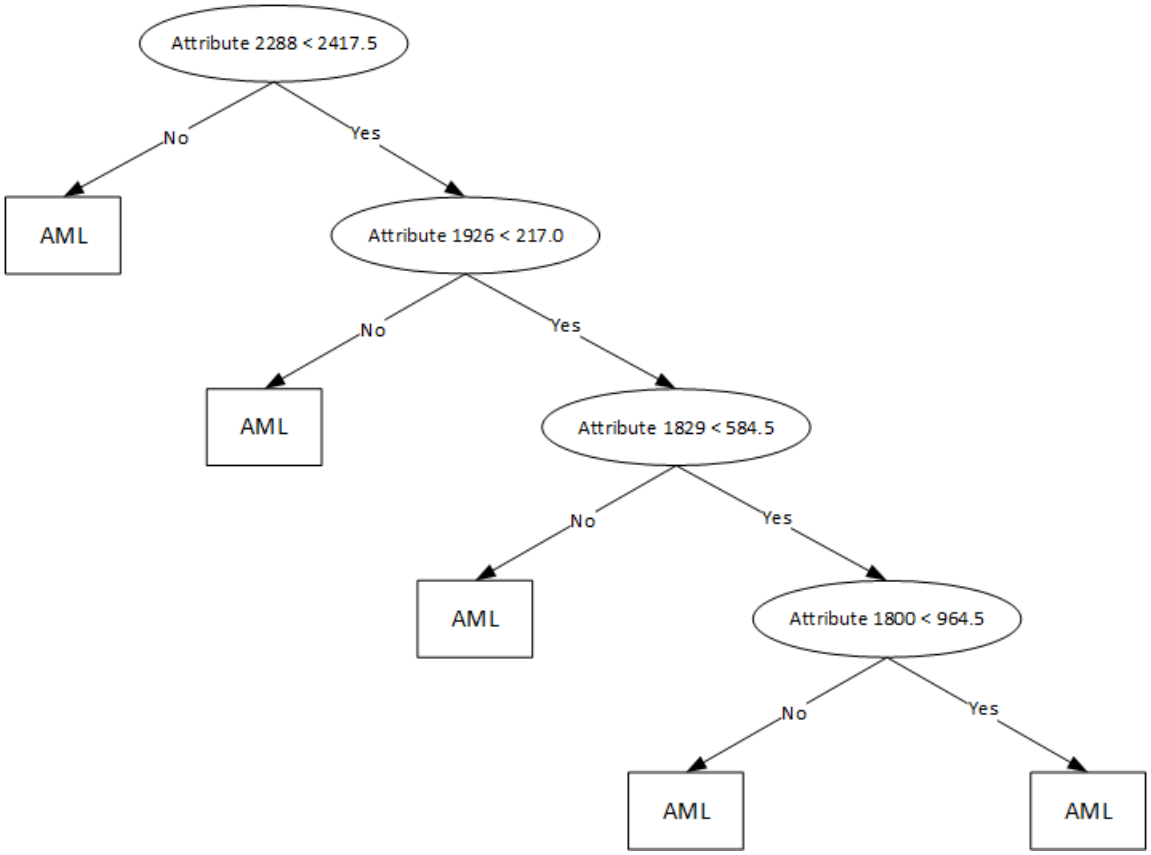

Fig. 6: Decision tree obtained from FDT classifier for the Leukemia data set

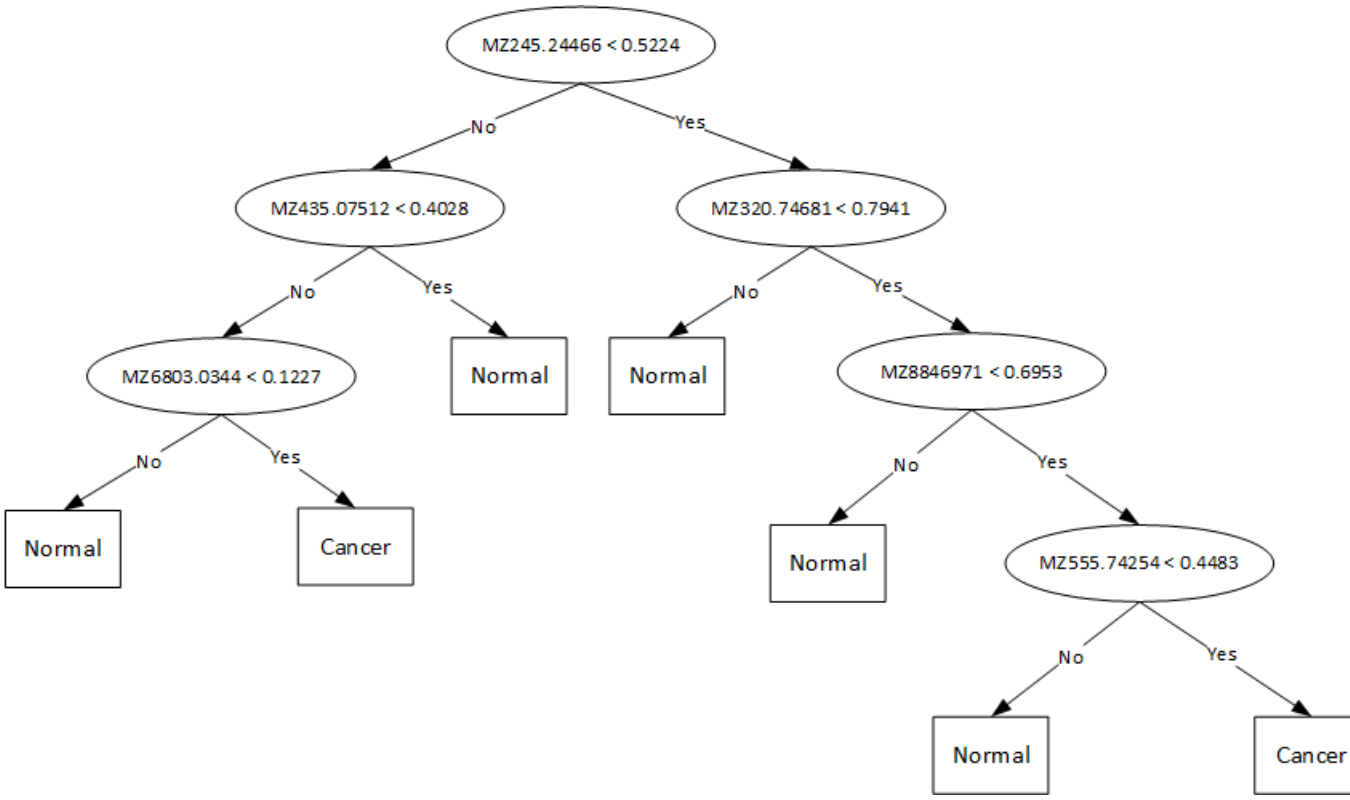

Fig. 8: Decision tree obtained from FDT classifier for the Ovarian cancer data set

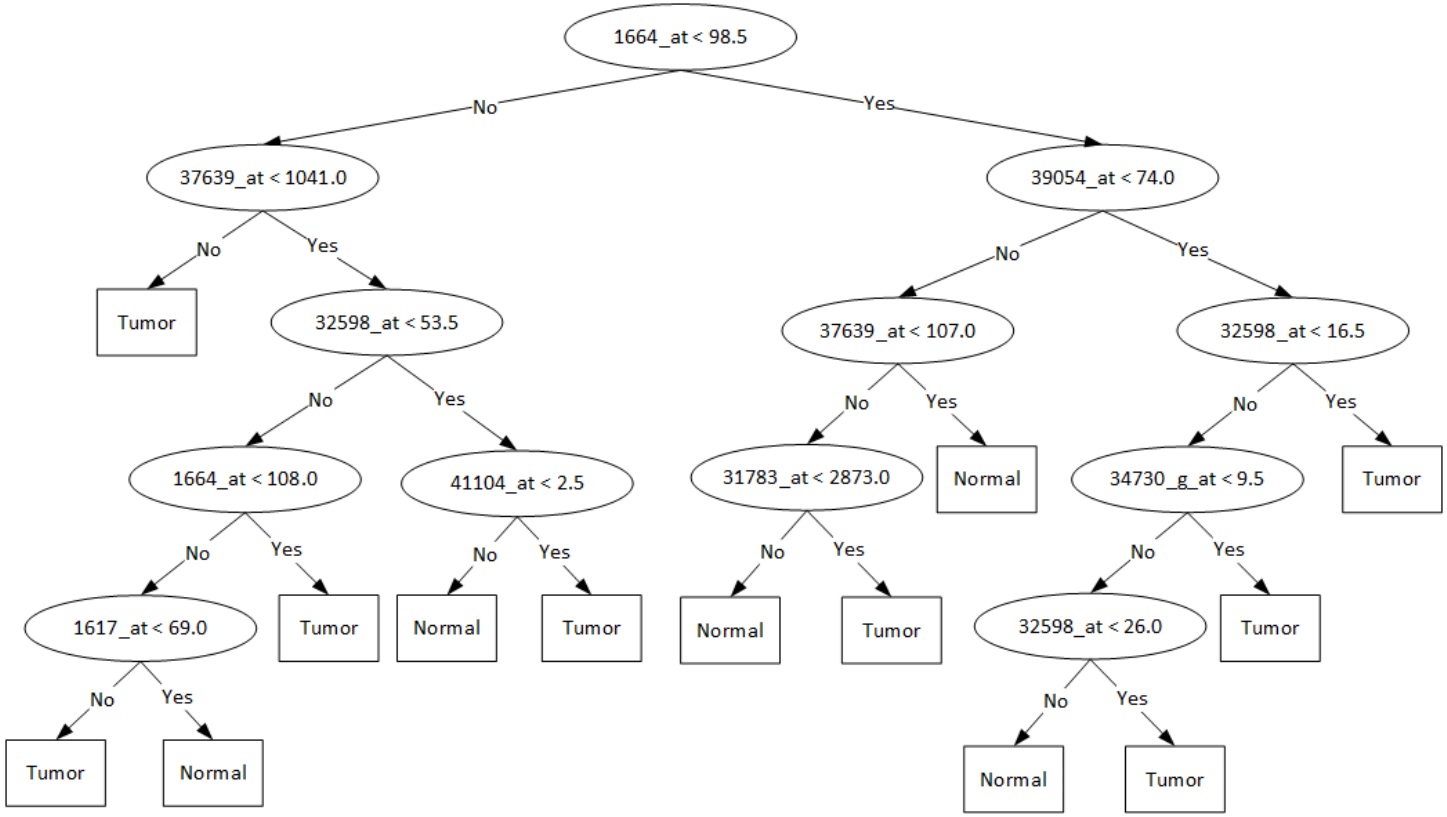

Fig. 9: Decision tree obtained from FDT classifier for the Prostate data set

5 Conclusion

This chapter investigated a fuzzy decision tree implementation applied to the classification of gene expression data. Five high-dimensionality cancer data sets were analyzed and compared with a classical decision tree algorithm as well as other well-known data mining algorithms.

The results revealed that comparing FDT with J48, the FDT algorithm outperformed J48 in terms of accuracy on four out of the five data sets when applied to the classification using the full data sets, and 3 out of 5 times when applied to the reduced data sets after feature selection was applied. In general, higher values of accuracy, sensitivity, and specificity were achieved on the preprocessed data sets as has been shown in past literature.

Other measures of sensitivity and specificity were also in favor of FDT. The AUC values for FDT were also calculated and revealed that, in general, higher AUC values are achieved when the preprocessed data sets were investigated. In addition, the data sets, both full and reduced feature set, were run with common data mining algorithms and the support vector machine algorithm outperformed all other data mining algorithms achieving 100% accuracy on some data sets. This implies that the decision tree algorithms (both FDT and J48) are not the best choice when analyzing the five gene cancer data sets when accuracy is the only concern.

Further analyzing the complexity of the resulting models comparing the overall best-performing SVM algorithm with the FDT algorithm revealed that the model of FDT is many times less complex since only a fraction of features are used for FDT as compared to SVM, which uses all features. The compactness of the resulting decision tree model of FDT as well as the comprehensibility of the model are the strengths of the decision tree algorithms including the implemented FDT algorithm.

To summarize, the benefits of the decision tree model are: (1) in-build feature selection, (2) nonlinear relationships between parameters do not affect the tree performance, and (3) easy to interpret and explain.

Future work includes the evaluation of the FDT algorithm on larger gene expression data sets once they become available. Furthermore, a possible improvement of the FDT algorithm with, for example, another algorithm such as neural networks could be investigated.

Acknowledgements The authors would like to thank Gongyi Xia for the drawing of the figures of the decision trees.

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999, June). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*, *96*(12), 6745–6750.

An, S., & Hu, Q. (2012). Fuzzy Rough Decision Trees. In J. Yao et al. (Eds.), *Rough sets and current trends in computing* (Vol. 7413, p. 397-404). Springer Berlin Heidelberg.

Biswal, M., & Dash, P. (2013, Nov). Measurement and Classification of Simultaneous Power Signal Patterns With an S-Transform Variant and Fuzzy Decision Tree. *Industrial Informatics, IEEE Transactions on*, *9*(4), 1819-1827.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Wadsworth, Belmont. In *Classification and regression trees.*

Chang, P.-C., Fan, C.-Y., & Dzan, W.-Y. (2010). A CBR-based fuzzy decision tree approach for database classification. *Expert Systems with Applications*, *37*(1), 214 - 225.

Chang, R. L., & Pavlidis, T. (1977, Jan). Fuzzy decision tree algorithms. *Systems, Man and Cybernetics, IEEE Transactions on*, *7*(1), 28-35.

Chen, M., & Ludwig, S. A. (2013). Fuzzy decision tree using soft discretization and a genetic algorithm based feature selection method. In *World congress on nature and biologically inspired computing.*

Cuperlovic-Culf, M., Belacel, N., & Ouellette, R. J. (2005). Determination of tumour marker genes from gene expression data. *Drug Discovery Today*, *10*(6), 429 - 437.

Delen, D., Walker, G., & Kadam, A. (2005, June). Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods. *Artif. Intell. Med.*, *34*(2), 113–127.

Frank, A., & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. In *University of california, school of information and computer science, irvine, ca.*

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... Lander, E. S. (1999, October). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, *286*(5439), 531-537.

Gordon, G., Jensen, R., Hsiao, L.-L., Gullans, S., Blumenstock, J., Ramaswamy, S., ... Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res*, *62*, 4963–4967.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, *46*(1-3), 389-422.

Hamdan, H., & Garibaldi, J. (2010, July). Adaptive neuro-fuzzy inference system (ANFIS) in modelling breast cancer survival. In *Fuzzy systems (fuzz), 2010 ieee international conference on* (p. 1-8).

Janikow, C. (1998, Feb). Fuzzy decision trees: issues and methods. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, *28*(1), 1-14.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., ... Meltzer, P. S. (2001, jun). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.

Nat Med, *7*(6), 673–679.

Khan, M. U., Choi, J. P., Shin, H., & Kim, M. (2008, Aug). Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare. In *Engineering in medicine and biology society, 2008. embs 2008. 30th annual international conference of the ieee* (p. 5148-5151).

Lai, R. K., Fan, C.-Y., Huang, W.-H., & Chang, P.-C. (2009). Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems with Applications*, *36*(2, Part 2), 3761 - 3773.

Li, M. F., & Casey, S. F. (2004). Multi-class cancer subtype classification based on gene expression signatures with reliability analysis. {*FEBS*} *Letters*, *561*(13), 186 - 190.

Lu, Y., & Han, J. (2003). Cancer classification using gene expression data. *Information Systems*, *28*, 243–268.

Ludwig, S. A., Jakobovic, D., & Picek, S. (2015). Analyzing Gene Expression Data: Fuzzy Decision Tree Algorithm applied to the Classification of Cancer Data. In *2015 ieee international conference on fuzzy systems.*

Mitra, S., Konwar, K., & Pal, S. (2002, Nov). Fuzzy decision tree, linguistic rules and fuzzy knowledge-based network: generation and evaluation. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *32*(4), 328-339.

Olaru, C., & Wehenkel, L. (2003). A complete fuzzy decision tree technique. *Fuzzy Sets and Systems*, *138*(2), 221 - 254.

Padmapriya, B., & Velmurugan, T. (2014, Dec). A survey on breast cancer analysis using data mining techniques. In *Computational intelligence and computing research (iccic), 2014 ieee international conference on* (p. 1-4).

Palivela, H., Yogish, H., Vijaykumar, S., & Patil, K. (2013, Feb). Survey on mining techniques for breast cancer related data. In *Information communication and embedded systems (icices), 2013 international conference on* (p. 540-546).

Petalidis, L., Oulas, A., Backlund, M., Wayland, M., Liu, L., Plant, K., ... Collins, V. (2008). Improved grading and survival prediction of human astrocytic brain tumors by artificial neural network analysis of gene expression microarray data. *Molecular Cancer Therapeutics*, *7*(5), 1013–1024.

Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., ... Liotta, L. A. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, *359*(9306), 572 - 577.

Polaka, I., Tom, I., & Borisov, A. (2010). Decision Tree Classifiers in Bioinformatics. *J. Riga Technical University*, *42*, 118–123.

Quinlan, J. R. (1979). Discovering rules by induction from large collections of examples. In D. Michie (Ed.), *Expert Systems in the Microlectronic age* (p. 168-201). Edinburgh University Press.

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

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. H., Angelo,

M., ... Golub, T. R. (2001, December). 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.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., ... Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, *1*(2), 203 - 209.

Swets, J. A. (1996). Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers.

Ulfenborg, B., Klinga-Levan, K., & Olsson, B. (2013). Classification of Tumor Samples from Expression Data Using Decision Trunks. *Cancer Informatics*, *12*, 53–66.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. M., Augustinus, Mao, M., ... Friend, S. H. (2002, January). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*(6871), 530–536.

Wang, K.-J., Makond, B., & Wang, K.-M. (2013). An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data. *BMC Medical Informatics and Decision Making*, *13*(1).

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Yu, L., Han, Y., & Berens, M. E. (2012, January). Stable Gene Selection from Microarray Data via Sample Weighting. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, *9*(1), 262–272.