Analyzing Gene Expression Data: Fuzzy Decision Tree Algorithm applied to the Classification of Cancer Data

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

Domagoj Jakobovic and Stjepan Picek
Faculty of Electrical Engineering and Computing
University of Zagreb, Croatia
{*domagoj.jakobovic,stjepan.picek*}*@fer.hr*

***Abstract*—In data mining, decision tree algorithms are very popular methodologies since the algorithms have a simple inference mechanism and provide a comprehensible way to represent the model in the form of a decision tree. Over the past years, fuzzy decision tree algorithms have been proposed in order to provide a way to handle uncertainty in the data collected. Fuzzy decision tree algorithms have shown to outperform classical decision tree algorithms. This paper 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, although both do not reach the accuracy values of the best performing classifier.**

***Keywords*-Soft discretization, fuzzy decision tree, cancer data**

I. Introduction

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. One well known methodology is based on decision trees where 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. 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 [1].

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 paper is arranged as follows. Section II describes related work. The proposed approach is introduced in Section III. The experimental setup and results are demonstrated

in Section IV. In the final section (Section V), conclusions and future work are discussed.

II. 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 papers exploring cancer data research.

The development of fuzzy variants of decision tree induction has been around for quite a while [1][2], but they remain a topic of interest in recent applications. These approaches provide examples for the application of "fuzzification" to standard machine learning methods.

There are also many variations of fuzzy decision trees. Soft Decision Trees (SDT) are presented in [3], 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 [4], 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 [5] they are integrated with genetic algorithms for data classification in database applications, and in [6] for developing a financial time series-forecasting model, where they were also combined with a genetic algorithm.

In [7], 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 [8] 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.

In the scope of cancer data analysis, a survey with a comprehensive study of various cancer classification methods is given in [9]. The authors conduct an analysis of the efficiency of methods based on their speed, accuracy and ability to reveal biologically meaningful gene information. A general framework of sample weighting to improve the stability of feature selection methods is proposed in [10].

Experimentation with a multiclass classifier based on SVM (Support Vector Machine) algorithm is reported in [11]. 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 [12]. In [13], 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 [14]. 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 [15].

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 [16]. 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 [17]. The authors experiment with SVM algorithms based on recursive feature elimination.

III. 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. For example, the i^{th} branch has the following form:

IF $x_{i1} \in A_1^m$ AND ... AND $x_{ij} \in A_j^n$ THEN $c_i \in C_i^k$

where x_{ij} denotes the j^{th} attribute of the i^{th} branch; A_j^m denotes the m^{th} antecedent value of the j^{th} attribute; and c_i is the consequent of the i^{th} rule.

The fuzzy decision tree has been extended based on fuzzy set theory [18]. The Ruspini fuzzy set [19] F is characterized by a membership function $F(a) : U \rightarrow [0, 1]$, whereby $F(a)$ is the membership degree of F with a value of $a \in U$. Let $V = \{F_1, F_2, ..., F_m\}$ be a family of fuzzy sets of U, then,

$$\sum_{i=1}^{m} F_i(a) = 1, \forall a \in U. \tag{1}$$

The "cut-point" is determined by the fuzzy set pair A_1 and A_2 such that $A_1(a) + A_2(a) = 1$. The fuzzy class entropy in a data set S is:

$$E(S) = \sum_{j=1}^{k} p(c_j, S) \log p(c_j, S), \tag{2}$$

where $p(c_j, S) = \sum_{a_i \in c_j} (A_1(a_i) + A_2(a_i))$ is the proportion of records in S, which belongs to class c_j.

After the soft discretization step, the set S is partitioned into two subsets S_1 and S_2 given a threshold value. The class information entropy is calculated by the probability of the fuzzy partition as [20]:

$$E(S) = \frac{N^{S_1}}{N^S} E(S_1) + \frac{N^{S_2}}{N^S} E(S_2) \tag{3}$$

$$E(S_i) = -\sum_{j=1}^{k} p(c_j, S_i) \log p(c_j, S_i), i = 1, 2 \tag{4}$$

$$p(c_j, S_i) = \frac{N^{S_i c_j}}{N^{S_i}}, i = 1, 2 \tag{5}$$

where $N^S = \sum_{n=1}^{|S|} \sum_{i=1}^{2} A_i(a_n)$, $N^{S_i} = \sum_{n=1}^{|S|} A_i(a_n)$, $i = 1, 2$.

The fuzzy discretization process consists of four phases: sorting, evaluation, splitting, and completion.

1) Sorting Phase: During the sorting phase, the continuous values of a feature are sorted in either ascending or descending order. Since this sorting can be computationally expensive an efficient sorting algorithm should be chosen. Thus, quick-sort is used since it is an efficient sorting algorithm with a time complexity of $O(NlogN)$ [21]. For a data set S with N records, the records are sorted according to the value of A generating a sequence of ordered values $a_1, a_2, ..., a_N$.

2) Evaluation Phase: The subsequent step after sorting is to find the best "cut-point". A "cut-point" splits a range of continuous values into two parts. A list of candidate "cut-points" $T = (a_i + a_{i+1})/2$ are generated making use of the class boundary points. By using the fuzzy set pair A_1 and A_2, the "cut-points" can be "fuzzified" to generate candidate soft discretizations. In this paper, the evaluation function used to evaluate each candidate soft discretization uses the measure of entropy as given by Equation (3).

3) Splitting Phase: The intervals are split in a top-down strategy, which requires to evaluate "cut-points". In order to choose the best one and split the range of continuous values into two partitions, the algorithm runs recursively for each part until a stopping criterion is satisfied.

4) Completion Phase: A stopping criterion specifies when the discretization process is completed. In particular, a threshold value $\theta \in [0.1, 0.2]$ is predefined and is applied as follows. If the truth level of a branch $\frac{N^{S_i}}{N^S}$ is greater than θ, then the truth level of the branch belonging to the j^{th} class is calculated by:

$$\delta_{i,j} = \frac{\sum_{a_k \in c_j} A_i(a_k)}{N^{S_i}}, i = 1, 2. \tag{6}$$

Otherwise, if the truth level of a branch $\frac{N^{S_i}}{N^S}$ is less than θ, then the corresponding branch is deleted. Another second stopping criterion is used, a predefined maximum value of δ called $\mu \in [0.8, 0.9]$. If the maximum δ value is greater than μ, the corresponding branch search is terminated as a leaf. This leaf is then assigned as the class c_j. Otherwise, the data set S is partitioned into S_1 and S_2 until the above criteria (either $\frac{N^{S_i}}{N^S} \geq \theta$ or $\delta \geq \mu$) are satisfied.

Generally, the FDT classifier starts by sorting the continuous values of a feature. Afterwards, it generates a possible candidate "cut-point", and "fuzzifies" the "cut-point" by using an entropy evaluation function to check whether the candidate's "cut-point" is satisfied or not. The algorithm keeps recursively checking until the best "cut-point" is found, and repeats the generation of the soft discretization for the other attributes. When all attributes have been soft discretized, the attribute of minimum value is selected to generate two child branches and nodes. This process repeats until one of the stopping criteria is met.

IV. Experiments and Results

The FDT was implemented in Java as outlined in the previous section. The decision tree algorithm used for comparison is WEKA's J48 decision tree implementation [22]. 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 a following subsection. 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.

A. Data Sets

The data sets[1] that have been chosen for this investigation are listed in Table I. All data sets contain gene data

[1] http://levis.tongji.edu.cn/gzli/data/mirror-kentridge.html

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). 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.

B. Evaluation measures

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

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}. \tag{7}$$

$$\text{Sensitivity} = \frac{TP}{TP + FP}. \tag{8}$$

$$\text{Specificity} = \frac{TN}{FP + TN}. \tag{9}$$

Another measure used to evaluate medical data sets is the Receiver Operating Characteristic (ROC) [28] 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:

$$\text{AUC} = \frac{1 - (1 - \text{Specificity}) + \text{Sensitivity}}{2}. \tag{10}$$

C. Comparison algorithms

The implemented FDT algorithm is compared with a classical decision algorithm known as J48 [29], which is implemented in WEKA. J48 is an extension of the C4.5 and the earlier ID3 algorithm [30].

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.

Table II
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 III
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

D. Experimental Results

To show an example of a decision tree generated by our FDT classifier applied to the Ovarian data set with the reduced feature set, we display the fuzzy decision tree in Figure 1.

Table II 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 III shows the same measures as Table II, 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, improves values can also be observed. Therefore, we can conclude that overall the feature reduction method improved the accuracy.

Table IV 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 2 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

Table I
DETAILS OF BINARY DATA SETS USED FOR EXPERIMENTS

Data set name	# of features	# of instances	class balance	Data set size	Short description	Reference
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	[23]
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	[24]
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	[25]
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")	[26]
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	[27]

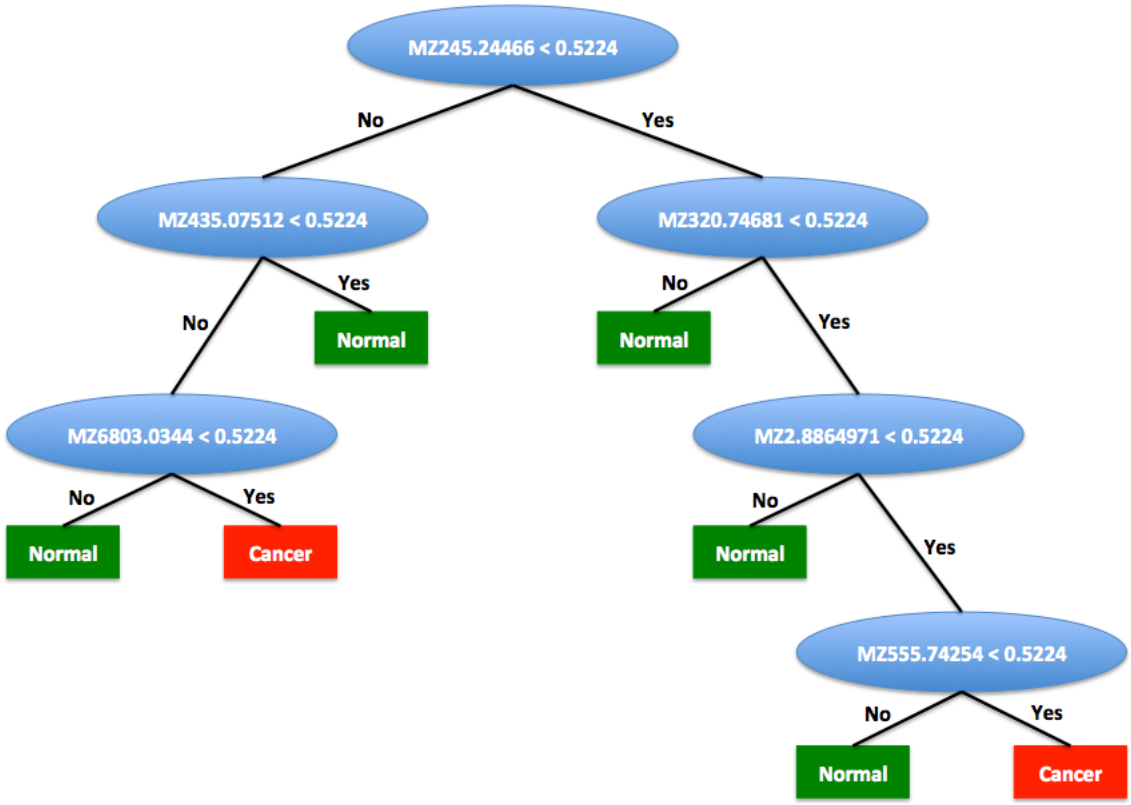

Figure 1. Decision tree resulting from the FDT classifier for the Ovarian data set

Table IV
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

Prostate cancer data set.

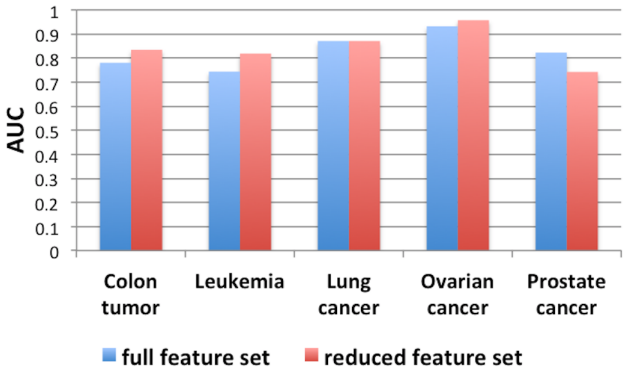

Figure 2. Comparison of AUC values for different data sets with full and reduced feature set

Table V 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). 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.

V. CONCLUSION

This paper 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.

Future work will include the comparison of different fuzzy-based classifiers to further analyze the data sets of the gene expression data also focusing on other accuracy measures.

REFERENCES

[1] R. L. Chang and T. Pavlidis, "Fuzzy decision tree algorithms," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 7, no. 1, pp. 28–35, Jan 1977.

[2] C. Janikow, "Fuzzy decision trees: issues and methods," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 28, no. 1, pp. 1–14, Feb 1998.

[3] C. Olaru and L. Wehenkel, "A complete fuzzy decision tree technique," *Fuzzy Sets and Systems*, vol. 138, no. 2, pp. 221 – 254, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165011403000897

[4] S. An and Q. Hu, "Fuzzy rough decision trees," in *Rough Sets and Current Trends in Computing*, ser. Lecture Notes in Computer Science, J. Yao, Y. Yang, R. Sowiski, S. Greco, H. Li, S. Mitra, and L. Polkowski, Eds. Springer Berlin Heidelberg, 2012, vol. 7413, pp. 397–404. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-32115-3_47

[5] P.-C. Chang, C.-Y. Fan, and W.-Y. Dzan, "A cbr-based fuzzy decision tree approach for database classification," *Expert Systems with Applications*, vol. 37, no. 1, pp. 214 – 225, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095741740900428X

[6] R. K. Lai, C.-Y. Fan, W.-H. Huang, and P.-C. Chang, "Evolving and clustering fuzzy decision tree for financial time series data forecasting," *Expert Systems with Applications*, vol. 36, no. 2, Part 2, pp. 3761 – 3773, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417408001474

[7] M. Biswal and P. Dash, "Measurement and classification of simultaneous power signal patterns with an s-transform variant and fuzzy decision tree," *Industrial Informatics, IEEE Transactions on*, vol. 9, no. 4, pp. 1819–1827, Nov 2013.

Table V
RESULTS OF COMPARISON OF FDT WITH OTHER WEKA ALGORITHMS IN TERMS OF ACCURACY

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

[8] S. Mitra, K. Konwar, and S. Pal, "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*, vol. 32, no. 4, pp. 328–339, Nov 2002.

[9] Y. Lu and J. Han, "Cancer classification using gene expression data," *Information Systems*, vol. 28, pp. 243–268, 2003.

[10] L. Yu, Y. Han, and M. E. Berens, "Stable gene selection from microarray data via sample weighting," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 9, no. 1, pp. 262–272, Jan. 2012.

[11] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub, "Multiclass cancer diagnosis using tumor gene expression signatures." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 26, pp. 15 149–15 154, Dec. 2001.

[12] L. M. Fu and C. S. Fu-Liu, "Multi-class cancer subtype classification based on gene expression signatures with reliability analysis," {*FEBS*} *Letters*, vol. 561, no. 13, pp. 186 – 190, 2004.

[13] M. Cuperlovic-Culf, N. Belacel, and R. J. Ouellette, "Determination of tumour marker genes from gene expression data," *Drug Discovery Today*, vol. 10, no. 6, pp. 429 – 437, 2005.

[14] L. Petalidis, A. Oulas, M. Backlund, M. Wayland, L. Liu, K. Plant, L. Happerfield, T. Freeman, P. Poirazi, and V. Collins, "Improved grading and survival prediction of human astrocytic brain tumors by artificial neural network analysis of gene expression microarray data," *Molecular Cancer Therapeutics*, vol. 7, no. 5, pp. 1013–1024, 2008.

[15] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nat Med*, vol. 7, no. 6, pp. 673–679, jun 2001.

[16] L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, no. 6871, pp. 530–536, Jan. 2002.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[18] P. Estevez, M. Tesmer, C. Perez, and J. Zurada, "Normalized mutual information feature selection," *Neural Networks, IEEE Transactions on*, vol. 20, no. 2, pp. 189–201, Feb 2009.

[19] E. Ruspini, "A new approach to clustering," *Information and Control*, vol. 15, pp. 22–32, 1969.

[20] L. A. Zadeh, "Probability measures of fuzzy events," *Journal of Mathematical Analysis and Applications*, vol. 23, pp. 421–427, 1968.

[21] S. S. Skiena, *The Algorithm Design Manual*, 2nd ed. Springer Publishing Company, Incorporated, 2008.

[22] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[23] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "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*, vol. 96, no. 12, pp. 6745–6750, Jun. 1999.

[24] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999.

[25] G. J. Gordon, R. V. Jensen, L. li Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Res*, vol. 62, pp. 4963–4967, 2002.

[26] E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *The Lancet*, vol. 359, no. 9306, pp. 572 – 577, 2002.

[27] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203 – 209, 2002.

[28] J. A. Swets, "Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers," 1996.

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

[30] ——, "Discovering rules by induction from large collections of examples," in *Expert Systems in the Microlectronic age*, D. Michie, Ed. Edinburgh University Press, 1979, pp. 168–201.