30th Annual International IEEE EMBS Conference
Vancouver, British Columbia, Canada, August 20-24, 2008

Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare

Muhammad Umer Khan, Jong Pill Choi, Hyunjung Shin and Minkoo Kim

***Abstract*—Data analysis systems, intended to assist a physician, are highly desirable to be accurate, human interpretable and balanced, with a degree of confidence associated with final decision. In cancer prognosis, such systems estimate recurrence of disease and predict survival of patient; hence resulting in improved patient management. To develop such a prognostic system, this paper proposes to investigate a hybrid scheme based on fuzzy decision trees, as an efficient alternative to crisp classifiers that are applied independently. Experiments were performed using different combinations of: number of decision tree rules, types of fuzzy membership functions and inference techniques. For this purpose, SEER breast cancer data set (1973-2003), the most comprehensible source of information on cancer incidence in United States, is considered. Performance comparisons suggest that, for cancer prognosis, hybrid fuzzy decision tree classification is more robust and balanced than independently applied crisp classification; moreover it has a potential to adapt for significant performance enhancement.**

I. Introduction

ACCORDING to National Cancer Institute of United States, estimated number of breast cancer cases, registered for the year 2007 is 180510, while the estimation of deaths exceeds 41000 [1]. Approximately, at the rate of one in three cancers diagnosed, breast cancer is the most frequently diagnosed cancer in women in America. Surgical biopsies confirm malignancy with high level of sensitivity, but are considered costly and can affect patient's psychology as well [2]. If malignancy is confirmed, physicians get indulged into prognosis. Surgery, radiation, chemotherapy, hormone therapy or any combination of them are considered to be the successful treatment methods. But again, selection of treatment method without considering resulting tumor behavior can lead to severe consequences. Towards these considerations, there is a growing trend of personalized predictive medicine using less invasive machine learning techniques. Prognosis helps in establishing a treatment plan by predicting the outcome of a disease. There are three predictive foci of cancer prognosis: 1) prediction of cancer susceptibility (risk assessment), 2) prediction of cancer recurrence and 3) prediction of cancer survivability. Focus of this paper is prediction of survivability, of a particular patient suffering from breast cancer, over a particular time period after the diagnosis. For this research effort, survival is considered as any incidence of breast cancer where the person is still living after 1825 days (5 years) from the date of diagnosis, as recommended in [3,4].

In the present research project, we surveyed various research efforts [5-8] in the application of different machine learning techniques to cancer prognosis. Some of the obvious trends which account for the motivation of experiments presented in this manuscript include:

- Fuzzy logic has been rarely used in cancer prognosis. Being non-crisp, it can act as a natural ally of a physician in prognostic decision making process.
- About 70% of all reported studies use Neural Networks which yield "Black Box" models for physicians to interpret.
- Majority of reported studies used machine learning techniques independently without considering potential in those techniques to cooperate with each other in a hybrid model.
- Lack of attention paid to data size. Data sets considered are not sufficiently large that can be reasonably partitioned into disjoint training and test sets.

Being motivated by above mentioned trends, we propose to investigate a hybrid scheme based on fuzzy decision trees, as an efficient alternative to crisp classifiers that are applied independently. An important aspect of this model is an interesting simultaneous cooperation between Fuzzy Logic and Decision Trees. This cooperation tries to soften the accuracy/interpretability tradeoff. In [5], fuzzy rules, for cancer diagnosis, are generated by randomly selecting data instances from training data, and performing rigorous genetic search evolving different models and then selecting the best ones. We have chosen to investigate a different alternative by using decision trees, in step-1, to learn a set of crisp rules to avoid complexities and optimization errors due to random selection of training records. In [6], among neural networks, decision trees and logistic regression, decision

Manuscript received April 16, 2008. This research is supported by Foundation of ubiquitous computing and networking (UCN) project, the Ministry of Knowledge Economy (MKE) 21st Century Frontier R&D Program in Korea and a result of subproject UCN 08B3-S2-10M

Muhammad Umer Khan is a MS candidate at Graduate School of Information and Communication Engineering, AJOU University, Suwon, Republic of Korea (email: umer@ajou.ac.kr).

Jong Pill Choi is a Professor at Department of Medical Informatics, School of Medicine, AJOU University, Suwon, Republic of Korea (email: cjp@ajou.ac.kr).

Hyunjung Shin is a Professor at Department of Industrial & Information Systems Engineering, AJOU University, Suwon, Republic of Korea (email: shin@ajou.ac.kr).

Minkoo Kim is a Professor at Graduate School of Information and Communication Engineering, AJOU University, Suwon, Republic of Korea (email: minkoo@ajou.ac.kr).

trees proved to be the best classifier for cancer prognosis using SEER data. Now the problem with decision tree algorithms is that, the decision boundaries at each node are sharp (for continuous valued attribute), due to which even small changes in attribute values may result in misclassifications [9]. Therefore, decision boundaries need to be softened and there should be a gradual transition between attribute values. Here comes the role of fuzzy logic in decision trees, in step-2, to generate fuzzy decision trees (FDTs).

Rest of the paper is divided in the following way: section II describes the data source and the modeling of FDTs, section III presents performance evaluation and finally section IV concludes this manuscript.

II. Materials and Methods

A. Prognostic and Predictive Factors in Breast Cancer

Survival of patients with breast cancer depends on two different types of prognostic factors: 1) Chronological [indicators of how long the cancer has been present (e.g. tumor size)], 2) Biological [indicators of metastatic aggressive behavior of a tumor (e.g. tumor grade)] [10]. They determine, either or not a particular tumor might respond to a specific therapy. Definitions and effect of some of the most important prognostic factors is given below:

- Lymph node status: Lymph nodes, where cancer cells get accumulated. Both number of nodes and level of involvement worsen the prognosis.
- Stage: Defined by the size of tumor and its spread. Survival is inversely proportion to size of tumor.
- Grade: How does the tumor looks like and its resemblance to more or less aggressive tumors.

Figure-1 shows ranking of survivability attributes in terms of their decisive strength, calculated using information gain (IG) applied to breast cancer data, as described in subsection-C.

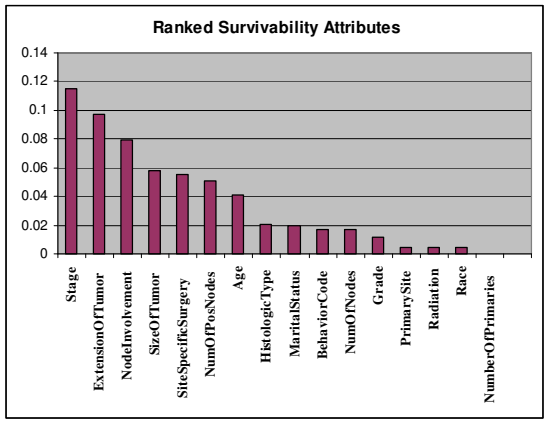

Fig. 1. Ranked Survivability Attributes

B. Data Source

In this research work, SEER (Surveillance, Epidemiology, and End Results) data (1973-2003) is used for breast cancer prognosis. Files were requested through website (www.seer.cancer.gov) of SEER program which is a part of Surveillance Research Program at National Cancer Institute. The data set is considered to be the most comprehensive source of information on cancer incidence in USA and SEER program claims quality and completeness of data.

Initially, there were 433,272 records each with 86 variables. Each record represents one incidence of cancer. Considering the aim of survival prediction, a binary target variable is created with values 0 (did not survive) and 1 (survived). Although much of the time in this research work was spent on data cleansing and preprocessing, only a brief description is given here. Records containing missing and inconsistent data, along with those in which cause of death (in 5 years after diagnosis) was other than cancer, were removed. Moreover, variables which are redundant (like recodes) and irrelevant (some variables are common in all cancer types) were removed. After an exhaustive preprocessing, final data set with 162500 records, 16 predictor variables and 1 target variable, was constructed. Following are the predictive variables used in our work:

TABLE1
PREDICTOR VARIABLES FOR SURVIVAL MODELING

Symbol	Nominal Variable Name	Num of Distinct Values
B	Race	28
C	Marital Status	9
D	Primary Site	9
E	HistologicTypeICD	44
F	Behavior Code	2
G	Grade	5
I	Extension of Tumor	12
J	Node Involvement	10
M	Site Specific Surgery	22
N	Radiation	9
O	Stage	9

Symbol	Numeric Variable Name	Mean	Std.Dev	Range
A	Age at Diagnosis	61.105	14.165	20-106
K	Num of Pos Nodes	24.376	41.238	0-99
H	Tumor Size	103.168	273.144	0-999
L	Num of Nodes	14.033	16.89	0-99
P	Num of Primaries	1.302	0.565	1-6

C. Decision Trees

As mentioned above, to generate fuzzy rules, we opted to use decision trees in the first step of modeling. We used binary C4.5 for all the experiments mentioned in this manuscript. It works as follows: to partition the data at each stage of tree, a test is performed to select an attribute with lowest entropy. Information gain (IG) is used as a measure of entropy (H) difference when an attribute contributes the additional information about class C [11].

$$H(C) = -\sum p(c)\, logp(c) \quad , c \in C \qquad (1)$$

$$H(C|X_i) = -\sum p(x) \sum p(c|x)\, logp(c|x), \quad x \in X_i, c \in C \qquad (2)$$

$$IG_i = H(C) - H(C|X_i) \qquad (3)$$

In equation (1), p(c) is the probability that an arbitrary sample belongs to class 'c'. Equation (2) shows the entropy after observing the attribute X_i for the class 'c' and p (c|x) is the probability that a sample in attribute branch X_i belongs to class 'c'. Table-2 shows different decision tree models, which we generated during experiments.

.

TABLE 2:
COMPLEXITY VS ACCURACY

	No. of Rules	Accuracy
Model-1	17	82%
Model-2	21	83.5 %
Model-3	40	84 %

And at 500 rules, accuracy was 95%. Due to sharp decision boundaries of crisp decision tree rules, this natural tradeoff gets harder. For a physician involved in prognostic decision making, both accuracy and interpretability are a must. So we decided to choose interpretability first and leave the accuracy and decision confidence for the second stage. Hence model-1 is chosen to apply fuzzification and inference in next step. Following are the first (R_1) and last (R_{17}) rules of this model.

R_1: ***If ExtensionOfTumor <= 40.0 AND AgeAtDiagnosis <= 73.0 AND NodeInvolvement <= 1.0 then IsSurvival=1***

- -

R_{17}: ***If ExtensionOfTumor > 40.0 AND SiteSpecificSurgery > 10.0 AND AJCCStage3ed > 32.0 then IsSurvival=0***

D. Fuzzy Decision Trees (FDTs)

In some previous studies [12, 13] on FDTs, proposed approaches focus on modification of decision tree pruning algorithm and require fuzzy parameters to be set by domain experts. We opted to fuzzify already generated decision tree nodes to relax the sharp decision boundaries. A similar approach is used in [7].

1) Fuzzification: An attribute or decision node is represented by a fuzzy set using a pair of complimentary membership functions. Although there can be many smart ways to initially specify the domain, lower bound 'δ_i' and upper bound 'δ_n' of a membership function, we stick to a common and simplified domain specification. Since decision threshold 'd_t' is already generated at each node of DT and remains fixed, domain delimiters can be calculated as:

$$\boldsymbol{\delta_i = d_t - f * \sigma} \quad \text{and} \quad \boldsymbol{\delta_n = d_t + f * \sigma} \qquad (4)$$

here δ_i and δ_n are lower and upper bounds of membership function, respectively. 'σ' is the standard deviation of the domain attribute. '*f*' is the fuzzification applied around decision threshold 'dt'. For our experiments, '*f*'=2 gave the optimal results. Usually, it is chosen in the range of 0 to 5. This is because larger values of '*f*' would introduce too much fuzzification and decision making process would become unclear. Linear, sigmoid, convex and concave membership functions were used to fuzzify each decision tree node. Sigmoid and linear produced good results.

$$\text{Linear } (\delta_i, \delta_n, x) = \begin{cases} 0 \rightarrow & x \le \delta i \\ x - \delta i \,/\, \delta n - \delta i \rightarrow & \delta i \le x \le \delta n \\ 1 \rightarrow & x \ge \delta n \end{cases}$$

$$\text{Sigmoid } (x; \delta_i, ,\delta_n, \beta) = \begin{cases} 0 \rightarrow & x \le \delta i \\ 2\left((x - \delta i) \,/\, (\delta n - \delta i) \right)^2 \rightarrow & \delta i \le x \le \beta \\ 1 - 2\left((x - \delta n) \,/\, (\delta n - \delta i) \right)^2 \rightarrow & \beta \le x \le \delta n \\ 1 \rightarrow & x \ge \delta n \end{cases}$$

Here β is usually half membership point ($\delta_i + \delta_n$)/2.

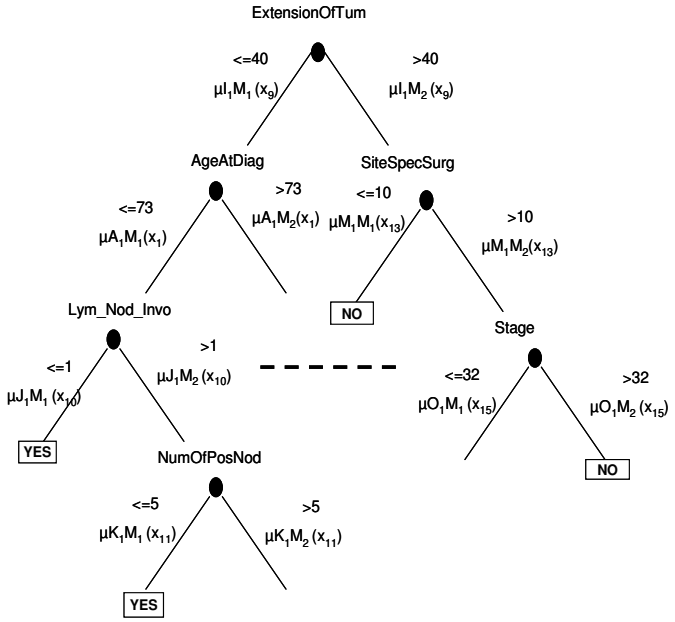

Fig. 2: First and Last Rules of Model-1 in Fuzzified Form

Figure-2 shows, how an attribute at each node is fuzzified using left and right complimentary membership functions (M1,M2).The above mentioned crisp rules (R_1, R_{17}) can be obtained in fuzzified form by traversing the left most and right most paths of the tree in figure-2. An important feature of this approach is that, it preserves the decision thresholds and symbolic structure obtained from induced tree.

2) Fuzzy Inference: The approach used is very simple and interesting in a way that, for classifying an example, all the rules contribute to some degree. A brief description is given below: (for classification of an incoming record)

- For each path (or rule) of the fuzzy decision tree, a cumulative membership grade is calculated by applying an intersection operator to the set of individual membership function values at each branch on that path. For example, cumulative membership grade of first rule R_1(left most path from root to leaf, in figure-2) is computed as:

  $\boldsymbol{\mu(R_1) = \cap \{\mu I_1M_1(x_9), \mu A_1M_1(x_1), \mu J_1M_1(x_{10})\}}$

- Now we have 17 cumulative membership grades. Each of such grades is multiplied by the rule weight or certainty factor CF_i. CF_i is calculated as [8]:

  1) When consequent class is YES (or 1)

$$CFi = \beta_{ClassYES}(R_i) - \beta_{ClassNO}(R_i) \,/\, \beta_{ClassYES}(R_i) + \beta_{ClassNO}(R_i)$$

  2) When consequent class is NO (or 0)

$$CFi = \beta_{ClassNO}(R_i) - \beta_{ClassYES}(R_i) \,/\, \beta_{ClassYES}(R_i) + \beta_{ClassNO}(R_i)$$

$$\text{Where } \beta_{ClassK}(R_i) = \sum_{x \in ClassK} \mu_i(x), k = YES, NO \qquad (5)$$

In simple words $0 \le CF_i \le 1$, means when all compatible patterns with rule R_i (those with $\mu_i(x) > 0$ for R_i) belong to the same class, CF_i equals one. This weight is calculated for each rule using training data. Rule weight has a great significance in fuzzy inference and here it is used as an alternative of Genetic Algorithms for parameter optimization.

- Finally, all the products ($\mu(R_i)*CF_i$) are combined using *union operator*, and a rule (e.g. Zadeh) or a class (e.g. Yager) with maximum membership grade, will decide the class of incoming record.

Decision $= \cup \{\mu(R_1)*CF_1 , \mu(R_2)*CF_2 , \ldots . \mu(R_{17})*CF_{17}\}$

We used Yager and Zadeh [11] inference operators for intersection and union.

III. Performance Evaluation

Experiments were performed using WEKA and Java on a Pentium PC at 1.7GHz with 1.5GB RAM. Execution time for calculating decision tree with different kernel functions varied for 6 to 12 seconds. Out of 162500 records, 30000 records as training and 10600 as test data were obtained using uniform random selection, taking into account the overlapping factor. In this research work, we used three performance measures: *accuracy*, *sensitivity* and *specificity*;

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

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad \text{Specificity} = \frac{TN}{TN + FP}$$

where TP, TN, FP and FN denotes true positives, true negatives, false positives and false negatives, respectively. Following table shows the class-wise statistics for performance evaluation of decision trees and fuzzy DTs.

TABLE2
PERFORMANCE STATISTICS OF DTS AND FDTS

Classes	YES		NO	
No. of Instances	8424		2238	
Classifier	DTs	FDTs	DTs	FDTs
True Positives	8000	8250	760	865
True Negatives	760	865	8000	8250
False Positives	1478	1373	330	174
False Negatives	424	174	1418	1373
Sensitivity	0.95	0.98	0.35	0.39
Specificity	0.34	0.39	0.96	0.99
Accuracy	82%	85%	82%	85%

Following are the average error rates on unseen test set, for different fuzzy membership functions and inference techniques in FDTs.

TABLE3
AVERAGE ERROR RATE ON UNSEEN DATA

Inference Tech.	Linear	Sigmoid	Convex	Concave
ZADEH	16.2	15.8	17.5	17.8
YAGER	15.5	14.8	16.7	17

IV. Discussion

In this paper, we have shared our experiences of investigating intelligent machine learning techniques for breast cancer prognosis analysis. We analyzed the possible potential of fuzzy logic based classifiers, and came up with a conclusion that they can be the natural allies of a physician involved in predictive medicine. Moreover, they can proficiently manage contrasting requirements of accuracy, interpretability and balance in decision. When we say balance, obviously it is not crisp. Interesting cooperation between DTs and FDTs helps to realize this aim.

After these experiments, we outlined some future dimensions which can help FDTs to prove their potential as a strong classifier and predictor in cancer prognosis. Optimization through rule weights or genetic algorithms; an analysis is required, since rule weights, domain delimiters and inference parameters are the key players affecting accuracy. In this work, we used rule weights for parameter optimizations [8]. Cooperation among rules in decision making process can also be a good area research in this perspective.

We are committed to explore the strengths of FDTs for personalized predictive medicine, which is indeed a growing trend in personalized healthcare.

Acknowledgment

We acknowledge UCN project of "21st Century Frontier R&D Program in Korea" to support this research. Hyunjung Shin gratefully acknowledges support from Post Brain Korea 21 and the research grant from AJOU University.

References

[1] http://seer.cancer.gov/csr/1975_2004/results_single/sect_01_table.01.pdf
[2] Ilias, Elias and Ioannis, *An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers*, Springer Science+Business Media, LLC 2007
[3] Cox DR. Analysis of survival data. London: Chapman & Hall;1984.
[4] H. Brenner, O.Gefellor, "A computer program for period analysis of cancer patient survival. Eur J Cancer 2002;38(5):690—5.
[5] C. Andres, P. a-reyes, "A fuzzy-genetic approach to breast cancer diagnosis. Artificial Intelligence in Medicine 17 (1999) 131–155
[6] D. Delen, G.Walker, A. kadam, "Predicting breast cancer survivability: a comparison of three data mining methods", Artificial Intelligence in Medicine (2005) 34, 113—127
[7] K. Crokett, Z.Bander, D. Mclean, "On constructing a fuzzy inference framework using crisp decision trees", Fuzzy Sets and Systems 157 (2006) 2809 – 2832
[8] H. Ishibuchi, T.Nakashima, "Effect of rule weights in fuzzy rule-based classification systems", IEEE Trans. Fuzzy Systems, VOL.9, NO.4, August
[9] J. Quinlan, Induction of Decision Trees, Machine Learning, vol. 1, Kluwer Academic Press, Dordrecht, 1986 pp. 81–106
[10] Bundred NJ. Prognostic and predictive factors in breast cancer. Cancer Treatment Rev 2001;27:137—42
[11] Ian H. Witten and Eibe Frank. Data Mining: Practiacal Machine Learning Tools and Techniques, 2nd Edition. San Fransisco:Morgan Kaufman; 2005.
[12] L. Sison, E. Chong, Fuzzy modeling by induction and pruning of decision trees, IEEE Symposium on Intelligent Control U.S.A., 1994,pp. 166–171
[13] M. Umano, H. Okamoto, "Generation of fuzzy decision trees by fuzzy ID3 algorithm and its application to diagnosis"