

DATA MINING BASED DECISION-MAKING APPROACH FOR PREDICTING SURVIVAL OF KIDNEY DIALYSIS PATIENTS

Andrew Kusiak¹, Shital Shah¹, and Bradley Dixon²

¹Intelligent System Laboratory, 3221 SC

²University of Iowa Hospital and Clinic, E300D GH

The University of Iowa, Iowa City, Iowa, 52246-1527

Tel: 1-319-335-5934; Fax: 1-319-335-5669

andrew-kusiak@uiowa.edu

<http://www.icaen.uiowa.edu/~ankusiak>

Abstract: Dialysis care is particularly complex and multiple factors may influence patient survival. The cost of such treatment for end stage kidney disease is high and needs attention for reducing it. Individual patient survival may depend on an intricate interrelationship between various demographic and clinical variables, medications, medical interventions and the dialysis treatment prescription. In this research, a data mining approach is used to extract knowledge regarding the interactions between the features and the outcome. There exist a complex and contradictory relationships among data mining rules that are difficult to interpret and implement. To resolve these conflicts a decision-making algorithm is developed using sixteen different classifiers. The decision-making algorithm employs simple and weighted voting schemes. Thus in this paper, a hybrid data mining enhanced decision making approach is used for predictions of an individual patient surviving beyond the median survival time. The concepts introduced in this research have been applied and tested using data collected at four dialysis sites.
Copyright © 2003 IFAC

Keywords: Decision Making, Data Mining, Dialysis, Survival, and Predictions.

1. INTRODUCTION

Approximately 370,000 Americans undergo dialysis, at an annual cost of \$11.1 billion (USRDS, 2002). More than 260,000 Americans suffer from chronic renal failure and around 50,000 people die each year due to kidney failure (USRDS, 2002; Cooper, 2002). Chronic renal failure occurs when the kidneys are operating at less than 50% of normal capacity (KDOQI, 2002). End-stage renal disease (ESRD) occurs when the kidneys are working at less than 10%-15% of normal capacity (KDOQI, 2002; NIDDKD, 2002). At this stage, either transplantation or repetitive kidney dialysis becomes necessary for survival. Two modalities of dialysis treatment are available, hemodialysis (HD) and peritoneal dialysis. The median life span for a person on dialysis in the US is slightly more than 3 years (USRDS1, 2002). Much of this excess mortality can be explained by the associated health conditions such as diabetes that lead to kidney failure. Understanding factors that are predictive of survival of a given patient may allow for targeted interventions for high-risk patients and may suggest improvements areas.

Data mining offers tools for individual patient based decisions rather than population-based. It provides algorithms and tools for identifying valid, novel, potentially useful, and ultimately understandable patterns from data (PAKDD, 2002; Fayyad, 1997). Although these algorithms have proven to be highly effective in predicting the outcomes, there are complex and contradictory relationships among the rules that are difficult to interpret and implement. Due to the conflicting nature of rules, development of a decision-making method is warranted. In this paper, a data mining aided by decision-making approach is used for predictions of an individual patient surviving beyond the median survival time.

2. DATA COLLECTION AND DECISION CATEGORY FORMULATION

Data collection was performed at four satellite locations of the University of Iowa Hospitals and Clinics (UIHC) in a routine process and was provided for this research project. Data collection was carried out for clinical variable, demographic,

dialysis solution concentrations, etc. Averages were computed for every patient for all variables. Data transformation was performed so as to explore the interactions between the features. In an effort to improve classification accuracy, insignificant parameters and patient data were removed from the dataset.

All patients with less than fifteen visits were removed from the dataset. Some constant chemical values for the dialysis solution common to all patients were discarded. All parameters with continuous numerical values were also discretized into ten intervals. The total number of patients and features considered for data mining were 114 and about 65 respectively.

Three decision categories were developed for data mining and decision making purposes. The category "Above median" consisted of patients who have either survived more than three years on dialysis and are still alive or deceased (Figure 1). The second category "Below median" consisted of patients who survived less than three years on dialysis (Figure 1). The third category "Undetermined" consisted of patients who are alive and have not yet completed three years on dialysis. The group of patients in the "Undetermined" category (Figure 1) could land in either categories of above or below median and were excluded from datasets for data mining.

The goal of this research was to demonstrate the applicability of data mining and decision-making to predict patient's survival. Thus the median survival time (three years in this research) could be modified to any other number of years (e.g., six years or nine years). In fact, the decision could refer to multiple intervals (e.g., less than 1 year, 1 - 2 years, 2 - 3 years, 3 - 4 years, etc). These multiple interval decisions can also be effectively mined to generate decision rules for each decision interval.

Survived Years	Alive	Deceased
Three Plus	Above median	
Less than three	Undetermined	Below median

Figure 1. Decision value formulation.

3. DECISION MAKING ALGORITHMS

Decision making algorithms was developed using partial datasets mining. The features for partial datasets were grouped based on medical relevance between the features (e.g., cardiovascular aspects, acid-base equilibrium, all chemicals, with continuous values, used in dialysis solutions formed one dataset, different systolic and diastolic blood pressures formed another dataset). Eight different combinations were created to form trial datasets (Table 1). These eight datasets were mined separately using rough set based (RS) (Pawlak, 1991) and decision-tree based (DT) (Quinlan, 1992) data mining algorithms.

The classification accuracy from 10 fold-cross validation scheme was above 75-80%. Each data subset produced two sets of rules (classifiers), one each from the two data mining algorithms. Thus in all there were sixteen classifiers capable of predicting the outcomes for new patients. These classifiers were developed to perform multi-angle, highly reliable (parallel redundancy concept in systems engineering), robust, accurate decisions (predictions). This approach forms a hybrid decision-making and data mining scheme.

Table 1. Partial datasets.

Trial Set	Parameter Description
T1	All continuous values
T2	All discrete values
T3	All Features
T4	Excludes all blood pressures except the continuous value differences
T5	Excludes all blood pressures except the discrete value differences
T6	All continuous blood pressure values except differences
T7	Continuous chemical values only
T8	Discrete chemical values only

The sixteen classifiers are analogous to sixteen experts providing judgment on a single patient's health. If all the experts concur then it would be likely a perfect diagnosis, but this rarely happens. Thus the prediction generated from the sixteen classifiers may not concur, as it is based on subset of features. A decision-making scheme was developed to handle such situations (Figure 2).

Simple voting scheme was developed in which each classifier was provided with one vote and the votes were counted for each decision outcome. Thus the decision outcome with maximum number of votes was the winner and the new record was assigned the winner outcome. A conservative approach was applied while handling cases where there was a tie. The outcome "Below median" was assigned, meaning that the patient would not survive above three years.

Features considered for generating classifiers may not have any correlation with the decision and they may be medically insignificant. Also, the quality of rules defining the classifiers may vary. The later two were taken into account by a weighted voting scheme. In this scheme the classification quality of rules was used as weight and the weight were added if more than one rule matches the outcomes (Table 2). Weighted voting scheme was employed to refine the outcomes and provide appropriate representation for strong rules versus weak rules. Thus the decision outcome with maximum number of weighted votes was the winner and the new record was assigned the winner outcome. Same conservative approach as above was applied in cases of a tie.

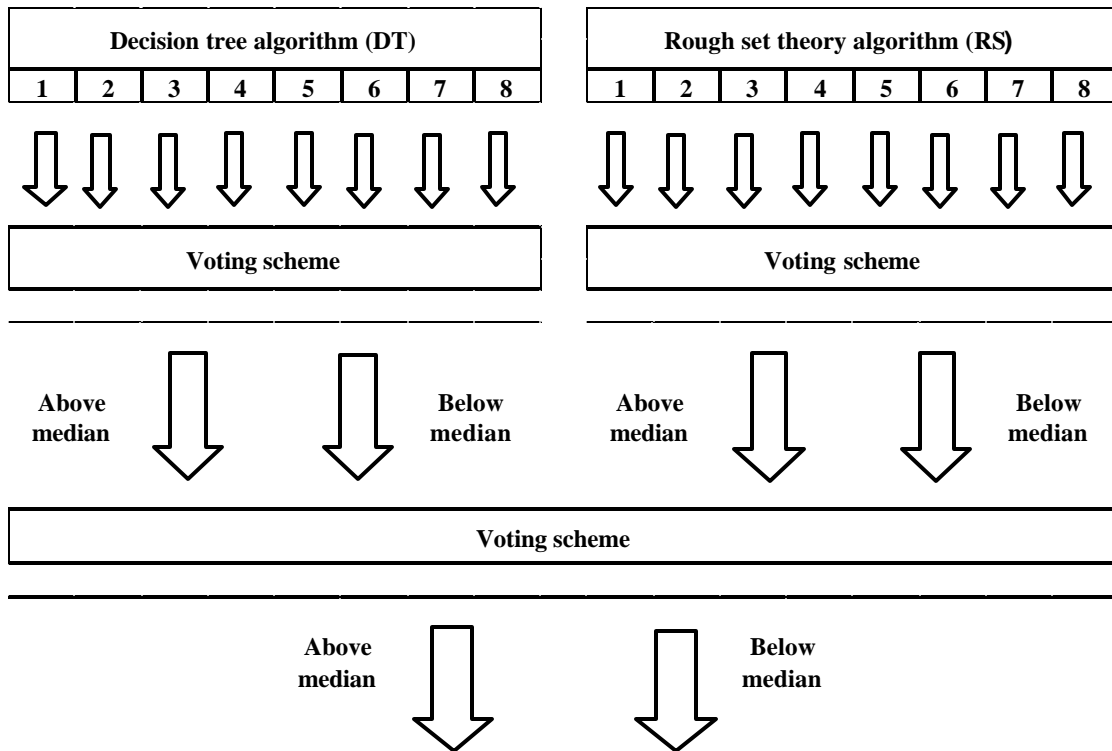


Figure 2. Decision-making scheme.

Table 2. Sample of weighted voting scheme.

Rule	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	Weight		Score	AD	ED
Case/CA	20.37	29.63	37.04	20.37	7.41	9.26	30	30	26	24	14	20	BM	AM			
1	BM	N	N	N	N	N	N	N	N	N	N	AM	20.37	20	Approx.	BM	?
2	N	N	N	N	N	N	N	N	AM	N	N	AM	0	46	Exact	AM	AM
3	N	N	BM	N	N	N	N	N	N	N	N	N	37.04	0	Exact	BM	BM
4	N	N	N	N	N	N	N	N	N	N	N	N	0	0	No Rule	?	?
5	BM	N	N	N	N	N	N	N	N	N	N	N	20.37	0	Exact	BM	BM
6	N	N	N	N	N	N	N	N	N	N	N	N	0	0	No Rule	?	?
7	N	N	N	BM	N	N	N	N	AM	N	N	AM	20.37	46	Approx.	AM	?
8	N	N	BM	N	N	N	N	N	N	N	N	N	37.04	0	Exact	BM	BM
9	BM	N	BM	N	N	N	N	N	N	N	N	N	57.41	0	Exact	BM	BM
CA: Classification Accuracy							BM: Below_med					AM: Above_med					
AD: Approximate Decision							ED: Exact Decision					N: Not applicable					

Table 3. Results produced by DT and RS algorithm.

DT						RS				
No.	A_M	B_M	Prediction	Actual	Result	A_M	B_M	Prediction	Actual	Result
1	5	3	A_M	A_M	Match	0	2	B_M	A_M	Fail
2	5	3	A_M	A_M	Match	8	0	A_M	A_M	Match
3	4	4	B_M	A_M	Fail	4	3	A_M	A_M	Match
4	4	4	B_M	B_M	Match	2	2	B_M	B_M	Match
5	4	4	B_M	B_M	Match	1	4	B_M	B_M	Match
6	4	4	B_M	A_M	Fail	1	4	B_M	A_M	Fail
7	2	6	B_M	A_M	Fail	1	2	B_M	A_M	Fail
8	1	7	B_M	B_M	Match	3	0	A_M	B_M	Fail
9	1	7	B_M	B_M	Match	0	8	B_M	B_M	Match

4. PREDICTIONS

Some patients in the "in determinate" category (Figure 1) actually crossed in the above or below median categories at the end of the study and were used for verification and validation of sixteen classifiers and testing the decision-making algorithm. Survival predictions can be performed on patients with at least 15-20 visits, i.e., based on about month dialysis data. Survival predictions for nine test cases were performed by pretending that the outcome was not known for these patients via truncating (six months before the test case either completed three years or deceased) the data after certain time into dialysis treatment. The results from the DT algorithm for test set of nine previously unseen cases are shown in Table 3. The classification accuracy is 66.67% meaning that for six out of nine cases the predictions were accurate. There were four borderline cases where decisions (votes) resulted in a tie and the assigned decision was "Below median". After analyzing these cases it was observed that they in fact were nearly approaching the three-year survival and there was a good chance of above the median survival. The prediction made by the knowledge generated with the RS algorithm (simple voting as well as weighted voting scheme) for the nine test cases are shown in Table 3. The prediction accuracy is 56% meaning that for five out of nine cases the predictions were accurate. The predictions were exact, meaning the outputs were predicted with 100% confidence for 56% cases. The rule generated sometimes conflicted with their outcome for certain cases. This was handled either by placing a "?" for such cases, forming exact RS or

applying voting scheme for the rule and predicting the outcome, forming approximate RS. For cases where the survival time cannot be predicted with high accuracy, imply that a transplant in immediate future would be necessary.

Interestingly cases 2 and 9 predicted the same outcome for each and every trial dataset. This implies that these cases were 'Most invariant'. Thus they form 'Signatures' (Kusiak, 2002) for that outcome and it is desired that other patients conform to the set of feature values of these signatures. This could be achieved, wherever possible through interventions. The overlap of cases (Table 4) was analyzed from the results generated by DT and RS algorithms. Thus for cases 2, 4, 5, 9 accurate predictions were made irrespectively of the type of the algorithm used. For cases 6 and 7 erroneous predictions have been reached.

The knowledge generated by the DT algorithm produced superior prediction accuracy, but limited numbers of features were considered (Figure 3). While the RS algorithm, resulted in higher classification accuracy (see Appendix 1) and larger number of features and exact rules. There is a tradeoff (Figure 3) between overall prediction accuracy (higher for DT) and individual classification accuracy (higher for RS). There is a need for larger datasets to obtain statistically sound prediction accuracy. As more new test cases become available the confidence of the classifiers and decision-making model will grow.

Table 4. Overlap analysis.

		Prediction			Overlap	Partial Overlap
		DT	RS Exact	RS Approx.		
Correct	Above-med	1, 2	2, 3	2	2	-
	Below_med	4, 5, 8, 9	4, 5, 9	4, 5, 9	4, 5, 9	-
	Above-med	3, 6, 7	1, 6, 7	1, 3, 6, 7	6, 7	1, 3
Wrong	Below_med	-	8	8	-	8

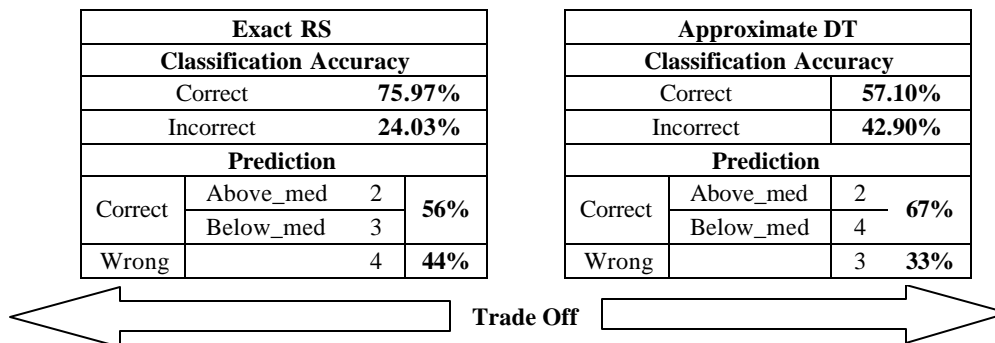


Figure 3. Decision-maker's dilemma.

After analyzing the decision rules generated by RS and DT algorithms, significant features were identified based up on the rules of high strength. The identified significant features for dialysis treatment are diagnosis, total dialysis time, potassium, calcium and sodium levels, deviation from target weight (Bellazzi, 2002), arterial pressure, post- dialysis pulse rate supine, difference between post and pre supine blood pressures, etc.

5. CONCLUSIONS

The most significant result obtained from this research was to demonstrate that data mining and decision making are useful for survival prediction of dialysis patients. The potential for making accurate decisions for individual patients is enormous and the classification accuracy is high enough to warrant use of additional resources and conduct further research. The decision tree algorithm produced correct predictions 67% of the time using equally weighted parameter sets. The rough set algorithm produced correct predictions 56% of the time using simple as well as weighted voting schemes. The decision making process can be enhanced using other approach such as goal programming, genetic programming, etc. A hybrid expert decision making system can be built by using the rules generated by the data mining algorithms and the framework of decision-making algorithm.

References

- Bellazzi, R., Larizza, C., Magni, P., Bellazzi, R., Cetta, S. (2002). Intelligent Data Analysis Techniques for Quality Assessment of Hemodialysis Services. Available: <http://magix.fri.uni-lj.si/idamap2001/papers/bellazzi.pdf>, Accessed on April 30, 2002.
- Cooper, J. (2002). U.S. incidence of kidney failure is the highest in the world. *The Medical Reporter*. Available: <http://medicalreporter.health.org/tmr0799/kidney.html>, Accessed on April 30, 2002.
- Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Eds, (1997). *Advances in Knowledge Discovery and Data Mining*, MIT Press, Cambridge, MA.
- KDOQI, (2002). K/DOQI Clinical Practice Guidelines for Chronic Kidney Disease: Evaluation, Classification, and Stratification. *National Kidney Foundation*. Available: <http://www.kidney.org/professionals/doqi/kdoqi/toc.htm>, Bethesda, MD, Accessed on December 03, 2002.
- Kusiak, A. (2002). A Data Mining Approach for Generation of Control Signatures, *ASME Transactions: Journal of Manufacturing Science and Engineering*, vol. 124, No. 4, pp. 923-926.
- NIDDKD, (2002). National Institute of Diabetes & Digestive & Kidney Diseases, National Kidney and Urologic Diseases Information Clearinghouse. Your Kidneys and How They Work. NIH Publication No. 02-4241, February 2002. Available: www.niddk.nih.gov/health/kidney/pubs/yourkids/index.htm, Accessed on April 30, 2002.
- PAKDD (2002). PAKDD Workshop. Toward the Foundation of Data Mining. Taipei, Taiwan. Available: www.mathcs.sjsu.edu/faculty/tylin/pakdd_workshop.html, Accessed on April 30, 2002.
- Pawlak, Z. (1991) *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer, Boston, MA.
- Quinlan, R. (1992). *C 4.5 Programs for ML*, Morgan Kaufmann, San Mateo, CA.
- USRDS (2002). U.S. Renal Data System, USRDS 2002 Annual Data Report: Atlas of End-Stage Renal Disease in the United States. *National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases*. Bethesda, MD. 2002. Available: <http://www.usrds.org/atlas.htm>, Accessed on 2002, December 03.
- USRDS1 (2002). The United States Renal Data Systems. Available: www.usrds.org, Accessed on April 30, 2002.

Appendix 1: Sample Cross-validation Results

The one-out-of n ($n=114$) cross-validation results for the 114 patient dataset are shown in Table 5 and Table 6. The k ($k=10$) fold cross-validation results are shown in Table 7 and Table 8.

Table 5. The confusion matrix.

	Below_med	Above_med	None
Below_med	27	15	0
Above_med	3	69	0

Table 6. Classification accuracy.

	Correct	Incorrect	None
Below_med	64.29	35.71	0
Above_med	95.83	4.17	0
Average	84.21	15.79	0

Table 7. The confusion matrix.

	Below_med	Above_med	None
Below_med	22	20	0
Above_med	4	68	0

Table 8. Classification accuracy.

	Correct	Incorrect	None
Below_med	52.83	47.17	0
Above_med	95.75	4.25	0
Average	75.97	24.03	0