International Journal of Engineering & Technology, 3 (3) (2014) 327-332
©Science Publishing Corporation
www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v3i3.2334
Research Paper

One-year renal graft survival prediction using a weighted decision tree classifier

Dalia M. Atallah [1]*, Ali I. Eldesoky [2], Amira Y. H. [3], Mohamed A. Ghoneim [4]

[1] *Electronic and Computer Engineer - Urology and Nephrology Center – Mansoura University – Mansoura – Egypt*
[2] *Professor of computer and control department - Faculty of Engineering – Mansoura University – Mansoura – Egypt*
[3] *Lecture of computer and control department - Faculty of Engineering – Mansoura University – Mansoura - Egypt*
[4] *Professor of urology - Urology and Nephrology Center – Mansoura University – Mansoura - Egypt*
**Corresponding author E-mail: daliaat@hotmail.com*

Abstract

This study introduces a weighted decision tree algorithm for prediction of graft survival in renal transplantation using preoperative patient's data. The objective was to identify the preoperative attributes that affect the graft survival. Between the years 2000-2009, renal allotransplantation was carried out for 889 patients at Urology and Nephrology Center which is the subject matter of this study. The ID3 algorithm was chosen to build up the decision tree using the weka machine learning software. A modification was made on ID3 to refine the results. A weighted vector was introduced. The element of such a vector represents the weight of each attribute which was obtained by trial and error. The results indicated that the weighted algorithm was successful in predicting the graft survival after one year and identifying the attributes affecting graft survival.

Keywords: *Decision Tree, Data Mining, ID3 Algorithm, Graft Survival, Kidney Transplantation*

1. Introduction

Medical research increasingly focuses on prediction of outcomes following surgical operations or clinical interventions. "Outcome" studies usually evaluate whether a particular data set in a particular clinical setting can predict the occurrence of a certain outcome. One of the strengths of this approach is that it allows analysis of outcomes that were previously understudied due to their complexity. Renal transplantation is a perfect target for such studies because it has a significant impact on healthcare costs and patient wellbeing.

Kidney transplantation is the only feasible hope for patients with endstage renal disease and is widely considered as a potential life saver leading to acceptable life style. In Egypt, live kidneys donors are the only available source for kidney transplantation and mostly limited to first and second degree relatives. Consequently, the number of patients with end stage renal disease requiring renal transplantation outnumbered the possible availability of living donation. This results in a state of chronic shortage in organs available for renal transplantation. Accordingly, the prognosis of kidney transplantation and the prediction of the long term survivability of the transplanted kidney are of paramount importance for planning medical research and of considerable help for decision makers in the medical area. The design of statistical survival prediction models is very complex and even more difficult to control in order to effectively predict outcome of organ transplantation. Compared to the statistical models, the data mining techniques provide much faster and promising solution. Kidney transplantation procedures consist of a large number of variables that may have nontrivial impact on modeling the prognosis of the grafts / patients [1].

Data mining is an interactive process of discovering models or patterns in large datasets and transforms a large collection of data into knowledge. The models have to be valid, novel, potentially useful, and ultimately understandable [2], [3]. The use of this computer-based information management system in medical institutions promotes digitalization of medical information, expands the information capacity in the hospital database and helps to generate knowledge enriched environment. This improves medical diagnosis, treatment, and medical research [4]. Of these data mining tools, the decision tree is the most powerful and popular decision support tool of machine learning in classification

problems [5]. It represents a rule set which categorizes data according to attributes. It is tree shaped structures that represent series of roles that lead to sets of decisions. There have been many decision tree algorithms.
Iterative Dichotomiser 3 (ID3) is one of the decision tree algorithms. ID3 is popular decision tree algorithm for classification in data mining. The advantage of ID3 is its time of construction, and computation is relatively small [4], [6]. Previous investigators have carried out new modifications and improvements in ID3 algorithm to overcome the deficiencies inherent in ID3 and to improve its performance [7], [8]. In a study by Maduskar and Kelkar [9] a new modified decision tree algorithm based on ID3 was proposed, and experimental results showed that the proposed algorithm perform well with better accuracy than the conventional ID3 algorithm.
Studies on medical data mining are increasingly published. Prediction models created by enhancing performance from classification algorithms can be used to develop evidence based adverse drug events monitoring systems in women admitted for labor and delivery [10]. Intelligible support vector machines provided a promising tool for prediction of diabetes with accuracy of 94%, sensitivity of 93%, and specificity of 94% [11]. In another study carried out on renal transplant by Jiakai et al. [12] using the Bayes net classifier for the graft status demonstrated very high prediction accuracy and true positive values for all classes suggesting that it can be employed in a clinical setting. Additionally, Karaolis et al. proposed a data mining system to extract rules for coronary heart disease events and these facilitated the grouping of risk factors into high and low risk factors that were associated with an event risk [13]. A data mining framework for DNA sequence biological data sets has been applied to the hepatitis B virus DNA by Ng et al. [14]. They developed a framework for markers discovery incorporating two algorithms. Both classifiers can explicitly give the importance of the markers and their interactions and have shown good performance in cancer prediction. Furthermore, the right ventricle support decision tree has exhibited the ability to replicate expert judgment with 85% sensitivity and 83% specificity [15]. Ravikumar et al. [1] carried out a study on renal transplantation proposing improved data mining based models for variable filtering and for prediction of graft status and survival period using the patient profile information prior to the transplantation. However, shortcoming of this study was the limited number of attributes used.
This study explored the preoperative data space to develop a weighted decision tree algorithm for predicting the graft survival in renal transplantation after one year and identifying the attributes that affect the graft survival.

2. Materials and methods

The materials of this study included data from a cohort of renal transplantation patients. These data were subjected to data mining that will be detailed.

2.1. Dataset and preprocessing

Renal transplantation patient's data were obtained from Urology and Nephrology Center, Mansoura University, Egypt from the years 2000 to 2009. The file was in SPSS format. There were 889 patients during this time period, and each patient record has 22 preoperative attributes. The patients who had missing data in any of the 22 attributes were excluded. A total of 726 patients with complete data are the subject matter of this study.
An additional step involved conversion of continuous value attributes to discrete ones so as ID3 algorithm can be applied. This has been carried out by SPSS 16 program. Discretized attributes are age of the recipient and age of the donor. Age of the recipient was divided into five values: <20, 20-, 30-, 40-, 50+. Age of the donor was divided into four values: <30, 30-, 40-, 50+.
A field named "graft survival" was added to obtain the survival status of the graft after one year following transplantation as success or failure. The success was for the patients with serum creatinine lower than three. The failure was for the patients with creatinine higher than three, graft failure for unclear reason, immunologic rejection, died with functioning graft, and technical failure. The file is converted to ARFF file that can be used in the weka program and then to Xml attribute relation file format (XRFF) file which represents the data in a format that can store comments, attribute and instance weights.

2.2. Weka

Waikato Environment for knowledge Analysis is a collection of state-of-the-art machine learning algorithms for data mining tasks which includes implementations of data pre-processing, classification, regression, clustering, association rules and visualization. It was developed at the University of Waikato in New Zealand. The system is written in Java and distributed under the terms of the GNU General Public License. It runs on almost any platform. It provides a uniform interface to many different learning algorithms, along with methods for pre- and postprocessing and for evaluating the result [16].

2.3. Data sampling

Since the prevalence of failures in the dataset was about 4% which means that the main class of the dataset is highly skewed toward success subjects (negative class), a processing step called resample weka filter with biasToUniformClass parameter set to 1.0 that resamples the data to infer a uniformly distributed new dataset was performed [12], [16], [17]. Table 1 shows the number of instances before and after resampling.

Table 1: Graft Survival Discrete Values.

Result	Number of instances before resampling	Number of instances after resampling
Failure	28	380
Success	698	346

2.4. Weighted algorithm

Algorithm: Decision tree using weighted ID3
Input:
Dataset D, which is a set of collected data and their associated class.
List of attributes, the set of selected attributes.
Weighted information gain attribute splitting selection method.
Output: a decision tree
Method:
Create a node X.
If collected data in the dataset are all of the same class, then return x as a leaf node labeled with the class.
If list of attributes is empty, then return X as a leaf node named with the majority class in the dataset.
Find the best attribute splitting criterion by applying weighted information gain as attribute splitting selection method, and label node X with this attribute.
Remove splitting attribute.
For each split S of splitting criterion partition the dataset and establish subtrees for each partition.
Let Ds be the collected data satisfying split S.
If Ds are empty then a leaf named with the majority class label in D to node X.
Else attach the node returned by generate decision tree (Ds, list of attributes, weighted information gain attribute splitting selection method) to node X.
End for.
Return node X.
In this study, ID3 was modified and used to establish the weighted decision tree. It uses weighted information gain as its attribute splitting selection method. The attributes with the highest weighted information gain is selected to build up the weighted decision tree. The weighted information gain is computed by using the following equations which is a modification in computing the information gain:

$$Info(D) = -\sum_{i=1}^{m} pi \log 2(pi) \quad (1)$$

$$InfoA(D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|}.Info(Dj) \quad (2)$$

$$\text{Gain (A)} = [\text{Info (D)} - \text{Info}_A \text{ (D)}]. \text{ W (A)} \quad (3)$$

Where: p_i = probability of i class in the dataset,
m = number of classes,
$|D_j|$ = probability of j value attribute in the dataset,
|D| = total number of instances,
V = number of attribute values,
W (A) = weight vector,
A = attribute.
The elements of the introduced weighted vector W obtained by trial and error. Figure 1 shows the stages of the proposed algorithm.

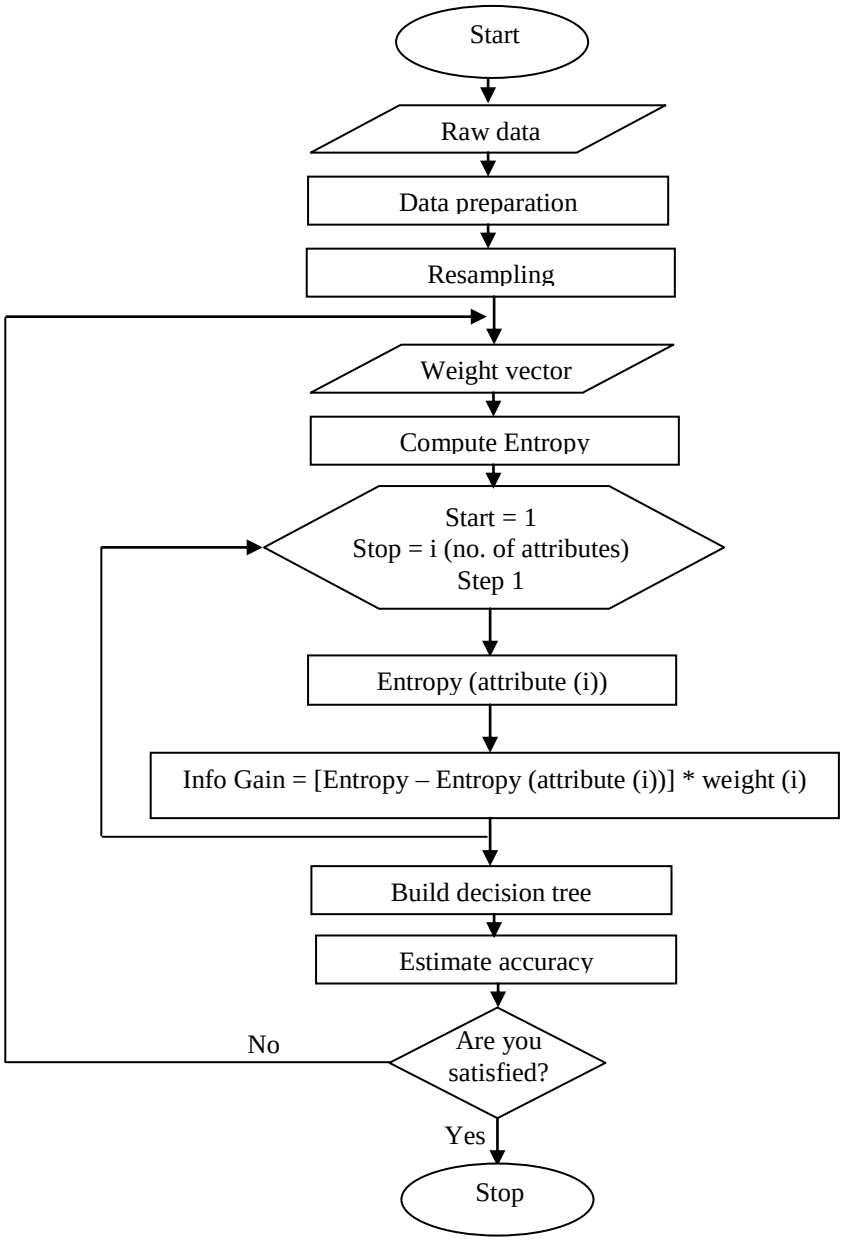

Fig. 1: Stages of the Proposed Algorithm.

3. Results

Data was prepared and resampled. The ID3 algorithm and weighted ID3 algorithm were trained and tested on the revised dataset (with uniform class distribution) that had 22 attributes and 726 instances. The decision tree algorithms were built on the dataset by means of 66% split to build the trees and 34% to test the trees. A separate decision tree was generated for each algorithm.

Simulation results are shown in Table 2. They demonstrated high prediction accuracies and kappa statistics for the weighted ID3 compared with ID3 for the dataset. There is no difference in the build time of the decision tree algorithms. The performance profile of the weighted ID3 gave true positive rate of 0.967 and 1 for the two class value of success and failure respectively as seen in Table 3. The confusion matrix for the weighted ID3 algorithm supports these findings (Table 4). The performance of the weighted ID3 demonstrated the ability to predict the status of the grafted kidney with higher accuracy for either success or failure.

Table 2: Results of Each Algorithm.

Method	Accuracy (%)	Build Time (Sec)	Kappa
Id3	92.31	0.02	0.92
Weighted ID3	97.98	0.02	0.97

Table 3: Performance Profile of Weighted ID3.

TP rate	FP rate	Precision	Recall	F-measure	ROC	Class
0.967	0	1	0.967	0.983	0.979	success
1	0.033	0.969	1	0.984	0.983	failure

Note: TP rate = true positive rate; FP rate = false positive rate; ROC = receiver operating characteristic.

Table 4: Confusion Matrix of Weighted ID3.

Success	Failure	← Classified as
116	4	Success
0	126	Failure

The whole 22 attributes used in this study and the weight vector for each attribute is shown in Table 5. Erythropoietin therapy has an important effect on graft survival. Age of the recipient, blood group similarity, prior blood transfusion, age of the donor, and HLA mismatching are the attributes that affect graft survival. The other attributes are of low importance.

Table 5: The Weight Vector for the Attributes.

Attribute name	weight	Attribute name	weight
Age recipient	1	Pre transplantation dialysis	0.5
Sex recipient	0.5	Type of dialysis	0.5
Original kidney disease	0.1	Age donor	1
Consanguinity	0.5	Sex donor	0.5
Recipient blood group	1.5	Donor blood group	0.5
Recipient antibodies	0.5	Donor antibodies	0.5
Blood group similarity	1	Donor schistosomiasis	0.5
Prior blood transfusion	1.5	HLA matching	0.5
Erythropoietin therapy	2	HLA mismatching	1.5
Pre transplantation hypertension	0.5	DR matching	0.5
Recipient schistosomiasis	0.5	DR mismatching	0.5

4. Conclusion

In this study, a weighted ID3 algorithm is introduced by proposing a weighted vector represent in the weight of each attribute. This vector was used in calculating the weighted information gain to find the best split of the decision tree. The modification was used to develop a decision tree to predict graft survival of renal transplantation after one year and to identify the attributes that affect the graft survival.

The weighted algorithm provided an improved performance. Evidence was provided that it is able to predict the graft survival after one year with a higher degree of accuracy than the ID3 algorithm. The weight vector identified the importance of each attribute. Erythropoietin therapy has an important positive influence on graft survival. Age of the recipient, blood group similarity, prior blood transfusion, age of the donor, and HLA mismatching are additional attributes that affect graft survival. Other attributes are of lower importance.

References

[1] A. Ravikumar, R. Saritha, and S. S. V. Chandra, "Recent Trends in computational prediction of renal transplantation outcomes," Inter. J. Comput. Appl., vol. 63, no. 12, pp. 33-37, Feb. 2013.

[2] D. T. Akomolafe, and A. Olutayo, "Using Data Mining Technique to Predict Cause of Accident and Accident Prone Locations on Highways," Amer. J. database Theory and appl., vol. 1, no. 3, pp. 26-38, Jan. 2012.

[3] D. Hand, H. Mannila, and P. Smyth, "Principles of data Mining," The MIT Press, 2001.

[4] R. Hu, "Medical data mining based on decision tree algorithm," Comput. inf. science, vol. 4, no. 5, Sep. 2011.

[5] D. Md. Farid, and Ch. M. Rahman, "Assigning Weights to training instances increases classification accuracy," Int. J. Data Min. Know. Management Process, vol. 3, no. 1, pp. 13-25, Jan. 2013.

[6] M. Slocum, "Decision making using ID3 algorithm," Insight: River Academic J., vol. 8, no. 2, 2012.

[7] N. Mathur, S. Kumar, S. Kumar, and R. Jindal ,"The Base Strategy for ID3 Algorithm of Data Mining Using Havrda and Charvat Entropy Based on Decision Tree," Inter. J. Inform. Elect. Eng., vol. 2, no. 2, pp. 253-258, Mar. 2012.

[8] L. Ramanathan, S. Dhanda, and S. Kumar, "Predicting students' performance using modified ID3 algorithm," Inter. J. Eng. Tech., vol. 5, no. 3, pp. 2491-2497, June-July 2013.

[9] V. Maduskar, and Y. Kelkar, "A new modified decision tree algorithm based on ID3," Int. J. Comput. Arch. Mob., vol. 1, no. 9, July 2013.

[10] L. M. Taft, R. S. Evans, C. R. Shyu, M. J. Egger, N. Chawla, J. A. Mitchell, S. N. Thornton, B. Bray, and M. Varner, "Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery," J. Biomed. Informat., vol. 42, pp. 356–364, Apr. 2009.

[11] N. H. Barakat, A. P. Bradley, and N. H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," IEEE Trans. Inf. Technol. Biomed., vol. 14, no. 4, pp. 1114–1120, July 2010.

[12] J. Li, G. Serpen, S. Selman, M. Franchetti, M. Riesen, and C. Schneider, "Bayes net classifiers for prediction of renal status and survival period," World Academy of Science, Eng. And Tech., no. 63, pp. 144-150, 2010.

[13] M. A. Karaolis, J. A. Moutiris, D. Hadjipanayi, and C. S. Pattichis, "Assessment of the risk factors of coronary heart events based on data mining with decision trees," IEEE Trans. Inf. Technol. Biomed., vol. 14, no. 3, pp. 559–566, May 2010.
[14] K. Leung, K. H. Lee, J. Wang, E. Y. T. Ng, H. L. Y. Chan, S. K. W. Tsui, T. S. K. Mok, P. C. Tse, and J. J. Sung, "Data mining on DNA Sequences of hepatitis B virus," IEEE Trans. Comput. Biol. Bioinf., vol. 8, no. 2, pp. 428-440, March - April 2011.
[15] Y. Wang, M. Simon, P. Bonde, B. U. Harris, J. J. Teuteberg, R. L. Kormos, and J. F. Antaki, "Prognosis of Right Ventricular Failure in Patients With Left Ventricular Assist Device Based on Decision Tree With SMOTE," IEEE Trans. Inf. Technol. Biomed., vol. 16, no. 3, pp. 383-385, May 2012.
[16] I.H. Witten, E. Frank, and Mark A. Hall, Data mining: Practical Machine learning Tools and Techniques, 3rd ed. USA: Morgan Kaufmann, 2011.
[17] Jiawei Han, Micheline Kamber and Jian Pei, Data mining concepts and techniques, 3rd ed. USA : Morgan Kaufmann, 2012.