Using T3, an improved decision tree classifier, for mining stroke-related medical data

Saraee, MH and Keane, J

http://dx.doi.org/10.1160/ME0317

Title	Using T3, an improved decision tree classifier, for mining stroke-related medical data
Authors	Saraee, MH and Keane, J
Type	Article
URL	This version is available at: http://usir.salford.ac.uk/18597/
Published Date	2007

USIR is a digital collection of the research output of the University of Salford. Where copyright permits, full text material held in the repository is made freely available online and can be read, downloaded and copied for non-commercial private study or research purposes. Please check the manuscript for any further copyright restrictions.

For more information, including our policy and submission procedure, please contact the Repository Team at: usir@salford.ac.uk.

© 2007 Schattauer GmbH

Using T3, an Improved Decision Tree Classifier, for Mining Stroke-related Medical Data

C. Tjortjis[1], M. Saraee[2], B. Theodoulidis[3], J. A. Keane[1]
[1]School of Computer Science, University of Manchester, Manchester, UK
[2]School of Computing, Science and Engineering, University of Salford, Salford, UK
[3]Manchester Business School, University of Manchester, Manchester, UK

Summary

Objectives: Medical data are a valuable resource from which novel and potentially useful knowledge can be discovered by using data mining. Data mining can assist and support medical decision making and enhance clinical management and investigative research.
The objective of this work is to propose a method for building accurate descriptive and predictive models based on classification of past medical data. We also aim to compare this method with other well established data mining methods and identify strengths and weaknesses.

Method: We propose T3, a decision tree classifier which builds predictive models based on known classes, by allowing for a certain amount of misclassification error in training in order to achieve better descriptive and predictive accuracy. We then experiment with a real medical data set on stroke, and various subsets, in order to identify strengths and weaknesses. We also compare performance with a very successful and well established decision tree classifier.

Results: T3 demonstrated impressive performance when predicting unseen cases of stroke resulting in as little as 0.4% classification error while the state of the art decision tree classifier resulted in 33.6% classification error respectively.

Conclusions: This paper presents and evaluates T3, a classification algorithm that builds decision trees of depth at most three, and results in high accuracy whilst keeping the tree size reasonably small. T3 demonstrates strong descriptive and predictive power without compromising simplicity and clarity. We evaluate T3 based on real stroke register data and compare it with C4.5, a well-known classification algorithm, showing that T3 produces significantly more accurate and readable classifiers.

Keywords

Knowledge-based systems, medical data mining, decision tree classification, computer-assisted decision making

Methods Inf Med 2007; 46: 523–529

doi:10.1160/ME0317

1. Introduction

Nowadays, many organisations generate and collect huge amounts of data, which traditionally used to be analysed manually. However, many hidden and potentially useful relationships may not be identified by the analyst. This explosive growth of data requires an automated way to extract useful knowledge. Data mining aims to discover novel, interesting, and useful knowledge and patterns from databases. The discovered knowledge can then be applied in the corresponding domain to increase working efficiency and improve the quality of decision making [1]. Data mining applications have already been proven to provide benefits to many areas of medicine, including diagnosis, prognosis and treatment [2].

Classification is a data mining technique which addresses the problem of constructing a predictive model for a class attribute given the values of other attributes and some examples of records with known classes [3]. Decision trees are one of the most well-established classification methods. Their intuitive nature matches the user's conceptual model without loss of accuracy [4].

When it comes to selecting a decision tree classifier though, the task is not an easy one, as a clear winner simply does not exist [5]. Important qualities for such a selection are classification and generalisation *accuracy* and tree size. In this paper we describe T3 [6], an improved version of an existing algorithm T2 [7]. T3 builds decision trees of depth at most three. T3 builds trees larger than T2 and adopts a less greedy approach in the tree-building phase, resulting in stronger predictive power without compromising simplicity and clarity. This is demonstrated using a comprehensive comparative study based on real medical data [8].

The remainder of the paper is organised as follows: in Section 2, we discuss the problem and present the data set from the medical domain; Section 3 gives a brief description of C4.5, a popular classification algorithm, and T3 which improves the tree-building process; comparison and evaluation of the performance of the two algorithms, based on extensive experiments on the medical data set is done in Section 4; discussion, conclusions and future work are presented in Section 5.

2. Description of the Problem Domain

According to the World Health Organisation international collaborative study [9], the definition of "stroke" is as follows: "Rapidly developing clinical signs of focal or global disturbance of cerebral function, lasting more than 24 hours or leading to death with no apparent cause other than that of vascular origin."

Worldwide, there are about 4.6 million deaths from strokes each year [10] and this makes it the leading cause of death in virtually all countries. There are 550,000 hospitalisations and 150,000 deaths attributable to stroke in the United States alone [11]. In 1990, strokes accounted for more than 76,000 deaths in the UK, 12% of the total (9% for males and 15% for females), and some 92% were in people over 65 years old [8]. Over 30% of stroke patients will die within the first year from the event [12]. The majority of stroke patients survive, but only one third make a good recovery and thus, strokes are a major cause of chronic disability which results in lower quality of life and increased expenses for nursing care and so-

Received: March 31, 2005; accepted: July 25, 2006
Downloaded from www.methods-online.com on 2011-10-17 | IP: 88.104.15.66
For personal or educational use only. No other uses without permission. All rights reserved.

cial support. Strokes are a large contributor to the total cost of health care, 5% of all such expenditure [8] whereas the total cost of strokes was estimated to be $30 billion in the US alone in 1993 [12]. People most likely to have a stroke are older people as well as those with certain medical problems, like hypertension and diabetes; lifestyle factors such as diet, drinking alcohol, smoking and exercise also affect the risk of stroke [13].

The case study we use in this paper has been developed in cooperation with the Medical School of the University of Manchester as part of a previous project [8, 14]. The problem addressed in that project was that of building models, that use data related to the medical history of people, which may describe and predict when stroke occurs and what the consequences of a stroke might be.

Real data selected from a computer-based stroke register were used. The stroke register was collected from the geographical area of East Lancashire, U.K., from notes completed by the patients themselves and also by the general practitioners (GPs) who interviewed the patients. The information collected is related to stroke, personal habits, family and personal medical history, and area of residence.

Patients, deceased or alive, were included as cases if they met the following criteria:

- age less than 80 years at the time of stroke,
- ordinary residence in the study area and registered with a participating GP,
- first-ever stroke occurred during the defined study interval

Patients with a first-ever stroke under 80 years old at the time of stroke were selected as cases because strokes for patients under 80 years old can be regarded as preventable.

Subjects were included as "controls" if they met all the following criteria:

- alive on the date of event (the index date) of the corresponding case,
- ordinary residence in the study area and registered with a participating GP,
- without a history of stroke before the index date

These were considered the most important attributes by the GPs and the data mining experts because of previous experience with similar case studies.

The GP practice computerised age-sex register was used to search for all patients of the same sex and age within two years of the date of birth of the index case in the same surgery, giving a list of subjects who might have been eligible for controls. Two were randomly selected from the list and their notes were found. If they did not have a documented history of stroke, they were selected as controls.

All relevant details before the index date were collected from the notes of both index cases and controls by the research nurse or a study physician using a standardised data collection form. All systolic and diastolic pressure readings with their dates of measurement, lifestyle advice, details of hypertension treatment, and other data were obtained from the practice notes. Thus there were 70 attributes per patient.

East Lancashire, England, includes the towns of Blackburn and Burnley, and the surrounding semi-rural and rural areas covering a total population of 534,287 in the 1995 general practice register. There were 118 general practices with a total of 276 GPs. The final version of the stroke register that was developed consists of 795 patients, one third (265) of whom in fact suffered from stroke, i.e. "cases", and two thirds are people who did not, i.e. "controls".

Some who had a stroke died and some survived it. Thus, we decided to build two models, one for stroke and one for surviving stroke. Out of the original 70 attributes presented in the database, 27 were selected to be used for building the model. There were a further ten attributes derived from the original 70 following doctors' advice and guidelines. The rest were deemed irrelevant (like postcode for example) or strongly correlated to others (e.g. age and date of birth). A comprehensive list of all these attributes is included in Table 1.

The data set mainly used for experimentation was *Med_123* which consists of 795 records of people, 265 of whom had suffered a stroke in the past. "Stroke" is the two-valued class attribute. Thus, the set has a baseline error of 33.3%, i.e. 66.6% of the records belong to the majority class, (no stroke). The set consists of 37 attributes, 11 being continuous, four binary, 20 ternary, and two having at least five different values. There are many missing values however (15,980 in total, meaning that 52% of the attribute values are missing). Only 5% of hypertension data were missing from the notes whereas the proportion of missing information for medical history was large. However, we have analysed the effect of these missing values, which illustrated their potential risks. It should be noted that values indicated as "not mentioned" were considered as "missing" values. A set called *above50* with 13 attributes and 2252 missing values in total was derived by excluding all the attributes which have more than 50% missing values. The attributes in *above50* are: ethnic, sex, age, source of referral, smoking status, alcohol any, alcohol status, height, weight, CP100at_3m, CP100_3m, asbp1_at, and adbp1_at. The baseline error for *above50* is also 33.3%.

We also used a *med_newlive* set that had the same attributes, being different only in that the class attribute was this time "lived" ('lived after stroke or not within two months of the date of the stroke diagnosis') instead of "stroke". Because of some missing values for that attribute, we had to exclude 124 records from the data set, thus leaving 671 with which to work. We used both cases and controls for the classification process because these relate to all the possible values of the classifying attributes. The number of missing values for this set is 13,275. The baseline error for *med_newlive* is 11.9%. Table 2 shows the properties of the data sets used.

3. Classification Algorithms Description

The aim behind the experimentation was that some of the attributes might be predictive of the class attribute ("stroke" in the first two data sets, "lived" in the third). Building classification trees is a useful method suitable for this type of study. However, there are several algorithms that produce such trees; C4.5 [15] and SLIQ [16] are reported to be of the best performing in this area. T3 is an algorithm that results in

Downloaded from www.methods-online.com on 2011-10-17 | IP: 88.104.15.66
For personal or educational use only. No other uses without permission. All rights reserved.

small but accurate trees [6]. T3 outperforms SLIQ and achieves results comparable to C4.5 when using public domain data sets [17]. Thus we decided to make a comparative analysis using C4.5 and T3, the main features of both are described below.

3.1 C4.5

C4.5 adopts a depth-first strategy, using the *gain ratio*, as a split criterion, resulting in reduced bias in favour of many-valued attributes, as compared to algorithms that use the *information gain* [15]. The algorithm, however, was also considered to be biased in favour of continuous attributes, this being the reason for improvements proposed later [18]. Splits in continuous values are binary, dividing the search space into two disjoint parts.

Unknown values are not treated as extra ones. They are ignored in the training phase resulting in classification errors. The probability of various possible results is estimated in the case of unknowns during testing. C4.5 employs a *pruning* technique that replaces subtrees with leaves, thus reducing overfitting. The accuracy achieved by C4.5 in a number of data sets was high compared to other algorithms [15, 18]. For our experiments we used both the pruned and the unpruned version of an implementation of C4.5 written in C++ running in UNIX [19].

3.2 T3

T3 [20] is an algorithm based on T2 [7]. T3 calculates optimal decision trees up to depth 3 by using two kinds of decision splits on nodes. Discrete splits are used on discrete attributes, where the node has as many branches as there are possible attribute values. Interval splits are used on continuous attributes where a node has as many branches as there are intervals. The number of intervals is restricted to be either at most as many as the number of existing classes plus one, if all the branches of the decision node lead to leaves, or to be at most 2 otherwise. The attribute value "unknown" is treated as a special attribute value. Each decision node (discrete or continuous) has

Table 1 Attribute description

Attribute	Type	Description and Values
Ethnic	ternary	1. Caucasian, 2. Indian subcontinent, 3. Other
Sex	binary	1. Male, 2. Female
Age	continuous	Numeric attribute. Age at the time of stroke
Admission	six values	Specifies when a patient was admitted to the hospital. 1. Within 24 hours, 2. Within a week, 3. Within a month, 4. More than one month, 5. Not admitted, 6. Already in hospital at time of stroke
Most recent stroke diagnosis recorded	six values	1. Unknown type of stroke, 2. Unspecified cerebral infarction, 3. Embolic infarction, 4. Thrombotic infarction 5. Subarachnoid haemorrhage, 6. Infracerebral haemorrhage 9. Other
Source of referral	ternary	1. GPs, 2. Clinical Audit Team, 3. District Liaison Team
Aspirin since	ternary	Was aspirin prescribed prior the stroke (1. Yes, 2. No, 9. N/A)
Past Angina	ternary	Suffered from Angina (1, 2, 9)
Past T.I.A	ternary	Suffered from past Transient Ischemic Attack (1, 2, 9)
Past M.I.	ternary	Suffered from previous Myocardial Infarction (1, 2, 9)
Past A.F.	ternary	Suffered from past Atrial Fibrillation (A.F.) (1, 2, 9)
Past others	ternary	Suffered from other cardiovascular disease (1, 2, 9)
Past Known Diabetes	ternary	Known diabetes of the patient (1, 2, 9)
Post Found Diabetes	ternary	Diabetes found post the stroke (1, 2, 9)
Past Renal Failure	ternary	Suffered from Renal failure (1, 2, 9)
Past Obesity	ternary	Suffered from Obesity (1, 2, 9)
Past Migraine	ternary	Suffered from Migraine (1, 2, 9)
Post Obesity	ternary	Suffered from Obesity post the stroke(1, 2, 9)
Past P.E.	ternary	Suffered from Pulmonary Embolism (P.E.) (1, 2, 9)
Past V.T.	ternary	Suffered from Venous Thrombosis (V.T.) (1, 2, 9)
History Any	ternary	Any history of a stroke in a first degree relative (1, 2, 9)
History M.I	ternary	Any history of a Myocardial Infarction (MI) in a first degree relative (1, 2, 9)
Smoking Status	binary	Smokes 1.Yes, 2. No
Alcohol Any	ternary	Any record of alcohol consumption prior to the event? (1, 2, 9)
Alcohol Status	ternary	The patient drinks alcohol: 1. Never, 2. occasionally, 3. regularly
Height	continuous	The height of a patient in centimetres.
Weight	continuous	The weight of a patient in kilos.
Derived attributes	**Type**	**Description and Values**
CP100at_3m	binary	Patient whose BPs (SBP/DBP) have been raised (SBP>=160 and DBP>=95 mmHg) on >= 2 times within 3 months, (Yes=1, No=0)
CP100_3m	binary	Patient whose BPs (SBP/DBP) have been raised (SBP>=160 or DBP>=95 mmHg) on >= 2 times within 3 months, (Yes=1, No=0)
asbpr_ra	continuous	Avg. SBP from first recorded to first raised BP, not including this.
adbpr_ra	continuous	Avg. DBP from first recorded to first raised BP, not including this
asbpt_s	continuous	Avg. SBP from first treatment to when this treatment stopped
adbpt_s	continuous	Avg. DBP from first treatment to when this treatment stopped
asbps_re	continuous	Avg. SBP from last treatment to last recorded BP before stroke
adbps_re	continuous	Avg. DBP from last treatment to last recorded BP before stroke
asbp1_at	continuous	Avg. SBP from first raised BP to last recorded BP for all subjects
adbp1_at	continuous	Avg. DBP from first raised BP to last recorded BP for all subjects

Downloaded from www.methods-online.com on 2011-10-17 | IP: 88.104.15.66
For personal or educational use only. No other uses without permission. All rights reserved.

Table 2 Data sets and their properties used in experimentation

				Attributes				
					No of distinct values			
Dataset	No of Records	Missing values	Classes	Continuous	2	3	>= 5	Total
Med_123	795	15980	2	11	4	20	2	37
above50	795	2252	2	5	4	4	0	13
med_newlive	671	13275	2	11	4	20	2	37

an additional branch, which takes care of unknown attribute values. In fact this way of treating unknown attributes is reported to perform better than that of C4.5 [21].

In the tree-building phase, at each node, all attributes are examined, in order to select one on which a split will be performed for the node. When the attribute is discrete, the relative frequencies of all of its possible values are calculated. For continuous attributes, the same approach would be inefficient because of the number of possible values and the resulting low frequencies of them. For that reason, local discretisation is used. Finally a split is performed on an attribute if it results in maximum accuracy. Consequently T3 produces a tree hierarchy, which determines how important is an attribute in the classification process, in contrast to C4.5 which uses the gain ratio.

To carry out this local discretisation of a continuous attribute, its values have to be partitioned into multiple intervals. The set of intervals that minimises the classification error is found by a thorough exploration instead of heuristically applying recursive binary splitting. The search for these intervals is computationally expensive, so T3 restricts decision trees to three levels of tests, where only the third level employs non-binary splits of continuous attributes.

T3 does not use a pruning technique. Instead it uses a parameter called *Maximum Acceptable Error* (MAE). MAE is a positive real number less than 1, used as a stopping criterion during building the tree. T2 was observed to use a greedy approach when building the tree, thus further splitting at a node would stop only if the records already classified in this node belonged to a single class.

However, this greedy approach is not optimal, because minimising the error in the leaf nodes does not necessarily result in minimising the overall error in the whole tree. In fact, it was proved that a strategy choosing locally optimal splits necessarily produces sub-optimal trees [22]. It should be noted here that *classification error* indicates how many instances of a training set have been incorrectly classified, while *generalisation error* indicates how many instances of a testing set have been incorrectly classified. Furthermore even minimising classification error does not always cause minimisation of the generalisation error due to overfitting.

By introducing MAE, we allow the user to specify the level of purity in the leaves and to stop further building of the tree at a potential node split. We set MAE to have four distinct values, namely 0.0, 0.1, 0.2 and 0.3, meaning that splitting at a node stops even if the error in that node is equal to or below a threshold of 0, 10, 20 or 30% respectively.

More precisely, building the tree would stop at a node in two cases. In the first case, building stops when the maximum depth is reached, i.e. 3 when T3 is used or 2 when T2 is used. In the second case, building stops at that node only when all the records remaining there to be classified belong to the same class in a minimum proportion of 70, 80, 90 or 100%. We used eight different versions of T3 in our experiments, four with maximum depth two and MAE set to 0, 0.1, 0.2 and 0.3 respectively; and four with maximum depth three and MAE set to 0, 0.1, 0.2 and 0.3 respectively. For the rest of this paper, we use the following naming convention: Tx.y is the version of T3 with maximum depth x and MAE set to 0.y. The T3 implementation used for the experiments was written in Visual C++ running on Windows [23].

4. Experimental Results and Evaluation

Initially we ran T3 and C4.5 against *Med_123* and it was shown that T3 outperformed C4.5. The best tree was built by T3.0 of size 39 with 0% classification error. The pruned version of C4.5 resulted in a trivial tree (size 1) and the unpruned version in a tree of size 226 with 25.5% classification error. Thus, the model built by C4.5 was both very complicated and inaccurate. The model built by T2.y was simpler than the one built by T3.y resulting in only one misclassification error that seems to be due to noise. It was also observed that "admission to hospital" and "most recent stroke" were two attributes appearing high in the tree hierarchy, alongside "sex" and "age".

We then split the data set randomly into a training set of 530 records and a test set of 265 records. The distribution was equal, resulting in a training set of baseline error of 33.2%, and a test set of 33.6%. The results were similar to the ones obtained previously. This time, T3.2 built the best (i.e. having a minimum sum of classification and generalisation error) tree of size 54, with 0% classification and 0.4% generalisation error. C4.5 unpruned built a tree of size 186 with 20.9% classification and 34.7% generalisation error; pruning resulted once more in a trivial tree of one node. The models built this time by T3 were nearly identical to the ones built previously. High accuracy was preserved and remained maximal even for prediction. C4.5 did slightly better than before, but still much worse than T3.

As mentioned above, "admission to hospital" and "most recent stroke" were attributes important for the classification process but were deemed to be strongly related to the class attribute. In fact, these two attributes were missing in all of the control data, which explains the perfect classification. Thus we decided to exclude these attributes, in search for more "interesting" rules. We then carried out experiments with

Downloaded from www.methods-online.com on 2011-10-17 | IP: 88.104.15.66
For personal or educational use only. No other uses without permission. All rights reserved.

two more data sets, (*Med_123-ah* and *Med_123-mrs*), produced by excluding one of each of the two attributes at a time. Results show that excluding just one of the two attributes would simply increase the "importance" of the other, thus eventually we decided to exclude both attributes in a data set called *Med_123-am*. In this case, the most important attributes ensued to be "alcohol status", "history any", "past MI", "asbpt_s" and "age".

Overall, we observed that, for *Med_123-ah*, the best tree was built by T3.0 (size 19 with 0% classification error), while C4.5-unpruned resulted in a tree of size 194 with 26.3% classification error. Similarly for *Med_123-mrs*, the best tree was built by T3.0 (size 30 and 0.1% classification error), while C4.5-unpruned resulted in a tree of size 226 and 25.4% classification error. Finally, for *Med_123-am* T3.0 resulted in a tree of size 50 with 14.6% classification error, while C4.5-unpruned resulted in a tree of size 194 with 26.2% classification error. The results are summarised in Table 3.

We then used another data set: *med_newlive* which has 671 records of 37 attributes and results still favoured T3. In fact T3.0 built a tree of size 92 with 5.1% classification error, while C4.5-unpruned resulted in a tree of size 58 with 11.2% error which is twice as bad as T3. The most important attributes here ensued to be "smoke status", "most recent stroke", "admission to hospital" and "alcohol status".

By splitting the set into a training subset of 447 records and a test subset of 224 records (*med_newlivet*) we achieved the following results: best version of T3 was T3.0, which built a tree of size 85, having the classification and generalisation error 3.4% and 16.5 % respectively. C4.5-unpruned built a tree of size 49 with 8.3% and 14.7% error.

By analysing the results acquired from *med_newlive* and *Med_123* sets, we hypothesized that T3 achieved here much higher accuracy than C4.5, contrary to evidence from previous work where performance was comparable [6], because of the large amount of missing values in the data sets.

To verify this hypothesis, we excluded from *Med_123* all the attributes having more than 50% missing values, resulting in a set we named *above50*, which has 13 attributes and 2252 missing values in total (for results see Table 4). The 50% threshold was selected in order to exclude attributes that appeared to be missing from records belonging to "controls" rather than "cases". Our hypothesis was verified as C4.5 unpruned resulted in the lowest classification error (21.0 %) while T3.0 did not do any better than 23.6%. However, the size of the tree for T3.0 was only 76 as compared to 220 for C4.5 unpruned. The most important attributes here ensued to be "smoke status", "alcohol status" and "source of referral".

Once again we split the set into training and test sets (*above50t*). This time results were comparable as C4.5 unpruned resulted in a tree of size 144, with 21.9% classification error and 38.5% generalisation error while T3.0 built a tree of size 73, with 21.3% and 40.8% classification and generalisation error respectively.

5. Discussion and Conclusions

This work highlights the strengths and weaknesses of C4.5, a well-established decision tree classification algorithm, and T3, an improved version of another algorithm, when using real life medical data sets. From the experiments reported in Section 4, it is clear that T3 can perform very well, especially if the objective is known. If simplicity is the main quality we are looking for then small tree size is required. When a good model of existing collected data is sought, then low classification error is the objective.

Table 3 Experimental results (part 1)

Data set	Med_123 37 att 795 rec		Med_123t 37 att 795 (530/265) rec			Med_123-ah 36 att 795 rec		Med_123-mrs 36 att 795 rec		Med_123-am 35 att 795 rec	
	tree size	class. error %	tree size	class. error %	gen. error %	tree size	class. error %	tree size	class. error %	tree size	class. error %
T2.0	**18**	0.1	**18**	0.2	0.8	**12**	0.1	**15**	0.3	**10**	16.9
T2.1	18	0.1	18	0.2	0.8	14	0.1	15	0.3	10	16.9
T2.2	18	0.1	18	0.2	0.8	14	0.1	15	0.3	10	16.9
T2.3	18	0.1	18	0.2	0.8	14	0.1	15	0.3	10	16.9
T3.0	39	**0**	39	0	0.8	19	**0**	30	**0.1**	50	**14.6**
T3.1	76	0	76	0	1.1	41	0	56	0.1	50	14.6
T3.2	56	0	54	**0**	**0.4**	41	0	56	0.1	50	14.6
T3.3	49	0	50	0	1.5	41	0	52	0.1	44	15
C4.5unpr.	226	25.5	186	20.9	34.7	194	26.3	226	25.4	194	26.2
C4.5pr.	1	33.3	1	33.2	33.6	1	33.3	1	33.3	1	33.3
C4.5pr.	1	33.3	1	33.2	33.6	1	33.3	1	33.3	1	33.3

N.B. "**bold**: best value of column"

Downloaded from www.methods-online.com on 2011-10-17 | IP: 88.104.15.66
For personal or educational use only. No other uses without permission. All rights reserved.

Table 4 Experimental results (part 2)

Data set	med_newlive 37 att 671 rec		med_newlivet 37 att 671 (447/224) rec			above50 13 att 795 rec		above50t 13 att 795 (530/265) rec		
	tree size	class. error %	tree size	class. error %	gen. error %	tree size	class. error %	tree size	class. error %	gen. error %
T2.0	**26**	8.8	**24**	6.5	15.2	20	29.3	**21**	28.7	**35.5**
T2.1	26	8.8	24	6.5	15.2	20	29.3	21	28.7	35.5
T2.2	1	11.9	1	10.7	14.3	20	29.3	21	28.7	35.5
T2.3	1	11.9	1	10.7	14.3	20	29.3	13	29.4	35.1
T3.0	92	**5.1**	85	**3.4**	16.5	76	23.6	73	**21.3**	40.8
T3.1	66	5.8	60	3.8	18.3	76	23.6	73	21.3	40.8
T3.2	1	11.9	1	10.7	14.3	76	23.6	69	21.7	40.8
T3.3	1	11.9	1	10.7	14.3	63	24.9	50	24.3	37.4
C4.5unpr.	58	11.2	49	8.3	**14.7**	220	**21.0**	144	21.9	38.5
C4.5pr.	1	11.9	1	10.7	14.3	**7**	31.4	23	29.2	38.1

N.B. "**bold**: best value of column"

Finally, when strong predictive power is required, it is low generalisation error that is required.

The best performing algorithm in terms of classification error was T3.0 in eight out of nine cases. It resulted in 6.69% average classification error, while C4.5's best version (unpruned) resulted in 20.06% error on average. The best performing algorithm in terms of generalisation error, excluding trivial trees of size one, was T2.0, with 17.17% error on average, while C4.5's best version (unpruned) resulted in 29.3% error on average. Finally, the best performing algorithm in terms of tree size, again excluding trivial trees, was T2.0, which generated trees of average size of 19.25, while T3.0 and C4.5 unpruned generated trees of average size of 56.62 and 162.87 respectively.

We also notice that C4.5 pruned resulted in trivial trees in seven out of nine cases, and even when it did not (*above50* and *above50t*), accuracy was still significantly low. On the other hand, C4.5 unpruned produced much larger trees than any version of T3 in seven out of nine cases. This means that models built by C4.5 are nearly always more complex despite accuracy being lower when compared to T3. As for T3, we can conclude that when building trees of size 2, classification error is high because of the simplicity of the model while generalisation error is rather low, because of the fact that it avoids overfitting the data.

A well known issue in the context of classification is where the induced tree may overfit the data and essentially learn the training set too well, thus resulting in poor performance of the classifier on a test set. Established approaches to avoid overfitting are to either pre-prune the tree, i.e. halt tree construction at some depth, or post-prune the tree, i.e. remove branches from a fully grown tree. In the context of the work here and the results obtained using T3, this tree-pruning occurs naturally as with T3 the depth of the tree is limited.

As it is clear from the results, whereas the error was low in the original data sets (0.1% at most) by excluding two attributes we reached an error of 14.6%, showing that these two have a strong predictive power. This feature however was not captured by C4.5.

Consequently T3 provided an accurate and predictive model of the data. It also highlighted the important and/or correlated attributes for further investigation. For example, all attributes known to affect the risk of stroke [13], such as age, hypertension, alcohol and smoking were picked up by T3. Other factors identified by T3, such as history of stroke in first degree relatives or history of myocardial infarction can be verified by experts, while others can be either attributed to the nature of the data (most recent stroke, admission to hospital) or can help to formulate new hypotheses (sex, source of referral). An alternative in order to identify correlated attributes would be the use of statistics and feature selection as a pre-processing step.

A separate issue highlighted by this work is that of how to handle missing values. C4.5 adopts the rather drastic solution of ignoring these, while T3 takes these into account as a separate attribute value and achieves higher accuracy. Both approaches have merits and one could argue that neither should be used without consulting the domain experts. On the one hand, excluding missing values, as these appear in older or less relevant data, is the solution adopted in [24, 25], where missing data ranged between 40% and 56.6%. On the other hand, one should recognise that missing data is an eventuality of most real data sets and choosing to ignore this can only weaken the benefits from data mining and statistical analysis. According to Hand et al.: "It is a rare data set that does not have such problems ... Of course, all of these problems also arise in standard statistical applications (though perhaps to a lesser degree with small, deliberately collected data sets) but basic statistical texts tend to gloss over them" [26]. Furthermore, Kurgan et al.

Downloaded from www.methods-online.com on 2011-10-17 | IP: 88.104.15.66
For personal or educational use only. No other uses without permission. All rights reserved.

recognise that: "In spite of the very high number of missing values, our system generated interesting results. They included finding some relationships that were already known to the experts, and which validated correctness of the approach. In addition, the system generated new and clinically important knowledge about the disease" [25]. For a more comprehensive discussion about treating missing values in medical studies the interested reader is referred to [27]. Here, the authors argue that excluding missing values might bias results and conclude that "there are fundamental limits on the ability of statistical methods to compensate for such problems".

However, one should not forget that even though data mining can provide assistance in making the diagnosis or prescribing the treatment, it cannot replace the physician's intuition, experience and interpretive skills [2, 27]. In short, data mining is not aiming to replace medical professionals and researchers, but to complement and support their efforts to save human life [24].

To conclude, T3 has given very encouraging results when using real data; however further evaluation would be useful, using different data sets to verify these findings. For example a comparison of fuzzy inference, logistic regression, and classification trees (CART) concluded that the accuracy rates achieved (less than 84%) are not sufficiently large to justify use of these methods in daily practice from a clinical point of view [28]. A recent comparison of neural networks, decision trees and logistic regression, the most commonly used statistical method, indicated that the decision tree (C5 an improved version of C4.5) is the best predictor [24]. T3 could demonstrate higher predictive ability as demonstrated by the results presented here. Another study used a combination of logistic regression and decision trees for the prediction of mortality for intensive care patients. The study showed that the hybrid model provides better prognostic performance than the global logistic regression model [29].

Other ideas that would improve T3, as a full-scale medical decision support system, would be the introduction of a pre-processing stage to eliminate attributes which are irrelevant or predict missing values in case there are many. Feedback from users who are experts in their domain should also be collected to fine-tune the system especially in the light of specialised evaluation of the decision trees themselves. Interesting or even novel patterns might be observed by medical experts. Finally, a user-friendly interface that would suggest which version is to be used depending on the nature of the data and the task in hand would enhance the systems performance.

Acknowledgements

The authors wish to thank the staff, particularly Dr X. Du and Dr J. K. Cruickshank, of the Medical School, University of Manchester for their help, and for providing data and explanations. We would also like to thank Prof. B. Richards of the School of Informatics for his valuable comments. Last but not least we wish to thank the anonymous reviewers for their very helpful comments that have improved the paper.

References

1. Pfaff M, et al. Prediction of Cardiovascular Risk in Hemodialysis Patients by Data Mining. Methods Inf Med 2004; 43: 106-13.
2. Richards G, et al. Data mining for indicators of early mortality in a database of clinical records. Artif Intell Med 2001; 22: 215-31.
3. Kamber M, et al. Generalisation and Decision Tree Induction: Efficient Classification in Data Mining. In Proc. Int'l Workshop Research Issues Data Engineering (RIDE'97) 1997; pp 111-20.
4. Ganti V, Gehrke J, Ramakrishnan R. Mining Very Large Databases. IEEE Computer, Special issue on Data Mining, 1999; 32 (8):38-45.
5. Kohavi R, Sommerfield D, Dougherty J. Data Mining using MLC++: A Machine Learning Library in C++. Tools with AI, 1996; pp 234-45.
6. Tjortjis C, Keane JA. T3: an Improved Classification Algorithm for Data Mining. Lecture Notes Computer Science Series, Springer-Verlag, Vol. 2412, 2002; pp 50-55.
7. Auer P, Holte RC, Maass W. Theory and applications of agnostic PAC-learning with small decision trees. In Proc. 12th Int'l Machine Learning Conf 1995; pp 21-9.
8. Du X, et al. Case-control study of stroke and the quality of hypertension control in North West England. BMJ 1997; 314: 272.
9. Tunstall-Pedoe H. Monitoring trends in cardiovascular disease and risk factors: the WHO MONICA Project. WHO. Chron 1985; 39: 3-5.
10. Bonita R, Beaglehole R. The enigma of the decline in stroke deaths in the United States: The search for an explanation. Stroke 1996; 27: 367-70.
11. Taylor TN, et al. Lifetime cost of stroke in the United States. Stroke 1996; 27: 1459-66.
12. Thorvaldsen P, et al. Stroke incidence, case fatality, and mortality in the WHO MONICA Project. Stroke 1995; 26: 361-7.
13. The Stroke Association, Preventing a stroke, Scriptographic Publ., www.stroke.org.uk, 2006.
14. Saraee M, Theodoulidis B. Knowledge discovery in temporal databases: the initial step. In: Proc. 4th Int'l Conf. Deductive Object-Oriented Databases 1995; pp 17-22.
15. Quinlan JR. C4.5: Programs for ML. Morgan Kaufmann, 1993.
16. Mehta M, et al. SLIQ: A Fast Scalable Classifier for Data Mining. In Proc. 5th Int'l Conf. Extending Database Technology 1996; pp 18-32.
17. UCI Machine Learning Repository. http://www.ics.uci.edu/~mlearn/MLRepository.html data sets converted to MLC++ format, http://www.sgi.com/tech/mlc/db/ (last accessed 12/05).
18. Quinlan JR. Improved Use of Continuous Attributes in C4.5, Journal AI Research 1996; 4: 77-90.
19. SGI MLC++ sources, http://www.sgi.com/tech/mlc/alpha/MLC1.3.1-src.tar.Z (last accessed 12/05).
20. Tjortjis C. T3: a classification algorithm for data mining. MPhil thesis, UMIST, UK, 1999.
21. Quinlan J R. Unknown attribute values in induction. In: Proc. 6th Int'l Machine Learning Workshop 1989; pp 164-8.
22. Murthy S, Saltzberg S. Decision Tree Induction: How effective is the Greedy Heuristic? In: Proc 1st Int'l Conf. KDD and DM 1995; pp 15-61.
23. T3, available at http://www.co.umist.ac.uk/~christos/TMiner.exe (last accessed 12/05).
24. Delen D, et al. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005; 34 (2): 113-27.
25. Kurgan L, Cios KJ, Sontag M, Accurso FJ. Mining the Cystic Fibrosis Data. In: Zurada J, Kantardzic M. (eds.). Next Generation of Data-Mining Applications, IEEE Press, 2005; pp 415-44.
26. Hand DJ, Mannila H, Smyth P. Principles of Data Mining. The MIT Press, 2001; p21.
27. Peduzzi P, Henderson W, Hartigan P, Lavori P. Analysis of Randomized Controlled Trials. Epidemiologic Reviews 2002; 24 (1): 26-38.
28. Schwarzer G, et al. Comparison of Fuzzy Inference, Logistic Regression, and Classification Trees (CART). Prediction of Cervical Lymph Node Metastasis in Carcinoma of the Tongue. Methods Inf Med 2003; 42: 572-7.
29. Abu-Hanna A, de Keizer N. Integrating classification trees with local logistic regression in intensive care prognosis. Artif Intell Med 2003; 29: 5-23.

Correspondence to:
C. Tjortjis
School of Computer Science
University of Manchester
P.O. Box 88
Manchester M60 1QD
UK
E mail: christos.tjortjis@manchester.ac.uk

Downloaded from www.methods-online.com on 2011-10-17 | IP: 88.104.15.66
For personal or educational use only. No other uses without permission. All rights reserved.