Soft Comput (2011) 15:1129–1136
DOI 10.1007/s00500-010-0587-x

FOCUS

Case study of inaccuracies in the granulation of decision trees

Salman Badr · Andrzej Bargiela

Published online: 10 March 2010

Abstract Cybernetics studies information process in the context of interaction with physical systems. Because such information is sometimes vague and exhibits complex interactions; it can only be discerned using approximate representations. Machine learning provides solutions that create approximate models of information and decision trees are one of its main components. However, decision trees are susceptible to information overload and can get overly complex when a large amount of data is inputted in them. Granulation of decision tree remedies this problem by providing the essential structure of the decision tree, which can decrease its utility. To evaluate the relationship that exists between granulation and decision tree complexity, data uncertainty and prediction accuracy, the deficiencies obtained by nursing homes during annual inspections were taken as a case study. Using rough sets, three forms of granulation were performed: (1) attribute grouping, (2) removing insignificant attributes and (3) removing uncertain records. Attribute grouping significantly reduces tree complexity without having any strong effect upon data consistency and accuracy. On the other hand, removing insignificant features decrease data consistency and tree complexity, while increasing the error in prediction. Finally, decrease in the uncertainty of the dataset results in an increase in accuracy and has no impact on tree complexity.

S. Badr · A. Bargiela (✉)
School of Computer Science, Faculty of Science,
University of Nottingham, Malaysia Campus,
43500 Semenyih, Malaysia
e-mail: abb@Cs.Nott.AC.UK

S. Badr
e-mail: salmanbadr@gmail.com

Keywords Rough set-based decision trees · Granulation · Accuracy · Complexity · Uncertainty · Attribute reduction

1 Introduction

Cybernetic systems focus on the representation and understanding of systems that have goals or decisions and participate in circular cause and effect relationships. The structuring of goals and decisions in a cause and effect relationship is something natural to decision trees, therefore, they can play a vital role in the application of machine learning to understand and develop cybernetic systems.

Because decision trees can get overly complex, it is vital that their complexity is controlled. Traditionally, the generalization of a complex decision tree has revolved around implementing pruning methodologies (Kweku-Muata 2007), such as reduced error pruning (REP), pessimistic error pruning, minimum error pruning (MEP), critical value pruning, cost-complexity pruning and error-based pruning (EBP). Studies comparing these techniques (Fierens et al. 2005; John 1989) have revealed that MEP and EBP tend to under-prune, whereas CVP and REP perform well in pruning and produce accurate results.

Apart from pruning, other researchers have focused their attention upon adjusting the input parameters used in the construction of C4.5 (Tusar 2007; Cherkauer and Shavlik 1996) or by generating a subset of rules from a set of numerous trees (Hall et al. 1998).

As granular computing groups entities based on their similarity, functional adjacency, indistinguishability or coherence (Bargiela and Pedrycz 2003), therefore it can be used as a powerful tool to reduce the complexity associated with large decision trees. Rough sets provide data reduction capabilities coupled with a straightforward interpretation

(Pawlak 1991) and hence used in the construction of decision trees as well as in the pre-processing of data.

With respect to the use of rough set theory in the construction of decision trees, (Han and Kim 2008) developed the entity attribute decision tree and the reduct attribute decision tree. Then, they showed their advantages of accuracy and rule simplification when compared with ID3 and C4.5.

Rough set theory has also been used to calculate the degree of dependence between the conditions and decision attribute. The condition with the highest significance is used as a splitting criteria in the construction of the decision tree (Huang et al. 2007; Wang and OU 2008).

Zhou et al. (2008) utilized rough sets to find redundant data and remove noise as a part of data pre-processing. Their comparisons with C4.5 showed that removal of irrelevant redundant attributes can increase the prediction accuracy of a decision tree. Similar works were carried out by (Yellasiri et al. 2005), which showed that rough sets which used pre-processed decision trees produced the highest accuracy when compared with ID3, C4.5 and CART.

The originality of the research reported here lies in the application of granular computing in real world cases: nursing home research has mostly relied on regression techniques and did not utilise rough sets and decision trees in prediction. In addition, it diverges from past research on rough sets and decision tree pre-processing by calculating significance factor and consistency rather than reducts and core for feature selection and reduction. Instead of comparing algorithms, we focus on exploring the relationships that exist between tree complexity, data uncertainty and accuracy.

2 Data source

2.1 Data description

The dataset used in the construction of decision trees was extracted from the Nursing Home Compare Database, which is available to download from the US Official Medicare Website (http://www.medicare.gov/default.asp). There are four parts to the database:

1. NHCAboutNH
2. NHCInspRes
3. NHCResidents
4. NHCStaff

NHCAboutNH provides information about the physical characteristics of a nursing home, such as their location, number of beds, ownership type, etc. NHCInspRes records the inspection results carried out by federal and state agencies once during a 15-month period or based on a complaint. NHCResidents provides description of a nursing home quality measures scores and finally, NHCStaff consists of the number of hours spent by various certified nurses, such as RN (registered nurses), CNA (certified nursing assistants) and LPNLVN (licensed practical nurse, licensed vocational nurse).

After combining all the tables, a total of 18 attributes were extracted. The selection process was guided by two main principles: (1) the attributes should be relevant and understandable to the residents searching for a high-quality nursing home and (2) the attributes have been shown to be significant through prior independent research done of deficiencies.

Attributes selected based on the criteria of resident interest were influenced by the Nursing Home Checklist available at the Medicare Website (http://www.medicare.gov/NHCompare/), which are shown in Table 1. Attributes selected based on the past research are shown in Table 2.

The majority of these attributes is self-explanatory. Few of them require further elucidation. *QISSurvey* indicates whether a comprehensive survey of the medical records and direct observations of the care of a large sample of residents was carried out or not. *SFF* lists whether a nursing home has a record of persistently performing poorly. *Severity* is an aggregation of the scope and level of harm of a deficiency and is represented as an alphabet letter, with 'A' being the least severe and 'L' being the highest. It was not directly used in the dataset, but functioned as a threshold value to limit the size of the dataset. *NumberRNHours, NumberCNAHours* and *NumberLPNLVNHours* are calculated by taking an average of the total number of hours worked by the respective nurse each day at the nursing home per resident, 2 weeks prior to the inspection and dividing the average by the total number of residents. *TotNumLicensedStaffHours* is an average of the total number of hours worked by the licensed staff each day at the nursing home per resident.

Table 1 Attributes selected based on the resident interest

Number	Attribute name
1	CertifiedNumberOfBeds
2	TotalNumberOfResidents
3	PercOfOccupiedBeds
4	LocatedWithinAHospital
5	CCRC (continuing care retirement community)
6	ResidentAndFamilyCouncils
7	QISSurvey (quality of care indicators)
8	SFF (special focused facility)
9	Severity
10	CategoryDesc (description of deficiency)

Table 2 Attributes selected based on the past research

Number	Attribute name
1	State
2	CategoryDescription (records the health plan of a nursing home)
3	TypeOfOwnership
4	MultiNursingHomeOwnership
5	NumberRNHoursPerResPerDay
6	NumberCNAHoursPerResPerDay
7	NumberLPNLVNHoursPerResPerDay
8	TotNumLicensedStaffHoursPerResPerDay

Records were retrieved using *Severity* set to >= 'A'. As each nursing home received multiple records, many duplicate instances of nursing homes were discovered. To remove this duplication and prevent the default hypothesis, only nursing homes having 'Electrical Deficiencies' were selected, resulting in 15,675 records and dividing the dataset into almost equal binary decisions: 8,177 records were 'No' and 7,498 had 'Yes'.

2.2 Cleaning skewness

(Fierens et al. 2005) showed that C4.5 is most affected by skewness, therefore the data were normalized. Three respective mathematical transformations: logarithmic, square root and inverse were applied on the skewed numerical variables to compare and select the transformation that provides the most normalized results. Among these attributes, *PercOfOccupiedBeds* was negatively skewed; *NumberCNAHours* had a near normal distribution, while the remaining were all positively skewed as shown in Table 3.

When a logarithmic transformation was applied to *TotalNumberOfResidents*, the skewness reduced to −1.11, however, a square root transformation produced a lower skewness of 0.654. The P–P plot of square root transformation produced a far more linear graph, than logarithmic transformation, and was hence chosen as the preferred transformation for *TotalNumberOfResidents*. Inverse transformation resulted in an increase in skewness of −15.23, therefore, it was not considered.

Table 3 Skewness of numerical variables

Attribute name	Skewness
TotalNumberOfResidents	3.223
CertifiedNumberOfBeds	3.011
PercOfOccupiedBeds	−1.697
NumberRNHours	4.970
NumberLPNLVNHours	2.617
NumberCNAHours	0.538
TotNumLicensedStaffHours	4.266

Table 4 Summary of transformations carried out on numerical variables

Attribute name	Transformation
TotNumberOfResidents	Square root
CertifiedNumberOfBeds	Square root
PercOfOccupiedBeds	Logarithmic
NumberRNHours	Inverse
NumberLPNLVNHours	Logarithmic
NumberCNAHours	Untransformed
TotNumLicensedStaffHours	Logarithmic

Because *PercOfOccupiedBeds* was negatively skewed, therefore, it was reflected by adding one to the absolute value of the maximum and subtracting it from all values of *PercOfOccupiedBeds*. Both logarithmic and square root transformations produced satisfactory results, with skewness reduced to −0.488 and 0.654, respectively. However, the P–P plots showed that logarithmic transformation produced a curve nearer to normal distribution. In the case of inverse transformation, the skewness was −2.99, higher than the original, −1.697, therefore it was rejected.

NumberCNAHours was already near normal, but for completeness, the transformations were applied. The minimum value for *NumberCNAHours* was 0; therefore one was added to all values. The results showed that logarithmic transformation produced a skewness of −0.993, square root transformation produced a skewness of −0.109 and inverse transformation produced a skewness of −4.23. As inverse transformation significantly increased skewness and therefore it was not accounted. P–P plots showed that both logarithmic and square root transformations did not significantly change the skewness of *NumberCNAHours,* therefore it was left untransformed. Table 4 provides a summary of the transformation carried out on all the numerical variables.

2.3 Missing values and outliers

Imputation of the median of ten nearby values was selected to remove the missing values because median has been shown to have less bias than mean (Refaat 2007) and it was assumed that nursing homes in the same state would probably have the same characteristics. The number ten was taken as a threshold value. Table 5 lists the statistics for the dataset before and after imputation.

Because some attributes did not have ten nearby examples to compare; hence, the process was repeated again with the threshold value decreased to 1 and thus

Table 5 Summary of missing attributes before and after imputation with the median of ten nearby values

Attribute name	Before imputation	After imputation
PercOfOccupiedBeds	99	0
NumberRNHours	891	2
NumberLPNLVNHours	891	2
NumberCNAHours	891	2
TotNumLicensedStaffHours	891	2

successfully removing all instances of missing values. The issue of outliers was not dealt with on the basis that outliers can provide information about decisions, which cannot be received otherwise.

3 Experimental design

From a total of 15,675 records in the dataset, two-third of the data was used for training and the remaining for testing. The training set consisted of 9,385 instances, whereas the testing set contained 6,290 cases. To evaluate accuracy, the mean absolute error (MAE) was used because it treats all values as equal without exaggerating the affects of outliers (Wittien and Frank 2005).

The MAE is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{1}$$

where f_i is the prediction and y_i is the true value.

For measuring the complexity, the tree size was used, which is the number of nodes and data precision was measured using consistency factor (CF) from rough set theory. CF is a ratio of the union of the set of lower approximation and the total number of examples in a decision system, *U,* and is represented as $\gamma(C, D)$, where C is the condition and D is the decision. If $\gamma(C, D) = 1$ then the decision system is considered consistent, meaning that it only contains unique rules. CF is formally written as:

$$\gamma(C, D) = \frac{|\text{POS}_C(D)|}{|U|} \tag{2}$$

where

$$\text{POS}_C(D) = \underset{X \in U/D}{\cup} \underline{C}(X) \tag{3}$$

$\text{POS}_c(D)$ being a union of the lower approximation, and hence called as 'positive region' of the partition *U/D* with respect to *C*. The significance of an attribute, *a,* is a measure of the change in the CF when the attribute in question is removed from the decision system. In rough set theory, it is calculated as:

$$\sigma_{(C,D)}(a) = 1 - \frac{\gamma(C - \{a\}, D)}{\gamma(C, D)} \tag{4}$$

If it equals 0, then the attribute is dispensable. Otherwise, it shows the percentage of rules in the decision system that would become inconsistent after the attribute's removal. For example, for a given *a*, if $\sigma_{(C,D)}(a) = 0.75$ then with the removal of *a,* 75% of the cases would become inconsistent.

To measure correlation, the Pearson correlation was used and scatter plots were used to confirm the relationship. The Pearson correlation test is a measure of the supposed linear relationship between two variables and returns a value between -1 and $+1$, indicating negative and positive correlation, respectively.

J48, which is the Weka implementation of the C4.5 algorithm, was used to build the decision trees. All of the default options were kept without change. An initial experiment was conducted on the complete dataset without any granularization. The results serve as a benchmark to compare with the other experiments. It led to a tree size of 1,649 having MAE of 0.4244 and having CF of 1.00.

3.1 Attribute binning

To enhance understanding, data were binned with the purpose of observing the affect of feature grouping on tree size, data consistency and error rate. Using information from the US Census Bureau on state and region division was used to group *State* to *Region*; http://www.census.gov/geo/www/us_regdiv.pdf.

For numerical attributes, values that fell within ± 1 standard deviation of the mean were binned into 'Average'. Values above and below ± 1 standard deviation was grouped into 'high' and 'low', respectively. After binning eight attributes, the tree size dramatically decreased from an initial size of 1,649 to 580 as shown in Table 6. Pearson correlation tests performed on the relationship between

Table 6 Summary of attribute binning on tree complexity, accuracy and consistency factor

Model	Attributes binned	Tree size	MAE	CF
1	0	1,649	0.4244	1.0
2	1	1,278	0.4654	1.0
3	2	1,182	0.464	1.0
4	3	1,218	0.4626	0.9998
5	4	1,117	0.4626	0.9996
6	5	1,116	0.4652	0.9996
7	6	1,105	0.4657	0.9982
8	7	751	0.4662	0.8643
9	8	580	0.4671	0.4052

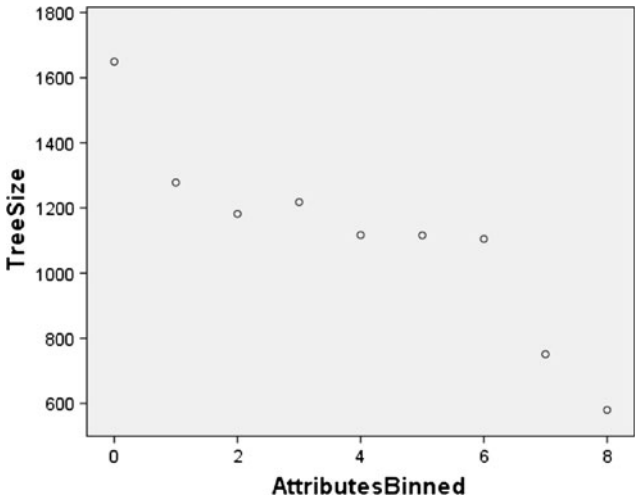

Fig. 1 Scatter plot showing the relationship between the number of attributes binned and tree size

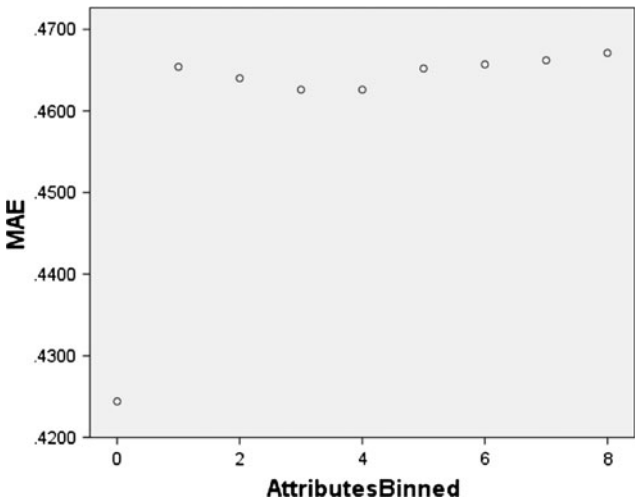

Fig. 2 Scatter plot showing the relationship between the number of attributes binned and MAE

attributes binning and tree complexity, MAE and CF revealed that a significant negative correlation exist between attribute binning and tree complexity (Fig. 1). The value for the Pearson correlation test was −0.914 with a *P* value of 0.01. However, no strong relationship was found between attribute binning and MAE and CF (Figs. 2, 3). The Pearson correlation result was 0.603 and −0.648, respectively.

Changing the order of the attributes binned will not have an effect on the above correlations because changing order does not affect the statistics of a particular attribute, as all the respective values are kept intact. The reason for the decrease in tree size is that with the binning of features the distribution of values decreases and correspondingly the number of splits in the decision tree. The concept can

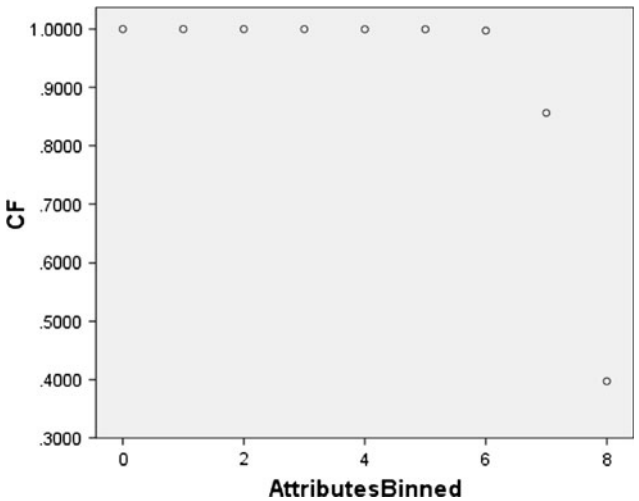

Fig. 3 Scatter plot showing the relationship between the number of attributes binned and CF

be illustrated in the case of grouping *States* to *Regions*, where the 50 values of *States* were grouped into five *Regions*. In the case of *State*, there existed 50 possible splits for the tree, while for *Regions* there existed only five possible splits.

A possible explanation for why binning did not have strong association with consistency lies in the fact that a single non-binned numerical attribute with high distribution can make a difference in the uniqueness of records, even though the remaining of the attributes were binned. For example, in Model 8 before binning *LogTotLicensedStaffHours* the CF was 0.8643. Afterwards, it fell to 0.4052.

3.2 Removing insignificant attributes

Table 7 lists the significance for the attributes after performing data binning. All attributes below 0.1 significance threshold were removed, one by one, and the respective change in consistency, complexity and error were observed. The reason for this test was to observe the behavior of attribute reduction on decision tree. A step by step removal of attributes allows a comparison on the number of attributes and different significant values upon the accuracy and complexity of the decision tree. For example, with a significance value of 0.03 only two attributes, *LocatedWithinAHospital* and *SFF*, are removed.

The results from the test are listed in Table 8. A preliminary observation of Table 8 reveals that as more attributes were removed the tree size and CF decreased, whereas the MAE increased. These observations were confirmed by the Pearson correlation tests and scatter plots (Figs. 4, 5, 6). The Pearson correlation value for the affect on tree size was 0.917, for MAE it was −0.931 and for CF it was 0.954. For each case, *P* value was <0.01.

Table 7 Attributes having the most significance on the consistency of the dataset are listed in descending order

Attribute name	Significance
TypeOfOwnership	0.32278
Region	0.28688
BinLogPercOfOccupiedBeds	0.18296
ResidentAndFamilyCouncils	0.17399
BinNumberCNAHours	0.17194
MultiNursingHomeOwnership	0.14486
BinInvRNHours	0.13290
BinLogLPNLVNHours	0.12045
BinLogTotLicensedStaffHours	0.07463
BinSqrtCertifiedBeds	0.07432
BinSqrtTotNumRes	0.07369
CategoryDescription	0.05983
CCRC	0.05873
QISSurvey	0.03763
LocatedWithinAHospital	0.02582
SFF	0.01763

Table 8 Summary of attribute removal on tree complexity, accuracy and consistency factor

Model	Attributes	Tree size	MAE	CF
9	18	580	0.4671	0.4052
10	17	568	0.4669	0.3980
11	16	620	0.4681	0.3876
12	15	567	0.4692	0.3716
13	14	481	0.4708	0.3466
14	13	448	0.47	0.3159
15	12	456	0.4698	0.2826
16	11	314	0.4719	0.2073
17	10	252	0.4717	0.1704

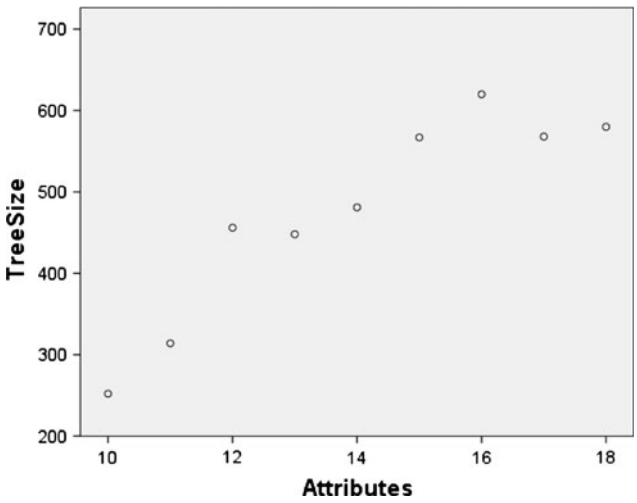

Fig. 4 Scatter plot showing the relationship between the number of attributes and tree size

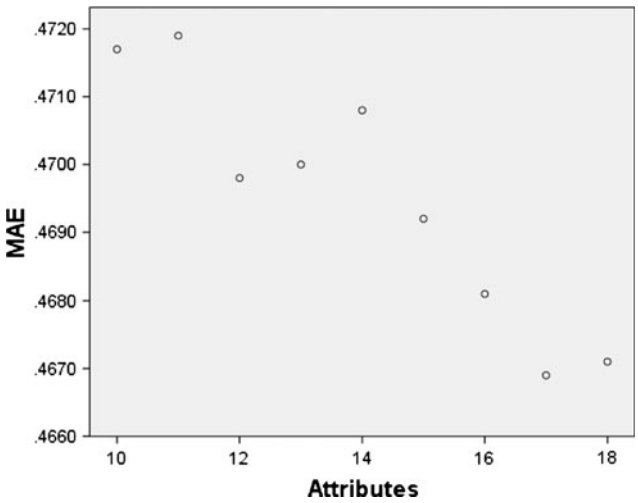

Fig. 5 Scatter plot showing the relationship between the number of attributes and mean absolute error

The decrease of the tree size as a result of the reduction of the number of attributes is fully expected because the branches in a decision tree are built using gain information from attributes. The fewer attributes there are in the dataset the smaller is the number of branches in the decision tree. The reason for the decrease of consistency is that the larger number of attributes provides greater possibility to differentiate records. Decrease of consistency may also be contributing to an increased error. This possibility will be explored next.

3.3 Increasing data consistency

Cases that were in the boundary region before attribute reduction specified in Set 3 were cumulatively removed, and the model was tested to compare the effect of increase in data consistency on prediction error and tree complexity. Because consistency in rough set theory is described as a ratio of the set of lower approximations and the total instances, removing boundary cases increases the consistency by decreasing the total number of instances, while maintaining the lower approximation. After binning attributes, a total of 9,324 records were in the boundary region. They were removed at an increment of 1,036. Each time 60% of the data was used to train and the remaining 40% for testing.

The results from increasing data consistency through the removal of boundary cases, Table 9, did not show any significant correlation with tree complexity. In Model 9, when the CF was 0.4052 the tree size was 580 and after reaching 1.0, the tree size became 573, which is not a

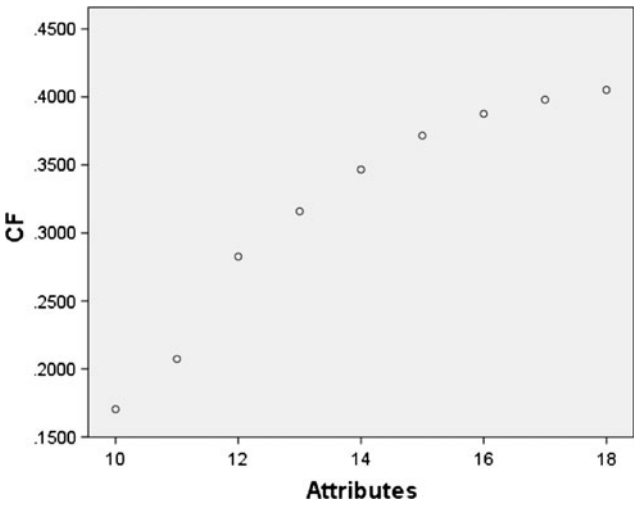

Fig. 6 Scatter plot showing the relationship between the number of attributes and consistency factor

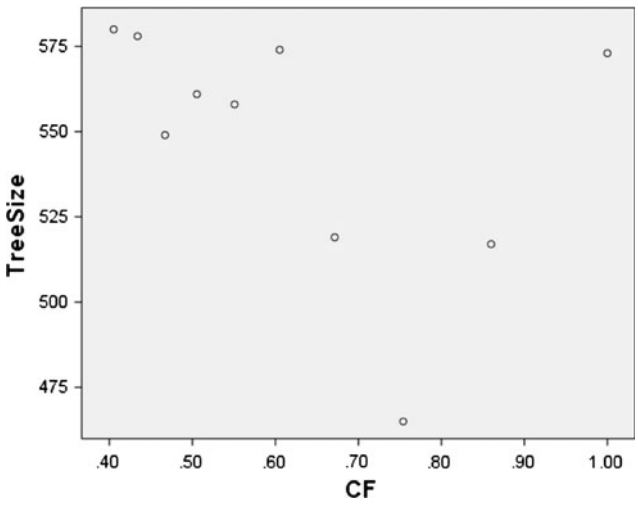

Fig. 7 Scatter plot showing the relationship between consistency factor and tree size

Table 9 Summary of increase in consistency factor, tree complexity and accuracy

Model	Number of instances	CF	Tree size	MAE
9	15,675	0.4052	580	0.4671
18	14,639	0.4339	578	0.4656
19	13,603	0.4670	549	0.461
20	12,567	0.5054	561	0.462
21	11,531	0.5508	558	0.4602
22	10,495	0.6052	574	0.4573
23	9,459	0.6715	519	0.4531
24	8,423	0.7540	465	0.4501
25	7,387	0.8599	517	0.437
26	6,351	1.0	573	0.4246

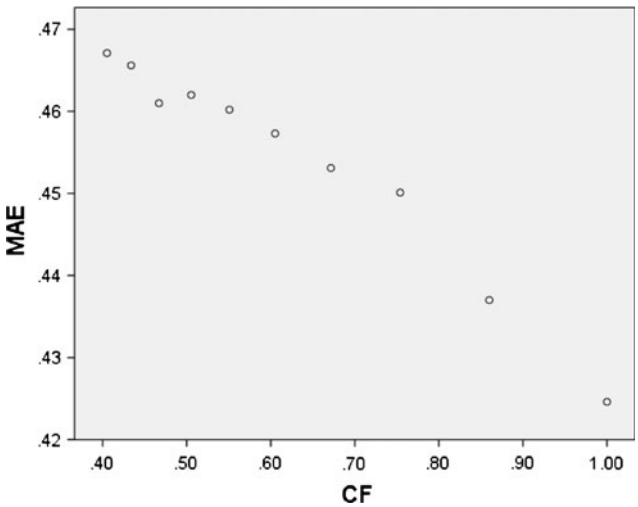

Fig. 8 Scatter plot showing the relationship between consistency factor and mean absolute error

significant difference. Pearson correlation test had a value of -0.397, with a $P > 0.01$. This was confirmed in the scatter plot (Fig. 7). In contrast, strong negative correlation was discovered between CF and MAE (Fig. 8). The Pearson correlation test had a value of -0.979 with a $P < 0.01$. This result shows that as concepts become more vague; the accuracy in prediction falls.

4 Conclusion

Within cybernetic systems, a trade-off has to be made in reducing the complexity of information, maintaining its consistency and generating accurate predictions. Complexity, consistency and accuracy exhibit complex interactions and relationships between one another and the type of granulation carried out can increase or decrease these factors.

Using concepts of consistency and significant factor from rough set theory, the effect of granulation upon tree complexity, data consistency and accuracy was studied in the case of predicting nursing home deficiences. Three forms of granulation were performed: attribute binning, removing insignificant attributes and removing inconsistent records. The results showed that attribute binning decreases tree size and has no significance on accuracy or data consistency. A decrease in insignificant attributes decreases tree complexity and data consistency, while increasing prediction error. Finally, removing inconsistent records decreases error, while having no significance on tree complexity.

This case study research points to two possible extensions of information processing paradigms in cybernetic

systems. Fuzzy sets can be employed in grouping of attributes and secondly, the effect of the number of groups on tree complexity and accuracy can be studied.

References

Bargiela A, Pedrycz W (2003) Granular computing: an introduction. Kluwer Academic Publishers, Dordrecht

Cherkauer KJ, Shavlik JW (1996) Growing simpler decision trees to facilitate knowledge discovery. In: Proceedings of the second international conference on knowledge discovery and data mining, pp 315–318

Fierens D, Ramon J, Blockeel H, Bruynooghe M (2005) A comparison of approaches for learning first-order logical probability estimation trees. LNCS 3720:556–563

Hall LO, Chawla N, Bowyer KW (1998) Decision tree learning on very large data sets. IEEE Int Conf Syst Man Cybern 3:2579–2584

Han SW, Kim JY (2008) A new decision tree algorithm based on rough set theory. Int J Innov Comput Inf Control 4:2749–5757

Huang L, Huang M, Guo B, Zhang Z (2007) A new method for constructing decision tree based on rough set theory. IEEE Int Conf Granular Comput 241–244

John M (1989) An empirical comparison of pruning methods for decision tree induction. Mach Learn 4:227–243

Kweku-Muata O-B (2007) Post-pruning in decision tree induction using multiple performance measures. Comput Oper Res 34:3331–3345

Pawlak Z (1991) Rough sets: theoretical aspects of reasoning about data. Kluwer Academic Publishers, Dordrecht

Refaat M (2007) Data Preparation for Data Mining Using SAS, Morgan Kaufmann

Tusar T (2007) Optimizing accuracy and size of decision trees. In: Proceedings of the sixteenth international electronical and computer science conference-ERK 2007, pp 81–84

Wang C, Ou F (2008) An algorithm for decision tree construction based on rough set theory. In: International conference on computer science and information technology, pp 295–298

Wittien IH, Frank E (2005) Data mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers, California

Yellasiri R, Rao CR, Reddy V (2005) Decision tree induction using rough set theory-comparative study. J Theor Appl Inf Technol 3:110–114

Zhou X, Zhang D, Jiang Y (2008) A new credit scoring method based on rough sets and decision tree. LNCS 5012:1081–1089