Using Supervised and Unsupervised Techniques to Determine Groups of Patients with Different Doctor-Patient Stability

Eu-Gene Siew[1], Leonid Churilov[2], Kate A. Smith-Miles[3], and Joachim P. Sturmberg[4]

[1] School of Information Technology, Monash University, Sunway Campus, Jalan Lagoon Selatan, 46150 Bandar Sunway, Selangor D.E.
siew.eu-gene@infotech.monash.edu.my
[2] National Stroke Research Institute, Heidelberg Heights, Victoria 3081, Australia
lchurilov@nsri.org.au
[3] Faculty of Science and Technology, Deakin University, Burwood, Victoria 3125, Australia
Kate.Smith-Miles@deakin.edu.au
[4] Department of General Practice, Monash University, Victoria 3800, Australia
jp.sturmberg@bigpond.com

Abstract. Decision trees and self organising feature maps (SOFM) are frequently used to identify groups. This research aims to compare the similarities between any groupings found between supervised (Classification and Regression Trees - CART) and unsupervised classification (SOFM), and to identify insights into factors associated with doctor-patient stability. Although CART and SOFM uses different learning paradigms to produce groupings, both methods came up with many similar groupings. Both techniques showed that self perceived health and age are important indicators of stability. In addition, this study has indicated profiles of patients that are at risk which might be interesting to general practitioners.

Keywords: Doctor-patient stability (MCI), Classification and Regression Trees (CART), Self Organising Feature Maps (SOFM or SOM), supervised learning, unsupervised learning.

1 Introduction

Two of the most popular methods for classifying and clustering data are decision trees and self-organising maps [1]. Classification and regression tree (CART) is used to classify the data while Kohonen's self-organising map (SOFM) is to cluster. These techniques differ in their approach to grouping patients in that CART [2] uses a supervised learning approach that requires a target variable to guide its groupings, whereas SOFM [3] uses an unsupervised learning approach by grouping patients without the need to specify the desired output.

This paper compares the similarities between any groupings found between supervised and unsupervised techniques, and to identify insights into factors associated with doctor-patient stability.

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 715–722, 2008.

Long-term doctor-patient stability is an important aspect to achieving continuity of care [4] Continuity of care has many benefits. It has been shown to build patients' personal trust in doctors [5], to increase the knowledge of doctor and patient about each other which in turn promotes an increased understanding of the social context of the patient [6] and has been shown to be vital to patient satisfaction [7].

Overall, research in this area of doctor-patient stability mostly treats patients as a single homogenous group [8, 9]. The factors are typically treated as having a one-to-one linear relationship to the outcome doctor-patient relationship stability variable.

An additional objective of this paper is to investigate the groups of patients produced by CART and SOFM and to evaluate these groups in terms of predicting doctor-patient stability.

This paper is organized as follows: Section 2 describes the study design; research methodology is discussed in Section 3; Sections 4 and 5 present SOFM and CART results respectively; Section 6 demonstrates how the results are validated; Section 7 describes the key profiles and comparisons between SOFM and CART, while the conclusions are made in Section 8.

2 Study Design

The data is obtained from a survey of randomly selected general practices in the NSW Central Coast, Australia. This region is estimated to have up to 230,000 people and ranks as top ninth highest population in Australia [10]. The practices in the area (n=93), were stratified into five classes according to their size, which is categorized into solo, 2, 3 to 4, and 5 and over, doctors. 100 consecutive patients are selected from the five practices of each of the five classes. In total, twenty of the sixty-one doctors (which constitute about 33 per cent) agreed to participate. Due to the high demand placed on doctors and their patients, eight doctors who initially agreed withdrew from the study. Information about 1,122 patients and their respective doctors was collected. Data collection occurred between February and November 1999.

Table 1. Data dictionary and average characteristics of the items contained in the questionnaire

Section	Abbreviation	Mean	Everywhere except for the Age variable, the score of "1" means that…
Pre-consultation items	Time	0.75	Doctor always has enough time for me
	Age	49.98	Age (years)
	Knowdo	0.68	Knows doctor well
	Health	0.74	Patient perceived to be in excellent health
	Psysym	0.52	Psychological distress
	Soc	0.48	Social distress
	Morbidity	0.15	Poly-morbidity
Consultation items	Condif	0.28	Most difficult consultations
	MCI	0.68	Most stable doctor-patient relationship
	Consl	0.87	Longer Consultations
Post consultation items	Commun	0.6	Excellent communication with doctor
	Enable	0.31	Highest enablement
	Satisf	0.8	Highest satisfaction

The questionnaire is divided into three parts: the first were answered by the doctor, the second by the patients before consultation and the final part by the patients after consultation. The questionnaires obtained information about the health service environment, the doctor's characteristics and perceptions about the patient, patient characteristics, information about the consultation process and the outcome. Only relevant variables are shown in Table 1 which indicates the average mean values of each questionnaire variables and describes the abbreviation used.

Doctor-patient stability variable is measured as modified continuity index (MCI). MCI is developed by Godkin and Rice and it indicates the frequency and intensity of the relationship by dividing the number of different doctors visited by the number of visits in a time period [11]. It is a continuous number between 0 and 1 and is the frequency visit to a dominant doctor over the number of visits in a year. Values close to 0 would indicate poor doctor-patient stability and 1, high doctor-patient stability.

3 Research Methodology

The research design contains the following stages:

Stage 1: Application of CART and SOFM using the training data sets.

At the first stage, about 20 per cent of patients were randomly allocated into evaluation set and the rest into training set. Both data mining techniques (CART and SOFM) were applied separately to group the general practice patients based on demographics and clinical variables using the training data set. For SOFM, a software package called Viscovery was used to model the data [12].

Stage 2: Validation of the CART and SOFM models using evaluation set (holdout sample).

The models generated in the training set were then applied to the relevant evaluation set. If the models were generalisable then the performance of the evaluation sets were analogous to the training period. To make the comparison, Mean absolute deviation (MAD) [13] and the coefficient of multiple determination (R^2) were used.

Stage 3: An analysis and comparison of the results from supervised and unsupervised data mining techniques.

4 SOFM Clusters

This section describes the application of SOFM onto the training data set of which SOFM generated 10 clusters (Stage 1) which were then renumbered in ascending order of MCI. There were three broad groups of clusters: those with MCI of 0.5 to 0.6, MCI of 0.6 to 0.76 and MCI over 0.76. When the clusters were created, the stability variable was left out as inputs. This was to see how well other variables were able to predict stability.

There seems to be a strong correlation between age and stability. If the patient's age is between 30 to 38, they are likely to group in Cluster 1 to 4 and are likely to have MCI between 0.5 to 0.58.

There also seems to be a separation between two main groups of patients. Those groups (Cluster 5-10) whose average stability is 0.73 and above (high stability) and those (Cluster 1-4) whose average stability is 0.58 and below (low stability). The following describes those clusters:

Table 2. Summary of SOFM clusters based on MCI

Variables	Low stability				High stability					
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Age	31	35	39	32	53	63	63	57	69	69
Health	Good	Good	Good	Good				Poor	Good	
Morbidity		Complex	Social		Psychological			Complex		Physical
Condif	Difficult	Easy		Easy			Difficult	Easy	Easy	Easy
MCI	0.53	0.54	0.55	0.58	0.73	0.76	0.76	0.81	0.81	0.83
Consl	Shortest		Long		Longest	Short				Long
Commun		Good	Good	Poor		Poor		Good		Good
Enable	Lowest	Good	Poor	Poor		Poor				Good
Satisf		Good	Good	Poor		Poor	Good	Good	Good	Good
Cluster Total	47	114	130	63	44	54	96	84	106	73

After the profiles of the stability clusters were examined, SOFM was used as a prediction tool. The MAD and R^2 in the training set were 0.1622 and 0.2356 respectively.

5 CART Results

This section describes the results obtained from applying CART onto the training data set (Stage 1). A separate study was conducted on the patient stability variable for which CART generated 7 terminal nodes. Figure 1 shows the CART tree diagram of stability when it was run on the training data set. Terminal node 1 has the lowest stability, as measured by the MCI index, while Terminal node 7 has the highest.

CART uses three patient variables: patient’s age, knowledge of the doctor and perception of their health. In general, patients whose average age is 46.5 and less tend to have lower doctor-patient stability. Patients who consider themselves in poor health tend to have more stable doctor-patient relationships. In addition, younger and healthier patients (represented in Terminal node 1) have a lower MCI score compared with younger but not-healthier patients (represented in Terminal node 4).

Patients with a high level of knowledge of their doctor are correlated with high stability. Terminal node 5 (with good knowledge of their doctor) has a higher stability score than Terminal node 2 (with poor knowledge of their doctor) even though both represent ages between 46.5 to 64.5 and good self-perceived health.

Like SOFM, there also seems to be a separation between two main groups of patients which are those groups (Cluster 5-10) whose stability is above 0.69 and those (Cluster 1-4) whose stability is 0.69 and below. Age, health and knowledge of doctor

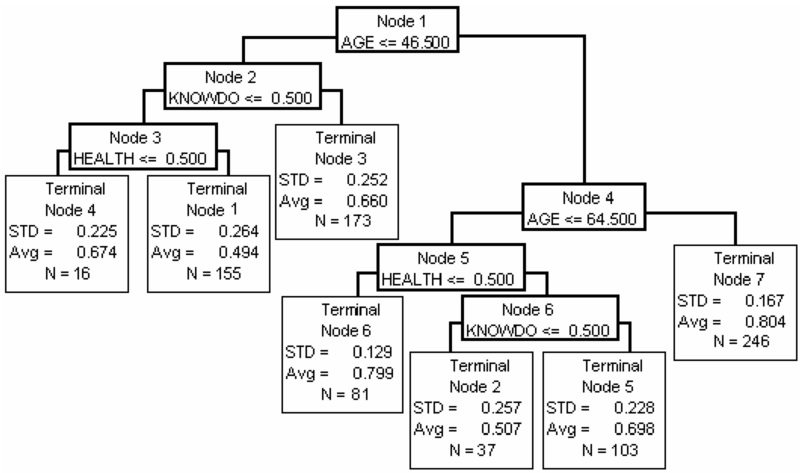

Fig. 1. CART tree of doctor-patient stability in primary care

variables can also be used to separate those 2 groups. Patients age 46.5 years and below, and also those who consider their health to be good but do not have a good knowledge about their doctor has a low stability. Once the profiles of the CART nodes were examined, CART was used as a prediction tool. The MAD and R^2 in the training set were at 0.1227 and 0.3591 respectively.

6 Validation of SOFM and CART Groupings

A comparison was made between the training and evaluation data sets to establish the generalisability of both the SOFM and CART grouping models of patient stability. A significantly higher MAD and lower R^2 in the evaluation set, compared to the training set, would indicate poor applicability of the model. Table 3 shows the comparison between MAD and R^2 of training and evaluation set of both SOFM and CART. Using an unpaired two-tailed t-test and the alpha level of .05, the null hypothesis that the MAD of the training and evaluation sets is statistically similar could not be rejected. The p-value was above 0.05 for both SOFM and CART.

Both these measures provide evidence that the SOFM model of clusters and the CART rules produced could be generalisable in creating patient groups that reflect doctor-patient stability.

Table 3. Comparison of MAD and R^2 for stability variable

		MAD			R^2	
Variable	*Method*	**Training set**	**Evaluation set**	*2 tail P-value*	**Training set**	**Evaluation set**
Stability	**SOFM**	0.1622	0.1827	0.0684	0.2356	0.2438
	CART	0.1227	0.1346	0.1394	0.3591	0.3977

7 Similarities between Supervised and Unsupervised

Guthrie and Wyke state that for some groups of patients stability is more important [14]. They list an example of a more serious morbidity group of patients that requires higher doctor-patient stability compared with the healthier groups. Thus, the interesting groups would be patients who have serious morbidity but for some reason choose not to have a usual general practitioner. These SOFM groupings are:

Cluster 1: Young patients (average 31.5 years) with complex morbidity and poor communication with their doctor, who have the lowest doctor-patient stability. They are the least enabled and most dissatisfied with their consultations. They represent the highest proportion that judge themselves in excellent health. Doctors find consultations with this group the most difficult.

Cluster 4: Young patients (average 32.0 years) in social distress who are dissatisfied with their consultations. This cluster has one of the lowest rates of enablement and satisfaction. Although they consider themselves in good health, they are oblivious to their social distress and are unable to understand and communicate with their doctor. They also have the second shortest consultation times and felt their doctor does not spend enough time with them.

Cluster 6: Older patients (average 62.6 years) who have negative attitudes towards holistic health care with combinations of morbidity. They have problems communicating with their doctor and have amongst the shortest consultations. They feel not enabled by and were dissatisfied with their consultation.

Those findings are to some extent consistent with CART which ranks age, self perception of health and social morbidity highly as important primary or surrogate splitters.

In addition, a comparison using Cohen Kappa[15] seems to show that SOFM and CART produced similar groupings. As mentioned earlier in Section 4 and 5, both CART and SOFM came up with two groupings of high and low stability groupings. An average of MCI 0.69 can be considered as a threshold for the broad groupings.

An assessment of inter-rater reliability using Cohen Kappa is shown in the tables below. Both SOM and CART are considered as the "raters" of the two categories based on high doctor-patient stability (above MCI 0.69) and low doctor-patient stability (MCI 0.69 and below).

Table 4. Degree of agreement between SOFM and CART

	SOFM	**Clusters 1-4**	**Clusters 5-10**	**Total**
CART		Avg MCI <0.69	Avg MCI >0.70	
Nodes1-4	Avg MCI <0.69	**325**	56	381
Nodes5-7	Avg MCI >0.70	29	**401**	430
Total		354	457	811

Table 5. Expected values in each cell if it were due by chance

	SOFM	Clusters 1-4	Clusters 5-10	Total
CART		Avg MCI <0.69	Avg MCI >0.70	
Nodes1-4	Avg MCI <0.69	**166.31**	214.69	381
Nodes5-7	Avg MCI >0.70	187.69	**242.31**	430
Total		354	457	811

The total agreement between CART and SOFM is 726 compared with the agreement if it were due to chance of 408.61. The Cohen Kappa in this case is 0.79 which seems to indicate that both SOFM and CART is similar to broadly group doctor-patient stability. If the groupings were due to chance, Cohen Kappa would be 0.

8 Summary and Conclusions

This paper has discussed the use of CART and SOFM to classify patients according to their stability of doctor-patient relationship. The contribution of this research is to identify groups of patients that are at different levels of stability. By doing this, it reveals key variables and profiles that are associated with the stability outcome and highlight high risk groups. There were groupings of patients with combinations of morbidity who, for some reason, consider themselves to be in good health. They do not have a principal general practitioner who can provide continuous care for them.

In addition, this research compares the performance of supervised and unsupervised learning. Both are able to come up with similar groupings based on Cohen Kappa and key attributes which are age, self perception of health and social morbidity.

There are limitations to this research. It is arguable whether the results could be applied outside the New South Wales Central Coast. In general, the central coast region tends to have patients who are predominantly native speakers of English, with less social mobility and with less availability of doctors.

Furthermore, the data on doctors and the general practice are limited. There are only twelve doctors that took part in the survey. A larger sample size of doctors and general practice would enable more association of their variables to patients. It is also probable that particular groups of doctors or even patients may be omitted from the final results because they did not participate in the survey when sampled.

Future research might include open ended questions targeting dissatisfied patients and in particular those unable to communicate with their doctors. These questions may elicit the reasons underlying the poor communication, such as poor doctor training, patient not being able to voice their opinion and doctors who felt rushed to complete as many consultations as possible.

References

1. Goebel, M., Gruenwald, L.: A Survey of Data Mining software Tools. SIGKDD Explorations 1(1), 20–33 (1999)
2. Steinberg, D., Phillip, C.: CART–Classification and Regression Trees, Salford Systems, San Diego, CA (1997)
3. Deboeck, G., Kohonen, T.: Visual Explorations in Finance with Self-Organizing Maps. Springer, London (1998)
4. Freeman, G., Hjortdahl, P.: What future for continuity of care in general practice? British Medical Journal 314(7342), 1870–1873 (1997)
5. McWhinney, I.R.: Core values in a changing world. British Medical Journal 316, 1807–1809 (1998)
6. Gulbrandsen, P., Hjortdahl, P., Fugelli, P.: General practitioners' knowledge of their patients' psychosocial problems: multipractice questionnaire survey. British Medical Journal 314, 1014 (1997)
7. Mansour, A., al-Osimy, M.H.: A study of satisfaction among primary health care patients in Saudi Arabia. Journal of Community Health 18, 163–173 (1995)
8. Hjortdahl, P., Laerum, E.: Continuity of care in general practice: effect on patient satisfaction. British Medical Journal 304(6837), 1287–1290 (1992)
9. Hjortdahl, P., Borchgrevink, C.F.: Continuity of care: influence of general practitioners' knowledge about their patients on use of resources in consultations. British Medical Journal 303(6811), 1181–1184 (1991)
10. New South Wales, http://www.geocities.com/lockstar/newsouthwales.html
11. Godkin, M.A., Rice, C.A.: A measure of continuity of care for physicians in practice. Family Medicine 16(4), 136–140 (1981)
12. VISCOVERY software, http://www.eudaptics.com
13. Weiss, S.M., Zhang, T.: Performance Analysis and Evaluation. In: Ye, N. (ed.) The handbook of Data Mining, pp. 425–440. Lawrence Erlbaum Assoc, Mahwah (2003)
14. Guthrie, B., Wyke, S.: Does continuity in general practice really matter? British Medical Journal 321, 734–736 (2000)
15. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46 (1960)