SOM for classifying data sets with missing values - Application to clinical data of bladder cancer patients -

Takashi Yamaguchi Kenneth J. Mackin

Department of Information Systems,
Tokyo University of Information Sciences
1200-2 Yatou-cho, Wakaba-ku, Chiba, 265-8501 Japan
(Tel : 043-236-1329)
(mackin@rsch.tuis.ac.jp)

Abstract: In this paper we investigate applying SOM(Self-Organizing Maps) for classification and rule extraction in data sets with missing values, in particular from real clinical data of bladder cancer patients. For this experiment, we used real data of bladder cancer patients.

When using input data with missing values for SOM, the missing value is either interpolated in the preprocessing stage, or the missing value is replaced with a specific value or property that marks it as a missing value. In either case, there is a possibility some rules can be extracted from data with missing values. On the other hand, these data can have a negative influence for the classification for data sets for which missing values should be neglected.

In this research we propose a method where SOM is trained using an input vector in which the properties for the missing values are excluded. The influence of information on the missing values can be reduced by using the proposed method. Through computer simulation, we showed that the proposed method gave good results in classification and rule extraction from clinical data of bladder cancer patients.

Keywords: Self-Organizing Maps, classification, missing values, clinical data, bladder cancer,

I. INTRODUCTION

Bladder cancer is a cancer that has the highest mortality rate following prostate cancer in the field of urology. Because a tumor marker like PSA(Prostate Specific Antigen) for prostate cancer does not exist, an effective method for prediction has not been established. We aim at analyzing and extracting rules from clinical data of bladder cancer patients. The final goal of the research is to establish a reliable prediction method.

For this experiment, we used real data of bladder cancer patients provided by Kitasato University hospital. The clinical data consists of properties with both continuous values and discrete values. A characteristic of this clinical data is that there are a large number of missing values.

In this paper we investigate applying Self-Organizing Maps (SOM) [1] for classification and rule extraction in data sets with missing values, in particular from clinical data of bladder cancer patients.

Samuel Kaski [4], T. Samad and S.A. Harp [3] proposed a method where SOM is trained using an input vector in which the properties for the missing values are excluded. But this method has a possibility that topological information is not correctly learned.

In this research we propose a method for training the data set without missing values at the early stage of training, and training with the data set including missing values afterwards

II. METHOD

SOM is a type of artificial neural networks. It is trained using unsupervised learning to produce low dimensional representation of the training samples while preserving the topological information of the input space.

There are 2 steps in SOM training; determining the winner unit and updating the weight vector. The winner unit is a unit on the competitive layer with the weight vector in which the distance is best matched for given input vector. In the step of determining the winner unit, it is kth unit that minimizes the degree of match difference m^k is selected. When distance measure used Euclidean distance, the degree of match difference m^k of weight vector w_k to an input vector x_i is defined as follows.

$$m_k = \sum_{i=1}^{n} (x_i - w_{ki})^2 \qquad (1)$$

Where n is the dimension of the input vector, x_i is the value of the ith input, and w_{ki} is the value of weight between the ith input and kth competitive layer's unit.

In the step of updating the weight vector, the weight vectors of the winner unit and its neighbors on the competitive layer is updated. The weight modification is show as follows.

$$\Delta w_{ki}(t) = h_{ci}(t) \cdot \left(x_i(t) - w_{ki}(t)\right) \qquad (2)$$

$h_{ci}(t)$ is called the neighborhood function. For the neighborhood function, we used the following Gaussian type function.

$$h_{ci} = \alpha(t) \cdot \exp\left(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right) \qquad (3)$$

Where r_i and r_c are coordinates in the competitive layer, and $\alpha(t)$ and $\sigma(t)$ are monotonic decreasing parameters.

When using input data with missing values for SOM, the missing value is either interpolated in the preprocessing stage, or the missing value is replaced with a specific value or property that marks it as a missing value. In either case, there is a possibility some rules can be extracted from data with missing values. On the other hand, these data can have a negative influence for the classification and the rule extraction for data sets for which missing values should be neglected.

Samuel Kaski [4], T. Samad and S.A. Harp [3] extended the SOM method for more flexibly using input with missing values. The extended SOM method is trained using a vector in which the properties for the missing values are excluded.

The degree of match m^k for determining the winner unit is show as follows.

$$m_k = \sum_{i \in P_t} \left(x_i - w_{ki}\right)^2 \qquad (4)$$

Where P_t is the set of input units in which input values x_i are not missing value at time t.

The weight modification is show as follows.

$$\Delta w_{ki}(t) = \begin{cases} h_{ci}(t) \cdot \left(x_i(t) - w_{ki}(t)\right) & if \quad i \in P_t \\ 0 & otherwise \end{cases} \qquad (5)$$

The influence of information on the missing values can be reduced by using this method.

When the ith input unit has missing value, the following expression (6) is derived from expression (5).

$$x_i(t) = w_{ki}(t) \qquad (6)$$

This means that the missing input value x_i is interpolated by the corresponding weight value w_{ki} at time t. Because the weight space is not necessarily corresponding to the input space, this method has a possibility that topological information is not correctly learned when data with the missing values is input at the early stage of training.

In this research we proposes a method where SOM is trained using the data set without missing values at the early stage of training, and the data set including missing values is used for training in the later stages.

$$x = \begin{cases} \in \Re'^n \in \Re^n & if \quad t < \tau \\ \in \Re^n & otherwise \end{cases} \qquad (7)$$

$\Re^n$ is the set of n dimensional input vectors is including missing values, $\Re'^n$ is the set of n dimensional input vectors that dose not including missing values, and τ is the parameter that defines the time until training with missing values. There are 2 stages in the learning process of SOM, the early stage and the latter stage. The map is adapted in the early stage, and the map is fine-tuned in the latter stage. In the proposed method, we aim at reducing the influence of the missing values by not using the missing data set in the early stage of training.

III. RESULT

1. Clinical data of bladder cancer patients

For this experiment, we used real data of bladder cancer patients provided by Kitasato University hospital. This data of bladder cancer patients based on the clinical record for transurethral resection for bladder tumor (TUR-Bt). A characteristic of this clinical data is that there a large number of missing values. 83% of the instances include missing values.

The clinical data consists of 145 patient instances, and has 28 properties with both continuous values and discrete values. Each instance is data of a unique patient, and there is no time series relation between each instance. 28 properties consist of 22 sets of the Boolean value type [True, False], 4 continuous value types and 2 date types. The complete list of the properties is shown in Table.1.

It is assumed that an immediate correlation dose not exist in "Operation day" and "Last survival confirmation day", therefore is not used for the input to SOM. "Last state" is stored as Boolean values of 4 types, "survival", "cancer-specific death", "death by other cause" and "unknown". We examine the prediction by analyzing the correlation with "Last state" and the other properties.

In the data preprocessing, we prepared 2 data sets for standard SOM method and extended SOM method.

For the data set for standard SOM method, the missing continuous value was replaced by 0.5. For the missing discrete values, the Boolean value of the property "unknown" is set to True

For the data set for extended SOM method, a Boolean value data that marks the missing value corresponding to each property of each instance was separately prepared.

Next, the continuous value properties were normalized to [0.0-1.0], and Boolean value properties were converted to [0.0, 1.0], and input to SOM as numeric data. The data set for standard SOM method consisted of 145 instances with 109 properties. The data set for extended SOM method consisted of the same 145 instances with 91 properties.

Variables	Type of Value (number of Boolean values)
Gender	Set of Boolean values(2)
Age	Continuous
Frequency of TUR	Continuous
Position of tumor	Set of Boolean values (10)
Number of tumor	Set of Boolean values (3)
Form of tumor	Set of Boolean values (7)
Size of tumor	Set of Boolean values (4)
Histological tissue	Set of Boolean values (4)
Grade	Continuous
pT stage	Set of Boolean values (8)
pV stage	Set of Boolean values (4)
pL stage	Set of Boolean values (5)
pN stage	Set of Boolean values (6)
pM stage	Set of Boolean values (11)
INF	Set of Boolean values (4)
Position of pT4	Set of Boolean values (11)
Concomitant CIS	Set of Boolean values (3)
Concomitant prostate cancer	Set of Boolean values (3)
Concomitant urinary bladder cancer	Set of Boolean values (3)
BCG therapy	Set of Boolean values (3)
Neo adjuvant chemotherapy	Set of Boolean values (4)
Adjuvant chemotherapy	Set of Boolean values (4)
Radiotherapy	Set of Boolean values (3)
Postoperative Day	Continuous
Operation day	Date
Last survival confirmation day	Date
Last state	Set of Boolean values (4)
Type of urinary diversion	Set of Boolean values (4)

Abbreviations: TUR, transurethral resection; pT, primary tumor; pV, vein invasion; pL, lymphovascular invasion; pN, regional lymph nodes; pM, distant metastasis; INF, Infiltrative growth pattern; CIS, carcinoma in situ; BCG, Bacille Calmette-Guérin;,

Table.1. The list of properties

2. Experimental result

For this experiment we used following SOM. 30x20 competitive layer was initialized by random. For the determining of winner unit we use Euclidian distance. For neighborhood function we use the Gaussian type (α_1=1.0, α_f=0.999, σ_1=30, σ_f=0.999). The number of training steps T is T=5000.

In this experiment we compare following 4 different conditions, standard method, proposed method (τ=0), proposed method (τ=500), proposed method (τ=T). Proposed method (τ=0) is equal to the extended method proposed by T. Samad[3]. Proposed method (τ=T) is equal to training only instance that dose not include missing values. For the parameter τ=500 was used because strong adaptation occurs in the early stages (t< 1000) of SOM training.

To measure the SOM training accuracy, we computed root mean square (RMS) error of difference between each input vector and corresponding winner unit's weight vector. The number of training steps was T=10000. Fig.2 shows the result of simulation. We confirmed that RMS error is convergent at T=5000. In proposed method (τ=T), number of instances were insufficient for training correctly.

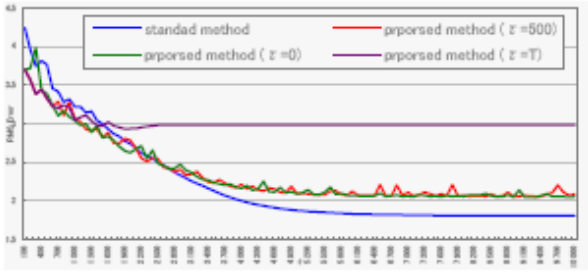

Fig.2. The result of training accuracy (RMS error)

We computed the map of unsupervised clustering using HC (hill-climbing) method based on the fixed kernel density estimation [2]. The parameters for these methods were selected based on prior experiments, so that the number of clusters becomes 20. Fig.3 shows a result of clustered map using HC method and fixed kernel density estimation. Table.2 shows a result of classes extracted from the map of unsupervised clustering. Fig.3 and Table.2 are example results for proposed method (τ=500). For the labels of each class extracted from result of clustering, we allocated the "Final state" of the largest number of instances in the cluster.

In the method which missing values were excluded, we confirmed that there was a lot of difference in the results between separate runs. It is thought that map initialization influences the differences of result. In the following experiment we calculate the average and variance of the experiment done 20 times for different map initialization.

To measure the accuracy of clustering, we computed means square quantization error (MSQE) between each

input vector and the weight vector of the cluster center for the cluster containing the corresponding winner unit. The cluster center is the competitive layer's unit that has highest value of density estimation.

To measure the accuracy of prediction for clinical data of the bladder cancer patients, we computed the classification accuracy based on the allocated class label. Table.3 shows the result of the classification accuracy.

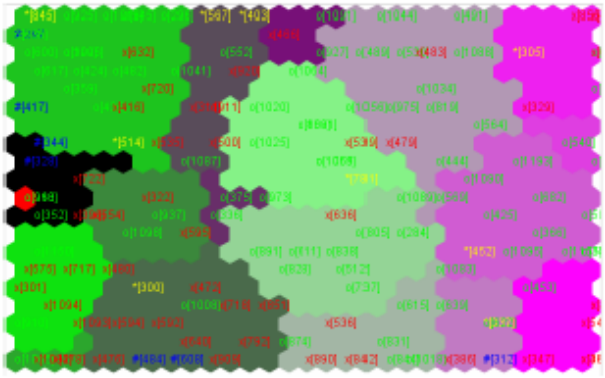

Fig.3. The result of clustered map using HC method and fixed kernel density estimation (τ=500),

Class	Number of instances Total	o	x	*	#	Label
0	7	3	2	0	2	o
1	8	2	6	0	0	x
2	22	14	4	2	2	o
3	3	1	2	0	0	x
4	5	2	3	0	0	x
5	14	2	9	1	2	x
6	8	2	4	2	0	x
7	3	3	0	0	0	o
8	3	2	1	0	0	o
9	10	8	1	1	0	o
10	10	9	1	0	0	o
11	9	6	3	0	0	o
12	13	11	2	0	0	o
13	4	1	1	1	1	o
14	10	9	0	1	0	o
15	2	2	0	0	0	o
16	7	2	4	1	0	x
17	7	1	6	0	0	x

o = survival, x = cancer-specific death,
* = death by other cause, # = unknown

Table.2. The result of classes extracted from the map of unsupervised clustering (τ=500),

	MSQE Average	MSQE Variance	Classification accuracy Average	Classification accuracy Variance
Standard method	3.03	0.0033	64.87	8.41
Proposed method τ=0	2.94	0.0030	67.32	3.66
Proposed method τ=500	2.87	0.0029	68.85	2.74
Proposed method τ=5000	3.26	0.0069	64.89	14.73

Table.3. The result of the classification accuracy.

The results of Table.3 showed that proposed method (τ=0) and proposed method (τ=500) gave better results than standard method. The results of classification accuracy showed that the average of proposed method (τ=0) and proposed method (τ=500) are almost equal, and the variance is lower in proposed method (τ=0) then proposed method (τ=500).

VI. CONCLUSION

In the past researches S.F. Shariat [5], B.H. Bochner [6], and P.I. Karakiewicz [7] showed that the result of prediction accuracy is 60-70% for prediction of bladder cancer patients. In this paper we showed that the proposed SOM method gave competitive results of 68.85%.

In this experiment, there is a possibility that the number of instance in data sets is too small. Note that it is necessary to increase the instances not including the missing values.

For SOM training with missing data, we showed that the proposed method can correctly learn by reducing the influence of information on the missing values. We confirmed that the influence of the map initialization was less in the proposed method. The proposed method could be extended to train using data sets including outlier values. The parameter τ of the proposed method is not optimized in this research, and parameter τ needs to be optimized in future works.

REFERENCES

[1] Teuvo Kohonen (1984), Self-organizing maps, Springer-Verlag, Berlin Heidelberg
[2] Marc M Van Hulle (1981), Faithful representations and topographic maps, John Wiley and Sons, New York
[3] Tariq Samad and Steven Alex Harp (1992), Self-organization with partial data, Network: Computation in Neural Systems, 3(2):p.205-p.212
[4] Samuel Kaski (1997), Data exploration using self-organizing maps, Acta Poly-technica Scandinavica, Mathematics, Computing and Management in Engineering Series No.82 DTech Thesis, Helsinki University of Technology, Finland
[5] Shahrokh F. Shariat et al (2006), Nomograms provide improved accuracy for predicting survival after radical cystectomy, Clinical cancer research, 12-22, p.6663-p.6676
[6] B.H. Bochner, MW Kattan, KC Vora (2006), Postoperative nomogram predicting risk of recurrence after radical cystectomy for bladder cancer, Journal of clinical oncology, Vol.24-24, p.3967-p.1457
[7] P.I. Karakiewicz et al (2006), Precystectomy nomogram for prediction of advanced bladder cancer stage, European urology, Vol.50-6, p.1254-p.1262