

Method Development of Multi-Dimensional Accident Analysis Using Self Organizing Map

Hitoshi Uno, Yusuke Kageyama, Akira Yamaguchi and Tomosaburo Okabe
Nissan

ABSTRACT

Implementation of appropriate safety measures, either from the viewpoint of a vehicle or the society or the infrastructure, it is an important issue to clearly understand the multi-dimension complicated real world accident scenarios. This study proposes a new method to easily capture and to extract the essence of such complicated multi-dimension mutual relationship by visualizing the results of SOM (Self Organizing Map).

The FARS data from 2010 is used to generate a dataset comprised of 16,180 fatal passenger car drivers and 48 variables. The 16,180 fatal drivers were clustered using hierarchy cluster analysis method and mapped into a two-dimensional square with one dot representing one fatal driver using SOM.

From the SOM assessment of the 16,180 fatal drivers, five clusters were created, and they are characterized as follows: Cluster 1 (Interstate highway accidents), Cluster 2 (Drunk speeding), Cluster 3 (Non speeding lane departure), Cluster 4 (Vehicle to vehicle) and Cluster 5 (Intersection).

The number of fatalities in Clusters 1 and 2 could be possibly reduced by application of CA (Crash Avoidance) technologies and stricter enforcement of traffic laws. For Clusters 3, 4 and 5, reduction by CA technologies and stricter enforcement would be more difficult, because (i) the majority of the drivers would be respecting the laws and (ii) the road environments and vehicle to vehicle situations are more complicated for current CA technologies. The remaining crash scenarios related to Clusters 3, 4 and 5 are head on collision / impact to fixed object on minor roads and side impact in urban areas.

CITATION: Uno, H., Kageyama, Y., Yamaguchi, A. and Okabe, T., "Method Development of Multi-Dimensional Accident Analysis Using Self Organizing Map," *SAE Int. J. Trans. Safety* 1(1):2013, doi:10.4271/2013-01-0758.

INTRODUCTION

Focusing on the history of occupant safety improvement, there has been significant technology evolution within the past several decades, such as the seatbelt, airbag, body structure and CA technologies. With the introduction of CA technologies, the strategy to improve occupant safety has become more complicated while considering the priority for technologies to reduce accident fatalities, the corresponding time frame and phenomena are different for CA and passive safety technologies. When conducting accident analysis to consider such priority, it is difficult to consider attributes that caused the accident and passive attributes at the same time, as the number of combinations of these attributes will be astronomical.

On the other hand, when focusing on a specific perspective such as the effect of seatbelt usage, only a few factors can be analysed simultaneously, and its relationship with other important factors are difficult to grasp. In addition, the outcome of such analysis will be highly dependent on the

choice of variables selected by the analyst. Hence, the present study focuses on a generalized method, by utilizing SOM to visualize the multi-dimension accident scenario.

A Self Organizing map (SOM) is a type of artificial neural network that is trained using unsupervised learning to map mutual relationship into a two-dimensional representation. This can prevent the analyst's arbitrary perspective.

Further, SOM is useful for low-dimensional visualization of high-dimensional input data, by using a neighborhood function to preserve the topological properties of the input space. This can preserve all characteristics of a large dataset (48 variables $\times$ 16,180 cases) used in this study and one can visually percept all 48 variables and their mutual relationships across the clusters at a glance.

METHODOLOGY

Data Set

This study uses the FARS 2010 data base maintained by NHTSA and UMTRI. Each traffic accident in FARS includes at least one fatality that occurred on a traffic way. Data key of FARS 2010 Occupant (FARS10OC) consists of driving scenarios including road environment, vehicle / driver relating information and occupant characteristics. As the objective of this study is to understand accident scenarios of general cars, the data set conditions are filtered as fatal passenger car and light truck drivers shown in [Table 1](#). From the total 69,124 fatal accident occupants in the FARS10OC data base, 16,180 passenger car fatal drivers can be extracted.

Table 1. Data set used in this study

2010 FARS Occupant cases (FARS10OC)	69,124
AUX: VEHICLE BODY TYPE = Passenger Car (1) + Light truck (2-5)	
OCCUPANT TYPE = Driver (1)	
OCCUPANT INJURY SEVERITY = fatal (4)	
Selected cases	16,180

Selection of variables

The first step is to narrow down the whole set of 566 variables to a subset of fewer meaningful variables. Without any loss of information, the fewer variables it will help to perceive the multi-dimension accident scenario more accurately. The 566 variables are narrowed down in the following rule by excluding, those semantically lower order variables, the variables having low frequency and similar variables.

(1). Exclude low-semantic variables in describing the accident. e.g. CASE NUMBER, COUNTY ID, VIN, ACCIDENT DATE-YEAR, etc.

(2). Exclude variables which attribute values having low-frequency occurrence of under 5%. e.g. CRASH RELATED FACTOR has 99.9% of value "0: none"

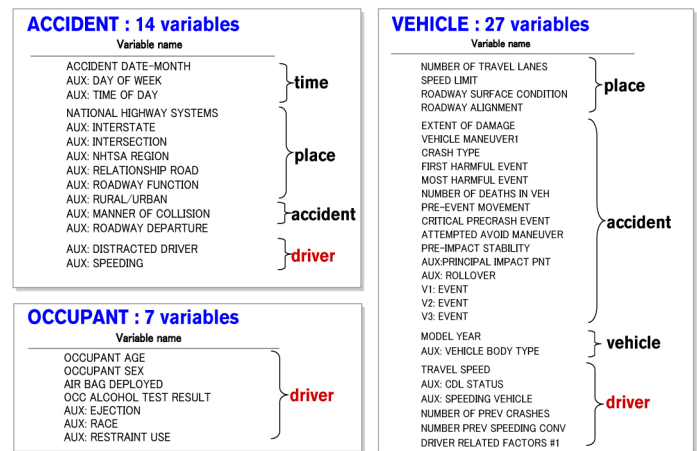
(3). Variables that have similar values with high coherence are grouped, to avoid over represented contribution. e.g. BODY TYPE, VIN BODY TYPE, VIN VEHICLE TYPE, etc. They are grouped by using only one representative variable.

Parameter selection

Based on the above rule, excluding similar variables and unimportant variables, finally a set of 48 variables were selected as shown in [Table 2](#).

Table 2. Selected variables

Total number of FARS10OC variables	566
(1) Low-semantic variables	346
(2) Low-frequency variables	59
(3) Similar variables	113
Selected variables	48



Following are example of the excluded variables

(1). Low-semantic

2 CASE NUMBER
3 SEQUENCE NUMBER
4 VEHICLE NUMBER
6 PERSON NUMBER
7 CITY
8 COUNTY
10 ACCIDENT DATE-DAY
11 ACCIDENT DATE-YEAR
13 ACCIDENT DATE-MINUTE
23 MILEPOINT
37 EMS NOTIFIED-MINUTE
39 EMS ARRIVAL-MINUTE

(2). Low-frequency

5 RECORD TYPE	100%
21 TRAFFICWAY IDENTIFIER #1	100%
22 TRAFFICWAY IDENTIFIER #2	100%
64 AUX: CRASH INJURY TYPE	100%
347 VEHICLE FACTOR3	100%
348 VEHICLE FACTOR4	100%
349 VEHICLE FACTOR5	100%
351 VEHICLE MANEUVER2	100%
433 VIOLATIONS CHARGED #6	100%
434 VIOLATIONS CHARGED #7	100%
435 VIOLATIONS CHARGED #8	100%
452 DRIVER VISION OBSCURED3	100%

(3). Similar variables

12 ACCIDENT DATE-HOUR
38 EMS ARRIVAL-HOUR
40 EMS HOSPITAL-HOUR
36 EMS NOTIFIED-HOUR
32 LIGHT CONDITION

12 ACCIDENT DATE-HOUR



Documents
Document Sharing Hub

Self Organizing Map

A Self Organizing map (SOM) is a type of artificial neural network that is trained using unsupervised learning to map mutual relationship into a two-dimensional representation. This can prevent the analyst's arbitrary perspective.

Further, SOM is useful for low-dimensional visualization of high-dimensional input data, by using a neighborhood function to preserve the topological properties of the input space. This can preserve all characteristics of a large dataset (48 variables $\times$ 16,180 cases) used in this study and one can visually percept all 48 variables and their mutual relationships across the clusters at a glance.

Vector Quantization (VQ)

Vector quantization is a classical quantization technique [1], [2]. Thus, this paper will only briefly explain the statistical outline to understand the results.

Vector quantization is to find the discrete approximate of the input vector x of the vector space R_n , by using infinite codebook vector $m_i \in R_n, i=1,2,3,...,k$. The approximate of x , is to find the most nearest codevector m_c to x with Euclidean distance as in [equation \(1\)](#). If the most appropriate value m_i is chosen, the square quantization error will be the minimum, as in [equation \(2\)](#).

$$\|x - m_c\| = \min_i \{\|x - m_i\|\} \quad (1)$$

$$E = \int \|x - m_c\|^2 p(x) dx \quad (2)$$

Each, codevector is related with a nearest codevector, placed in a nearest neighbor region called a Voronoi region [3]. Voronoi is a space where the codebook vectors are fitted in the two-dimensional space side by side, like a mosaic, and the space is separated into a multiple domain, and each area is divided by a hyper-plane. Each separated area has a vector which has the most nearest vector to the surrounding vectors in the area. The mosaic separation lines, in other words the planes between the vectors, each consist from a Voronoi mosaic separation.

Application

In order to define one accident case as a 48 dimensional vector, a VQ value is calculated for all 16,180 accident cases. A VQ value would be a single scalar value, such as "0.8234", that represents the length of the 48 dimensional vector as a single value. The similarity of two accident cases (vectors) can be expressed by the difference of the VQ values.

VQ values are also calculated for groups which have similar VQ values. Hierarchy clustering was used to group similar values and it is a classical method that uses the difference of all variables to calculate the distance between each item. Hence, it is more accurate in representing the

distance of each item, rather than comparing two scalar vector values. The Voronoi regions were created by using the hierarchy structure of the hierarchy clustering results. For example, if there are three VQ's as VQa, VQb, VQc, and if these are clustered as one group in the hierarchy structure, the VQ(cluster a,b,c) for this group would be the center average of the 3 VQ's as shown in [equation \(3\)](#). By using hierarchy clustering to calculate the grouped VQ's, it is possible to maintain the cluster structure in the two-dimensional SOM square map.

$$VQ_a = ax_1 + a'x_2 + a''x_3 + \dots a'''x_m$$

$$VQ_b = bx_1 + b'x_2 + b''x_3 + \dots b'''x_m$$

$$VQ_c = cx_1 + c'x_2 + c''x_3 + \dots c'''x_m$$

$$VQ_{(a,b,c)} = (a + b + c) / mx_1 +$$

$$(a' + b' + c') / mx_2 + (a'' + b'' + c'') / mx_3 + \dots \quad (3)$$

Mapping

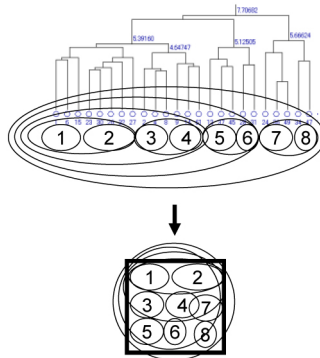
To map the items into a two-dimensional space, a square map of the size of all samples is prepared. From the calculated VQ values using [equation \(1\)](#) and [\(2\)](#) for each item, and also by maintaining the cluster hierarchy by [equation \(3\)](#), the items are plotted in the map by plotting each item with the closest VQ value. The mapping steps are shown below.

- Step1. Calculate VQ value for each of all of the cases
- Step2. Randomly plot one case in the middle of the map
- Step3. Randomly select another case and plot beside the first case
- Step4. Randomly select another case and plot beside the case with the nearest VQ value
- Step5. Repeat Step4, repeat until there are no cases left.

As the items are plotted into the two-dimensional square maintaining the hierarchy structure, the clustering results would always be maintained as shown in [Figure 1](#), even after SOM method is applied to the same data for a number of times.

Creating of SOM

As each VQ and accident number have a one-to-one correspondence, for each variable, the accident number is replaced by the variable value with color gradation, from the smallest value as blue to the largest value as red, as shown in [Figure 2](#).



Quantization vector map

0.584	1.006	0.584	1.848	1.427
0.271	0.163	1.006	1.848	1.427
0.163	0.163	-0.993	-0.993	-0.993
-0.993	0.163	-0.150	-0.993	-0.572
-0.993	-0.993	-0.993	-0.993	-0.993

Accident no. map

16	19	17	24	22
23	14	20	25	21
9	10	5	7	2
13	8	18	12	15
1	4	11	6	3

Quantized Data Matrix

Accident No.	Manner of collision	Speeding	Quantization vector	Manner of collision	Speeding
1	1	1	0.584	2	2
2	1	1	0.271	2	2
3	1	1	0.163	2	2
4	1	1	-0.993	2	2
5	1	1	-0.993	2	2
6	1	1	-0.993	2	2
7	1	1	-0.993	2	2
8	1	1	-0.993	2	2
9	1	1	-0.993	2	2
10	1	1	-0.993	2	2
11	1	1	-0.993	2	2
12	1	1	-0.993	2	2
13	1	1	-0.993	2	2
14	1	1	-0.993	2	2
15	1	1	-0.993	2	2
16	1	1	-0.993	2	2
17	1	1	-0.993	2	2
18	1	1	-0.993	2	2
19	1	1	-0.993	2	2
20	1	1	-0.993	2	2
21	1	1	-0.993	2	2
22	1	1	-0.993	2	2
23	1	1	-0.993	2	2
24	1	1	-0.993	2	2
25	1	1	-0.993	2	2

$$r(xy) = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4)$$

AVA Documents
Document Sharing Hub