Using Affinity Set on Mining the Necessity of Computed Tomography Scanning

Yuh-Wen Chen [1], Moussa Larbani [2], Tzung-Hung Li [1] and Chao-Wen Chen [3]

Abstract—**Computed tomography (CT) is a medical imaging method of tomography. Digital geometry processing is used to generate a three-dimensional image of the inside of a patient from a large series of two-dimensional X-ray images taken around a single axis of rotation. The scanning of CT has become an important tool in medical imaging to supplement X-rays and medical ultrasonography. Although it is expensive, it is the best tool to diagnose a large number of different disease entities; especially, for the trauma patients in emergency room. In this study, the trauma patients, who were treated by the CT scanning are collected in order to discover the critical knowledge; that is, what characteristics of trauma patients would lead to the necessity of CT scanning? The data mining model of affinity set and neural network (NN) are both used for resolution and comparison. Finally, studying results show that he affinity model performs better than the NN model, but the collected data lacks the explanatory power in practices. Thus, a further research is necessary.**

I. INTRODUCTION

CURING disease, maintaining health, and saving lives is the doctor's mission. People in traditional society believe the doctor's expertise is indisputable, trustworthy, and unmistakable. However, the limited time for doctors in the emergency room (ER) leads to the inevitable risk of delayed diagnosis [5, 6, 8]; thus, some tiny symptoms inside the human body could be ignored. Computed tomography (CT) is an effective imaging method of tomography to detect the inner bleeding of trauma patients in advance, but, it is really expensive. CT scanning combines the special x-ray equipment with sophisticated computers to produce multiple images or pictures of the inside of the human body. These cross-sectional images of the area being studied can then be examined on a computer monitor or printed. CT scanning of the head is typically used to detect, e.g., bleeding, brain injury and skull fractures in patients with head injuries, bleeding caused by a ruptured or leaking aneurysm in a patient with a sudden severe headache, a blood clot or bleeding within the brain shortly after a patient exhibits symptoms of a stroke, etc [12, 14]. In addition, CT scanning is also valuable to evaluate the extent of bone and soft tissue damage in trauma patients, and planning surgical reconstruction, diagnose diseases of the temporal bone on the side of the skull, which may be causing hearing problems, determine whether inflammation or other changes are present in the paranasal sinuses, etc. As the available resources of CT scanning are constrained in hospitals, what type of trauma patients has the highest priority to be treated by CT scanning should be explored and arranged [15]. Therefore, this study collects the patient data of CT scanning, and uses the affinity set to discover the knowledge/know-how above, which are valuable in doctor training.

This paper is organized as follows to solve the medical data mining problem [2, 9]: section 2 introduces the basic concepts and definitions of affinity set; after that, the affinity data-mining model is proposed. Section 3 uses the actual samples of CT scanning from the Kaohsiung Medical University hospital of Taiwan to validate our affinity data mining idea, deriving key attributes for the necessity of CT scanning. In addition, the performance of affinity model and neural network model are compared. Finally, section 4 gives conclusions and recommendations based on our current mining results.

II. PREPARATION FOR STUDY

Here, the basic concepts and definitions of affinity are simply reviewed [4, 5, 7, 11].

A. Affinity Set

Although medical data is inevitably combined with noise and vagueness [1, 2, 3], so far as we know, in literatures, there is no theory dealing with affinity as a vague and time-dependent concept, and little scholarly awareness that such a simple affinity idea could be developed for valuable models in medical data mining. Fuzzy set theory could the best tool for representing vague and imprecise concepts so far; however, the affinity set proposed here is not merely a fuzzy set because assuming any type of membership function here is unnecessary. Instead, this work uses the closeness or distance from Topology [10] between any two objects to develop the useful model in information sciences. The interesting and innovative idea in this study is using a decision maker's perception of distance/closeness to form his or her preferred affinity set. This new relation theory: affinity set theory is quite general, not only able to describe the similarity between objects, but also able to represent general

Yuh-Wen Chen is the associate Prof. of Da-Yeh University, Taiwan (corresponding author to provide phone: +886-4-8511888 ext 4120; fax: +886-4-8511270; e-mail: profcher@ mail.dyu.edu.tw). Mr. Tzung-Hung Li is his master student.

Moussa Larbani is the associate Prof. of Department of Business Administration, Kulliyyah of Economics and Management Sciences, IIUM University, Malaysia (e-mail: moussa.larbani@gmail.com).

Chao-Wen Chen is the trauma doctor of emergency room in Hospital of Kaohsiung Medical University (e-mail: ytljwc@yahoo.com.tw).

978-1-4244-3541-8/09/$25.00 ©2009 IEEE

relationships between objects, e.g., closeness, belongingness, equivalence, …, etc, so that a decision maker can easily use this simple concept of modeling in information sciences.

Definition 2.1. By affinity set we mean any object (real or abstract) that creates affinity between objects.

Some examples are given to clarify our idea.

Example 2.1. An institution or company is an affinity set, for it is an object that creates affinity between people that make them work together.

Definition 2.2. Let *e* and A be a subject and an affinity set, respectively. Let I be a subset of the time axis $[0, +\infty[$. The affinity between *e* and A is represented by a function

$$M_A^e(.): I \to [0,1]$$
$$t \to M_A^e(t).$$

The value $M_A^e(t)$ expresses the degree of affinity between the subject *e* and the affinity set A at time *t*. When $M_A^e(t)= 1$ this means that affinity degree of *e* with affinity set A is at the maximal level at time *t*; but $M_A^e(t)= 1$ doesn't mean that *e* belongs to A, unless the considered affinity measurement $M_A^e(t)$ is a function of belongingness degree. When $M_A^e(t)= 0$ this means that *e* has no affinity with A at time *t*. When $0 < M_A^e(t) < 1$, this means that *e* has partial affinity with A at time *t*. Here we emphasize the fact that the notion of affinity is more general than the notion of membership or belongingness: the later is just a particular case of the former.

Definition 2.3. The universal set, denoted by U, is the affinity set representing the fundamental principle of existence. We have

$$M_U^e(.): [0,+\infty[\to [0,1]$$
$$t \to M_A^e(t)$$

and $M_U^e(t)=1$, for all existing objects at time *t*, and for all times *t*.

In other words the affinity set defined by the affinity "existence" has complete affinity with all previously existing objects, that exist in the present, and that will exist in the future. In general, in real world situations, some traditional referential set V, such as that when an object *e* is not in V, $M_A^e(t)= 0$ for all *t* in $I \subset [0,+\infty[$, can be determined. In order to make the notion of affinity set operational and for practical reasons, in the remainder of the paper, instead of dealing with the universal set U, we only discuss affinity sets defined on a traditional referential set V. Thus, in the remainder of the paper when we refer to an affinity set, we assume that sets V and I are given.

Definition 2.4. Let A be an affinity set. Then the function defining A is

$$F_A(.,.): V \times I \to [0,1] \quad (1)$$
$$(e, t) \to F_A(e, t)= M_A^e(t)$$

An element in real situations often belongs to a set for some time and not in that set other times. Such behavior can be represented using the affinity set notion. The behavior of affinity set A over time can also be investigated through its function $F_A(.,.)$.

Definition 2.5. Let A be an affinity set and $k \in [0,1]$. We say that an element *e* is in the *t*-*k*-Core of the affinity set A at time *t*, denoted by *t*-*k*-Core(A), if $M_A^e(t) \geq k$, that is,

$$t-k-Core(A) = \{e \mid M_A^e(t) \geq k\} \quad (2)$$

when k=1, *t*-*k*-Core(A) is simply called the core of A at time *t*, denoted by *t*-Core(A).

Definition 2.6. An observation period is defined as the continuous or discrete period analyzing the behavior of an element *e* of V with respect to an affinity set A: an illustration is given in Figure 1 below.

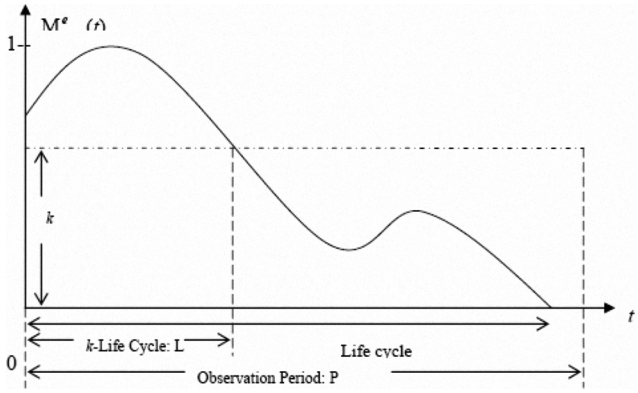

Figure 1. Illustration of the affinity between an element *e* and an affinity set A over an observation period P.

The affinity of element *e* with respect to affinity set A in real-world situations often depends implicitly on other variables than time. These variables generally express condition or constraint variability that affects affinity evaluation. Studying element *e* behavior with respect to time and other variables may be practically desirable. A decision maker may even study element *e* behavior at a fixed time with respect to other variables. This section extends the affinity set definition to the case where desired variables appear explicitly. This definition makes it possible to study *e* affinity behavior over time and with respect to other variables as well.

Definition 2.7. Let *e* and A be an element and an affinity set, respectively. Assume that the affinity of *e* with respect to A depends on some variable *w* that takes its values in a traditional set W. In order to make the variable *w* appear in the affinity definition between *e* and A, we introduce the following affinity

$$M_A^e(.): I \times W \to [0,1]$$
$$(t,w) \to M_A^e(t,w)$$

The value $M_A^e(t,w)$ expresses the degree of affinity between element *e* and A at time *t* with respect to *w*.

Definition 2.8. Let A be an affinity set depending on a variable $w \in W$. Then the function defining A is defined by

$$R_A(.,.,.): V \times I \times W \rightarrow [0,1]$$

$$(e, t, w) \rightarrow R_A(e, t, w) = M_A^e(t, w)$$

where V is the traditional referential set.

Definition 2.9. Let A be an affinity set depending on a variable $w \in W$ and $k \in [0,1]$. We say that an element e in V is in the (t, w^0)-k-Core of A at time t when $w = w^0$, denoted by (t, w^0)-k-Core(A) if $M_A^e(t, w^0) \geq k$, that is,

$$(t, w^0)\text{-}k\text{-Core(A)} = \{e \mid M_A^e(t, w^0) \geq k\}$$

When $k=1$, (t, w^0)-k-Core(A) is simply called the core of A at time t when $w = w^0$ and denoted by (t, w^0)-Core(A).

B. Affinity Data Mining

Here, following the spirit of affinity, a simple data mining model is proposed.

Definition.2.10 Let X be a set endowed with a distance $d(x, y)$, i.e. (X, d) is a metric space [10]. Let V be a subset of X. An affinity set A in V is given by

$$A = (d', B, V)$$

Where d' is defined by

$$d': V \rightarrow [0,1]$$

$$.e \rightarrow d'(e, B) = 1 - \alpha\, d(e, B)$$

where d' is the affinity, the set B is called the core of the affinity set A, $d(e,B)$ is defined by

$$d(e,B) = \min_{z \in B} d(e, z)$$

Here please note that there is a difference between $d(e, B)$ and $d(x, y)$, although the same notation is used "d". Indeed, $d(e,B)$ is the distance between an element e of V and the subset B of V, while $d(x, y)$ is a distance between two elements x and y of V. Please do not confuse these two.

$\alpha = \dfrac{1}{\max\limits_{(x,y) \in V \times V} d(x, y)}$, that is α is the maximal distance between elements of V.

Procedure 2.1

1) Define the metric space (X, d)
2) Determine the referential set V
3) Determine the core B of the affinity set
4) Use the affinity d' as defined

$$d': V \rightarrow [0,1]$$

$$e \rightarrow d'(e, B) = 1 - \alpha\, d(e, B)$$

Then computing the k-core (A) is easy once the k value is given. Next, we introduce an actual example to see how this idea works.

III. ACTUAL EXAMPLE OF CT SCANING PROBLEM

The objective of this research is to find key/core attributes leading to the necessity/effectiveness of CT Scanning. Doctors give 174 male samples and 82 female samples of clinical data in 2008 (from Jan. to Dec.). After that, samples are separated into two parts: training and validation. The training/validation rate is designed as 80%/20%, 70%/30% and 60%/40% of data for male samples and female samples, respectively. Doctors suggest seven possible influential attributes/causes $\{x\}$ leading to the necessity of CT Scanning (y), which are shown in Table 1. Here a rule is defined as a vector of $(x_1, x_2, x_3, x_4, x_5, x_6, x_7, y)$ in the metric space of Definition 2.10. The value of each x_i (i=1,2,...,7) and y are randomly selected from the attribute domain in Table 1. If any x_i (i=1,2,...,7) has the value of zero, then this means this attribute x_i won't be included to form a rule. If a randomly generated rule is found in the training set once, then the occurrence frequency of rule is one; if a randomly generated rule is found in the training set twice, then the occurrence frequency of rule is two;...,etc. The affinity degree d' is simply defined as the occurrence frequency of a rule divided by the number of all samples in the training set. For example, consider the training/validation rate of 80/20 for male data and y=1, the rules with the higher affinity degrees and with at least two explanatory variables $\{x\}$ are shown in Table 2. In other words, one minus the miss rate of a rule is the affinity degree of rule.

Table 1 Classification of Attributes

Attribute variable	Values of attribute
x_1: Patient's age	"1": under 6 years old; "2": 6 - 17 years old; "3": 18 - 40 years old; "4": 41 - 65 years old; "5": over 65 years old
x_2: Triage	"1": resuscitation, injuries require immediate medical care. "2": emergency, injuries require surgery within 10 minutes. "3": urgent, injuries require surgery within 30 minutes.
x_3: Number of trauma	"1": only one; "2": two; "3": over than two
x_4: Glasgow coma scale (GCS)	"1": 13-15; "2": 9-12; "3": 3-8
x_5: Breathe	"1": 10-24 times per minute, normal; "2": else, abnormal.
x_6: Blood-pressure	"1": 90-140 mmHg, normal; "2": else, abnormal.
x_7: Pulse	"1": 60-100 Times per minute, normal; "2": else, abnormal.
y: Diagnosis	"1": The CT scanning finds something (effective). "0": The CT scanning finds nothing (ineffective).

Table 2 $M_A^e(w^0)$ of Cardinal $\{x_i\}$=2 for y=1

Rule	Combination $\{x_i\}$	Value	$M_A^e(w^0)$
1	x_5, x_7	1,1	0.6087
2	x_4, x_5	1,1	0.5580
3	x_4, x_7	1,1	0.4420
4	x_4, x_5, x_7	1,1,1	0.4420
5	x_3, x_4, x_7	1,1,1	0.2391

Finally, the confusion matrix [13] is used to test the performance between our affinity model and neural network model. In the field of artificial intelligence, a confusion matrix is a visualization tool typically used in supervised learning. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. One benefit of a confusion matrix is that it is easy to see if the system is confusing two classes (i.e. commonly mislabelling one as another). For example, the following table shows the confusion matrix for a two class classifier. The entries in the confusion matrix have the following meaning in the context of our study: *a* is the number of correct predictions that an instance is negative, *b* is the number of incorrect predictions that an instance is positive, *c* is the number of incorrect of predictions that an instance negative, and *d* is the number of correct predictions that an instance is positive [13].

Table 3 Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	*a*	*b*
	Positive	*c*	*d*

Several standard terms have been defined for the 2 class matrix:

- The *accuracy* (*AC*) is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$AC = \frac{a+d}{a+b+c+d}$$

- The *recall* or *true positive rate* (*TP*) is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$TP = \frac{d}{c+d}$$

- The *false positive rate* (*FP*) is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation:

$$FP = \frac{b}{a+b}$$

- The *true negative rate* (*TN*) is defined as the proportion of negatives cases that were classified correctly, as calculated using the equation:

$$TN = \frac{a}{a+b}$$

- The *false negative rate* (*FN*) is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation:

$$FN = \frac{c}{c+d}$$

- Finally, *precision* (*P*) is the proportion of the predicted positive cases that were correct, as calculated using the equation:

$$P = \frac{d}{b+d}$$

The computation results are summarized in the following Table 4.

Table 4 Computation Results between Affinity Model and NN Model

Training/Validation Rate	Affinity Model	NN Model
80/20	Male: *AC*=94.1%;*P*=96.9%;*TP*=97.1% Female: *AC*=88.2%;*P*=92.8%;*TP*=92.9%	Male: *AC*=79.4%;*P*=96.4%;*TP*=81.8% Female: *AC*=70.6%;*P*=80.0%;*TP*=85.7%
70/30	Male: *AC*=94.2%;*P*=96.1%;*TP*=98.1% Female: *AC*=88.0%;*P*=95.0%;*TP*=90.9%	Male: *AC*=61.5%;*P*=94.1%;*TP*=64.0% Female: *AC*=88.1%;*P*=88.0%;*TP*=100.0%
60/40	Male: *AC*=90.2%;*P*=95.4%;*TP*=93.9% Female: *AC*=85.3%;*P*=92.5%;*TP*=89.3%	Male: *AC*=67.1%;*P*=93.9%;*TP*=69.7% Female: *AC*=75.3%;*P*=82.7%;*TP*=85.7%

It is clear the affinity model has a better exploitation power. Thus, the rules of affinity model are used for explaining the necessity of CT scanning. However, the currently collected data on hand lacks the actual link to practices. Although the explanation power of affinity model is high, its generated rules, e.g., see Table 2: if Breath is normal and Pulse is normal then CT scanning is necessary/effective, such a rule is quite unreasonable/unlogical. This is because the data on hand miss some important attributes/features of CT patients; thus, using more appropriate explanatory variables $\{x\}$ for y is necessary in the near future.

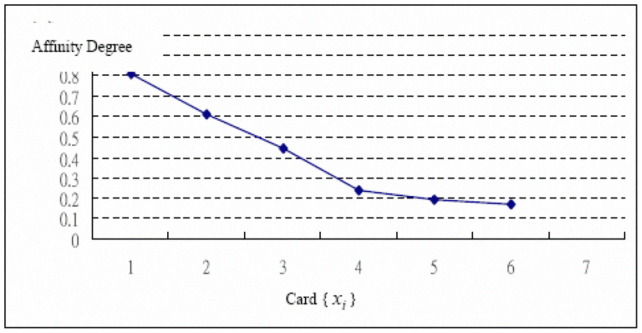

Figure 2 Affinity Degree of Rule Combination $\{x_i\}$ for Training/Validation Rate = 80/20 and Male Data

IV. CONCLUSIONS AND RECOMMENDATIONS

This is not a 100% successful data-mining attempt by affinity set, although the explanatory power of affinity model is better than that of NN model. The currently collected data of CT patients misses some important attributes/features such that the mining results are not 100% satisfied. However, the affinity models had validated for their values and superiority in some literatures [4, 5, 7, 11]. Other mapping/projection methods inspired from Topology [10] may also generate effective rules for this research. In other words, the rule base V defined in this study is not unique and the only one. When the computation/generation of rules becomes a tough task, evolutionary algorithms may be valuable.

Readers should also recognize that: (a) the affinity model is quite simple, and (b) no fuzzy membership functions are

assumed in this study; thus, we don't think our affinity data mining model is fuzzy. Finally, this study hints that efficient communication between patients and doctors is necessary/emergent in ER, because these efforts will improve the data quantity and quality before data mining.

ACKNOWLEDGMENT

This research is funded by National Science Council of Taiwan (96-2416-H-212-002-MY2), and authors would like to appreciate Prof. Jerzy Michnik, Miss Chih-Min Shen and Mr. Cheng-Yen Hsieh for their earlier efforts in this area.

REFERENCES

[1] Aguilar-Ruiz, J.S., Costa, R. & Divina, F., "Knowledge discovery from doctor-patient relationship," *Proceedings of SAC* '04, March 14-17, 2004, Nicosia, Cyprus.

[2] Berman, J.J, "Confidentiality issues for medical data miners," *Artificial Intelligence in Medicine*, Vol. 26, pp.25-36, 2002.

[3] Bratko, I. and Konomenko, I., "Learning diagnostic rules from incomplete and noisy data.," In: Phelps, B. (ed) *AI Methods in Statistics*. London, Gower Technical Press, 1987.

[4] Chen, Y. W. and Larbani, M., "Developing the Affinity Set and Its Applications," *Proceeding of the Distinguished Scholar Workshop by National Science Council*, Jul. 14-18, 2006, Taipei, Taiwan.

[5] Chen Y.W., Larbani, M., Hsieh, C-Y, and Chen, C-W, "Introduction of Affinity Set and Its Application in Data Mining Example of Delayed Diagnosis," *Expert Systems With Applications*, to appear.

[6] Kohn, L. T., Corrigan, J. M. and Donaldson, M. S., *To Err is Human: Building a Safer Health System*, 1999, Washington D. C.

[7] Larbani M. and Chen Y.W, "Affinity Set and Its Application," In: Trzaskalik (ed.), *Multiple Criteria Decision Making '07*, Publisher of The Karol Adamiecki University of Economics in Katowice, Poland, pp. 117 – 134, 2008.

[8] Leape L. L., Brennan, T. A., Laird, N., Lawthers, A. G., Localio, A. R., Barnes, B. A., Hebert L., Newhouse, J. P., Weiler, P. C., and Hiatt, H., "Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I," *The New England Journal of Medicine*, Vol. 324, pp. 370-376, 1991.

[9] Marisa S., Viveros J., Nearhos, P. and Rothman, M. J., "Applying Data Mining Techniques to a Health Insurance Information System," *Proceeding of the 22nd Very Large Data Bases Conference*, Sep. 3-6, 1996, Mumbai, India.

[10] Mendelson, B, *Introduction to Topology*, Dover Publications, 1990.

[11] Michnik, J., Michnik, A. and Pietuch, B, "Credit scoring model based on the affinity set," *Proceeding of the 10th International Conference on Enterprise Information Systems*, 14 - 16, June 2008, Barcelona, Spain.

[12] Pickhardt, P. J., Choi, J. R., Hwang, I., Butler, J. A., Puckett, M. L., Hildebrandt, H. A., Wong, R. K., Nugent, P. A., Mysliwiec, P. A. and Schindler W. R., "Computed Tomographic Virtual Colonoscopy to Screen for Colorectal Neoplasia in Asymptomatic Adults," *The New England journal of Medicine*, Vol. 349, pp. 2191-2200, 2003.

[13] Provost, F. J. and Kohavi, R., "On Applied Research in Machine Learning," *Machine Learning*, Vol. 30, pp. 127-132, 1998.

[14] Rao, P. M., Rhea, J. T., Novelline, R. A., Mostafavi,A.A. and McCabe, C.J., "Effect of Computed Tomography of the Appendix on Treatment of Patients and Use of Hospital Resources," *The New England journal of Medicine*, Vol. 338, pp. 141-146, 1998.

[15] Rumberger, J. A., Behrenbeck, T., Breen, J.F., and Sheedy, P.F., "Coronary calcification by electron beam computed tomography and obstructive coronary artery disease: a model for costs and effectiveness of diagnosis as compared with conventional cardiac testing methods," *Journal of the American College of Cardiology*, Vol. 33, pp. 453-462, 1999.