Journal of the Nigerian Association of Mathematical Physics
Volume **15 (November, 2009), pp 355 - 362**

Embedded confusion matrix in Fuzzy logic for system appraisal

[1]J.O. Omolehin, [2]L. B. Asaju, [3]M. Olagunju, [4]A. O. Ameen and [5]K. Rauf
[1,4,5]Mathematics Department,
[2,3]Computer Science Department,
University of Ilorin, Ilorin, Nigeria

Abstract

Fuzzy logic methods give effective system appraisal solution to some classification problems as against some techniques on probability prediction of solution. This paper used fuzzy logic approach as a comparative method to visual encoding technique.

Keywords
Cleaning data, data mining, filter data, visual encoding or confusion matrix.

1.0 Introduction

Data mining, sometimes called data or knowledge discovery [1] is the process of analyzing data from different perspectives and summarizing into useful information. Data mining [1, 2, 3] has a lot of application which include:

(i) In politics, it can be used for identification of potential voters in an election
(ii) In the field of medicine, data mining can be used for diagnosis purpose
(iii) In banking system, it can also be used to determine the rate of fraud in the system
(iv) Data mining is used in the field of telecommunication to determine user's behavior.

Arising from the usefulness of data mining enumerated above, it is our concern in this paper to apply one of the techniques of data mining to solve the problem that normally occur during the grading of students result. The techniques used will help a lot to measure both the students and teachers performances.

1.1 Data mining process

There are several processes that can be used in data mining. This paper presents Cross-Industry Standard Process for Data Mining which by convention abbreviated CRISP-DM [1, 2, 3 and 5] is widely used in industry. CRISP-DM consists of six phrases which include:

(a) Business preparation
(b) Data Understanding
(c) Data preparation
(d) Model building

[1]Corresponding author:
[1]e-mail address: omolehin_josephi@yahoo.com
[1]Telephone: +234-0803-357-8643

[5]Telephone: +234-0803-396-5848

(e) Testing evaluation
(f) Deployment

These phrases can be presented diagrammatically as shown in figure 1.1.

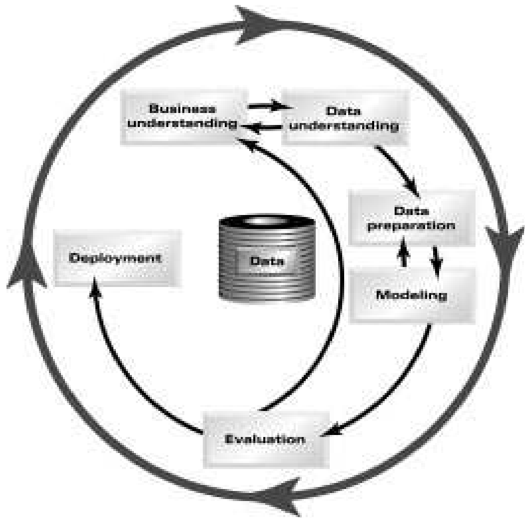

Figure 1.1 Crisp-dm

2.0 CRISP-Dm process

The CRISP-Dm process consists of six steps [1] and these steps itemized and discussed as follows:

2.1 Business Understanding

This includes managerial decisions which include the types of customers and profits. In general, this stage is the most important stage in data mining because it deals with objectives plan of the management that is responsible for data mining exploration. See [6 and 7].

2.2 Data Understanding

It considers data requirements. These steps can include initial data collection, data description, data exploration and verification of data quality.

2.3 Data Preparation

This involved cleaning of data and data transformation in preparation for data modeling

2.4 Modeling

This involves tests of model to process the data collected.

2.5 Evaluation

This involves tests of model with learning data

2.6 Deployment

This deals with development of hypothesis or knowledge discovery from the modeling that was developed.

3.0 Materials and methods

Data provided for analysis in this paper are the list of students' grades in CSC 201 and CSC 202 courses in the Department of Computer Science, University of Ilorin, Ilorin, Nigeria in the 2007/2008 academic session. The method of analysis considered utilizes the confusion matrix and fuzzy logic.

A confusion matrix, also known as coincidence matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. A typical confusion matrix is presented in Table 3.1.

Table 3.1: Confusion matrix diagram

	True Class		
	Positive	**Negative**	**Total**
Predictive Class	True positive count (TP) or True Pass	False positive count (FP) False pass	TP + FP
	False negative count (FN) or False Pass	True negative count (TN) True fail	FP + TN
Total	TP + FN	FP + TN	TP + FN + FP + TN

Table 3.2: Grades in CSC 201

S/N	MAT. NO	SCORE	S/N	MAT. NO.	SCORE	S/N	MAT. NO	SCORE	S/N	MAT. NO.	SCORE
1	02/55EC061	40	26	04/55EC087	33	51	05/55EC020	42	76	05/55EC050	53
2	03/10AC078	43	27	04/55EC090	50	52	05/55EC021	48	77	05/55EC051	40
3	03/30GB047	14	28	04/55EC092	48	53	04/55EC022	33	78	05/55EC052	51
4	03/30GC005	46	29	04/55EC095	61	54	05/55EC022	40	79	05/55EC053	40
5	03/30GC126	42	30	04/55EC113	40	55	05/55EC024	60	80	05/55EC054	40
6	03/30GD050	41	31	04/55EC116	64	56	05/55EC025	42	81	05/55EC056	40
7	05/55EC095	42	32	04/55EC147	63	57	05/55EC026	53	82	05/55EC057	50
8	03/55EC122	32	33	04/55EC148	45	58	05/55EC026	40	83	05/55EC060	46
9	04/30GA021	41	34	04/55EC151	50	59	05/55EC028	42	84	05/55EC061	53
10	04/30GC040	28	35	04/55EC157	30	60	05/55EC029	47	85	05/55EC063	54
11	04/30GC045	1	36	05/55EC001	61	61	05/55EC030	50	86	05/55EC064	53
12	04/30GC052	54	37	05/55EC002	48	62	05/55EC032	32	87	05/55EC065	51
13	04/30GC112	46	38	05/55EC003	50	63	05/55EC033	13	88	05/55EC066	42
14	04/30GD046	32	39	05/55EC006	43	64	05/55EC035	60	89	05/55EC067	45
15	04/30GD048	31	40	05/55EC007	42	65	05/55EC036	63	90	05/55EC068	47
16	04/30GD131	51	41	05/55EC009	51	66	05/55EC037	65	91	05/55EC069	60
17	04/55EC010	53	42	05/55EC010	47	67	05/55EC038	60	92	05/55EC070	52
18	04/55EC027	44	43	05/55EC012	32	68	05/55EC040	70	93	05/55EC071	43
19	04/55EC037	44	44	05/55EC013	45	69	05/55EC041	55	94	05/55EC072	46
20	04/55EC064	40	45	05/55EC014	45	70	05/55EC044	70	95	05/55EC073	50
21	04/55EC067	33	46	05/55EC015	46	71	05/55EC045	43	96	05/55EC074	40
22	04/55EC069	30	47	05/55EC016	48	72	05/55EC046	42	97	05/55EC075	61
23	04/55EC074	40	48	05/55EC017	40	73	05/55EC047	27	98	05/55EC076	46
24	04/55EC080	32	49	05/55EC018	50	74	05/55EC048	65	99	05/55EC078	50
25	04/55EC083	24	50	05/55EC010	43	75	05/55EC049	51	100	05/55EC079	61

Table 3.3: Grades in CSC 202

S/N	MAT. NO	SCORE	S/N	MAT. NO.	SCORE	S/N	MAT. NO	SCORE	S/N	MAT. NO.	SCORE
1	02/30GC153	41	26	04/55EC078	43	51	05/55EC017	40	76	05/55EC050	52
2	02/30GC184	40	27	04/55EC083	82	52	05/55EC018	63	77	05/55EC051	47
3	03/10AC034	52	28	04/55EC095	62	53	05/55EC019	56	78	05/55EC052	45
4	03/10AC078	50	29	04/55EC096	52	54	05/55EC020	52	79	05/55EC053	40
5	03/10AC178	40	30	04/55EC101	45	55	05/55EC021	62	80	05/55EC054	40
6	03/30GA009	32	31	04/55EC113	36	56	05/55EC023	43	81	05/55EC056	62

7	03/30GB006	32	32	04/55EC120	46	57	05/55EC024	34	82	05/55EC057	45

Table 3.3: Grades in CSC 202 (contd)

S/N	MAT. NO	SCORE	S/N	MAT. NO.	SCORE	S/N	MAT. NO	SCORE	S/N	MAT. NO.	SCORE
8	03/30GC043	46	33	04/55EC141	42	58	05/55EC025	51	83	05/55EC058	59
9	03/30GD050	51	34	04/55EC148	70	59	05/55EC026	40	84	05/55EC061	53
10	03/30GD053	47	35	04/55EC151	40	60	05/55EC028	50	85	05/55EC064	62
11	03/55EC028	31	36	04/55EC157	51	61	05/55EC029	60	86	05/55EC065	57
12	03/55EC116	40	37	05/55EC001	65	62	05/55EC030	65	87	05/55EC066	45
13	04/30GA021	51	38	05/55EC002	46	63	05/55EC032	36	88	05/55EC067	60
14	04/30GC040	45	39	05/55EC003	64	64	05/55EC035	54	89	05/55EC068	44
15	04/30GC052	51	40	05/55EC004	40	65	05/55EC036	43	90	05/55EC069	72
16	04/30GD046	46	41	05/55EC006	52	66	05/55EC037	85	91	05/55EC070	61
17	04/30GD078	50	42	05/55EC007	47	67	05/55EC038	62	92	05/55EC071	47
18	04/30GD 131	57	43	05/55EC009	51	68	05/55EC040	62	93	05/55EC072	54
19	04/55EA084	47	44	05/55EC010	50	69	05/55EC041	61	94	05/55EC073	60
20	04/55EC024	51	45	05/55EC011	63	70	05/55EC044	86	95	05/55EC074	42
21	04/55EC030	44	46	05/55EC012	35	71	05/55EC045	45	96	05/55EC075	56
22	04/55EC037	45	47	05/55EC013	44	72	05/55EC046	49	97	05/55EC076	60
23	04/55EC038	68	48	05/55EC014	40	73	05/55EC047	31	98	05/55EC078	45
24	04/55EC064	63	49	05/55EC015	55	74	05/55EC048	78	99	05/55EC079	60
25	04/55EC076	44	50	05/55EC016	51	75	05/55EC049	43	100	05/55EC008	43

4.0 Modeling process

In order to develop a reasonable model [4], the under-listed assumptions are necessary. The number of students is fixed 100. Student must be classified by his or her grades e.g. pass or fail etc. Based on the above assumptions, model is developed by defining the following:

$$F(a,b,c,) = a_0 \leq a \leq a_1,\ b_0 \leq b \leq b_1,\ c_0 \leq c \leq c_1,\ d_0 \leq d \leq d_1 \quad (4.1)$$

where

$$a_0 = 0,\quad a_1 = 39,\quad b_0 = 40,\quad b_1 = 49,\quad c_0 = 50,\quad c_1 = 69,\quad d_0 = 70,\quad d_1 = 100 \quad (4.2)$$

4.1 Numerical Example

***Example* 4.1**

Table 4.1: The results of CSC 201 for 100 students

Scores	Classification	Number
0 – 39	True negative or true fail	17
40 – 49	False negative or false fail	46
50 – 69	False positive or false fail	35
70 – above	True positive or true pass	2

Table 4.2: Confusion matrix table for the results of CSC 201 for 100 students

	Positive	Negative	Total
Positive	TP 2	FP 35	37
Negative	FN 46	TN 17	63
Total	48	52	100

With the confusion matrix in Table 4.2, probability of true positive rate, true negative rate and accuracy rate are calculated in what follows:

$$\text{Probability of true positive rate} = \frac{TP}{TP + FN} = 0.042$$

$$\text{Probability of true negative rate} = \frac{TN}{TN + FP} = 0.33$$

$$\text{Probability of Accuracy rate} = \frac{TP + TN}{TP + TN + FP + FN} = 0.19$$

***Example* 4.2**

Table 4.2: The results of CSC 202 for 100 students

Scores	Classification	Number
0 – 39	True negative or true fail	8
40 – 50	False negative or false fail	40
50 – 69	False positive or false fail	46
70 – above	True positive or true pass	6

Table 4.3: Confusion matrix table for the results of CSC 202 for 100 students

	Observation Class		Total
Positive	TP 6	FP 46	52
Negative	FN 40	TN 8	48
Total	46	54	100

With the confusion matrix in Table 4.3, probability of true positive rate, true negative rate and accuracy rate are calculated in what follows:

$$\text{Probability of true positive rate} = \frac{TP}{TP + FN} = 0.13$$

$$\text{Probability of true negative rate} = \frac{TN}{TN + FP} = 0.15$$

$$\text{Probability of Accuracy rate} = \frac{TP + TN}{TP + TN + FP + FN} = 0.14$$

5.0 Fuzzy Method

Fuzzy methods provides effective solution to simplification problems as against some techniques such as data mining which focus on probability predictions of solution to problems. In this paper the fuzzy methods provides an effective system appraisal as against data mining approach and is used as a comparative method to visual encoding technique.

In the computer environment, we have become used to the classical two valued logic of yes or no, true or false, on or off, good or bad etc. In classical set theory it is a question of an element being a member of a set or its complement. The human brain reasons with vague assertion that involves uncertainty or vague judgments such as “the air is cool”, or “that speed is fast” or “that man is tall” or “she is old”. Fuzzy logic is an area of research, which departs from the all or nothing logic. It redefines the yes or no idea. It is a system-based research topic [10 and16].

Fuzzy logic is based on the theory that a particular element can belong to more than one set at the same time. Fuzzy logic emerged from the work done on Fuzzy set theory in 1965 by Professor Lofi Zadeh of the University of California. Before the introduction of fuzzy systems, it was generally believed that an element that is not in a given set is actually not a member of that set but the new idea redefines this concept. This concept extends the conventional "crisp" set theory to classes with indistinct boundaries, which represent a large majority of natural categories and concepts. An element can belong to different categories with varying degree of membership. For example, water can be 25% cold and 75% hot. The only constraint here is that the degree of membership of an object cannot be more than the extreme

situation which exists when water is 0% cold or 100% hot, thus reducing it to the conventional set theory.

Fuzzy logic is not just restricted to just two categories as illustrated above, it can be applied to any number of the categories. For example, an element *x* can belong to set *A* with membership function a, to set B with membership function b, to set C and membership function c and so on. However it is important to keep it in mind that the sum a,b,c etc should equal unity.

5.1 Fuzzy set operation

The following rules which are common in classical set theory are also applicable in fuzzy set theory:

Sum of two sets, Product of two sets, Intersection, Complement, Containment, Equality, Associativity, Commutativity and Distributivity and De Morgan's law [14]. A more detailed discussion of these and other notions may be found in [15] and [16]. Fuzzy systems input undergo three transformations viz: Fuzzification, Rulebase and Defuzzification process

5.1.1 Fuzzification

This is a process that uses predefined membership functions that maps each system input into one or more degree of membership(s).

5.1.2 Rulebase

Rule (Predefined) is evaluated by combining degrees of membership to form output strengths.

5.1.3 Defuzzification

This is a process that computes system outputs based on strengths and membership functions. The two most popular Defuzzification methods are the Mean-Of-Maximum (MOM) and the Centre of Area (COA) methods. For MOM, the crisp output Δq is the mean value of all points ωi whose membership values $\mu c(\omega i)$ are maximum. In the case of discrete universal set W, MOM is defined by

$$X = \sum_{i=1}^{n} \frac{\omega_i}{n}$$

where

$$\{\omega_i \mid \mu_c(\omega_i) \leq \mu_c(\omega_j), \omega_i, \omega_j \in W, \omega_i \neq \omega_j \}$$

and *n* is the number of such support values. As for COA, the crisp output Δq is the centre of gravity of distribution of membership function μc. In the case of the discrete universal set W, COA is defined as in [15]:

$$X = \frac{\sum_{i=1}^{n} (\mu_c(\omega_i) x \omega_i)}{\sum_{i=1}^{n} \mu_c(\omega_i)}$$

Where *n* is the number of elements of the fuzzy set C, and $\omega \in W$. In this model, the COA method is used for Defuzzification. For continuous form see [12 and 13].

We apply the method of COA to our data in tables 3.1 and 3.2 and their centres of gravities are 29.64 and 36.99 respectively. See [8].

6.0 Computation of results

The computer program for the confusion matrix approach was developed using C++. See [7, 9 and 11]. The program is used to generate iterative values for true positive rate, true negative and associated accuracy.

The computer program outputs are in agreement with manual computation and easy for various values of true positive rate, true negative rate and associated accuracy. The program outputs are given below:

Table 6.1: The calculated result

NUM	TN	FN	FF	TP	TPR	TNR	AR
1	17	46	38	2	0.042	0.33	0.19
2	8	40	46	6	0.13	0.15	0.14

However, the program for Fuzzy system approach to calculate the centre of gravity was written in FORTRAN language.

7.0 Analysis of results

It can be observed that the confusion matrix system did not take into account what happens at the boundaries. This makes the method not to be absolutely reliable. The fuzzy system considers the degree of belongingness of any particular score with other members of the group. This makes fuzzy system more reliable than the confusion matrix system. However, the two methods show that the performances in both subjects are not satisfactory.

8.0 Conclusion

The performance is better in CSC 202 than CSC 201 because the Centre of Gravity of CSC 202 is 36.99 and that of CSC 201 is 29.64 as reported by fuzzy system. The confusion matrix system did not take into account the appraisal between CSC 202 and CSC 201 as an entity. More work is still on going.

References

[1] H. Havenstein (2006): IT efforts to help determine election successes, failure: Dems deploy data tools; GOP expands micro targeting use, computer world 40: 45; // September, 2006, 1, 16.

[2] T. G. Roche (2006): Expert increases adoption roles of certain types of ETRC, EMRS, managed Health care executive 16: 4, 58.

[3] N. Swartz (2004): IBM, mayo Clinic to mine medium data, the Information Management Journal 38: 6, Nov/Dec 2004.

[4] S. S. Wenz et. al (2006) The study and verification of mathematical modeling for customer purchasing behavior, Journal of Computer Information Systems 47 : 2, 46 – 57.

[5] R. M. Rejesus, B. B. Littla, A. C. Lovell (2004): Using data mining to detect crop insurance fraud; is there role for social sciences? Journal of Financial Crime 12 : 1, 24 – 32.

[6] G. S. Linoff (2004): Survival data mining for customer insight, intelligent enterprise 7 : 12, 28 – 33.

[7] John R. Hubbard (2006): Programming with C++, Published by Tata McGraw-Hill Publishing Company Limited, New Delhi.

[8] Omolehin, J.O., Oyelade, J.O., Ojeniyi, O.O. and Rauf, K. (2007)", A Comparison of fuzzy and Deterministic Models for assessing Students' Performance. Nig. Journal of pure and Applied sciences, 22, pp.1-7.

[9] Bron Goffried (1996): Programming with C, Published by McGraw – Hill, New York.

[10] D Hand, H. manila, P. Smyth (2001): Principle of Data mining. MIT Press Cambridge, M. A.

[11] Bjarie, S. (1986): The C++ Programming Language Addison Wesley, California.

[12] Lee,H.J and Lee, K.H (2001): Comparison of Fuzzy values on continuous Domain, Fuzzy set and systems, vol. 118, No 3, 411 – 428.

[13] Kim, J. K., Lee, K. H. and Yoo, S. W. (2001): Fuzzy Bin Packling Problem, Fuzzy set and systems, vol. 120, No. 3, 429 – 434.

[14] Zadeh, L. A. (1965): *Fuzzy Sets*. Information and Control 8, pp. 338-353.

[15] Zimmermann, H.J. (1987): Fuzzy Sets, Decision Making, and Expert Systems. Kluwer Academic Publishers, Boston, Dordrecht, Lancanster.

[16] Zimmermann, H. J., Becker, K., Kasmacher, H., Juffernbruch, K., Rau, G. and Kalff, G. (1993): An Intelligent Alarm System for Decision-Support in Cardioanaesthesia, Knowledge Base and User Interface. First European Congress on Fuzzy and Intelligent Technologies, Aachen, pp. 1023-1026.