Confusion Theory and Assessment

G. Donald Allen[1], Dianne Goldsby[2]

[1] Department of Mathematics, Texas A&M University
College Station, Texas 77843, USA

[2] Department of Teaching Learning and Culture, Texas A&M University
College Station, Texas 77843, USA

Abstract

Confusion theory and their associated confusion matrices have been principally used to train and evaluate machine learning. A confusion matrix is also known as a contingency table or an error matrix. They have been used to measure satellite classification of landscape types, for machine recognition of alphabetic characters, and for general pattern recognition. In this paper we will use confusion matrices as an assessment tool for student learning and understanding. This will be approached by evaluating whether the subjects know in which category a given problem resides. The application of confusion theory to student learning seems to be completely new in aspects of assessment of learning.

Keywords: *Confusion matrices, mathematical misconceptions exam, mathematics learning.*

1. Introduction

The goal of this paper is to examine responses to a survey using the tools of confusion theory. The particular type of survey can be of the preference type, most particularly of the classification type, wherein respondents are asked to select to the best choice of a fixed number of possible choices. For each item, the choices remain the same. For example, we could ask respondents to classify a problem by syllabus objective. Or we could ask respondents to classify a teaching situation by pedagogical method. The goal is not to score individuals, such would be done on a test, but to score the group collective as to how accurate it selects the measures presented. In each survey, there is a key, constructed by the test designer. In the sense of machine learning, the class collective is regarded as the "machine" and the variations of responses are studied.

What we will study in particular is whether students understood what was the principle topic or objective in a set of precalculus problems. For example, we might ask for a given problem whether the topic was percentages, fractions, or any of another seven other topics. What is unique about this study is first, the use of confusion theory, not used in assessment literature and research to our knowledge, and second, we did not ask students to solve the problems, but rather to indicate what the problem was about. This is believed to be an indicator of their knowledge about the problem.

2. Confusion Theory

This section will be set off as an introduction to the basics of confusion theory and confusion matrices [5]. In it we will include a few new measures of confusion measurement. A simple confusion matrix is shown in Table 1. In it you see the subjects being classified and the actual identifications. The columns are the counts of what the classifier selects. So for example, there were 13 images of a robin administered. The classifier selected 11 of them as robins, one as an oriole, and 1 as a meadowlark.

Table 1 – A simple confusion matrix

		Predicted			Row Sums
		Robin	Oriole	Meadowlark	
Actual	Robin	11	1	1	13
	Oriole	3	9	8	20
	Meadowlark	3	1	1	5
	Column Sums	17	11	10	38

The classifier is given an actual image of a Robin, Oriele, or Meadowlark, and then classifies it according to the internal program or algorithm. These are actual counts. The diagonal elements are referred to the **true positives**, and the strictly lower triangular entries are called **false positives**. That is, for example a Robin is predicted but in three cases each for Orioles and Meadowlarks. The entries in the upper triangular portion of the matrix are called **false negatives**. For example, two actual robin images were negatively classified respectively as an Orioles and a Meadowlark, respectively. By totaling all the entries it is observed that the classifier was required to identify a total of 38 objects, of which 13 were Robins, 20 were Orioles, and 5 were Meadowlarks. The values on the diagonal 11, 9, and 1, are the true positives. The most natural question is this: Is this classifier any good? We need a measure that will help determine this. Notice we did not use the term "metric" because the measure should really be independent of scale changes.

There are interesting studies for machine identification of alphabetic characters [2], and satellite photographic data [1], [3], [12], for examples, using confusion analysis. For a general background see [13].

Any nonnegative, nonzero matrix can be regarded as a confusion matrix if it refers to a classification scheme of select objects. Let C be an $m \times n$ confusion matrix with entries denoted by c_{ij}, $i = 1..m$, $j = 1...n$. Normally, the rows are the specific items to be classified and the columns pertain to what the classifier has determined. Rows and columns may have the same designation, or the rows may be objects into which a classification (columns) is to be determined. A confusion matrix can be counts or decimal valued. If they are probabilities C is called row stochastic.

Why should the confusion matrix be permitted to be non square? The reason is simply a consequence of the classifier algorithm classifying a object as not among the objects presented. Thus in the confusion matrix above, had the classifier selected a robin image as a bluebird, the matrix would become 3×4.

We define the overall **accuracy** of the classifier by

$$A_C = \frac{\sum_{i=1}^{m} c_{ii}}{\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij}}$$

This is simply the ratio of the number of correct identifications to the total number of classifications. It is obvious that $0 \le A_C \le 1$. For the example above $A_C = 0.553$. As you can see from the numbers, the classifier that produced this data is not very good. It makes many false positives. In practice, classifiers with higher values of A_C are preferred, if they can be found or derived. Indeed, you want to select the best classifier for a given grouping. In using more than one classifier for further combination, other factors also need to be considered.

Definition. Given a confusion matrix C a **Confusion Measure** of accuracy, $A(C)$, should satisfy the following properties.

(i) $A(C)$ is invariant under scale changes, that is, $A(aC) = A(C)$ for every positive constant a.

(ii) $0 \le A(C) \le 1$, with the conditions that $A(C) = 0$ if and only if C has zeros on the diagonal $A(C) = 1$ if and only if C is a diagonal matrix.

(iii) For any two confusion matrices of the same size

$$A(C_1 + C_2) \le \max(A(C_1),\ A(C_2))$$

Generally speaking, most accuracy measures satisfy (i) and (ii). However, (iii) requires something more like a norm structure for $A(C)$. In words (iii) tells us if the classifier is run twice, the combined confusion matrix is less accurate that the better of the individual accuracy measures.

In the following C is an $m \times n$ confusion matrix. We have already defined A_C as above. Now let us refine our definitions a bit. Define the row sums of C by

$$R_{iC} = \sum_{j=1}^{n} c_{ij}$$

Following the manner in which A_C was defined, we define the accuracy of classifier for item i is defined to be

$$A_{iC} = \frac{c_{ii}}{R_{iC}}$$

We define the average accuracy of the classifier to be

$$A_C^a = \frac{1}{n} \sum_{i=1}^{m} A_{iC}$$

If all the row sums are the same, then $A_C^a = A_C$. At times the confusion matrix comes to us in the form of percentages of each classifier, making the rows have all the same sums. For our example in Table 1

Local Sums			Local Accuracies				
R_{1C}	=	13	A_{1C}	=	$\frac{11}{13}$	=	0.846
R_{2C}	=	20	A_{2C}	=	$\frac{9}{20}$	=	0.450
R_{3C}	=	5	A_{3C}	=	$\frac{1}{5}$	=	0.200

The average of these is

$$A_C^a = \frac{1}{3}\left(\frac{11}{13} + \frac{9}{20} + \frac{1}{5}\right) = 0.499$$

This value is lower than A_C and reflects better the poor classification of treatments 2 and 3.

Another measure of accuracy is the **geometric mean** of the A_{iC}. This is given by

$$A_C^g = \left(\prod_{i=1}^{m} A_{iC}\right)^{\frac{1}{m}}$$

Note the number of local averages is the same as the root. For our example,

$$A_C^g = \left(\frac{11}{13}\cdot\frac{9}{20}\cdot\frac{1}{5}\right)^{\frac{1}{3}} = 0.423$$

This further amplifies the poor classification of treatments 1 and 3. We know that the geometric mean is always smaller than the arithmetic mean, that is

$$A_C^g \le A_C^a$$

This of course holds for any finite set of positive numbers. As a matter of note to be taken up later, we have that A_C and A_C^a satisfy confusion measure properties (i) and (ii) above. However, if any of the diagonal entries of C is zero, then $A_C^g = 0$. Our next example is to generalize the the accuracy measure A_C to other norm-type functions. Define the p -measures A_C^p by

$$A_C^p = \frac{\left(\sum_{i=1}^{m}(c_{ii})^p\right)^{1/p}}{\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}} \quad 1 \le p < \infty$$

In the special $p = \infty$ case we define

$$A_C^\infty = \frac{\sup_{i=1..n}|c_{ii}|}{\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}}$$

In fact, we have the

Theorem. All of the measures, A_C^p for $1 \le p \le \infty$ satisfy all three properties of a Confusion Measure.

The proof is relatively straightforward and is omitted, though we mentioned it follows from the standard norm-type arguments. Even though all our confusion matrix values are non negative, we've used the supremum (actually the maximum in the case of finite sets) with the absolute values. It is a fact that

$$A_C^p \ge A_C^q \quad \text{if} \quad 1 \le p \le q \le \infty$$

This implies the larger the value of p in this class of measures, the less the measured accuracy.

Kappa. The next main measure for accuracy, κ, sometimes referred to as the Cohen κ, attempts to compensate for what the average confusion matrix may be given a random matrix with the same row and column sums [4], [6]. Its definition presents a challenge because the “random” matrix is rather manufactured for convenience of application. Define the $m \times n$ **expectation** matrix E by

$$E_{ij} = R_i C_j / \mathrm{T}, 1 \le i \le m, 1 \le j \le n$$

where

$$T = \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}$$

$$R_i = \sum_{j=1}^{n} c_{ij}, \quad 1 \le i \le m$$

$$C_j = \sum_{i=1}^{m} c_{ij}, \quad 1 \le j \le n$$

Now suppose that E is the average matrix among all confusion matrices *having these same row and column sums*. Clearly, if we restrict ourselves to integer entries, there will be a finite number of such matrices. E is then just the arithmetic average of all such matrices. In some way then E is the "expectation matrix" given the row and column sums of our original confusion matrix C. It will be the numerical average over all such matrices. Then, in some fashion, $C - E$ expresses how the original confusion matrix differs from what might be expected by pure chance. Considering A_C, which you might also note appears to be the probability of a correct classifier, and A_C as the same for E we define the **kappa** measure of C to be

$$\kappa(C) = \frac{A_C - A_E}{T - A_E}$$

where $T = \sum_{i=1}^{n} R_i = \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}$. The only property of $\kappa(C)$ as a Confusion Measure (above) is that $\kappa(C) \le 1$. It is not even positive. In fact, when

$\kappa(C) < 0$ we can be assured that C is a terrible classifier. Here is a list of the quality of classifications as measured by κ, though this is a matter of interpretation. What kappa indicates a good or poor classifier? We have this ad hoc interpretation [4].

Table 2 – kappa interpretation

κ	Agreement
< 0	Poor
[0,02]	Slight
[0.2,0.4]	Fair
[0.4,0.6]	Moderate
[0.6,0.8]	Substantial
[0.8,1.0]	Almost Perfect

This make computation possible as the actual expected matrix is difficult to compute from probabilistic considerations. If the row and column sums are equal to one, the marginal totals are then proportions. These, in turn, can be interpreted as probabilities. This interpretation is that E_{ij} is taken as the joint probability of a classification of i as a j [5]. The question is what is this "expectation" matrix in relation to the true expected matrix as described above? The true expectation matrix, namely the probabilistically developed matrix, is difficult to compute. Indeed, it is rather difficult to compute. Even in the 2×2 case there is no simple formula. For the expectation matrix associated with Table 1, we have

Table 3 – Expectation Matrix

		Predicted			Row Sums
		Robin	Oriole	Meadowlark	
Actual	Robin	5.82	3.76	3.42	13
	Oriole	8.95	5.79	5.26	20
	Meadowlark	2.24	1.45	1.32	5
		17	11	10	38

And the associated accuracy and kappa we have in

Table 4- Kappa Calculation

Row Sums	Diagonal Entries	Item Accuracies
13	5.82	0.45
20	5.79	0.29
5	1.32	0.26
38	12.92	
Accuracy		0.34
Average Item Accuracies		0.33
Kappa		0.332

From Table 3, we assess the "birds" classifier to be at best fair. We are now prepared to apply this analysis to the assessment of student knowledge and learning.

3. The Study

The survey. We have taken a 25 question test on misconceptions in algebra and arithmetic and classify them as to what principle mathematics topic it fits best [11]. Misconceptions in mathematics have been studied for years, but persist in all elementary courses [8], [9], [10]. All questions were fairly typical of the subject, but the distractors for the multiple choices were specifically chosen so that a student pursuing the solution incorrectly would likely find their incorrect answer in the list. In the present survey, students read only the question, seeing none of the distracters. Instead, we are gaining the collective opinion of many students, all with very similar training, all currently enrolled in the same course, all with the same math background of only the problem type. It requires fewer problems, and moreover gives a collective opinion. We are not asking students to solve these problems, but just put them in the correct category. We intend to measure

1. Respondent (i.e. student) confusion about selecting the dominant problem objective
2. Examining how the key would change if the student voting majority determined the key. This would be (added this) recalibrating confusion by using a voter preference method.
3. Examining a range of measures of confusion.

The 63 participants in the survey consisted of 10 sophomores, 38 juniors and 15 seniors at a large south central Texas university. They were currently enrolled in a problem solving course required for all middle grades certification students. This course incorporated modeling pedagogical instructional strategies for problem solving but was not a methods course. Most students had taken or were taking mathematics for teachers course and had completed other math courses. The two sections of the course had 3 male students and 29 females and 3 males and 28 females, respectively. All had previously taken standard one semester business calculus and finite mathematics courses.

There were nine possible problem classifications from which a student could select.

Table 5 – Problem Classifications

Number of Items	Question Objectives	Code
4	Exponents, roots, radicals	1
3	Order of Operations	2
2	Magnitude of negative numbers	3
3	Factoring/Distributive property	4
5	Percentages	5
2	Exact vs. Approximate	6
3	Absolute Value, Inequalities	7
2	Fractions	8
1	Translational	9

To contain tables of results, we assigned a code to each classifier or objective. These nine categories comprise the fundamental syllabus of an algebra/precalculus course [8]. Given in another year to a different but similar population, the average grade for the actual test was about 50%. It is expected that this group would have performed similarly. Moreover, this is consistent with data taken over large populations of students. The highest any group of students scored on the test was about 75% for freshmen mathematics majors. For the classification survey, the score in correctness of classification was about 71%. This indicates that while a student may know what type of problem is presented, he/she may not be able to solve it. Typical questions are shown below. The "experts" were the authors in deciding which problem should be classified into which category. We also explore the use of a *voters* key concept, to be defined below.

Table 6 – Typical Questions

1. Simplify $\sqrt{(-8)^2}$	4. Simplify $\frac{2+6\sqrt{18}}{2}$
9. Simplify $\frac{5x}{12}-\frac{8y}{15}$	15. Convert 87.49 into a percentage.
18. Solve $(x+6)(x-4)=11$	25. Convert 16.2% into a fraction in simplest form.

The problem for which the students performed overwhelmingly the worst on the actual test was number 11: *For every 10 families, there were 7 pets. Write an equation representing this relationship, where F is the number of families and P is the number of pets.*

4. Results

In the Tables 7-9, we have suppressed the row and column sums. There were a total of 1566 responses, not exactly 25x63=1575, as some did not respond to select questions. However, this very high total response rate was an unexpected return since students could simply omit any question. The expectation matrix is shown in the next table.

Table 7- Confusion Matrix for Study

Confusion matrix										
		Predicted								
	Code	1	2	3	4	5	6	7	8	9
Actual	1	196	33	0	4	0	0	0	17	0
	2	0	73	16	1	0	13	0	0	21
	3	0	0	20	0	0	0	1	42	0
	4	0	21	0	72	0	0	0	31	1
	5	0	1	0	1	280	0	2	27	4
	6	0	1	0	5	25	22	0	69	4
	7	3	8	0	12	0	2	87	0	12
	8	0	0	0	3	1	0	2	56	1
	9	2	13	0	21	0	4	6	59	20

Table 8 – Expectation Matrix for Study

Expectation Matrix										
		Predicted								
	Code	1	2	3	4	5	6	7	8	9
Actual	1	34	35	13	27	49	7.8	26	48	11
	2	25	26	9.4	20	37	5.9	19	36	8.2
	3	17	18	6.4	14	25	3.9	13	24	5.6
	4	25	26	9.4	20	37	5.9	19	36	8.2
	5	51	53	19	41	74	12	39	73	17
	6	8.4	8.8	3.2	6.8	12	2	6.5	12	2.8
	7	25	26	9.4	20	37	5.9	19	36	8.2
	8	17	18	6.4	14	25	3.9	13	24	5.6
	9	8.3	8.7	3.1	6.7	12	1.9	6.4	12	2.7

We summarize the survey as a classifier in the next table.

Table 9 – Summary

Accuracy	0.7120
Item Accuracy	0.6408
Difference	0.0712
Kappa	0.6737

It should come as no surprise that the accuracy is about the same as the average class grade. This is because the accuracy is the total number correct divided by the number of answers given. But, we were not seeking a grade but rather the confusion about problem types. This is well revealed by examining Category 8: Fractions. Many students confused a problem not deemed to be principally one of fractions as strictly a fractions problem. The value of kappa, by Table 2, indicates substantial agreement with the expert classifier.

At this point, we consider the voter's key concept. That is we took the majority (plurality) of what student viewed as the most significant classifier. This was performed to look for internal consistency among the group's performance. This changed the key somewhat and therefore the respective confusion information. This is displayed below. Note that none of the students classified any problems in category 6: Exact vs. Approximate. This is one of the troublesome aspects of teaching mathematics in the early grades, the apparent inability of students to distinguish between the two concepts. This condition persists into the collegiate level.

Table 10 – Voters Confusion

Voters Confusion matrix										
		Predicted								
	Code	1	2	3	4	5	6	7	8	9
Actual	1	196	33	0	4	0	0	0	17	0
	2	0	124	0	1	0	0	0	1	0
	3	0	4	42	0	0	7	3	1	6
	4	10	18	0	121	0	4	7	6	21
	5	0	1	0	1	280	0	2	27	4
	6	0	0	0	0	0	0	0	0	0
	7	3	9	1	12	0	3	147	0	12
	8	1	20	20	29	26	22	3	250	5
	9	0	10	16	1	0	13	0	0	21

Table 11 – Voters Expectation

Voters Expectation Matrix										
		Predicted								
	Code	1	2	3	4	5	6	7	8	9
Actual	1	33.5	35	12.6	27	48.9	7.83	25.9	48.2	11
	2	16.9	17.6	6.36	13.6	24.6	3.95	13	24.3	5.56
	3	8.45	8.82	3.18	6.8	12.3	1.97	6.52	12.2	2.78
	4	25.1	26.2	9.44	20.2	36.6	5.85	19.4	36.1	8.24
	5	42.3	44.1	15.9	34	61.6	9.86	32.6	60.8	13.9
	6	0	0	0	0	0	0	0	0	0
	7	25.1	26.2	9.44	20.2	36.6	5.85	19.4	36.1	8.24
	8	50.5	52.6	19	40.6	73.5	11.8	38.9	72.6	16.6
	9	8.19	8.54	3.08	6.59	11.9	1.91	6.31	11.8	2.69

The corresponding summary is somewhat different from the other case.

Table 12 – Summary

Accuracy	0.7546
Item Accuracy	0.6407
Kappa	0.7237

Here it is of note the accuracy is somewhat higher, indicating that there was definite consistency in how students classified the problems. That should be considered favorable. The value of kappa also indicated the students are even more internally consistent with each other than with the expert. Note as well that a chi-squared test between the voters confusion matrix and confusion matrix reveals no difference between the two matrices. In this sense, the confusion analysis is more sensitive than, at least, simple statistical analysis.

5. Conclusions

This is a study to demonstrate that a tool heretofore used substantially in the domain of engineering can be configured to address the learning of another kind of machine, people.

One of the reasons for the relatively high agreement with expert opinion is that the material had previously been learned. In that context, it proved that learning did take place. Another place for this type of study would be applied to more formative learning when students are in the process of learning. For example, in calculus II, such a test could be given in advance of the major exam inquiring whether students understood which integration method to apply. In a physics course, the pre-test could examine which of the inventory of forces a particular situation indicated. In engineering courses the objective of the pre-test could relate to design considerations.

In a mathematics problem solving course for teachers the test could be used to determine conceptual understanding of topics in order to align correctly content and standards. The experience of such an activity could emphasize to students the necessity of not only solving problems but also correctly classifying them as to topic and concepts as they will be teaching in classrooms in which standards alignment is required. A teacher needs to understand and recognize the essential mathematics topic to be able to select meaningful instructional tasks and generalizable strategies and to make connections to other topics and concepts. Making mathematical relationships explicit have been found to result in conceptual understanding [14].

One particularly and exciting aspect of such classifier tests is they take only minutes to complete. One may argue that if a student doesn't even know what a given problem is about, their prospects of actually solving it are diminished.

In general such a test could comprise a strong predictor of success and learning ability. The type of information obtained about student learning and understanding from confusion analysis is new.

Finally, it could be argued it would have been better to get an individual to classify several hundred problems. But then we would need several hundred non-ambiguous problems and a person willing to make such the classification key. That achieved, we give the survey to a single person. What is achieved is basically how well one person (the expert) compares with the classifier (the respondent).

References

[1] Stehman, Stephen V. Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment 62 (1), 1997, 77--89.

[2] J. T. Townsend. "Theoretical analysis of an alphabetic confusion matrix." In: Attention, Perception, & Psychophysics, 9(1), 1971.

[3] Stehman, S.V. and R.L. Czaplewski. Design and analysis for thematic map accuracy assessment: Fundamental principles. Remote Sensing of Environment, 64, 1988, 331-344.

[4] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. Biometrics, 33(3):159-174, 1977.

[5] Y Bishop, S Fienberg, and P Holland. Discrete multivariate analysis: theory and practice. MIT Press, Cambridge, MA, 1975.

[6] Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 1960, 37-46.

[7] Barnett, Raymond A., Michael R. Ziegler, and Karl E. Byleen. *College Algebra*, 8th edition,McGraw-Hill, 2008.

[8] Resnick, L. Mathematics and science learning: A new conception. Science, 220, 477-478, 1983.

[9] Mestre, J. Why should mathematics and science teachers be interested in cognitive research findings? Academic Connections, 1987, pp. 3-5, 8-11. New York: The College Board.

[10] Pines (Eds.), A. L., Towards a taxonomy of conceptual relations. In L. West and A. L. Pines (Eds.) Cognitive structure and conceptual change (pp.101-116). New York, Academic Press, 1985.

[11] Allen, G. Donald, Scarborough, Sherri, and Goldsby, Dianne, Misconceptions exam for methodology. Internet: http://disted6.math.tamu.edu/confusion/confusion.html

[12] Smith, J.H., S.V. Stehman, J.D. Wickham, and L. Yang. Effects of landscape characteristics on land-cover class accuracy. Remote Sensing of Environment 84, 2003, 342-349.

[13] USGS, Accuracy Assessment of 1992 National Land CoverData, 2012,Internet: http://landcover.usgs.gov/accuracy/index.php

[14] Bay-Williams, J.M. Influences on student outcomes: Teachers' classroom practices. In D. Lambdin (Ed.), Teaching and learning mathematics: Translating research to the classroom. (pp. 31-36). Reston, VA: NCTM, 2010.

Author Biographies

G. Donald Allen has BS, MS, and PhD degrees from the University of Wisconsin, the most recent in 1971. He has been employed at Texas A&M University through most of the time since. Currently, he is professor of mathematics and Director of the Center for Technology-Mediated Learning in Mathematics. With more than US$15m in grants over all, he is author of more than 60 papers in mathematics and mathematics education, three books, in addition to dozens of shorter papers on mathematics and educational topics, and has given more than 200 talks, workshops, and webinars over the past two decades. Allen works in mathematical modeling, mathematics education, assessment, and has experience in functional analysis. Allen sits on the executive steering committee of International Conference on Technology in Collegiate Mathematics (ICTCM), and is a member of both the Athens Institute for Education and Research (ATINER) and the US-Finland SAVI (Science Across Virtual Institutions).

Dianne S. Goldsby has a BS and M Ed in mathematics education and a PhD in Curriculum and Instruction. She is a teaching award winning Clinical Professor at Texas A&M University in the Teaching, Learning and Culture Department and Coordinator of the department online M Ed programs. She has over 60 presentations at international, national and regional conferences, has written chapters in books and articles in peer-reviewed journals, and worked on grants totaling US$1.5 million. She has been an associate editor of SSMJ and serves as a reviewer for math education journals. Goldsby works in the area of teacher perceptions of mathematics and their perceptions of the teaching and learning of mathematics.