Confusion Theory for Pedagogical Decisions

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

[1] Department of Mathematics, Texas A&M University
College Station, Texas 77843, USA
gdonaldallen@gmail.com

[2] Department of Teaching Learning and Culture, Texas A&M University
College Station, Texas 77843, USA
dgoldsby@tamu.edu

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.

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 is the principle method by which a particular topic in middle school math should be taught. There are about nine general methods used in middle school, and really by us all. They are: (a) Rules, (b) Exploration/Inquiry, (c) Guided Invention, (d) Mental Math, (e) Examples, (f) Models, (g) Group Learning, (h) Theory, and (i) Direct Instruction . Remarkably, there was little agreement among students or faculty creating keys on what method to apply to what type of problems. Some of the methods, such as theory were not considered for any topic. We attempt to make some sense of this, though ultimately, the confusion of student may reflect fundamental confusion on the use of methods in the teaching profession, in

general. 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 to teach the topic, but rather to describe how they might teach it. . This is believed to be an indicator of their teaching concepts – or PK, pedagogical knowledge.

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, Oriole, 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, 13 of which 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 a $m \times n$ 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 an 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

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

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, $A\ aC \quad A\ C$ for every positive constant a.

(ii) $0 \le A(C) \le 1,$ with the conditions that $A\ C \quad 0$ 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

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

We define the average accuracy of the classifier to be

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

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

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

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

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

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

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

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

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

$$C \quad \frac{A_C \quad A_E}{T \quad A_E}$$

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 $\kappa(C) \leq 1$. 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	Meadowlar	
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 – Part I

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.

These results have been previously reported [15]. They show that students were in general agreement on what techniques were needed to solve various math problems. Both accuracy and kappa were very high.

4. The Study – Part II

In this section, we changed the goal to examine what pedagogy students might select as teaching methods. We gave this pedagogy survey to 83 pre-service teachers. The goal was for the students to identify the proper pedagogy to teach a given topic. All questions are included in Appendix A. Typical questions include the likes of

1. I know a pound of coffee costs $2.50. How much does six ounces cost? How do I teach this?
2. Every student understands fractions by informal knowledge. How do I reinforce this?
3. What is the best way to teach the idea of a linear relationship?
4. What is the best way to teach computations with decimals?

The responses were available were Exploration/Inquiry, Guided Invention, Mental Math, Examples, Models, Group Learning, Theory, and Direct Instruction. For each question, a key was constructed by the authors. In addition, other instructors were asked to take this survey, and from their responses we constructed other keys. Note, unlike a mathematics test per se, this was not a test, and a variety of responses are possible. Essentially, there are multiple correct, or at least acceptable, responses. For the data below we add the code-key for the responses in Table 5. Included as well is the number of items with the given code-key.

Table 5 – Code Key and Numbers of Items

Response	Code	Number Items
Rules	1	1
Exploration/Inquiry	2	2
Guided Invention	3	4
Mental Math	4	1
Examples	5	5
Models	6	10
Group Learning	7	0
Theory	8	0
Direct Instruction	9	2

The results are interesting. In Table 6 we show the responses compiled as a confusion matrix. Results were similar for all the keys we tried, i.e. no matter what instructor indicated their preference.

As is evident most students selected using models for almost every type of question, indicating that modeling was a significant factor in their thinking. In addition to modeling, clearly exploration, guided intervention, and examples were strongly favored. Rules, group learning, theory, direct instruction, and mental math were not in their thoughts. Consequently, the accuracy was very low as shown in Table 7. Even the kappa measure was low, indicating only slight-fair agreement.

Table 6 – Basic Confusion Matrix

Confusion matrix for problem types											
		Predicted									
		1	2	3	4	5	6	7	8	9	Row Sum
Actual	1	0	0	0	0	0	0	0	0	0	0
	2	58	98	59	8	158	143	41	28	70	663
	3	18	28	13	16	28	12	28	3	18	164
	4	0	0	0	0	0	0	0	0	0	0
	5	46	2	7	0	12	3	3	1	9	83
	6	114	123	68	20	260	309	63	26	173	1156
	7	0	0	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	0	0	0
	9	0	0	0	0	0	0	0	0	0	0

Table 7 - Accuracy

	Accuracy	**0.2091**
	Item Accuracy	**0.1597**
	Difference	**0.0494**
	Kappa	**0.0355**

The issue at hand for this study is what happened. Why didn't we achieve at least a plurality of key responses for most of the questions? Here are a few possible reasons.

- The questions were vague
- There were multiple correct answers
- We did not clarify at which stage of the learning process the student was
- Items 2, 3, 5 and 6 are strongly favored in pre service teacher education

Probably, all of them figured into the mix, though there was some improvement by grouping responses.

In conclusion, we may note

- Can be given as a pretest to identify whether a student even knows how to proceed.
- Can be used to identify vague questions, weak predicates. (i.e., to identify actual confusion).
- Can be used as a formative instrument to identify possible problems.

From our earlier study, pre service teachers are fairly certain of what type of problem they are considering. However, they are not at all clear on how to teach the topic. It must be noted these students had not yet been in a mathematics methods course which would focus on appropriate pedagogical methods for the topics. In addition, they were just beginning field placements in middle grade classrooms. With the current focus on understanding of topics rather than just rote procedural knowledge [16], pre-service teachers are engaged in more hands-on methods of instruction, rather than direct instruction methods. This reform approach to teaching and learning mathematics has resulted in a need for a change in beliefs about mathematics beyond just learning methods and materials [17]. This emphasis on strategies other than rules and direct instruction could be a result of the interactions they were experiencing in their course work, creating a need for these changing beliefs. In addition the instructor emphasized the need for multiple strategies in mathematics. This focus on multiple approaches and then the survey requesting a selection of one approach could have been responsible for some of the confusion.

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 Cover Data, 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.
[15] Allen, G. Donald and Goldsby, Dianne, (2014) Confusion Theory and Assessment, IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 10.
[16] National Council of Teachers of Mathematics, (2000) Principles and Standards for School Mathematics. Reston, VA: Author, 2000.
[17] Richardson, V. & Placier, P. (2001) Teacher change. In V. Richardson, (Ed.), Handbook of research on teaching (4th ed., pp. 905-947). Washington, DC: The American Educational Research Association.

Appendix A. Pedagogical Methods Classification. The first question has the response choices. They were the same for all questions.

1. What method should I use to teach integers (signed numbers)?

○ Rules	○ Exploration/Inquiry	○ Guided Invention
○ Mental Math	○ Examples	○ Models
○ Group Learning	○ Theory	○ Direct Instruction

2. What is the best way for students to learn the formulas for area and volume of figures and solids?

3. I need to show how to divide by a fraction. What is the best method to use?

4. What is the best way to have students learn to multiply by a fraction?

5. How do we develop the grouping-by-tens concept in place value?

6. How should students be taught to count rationally?

7. How do we connect the operations of addition and subtraction?

8. What is the best way to help students master the basic number facts?

9. How do we teach flexible methods of whole-number computation?

10. How do we connect the operations and multiplication and division?

11. What is the best way to teach computations with decimals?

12. How do we teach the correct meaning of the "=" sign?

13. I need to show students the nature of a unit measure.

14. I need to show students how to subtract fractions. What method do I use?

15. In showing students the order of operations the best method is

16. My curriculum requires I show students what a fraction means. What method should I use?

17. Every student understands fractions by informal knowledge. How do I reinforce this?

18. I know a pound of coffee costs $2.50. How much does six ounces cost? How do I teach this?

19. On Monday, baseball player Joe hits 3 for 4. On Tuesday he hits 1 for 3. Over both days he hits 4 for seven. How should this be explained?

20. I am teaching the concept of an unknown, say x. How do I teach this?

21. What is the best way to teach the idea of a linear relationship?

22. What is the best way to teach the idea of a nonlinear relationship.

23. I have a complex area created by positioning a group of rectangles.

24. How should I teach how to find the total area?

25. I need to teach the multiplication of two digit numbers. How do I teach this?