Abstract: This paper describes the use of fuzzy logic to generate norm-referenced student assessments. Formative assessment is described, with comparison between criterion-referenced and norm-referenced evaluations. The proposed fuzzy evaluation is described with ratings ‘excellent’, ‘good’, ‘average’, ‘below average’, and ‘weak’, followed by description of the implementation prototype system. Finally the results of testing the prototype using a sample class section are given.

Key-Words: formative assessment, fuzzy system, formative evaluation, student assessment, fuzzy logic, norm-referenced evaluations.

1 Introduction

The current learning environment is evolving with the introduction new methods and the use of technology in the classroom. The learning process has become more systematic and focused on the learning outcomes to be delivered to the student, making it easier to formalise the lessons and ensure that the students are learning the right knowledge and skills as intended in the syllabus of a subject or course. The focus is on the student performance – “the best way to infer learning is by measuring changes in students’ performance” [1].

The assessment process is to provide an indication to the student on how much have been learnt, and more importantly how much that was not learnt or understood to be improved upon in subsequent learning process. For the teacher or lecturer, the assessment can provide an indication on the effectiveness of the teaching process and whether remedial course would be required to improve the learning of the subject.

It is important for assessment to be accurate as it supports the decision making by the teachers. The decisions can be instructional management decisions, selection decision, placement decision or classification decision [2] that has significant impact on the student.

Furthermore, in e-learning systems, assessment of students while learning has become an important feature to provide better delivery of learning contents. Since the students are doing self-learning, accurate assessment is necessary to support the student in learning more effectively.

2 Formative Assessment

Formative evaluation is the evaluation on the quality of a student’s achievement of learning targets while the student is in the process of learning, whereas summative evaluation is the evaluation on the quality of a student’s achievement of learning targets after the instructional process is complete [2].

There are many evaluations that can be conducted for a subject. Some of the different types of evaluations are:

- Test
- Quiz
- Assignment/Project
- Tutorial
- Laboratory exercises

Each evaluation gives a score that measures the performance of the student on that particular evaluation. The formative assessment is obtained through the combination of the evaluations, usually with weights denoting different significance of each evaluation.

For example, the formative evaluation for student for student $x$ would be:

$$ f(x) = w_1e_1(x) + w_2e_2(x) + ... + w_ne_n(x) $$

$$ = \sum_{i=1}^{n} w_i e_i(x) \quad (1) $$

where $e_n(x)$ is evaluation n for the student and $w_n$ is the corresponding weight of the evaluation.
Using this formula, each student will obtain a specific score that would represent the formative assessment for the student. For example, a student with the scores in Table 1 will have a formative assessment score of

\[ f(x) = 0.2(60) + 0.1(70) + 0.5(50) + 0.2(75) = 69 \]

Table 1 Example of weights and scores for formative evaluations

<table>
<thead>
<tr>
<th>Evaluation No. (i)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight ((w_i))</td>
<td>0.2</td>
<td>0.1</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>Score ((e_i))</td>
<td>60</td>
<td>70</td>
<td>50</td>
<td>75</td>
</tr>
</tbody>
</table>

2.1 Criterion-referenced Evaluation

Criterion-referenced tests indicate the student level of proficiency or mastery of some skill or set of skills, giving a “grade” from a standard of mastery called a criterion. For example, the formative assessment score of 69 may be given a grade such as “B” or “Credit” in criterion-referenced evaluation. A typical grade function \(g(x)\) can be as follows:

\[
g(x) = \begin{cases} 
  g_1 & \text{for } f(x) \geq a_1 \\
  g_2 & \text{for } a_1 \geq f(x) \geq a_2 \\
  \vdots \\
  g_d & \text{for } f(x) \leq a_d 
\end{cases}
\]

where \(G = \{g_1, g_2, \ldots, g_d\}\) is the set of grades and \(\{a_1, a_2, \ldots, a_d\}\) are the boundary values between the different grades.

2.2 Norm-referenced Evaluation

Norm-referenced tests indicate the student performance as compared to other students, giving a “place” or “rank” for each student. For norm-referenced evaluation, the score will be compared against other student scores and given a ranking or position, for example using percentile ranking [3]. Percentile ranking is a statistical measure that is calculated based on the percentage of students ranked below the student currently being ranked. We can calculate \(p(x)\) as follows:

\[
p(x) = \frac{100(e - l)}{2n} 
\]

where \(e\) is the number of students with the same score, \(l\) is the number of students with lower score and \(n\) is the total number of students.

3 Fuzzy Evaluations

In writing about qualitative formative evaluations, Sadler [4] remarked that some of the criteria used in appraisal are “fuzzy rather than sharp”. It was recognised that between different criteria, it is possible to have continuous gradation from one state to another rather than an abrupt transition. Rather than relying on specific scores to assess a student’s performance, it is a common practice to assess students in terms such as ‘excellent’, ‘good’, ‘average’, ‘weak’, etc. These forms of categorizing students are most suitable to apply fuzzy logic because they are usually not crisply defined with specific score boundaries. Thus, we can create fuzzy sets to represent groups of students that are described with these terms. These terms can be used by teachers to make decisions in the learning process, such as deciding on the next learning activity to be done.

3.1 Linguistic Variables

The first step in designing the fuzzy model is to determine the linguistic variables. In making decision about a student’s performance, the following are the terms used:

<table>
<thead>
<tr>
<th>Term (Linguistic variable)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>Students who performed very well compared to the other students – can self-study with minimal supervision</td>
</tr>
<tr>
<td>Good</td>
<td>Students who performed reasonably well compared to other students – can self-study but may need guidance occasionally</td>
</tr>
<tr>
<td>Average</td>
<td>Students who are in the middle of the class – may need some guidance sometimes</td>
</tr>
<tr>
<td>Below Average</td>
<td>Students who performed not so well – may need guidance and attention to improve</td>
</tr>
<tr>
<td>Weak</td>
<td>Students who performed worse than the other students – may need extra guidance and attention</td>
</tr>
</tbody>
</table>

In criterion-referenced evaluation, these terms will be determined by specific boundary values for the scores as in Equation 2. For norm-referenced evaluation, these terms will be determined by the range for the percentile ranking e.g. ‘Excellent’ will be for students with percentile rankings above 80% and ‘Average’ will be for percentile rankings between 40% and 60%. However, these boundaries...
are usually fuzzy and these terms can be represented better with fuzzy logic.

### 2.2 Membership Functions

Fig. 1 illustrates the membership functions for the different terms used to rate the students. For a student, the percentile ranking \( p(x) \) is used to derive the different membership values \( F_e(x) \), \( F_g(x) \), \( F_a(x) \), \( F_b(x) \), and \( F_w(x) \) for the ratings ‘Excellent’, ‘Good’, ‘Average’, ‘Below Average’, and ‘Weak’ respectively. The equations for the membership functions are given in Equations (4) to (8), to be used to fuzzify the percentile ranking of the student.

\[
F_e(x) = \begin{cases} 
1 & \text{for } 90 \leq p(x) \leq 100 \\
\frac{p(x)-80}{10} & \text{for } 80 \leq p(x) \leq 90 \\
0 & \text{otherwise} 
\end{cases} \quad (4)
\]

\[
F_g(x) = \begin{cases} 
\frac{p(x)-80}{-10} & \text{for } 70 \leq p(x) \leq 80 \\
1 & \text{for } 60 \leq p(x) \leq 70 \\
\frac{p(x)-60}{10} & \text{for } 50 \leq p(x) \leq 60 \\
0 & \text{otherwise} 
\end{cases} \quad (5)
\]

\[
F_a(x) = \begin{cases} 
\frac{p(x)-60}{-10} & \text{for } 60 \leq p(x) \leq 70 \\
1 & \text{for } 40 \leq p(x) \leq 60 \\
\frac{p(x)-30}{10} & \text{for } 30 \leq p(x) \leq 40 \\
0 & \text{otherwise} 
\end{cases} \quad (6)
\]

\[
F_b(x) = \begin{cases} 
\frac{p(x)-30}{-10} & \text{for } 30 \leq p(x) \leq 40 \\
1 & \text{for } 20 \leq p(x) \leq 30 \\
\frac{p(x)-10}{10} & \text{for } 10 \leq p(x) \leq 20 \\
0 & \text{otherwise} 
\end{cases} \quad (7)
\]

\[
F_w(x) = \begin{cases} 
1 & \text{for } 0 \leq p(x) \leq 10 \\
\frac{p(x)-10}{-10} & \text{for } 10 \leq p(x) \leq 20 \\
0 & \text{otherwise} 
\end{cases} \quad (8)
\]

### 3.3 Aggregation

In most learning environment, each student would have many evaluations while learning. Thus each student would have a set of evaluations \( \{p_1(x), p_2(x), \ldots, p_n(x)\} \) where \( n \) is the number of evaluations. When each evaluation is fuzzified, the set of evaluations obtained are shown in Equation (9). This is similar to Ma [5] for criterion-referenced evaluations.

\[
E = \left\{ e_{11}(x), e_{12}(x), \ldots, e_{1d}(x), e_{21}(x), e_{22}(x), \ldots, e_{2d}(x), \ldots, e_{n1}(x), e_{n2}(x), \ldots, e_{nd}(x) \right\} \quad (9)
\]

Each value \( e_{ij}(x) \) is derived using the corresponding membership function \( F_j(x) \) given in Equations (4) to (8).

To obtain a value representing a particular rating, the membership values for the rating are aggregated with the following function:

\[
f_j(x) = \frac{\sum_{i=1}^{n} e_{ij}(x)}{n} \quad (10)
\]

The set \( \{f_1(x), f_2(x), f_3(x), f_4(x), f_5(x)\} \) represents the aggregated membership values for the ratings ‘Excellent’, ‘Good’, ‘Average’, ‘Below Average’, and ‘Weak’ respectively.

### 3.4 Defuzzification

Defuzzification is the process of identifying the representative rating for the student. The rating can be determined by selecting the highest membership value from the set of aggregated membership values. Thus the rating \( r(x) \) for the student is determined as follows:

\[
r(x) = r_p \text{ where } f_p(x) = \max(f_j(x)) \quad (11)
\]
4 Prototype

A prototype system was developed to evaluate students based on the fuzzy model described above. The main functions of the prototype are:

- read the scores for the student evaluations
- sort and rank the scores, and generate the percentile rankings
- fuzzify the percentile rankings
- aggregate the membership values
- defuzzify the membership values

Figure 2 shows the main screen of the prototype that displays the student list and the evaluations for each student.

![Figure 2 Main screen of prototype system](image)

To generate the membership values given in Equation (9), the formative evaluations are sorted and ranked, and then fuzzified. This is done in the screen shown in Figure 3.

![Figure 3 Screen for formative evaluations](image)

After generating the membership values, the aggregation is done in the screen shown in Figure 4. The defuzzification process is then done to generate the rating for each student, as in Equation (11).

![Figure 4 Screen for fuzzy aggregation](image)

4.1 Validation of results

To check the correctness of the rating, a comparison is done against the summative evaluation that is rated using the same fuzzy model. The rating from the summative evaluation is assumed to be the actual level of the student, to validate the rating ‘predicted’ using the fuzzy system. Figure 5 illustrates this process.

![Figure 5 Process to validate results of fuzzy system](image)

5 Results

The system is tested using a class section for Software Engineering subject with 69 students. The students in this subject have 4 evaluations:

- A test
- An assignment
- A report
- Tutorial exercises

The fuzzy model in Figure 1 is used to generate the results. Table 3 shows the results for the different ratings, giving the number of students for each rating and the number of students that have matching ratings for the summative evaluations. The rating with the highest percentage of correctness is the ‘Excellent’ rating, whereas the rating with the lowest percentage of correctness is the ‘Weak’ rating.
Table 3 Results of fuzzy evaluations

<table>
<thead>
<tr>
<th>Rating</th>
<th>No. of students rated</th>
<th>No. of matched ratings</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>11</td>
<td>6</td>
<td>54.55</td>
</tr>
<tr>
<td>Good</td>
<td>17</td>
<td>8</td>
<td>47.06</td>
</tr>
<tr>
<td>Average</td>
<td>19</td>
<td>6</td>
<td>31.58</td>
</tr>
<tr>
<td>Below Average</td>
<td>17</td>
<td>4</td>
<td>23.53</td>
</tr>
<tr>
<td>Weak</td>
<td>5</td>
<td>1</td>
<td>20.00</td>
</tr>
<tr>
<td>Total</td>
<td>69</td>
<td>25</td>
<td>36.23</td>
</tr>
</tbody>
</table>

The results are further analyzed for the wrong matches, to identify possible trends in the ratings. The students deemed ‘improved’ are the students with higher rating in summative evaluation compared to the rating in formative evaluation. The students deemed ‘declined’ are the students with lower rating in summative evaluation compared to the rating in formative evaluation. Table 4 shows the results for test data, which is fairly balanced between the improved and declined students. As it is normal that some students will improve and some will decline, this balance shows that there is less bias in the fuzzy model.

Table 4 Changes in student ratings

<table>
<thead>
<tr>
<th>Rating</th>
<th>No. of improved students</th>
<th>No. of declined students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Good</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Average</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>Below Average</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Weak</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>23</td>
</tr>
</tbody>
</table>

7 Conclusion

The proposed norm-referenced fuzzy evaluations were used to generate ratings for the student performance level, with overall correctness of 36.23%. The best results are for the ‘Excellent’ rating with 54.55% correctness, whereas the worst results are for the ‘Weak’ rating with 20% correctness.

Further investigation into different models needs to be done to improve the results. Other factors such as the number of formative evaluations and their weights may need to be evaluated as well.

References:

6 Further Work

The prototype system was tested using a fuzzy model with trapezoidal membership functions. This model can be further modified by changing the shape of the functions and tested to obtain the best model. Other models such as the triangular membership functions can also investigated to find the model that gives the best results.

Alternative aggregation methods may be investigated such as the use of neural networks and genetic algorithms to improve the fuzzy system. Other factors such as changing the number of ratings and evaluations would need to be investigated also.