Johanyák, Z. C.: Survey on Four Fuzzy Set Theory Based Student Evaluation Methods, Proceedings of Kecskemét College, Faculty of Technology (GAMF), Kecskemét, XXIII. (2008), HU ISSN 1587-4400, pp. 121-130. http://johanyak.hu

Survey on Four Fuzzy Set Theory Based Student Evaluation Methods[1]

Johanyák, Zsolt Csaba[2]

1. Introduction

The evaluation of students' learning achievement in case of narrative written response tests cannot be fully automated, and therefore the results significantly depend on human judgement. Even a trained expert evaluator has to face often situations when he or she cannot rank unambiguously the response given by the student in one or another grade category or score value. In such circumstances it could be extremely helpful the support of a tool that allows a rater to express the vagueness in his or her judgement.

Recently several fuzzy set theory based evaluation methods have been published, which support the whole evaluation process or a part of it. This paper surveys four of these methods emphasizing their advantages and drawbacks.

The rest of this paper is organized as follows. Section 2 introduces a condition set on fuzzy evaluation methods containing the features considered as important ones. The four methods are presented in section 3. The conclusions are discussed in section 4.

2. Conditions on methods supporting the students' learning achievement evaluation

In order to evaluate and compare the different fuzzy set based evaluation methods we defined the following condition set.

1. The method should not increase the time needed for the assessment compared to the traditional evaluation techniques.
2. The method should help the grader to express the vagueness in his/her opinion.
3. The method should be transparent and easy to understand for both parties involved in the assessment process, i.e. the students and the graders.

[1] Lektorált tanulmány

[2] PhD, főiskolai docens, KF GAMF Kar, KSII, Informatika Szakcsoport

4. The method should ensure a fair grading, i.e. it should be beneficial for all students.
5. The method should allow the teacher to express the final result in form of a total score or percentage as well as in form of grades using a mapping between them that is prescribed by the university.
6. The method should be easy implementable in software development terms.
7. The method should be compatible with the traditional scoring system, i.e. when the grader provides crisp scores for each response the total score and the final grade should be identical with the one calculated by the traditional way.

3. Fuzzy student evaluation methods

3.1. FEM and GFEM

The key idea of the Fuzzy Evaluation Method (FEM) proposed by Biswas [2] is that each question in the student answerscript is evaluated independently with a discrete fuzzy set containing membership values for six uniformly distributed predefined points (X) of the traditional percentage based evaluation scale [0,100]

$$X = \{ 0, 20, 40, 60, 80, 100 \}. \tag{1}$$

The resulting fuzzy set is compared to all of the so called Standard Fuzzy Sets (SFSs). The SFSs are defined on the same universe of discourse [0,100] corresponding to the grading standard of the university. Each SFS corresponds to a traditional grade (e.g. Excellent). The comparison is made by the means of a similarity degree that is calculated by

$$S_i\left(\overline{E_i}, \overline{SFS}_j\right) = \frac{\overline{E_i} \cdot \overline{SFS_j}}{\max\left(\overline{E_i} \cdot \overline{E_i}, \overline{SFS_j} \cdot \overline{SFS_j}\right)}, \tag{2}$$

where the index i denotes the ordinal number of the question, $\overline{E_i}$ is the vector containing the membership values of the evaluation and $\overline{SFS_j}$ is the vector describing the j^{th} standard fuzzy set, and "." denotes the dot product. Further on, the degree corresponding to the SFS with maximum similarity will represent the evaluation of the actual question.

After processing all the questions the evaluator determines a total score by calculating the weighted average of the representative values (midpoints) of the fuzzy sets corresponding to the individual grades assigned to the questions by

$$TS = \frac{\sum_{i=1}^{n} \left(T(Q_i) \cdot P(g_i) \right)}{100}, \tag{3}$$

where

$$\sum_{i=1}^{n} T(Q_i) = 100\,, \tag{4}$$

where the index *i* denotes the ordinal number of the question, *n* is the total number of questions, Q_i is the question, $T(Q_i)$ is the weight of the question, g_i is the degree assigned to the question, $P(g_i)$ is the representative value of the degree, and "." symbolizes the dot product.

Biswas also suggested a generalized version of its method called Generalized Fuzzy Evaluation Method (GFEM) [2]. GFEM evaluates each answer from four different aspects, namely the accuracy of information, the adequate coverage, the conciseness, and the clear expression. The arithmetic mean of the midpoints of the fuzzy sets representing the four grades assigned will represent the evaluation of the given question expressed with marks between 0 and 100

$$E_i = \frac{\sum_{k=1}^{4} P(g_{ik})}{4}, \tag{5}$$

where *k* identifies the point of view. One calculates the total score (*TS*) as a weighted average of the individual marks

$$TS = \frac{\sum_{i=1}^{n} (T(Q_i) \cdot E_i)}{100}. \tag{6}$$

The applied weighting is the same as in the case of FEM.

The advantage of FEM and GFEM is their easy-to-understand and easy-to-implement character. Their disadvantage is that they determine separate grades for each question applying a rounding to the most similar grade, which introduces an error in each evaluation step. The error summarizes in course of the evaluation of the answerscript and at the end it can lead to a quite strange final result. The use of the midpoints in the total score calculation is a quasi defuzzification before the final aggregation, which also can mislead the evaluation. Besides, the relation between the SFSs and the values of the midpoints is not defined clearly. However, the SFS based concept can soften the difference between the final scores given by independent evaluators owing to the feature that slightly differing evaluations can result in the same grade. Thus we can summarize that the FEM-GFEM method pair satisfies conditions 2, 3, 5, 6, and 7.

3.2. Chen-and-Lee's methods

The method proposed by Chen and Lee [3] (further on we will refer to it as CL method) has several similar elements to FEM. However, they use a slightly different terminology. The method defines a finer resolution of the scoring scale,

which is in this case the interval [0,1] by using eleven so called satisfaction levels (SL) that are crisp similar to the traditional grade based evaluation. Here one uses an extended grade sheet for the evaluation's documentation, which contains for each question eleven cells that have to be filled in by the evaluator with values between 0 and 1. They describe in what amount the answer given by the student belongs to the predefined satisfaction levels. They can be considered also as membership values. After filling in the eleven cells of the current row a degree of satisfaction $D(Q_i)$ is calculated for the current question Q_i by

$$D(Q_i) = \frac{\sum_{j=1}^{11} y_{ij} \cdot T(SL_j)}{\sum_{j=1}^{11} y_{ij}}, \tag{7}$$

where y_{ij} is the membership value assigned for the j^{th} satisfaction level SL_j, and $T(SL_j)$ is the upper bound of the score interval corresponding to SL_j.

Finally, the total score of the student is calculated as a weighted average of the individual degrees of satisfaction

$$TS = \sum_{i=1}^{n} s_i \cdot D(Q_i), \tag{8}$$

where the weights have to satisfy the equation

$$\sum_{i=1}^{n} s_i = 100. \tag{9}$$

Chen and Lee also published in [3] a generalized version of their method (further on we will refer to it as LCG method). The applied approach is similar to GFEM; it uses the same four criteria for evaluation of each question from different points of view. Thus one calculates four degrees of satisfaction for each question. The overall mark $P(Q_i)$ of the response is calculated as a weighted average of the four degrees of satisfaction

$$P(Q_i) = \frac{\sum_{k=1}^{4} w_k \cdot D(Q_i, k)}{\sum_{k=1}^{4} w_k}, \tag{10}$$

where w_k is the weight of the k^{th} criteria, and $D(Q_i, k)$ is the degree of satisfaction of the k^{th} criteria. CLG determines the total score by substituting $P(Q_i)$ for $D(Q_i)$ in (8).

The CL and CLG methods are in several ways similar to the FEM-GFEM pair. They introduce improvements by a finer resolution of the scoring interval and by allowing the weighting of the four criteria.

The price for this advantage is the increased workload of the evaluator having to fill in more gaps. These modifications also increase the computational need, however, this not a great problem owing to the fact that the methods are applicable in practice only when a software support is ensured. It can be considered as a drawback of the method that it uses the upper endpoint of the satisfaction level intervals as representative value regardless of the evaluator's actual opinion, which can result in same scores for significantly different answerscripts. Thus we can summarize that the method fulfils conditions 2, 3, 5, and 6.

3.3. Wang-and-Chen's methods

Wang and Chen [6] published a new method (further on we will refer to it as WC method) and its generalized version (further on we will refer to it as WCG method) that extend the CL-CLG method pair by introducing the degree of optimism ($\lambda \in [0,1]$) that characterizes the evaluator, and by using type-2 fuzzy numbers for the definition of the membership in each satisfaction level (y_{ij} in (7)). However, the later modification does not have any effect on the final score of the students because the authors suggest the application of isosceles triangle shaped membership functions, and they defuzzify the value before any further utilization. The applied defuzzification method returns the vertex of the triangle as the crisp value.

Thus the modification of the CL-CLG methods consists only in an alternate calculation of the degree of satisfaction $D(Q_i)$ of the question Q_i by

$$D(Q_i) = \frac{\sum_{j=1}^{11} y_{ij} \cdot \left[\lambda \cdot T(SL_j) + (1-\lambda) \cdot L(SL_j)\right]}{\sum_{j=1}^{11} y_{ij}}, \qquad (11)$$

where $L(SL_j)$ is the lower bound of the score interval corresponding to SL_j.

We can summarize that the WC-WCG method pair introduced a slight improvement of the method pair CL-CLG by enabling a way of tuning of the score based on the characterization of the grader. Increasing the level of optimism of the evaluator the total score also increases because in case of each satisfaction level a higher a higher value will be taken into consideration in (11). The result is identical with the one obtained by CL (CLG) only in case of maximum optimism ($\lambda = 1$) of the grader. However, this tool enables only a coarse tuning of the total score; the position of the representative point of the satisfaction levels cannot be modified on per question basis.

Beside the increased computational complexity the major drawback of the WC-WCG method pair is that the authors did not specify how is determined the value of λ. It is supposed to be a self evaluation of the grader but further information is not given. Thus we can summarize that the method fulfils conditions 2, 3, 5, and 6.

3.4. Bai-and-Chen's method

In order to reduce the subjectivism in student evaluation Bai and Chen (further on we will refer to it as BC method) suggested a quite complex solution in [1]. However, their method addresses only a part-task of the evaluation, namely the ranking of the students that obtained the same total score.

The BC method is applied as a follow-up of a conventional scoring technique. First, in case of each student ($S_j, 1 \le j \le n$) each question ($Q_i, 1 \le i \le m$) is evaluated independently by an accuracy rate a_{ij}, where $a_{ij} \in [0,1]$. Then, the evaluator calculates a total score for the student by

$$TS_j = \sum_{i=1}^{m} a_{ij} \cdot g_i \,, \tag{12}$$

where g_i is the maximum achievable score assigned to the question Q_i ($\sum_{i=1}^{m} g_i = 100$).

In order to rank the students having the same total score Bai and Chen propose an adjustment of their scores. The adjustment is based on introduction of new aspects in the evaluation, i.e. the importance and the complexity of the questions, which are based on fuzzy sets determined by the evaluator or by domain experts. The measurement part of the evaluation is also extended by including the time necessary for answering the individual questions divided by the maximum time allowed to solve the question (answer-time rate, $t_{ij} \in [0,1]$). Although it is used only in cases when two or more students achieve the same total score, the answer-time rate has to be measured for each student during the exam because it can not be obtained posteriorly.

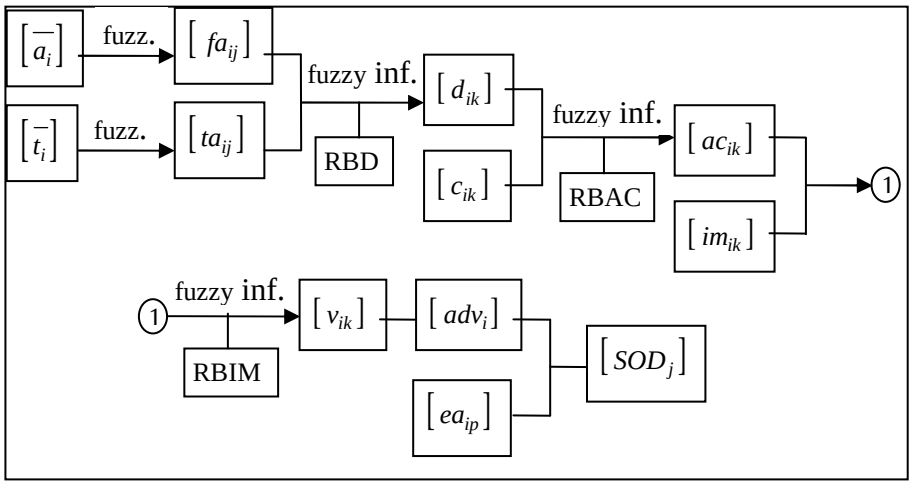

Figure 1. Block diagram of the BC method

The modified scores are determined in six steps applying a three-level fuzzy reasoning process whose block diagram is presented in figure 1. After calculating the average of the accuracy rates ($\overline{a_i}$) and the average of the answer-time rates ($\overline{t_i}$) for each question these are fuzzyfied by calculating their membership values in the corresponding predefined partitions resulting in the fuzzy grade matrices $[\ fa_{ik}]$ and $[\ ft_{ik}]$.

In the second step of the method one determines the fuzzy difficulty ($[\ d_{ik}]$) of each question using a special kind of fuzzy reasoning applying a predefined rule base (*RBD*) and a weighted average of the previously calculated membership values. The third step of the method concentrates on the calculation of the answer-cost of each question (a_{ik}) from the difficulty and the complexity values. The complexity of each question (c_{ik}) is expressed as membership values in the five sets of the predefined complexity partition. The $[\ c_{ik}]$matrix is defined by domain experts. This step uses the same fuzzy inference model as the previous one applying a predefined rule base (*RBAC*).

The fourth step of the method calculates the adjustment values (v_{ik}) of each question from the answer-cost and the importance values. The importance of each question (im_{ik}) is expressed as five membership values in the five sets of the predefined importance partition. The $[\ im_{ik}]$ matrix is defined by domain experts. This step uses the same fuzzy inference model as the previous one applying a predefined rule base (*RBIM*). Next, one calculates the final

adjustment value (adv_i) for each question as a weighted average of the individual adjustment values (v_{ik}) corresponding to the question.

In step 5 a new grade matrix ($\left[ea_{ip} \right]$) is constructed that contains only that k columns of the original accuracy rate matrix, which correspond to the students having the same total score.

The modified score values of each student ($SOD_j, 1 \le j \le n$) are calculated in the last step by

$$SOD_j = \sum_{p=1}^{k} \left[\left[\sum_{\substack{i=1 \\ i \neq j}}^{m} \left(ea_{pj} - ea_{pi} \right) \right] \cdot g_p \cdot \left(0.5 + adv_p \right) \right]. \tag{13}$$

The main advantages of the method are that it does not increase the time needed for the evaluation and it allows the evaluators to make a ranking among students achieving the same score in the traditional scoring system. However, one has to pay a too high prise for this result. In course of the exam preparation two matrices have to be defined by domain experts, one describing the complexity $\left[c_{ik} \right]$ and one describing the importance $\left[im_{ik} \right]$ of each question. It introduces redundancy in the evaluation process because these aspects presumably already have been taken into consideration in course of the definition of the vector $\left[g_i \right]$.

Thus it is hardly avoidable the occurrence of cases when the achievable score of a question is not in accordance with its complexity and importance evaluation. Besides, the level of subjectivity is also increased by the fact that seven weights have to be determined by domain experts as well and there is no formalized way to determine their optimal values. Another drawback of the method is that it does not allow the evaluator to give a fuzzy set as evaluation.

The real novel aspect of the evaluation is the answer-time rate. However, it is not clear how the base time for each question is defined. Besides, it seems not too efficient to measure the answer time for each student for each question and then to use it in case of students having the same total score unless it can be done by software automatically. Thus the BC method is not applicable in case of non computer-based exams. We can summarize that it fulfils conditions 1, 4, 5, and 6.

4. Conclusions

Recently several fuzzy set theory based student evaluation methods have been proposed that aim the reduction of the effects of the subjectivism in the teacher's

judgement and the assistance of the evaluator in expressing the vagueness in his or her decisions.

In the first part of this paper a set of conditions on fuzzy evaluation methods is proposed in order to ease the examination of the methods on this field. The first three methods surveyed share several common features and contain only relative small differences. Despite their advantages usually they increase the time need of the assessment and sometimes they are not beneficial for all students. The last method (BC) provides benefits in both above mentioned fields, which is however compensated by its excessive complexity. Further research plans include the development of fuzzy evaluation methods that apply fuzzy control and iterative learning control .

Acknowledgement

This research was supported by Kecskemét College GAMF Faculty grant no: 1KU16/2008), and the National Scientific Research Fund Grant OTKA K77809.

REFERENCES

[1] Bai, S.M., Chen, S. M.: Evaluating students' learning achievement using fuzzy membership functions and fuzzy rules, Expert Systems with Applications, 34 (2008), pp. 399-410.

[2] Biswas, R.: An application of fuzzy sets in students' evaluation. Fuzzy Sets and System, 74(2), 1995, pp. 187–194.

[3] Chen, S. M., and Lee, C. H.: New methods for students' evaluating using fuzzy sets. Fuzzy Sets and Systems, 104(2), 1999, pp. 209–218.

[4] Precup, R.E.,Preitl, S., Tar, J. K., Tomescu, M. L., Takács, M., Korondi, P., and Baranyi, P.: Fuzzy control system performance enhancement by Iterative Learning Control, IEEE Transactions on Industrial Electronics, vol. 55, no. 9, Sep. 2008., pp. 3461-3475.

[5] Saleh, I., and Kim, S.: A fuzzy system for evaluating students' learning achievement, Expert Systems with Applications, 36 (2009), pp. 6236-6243.

[6] Wang, H.Y., and Chen, S.M.: New methods for Evaluating the Answerscripts of Students Using Fuzzy Sets, Advances in Applied Artificial Intelligence, Lecture Notes in Computer Science, Vol. 4031, 2006, pp. 442-451.

Négy fuzzy hallgató-értékelési módszer vizsgálata

Dr. Johanyák Zsolt Csaba

Összefoglalás

A hallgatók munkájának értékelése nem automatizálható teljes mértékben kifejtős feladatok esetén, így az eredmények jelentős mértékben függhetnek szubjektív emberi döntésektől. Ilyen körülmények között különösen hasznos lehet egy olyan eszköz, ami lehetővé teszi az osztályozó véleményében rejlő bizonytalanság kifejezését.

A cikk egy hét pontból álló követelményrendszer felállítását követően négy olyan módszert mutat be, vizsgál meg és értékel, amelyek különböző fuzzy megközelítést alkalmazva támogatják az értékelő munkáját.

Untersuchung von vier Fuzzy Methoden für Bewertung der Studenten

Dr. Johanyák, Zsolt Csaba

Zusammenfassung

Die Bewertung der schriftlichen Prüfungen kann nicht voll automatisiert werden, und daher die Ergebnisse sind erheblich abhängig von dem menschlichen Urteil. Sogar ein ausgebildeter Sachverständiger begegnet Situationen, wenn er kann nicht eindeutig die Antwort in der einen oder anderen Kategorie ordnen. Unter solchen Umständen könnte es extrem nützlich sein die Unterstützung eines Werkzeugs, das einem Prüfer erlaubt, die Unbestimmtheit in seinem Urteil auszudrücken.

Dieser Artikel definiert Anforderungen an Fuzzy Bewertungsverfahren sowie vorstellt und examiniert vier solche Methoden betonend ihre Vor- und Nachteile.