Fuzzy Logic Schemes of Schools Grouping

Ivan Borshukov, Olga Georgieva
Department of Software Engineering, FMI
Sofia University "St. Kliment Ohridski"
Sofia, Bulgaria
o.georgieva@fmi.uni-sofia.bg

***Abstract*— The paper presents and analyses different fuzzy rule-base schemes for a model that finds groups of schools according to the scores presented by their students. The model reflects the existing uncertainty in the data by applying information fusion concept. The expert knowledge and statistical analysis are useful for definition of the model structure, whereas fuzzy values account for the imprecision in the data. By the discussed schemes having the current scores of a certain school we are able to predict its learning deficit.**

Keywords-Big Data, Decision support, Education, Fuzzy Rule-Based model

I. INTRODUCTION

Nowadays human activities, industrial processes and research lead to data collection with volumes that are growing exponentially. The result of data explosion is apparent in all domains of the daily life with user-generated content of around 2.5 quintillion bytes every day [4]. This imposes data processing on an unprecedented scale that leads to new electronic and industrial products. For its part they impose new software services and finally new business processes [1].

Big Data Value Association [3] surveys show that the gains from data value chain that aims to discover models, correlations, deviations and other facts and events that are hidden in the data is of a large importance for the industry and production, science achievements and society prosperity [2]. However, the value refers to the final product created on the data processing is not a trivial task. The efforts are directed to individual solution, taking into account all available information and existing methods for its description, processing, formalization and logic conclusions. Handling today's highly variable and real-time datasets requires not a standard solution and supposes to look for an application enable to use different knowledge and combine technologies. The need to discover new relations within the data as well as with data of different areas perhaps with different structure and format is a permanent task with increasing importance [5,7].

A specific part of this data explosion is due to the collection and exploitation of people, public authorities, and public registries. The Governance sector disposes large amount of data but at the same time the sector lacks of timely and suitable decision support. These data are good basis for establishment of new ICT services and networks to facilitate access, navigation, searching and reuse of data for citizens and could increase efficiency in the public administrations processes. The exploitation of these data is of big importance for the society as this creates new knowledge and supports Data Driven Government (Public Services based on Open Data). The challenge in their exploitation is a result of the real-time manner of data provision. In addition the data are imperfect due to an objective prerequisite for existing data inaccuracy.

The social Big Data processing technologies, describing social media characteristics of Big Data and current trends in the field are discussed in [6]. The focus is on the significance of Information Fusion (IF) – a method for combining information into a new set of information in order to remove the data uncertainty. The challenges of applying IF to social Big Data are related to data diversity, common referencing and data association, trust/reliability, real-time fusion of streaming data and data access. Several future research directions, including format unification, data sharing, data imperfection, central data union, data integration and security, are identified. The education process could be properly governed if all factors and reflections to the society are known.

The newly developed model of education data [13] accepts the concept of Information Fusion and proposes a specific solution of a schools' grouping based on data for their assessment scores. It groups schools according to the students' learning deficits. The model accounts for different uncertainties in the existing data by applying soft computing approach for model building. The goal of the present paper is to extend this intelligent solution. Modeling the education open data provided by the primary and secondary schools is extended at exploring different fuzzy logic schemes to give guidance for using appropriate scheme that can help for decision support of Governance sector in forming the education policy. For this purpose the models are considered to be a part of the data driven solution for automation and future integration in a common platform for informed and timely decisions on e-Governance big data (ITDGate).

The rest of the paper is organised as follows. Section II describes the information that was accounted for. Section III presents the fuzzy logic schemes as models for schools grouping. Section IV presents a comparative analysis of the

fuzzy schemes. Section V summarizes the paper results and gives directions for the future work.

II. INFORMATION FOR THE MODEL BUILD

Along the existing data certain information about the dependences, concomitant processes of data collection and processing are important factor of choice for the model type and structure. These are considered as restrictions to account for.

For instance, there is imprecision due to unknown dependencies between the key variables that determine schools groups. Some of these variables are provided in the data sets but others have to be revealed additionally. On the other hand, the dependences between the variables are not known. The complexity of the relationship does not allow to use an analytical function for group calculation.

A part of the existing uncertainty is due to the human factor appeared during the process of data collecting. All data are loaded to the data warehouses by people working at the education institutions and the reliability of the data accuracy is on their responsibility.

A restriction requirement is the need to search for a simple calculation schemes that ensures further implementation in the foreseen Data Driven Government platform.

All this argues to imply a rule-base system than a pre specified analytical model function. A soft computing solution where the ambiguity could be easily covered is an advantage. It could be summarized that a fuzzy-rule base with fuzzy formality scheme of calculation could be elaborated as an effective solution of the schools grouping.

A. Data

We develop the schools' score model using open data from national external assessments (NEAs) after IV and VII class, and the state matriculation exams (SMEs) in all schools in Bulgaria. The assessment covers two subjects – Bulgarian language and literature (BLL) and Mathematics (Maths). The data provide also information for the average score, the number of students in the school as well as the number of the students that participated in each exam.

Several procedures have been applied in advance in order to purify the raw data and to obtain a form appropriate for further analysis and interpretation.

- *Data collection* from variety of data sources that are the data of each school;
- *Data filtering* – the data for special schools is removed, since the educational abilities of their students are not assessed. Thus, the elaborated model does not cover the special schools for students who have specific educational needs due to severe learning difficulties, physical disabilities or behavioural problems.
- *Data cleaning* – missing or inconsistent data have been replaced with appropriate values.
- *Data formatting* – the data is formatted according to strong data type format in order to allow automated processing.

B. Expert information

The idea is to assess the schools groups based not only the data but accounting for the existing expert information. The indicators (Table I) that relate to the students results from NEAs and SMEs are already used by the experts in the education sector. They use them to provide a heuristic estimation of the schools' groups. However, this assessment possesses a certain level unreliability.

A problem we have to tackle is the existing large amount of variables and by that having different strength to the formed groups. In order to reduce this amount we can rely on the expert knowledge. We can choose or form the model variables among the indicators used by the experts (Table I). For instance, according to this knowledge the indicators are divided on primary and secondary, which reflects the contribution of the respective indicator to the assessment of the schools' education abilities. Thus, due to the side effect of correlation between the indicators we incorporate in the model description only primary indicators. By reducing the use of secondary indicators we are able to reduce the complexity and to obtain more distinguishable and informative data space.

This knowledge holds a sort of uncertainty that is easily grasped by the elaborated fuzzy logic schemes.

TABLE I. INDICATORS OF SCHOOLS' RATE

No	Title	Type	Metric
1	NEA after IV class (average grate point on BLL)	secondary	number
2	NEA after IV class low grades (BLL)	primary	percent
3	NEA after IV class (average grate point on Maths)	secondary	number
4	NEA after IV class low grades (Maths)	primary	percent
5	NEA after VII class (average grate point on BLL)	secondary	number
6	NEA after VII class low grades (BLL)	primary	percent
7	NEA after VII class (average grate point on Maths)	secondary	number
8	NEA after VII class low grades (Maths)	primary	percent
9	SME (average grate point on BLL)	secondary	number
10	SME low grades (BLL)	primary	percent
11	SME (average grate point on Maths)	secondary	number
12	SME low grades (Maths)	primary	percent

C. Statistical analysis

The knowledge acquired by statistical data analysis is another source of information to take into account. It is helpful source in defining the number and value of the model variables.

A large difficulty in model of schools grouping comes from the fact that a significant part of the schools conduct more than one of the mentioned exams. A comprehensive strategy accounting for all school assessments is need. For this purpose the pre-processing statistical analysis of raw data is summarized. The analysis is given in detail in [13] and here it is briefly presented.

The ratio of the number of examined students in BLL and Maths for both NEA exams is approximately equal as the correlation coefficients are above 0,9. Number of students in SMEs examined by BLL is quite larger than those by Maths due to the fact that Maths is not mandatory exam for SME assessment. However, their correlation coefficient is still positive at a moderate level of 0,539. A significant lack of correlation is observed by analyzing the number of students examined by NEA after IV class and SME (correlation is -0,1213) and at NEA after VII class and SME (correlation is 0,015). The correlation between NEA after IV and VII class is high. These results are due to the specificity of the education system as most of primary schools do not serve for secondary education.

It could be concluded that the data of the three assessments are not exclusive but complimentary. Appropriate description of this dependency is disjunction operation and respective realization in a fuzzy sentence is by a t-conorm operator.

The data for the percentage low grades for each assessment is preferred instead the assessments value itself. The expert knowledge is that these are primary indicators (Table I).

As the correlation coefficients of the percentage low grades for BLL and Maths for each class is positive we could take their average value as a model input variable (Table II). The effect of this operation is reduction of the dimensionality of the data space, which reduces the model complexity and ambiguity.

TABLE II. CORRELATION COEFFICIENTS OF THE PERCENT LOW GRADES OF BLL AND MATH FOR EACH EDUCATION CLASS

% low grades of BLL and Maths	Correlation coefficient
NEA after IV class	0,5652
NEA after VII class	0,6868
SME	0,1065

Data for the average score (AS) of a school for the education year is of a major importance for the education deficit level. The meaning of this input variable compliments the other input variables (the average of the percent of low grades). Thus, their relation is described by a conjunction operator.

As could be seen from the histogram on Figure 1 AS data are between marks 3 and 6. It is reasonable to determine fuzzy values of the variable on the entire diapason of change by accounting for the observed distribution.

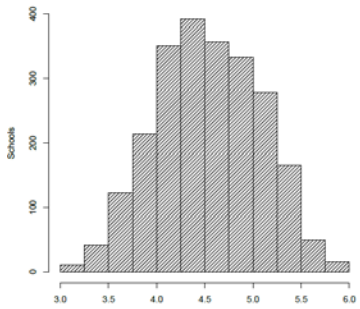


Fig. 1. Distribution of average learning success of the schools

III. FUZZY LOGIC SCHEMES FOR MODEL DESCRIPTION

A. Number of Rules

The interested stakeholder that is government authority in the education sector needs seven schools' groups related to the level of the education deficit. Each group is defined by the students' rate of as follows (Table III). This constrains our model to seven rules each for a predefined group.

TABLE III. SCHOOLS GROUPS

Group	Score
I	highest
II	high
III	significant
IV	average
V	low
VI	lower
VII	minor

B. Fuzzy Variables

Generally, there are two different strategies to partition the input data space in subspaces that guarantees fine and interpretable description. Each subspace determines the area where respective rule fulfills. First, it is grid partition of the input data space [9,12] and second it is clustering of the input data space [10,11]. Our previous investigation on this data base shows that clustering analysis is not very effective [8]. The seven groups are not clearly identified by clustering the space formed by indicators data of Table I. For this reason here we explore the grid partition of the data space as an alternative approach to form the subspace of each rule.

For this each antecedent variable is treated as a linguistic variable with appropriate fuzzy values. According to the considerations given in the previous section the input variables that are incorporated in the model for schools grouping are:

- **NEA4%** is average of the percent of low grades of BLL and Maths of NEA after IV class of a school
- **NEA7%** is average of the percent of low grades of BLL and Maths of NEA after VII class of a school
- **SME%** is average of the percent of low grades of BLL and Maths of SME of a school
- **AS** is average score that shows the average learning success of a school.

The three variables **NEA4%**, **NEA7%** and **SME%** are considered as a linguistic variables with three fuzzy values of their term set T(NEA4%)=T(NEA7%)=T(SME%)={**low**, **medium**, **high**} defined at the universe of discourse U=[0, 100]. The interpretation of the values is as follows:

- value **A1 = low** corresponds to a grade below 10%
- value **A2 = medium** is a grade around 15%
- value **A3 = high** is a grade above 20%.

A trivial form of triangular membership function is accepted for description of the respective fuzzy values. The membership function spreading (Figure 2) is dictated by the fact that for the three variables most of the values are in the range of 0% to 5%. After this range the occurrence frequency decreases drastically and uniformly.

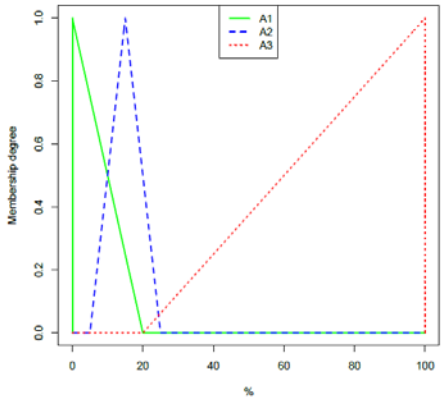


Fig. 2 Fuzzy membership functions of **NEA4%, NEA7%** and **SME%**

The linguistic variable AS has a term set T(AS)={**low**, **medium**, **high**} over universe of discourse U=[3,6]. The values are presented by a triangular membership functions (Figure 3) as:

- value **B1 = low** is assessment close to 3
- value **B2 = medium** is assessment about 4,5
- value **B3 = high** is assessment close to 6.

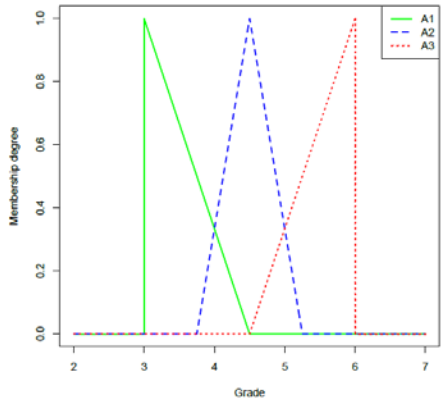


Fig. 3 Fuzzy membership functions of **AS**

The model output (consequent) variable is the linguistic variable **SCORE** defined by a term set of seven values T(**SCORE**)={**minor, lower, low, average, significant, high, highest**} over the universe of discourse U = [0, 100]. The seven fuzzy values interpret the respective education level of a school group (Table III). As we do not have advance information about their values an uniformly spread triangular fuzzy sets is accepted (Figure 4).

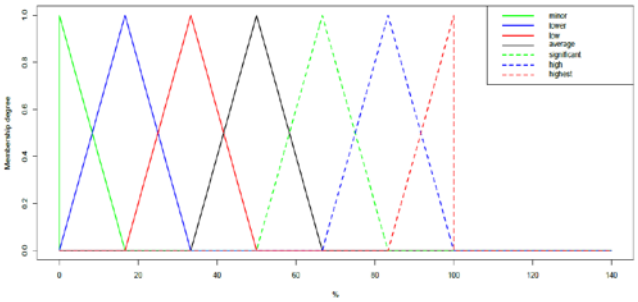


Fig. 4 Fuzzy membership functions of **SCORE**

C. Fuzzy Rule-Based Systems

Complexity of the rule structure is determined by the rules antecedent part where the input variable values are connected both conjunctively and disjunctively. Another source of complexity is the large dimensionality of the rule base. The full number of rules is 81 as we have 4 variables with three values each. However, not all of these rules are activated as the data are not spread over the whole data space. Thus, we can extract only those rules that are meaningful of the expert experience that best fits the seven groups:

Rule Base 1:

Rule 1: IF (**NEA4%** is **low** or **NEA7%** is **low** or **SME%** is **low**) and **AS** is **high** THEN the school **SCORE** is **highest** (Group I).

Rule 2: IF (**NEA4%** is **low** or **NEA 7%** is **low** or **SME%** is **medium**) and **AS** is **high** THEN the school **SCORE** is **high** (**Group II)**.

Rule 3: IF (**NEA4%** is **low** or **NEA7%** is **medium** or **SME%** is **medium**) and **AS** is **medium** THEN the school **SCORE** is **significant (Group III)**.

Rule 4: IF (**NEA4%** is **medium** or **NEA7%** is **medium** or **SME%** is **medium**) and **AS** is **medium** THEN the school **SCORE** is **average (Group IV)**.

Rule 5: IF (**NEA4%** is **medium** or **NEA7%** is **medium** or **SME%** is **high**) and **AS** is **medium** THEN the school **SCORE** is **low** (**Group V**).

Rule 6: IF (**NEA4%** is **medium** or **NEA7%** is **high** or **SME%** is **high**) and **AS** is **low** THEN the school **SCORE** is **lower (Group VI**).

Rule 7: IF (**NEA4%** is **high** or **NEA7%** is **high** or **SME%** is **high**) and **AS** is **low** THEN the school **SCORE** is **minor** (**Group VII**).

The criticism of the **Rule base 1** is its seven rules that do not cover the entire data space. Merging the rules of the full

rule base system is a possible solution however there is no any information which specifies this merge.

The other approach in the rule base forming is to simplify the antecedent part of the rules. In case the school has more than one assessment (NEA4%, NEA7%, SME%), the one from the highest class is taken considering to be the most representative grade for that school. The second rule base system has significantly less rules. The nine rules are easily adapted to describe the seven score groups:

Rule Base 2:

Rule 1: IF average of the percent of low grades of BLL and Maths of the highest class is **low** and **AS** is **high** THEN the school **SCORE** is **highest** (Group I).

Rule 2: IF average of the percent of low grades of BLL and Maths of the highest class is **low** and **AS** is (**medium or low)** THEN the school **SCORE** is **high** (**Group II)**.

Rule 3: IF average of the percent of low grades of BLL and Maths of the highest class is **medium** and **AS** is **high** THEN the school **SCORE** is **significant (Group III)**.

Rule 4: IF average of the percent of low grades of BLL and Maths of the highest class is **medium** and **AS** is **medium** THEN the school **SCORE** is **average (Group IV)**.

Rule 5: IF average of the percent of low grades of BLL and Maths of the highest class is **medium** and **AS** is **low** THEN the school **SCORE** is **low** (**Group V**).

Rule 6: IF average of the percent of low grades of BLL and Maths of the highest class is **high** and **AS** is (**medium or high)** THEN the school **SCORE** is **lower (Group VI**).

Rule 7: IF average of the percent of low grades of BLL and Maths of the highest class is **high** and **AS** is **low** THEN the school **SCORE** is **minor** (**Group VII**).

Disadvantage of this fuzzy scheme is that classification is done by not entire but restricted information of a school.

Different fuzzy operations as disjunction operation ("or"), conjunction operator ("and"), implication for "if-then" could be realized by different calculation formulas of t-conorms, t-norms and implication, respectively [9,12]. These possibilities are further explored to both fuzzy rule bases.

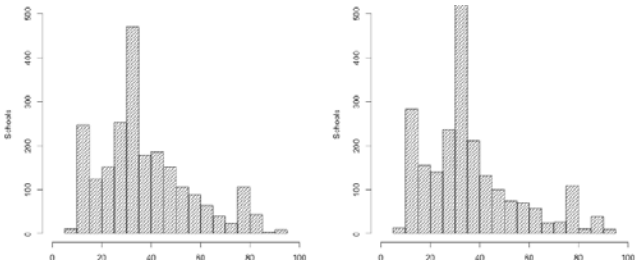
IV. Experimental analysis

Using the fuzzy logic schemes we calculate the grouping of each particular school. For **Rule Base 1** the trivial solution of Mamdani rule base inference mechanism is applied (Table IV, version 1). Disjunction is calculated by fuzzy maximum operator, conjunction – by fuzzy minimum operator and max-min rule of inference for composition rule of inference is implemented. The method of gravity center that is applied for defuzzyfication of the obtained fuzzy output value and to determine the school score is implemented for all fuzzy rule schemes. The school is classified to a group to which the calculated school score has a maximal membership degree. The classification was conducted for the whole data having in total 2334 records. For the system given by **Rule Base 1** that has reduced number of rules, a large part of the schools were successfully classified (Figure 5). Only 71 schools are not classified due to the lack of rule coverage.

TABLE IV DIFFERENT REALIZATION SCHEMES FOR RULE BASE 1

No version	and	or	implication	union	Min score	Max score	Unclassified
1	min	max	min	max	6,574	92,427	71
2	prod	max	prod	max	5,50	92,99	71

In order to deal with unclassified data we have to increase the system rules. For instance, there is a single case of a basic school that presents values: high NEA4%=33%, NEA7%=0%, SME%=0% and medium AS = 4,50, which is not presented by the rule base, but typically it could be determined by Group V.



a) Scheme No1 (Table IV) b) Scheme No2 (Table IV)

Fig. 5 Score distribution of classified schools

The second version of this scheme (Table IV, version 2) uses product operator as the t-norm and implication operation whereas maximum operator is still applied for disjunction and rule union. As could be seen the minimal and maximal score values obtained by version 2 include the corresponding values of the version 1 scheme. Thus, version 2 scheme is more optimistic as it is spread wider over the discourse values.

The **Rule Base 2** covers the whole data space by providing classification of each school.

The validation of the different fuzzy schemes is difficult to be verified as there is no a reference model to validate the grouping. It could be done heuristically by analysis of the grouping of schools that are known by the society and predictable for their learning deficit. For instance, the first school of Table V is a well-known state leading school having excellent students. The model recognizes this fact by classifying it in Group I. The second school is primary one and it is reasonably classified in a group of low deficit.

However, the two rule bases classified it in different groups – Group II and Group III, respectively. It is due to the applied different calculation schemes. The first one accounts for all three different assessments-NEA4%, NEA7% and SME% in order to calculate the school score, whereas the second scheme accounts for the assessment of highest class only. In this case it is NEA7%. By the same reason the rest two schools are classified in different groups. They are typical examples of schools in the middle of the presented learning deficit.

TABLE V. SCHOOLS' GROUPING

School No	NEA4%	NEA7%	SME%	AS	Score & Group by Rule Base 1 (version1)	Score & Group by Rule Base 2 (version1)
1	0	0	0	5,35	94 Group I	94 Group I
2	2,25	13,9	0	5,09	82,15 Group II	70.1 Group III
3	0	87,75	26,1	4,27	54,84 Group IV	16,1 Group VI
4	12,5	35,75	40	4,37	41,19 Group V	16,4 Group VI

The three peaks seen at the distribution of classified schools on Figure 5.a,b clearly show that three significant groups of schools are recognized by both rule bases. First, schools with score around 15 are group with large learning deficit. The second group with a score around 35 score is the largest group. The group of schools of a score around 80 comprises schools with small learning deficit. Comparing with the other groups it is relatively small group.

V. CONCLUSION

The paper presents and analyses different fuzzy rule-base schemes for a model that finds groups of schools according to the scores presented by their students. The model reflects the existing uncertainty in the data by applying information fusion concept. The expert knowledge and statistical analysis are useful for structure definition of the grouping model, whereas fuzzy values account for the imprecision in the data. The implemented fuzzy rule of inference is an easy calculated that enables the intended Data Driven Government platform for educational sector for open data services. In fact we propose a big data solution not only because the large data amount has to be processed but more to the perspective to further enlargement of the data set as the schools' rates and information about the education process are continuously collected. It is suitable for predictive analysis. Having the current scores of a certain school we are able to predict its learning deficit.

ACKNOWLEDGMENT

The authors acknowledge the financial support by the National Scientific Fund under project grant DN 02/11 and Science Fund of Sofia University "St Kl. Ohridski" under grand № 80-10-196/2017.

REFERENCES

[1] Communication from the Commission to the European Parliament, The Council, The European Economic and Social Committee and The Committee of the Regions: Towards a thriving data-driven economy.

[2] IBM, "Bringing big data to the enterprise", http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

[3] http://www.bdva.eu/

[4] https://www.researchgate.net/publication/262211098_Big-Data_Applications_in_the_Government_Sector

[5] M. Pospiech and C. Felden, "Big Data – A State-of-the-Art," Eighteenth Americas Conference on Information Systems, Seattle, Washington, August 9-12, 2012, pp. 1-12.

[6] I. Yaqoob, V. Chang, A. Gani, S. Mokhtar, I. Hashem, E. Ahmed, N. Anuar, and S. Khan, "Information fusion in social big data: Foundations, state-of-the-art, applications, challenges, and future research directions," in International Journal of Information Management, 2016, in press.

[7] A. Fertier at al, Adoption of big data in crisis management toward a better support in decision-making, Intelligent Decision Support in the Networked Society, ISCRAM 2016 Conference, Rio de Janeiro, Brazil, May 2016.

[8] D. Petrova-Antonova, O. Georgieva, S. Ilieva, Modelling of Educational Data Following Big Data Value Chain, CompSysTech 2017, June 2017, Russe, Bulgaria.

[9] G. Klir, B.Yuan, Fuzzy sets and fuzzy logic. Theory and applications, Prentice Hall, 1995.

[10] Babuska R. (1998) Fuzzy Modeling for Control, Kluwer Academic Publishers. Boston/Dordrecht/London

[11] Kruse R., J. Gebhardt, F.Kalwonn (1994) Foundations of Fuzzy Systems, Chichester: John Wiley and Sons.

[12] Yager R., D. Filev (1994) Essentials of Fuzzy Modeling and Control, New York: John Wiley & Sons.

[13] Borshikov I., O. Georgieva, 2017, Soft Computing Modeling of Schools Grouping via Score Data, EECS 2017 (accepted for presentation)