

Fuzzy Hoeffding Decision Trees for Learning Analytics

Gabriella Casalino¹, Pietro Ducange², Michela Fazzolari³ and Riccardo Pecori^{4,5}

¹University of Bari, Department of Computer Science, Via E. Orabona 4, Bari, Italy

²University of Pisa, Department of Information Engineering, Largo Lucio Lazzarino 1, Pisa, Italy

³National Research Council, Institute of Informatics and Telematics (IIT), Via Giuseppe Moruzzi 1, Pisa, Italy

⁴“Universitas Mercatorum” University, Faculty of Economics, Piazza Mattei 10, 00186, Rome, Italy

⁵National Research Council, Institute of Materials for Electronics and Magnetism (IMEM), Parco Area delle Scienze 37/A, 43124, Parma, Italy

Abstract

Virtual Learning Environments (VLEs) are online educational platforms that combine static educational content with interactive tools to support the learning process. Click-based data, reporting the students' interactions with the VLE, are continuously collected, so automated methods able to manage big, non-stationary, and changing data are necessary to extract useful knowledge from them. Moreover, automatic methods able to explain their results are needed, especially in sensitive domains such as the educational one, where users need to understand and trust the process leading to the results. This paper compares two adaptive and interpretable algorithms (Hoeffding Decision Tree and its fuzzy version) for predicting exam failure/success of students. Experiments, conducted on a subset of the Open University Learning Analytics (OULAD) dataset, demonstrate the reliability of the adaptive models in accurately classifying the evolving educational data and the effectiveness of the fuzzy methods in returning interpretable results.

Keywords

Fuzzy Models, Educational Data Streams, Hoeffding Decision Tree, Explainable Artificial Intelligence, Learning Analytics

1. Introduction

Learning Analytics refers to an iterative process aiming at collecting and analyzing educational data in order to generate new knowledge that can be used as feedback for all the involved stakeholders, to improve their tasks [1, 2]. It is an umbrella term covering different applications of statistical methods and analyses in the educational domain, sometimes overlapping also with proper Artificial Intelligence techniques [3]. Some examples are the exploitation of augmented reality insights [4], Internet of Things (IoT) data analysis [5, 6], robotics [7], fog computing [8], video and log processing [9], and information visualization [10], just to mention a few.

Particularly, the use of automatic techniques to analyze sensitive data, such as the educational ones, is gaining attention, since regulation is required. Specifically, automatic analyses must be


OLUD 2022: First Workshop on Online Learning from Uncertain Data Streams, July 18, 2022, Padua, Italy.

✉ gabriella.casalino@uniba.it (G. Casalino); pietro.ducange@unipi.it (P. Ducange); m.fazzolari@iit.cnr.it (M. Fazzolari); riccardo.pecori@unimercatorum.it (R. Pecori)

🆔 0000-0003-0713-2260 (G. Casalino); 0000-0003-4510-1350 (P. Ducange); 0000-0002-8562-6904 (M. Fazzolari); 0000-0002-5948-5845 (R. Pecori)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

explainable and trustworthy [11].

Fuzzy Logic plays an important role in explainability, since it is able to represent uncertain and vague concepts by using natural language. This leads to interpretable or explainable results, which are easier to understand for the domain experts than those returned by black-box algorithms [12, 13]. Indeed, fuzzy logic has been proven to be effective in the educational domain to solve different tasks such as user modeling [14], students' performance or engagement evaluation [15, 16, 17, 18], students' support systems [19], etc.

However, most of the Learning Analytics literature ignores *time*, which, on the contrary, is a critical factor, since the learning process is inherently incremental. An exception is the Deep Knowledge Tracing (DKT) methodology, which models the student's learning behavior from the analysis of previously solved tasks [20], through the use of Recurrent Neural Networks (RNNs) that are able to take into account the time. RNNs are also used to analyze sequential log-based information about the students [21]. However, these are still black-box methods that do not allow one to understand how the results are obtained. A first attempt at using incremental and interpretable methods for analyzing educational data can be found in [22].

In the last years, also thanks to the spreading of the COVID-19 pandemic, distance learning and the usage of Virtual Learning Environments (VLEs) have experienced a steep increase, becoming powerful tools to support higher education throughout the world. VLEs allow also one to continuously collect logs and information, non-stationary by nature, regarding how and when students interact with the educational platform.

In this work, we exploit the evolving nature of students' behaviors on VLEs, by using two stream-based classifiers, namely Hoeffding Decision Tree (HDT) [23] and its fuzzy version (FHDT) [24], to predict the students' outcomes in sequential semesters. Both algorithms lead to interpretable results, since they create incremental decision trees, adapting their structures to the incoming data, thus resulting in incremental sets of *IF-THEN* rules. Moreover, the fuzzy variant results in greater interpretability, given the intrinsic usage of linguistic terms associated with the fuzzy partitions themselves, and it is usually more robust and adaptable against the so-called concept drift, i.e., the evolving change in the distribution of features and labels values along with the continuously incoming instances.

In order to test the aforementioned evolving models, the Open University Dataset has been used, which reports click-stream interactions among students with a VLE.

The rest of the paper is structured as follows. Section 2 briefly details the considered subset of data and the adopted algorithms. Section 3 discusses the obtained results, while conclusions and future developments of our research are depicted in Section 4.

2. Materials and Methods

In this paper, a subset of the Open University Learning Analytics Dataset (OULAD), referring to the academic years 2013 and 2014, has been used¹ [25]. Since the aim of the analysis is to predict the students' outcomes based on the previous information, each semester has been considered as a temporal unit to derive four chronological ordered chunks, i.e., 2013 – I, 2013 – II,

¹Dataset: <https://zenodo.org/record/4264397#.X60DEkJk8E>

2014 – I, 2014 – II. Then, this data stream will be sequentially evaluated through the considered algorithms.

A total of 18 features, grouped into three semantic classes, i.e., demographic information, student performance, and the interactions with the VLE, have been used to describe the behavior of a single student for a given course. Moreover, an additional feature is used for the target class to represent the student's final outcome, which can assume two values: *PASS* and *FAIL*.

Regarding the classification models, we adopted HDTs and FHDTs, whose structure can be updated incrementally while new chunks of labeled semester data become available.

Both considered algorithms are trained by an incremental procedure, made of two main phases: i) the update of the statistics of the classes (binary outcomes of the students) in both the internal nodes and the leaves, and ii) the expansion of the tree if certain conditions on some parameters are fulfilled. The considered parameters are the *grace period*, the *tie threshold*, the *split confidence*, and the *minimum fraction* [26]. The split confidence is involved in computing the *Hoeffding bound*, a heuristic threshold allowing, with high probability, the choice of the attribute for each split as in the case of using an infinite number of instances.

FHDT differs from traditional HDT in the following two aspects [27]: i) the update of the statistics of a given node, ii) the use of the fuzzy Information Gain to choose the best splitting attribute. Concerning the statistics, a training instance in the FHDT can reach more than one node and leaf because of the fuzzy partition defined for each input attribute. The considered partition is strong and uniform, thus exactly two output branches are initialized at each split. The computed statistics at each node are the membership degree, the local fuzzy cardinality of the whole node, and the fuzzy cardinalities per class in a node. As regards the fuzzy Information Gain, the Hoeffding bound has been modified to consider a local fuzzy cardinality instead of the usual sum of the instances in a given leaf. More details on FHDTs, which ensure a good trade-off between their classification performance level, the overall model complexity, and their explainability, in turn, one of the current hot topic in the specialized literature [28, 29], can be found in [24].

3. Results

To evaluate the effectiveness of the proposed approaches, the experiments have been carried out in an incremental way, i.e., using the so-called *Test-the-train* approach. This implies that the stream of data is subdivided into chunks (4 in our case, each corresponding to a semester) and that each chunk is used as a test set in advance and, subsequently, as a part of the training set of the considered tree models.

Moreover, since the cardinality of the fuzzy sets used to describe each fuzzy variable is a critical parameter and could affect the results of FHDTs, two fuzzy models have been experimented, considering two different granularities for the fuzzy partition P_f for each input variable X_f . Particularly, we set the number of fuzzy sets T_f for each input variable to 3 and 5. Indeed, a lower number would have not been sufficient for representing the values of the variables, whilst a higher number would have led to more complex and less interpretable models.

The final voting strategy of both HDTs and FHDTs has been the Adaptive Naive Bayes one.

In order to evaluate the considered models, we focused on the following metrics: the Area

Under the Curve (AUC), and the number of leaves of the derived trees, to assess the classification performance and model complexity of the adopted predictive methods, respectively.

Table 1 reports the comparison of the HDT and the FHDTs models, in terms of classification performance and model complexity, for each semester (test set). We can observe that both HDT and FHDT have low classification performance for the first chunk, thus suggesting that the models are not able to correctly represent the incoming data.

Table 1

Comparison of HDT and FHDT in terms of classification performance and model complexity, for the tested chunks of the considered stream dataset.

Chunks	AUC			No. of Leaves		
	FHDT-3FS	FHDT-5FS	HDT	FHDT-3FS	FHDT-5FS	HDT
2013 – II	0.6336	0.7993	0.8072	17	37	96
2014 – I	0.9040	0.9154	0.8877	17	41	249
2014 – II	0.9028	0.9043	0.8916	19	41	362

However, when the third and the fourth chunks arrive the models are able to adapt their structures, thus leading to high and stable AUC values. For these chunks, models based on fuzzy logic return slightly better results than HDT.

Moreover, the FHDT models need a lower number of leaves if compared with the traditional HDT. This suggests that while the fuzzy models outperform the results given by HDT, they are also able to greatly reduce the complexity, thus resulting in higher interpretability.

Figure 1 shows the model obtained with the FHDT and 3 fuzzy sets per feature, on the training set at the end of the processing (i.e., after the third semester).

Each node represents a fuzzy feature, while the branches stand for the 3 values (low, medium, and high) associated with each fuzzy partition. As previously discussed, the model is compact and thus easy to understand. It can be further explained through IF-THEN rules, leading to the two target classes (PASS and FAIL). From the tree, IF-THEN rules can be easily derived, following the paths from the root to the leaves. To this aim, we consider a zero-order Takagi-Sugeno (TS) fuzzy model [30]. In this case, the antecedent of each rule is expressed through fuzzy sets defining the input variables in the nodes and their values on the branches, while the consequent is expressed through fuzzy singletons corresponding to output classes on the leaves. Formally, the TS fuzzy rules can be defined as:

$$\begin{aligned}
 R_k : & \text{ IF } X_1 \text{ is } A_{1,j_k,1} \text{ AND } \dots \text{ AND } X_{F_k} \text{ is } A_{F_k,j_k,F} \\
 & \text{ THEN } Y = f_k(\mathbf{X})
 \end{aligned} \tag{1}$$

where $j_{k,f} \in [1, T_f]$ identifies the index of the fuzzy set of partition P_f of input variable X_f used in the rule R_k . In the case of the zero-order TS model that we adopted in this work, we consider that the consequent part of each rule can assume only two values, namely PASS or FAIL. Furthermore, in our experiments, we used triangular uniform fuzzy partitions.

Some examples of the final extracted rules are reported in Table 2.

Rule 3, for example, suggests that students with a medium number of intermediate assessments will pass the exam. Instead, if the number of intermediate assessments is high, additional criteria

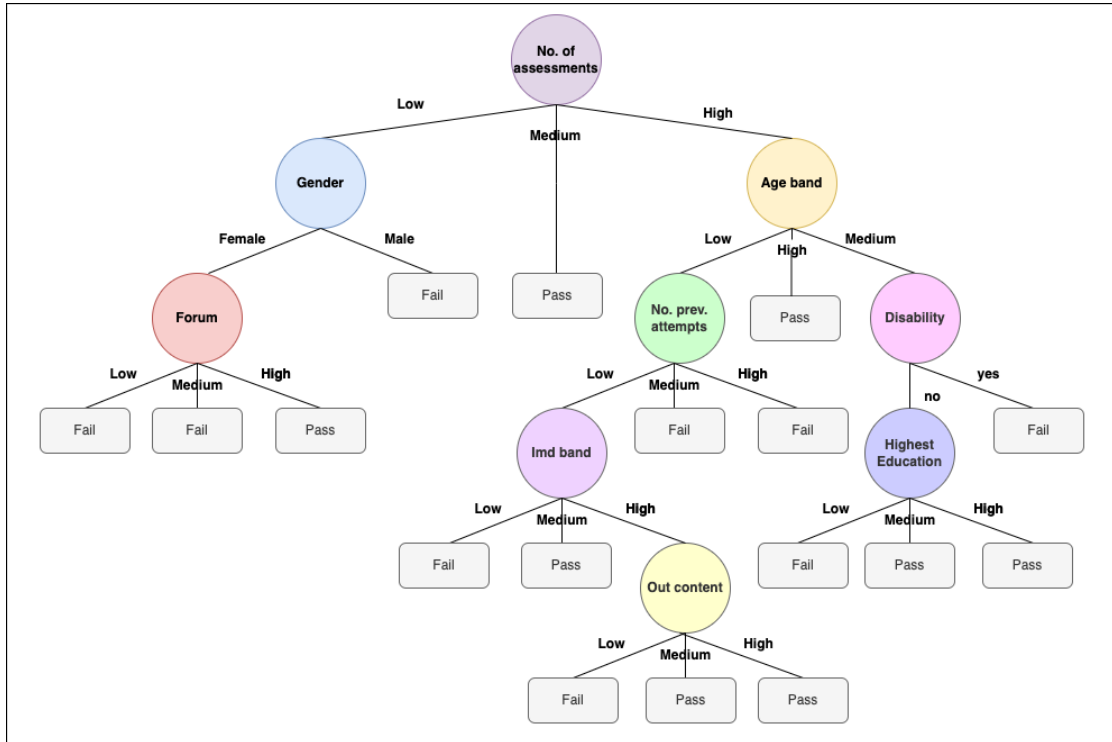


Figure 1: Final FHDT structure obtained at the end of the data stream when considering 3 fuzzy sets.

must be verified (e.g., if the age band is high then the student will pass the exam). Finally, if the number of intermediate assessments is low, then the attribute *Gender* plays a critical role in predicting the failure/success.

In domains such as the educational one, where the final stakeholders are no technicians, models that are easy to understand are preferable since their results are meant to be used as feedback to improve the course design, or the student's learning behavior, for example.

Table 2

Example of final rules extracted from the tree generated by the FHDT algorithm.

- | | |
|---|--|
| 1 | IF (No. of assessment is LOW) AND (Gender is Female) AND (Forum is HIGH) THEN PASS |
| 2 | IF (No. of assessment is LOW) AND (Gender is Male) THEN FAIL |
| 3 | IF (No. of assessment is MEDIUM) THEN PASS |
| 4 | IF (No. of assessment is HIGH) AND (Age band is HIGH) THEN PASS |
| 5 | IF (No. of assessment is HIGH) AND (No. prev. attempts is HIGH) THEN FAIL |

Beyond the results presented so far, a feature importance analysis has been carried out on the HDT model and its fuzzy variants. Figure 2 shows the most relevant features for the classification task, returned by the three algorithms. We can observe that three different subsets of features have been returned, but all the models identify the feature *Number of assessment* as

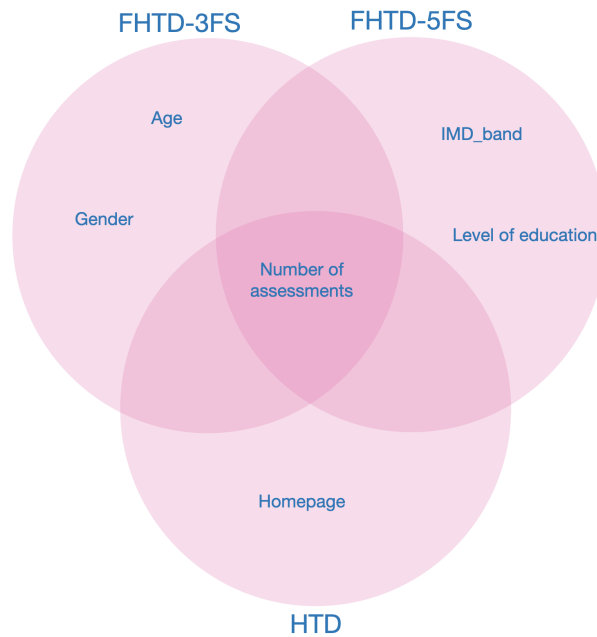


Figure 2: Feature importance analysis of the three models with the given dataset.

one of the most important one. This feature counts the number of intermediate assessments a given student has performed for a given course. It is related to the students' success/failure since a higher number of intermediate assessments suggests a constant study, that is more likely to lead to passing the exams. Also, it is interesting noticing that the two fuzzy models have identified features related to the demographic information as relevant, whilst the crisp model focused on the student's interaction with the VLE.

4. Conclusion

To conclude, in this preliminary work, a student-oriented subset of the Open University dataset has been incrementally analyzed to verify the effectiveness of the Hoeffding Decision Trees, and their fuzzy variants, to correctly predict the students' outcomes in a degree course. To this aim, information related to four semesters has been sequentially analyzed. Results have shown that the fuzzy algorithms are more able to incrementally adapt the structure of the learned model to the new incoming data. Moreover, they have been proven to be more interpretable and thus more suitable for the educational domain.

Finally, a feature importance analysis has been performed to identify the most relevant features for the predictive task. Whilst the tree algorithms do not agree on the set of the most important features, all of them identified the number of intermediate assessments as a discriminant feature.

Further analyses are necessary to better understand the influence of the different categories of features on the students' assessments. Also, a deeper study on the models' interpretability,

and how this characteristic could help in the adoption of automatic techniques in real scenarios, are needed. To this aim, domain experts will be involved in problem definition and analysis.

Acknowledgments

Gabriella Casalino acknowledges funding from the Italian Ministry of Education, University and Research through the European PON project AIM (Attraction and International Mobility), nr. 1852414, activity 2, line 1. Gabriella Casalino is a member of the INdAM GNCS research group. The contribution of Pietro Ducange has been partly funded by the Italian Ministry of University and Research (MIUR), in the framework of the Cross-Lab project (Departments of Excellence). Riccardo Pecori is a member of the INdAM GNCS research group.

References

- [1] W. Chango, J. A. Lara, R. Cerezo, C. Romero, A review on data fusion in multimodal learning analytics and educational data mining, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2022) e1458.
- [2] R. Pecori, V. Suraci, P. Ducange, Efficient computation of key performance indicators in a distance learning university, *Information Discovery and Delivery* 47 (2019) 96–105.
- [3] M. N. Sadiku, S. M. Musa, U. C. Chukwu, *Artificial Intelligence in Education*, iUniverse, 2022.
- [4] M. Farella, M. Arrigo, G. Chiazzese, C. Tosto, L. Seta, D. Taibi, Integrating xAPI in AR applications for Positive Behaviour Intervention and Support, in: *2021 International Conference on Advanced Learning Technologies (ICALT)*, IEEE, 2021, pp. 406–408.
- [5] G. Tripathi, M. A. Ahad, IoT in education: an integration of educator community to promote holistic teaching and learning, in: *Soft Computing in Data Analytics*, Springer, 2019, pp. 675–683.
- [6] M. A. Ahad, G. Tripathi, P. Agarwal, Learning analytics for IoE based educational model using deep learning techniques: architecture, challenges and applications, *Smart Learning Environments* 5 (2018) 1–16.
- [7] D. Schicchi, G. Pilato, A social humanoid robot as a playfellow for vocabulary enhancement, in: *2018 Second IEEE International Conference on Robotic Computing (IRC)*, IEEE, 2018, pp. 205–208.
- [8] R. Pecori, Augmenting Quality of Experience in Distance Learning Using Fog Computing, *IEEE Internet Computing* 23 (2019) 49–58. doi:10.1109/MIC.2019.2936754.
- [9] D. Schicchi, B. Marino, D. Taibi, Exploring Learning Analytics on YouTube: a tool to support students' interactions analysis, in: *International Conference on Computer Systems and Technologies' 21*, 2021, pp. 207–211.
- [10] D. Malandrino, A. Guarino, N. Lettieri, R. Zaccagnino, On the Visualization of Logic: A Diagrammatic Language Based on Spatial, Graphical and Symbolic Notations, in: *2019 23rd International Conference Information Visualisation (IV)*, IEEE, 2019, pp. 7–12.
- [11] H. Khosravi, S. B. Shum, G. Chen, C. Conati, D. Gasevic, J. Kay, S. Knight, R. Martinez-

- Maldonado, S. Sadiq, Y.-S. Tsai, Explainable Artificial Intelligence in education, *Computers and Education: Artificial Intelligence* (2022) 100074.
- [12] J. M. Alonso Moral, C. Castiello, L. Magdalena, C. Mencar, Toward explainable artificial intelligence through fuzzy systems, in: *Explainable Fuzzy Systems*, Springer, 2021, pp. 1–23.
- [13] J. M. Alonso, P. Ducange, R. Pecori, R. Vilas, Building Explanations for Fuzzy Decision Trees with the ExpliClas Software, in: *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2020, pp. 1–8. doi:10.1109/FUZZ48607.2020.9177725.
- [14] S. Ulfa, D. B. Lasfeto, I. Fatawi, Applying Fuzzy Logic to Customize Learning Materials in e-Learning Systems., *Ubiquitous Learning: An International Journal* 14 (2021).
- [15] M. Dhokare, S. Teje, S. Jambukar, V. Wangikar, Evaluation of Academic Performance of Students Using Fuzzy Logic, in: *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, IEEE, 2021, pp. 1–5.
- [16] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, O. Ragos, Fuzzy-based active learning for predicting student academic performance using autoML: a step-wise approach, *Journal of Computing in Higher Education* 33 (2021) 635–667.
- [17] S. K. Nagothu, P. B. Sri, R. Koppolu, Smart Student Participation Assessment Using Fuzzy Logic, *ICoCIST 2021* (2021) 673.
- [18] G. Casalino, G. Castellano, G. Zaza, Neuro-fuzzy systems for learning analytics, in: *International Conference on Intelligent Systems Design and Applications*, Springer, 2022, pp. 1341–1350.
- [19] P. Ardimento, M. L. Bernardi, M. Cimitile, G. De Ruvo, Learning analytics to improve coding abilities: a fuzzy-based process mining approach, in: *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, 2019, pp. 1–7.
- [20] G. Casalino, L. Grilli, P. Limone, D. Santoro, D. Schicchi, Deep learning for knowledge tracing in learning analytics: an overview, in: *First Workshop on Technology Enhanced Learning Environments for Blended Education-The Italian e-Learning Conference 2021*, volume 2817, CEUR-WS, 2021, pp. 1–10.
- [21] X. Song, J. Li, T. Cai, S. Yang, T. Yang, C. Liu, A survey on deep learning based knowledge tracing, *Knowledge-Based Systems* 258 (2022) 110036.
- [22] G. Casalino, G. Castellano, C. Mencar, Incremental and adaptive fuzzy clustering for virtual learning environments data analysis, in: *2019 23rd International Conference Information Visualisation (IV)*, IEEE, 2019, pp. 382–387.
- [23] N. Kourtellis, G. De Francisci Morales, A. Bifet, A. Murdopo, VHT: Vertical hoeffding tree, in: *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 915–922. doi:10.1109/BigData.2016.7840687.
- [24] P. Ducange, F. Marcelloni, R. Pecori, Fuzzy Hoeffding Decision Tree for Data Stream Classification, *International Journal of Computational Intelligence Systems* 14 (2021) 946–964. URL: <https://doi.org/10.2991/ijcis.d.210212.001>. doi:<https://doi.org/10.2991/ijcis.d.210212.001>.
- [25] J. Kuzilek, M. Hlosta, Z. Zdrahal, Open university learning analytics dataset, *Scientific data* 4 (2017) 1–8.
- [26] P. Domingos, G. Hulten, Mining high-speed data streams, in: *Proceedings of the Sixth*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '00, Association for Computing Machinery, New York, NY, USA, 2000, p. 71–80. URL: <https://doi.org/10.1145/347090.347107>. doi:10.1145/347090.347107.

- [27] R. Pecori, P. Ducange, F. Marcelloni, Incremental Learning of Fuzzy Decision Trees for Streaming Data Classification, in: Proceedings of the 11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019), Atlantis Press, 2019, pp. 748–755.
- [28] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information fusion* 58 (2020) 82–115.
- [29] K. Kaczmarek-Majer, G. Casalino, G. Castellano, M. Dominiak, O. Hryniewicz, O. Kamińska, G. Vessio, N. Díaz-Rodríguez, Plenary: Explaining black-box models in natural language through fuzzy linguistic summaries, *Information Sciences* 614 (2022) 374–399.
- [30] T. Takagi, M. Sugeno, Fuzzy identification of systems and its applications to modeling and control, *IEEE transactions on systems, man, and cybernetics* (1985) 116–132.