A Preliminary Many Objective Approach for Extracting Fuzzy Emerging Patterns

Angel Miguel Garcia-Vico, Cristobal J. Carmona(✉), Pedro Gonzalez, and Maria Jose del Jesus

Interuniversity Andalusian Institute on Data Science and Computation Intelligence, University of Jaén, 23071 Jaén, Spain
{agvico,ccarmona,pglez,mjjesus}@ujaen.es

Abstract. A preliminary many objective algorithm for extracting fuzzy emerging patterns is presented in this contribution. The proposed algorithm employs fuzzy logic together with an evolutionary algorithm. The aim is to expand the complex search space that we have in emerging pattern mining.

The experimental study presented in this paper faces this new proposal regarding an ensemble of one of the most used algorithms within supervised descriptive rule discovery. Results presents a set of patterns with a major interpretability and precision for the new proposal which could be interesting for experts in real-world applications.

Keywords: Many objective evolutionary algorithm · Emerging pattern mining · Fuzzy patterns

1 Introduction

Emerging pattern mining (EPM) is a data mining task that tries to find discriminative patterns whose support increases significantly from one class, or dataset, to another. EPM is halfway prediction and description because it describes a problem by discovering some relationships on the data by means of a target variable, typically used in classification. In fact, EPM belongs to the supervised descriptive rule discovery framework [5].

The quality of an emerging pattern (EP) can be determined by a wide range of quality measures [17]. In fact, there is no consensus in the literature about the most relevant quality measures to analyse the goodness of a supervised descriptive rule algorithm, but rather the quality is based on three fundamentally axis: interpretability of the sets of extracted patterns, balance between generality and reliability, and interest of the emerging patterns.

In this contribution, we present a preliminary approach for extracting emerging patterns through a many objective algorithm, the ManyObjective-EFEP algorithm. The proposal is based on soft computing techniques, in particular, it is an evolutionary fuzzy system (EFS) [22], an hybridization of fuzzy logic [28] and evolutionary algorithms [21]. The former allows us the obtaining of fuzzy

Á. Herrero et al. (Eds.): SOCO 2020, AISC 1268, pp. 100–110, 2021.
https://doi.org/10.1007/978-3-030-57802-2_10

emerging patterns which facilitate the analysis and understanding by the experts; the latter is an evolutionary algorithm based on NSGA-III [9] that allows us the use of a wide number of quality measures within the evolutionary search process without degrading its performance.

The paper is organized as follows: Sect. 2 presents the main concepts and properties of the EPM. In Sect. 3 the main characteristics of the EFSs are shown. Section 4 presents the ManyObjective-EFEP algorithm. Section 5 presents the experimental study carried out to determine the quality of the proposed method. Finally, the conclusions extracted from this work are depicted in Sect. 6.

2 Emerging Pattern Mining

EPM was defined such as the search for patterns whose support increase significantly from one dataset (D_1) to another (D_2) [20]. Specifically, D_1 contains examples for one class and D_2 examples for the remaining classes. A pattern is emerging if the growth rate (GR) is greater than a threshold $\rho > 1$ and it is defined as:

$$GR(x) = \begin{cases} 0, & IF\ Supp_{D_1}(x) = Supp_{D_2}(x) = 0, \\ \infty, & IF\ Supp_{D_1}(x) \neq 0 \wedge Supp_{D_2}(x) = 0, \\ \frac{Supp_{D_1}(x)}{Supp_{D_2}(x)}, & another\ case \end{cases} \tag{1}$$

EPs are usually represented by means of conjunctions of attribute-value pairs, or attribute-value pairs in disjunctive normal form (DNF), which represents the discriminative characteristics they want to describe. For the determination of D_1 and D_2, these patterns are usually labeled with the class or the dataset they try to describe. Generally, these patterns can be represented as rules in the following form [5]:

$$P : Cond \rightarrow Class \tag{2}$$

where $Cond$ represents the condition of the pattern and $Class$ is the value of the class.

The analysis of the descriptive behaviour of a pattern is key in EPM. For this purpose, a contingency table is usually calculated. In this contingency table, the number of examples covered or not covered by the patterns which belong or do not belong to the class of the pattern is calculated. An example is shown in Table 1.

By means of this table, several quality measures can be used from the EPM for the determination of a wide range of aspects. The most widely used quality measures in EPM are outlined in Table 2 [20].

Table 1. Contingency table of a pattern.

	Class	No class
Covered	p	n
Not covered	$\overline{p}$	$\overline{n}$
	P	N

Table 2. Quality measures used in EPM for the determination of the quality of a pattern.

Name	Abbreviation	Formula
Number of patterns	nP	–
Number of variables	nV	–
Confidence [13]	Conf	$\frac{p}{p+n}$
Weighted Relative Accuracy [5]	WRAcc	$\frac{p+n}{P+N}\left(\frac{p}{p+n} - \frac{P}{P+N}\right)$
Growth Rate [11]	GR	$\frac{p \cdot N}{P \cdot n}$
True Positive Rate [25]	TPR	$\frac{p}{P}$
False Positive Rate [16]	FPR	$\frac{n}{N}$

3 Evolutionary Fuzzy Systems for Extracting Emerging Patterns

A fuzzy system [28] augmented with a learning process based on evolutionary algorithms [12] is defined as evolutionary fuzzy systems (EFSs) as can be observed in [22]. In this definition two concepts are presented: fuzzy systems and evolutionary algorithms. The former are usually considered in the form of fuzzy-rule based systems (FRBSs), which are composed of "IF-THEN" rules where both the antecedent and consequent can contain fuzzy logic statements. Fuzzy systems are based on fuzzy logic [28], which already allow us to consider uncertainty, and also to represent the continuous variables in a manner which is close to human reasoning. In this way interpretable fuzzy rules consider continuous variables as linguistic ones, where values are represented through fuzzy linguistic labels (LLs) in fuzzy sets [24]. These fuzzy sets facilitate the application to real-world problems because the representation of continuous variables is very close to human reasoning, e.g. a variable such as *Age* could be represented with three linguistic labels such as *Small*, *Normal* and *Tall* making it possible to achieve better analysis.

On the other hand, evolutionary algorithms are stochastic algorithms for optimizing and searching. These algorithms were introduced by Holland [23]. Different computational models can be found within these types of algorithms

such as genetic algorithms [21,23], evolution strategies [27], evolutionary programming [15] and genetic programming [26], amongst others. The evolutionary algorithms imitate the principles of natural evolution to address optimization and learning problems. They are well suited to perform the EPM task due to their ability to reflect the interaction of variables in a rule-learning process also providing great flexibility in the representation [14].

EPM is a supervised descriptive rule discovery task that can be seen as an approximation problem in which the objective is the learning of the parameters of the model. In this task, the search space can be very complex and the search strategy used becomes a key factor. The use of evolutionary fuzzy systems is very well suited to this task because these types of algorithms perform a global search in the space in a suitable way, as can be observed in the real-world problems solved in the literature. For example, in Bioinformatics [1,7], Medicine [4], E-commerce [6] or Industry [2], amongst others.

4 ManyObjective-EFEP: ManyObjective Evolutionary Algorithm for Extracting Fuzzy Emerging Patterns

Throughout the literature, a wide number of quality measures have been presented both to guide the search process in order to find the best EPs and to measure the quality of these patterns, as can be observed in [17,18]. In fact, as we have presented in our previous review [20], the main purpose of an EPM algorithm is to find a good trade-off between generality, reliability and interest. This could lead us to employ a wide number of quality measures in the search process.

The main proposal of the ManyObjective-EFEP algorithm is to extract emerging fuzzy and/or crisp patterns, depending on the type of variables the problem contains, with a good trade-off between reliability and descriptive capacity through the use of a wide number of objectives in the evolutionary process. Specifically, this algorithm is based on the NSGA-III algorithm [9] where the main difference with respect to NSGA-II is that former uses a set of reference points to maintain the diversity of the Pareto points during the search. This results in a very even distribution of Pareto points across the objective space, even when the number of objectives is large.

ManyObjective-EFEP uses a "chromosome = rule" approach where only the antecedent is represented. In this way, an execution for each value of the class is performed in order to extract knowledge for all the classes. The algorithm is able to extract patterns following a DNF representation because it is the best one for the extraction of descriptive EPs [19]. DNF patterns are codified by means of a bit-vector genotype whose length is equal to the total number of features. The number of features is determined by the number of possible categories for nominal variables, while for numeric variables it is the number of LLs used. A fuzzy emerging pattern and its representation can be observed in Fig. 1. Note that the class must be fixed for a value beforehand. Therefore, it is necessary to execute the algorithm for each value of the class.

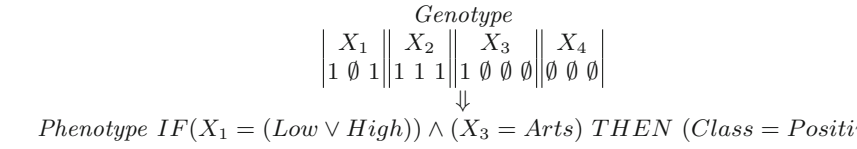

Fig. 1. Representation of a fuzzy DNF pattern with continuous and categorical variables in ManyObjective-EFEP.

In the final stage, the algorithm obtains a set of patterns for each value of the class where the repeated patterns are deleted. The operating scheme of Manyobjective-EFEP algorithms can be seen in Fig. 2.

```
BEGIN
    Create P_0 and reference points
    REPEAT
        Q_t ← ∅
        Generate (Q_t) through genetic operators on P_t
        R_t ← Join(P_t,Q_t)
        Non-dominationed-sort(R_t) based on five objectives
        Associate with reference points
        Apply niche preservation and save in P_{t+1}
        t ← t + 1
    WHILE (num-eval < Max-eval)
    RETURN F_1 without repeated
END
```

Fig. 2. The ManyObjective-EFEP algorithm.

5 Experimental Study

This section presents a summary about the experimental framework in Sect. 5.1, results of the experimental study and a complete analysis of the results are outlined in Sect. 5.2.

5.1 Experimental Framework

The experimental framework used for the evaluation of ManyObjective-EFEP is presented below:

– Algorithms and parameters. The ManyObjective-EFEP algorithm is compared in this paper with an adaptation of the well-known NSGA-II algorithm

Table 3. Algorithms and their parameters used in this experimental study.

Parameters
Population length = 51
Number of labels = 3
Number of evaluations = 10000
Crossover probability = 0.6
Mutation probability = 0.1
Objectives = TPR, FPR, WRAcc, Conf, Strength

[8]. Both algorithms are presented in the jMetal framework[1]. The parameters chosen for both algorithms are identical in order to perform a fair comparison, and they are summarized in Table 3.

- Quality measures in the search process. The main difference between both algorithms is considered with respect to the search process of the evolutionary algorithm. Specifically, for the NSGA-II algorithm we employ an ensemble of algorithms based on the seven possible combinations of the objectives considered in Table 3. In this way, we obtain seven versions of the NSGA-II where all extracted rules for each version are joined and repeated rules are deleted. On the other hand, the ManyObjective-EFEP is executed only once with the five objectives.
- Datasets. The study with datasets from the UCI repository [10] were employed for comparing the quality of the proposed method. They are presented in Table 4. For each data set, it is shown its name and its number of instances, attributes (the number of Real/Integer/Nominal attributes in the data) and classes (number of possible values of the output variable). In addition, the table shows if the corresponding data set has missing values or not (for data sets with missing values the table shows the number of instances without missing values, and the total number of instances between brackets).
- Experiment evaluation. As EPM tries to describe the underlying phenomena in data, an evaluation becomes necessary of the patterns extracted using unseen data. Therefore, this experimental study follows a five-fold stratified cross-validation schema in order to avoid as much as possible bias when creating the training-test partitions.
- Analysis of the quality. The quality measures analyzed in this study were presented in Table 2. These measures are key for the determination of the quality of the patterns extracted regarding the different aspects of EPM. In addition, the number of patterns (nP) and the average number of variables (nV) are analysed in order to determine the model complexity. It is important to remark that the value shown for GR represents the percentage of patterns whose GR in test is greater than one. This is because the domain of GR is $[0, \infty]$, so the average cannot be computed properly.

[1] http://jmetal.github.io/jMetal/.

Table 4. Datasets employed in this experimental study.

Name	♯ Attributes	(R/I/N)	♯ Examples	♯ Classes
appendicitis	7	(7/0/0)	106	2
Australian	14	(3/5/6)	690	2
automobile	25	(15/0/10)	150 (205)	6
bands	19	(13/6/0)	365 (539)	2
breast	9	(0/0/9)	277 (286)	2
car	6	(0/0/6)	1728	4
chess	36	(0/0/36)	3196	2
cleveland	13	(13/0/0)	297 (303)	5
coil2000	85	(0/85/0)	9822	2
contraceptive	9	(0/9/0)	1473	3
crx	15	(3/3/9)	653 (690)	2
dermatology	34	(0/34/0)	358 (366)	6
flare	11	(0/0/11)	1066	6
German	20	(0/7/13)	1000	2
glass	9	(9/0/0)	214	7
heart	13	(1/12/0)	270	2
hepatitis	19	(2/17/0)	80 (155)	2
housevotes	16	(0/0/16)	232 (435)	2
led7digit	7	(7/0/0)	500	10
letter	16	(0/16/0)	20000	26
lymphography	18	(0/3/15)	148	4
magic	10	(10/0/0)	19020	2
mammographic	5	(0/5/0)	830 (961)	2
marketing	13	(0/13/0)	6876 (8993)	9
monk2	6	(0/6/0)	432	2
nursery	8	(0/0/8)	12690	5
pageBlocks	10	(4/6/0)	5472	5
penbased	16	(0/16/0)	10992	10
pima	8	(8/0/0)	768	2
post-operative	8	(0/0/8)	87 (90)	3
ring	20	(20/0/0)	7400	2
saheart	9	(5/3/1)	462	2
satimage	36	(0/36/0)	6435	7
segment	19	(19/0/0)	2310	7
shuttle	9	(0/9/0)	58000	7
thyroid	21	(6/15/0)	7200	3
tictactoe	9	(0/0/9)	958	2
twonorm	20	(20/0/0)	7400	2
vehicle	18	(0/18/0)	846	4
vowel	13	(10/3/0)	990	11
wine	13	(13/0/0)	178	3
winequalityRed	11	(11/0/0)	1599	11
winequalityWhite	11	(11/0/0)	489	8
wisconsin	9	(0/9/0)	683 (699)	2
yeast	8	(8/0/0)	1484	10
zoo	16	(0/0/16)	101	7

5.2 Analysis of the Results Obtained

Due to the extension of the results obtained in this experimental study, the complete results are presented in a website[2]. In addition, the average results of the study are presented in Table 5.

Table 5. Average results extracted from the NSGA-II ensemble and ManyObjective-EFEP methods.

Algorithm	nP	nV	$WRACC$	$CONF$	GR	TPR	FPR
NSGA-II *Ensemble*	157.74	9.65	**0.538**	0.286	**0.426**	**0.270**	0.116
ManyObjective-EFEP	**45.65**	**9.91**	0.523	**0.380**	**0.426**	0.091	**0.015**

The results are analysed based on the three important axis for the supervised descriptive rule discovery tasks [3]:

- *Interpretability*: The ensemble of the different versions of NSGA-II algorithm obtains an elevated number of patterns, three times upper than the algorithm presented in this contribution. Throughout the literature, we are aware about the complexity to incorporate more than three objectives within the evolutionary process because the number of patterns grows very high. In this way, the new approximation keeps a number of pattern more reduced that is more relevant within supervised descriptive rule discovery. On the other hand, there is no difference in the number of variables of the patterns extracted with values very similar. However, it is important to note a high complexity in the knowledge extracted where expert would need to analyse results with a high number of rules and variables which would complicate the understanding of the problem.
- *Tradeoff between generality and reliability*: The generality is measured through the TPR where the percentage of examples covered for the class are calculated. In this way, the algorithm NSGA-II *Ensemble* obtains a value more interesting with a value three times upper than the ManyObjective-EFEP algorithm. However, the reliability of the patterns extracted is far below. In fact, the ratio between TPR and FPR (false positive rate) in the ManyObjective-EFEP algorithm is about six times upper so more precise patterns are extracted for this algorithm. Therefore, the values of confidence are higher in this one. On the other hand, the value in the GR is similar in both algorithms, i.e., the percentage of fuzzy emerging patterns is similar in both algorithms.
- *Interest*: This concept within supervised descriptive rule discovery is calculated through the $WRACC$ quality measure which is key as can be observed in [5]. The interest values obtained by the NSGA-II *Ensemble* algorithm in

[2] https://simidat.ujaen.es/papers/ManyObjectiveEFEP/.

this experimental study are very close to those obtained by ManyObjective-EFEP. This value is determined by the coverage of the rule that, as we have seen previously, is superior in the first algorithm.

6 Conclusions

This contribution presents a first approximation of a many objective algorithm for extracting fuzzy emerging patterns. The ManyObjective-EFEP algorithm combines soft-computing techniques such as fuzzy logic and the NSGA-III evolutionary algorithm. The complexity of the search process with the inclusion of a wide number of objectives in the evolutionary process is analysed in this study, where good results in reliability with interest are obtained but with a low values in generality. However, it is interesting to see how the number of patterns is reduced with respect to an ensemble approach.

As future work, we will study and continue with the analysis of the use of many objective evolutionary algorithms for EPM, because it is a complex space, and the tradeoff among a wide number of quality measures is desired.

Acknowledgement. This study was funded by the FPI 2016 Scholarship reference BES-2016-077738 (FEDER Founds).

References

1. Carmona, C.J., Chrysostomou, C., Seker, H., del Jesus, M.J.: Fuzzy rules for describing subgroups from influenza a virus using a multi-objective evolutionary algorithm. Appl. Soft Comput. **13**(8), 3439–3448 (2013)
2. Carmona, C.J., González, P., García-Domingo, B., del Jesus, M.J., Aguilera, J.: MEFES: an evolutionary proposal for the detection of exceptions in subgroup discovery. An application to concentrating photovoltaic technology. Knowl.-Based Syst. **54**, 73–85 (2013)
3. Carmona, C.J., González, P., del Jesus, M.J., Herrera, F.: Overview on evolutionary subgroup discovery: analysis of the suitability and potential of the search performed by evolutionary algorithms. WIREs Data Min. Knowl. Disc. **4**(2), 87–103 (2014)
4. Carmona, C.J., González, P., del Jesus, M.J., Navío, M., Jiménez, L.: Evolutionary fuzzy rule extraction for subgroup discovery in a psychiatric emergency department. Soft Comput. **15**(12), 2435–2448 (2011)
5. Carmona, C.J., del Jesus, M.J., Herrera, F.: A unifying analysis for the supervised descriptive rule discovery via the weighted relative accuracy. Knowl.-Based Syst. **139**, 89–100 (2018)
6. Carmona, C.J., Ramírez-Gallego, S., Torres, F., Bernal, E., del Jesus, M.J., García, S.: Web usage mining to improve the design of an e-commerce website: OrOliveSur.com. Expert Syst. Appl. **39**, 11243–11249 (2012)
7. Carmona, C.J., Ruiz-Rodado, V., del Jesus, M.J., Weber, A., Grootveld, M., González, P., Elizondo, D.: A fuzzy genetic programming-based algorithm for subgroup discovery and the application to one problem of pathogenesis of acute sore throat conditions in humans. Inf. Sci. **298**, 180–197 (2015)

8. Deb, K.: Multi-Objective Optimization using Evolutionary Algorithms. John Willey & Sons, Hoboken (2001)
9. Deb, K., Jain, H.: An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: solving problems with box constraints. IEEE Trans. Evol. Comput. **18**(4), 577–601 (2014)
10. Dheeru, D., Karra Taniskidou, E.: UCI machine learning repository (2017). http://archive.ics.uci.edu/ml
11. Dong, G., Li, J.: Efficient mining of emerging patterns: discovering trends and differences. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, pp. 43–52. ACM (1999)
12. Eiben, A.E., Smith, J.E.: Introduction to Evolutionary Computation. Springer, Berlin (2003)
13. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: an overview. In: Advances in knowledge discovery and data mining, AAAI/MIT Press, Menlo Park, CA, USA, pp. 1–34 (1996)
14. Fernández, A., García, S., Luengo, J., Bernadó-Mansilla, E., Herrera, F.: Genetics-based machine learning for rule induction: state of the art, taxonomy, and comparative study. IEEE Trans. Evol. Comput. **14**(6), 913–941 (2010)
15. Fogel, D.B.: Evolutionary Computation - Toward a New Philosophy of Machine Intelligence. IEEE Press, New York (1995)
16. Gamberger, D., Lavrac, N.: Expert-guided subgroup discovery: methodology and application. J. Artif. Intell. Res. **17**, 501–527 (2002)
17. García-Borroto, M., Loyola-Gonzalez, O., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A.: Comparing Quality Measures for Contrast Pattern Classifiers, pp. 311–318. Springer, Berlin Heidelberg (2013)
18. García-Borroto, M., Loyola-González, O., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A.: Evaluation of quality measures for contrast patterns by using unseen objects. Expert Syst. Appl. **83**, 104–113 (2017)
19. García-Vico, A.M., Carmona, C.J., González, P., del Jesus, M.J.: MOEA-EFEP: multi-objective evolutionary algorithm for extracting fuzzy emerging patterns. IEEE Trans. Fuzzy Syst. **26**(5), 2861–2872 (2018)
20. García-Vico, A.M., Carmona, C.J., Martín, D., García-Borroto, M., del Jesus, M.J.: An overview of emerging pattern mining in supervised descriptive rule discovery: taxonomy, empirical study, trends and prospects. WIREs: Data Min. Knowl. Disc. **8**(1), e1231 (2018)
21. Goldberg, D.E.: Genetic Algorithms in search, optimization and machine learning. Addison-Wesley Longman Publishing Co., Inc. (1989)
22. Herrera, F.: Genetic fuzzy systems: taxomony, current research trends and prospects. Evol. Intell. **1**, 27–46 (2008)
23. Holland, J.H.: Adaptation in Natural and Artificial Systems, 2nd edn. University of Michigan Press, Ann Arbor (1975)
24. Hüllermeier, E.: Fuzzy sets in machine learning and data mining. Appl. Soft Comput. **11**(2), 1493–1505 (2011)
25. Kloesgen, W.: Explora: a multipattern and multistrategy discovery assistant. Advances in Knowledge Discovery and Data Mining, pp. 249–271. American Association for Artificial Intelligence, Menlo Park, CA, USA (1996)
26. Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge (1992)

27. Schwefel, H.P.: Evolution and Optimum Seeking. Sixth-generation Computer Technology Series, Wiley (1995)
28. Zadeh, L.A.: The concept of a linguistic variable and its applications to approximate reasoning. Parts I, II, III. Inf. Sci. **8-9**, 43–80, 199–249, 301–357 (1975)