Applying Subgroup Discovery Based on Evolutionary Fuzzy Systems for Web Usage Mining in E-Commerce: A Case Study on OrOliveSur.com

C. J. Carmona, M. J. del Jesus and S. García

Abstract In data mining, the process of data obtained from users history databases is called Web usage mining. The main benefits lie in the improvement of the design of Web applications for the final user. This paper presents the application of subgroup discovery (SD) algorithms based on evolutionary fuzzy systems (EFSs) to the data obtained in an e-commerce Web site of extra virgin olive oil sale called http://www.orolivesur.com. For this purpose, a brief description of the SD process (objectives, properties, quality measures) and EFSs is presented. A discussion about the results obtained will also be included, especially focusing on the interests of the designer team of the Web site, providing some guidelines for improving several aspects such as usability and user satisfaction.

Keywords Evolutionary fuzzy system · Subgroup discovery · Web usage mining · www.OrOliveSur.com

1 Introduction

Application of data mining techniques in order to extract knowledge in a Web site automatically was considered by Etzioni [1] as Web mining which was classified in three domains with respect to the nature of data [2]: Web content mining, Web structure data, and Web usage mining.

C. J. Carmona (✉) · M. J. d. Jesus · S. García
Department of Computer Science, Building A3, University of Jaen, 23071 Jaen, Spain
e-mail: ccarmona@ujaen.es

M. J. d. Jesus
e-mail: mjjesus@ujaen.es

S. García
e-mail: sglopez@ujaen.es

F. Sun et al. (eds.), *Foundations and Applications of Intelligent Systems*,
Advances in Intelligent Systems and Computing 213,
DOI: 10.1007/978-3-642-37829-4_50,

This paper is focused on the extraction of useful information from Web usage data acquired using Google Analytics tookit in the Web site http://www.orolivesur.com, i.e., Web usage mining applied in an e-commerce Web site. The extraction process is performed through subgroup discovery (SD) algorithms, in particular, algorithms based on evolutionary fuzzy systems (EFSs): SDIGA, MESDIF, and NMEEF-SD.

SD [3, 4] is a broadly applicable data mining technique aimed at discovering interesting relationships between different objects in a set with respect to a specific property which is of interest to the user the target variable. The patterns extracted are normally represented in the form of rules and called subgroups.

In a previous work, [5] were analyzed concepts concerning to the access properties of the users as browser or source in http://OrOliveSur.com in the year 2010. However, in this paper, we perform an analysis related to search engine access with respect to three values: olive oil, organic, and brand, in the year 2012 from January to June. The main objective is to obtain information related to the time and pages visited by users depending on the access keyword.

Structure of this paper is organized as follows: Sect. 2 presents the SD data mining technique: definition, properties, quality measures, and algorithms used in this study, Sect. 3 presents the main information about the e-commerce Web site in which is based on this paper "http://www.OrOliveSur.com," in Sect. 4, the complete experimental study is presented, and finally, Sect. 5 presents concluding remarks about this study to the experts.

2 Subgroup Discovery

The main objective of SD task is to extract descriptive knowledge from the data concerning a property of interest [6, 7], where the main aim of these tasks is to understand the underlying phenomena with respect to an objective class and not to classify new instances.

In the following subsections, the properties of the SD task, quality measures, and algorithms used in this paper are depicted.

2.1 Properties

The concept of SD was initially introduced by Kloesgen [3] and Wrobel [4], and it can be defined as [8]:

> In subgroup discovery, we assume we are given a so-called population of individuals (objects, customer, . . .) and a property of those individuals we are interested in. The task of subgroup discovery is then to discover the subgroups of the population that are statistically "most interesting", i.e., are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest.

The main purpose of SD is based on the search of relations between different properties or variables with respect to a target variable, where it is not necessary to obtain complete but partial relations. Therefore, this technique uses the descriptive induction through supervised learning.

Relations are described in the form of individual rules. Then, a rule (R), which consists of an induced subgroup description, can be formally defined as [9]:

$$R : \text{Cond} \rightarrow \text{Target}_{\text{value}}$$

where $\text{Target}_{\text{value}}$ is a value for the variable of interest (target variable) for the SD task, and Cond is commonly a conjunction of features (attribute-value pairs) which is able to describe an unusual statistical distribution with respect to the $\text{Target}_{\text{value}}$.

Main elements for a SD approach are as follows: type of the target variable, description language, search engine, and quality measures. The study and configuration of these elements are very important in order to develop a new approach for SD task. In the following subsection, quality measures employed in this experimental study are shown.

2.2 Quality Measures

This element is key for extraction of knowledge because quality measures guide search process and allow to quantify quality of extracted knowledge. In addition, they show to the experts the quality of subgroups obtained. Throughout the specialized bibliography have been presented a wide number of quality measures [3, 7, 9–11]. Below, quality measures employed in this experimental study are presented:

- *Number of rules* (n_r): This measures the number of induced rules.
- *Number of variables* (n_v): This quality measure obtains the average of variables in the antecedent. The number of variables of the antecedent for a set of rules is computed as the average of the variables for each rule of that set.
- *Significance*: This measure indicates the significance of a finding, if measured by the likelihood ratio of a rule [3].

$$\text{Sign}(R) = 2 \cdot \sum_{k=1}^{n_c} n(\text{Target}_{\text{value}k} \cdot \text{Cond}) \cdot \log \frac{n(\text{Target}_{\text{value}k} \cdot \text{Cond})}{n(\text{Target}_{\text{value}k}) \cdot p(\text{Cond})} \quad (1)$$

where n_c is the number of values of the target variable, $n(\text{Target}_{\text{value}} \cdot \text{Cond})$ is the number of examples which satisfy the conditions and also belong to the value for the target variable, $n(\text{Target}_{\text{value}})$ is the number of examples for the target variable, and $p(\text{Cond}) = \frac{n(\text{Cond})}{n_s}$ is used as a normalized factor, n_s is the number of examples, $n(\text{Cond})$ is the number of examples which satisfy the conditions. It must be noted that although each rule is for a specific $\text{Target}_{\text{value}}$,

the significance measures the novelty in the distribution impartially, for all the values.

- *Unusualness*: This measure is defined as the weighted relative accuracy of a rule [12]. It can be computed as:

$$\text{Unus}(R) = \frac{n(\text{Cond})}{n_s}\left(\frac{n(\text{Target}_{\text{value}} \cdot \text{Cond})}{n(\text{Cond})} - \frac{n(\text{Target}_{\text{value}})}{n_s}\right) \tag{2}$$

The unusualness of a rule can be described as the balance between the coverage of the rule $p(\text{Cond}_i)$ and its accuracy gain $p(\text{Target}_{\text{value}} \cdot \text{Cond}) - p(\text{Target}_{\text{value}})$.

- *Sensitivity*: This measure is the proportion of actual matches that have been classified correctly [3]. It can be computed as:

$$\text{Sens}(R) = \frac{\text{TP}}{\text{Pos}} = \frac{n(\text{Target}_{\text{value}} \cdot \text{Cond})}{n(\text{Target}_{\text{value}})} \tag{3}$$

where *Pos* are all the examples of the target variable $(n(\text{Target}_{\text{value}}))$. This quality measure was used in [13] as *Support based on the examples of the class* and used to evaluate the quality of the subgroups in the receiver operating characteristic (ROC) space. Sensitivity combines precision and generality related to the target variable.

- *Confidence*: It measures the relative frequency of examples satisfying the complete rule among those satisfying only the antecedent. This can be computed with different expressions, e.g., [14]:

$$\text{Conf}(R) = \frac{n(\text{Target}_{\text{value}} \cdot \text{Cond})}{n(\text{Cond})} \tag{4}$$

In this paper, we use fuzzy confidence [13]. It is an expression adapted for fuzzy rules which are generated by algorithms used in this experimental study.

2.3 Evolutionary Fuzzy Systems

A EFS is basically a fuzzy system augmented by a learning process based on evolutionary computation, which includes genetic algorithms, genetic programming, and evolutionary strategies, among other evolutionary algorithms [15]. Fuzzy systems are one of the most important areas for the application of the fuzzy set theory [16]. Usually, this kind of systems considers a model structure in the form of fuzzy rules. They are called fuzzy rule-based systems (FRBSs), which have demonstrated their ability with respect to different problems like control problems, modeling, classification, or data mining in a large number of applications. FRBSs provide us a comprehensible representation of the extracted knowledge and moreover a suitable tool for processing the continuous variables.

Three algorithms based on EFSs for SD task have been presented:

- SDIGA is an evolutionary model for the extraction of fuzzy rules for SD [13]. The use of a knowledge representation based on fuzzy logic and the use of evolutionary computation as a learning process receive the name of evolutionary fuzzy system [17]. This type of systems for SD task has been applied in different real-world applications like [18–20].
- MESDIF is a multi-objective evolutionary algorithm for the extraction of fuzzy rules which describe subgroups [21]. The algorithm extracts a variable number of different rules expressing information on a single value of the target variable. The multi-objective evolutionary algorithm is based on the SPEA2 approach and so applies the concepts of elitism in the rule selection (using a secondary or elite population) and the search for optimal solutions in the Pareto front. In order to preserve the diversity at a phenotypic level, the algorithm uses a niches technique which considers the proximity in values of the objectives and an additional objective based on novelty to promote rules which give information on examples not described by other rules of the population.
- NMEEF-SD [22] provides from Non-dominated Multi-objective Evolutionary algorithm for extracting fuzzy rules in SD. This is an evolutionary fuzzy system whose objective is to extract descriptive fuzzy and/or crisp rules for the SD task, depending on the type of variables present in the problem. NMEEF-SD has a multi-objective approach based on NSGA-II, which is a computationally fast MOEA based on a non-dominated sorting approach, and on the use of elitism. The proposed algorithm is oriented toward SD and uses specific operators to promote the extraction of simple, interpretable, and high quality SD rules. The proposal permits a number of quality measures to be used both for the selection and for the evaluation of rules within the evolutionary process.

3 OrOliveSur.com an E-commerce Website Related to Organic Extra Virgin Olive Oil

OrOliveSur is a project born in the province of Jaén from Andalusia (Spain) in 2010. The main purpose is to announce to the world the treasure of its land, the extra virgin olive oil. This Web site is focused in the olive oil produced in a particular territory of Jaén: the Sierra Mágina Natural Park. Sierra Mágina is a protected area of 50,000 acres of natural park, made up of forested slopes, concealed valleys, and rugged mountain peaks. The highest peak, the Mágina Mountain is the highest in the Jaén province, standing at 2,167 m.

OrOliveSur's catalog presents a wide number of extra virgin olive oils focused on the Picual variety. This is the most extended olive grove variety at the world. In Spain, it represents 50 % of production. Most of it is to be found in Andalusia, especially in the province of Jaén. Its olive is large sized and elongated in shape,

with a peak at the end. The trees of this variety are of an intense silvery color, open, and structured.

Along 2 years, OrOliveSur has received both national and international orders from European Union countries (Spain, Denmark, Germany, Great Britain, France, etc.), and its visits and orders are increased every day. Moreover, the OrOliveSur

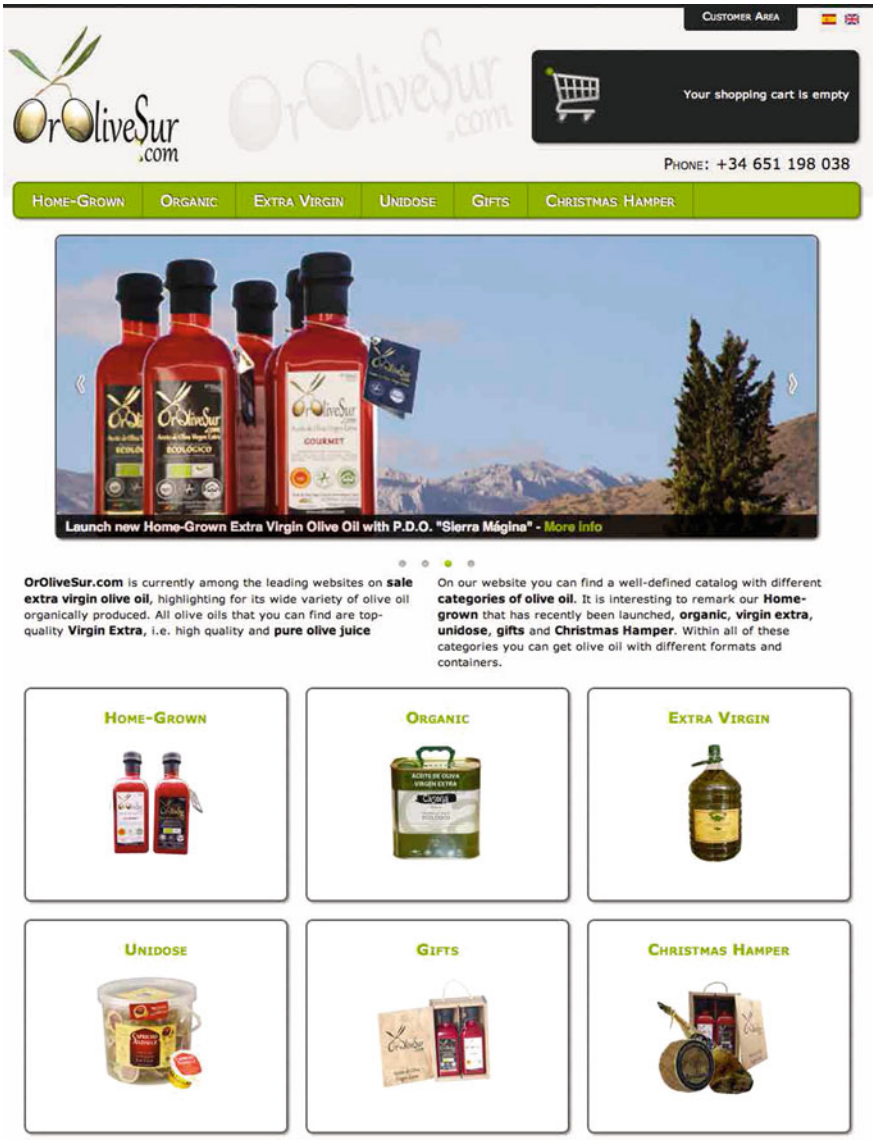

Fig. 1 Homepage from the e-commerce Web site http://OrOliveSur.com

Web site gives direct sales and clients can pay by transfer bank, PayPal, or credit card. In Fig. 1, the homepage of OrOliveSur is shown.

4 Experimental Study: Web Usage Mining Applied to OrOliveSur.com

The experimental study presented in this paper is focused on Web usage mining which was defined by Srivastava [23] as:

> *The process of applying data mining techniques to the discovery of usage patterns from Web data.*

Patterns are represented as a collection of pages or items visited by users. These patterns can be employed to understand the main features of the visitants behaviors in order to improve the structure of one Web site and create personal or dynamic recommendations about content of the web. The main purpose is to analyze the interaction of the users in the e-commerce Web site of http://OrOliveSur.com through Web usage mining techniques. With results obtained, recommendations about the design or the access of the users could be given in order to improve the e-commerce Web site.

Next, experimental framework is presented in Sect. 4.1, and results obtained are shown in Sect. 4.2.

4.1 Experimental Framework

Database has been obtained with the Webmaster tool *Google Analytics* from the period January 1 to June 30, 2012. Moreover, several filters have been applied in data set in order to obtain only instances where bounce rate is lower than 100.00 %. This value is the percentage of single-page visits or visits in which the person left your site from the landing page, i.e., we only obtain visits where users have been visited the Web site more than one seconds. In total, the data set is composed by 2.340 instances.

Variables analyzed in this experimental study are described below:

- Keyword: It is the keyword access to the Web site by the user. The complete keyword set has been classified in three categories. It is important to remark that keywords of the original data set can be found in different languages, and they are classified in a general category with English terms: olive oil, organic, brand. This variable is used as target variable in the experimental study.
  - Olive oil. This value contains all the generic keywords related to the olive oil like *buy olive oil*, *venta de aceite*, *organic olive oil*, *aceite de oliva virgen extra*, *huile d'olive*, *Oliven* öl *Extra Vergine*, and so on.

  - Brand. This keyword contains the entries related to any brand of the catalog as *La Casona*, *Verde Salud*, *Gamez-Piñar* or *OrOliveSur*, for example.
  - Organic. This keyword contains all generic keywords related to organic olive oil like *organic olive oil*, *buy organic olive oil*, and so on.

- New visitor (NV): It indicates if the user is a new or returning visitor.
- Page views (PV): It indicates the page views for users with the same browser, visitor type, keyword, and source.
- Unique page views (UPV): It presents the number of unique page views by users with the same browser, visitor type, keyword, and source.
- Time on site (TS): This feature indicates the time spent on Web site by users with the same browser, visitor type, keyword, and source.
- Average time on page (ATP): It shows the average time used by the user per page view.

In Table 1 are described parameters used by EFSs for SD described previously.

4.2 Results Obtained

In this section are presented results obtained with respect to the data set http://OrOliveSur.com. Table 2 shows the most important rules obtained in this experimental study. In addition, values of quality measures are presented in order to understand the quality of these rules.

In the experimental study have been obtained 40 rules in total for all algorithms. However, in this section have presented the most interesting rules for each algorithm executed. In summary, two rules for SDIGA algorithm (one for keyword → brand and keyword → olive oil), three for MESDIF algorithm (one for each value of the target variable) and two rules for NMEEF-SD (one for keyword → brand and keyword → olive oil).

Table 1 Parameters of algorithms employed

Algorithm	Parameters
SDIGA	Population size = 100, evalutions = 10,000, crossover probability = 0.60, mutation probability = 0.01, minimum confidence = 0.4, representation of the rule = canonical, linguistic labels = 3, objective1 = sensitivity, objective2 = unusualness, objective3 = fuzzy confidence, weight for objective1 = 0.4, weight for objective2 = 0.3 and weight for objective3 = 0.3
MESDIF	Population size = 100, evalutions = 10,000, elite population = 3 individuals, crossover probability = 0.60, mutation probability = 0.01, representation of the rule = canonical, linguistic labels = 3, objective1 = sensitivity and objective2 = unusualness
NMEEF-SD	Population size = 50, evalutions = 10,000, crossover probability = 0.60, mutation probability = 0.1, minimum confidence = 0.4, representation of the rule = canonical, linguistic labels = 9, objective1 = sensitivity and objective2 = unusualness

Table 2 Rules and results obtained by EFSs in this case of study

Algorithm	♯	Rule	Significance	Unusualness	Sensitivity	Fuzzy confidence
SDIGA	R1	IF TS = Medium THEN keyword = brand	8.285	0.007	0.997	0.466
	R2	IF TS = Low THEN keyword = olive oil	0.604	0.006	0.995	0.450
MESDIF	R3	IF TS = Medium THEN keyword = brand	8.285	0.007	0.997	0.466
	R4	IF PV = Low THEN keyword = olive oil	0.001	0.003	0.444	0.444
	R5	IF UPV = Low AND ATP = Medium THEN keyword = organic	0.233	0.001	0.232	0.232
NMEEF-SD	R6	IF TS = Medium THEN keyword = brand	8.285	0.007	0.997	0.466
	R7	IF PV = Low THEN keyword = olive oil	0.001	0.003	0.444	0.444

As can be observed in Table 2, there is a coincidence between the algorithms because rules $R1$, $R3$, and $R6$ are the same rule. This rule represents users that remain in the Web site during an interesting period of time in the Web site. These users accessed to OrOliveSur through a keyword related to a brand. However, with respect to the keyword olive oil, this variable obtains the linguistic label *Low*, i.e., when users access to the Web site through a keyword introduced in a search engine related to olive oil, they land in the Web site but they remain in small period. With rules $R2$ and $R7$ obtained by SDIGA and NMEEF-SD, respectively, we could consider that users cannot find information that they expected. Results obtained for the keyword organic are not very precise because in the database, there are few instances for this value. However, as can be observed in rule $R5$ is interesting to remark that despite of the number of page views is low, users remain in the Web site during a good period because the average time on page is medium.

With respect to the results obtained for each fuzzy subgroup, on the one hand, a good sensitivity for subgroups obtained for keyword brand with values close to 100 % and high values of significance can be observed. On the other hand, fuzzy confidence values are close to 50 % in keywords olive oil and brand. It is interesting to remark the use of few variables in subgroups to describe the problem.

5 Conclusions

This study presents the application of SD algorithm–based EFSs to the real-world application OrOliveSur: an e-commerce Web site related to olive oil and organic olive oil. The main objective is to extract unusual knowledge about users history associated with the Web site. SD algorithms allow extract knowledge with respect

to a target variable, where it is not necessary to obtain complete but partial relations. In this way, SD uses the descriptive induction through supervised learning.

The most interesting fuzzy subgroups obtained by SDIGA, MESDIF, and NMEEF-SD are presented. Conclusions proportioned to the Webmaster team with respect to the knowledge extracted are:

- The design must be reviewed with respect to the appearance and text of products in order to follow than users remain in the Web site during more time.
- A new categorization should be generated in order to facilitate the users different types of olive oils included in the Web site. In this way, users could explore more pages and perform more orders.
- Search Engine Optimization with respect to the keyword organic must be performed because there are few visits with respect to brand.

Acknowledgments This paper was supported by the Spanish Ministry of Education, Social Policy and Sports under project TIN-2008-06681-C06-02, FEDER Founds, by the Andalusian Research Plan under project TIC-3928, FEDER Founds, and by the University of Jaén Research Plan under proyect UJA2010/13/07 and Caja Rural sponsorship.

References

1. Etzioni O (1996) The World Wide Web: quagmine or gold mine. Commun ACM 39:65–68
2. Cooley R, Mobasher B, Srivastava J (1997) Web mining: information and pattern discovery on the World Wide Web. On tools with, artificial intelligence, pp 558–567
3. Kloesgen W (1996) Explora: a multipattern and multistrategy discovery assistant. In: Advances in knowledge discovery and data mining. American Association for, artificial intelligence, pp 249–271
4. Wrobel S (1997) An algorithm for multi-relational discovery of subgroups. In: Proceedings of the 1st European symposium on principles of data mining and knowledge discovery. Volume 1263 of LNAI. Springer, New York, pp 78–87
5. Carmona CJ, Ramírez-Gallego S, Torres F, Bernal E, del Jesus MJ, García S (2012) Web usage mining to improve the design of an e-commerce website: OrOliveSur.com. Expert Syst Appl 39:11243–11249
6. Lavrac N, Kavsek B, Flach PA, Todorovski L (2004) Subgroup discovery with CN2-SD. J Mach Learn Res 5:153–188
7. Herrera F, Carmona CJ, González P, del Jesus MJ (2011) An overview on subgroup discovery: foundations and applications. Knowl Inf Syst 29(3):495–525
8. Wrobel S (2001) Relational data mining. In: Inductive logic programming for knowledge discovery in databases. Springer, New York, pp 74–101
9. Lavrac N, Cestnik B, Gamberger D, Flach PA (2004) Decision support through subgroup discovery: three case studies and the lessons learned. Mach Learn 57(1–2):115–143
10. Gamberger D, Lavrac N (2003) Active subgroup mining: a case study in coronary heart disease risk group detection. Artif Intell Med 28(1):27–57
11. Kloesgen W, Zytkow J (2002) Handbook of data mining and knowledge discovery. Oxford
12. Lavrac N, Flach PA, Zupan B (1999) Rule evaluation measures: a unifying view. In: Proceedings of the 9th international workshop on inductive logic programming. Volume 1634 of LNCS. Springer, New York, pp 174–185

13. del Jesus MJ, González P, Herrera F, Mesonero M (2007) Evolutionary fuzzy rule induction process for subgroup discovery: a case study in marketing. IEEE Trans Fuzzy Syst 15(4):578–592
14. Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo A (1996) Fast discovery of association rules. In Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds.) Advances in knowledge discovery and data mining. AAAI Press, pp 307–328
15. Eiben AE, Smith JE (2003) Introduction to evolutionary computation. Springer, New York
16. Zadeh LA (1975) The concept of a linguistic variable and its applications to approximate reasoning. Parts I, II, III. Inf Sci 8–9:199–249,301–357,43–80
17. Herrera F (2008) Genetic fuzzy systems: taxonomy, current research trends and prospects. Evol Intell 1:27–46
18. Carmona CJ, González P, del Jesus MJ, Romero C, Ventura S (2010) Evolutionary algorithms for subgroup discovery applied to e-learning data. In: Proceedings of the IEEE international education, engineering, pp 983–990
19. Carmona CJ, González P, del Jesus MJ, Navío M, Jiménez L (2011) Evolutionary fuzzy rule extraction for subgroup discovery in a Psychiatric Emergency Department. Soft Comput 15(12):2435–2448
20. Carmona CJ, González P, del Jesus MJ, Ventura S (2011) Subgroup discovery in an e-learning usage study based on Moodle. In: Proceedings of the international conference of European transnational, education, pp 446–451
21. del Jesus MJ, González P, Herrera F (2007) Multiobjective genetic algorithm for extracting subgroup discovery fuzzy rules. In: Proceedings of the IEEE symposium on computational intelligence in multicriteria decision making. IEEE Press, pp 50–57
22. Carmona CJ, González P, del Jesus MJ, Herrera F (2010) NMEEF-SD: non-dominated multi-objective evolutionary algorithm for extracting fuzzy rules in subgroup discovery. IEEE Trans Fuzzy Syst 18(5):958–970
23. Srivastava J, Cooley R, Deshpande M, Tan P (2000) Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explorations, pp 12–23