Subgroup Discovery Applied to the e-Commerce Website OrOliveSur.com

C. J. Carmona[1], S. Ramírez-Gallego[1], F. Torres[2], E. Bernal[3], M. J. del Jesus[1] and S. García[1]
[1]*Department of Computer Science, University of Jaen, Jaen, Spain*
[2]*Department of Marketing, University of Jaen, Jaen, Spain*
[3]*Department of Economics, University of Jaen, Jaen, Spain*
{*ccarmona, ftorres, ebernal, mjjesus, sglopez*}*@ujaen.es*

Keywords: Subgroup Discovery, NMEEF-SD, Web Usage Mining, OrOliveSur.com.

Abstract: Subgroup discovery is a descriptive data mining technique whose main objective is the search for partial relations with unusual statistical characteristics with respect to a property of interest. In this paper, we present the application of a subgroup discovery technique in a users history data set associated to an e-commerce website called www.OrOliveSur.com which is related to sales of extra virgin olive oil and iberian products from Spain. The unusual knowledge is extracted using NMEEF-SD algorithm which is one of the most representative algorithm in this task throughout the literature. In order to apply this algorithm, information of website such as browser, source, keywords and so on is extracted through *Google Analytics* toolkit. Results obtained are discussed to provide advices and improve the design of the website.

1 INTRODUCTION

Electronic commerce is the buying and selling of products or services through electronic media, such as Internet and other computer networks. Nowadays, the amount of trade conducted electronically has grown extraordinarily due to the Internet. A high variety of commerce is made in this way (Soares et al., 2008), stimulating the creation and use of innovations such as electronic funds transfer, the supply chain management, marketing on Internet, online transaction processing, among others. Due to the concentration of olive oil cooperatives in Andalusia (Spain) in the last years, the literature proliferates on the export of olive products (Moral-Pajares and Lanzas-Molina, 2009), the use of e-commerce in the agricultural cooperatives and the adoption of Information and Communication Technologies as an essential toolkit in such export. This necessity arises to propose methodologies for intelligent data analysis, to enable the extraction of useful knowledge from the data. This is the concept of the Knowledge Discovery in Databases (KDD) (Han, 2005).

KDD in web mining was defined by Etzioni (Etzioni, 1996) as the use of data mining techniques to discover and extract knowledge in a website automatically, and by Cooley (Cooley et al., 1999) as the importance to consider the behaviour and preferences of the user. Web mining can be classified in three domains with respect to the nature of data (Cooley et al., 1997; Markov and Larose, 2007): web content mining, web structure data and web usage mining.

In the specialized literature, we found recent applications and consolidated reviews on the use of data mining in e-commerce. In (Schafer et al., 2001), the authors discussed different models of e-commerce recommendation and in (Hu and Liu, 2004) a methodology to extract information from customer questionnaires was provided. The extraction of predictive knowledge is used to set personalized recommendations in web use (Zhang and Jiao, 2007) and association rules are used for descriptive same task (Lazcorreta et al., 2008). Predictive and descriptive tasks can hybridize to achieve the same purpose (Kim et al., 2002) and the recommendation of time-varying products (Min and Han, 2005).

This paper is focused on web usage mining. In this way an specific methodology for extracting useful information from web usage data acquired using Google Analytics toolkit in the website www.OrOliveSur.com is applied: subgroup discovery task (Kloesgen, 1996; Wrobel, 1997). The main objective of this task is to obtain unusual knowledge and describe behaviour of different access to the website for users in order to increment the number of visits and orders in the website.

Structure of this paper is organised as follows: Section 2 presents the subgroup discovery data mining technique, Section 3 presents the main information about the e-commerce website in which is based this paper "www.OrOliveSur.com", in Section 4 the complete experimental study is presented and finally, Section 5 presents concluding remarks about this study to the experts.

2 SUBGROUP DISCOVERY

The concept of subgroup discovery was initially introduced by Kloesgen (Kloesgen, 1996) and Wrobel (Wrobel, 1997). It can be defined as (Wrobel, 2001):

> *"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."*

Considering this definition, the main property of this task is the search of partial relations where the majority of examples for the property of interest (or target variable) will be covered. In addition, the relations must be interesting with an unusual behaviour respect to the full data set.

In order to represent the knowledge, subgroup discovery employs a rule (R) which consists of an induced subgroup description. It can be formally defined as:

$$R : Cond \rightarrow TargetVar$$

where $TargetVar$ is a value for the variable of interest (target variable) for the subgroup discovery task (which also appears as *Class* in the literature), and $Cond$ is commonly a conjunction of features (attribute-value pairs) which is able to describe an unusual statistical distribution with respect to the $TargetVar$.

In Fig. 1 is represented a subgroup with two values for the target variable ($TargetVar = o$ and $TargetVar = x$). In this representation a subgroup for the first value of the target variable can be observed, where the rule attempts to cover a high number of objects with a single function as for example a circle. As can be observed the subgroup does not cover all the examples for the target value o even the examples covered are not positive in all the cases, but the function is uniform and simple.

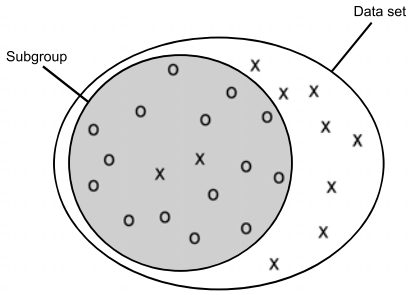

Figure 1: Representation of a subgroup discovery rule with respect to a value (o) of the target variable.

Throughout the literature have been presented a wide number of algorithms in the subgroup discovery task (Herrera et al., 2011), as for example proposals based on adaptations of classification algorithms, based on association rules algorithms or evolutionary fuzzy systems for subgroup discovery. This paper is focused in an evolutionary fuzzy systems called NMEEF-SD algorithm (Carmona et al., 2010) which is one of the most representative into subgroup discovery task.

3 OROLIVESUR.COM AN E-COMMERCE WEBSITE

OrOliveSur[1] 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 website 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 metres.

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 colour, open and structured. In addition, picual variety has excellent organoleptic properties because in stability and oleic acid obtains the best values with respect to other varieties like arbequina or hojiblanca, among others.

It is interesting to remark that users can find

[1] http://www.orolivesur.com

Figure 2: Homepage from the e-commerce website OrOliveSur.com.

an English (http://en.orolivesur.com) and Spanish (http://www.orolivesur.com) version. In Fig. 2 the homepage of OrOliveSur is shown.

Along two 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. The most important characteristic is that OrOliveSur offers a complete catalog with a lot of products and complete descriptions about these ones. Moreover, the OrOliveSur website gives direct sales and clients can pay with different types like transfer bank, PayPal or credit card.

Applying subgroup discovery algorithms in this data, the webmaster team can obtain information related to the main properties of user access with unusual behaviours with respect to a target variable such as source or keyword access for example.

4 NMEEF-SD APPLIED TO OROLIVESUR'S LOG DATA

Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery (NMEEF-SD) (Carmona et al., 2010) is an evolutionary fuzzy system (Herrera, 2008) whose objective is to extract descriptive fuzzy and/or crisp rules for the subgroup discovery task, depending on the type of variables present in the problem. This algorithm includes some quality measures in order to obtain rules with suitable values not only in the quality measures used but also in the rest of the most used quality measures in subgroup discovery. The best way to obtain solutions with a good compromise between several quality measures for subgroup discovery is through a MOEA approach. In this sense, NMEEF-SD has a multi-objective approach based on NSGA-II (Deb et al., 2002), a MOEA based on a non-dominated sorting approach, and on the use of elitism. NMEEF-SD is oriented towards the subgroup discovery task and uses specific operators to promote the extraction of simple, interpretable and high quality subgroup discovery rules. The algorithm permits a number of quality measures to be used both for the selection and the evaluation of rules within the evolutionary process and it also allows the use of different representation for rules (Carmona et al., 2009): canonical and DNF.

As the general objective of NMEEF-SD is to obtain a set of general and accurate rules, the algorithm includes components to enhance these characteristics. In particular, diversity is enhanced in the population using a new operator which performs a reinitialisation based on coverage. In addition, the algorithm can employ different niching techniques (Carmona et al., 2011a) as crowding distance, utility or knee-angle measure for the selection of the rules. In this study, a comparison among different measures promoting the diversity of the population is presented, in order to obtain the best compromise between the objectives of the MOEA. On the other hand, to promote generalisation, as well as the objectives considered in the evolutionary approach, the algorithm includes operators of biased initialisation and biased mutation. Finally, to ensure accuracy, in addition to the objectives, NMEEF-SD returns as its final solution those rules which reach a predetermined confidence threshold.

NMEEF-SD has shown its quality in real-world problems in different domains as education (Carmona et al., 2011c) or medical (Carmona et al., 2011b). The main purpose of the application of this algorithm in this data set is focused on the study of design in the e-commerce website of OrOliveSur.com through the obtention of unusual subgroups in a data set obtained with the webmaster toolkit *Google Analytics* from the period 1st January to 31st December for the year 2011. Among data we collect information related to:

- *Browser* name: IE, Firefox, Chrome, Android and so on.
- *Keyword* access: Olive oil, Iberian products, Brand, Gift, Other or Nothing.
- *Visitor type*: New or Returning.
- *New visits*.
- *Source* access: Direct, Mail, Search Engine, Social Network and so on.
- Page *views*.
- *Time* on site.
- Time per page ($time/page$).
- *Unique page views*.

Due to fact that the NMEEF-SD algorithm needs to select a target variable in order to obtain results, we employ as target variable different features: *Keyword*, *Visitor type* and *Source*, i.e. NMEEF-SD obtains different subgroups for each target variable selected with the main objective for describing a complete set of interesting relationships in data. With respect to the parameters used by NMEEF-SD algorithm can be observed in Table 1.

Table 1: Parameters used by NMEEF-SD algorithm.

Population size	50
Evaluations	10000
Crossover Probability	0.6
Mutation Probability	0.1
Minimum confidence	0.6
Rule representation	Canonical
Linguistic labels	9
Objective 1	Sensitivity
Objective 2	Unusualness

The most relevant subgroups obtained for NMEEF-SD algorithm with respect to the different property values together values of quality measures are shown in Table 2. This one describes rules obtained and the quality measures of significance ($SIGN$), unusualness ($UNUS$), sensitivity ($SENS$) and fuzzy confidence ($FCNF$). A complete description of these quality measures can be found in (Herrera et al., 2011).

As can be observed in results obtained by NMEEF-SD, there are a huge number of rules with high values in the majority of quality measures. Even

Table 2: Rules and results obtained by NMEEF-SD algorithm.

♯	Rule	SIGN	UNUS	SENS	FCNF
R1	IF source=E THEN keyword=olive oil	1949.707	0.117	0.999	0.483
R2	IF source=E THEN keyword=brand	1949.707	0.073	1.000	0.303
R3	IF time/page views=Low THEN keyword=nothing	3.920	0.001	0.999	0.448
R4	IF time=Low THEN keyword=nothing	11.175	0.005	0,982	0.486
R5	IF keyword=nothing AND page views=Very low AND unique page views=Very low THEN source=R	2216.810	0.090	0.996	0.373
R6	IF keyword=nothing AND unique page views=Very low THEN source=R	2265.863	0.089	0.999	0.368
R7	IF keyword=nothing AND page views=Very low AND page/visits=Very low THEN source=R	2216.810	0.090	0.996	0.372
R8	IF keyword=nothing AND unique page views=Very low AND unique page/visits=Very low THEN source=R	2265.863	0.089	0.999	0.368
R9	IF visitor-type=N AND unique page views=Low THEN source=E	90.077	0.038	0.658	0.653
R10	IF browser=IE AND page views=Low THEN source=E	137.419	0.057	0.575	0.709
R11	IF new visits=0 THEN visitor type=R	2819.825	0.229	1.000	1.000

though some rules like $R11$ is obvious because if visits are not news the consequence is because users are returning. However, this rule provides information about the good behaviour of the algorithm used.

It is interesting to remark that users that access directly to the website, i.e. without using any keywords as rules $R3$ and $R4$ show in the results, remain in the website during an acceptable time and time per page views is interesting. In addition, $R5$, $R6$, $R7$ and $R8$ show that reference websites like directories or blogs with external links to OrOliveSur are visits with low number of page-views and unique-page-views. In this way, webmaster must improve the description and image of OrOliveSur in these reference websites because it is probably that users does not find the information hoped.

Rule most interesting discovered by NMEEF-SD is the use of the browser Internet Explorer for the majority users that visit OrOliveSur through search engine as *Google* or *Yahoo*, for example. These users visit between 1 and 100 pages in the website. In this way, we recommend to the webmaster to analyse the design of the website to test that is correctly shown and designed in this browser in different versions.

5 CONCLUSIONS

In this paper, a study based on a subgroup discovery technique in order to extract unusual knowledge in a data set with information about users history associated to an e-commerce website is presented. These data are collected from the e-commerce website OrOliveSur.com which is related to the sell of extra virgin olive oil and iberian products from Spain. The main purpose is to discover interesting and unusualness information that allow to help to the webmaster team to improve the design of the website. To do so, NMEEF-SD algorithm is employed which is one of the most representative throughout the related literature. This real-world application is classified within web usage mining.

In general, knowledge discovered is related to the original point of user access where accesses performed through keywords are more interesting than by references websites. In this way, webmaster team must improve the description and image of OrOliveSur in these reference websites because it is probably that users do not find the information hoped.

Finally, an important recommendation is performed in order to analyse the design of the website with the browser IE in different versions because the majority visits are performed from this browser with high values of page views.

ACKNOWLEDGEMENTS

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

Carmona, C. J., González, P., del Jesus, M. J., and Herrera, F. (2009). An Analysis of Evolutionary Algorithms with Different Types of Fuzzy Rules in Subgroup Discovery. In *Proceedings of the FUZZIEEE*, pages 1706–1711.

Carmona, C. J., González, P., del Jesus, M. J., and Herrera, F. (2010). NMEEF-SD: Non-dominated Multiobjective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery. *IEEE Transactions on Fuzzy Systems*, 18(5):958–970.

Carmona, C. J., González, P., del Jesus, M. J., and Herrera, F. (2011a). Analysis of the Impact of Using Different Diversity Functions for the Subgroup Discovery Algorithm NMEEF-SD. In *Proceedings of the IEEE Int. Workshop on GEFS*, pages 17–23.

Carmona, C. J., González, P., del Jesus, M. J., Navío, M., and Jiménez, L. (2011b). Evolutionary Fuzzy Rule Extraction for Subgroup Discovery in a Psychiatric Emergency Department. *Soft Computing*, 15(12):2435–2448.

Carmona, C. J., González, P., del Jesus, M. J., and Ventura, S. (2011c). Subgroup discovery in an e-learning usage study based on Moodle. In *Proceedings of the ICEUTE*, pages 446–451.

Cooley, R., Mobasher, B., and Srivastava, J. (1997). Web Mining: Information and Pattern Discovery on the World Wide Web. *On Tools with Artificial Intelligence*, pages 558–567.

Cooley, R., Mobasher, B., and Srivastava, J. (1999). Data preparation for mining World Wide Web browsing patterns. *Knowledge and Information Systems*, 1:5–32.

Deb, K., Pratap, A., Agrawal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions Evolutionary Computation*, 6(2):182–197.

Etzioni, O. (1996). The World Wide Web: quagmine or gold mine. *Communications of the ACM*, 39:65–68.

Han, J. (2005). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc.

Herrera, F. (2008). Genetic fuzzy systems: taxomony, current research trends and prospects. *Evolutionary Intelligence*, 1:27–46.

Herrera, F., Carmona, C. J., González, P., and del Jesus, M. J. (2011). An overview on Subgroup Discovery: Foundations and Applications. *Knowledge and Information Systems*, 29(3):495–525.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Kim, J. K., Cho, Y. H., Kim, W. J., Kim, J. R., and Suh, J. H. (2002). A personalized recommendation procedure for Internet shopping support. *Electronic Commerce Research and Applications*, 1(3-4):301–313.

Kloesgen, W. (1996). Explora: A Multipattern and Multistrategy Discovery Assistant. In *Advances in Knowledge Discovery and Data Mining*, pages 249–271. American Association for Artificial Intelligence.

Lazcorreta, E., Botella, F., and Fernandez-Caballero, A. (2008). Towards personalized recommendation by two-step modified Apriori data mining algorithm. *Expert Systems with Applications*, 35(3):1422–1429.

Markov, Z. and Larose, D. T. (2007). *Data Mining The Web. Uncovering patterns in Web Content, Structure and Usage*. Wiley-Interscience.

Min, D. H. and Han, I. (2005). Detection of the customer time-variant pattern for improving recommender systems. *Expert Systems with Applications*, 28(2):189–199.

Moral-Pajares, E. and Lanzas-Molina, J. R. (2009). La exportacion de aceite de oliva virgen en Andalucia: Dinamica y factores determinantes. *Revista de estudios regionales*, 86(45-70).

Schafer, J. B., Konstan, J. A., and Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, 5(1-2):115–153.

Soares, C., Peng, Y., Meng, J., Washio, T., and Zhou, Z. H., editors (2008). *Applications of data mining in e-business and finance*. Frontiers in artificial intelligence and applications. IOS Press.

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*, pages 78–87. Springer.

Wrobel, S. (2001). *Inductive logic programming for knowledge discovery in databases*, chapter Relational Data Mining, pages 74–101. Springer.

Zhang, Y. and Jiao, J. (2007). An associative classification-based recommendation system for personalization in b2c e-commerce applications. *Expert Systems with Applications*, 33(2):357–367.