Analysing Concentrating Photovoltaics Technology Through the Use of Emerging Pattern Mining

A.M. García-Vico[1], J. Montes[2], J. Aguilera[2], C.J. Carmona[3](✉), and M.J. del Jesus[1]

[1] Department of Computer Science, University of Jaén, 23071 Jaén, Spain
{agvico,mjjesus}@ujaen.es
[2] Department of Electronics and Automatization Engineering, University of Jaén, 23071 Jaén, Spain
aguilera@ujaen.es
[3] Department of Civil Engineering, University of Burgos, 09006 Burgos, Spain
cjcarmona@ubu.es

Abstract. The search of emerging patterns pursues the description of a problem through the obtaining of trends in the time, or characterisation of differences between classes or group of variables. This contribution presents an application to a real-world problem related to the photovoltaic technology through the algorithm EvAEP. Specifically, the algorithm is an evolutionary fuzzy system for emerging pattern mining applied to a problem of concentrating photovoltaic technology which is focused on the generation of electricity reducing the associated costs. Emerging patterns have discovered relevant information for the experts when the maximum power is reached for the cells of concentrating photovoltaic.

Keywords: Emerging pattern mining · Concentrating photovoltaics · Evolutionary fuzzy system · Supervised descriptive rule discovery

1 Introduction

In data mining process there are two inductions clearly differentiated, predictive and descriptive induction. However, the last years there has been a great interest in the community about supervised descriptive rule discovery [18]. Latter includes a group of techniques for describing a problem through supervised learning such as subgroup discovery [4,14] or emerging pattern mining (EPM) [9], amongst others.

This contribution presents the application of the EPM technique to a real-world problem. The main objective of this data mining technique is to search for patterns with the ability to find large differences between datasets or classes. This property has led to the use of EPM in predictive induction with good results but not in descriptive induction although they were defined for this purpose. Specifically, the EPM algorithm employed in this contribution is the EvAEP algorithm

© Springer International Publishing AG 2017
M. Graña et al. (eds.), *International Joint Conference SOCO'16-CISIS'16-ICEUTE'16*, Advances in Intelligent Systems and Computing 527, DOI 10.1007/978-3-319-47364-2_32

which is an evolutionary fuzzy system (EFS) [13]. These systems are based on evolutionary algorithms [11] which offer advantages in knowledge extraction and in rule induction process. In addition, they use fuzzy logic [23] with the use of fuzzy sets with linguistic labels in order to represent the knowledge allowing to obtain representation of the information very close to the human reasoning [16].

The algorithm is applied to a Concentrating Photovoltaic (CPV) problem which is an alternative to the conventional Photovoltaic for the electric generation. CPV technology is based on using concentrated sunlight to produce electricity in a cheaper way by means of high efficiency multi-junction solar cells, specifically designed for this type of technology. The efficiency of this type of solar cells has experienced a fast evolution in the last decade and it has a very strong potential of increasing along next years. Despite of these expectations, several obstacles to develop CPV technology currently still remain, as the lack of CPV normalisation and standardisation, the lack of knowledge of the influence of the meteorological parameters on the performance of high efficiency multijunction solar cells. Therefore it is necessary to deepen in the study and knowledge of CPV technology. Results obtained in this contribution are very promising when maximum power is obtained by cells analysed.

The paper is organised as follows: Sect. 2 describes the background of the contribution with the presentation of EPM and CPV. Next, Sect. 3 shows the main properties and features of the algorithm EvAEP that is the first EFS for extracting EPs throughout the literature. Section 4 outlines the experimental framework, shows the results obtained, and an analysis about these results. Finally, some concluding remarks are outlined.

2 Background

2.1 Emerging Pattern Mining

The EPM was defined in 1999 [8] as itemsets whose support increase significantly from one dataset to another in order to discover trends in data. In this way, an itemset is considered as emerging when the growth rate (GR) is upper than one, i.e.:

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

where $Supp_{D_1}(x)$ is the support for the pattern x in the first dataset and $Supp_{D_2}(x)$ is the support with respect to the second dataset, i.e. $Supp_{D_1}(x) = \frac{count_{D_1}(x)}{|D_1|}$ and $Supp_{D_2}(x) = \frac{count_{D_2}(x)}{|D_2|}$. This concept could be generalised for one dataset with different classes, for example, $D_1 \equiv Class$ and $D_2 \equiv \overline{Class}$.

The most representative algorithm is DeEPs [19] which is based on the borders concept. A border is a pair of minimal and maximal patterns $< L, R >$ in order to represent all patterns within this border. As can be intuited the search space could become huge considering a complex problem. Therefore, authors have used different heuristics and techniques in order to reduce the search space and obtaining better results.

To facilitate to the experts and the community the analysis of the knowledge extracted, this is represented through the use of rules (R) with the following representation:

$$R : Cond \rightarrow Class$$

where $Cond$ is commonly a conjunction of attribute-value pairs as definitions mention, and $Class$ is the analysed value for the class, i.e. the class with a high support in front of the remaining classes for the dataset.

2.2 CPV Technology

Photovoltaic technology has experienced a major boost because it is a method of generating electrical power by converting solar radiation into direct current electricity using solar panels composed of a number of solar cells containing a semiconductor material. A variant of this technology is the CPV which is based on using concentrated sunlight to produce electricity in a cheaper way by means of high efficiency multi-junction solar cells, specifically designed for this type of technology. The efficiency of this type of solar cells has experienced a fast evolution[1]. In addition, CPV technology needs to use solar trackers, allowing an important increment of the energy generated by the system with a lower cost. Despite of these expectations, several significant obstacles to the development of CPV technology currently still remain:

- The lack of CPV normalisation and standardisation.
- The complexity and variety of solar cells.
- The lack of knowledge of the influence of meteorological parameters on the performance of high efficiency multi-junction solar cells. In fact, in real projects the productivity of this type of technology has been below expectations.
- The lack of detailed experimental and operational data about real outdoor performance.
- The development of complex regression models for performance.

The most interesting parameter to analyse is the Maximum Module Power (P_m). For each kind of solar cell, the manufacturer measures the CPV module under certain atmospheric conditions (called Standard Test Conditions, STC) and provides the $P_{m,STC}$. Nevertheless, in a real operation these conditions are not satisfied and the performance of the CPV module can be very different from that indicated by the manufacturer. It is known that P_m is highly influenced by atmospheric conditions, but it is necessary to know what happens with the combination of real atmospheric conditions. This knowledge can be very useful for predicting energy production in a certain period of time.

[1] http://www.nrel.gov/ncpv/images/efficiency_chart.jpg.

The *DNI* is considered as the main atmospheric parameter which influences the outdoor electric performance of a CPV module. Using the DNI as the integration value along the whole wavelength range for a specific photovoltaic device is a common practice. Nevertheless, we can consider the DNI value for each wavelength value, obtaining the *solar spectrum distribution*. As has been widely demonstrated, the DNI, as well as its spectral distribution, have an important influence on the electric performance of multi-junction solar cells. It is well known that the multi-junction solar cells temperature affect to their electric performance. In this sense, the temperature has an almost negligible positive effect on the short circuit current delivered by the multi-junction solar cell, and a negative predominant effect on both the open circuit voltage and P_M [12,17]. The same behaviour is observed when analysing the impact of the temperature (T_A) on the electric performance of CPV modules equipped with multi-junction solar cells [21]. However, the own disposition of the multi-junction solar cells inside the CPV module makes it very difficult to measure their temperature. In this work, T_A is considered as influential factor, given a direct relation between cell temperature and T_A [1,2]. The consideration of the wind (W_S) as one of the influential factors whose contribution must be added to the study because it can perform a positive refrigerating effect on the electric performance of a CPV system, cooling the multi-junction solar cells which compose the module down, and obtaining therefore a better behaviour [7]. However, high W_S values can also exert a negative effect of misalignment on the tracker [20], displacing the multi-junction solar cells from their optimum arrangement in the solar beam direct trajectory. Finally, for this contribution have been considered the incident global irradiance (G) and the spectral irradiance distribution of G described through the average photon energy (APE) and the spectral machine radio (SMR). Both parameters intend to define the shape of the solar spectrum in an easy way. The use of one or the other depends of the monitorised parameter during the experimental campaign.

3 EvAEP: Evolutionary Algorithm for Extracting Emerging Patterns

This section describes the algorithm Evolutionary Algorithm for extracting Emerging Patterns (EvAEP) presented in [6]. This algorithm is able to extract emerging fuzzy patterns in order to describe a problem from supervised learning. The main objective of the EvAEP is the extraction of an undetermined number of rules to describe information with respect to an interest property for the experts. It is important to note that a target variable is able to have different values or classes, in this way the algorithm obtain patterns for all values of the target variable because it is executed once for each value.

The algorithm is an EFS [13] that is a well-known hybridisation between a fuzzy system [23] and a learning process based on evolutionary computation [10]. EvAEP employs an evolutionary algorithm with a codification "Chromosome = Rule" where only the antecedent part of the rule is represented, and the antecedent is composed by a conjunction of pairs variable-value. Figure 1 represents the phenotype and genotype for a chromosome = rule in the EvAEP algorithm. As can be observed, the value 0 represents the absence of a variable in the representation of a rule.

$$\overset{Genotype}{\begin{array}{|c|c|c|c|} x_1 & x_2 & x_3 & x_4 \\ 3 & \emptyset & 1 & \emptyset \end{array}} \Rightarrow \overset{Phenotype}{\text{IF } (x_1 = 3) \text{ AND } (x_3 = 1) \text{ THEN } (x_{Obj} = ValorObjetivo)}$$

Fig. 1. Representation of a chromosome = rule for the EvAEP algorithm

On the other hand, if the variable has a continuous range, the algorithm employs a fuzzy representation with fuzzy sets composed by linguistic labels defined with uniform triangles forms.

The algorithm uses a mono-objective approach with an iterative rule learning (IRL) [22] which is executed once for each value of the target variable, i.e., for each class the best individuals are obtained in an iterative process. In this way, the algorithm iterates in order to obtain emerging patterns until a non-emerging pattern is obtained. Moreover, the algorithm stops if all instances for the class are covered for the patterns obtained previously or a pattern with null support is obtained.

The main operation scheme for the EvAEP algorithm is shown in Fig. 2.

The main elements of the algorithm are described in the following subsections.

```
BEGIN
Set of Emerging Patterns = ∅
repeat
   repeat
      Generate P(0)
      Evaluate P(0)
      repeat
         Include the best individual in P(nGen+1)
         Complete P(nGen+1): Cross and Mutation for individuals of P(nGen)
         Evaluate P(nGen+1)
         nGen ← nGen + 1
      until Number of evaluations is reached
      Obtain the best rule (R)
      Emerging Patterns ∪ R
      Marks examples covered by R
   until (GrowthRate(R) ≤ 1) OR (R has null support) OR (R not cover new ex-
   amples)
until Class = ∅
return Emerging Patterns
END
```

Fig. 2. Operation scheme for EvAEP

3.1 Biased Initialisation

EvAEP generates an initial population (P_0) with a size determined through external parameter. The objective of this function is to create a part of the individuals with a maximum percentage of variables being part of the rule. Specifically, the algorithm creates a population with 50 % of the individuals generated in a random way completely, and the remaining individuals of the population must have at least one variable with a value and a maximum number of variables (80 %) with values.

This operator allows the obtaining of a first population with a wide generalisation in order to explore the major area in the search space. In the evolutionary process the main idea is to maximise precision of the rules.

3.2 Genetic Operators

The population of the next generation is generated through some genetic operators widely used throughout the literature. The algorithm employs an elitism size equal to one, in this way the best individual is saved directly in the population of the next generation. The best individual is measured through an aggregation function such as:

$$NSup(R) * 0.5 + Fitness(R) * 0.5 \qquad (2)$$

where $NSup(R)$ is the number of examples covered for the rule R non-covered for the previous patterns obtained, divided by the number of remaining examples to cover for the class. The $Fitness$ is detailed in the following section. In case of tie, the individual with less number of variables is considered as the winner.

On the other hand, the algorithm employs the operator multi-cross point operator [15] and a biased mutation introduced in one algorithm of subgroup discovery [3].

3.3 Fitness Function

It is the key concept of the algorithm because the main objective is to search for emerging patterns with high values in confidence and precision, with the maximum generalisation possible, and finally, an interesting gain of accuracy for the community. The fitness employed by EvAEP is defined below:

$$Fitness(R) = \sqrt{TPr * TNr} \qquad (3)$$

where a geometric average for an individual is used in order to maximise the precision in the class and non-class in a balance way. The first component (TPr) is known as *True Positive rate* or sensitivity and it measures the percentage of examples correctly classified for the class, and the second (TNr) is known as *True Negative rate* where the measurement of examples non-covered correctly is considered.

4 Experimental Study

The data were obtained from one model of CPV modules, whose main characteristics are solar cells type $Multijunction - GaInP/Ga(In)As/Ge$ with 25 solar cells and a concentration factor of 550. The measures were acquired at the rooftop of the Higher Polytechnical School of Jaén during the period between March 2013 and November 2013, forming a whole dataset composed of 8780 samples. The characteristics of data collected, recorded every 5 min, are shown in Table 1. In summary, this system is able to simultaneously measure P_m of the CPV modules and the outdoor atmospheric conditions that influence the performance of the module.

Table 1. Characteristics of data collected by the Automatic Test & Measurement System

Variable	Name	Range	Unit
DNI	Direct normal irradiance	[600, 1040]	W/m^2
T_A	Ambient temperature	[6, 44]	°C
W_S	Wind speed	[0, 25]	m/s
G	Incident global irradiance	[650, 1370]	W/m^2
APE	Spectral irradiance distribution of G, described through average photon energy	[1.72, 2.28]	
SMR	Spectral irradiance distribution of G, described through spectral machine radio	[0.65, 1.25]	

It is important to remark that the experimentation process is performed through a separation between training and test dataset. In this case, we use 80 % (training) of the whole dataset to calculate the EPs. Otherwise, 20 % (test) of the whole dataset was used to validate the descriptive capacity of the proposed model.

For the type of solar module under study (Fig. 3), P_m values under 64.5 W are not significant, and the samples of the dataset with these values have been removed. P_m values have been discretised in three different intervals according to the $P_{m,STC}$ provided by the manufacturer for this kind of module (150 W) and the expert criteria. These intervals are depicted in Table 2, where P_m *range* are the values of the intervals defined on P_m (in W) and % P_m *range* is the percentage with respect to the maximum power established by the manufacturer (150W). In addition, the percentage of instances for each values is shown.

The results obtained by the algorithm EvAEP have been summarised in Table 3 where the rule (R) and its quality measures GR, TPr and FPr are shown. As we have mentioned previously, the GR is the growth rate of the rule, TPr is the true positive rate and FPr is the false positive rate that measures the ratio between the examples covered incorrectly and the number of examples for the non-class.

Fig. 3. Solar tracker at High Technical School of the University of Jaen

Table 2. Intervals defined by experts for P_m variable

Interval	P_m range	% P_m range	% of instances
2	[64.5, 93]	(43 %, 62 %]	20 %
3	(93, 121.5]	(62 %, 81 %]	68 %
4	(121.5, 150]	(81 %, 100 %]	12 %

Table 3. Results obtained in *Concentrating Photovoltaic Module* dataset

Rule	GR	TPr	FPr
R_1: IF $DNI =$ Very Low THEN $P_m = 2$	5.30	0.524	0.098
R_2: IF $DNI =$ Low THEN $P_m = 2$	75.89	0.362	0.004
R_3: IF $G =$ Medium THEN $P_m = 3$	1.35	0.889	0.656
R_4: IF $APE =$ Low THEN $P_m = 3$	1.05	0.993	0.942
R_5: IF $DNI =$ Medium THEN $P_m = 3$	4.33	0.397	0.091
R_6: IF $DNI =$ High AND $APE =$ Low AND $T_A =$ Medium AND Ws $=$ Medium THEN $P_m = 4$	3.87	0.037	0.009
R_7: IF $DNI =$ High AND $APE =$ Low THEN $P_m = 4$	9.92	0.134	0.013
R_8: IF $APE =$ Low AND $T_A =$ Low AND $W_S =$ Medium AND $SMR =$ High AND $G =$ High THEN $P_m = 4$	31.00	0.033	0.001
R_9: IF $APE =$ Low AND $T_A =$ Low AND $W_S =$ Low AND $SMR =$ High AND $G =$ High THEN $P_m = 4$	5.81	0.025	0.004
R_{10}: IF $APE =$ Low AND $T_A =$ Very Low AND $SMR =$ High THEN $P_m = 4$	21.31	0.046	0.002

As can be observed in the results obtained in this contribution could be performed different analysis conditioned to the power analysed:

- Low power: For this class are obtained the rules with the highest GR. In addition, the rules extracted have a good balance between TPr and FPr where the number of examples covered are in almost all cases of the class analysed.
- Medium power: In this class there is an emerging pattern that should be discarded, the rule 4 because is very general and the ratio between TPr and FPr is very similar and close to the 100 %, i.e. this rule covers all examples for the dataset both positive and negatives examples in the same ratio. On the other hand, rule 3 and 5 continue the tendency of the analysis where a major value for DNI is synonyms of major power.
- High power: For this class are obtained very specific rules but with good values in GR and relationships between TPr and FPr. As can be observed, all rules have a low value for APE which is a very interesting value, however it is important to note that this value is combined with a high value for DNI or SMR, rules 6, 7 or 8, 9, 10, respectively.

In general, a direct relationship between DNI and the P_m can be observed in the study. This assumption confirms the results obtained in a previous analysis performed through subgroup discovery [14] in the paper [5].

5 Conclusions

The CPV technology has been analysed from a new point of view in this contribution, the EPM data mining technique which is a descriptive induction based on supervised learning. The analysis has been performed through the EFS called EvAEP.

Results obtained in this study confirm the possible relationships between atmospheric variables and P_m suspected by CPV experts, as well as some new knowledge. The knowledge extracted on interval of $P_m = [43\,\%, 62\,\%]$ and $P_m = (62\,\%, 81\,\%]$ confirm the existing relations between DNI and P_m. However, there is a new interest obtained when the P_m is maximum because there is a relation in all rules, the values for APE which belongs to the linguistic label *Low*. In this way, there is an interesting further analysis with respect to the influence of this variable in the performance of the CPV module with maximum P_m and low values for APE.

Acknowledgment. This work was supported by the Spanish Science and Innovation Department under project ENE2009-08302, by the Department of Science and Innovation of the Regional Government of Andalucia under project P09-TEP-5045, and by Spanish Ministry of Economy and Competitiveness under project TIN2015-68454-R (FEDER Founds).

References

1. Almonacid, F., Pérez-Higueras, P., Fernández, E., Rodrigo, P.: Relation between the cell temperature of a hcpv module and atmospheric parameters. Sol. Energy Mater. Sol. Cells **105**, 322–327 (2012)
2. Antón, I., Martínez, M., Rubio, F., Núñez, R., Herrero, R., Domínguez, C., Victoria, M., Askins, S., Sala, G.: Power rating of CPV systems based on spectrally corrected DNI, vol. 1477, pp. 331–335 (2012)
3. Carmona, C.J., González, P., del Jesus, M.J., Herrera, F.: NMEEF-SD: non-dominated multi-objective evolutionary algorithm for extracting fuzzy rules in subgroup discovery. IEEE Trans. Fuzzy Syst. **18**(5), 958–970 (2010)
4. 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)
5. 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)
6. Carmona, C.J., Pulgar-Rubio, F.J., García-Vico, A.M., González, P., del Jesus, M.J.: Análisis descriptivo mediante aprendizaje supervisado basado en patrones emergentes. In: Proceedings of the VII Simposio Teoría y Aplicaciones de Minería de Datos, pp. 685–694 (2015)
7. Castro, M., Domínguez, C., Núez, R., Antón, I., Sala, G., A. K.: Detailed effects of wind on the field performance of a 50 kw CPV demonstration plant. In: AIP Conference Proceedings, vol. 1556, pp. 256–260 (2013)
8. Dong, G.Z., Li, J.Y.: Efficient mining of emerging patterns: discovering trends and differences. In: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 43–52. ACM Press (1999)
9. Dong, G.Z., Li, J.Y.: Mining border descriptions of emerging patterns from dataset pairs. Knowl. Inf. Syst. **8**(2), 178–202 (2005)
10. Eiben, A.E., Smith, J.E.: Introduction to Evolutionary Computation. Springer, Heidelberg (2003)
11. Eshelman, L.J., Schaffer, J.D.: Real-coded genetic algorithms and interval-schemata. In: Foundations of Genetic Algorithms 2, pp. 187–202. Kaufmann Publishers (1993)
12. Helmers, H., Schachtner, M., Bett, A.: Influence of temperature and irradiance on triple-junction solar subcells. Sol. Energy Mater. Sol. Cells **116**, 144–152 (2013)
13. Herrera, F.: Genetic fuzzy systems: taxomony, current research trends and prospects. Evol. Intel. **1**, 27–46 (2008)
14. Herrera, F., Carmona, C.J., González, P., del Jesus, M.J.: An overview on Subgroup Discovery: Foundations and Applications. Knowl. Inf. Syst. **29**(3), 495–525 (2011)
15. Holland, J.H.: Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor (1975)
16. Hüllermeier, E.: Fuzzy methods in machine learning and data mining: status and prospects. Fuzzy Sets Syst. **156**(3), 387–406 (2005)
17. Kinsey, G., Hebert, P., Barbour, K., Krut, D., Cotal, H., Sherif, R.: Concentrator multijunction solar cell characteristics under variable intensity and temperature. Prog. Photovoltaics Res. Appl. **16**(6), 503–508 (2008)
18. Kralj-Novak, P., Lavrac, N., Webb, G.I.: Supervised descriptive rule discovery: a unifying survey of constrast set, emerging pateern and subgroup mining. J. Mach. Learn. Res. **10**, 377–403 (2009)

19. Li, J.Y., Dong, G.Z., Ramamohanarao, K., Wong, L.: DeEPs: a new instance-based lazy discovery and classification system. Mach. Learn. **54**(2), 99–124 (2004)
20. Lin, C.-K., Fang, J.-Y.: Analysis of structural deformation and concentrator misalignment in a roll-tilt solar tracker. In: AIP Conference Proceedings, vol. 1556, pp. 210–213 (2013)
21. Peharz, G., Ferrer Rodríguez, J., Siefer, G., Bett, A.: Investigations on the temperature dependence of CPV modules equipped with triple-junction solar cells. Prog. Photovoltaics Res. Appl. **19**(1), 54–60 (2011)
22. Venturini, G.: SIA: a supervised inductive algorithm with genetic search for learning attributes based concepts. In: Brazdil, P.B. (ed.) ECML 1993. LNCS, vol. 667, pp. 280–296. Springer, Heidelberg (1993). doi:10.1007/3-540-56602-3_142
23. Zadeh, L.A.: The concept of a linguistic variable and its applications toapproximate reasoning. Parts I, II, III. Inform. Sci. **8-9**, 199–249, 301–357, 43–80 (1975)