

Predicción de potencia generada mediante sistemas fotovoltaicos: una aproximación desde la minería de datos

CASTILLO, Víctor*†, VENEGAS-TRUJILLO, Tiberio, SORIANO-EQUIGUA, Leonel y CONTRERAS-AGUILAR, Luis.

Facultad de Ingeniería Mecánica y Eléctrica, Universidad de Colima, Coquimatlán, Coquimatlán, Colima, México, C.P. 28400, Tel. 01 (312)316 11 65, Ext. 51451

Recibido Julio 15, 2016; Aceptado Septiembre 2, 2016

Resumen

Conforme crece la población mundial y se hace más accesible la adquisición de equipo eléctrico y electrónico, la demanda de la energía que haga funcional a ese equipo también aumenta. Ante esa problemática se han propuesto diversas tecnologías, siendo la fotovoltaica una de las más ampliamente adoptadas en los hogares. Aunque los sistemas fotovoltaicos se sustentan en una tecnología madura, la predicción exacta del nivel de producción es importante porque ayuda a estimar de forma precisa el tiempo en que ocurre el retorno de inversión de la misma. Aunque los métodos numéricos tradicionales ofrecen alternativas para predicción, el aumento de los volúmenes de datos sugiere que se empleen métodos alternos. La minería de datos ofrece una posible solución. Este artículo propone el uso de esta técnica para predecir la producción de energía en sistemas fotovoltaicos. Los resultados de esta evaluación sugieren niveles de exactitud aceptables, con la ventaja de poder usarse para hacer predicciones con datos provenientes de almacenes masivos. Lo anterior eventualmente beneficiará tanto a proveedores, como a consumidores, de los sistemas fotovoltaicos.

Predicción de potencia, árboles de regresión, minería de datos, sistemas fotovoltaicos

Abstract

Due to growing world population and cost reduction of electrical and electronic equipment, demand for energy that makes functional such equipment also increases. Faced with this problem it has been proposed several technologies, which photovoltaics is one of the most widely adopted in homes. Notwithstanding photovoltaic systems are based on a mature technology, the accurate prediction of the level of production is important because it helps to accurately estimate the time for return on investment. While traditional methods offer several choices for numerical prediction, increasing volumes of data requires alternative methods for predicting. Data mining offers a possible solution. This article proposes the use of this technique to predict the energy production in photovoltaic systems. The results of our evaluation suggest acceptable levels of accuracy, with the advantage of being used to make predictions with data from massive datasets. This will eventually benefit both suppliers and consumers of photovoltaic systems.

Power prediction, regression trees, data mining, photovoltaic systems

Citación: CASTILLO, Víctor, VENEGAS-TRUJILLO, Tiberio, SORIANO-EQUIGUA, Leonel y CONTRERAS-AGUILAR, Luis. Predicción de potencia generada mediante sistemas fotovoltaicos: una aproximación desde la minería de datos. Revista de Investigación y Desarrollo 2016, 2-5: 53-60

* Correspondencia al Autor (Correo Electrónico: victorc@uol.mx)

† Investigador contribuyendo como primer autor.

Introducción

Conforme crece la población mundial y se hace más accesible la adquisición de equipo eléctrico y electrónico, la demanda de la energía que haga funcional a ese equipo también aumenta. La OECD reporta que en los países miembros de esa organización, en 2014 se produjeron 10,361,778 GWh (OECD, 2016). Asimismo, en México, la Secretaría de Energía informa que en 2015 se produjeron 248,738.68 GWh (SENER, 2015).

La alta producción de energía, descrita previamente, también refleja la alta demanda de este insumo, así como la trascendencia de su proceso de producción y su impacto en el medio ambiente. Ante la misma, se han hecho esfuerzos para desarrollar fuentes alternativas de energía. Por ejemplo, en México se están realizando estrategias para mitigación de cambio climático. Entre otras alternativas, se proyecta que al 2030 la generación de electricidad solar fotovoltaica evitará la emisión de 0.21 MtCO₂ (CMM, 2009). Además de lo anterior, la energía fotovoltaica es importante en México porque este país se encuentra en lo que se conoce como el cinturón solar, en el cual se presenta una radiación solar mayor a 5KWh/m² al día (Lozano Cardona, 2013). Debido a las políticas públicas que promueven la producción de energías renovables en México, se estima que para 2040, aún con un incremento aproximado de 25% en la demanda de energía, en ese país disminuirán las tasas de emisión de PM_{2.5}, NO_x y SO₂, a valores cercanos a 80%, 70% y 50%, respectivamente (IEA, 2016).

Ante el escenario anterior se han realizado esfuerzos por administrar la producción y distribución de energía. Por ejemplo, considerando que la generación distribuida de energía por fuentes renovables no es constante, Barukčić *et al.* (2014) utilizan una estrategia de optimización multiobjetivo a fin disminuir pérdidas económicas.

Algunos autores proponen el uso de algoritmos genéticos para monitorear los disturbios transitorios en redes de transmisión, lo cual ayude a las compañías que generan energía, a proveer un servicio de calidad (Almeida & Kagan, 2011). Considerando que los algoritmos clásicos, que convergen en tiempo polinomial, tiempos de convergencia elevados cuando procesan grandes volúmenes de datos, en la actualidad existe investigación enfocada a la aplicación de algoritmos de minería de datos para estudiar fenómenos de producción de energía. A través de un enfoque de minería de datos se puede hacer un análisis en volúmenes de datos masivos y heterogéneos, asegurando tiempos de convergencia pertinentes. En ese sentido, Ferreira *et al.* (2011) proponen la utilización de algoritmos de agrupamiento para ayudar a los operadores del sector de producción de energía a planear la inversión y expansión de redes de generación futuras. Asimismo, Kusiak y Li (2010) recomiendan el uso de varios algoritmos de minería de datos para desarrollar modelos virtuales de turbinas de viento. No obstante, la propuesta del uso de enfoques de minería de datos para estudiar fenómenos de producción de energía eléctrica, no encontramos planteamientos sobre el uso de este enfoque para predecir producción de energía. La aplicación de la minería de datos para apoyar la predicción de ese proceso de producción es muy importante, ya que ayudaría a sustentar políticas públicas para la producción de energía basada en tecnologías verdes y, por otra parte, sería una herramienta importante para calcular en forma precisa retorno de inversión, que podría incentivar el uso de esas tecnologías en los hogares, aspecto que eventualmente tendría un impacto social relevante.

Aunque en la literatura se referencia trabajo que proyecta producción en plantas de energía renovable, esta utiliza lógica difusa (Rosato, Altilio, Araneo, & Panella, 2016), técnica que no corresponde al enfoque de minería de datos.

Debido a lo anterior, el objetivo del presente trabajo es apoyar la predicción de potencia generada mediante sistemas fotovoltaicos.

Este artículo está organizado como sigue. En la segunda sección describimos los métodos y herramientas analíticas que se utilizan para predicción numérica y se define la que se eligió para apoyar ese proceso. Posteriormente, en la tercera sección se detalla la metodología de investigación seguida en este trabajo. Entonces, en la cuarta sección se presenta los resultados obtenidos y se discute las implicaciones de los mismos. Finalmente, se puntualiza la conclusión del presente trabajo.

Predicción numérica

Como refieren Witten y Frank (2005), las aplicaciones de minería de datos se sustentan principalmente en cuatro estilos de aprendizaje, mismos que se aplican a los casos o ejemplos que integran un conjunto de datos. En el aprendizaje por *clasificación* se tiene un esquema de aprendizaje basado en un conjunto de ejemplos desde los cuales se aprende la forma de clasificar ejemplos no vistos. Por otra parte, en el aprendizaje por *asociación* se encuentran relaciones entre las características de los ejemplos de asociación, no sólo en aquellas que predigan el valor particular de una clase. En el aprendizaje por *agrupamiento*, se encuentran y conforman grupos de ejemplos. Finalmente, en la *predicción numérica* el resultado a predecir no es discreto, sino más bien una cantidad numérica.

Debido a que los ejemplos –es decir, datos que representan un valor de potencia, i.e. 5 Wh- de producción de energía son cantidades numéricas, nuestro interés se centra en predicción numérica.

Witten y Frank (2005) puntualizan que en minería de datos la predicción numérica se basa en estructuras de árbol. En una estructura de este tipo, una hoja podría representar dos cosas: 1) el valor de una clase que representa el valor promedio de las instancias que alcanzan esa hoja; o bien 2) un modelo de regresión lineal que predice el valor de la clase de instancias que alcanzan una determina hoja. En el primer caso el modelo se llama árbol de regresión y, en el último caso, árbol de modelo. Ambos tipos de árboles se construyen usando primero un algoritmo de inducción de árbol de decisión que ayuda a construir un árbol inicial. La única diferencia entre un árbol de regresión y uno de modelo es que en el de modelo cada nodo se reemplaza por un plano de regresión en un lugar de un valor constante.

De las técnicas de predicción numérica descritas previamente, los árboles de regresión son adecuados para predecir cantidades numéricas, como lo es la potencia eléctrica, razón por la cual, se describe a continuación.

Árboles de regresión

Loh (2011) establece que en un problema de clasificación se tiene una muestra de n observaciones en una variable de la clase Y que toma valores $1, 2, \dots, k$, y un número p de variables predictoras, X_1, X_2, \dots, X_p . La meta es encontrar un modelo para predecir valores de Y a partir de nuevos valores de X . Teóricamente, la solución consiste en particionar el espacio de X en k conjuntos disjuntos, A_1, A_2, \dots, A_k , tal que el valor predicho de Y es j si X pertenece a A_j , para $j = 1, 2, \dots, k$.

Si las X variables toman valores ordenados, Loh puntualiza que dos soluciones a ese problema podrían ser el análisis discriminante y la clasificación por el método del vecino más cercano. Asimismo, el autor explica que estos métodos generan conjuntos A_j con secciones lineales y no lineales, respectivamente, de fronteras que no son fáciles de interpretar si p es muy grande. Ante este problema, el método de árbol de clasificación proporciona conjuntos rectangulares A_j mediante la partición recursiva de una variable X a la vez en el conjunto de datos. De esta forma, los conjuntos de clasificación generados son más fáciles de interpretar. Sin embargo, un árbol de clasificación predice clases de objetos, no valores numéricos.

Por otra parte, un árbol de regresión ayuda a predecir cantidades numéricas. Éste es similar a un árbol de clasificación, excepto que la variable Y toma valores ordenados y en cada nodo se ajusta un modelo de regresión para generar los valores predichos de Y (Loh, 2011). Por lo anteriormente descrito, en el presente trabajo se utilizan árboles de regresión para predecir potencia generada en sistemas fotovoltaicos.

Metodología

Conjunto de datos

Los datos utilizados para la predicción numérica fueron captados en el periodo del 18 de mayo de 2014 al 18 de abril de 2016. El sistema fotovoltaico se interconectó a la red eléctrica de baja tensión. La tensión de operación del mismo fue de 220 Volts, a 2F-4H. Asimismo, la capacidad del sistema es de 1 kWp (un kiloWatt Punta). Los módulos fotovoltaicos son marca Solartec de 250 Watts, 60 celdas. Además, los microinversores son marca Enphase modelo M-215.

Del mismo modo, el sistema de monitoreo operó las 24 horas del día y registró parámetros eléctricos cada 5 minutos en las siguientes variables: potencia (W), energía (kWh), temperatura, frecuencia, corriente alterna, corriente directa, voltaje de CD y CA. El sistema se instaló en una casa con tarifa 1B residencial en la ciudad de Colima, México.

El conjunto de datos considerado para construir el árbol de regresión se constituye de tres columnas: *Energy*, *Month*, y *Day*. En la primera se almacena el total de energía generada en un día, la unidad es Wh. Por lo tanto, de acuerdo al periodo en que se hizo la lectura de datos en el sistema fotovoltaico, el conjunto de datos se compone de 702 filas o casos, las cuales representan el número de días de los que se tiene registro. Por otra parte, la columna *Month* almacena el mes relativo a cada una de las mediciones. Finalmente, la columna *Day* indica el día del mes en que se realizaron las mediciones. Los datos de las tres columnas son numéricos, enteros.

Generación del árbol de regresión

Para la generación del árbol de regresión se utilizó R (R_Foundation, 2016). R es un lenguaje y entorno de desarrollo para computación estadística. A fin de generar el árbol de regresión se usó el comando *tree* que se encuentra en el paquete del mismo nombre.

La figura 1 muestra un resumen de la construcción del árbol de regresión creado con el comando *tree* en R. Como se observa en la misma, se crearon 4 nodos a partir del conjunto de datos de potencia. Asimismo, puede verse que para la creación del árbol se tomaron en cuenta las columnas *Energy* y *Month* del conjunto de datos. De esta forma, el árbol de regresión generado muestra como criterio de ramificación el mes en que se toma la medición y, con base en éste, el modelo predice el monto de potencia generada.

Los resultados de predicción se detallan en la siguiente sección.

```
Regression tree:
tree(formula = Energy ~ Month, data = datos)
Number of terminal nodes: 4
Residual mean deviance: 669400 = 467200000 / 698
Distribution of residuals:
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-3876.0 -359.5  185.1    0.0  538.1 1402.0
```

Figura 1 Resumen del resultado de la generación del árbol de regresión.

Resultado y discusión

El modelo de árbol de regresión generado a partir de los datos se muestra en la figura 2, en la cual se observan los cuatro nodos terminales generados. Como lo describe el árbol de esa figura, el primer criterio para predecir la energía es determinar si la fecha de producción es antes de la mitad del mes de mayo. De ser así, si la fecha del mes se encuentra antes de mediados de febrero, la producción será de 3986 Wh, en caso contrario será de 4492 Wh. Por otra parte, si la fecha de producción de energía es a partir de la mitad del mes de mayo, se verifica que si la fecha es antes de mediados de agosto, la producción será de 4000 Wh, y en caso contrario, de 3747 Wh.

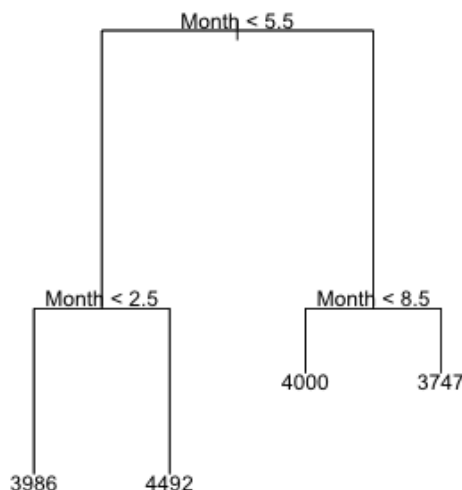


Figura 2 Árbol de regresión para la producción de potencia.

Los datos de la figura 2 pueden resumirse como la descripción de cuatro escenarios, correspondiendo a los cuatro nodos respectivos. Según el árbol de regresión generado, el escenario 1 se presenta si el periodo es desde el mes de enero hasta mediados de febrero, en el cual la producción será de 3986 Wh por día. Asimismo, el escenario 2 se manifiesta en el periodo de mediados de febrero a mediados de mayo, en el que la producción será de 4492 Wh diariamente. De la misma manera, a partir de mediados de mayo hasta mediados de agosto, se presenta el escenario 3, en el que la producción será de 4000 Wh al día. Finalmente, desde mediados de agosto hasta diciembre ocurre el escenario 4, en el que la producción será de 3747 Wh.

Para evaluar el comportamiento del árbol de regresión generado a partir los datos captados, se tomaron aleatoriamente bloques de seis casos ubicados en cada uno de los cuatro escenarios descritos previamente.

La tabla 1 muestra la desviación obtenida en los casos seleccionados para el escenario 1. Como se observa en la misma, existen valores precisos, donde la desviación de la potencia predicha es menor a 1% (ver caso 257 de esa tabla), y otros con márgenes de desviación notorios, como los presentados en los casos 254 y 258. El promedio del valor absoluto de las desviaciones es 22.1% para este escenario.

Caso	Potencia generada (Wh)	Día	Mes	Porcentaje de desviación
254	2579	26	1	-54.6
255	4103	27	1	2.9
256	4074	28	1	2.2
257	3953	29	1	-0.8
258	2733	30	1	-45.8
259	3159	31	1	-26.2

Tabla 1 Porcentajes de desviación para el escenario 1.

Por otra parte, la tabla 2 presenta resultados de desviación para datos incluidos en el escenario 2, descrito anteriormente. Como se observa en la misma tabla, para los casos seleccionados se presenta una desviación baja, con un mínimo de 0.2% y un máximo de 11.9%. En este escenario, el promedio del valor absoluto de las desviaciones es 5.9%.

Caso	Potencia generada (Wh)	Día	Mes	Porcentaje de desviación
290	5096	3	3	11.9
291	4402	4	3	-2.0
292	4279	5	3	-5.0
293	4785	6	3	6.1
294	4471	7	3	-0.5
295	4481	8	3	-0.2

Tabla 2 Porcentajes de desviación para el escenario 2.

Asimismo, se analizó el comportamiento para la predicción de potencia en el escenario 3, el cual se ilustra en la tabla 3. En ésta se observa que para los casos seleccionados, el porcentaje de desviación oscila entre el 1.5% y el 16.6%. En este escenario, el promedio del valor absoluto de las desviaciones es de 12.9%.

Caso	Potencia generada (Wh)	Día	Mes	Porcentaje de desviación
408	4192	29	6	4.6
409	4363	30	6	8.3
410	4797	1	7	16.6
411	4296	2	7	6.9
412	3942	3	7	-1.5
413	3481	4	7	-14.9

Tabla 3 Porcentajes de desviación para el escenario 3.

Finalmente, en la tabla 4 muestra la desviación de potencia predicha en el escenario 4. Como lo ilustra la tabla, para los casos seleccionados en la evaluación de este escenario, el porcentaje de desviación oscila entre el 0.3% y el 34.4%.

Para este escenario el promedio del valor absoluto de las desviaciones en los valores predichos es de 18.4%.

Caso	Potencia generada (Wh)	Día	Mes	Porcentaje de desviación
532	2785	31	10	-34.5
533	5097	1	11	26.5
534	3758	2	11	0.3
535	4184	3	11	10.4
536	4700	4	11	20.3
537	4578	5	11	18.2

Tabla 4 Porcentajes de desviación para el escenario 4.

De acuerdo a los resultados de evaluación descritos, el escenario que tuvo menos variación en la predicción fue el correspondiente al periodo desde mediados de febrero a mediados de mayo (escenario 2). Mientras que en el que se manifestó una variación más acentuada fue el relativo al periodo desde el mes de enero hasta mediados de febrero (escenario 1). Lo anterior podría explicarse debido a que en la región geográfica donde se levantaron los datos de potencia el clima es más estable en el periodo cubierto por el escenario 2. Por otra parte, el clima es menos estable en el periodo del año que circunscribe el escenario 1. La inestabilidad del clima es uno de los factores que hace complicado predecir el estado del tiempo, esa inestabilidad promueve la formación y desplazamiento de bancos de nubes, lo cual tiene un efecto directo en la producción de potencia en los sistemas fotovoltaicos. Otro hecho que podría explicar este comportamiento del modelo de predicción propuesto es la alta radiación solar que se presenta en el periodo circunscrito en el escenario 2, comparado con la mínima radiación que se presenta en el periodo considerado en el escenario 1 (Econotecnia, 2016), la cual influye en la nubosidad.

El presente trabajo tiene algunas limitaciones. Primero, éste no toma en cuenta aspectos secundarios que afectan la producción de potencia mediante sistemas fotovoltaicos, como la nubosidad. Sin embargo, como se describe previamente, la importancia del mismo radica en el modelo de árbol de regresión desarrollado, el cual es una primera aproximación a un modelo predictivo propio de la región donde se produjo la energía. También, es importante explicar que existen algoritmos para sistemas no lineales que podrían considerar más variables a fin de que el modelo predictivo sea más dinámico y que su comportamiento responda a múltiples variables, entre ellos se encuentran las redes neuronales y las máquinas de soporte vectorial (Steeb, 2005). Asimismo, otro aspecto importante que podría mejorar la exactitud del modelo predictivo propuesto es el de considerar la captación de datos en más puntos geográficos y durante un periodo más prolongado, pues en el presente estudio sólo se consideró un punto.

Conclusión

El presente trabajo esboza el diseño de un árbol de regresión como modelo predictivo de la producción de potencia mediante sistemas fotovoltaicos. Los árboles de regresión se generan mediante un algoritmo de aprendizaje de máquina propio de la minería de datos. De acuerdo al análisis del comportamiento del modelo diseñado, existe concordancia entre los valores de potencia predichos y los reales. Aún con estos resultados, el modelo podría mejorar si considerase otras variables que influyen en la producción de energía por medios fotovoltaicos, como lo es la nubosidad. Entre ese tipo de herramientas de cómputo están las técnicas no lineales, como las redes neuronales y las máquinas de soporte vectorial.

Los resultados del presente estudio son un paso adelante en la utilización de métodos de minería de datos que ayuden a predecir potencia en sistemas fotovoltaicos aún cuando éstos residan en almacenes de datos masivos. Lo anterior eventualmente beneficiará tanto a proveedores, como a consumidores de energía producida por medio de los sistemas fotovoltaicos.

Referencias

- Almeida, C., & Kagan, N. (2011). Using Genetic Algorithms and Fuzzy Programming to Monitor Voltage Sags and Swells. *IEEE Intelligent Systems*, 26(2), 46-53. doi:10.1109/MIS.2011.2
- Barukčić, M., Hederić, Ž., & Miklošević, K. (2014). *Multi objective optimization of energy production of distributed generation in distribution feeder*. Paper presented at the IEEE International Energy Conference (ENERGYCON), .
- CMM. (2009). *Centro Mario Molina. Modelos integrales de economía y cambio climático. La ruta de México hacia una economía sustentable de alta eficiencia energética y baja intensidad de carbón (Reporte I/VIII)*. Retrieved from México: http://transparencia.energia.gob.mx/estudios_opiniones.aspx
- Econotecnia. (2016). Datos de irradiación solar en la República Mexicana. Retrieved from <http://econotecnia.com/radiacion-solar.html>
- Ferreira, J., Ramos, S., Vale, Z., & Soares, J. (2011). A Data-Mining-Based Methodology for Transmission Expansion Planning. *IEEE Intelligent Systems*, 26(2), 28-37. doi:10.1109/MIS.2011.4
- IEA. (2016). *International Energy Agency. World Energy Outlook. Special Report*. Retrieved from <http://www.worldenergyoutlook.org/>

Kusiak, A., & Li, W. (2010). Virtual Models for Prediction of Wind Turbine Parameters. *IEEE Transactions on Energy Conversion*, 25(1), 245-252.
doi:10.1109/TEC.2009.2033042

Loh, W.-Y. (2011). Classification and regression trees. *Data Mining Knowl Discov*, 10.1002/widm.8, 14–23.

Lozano Cardona, W. (2013). *Energías renovables*. ProMéxico, Secretaría de Energía. Retrieved from México: http://mim.promexico.gob.mx/wb/mim/links_de_interes

OECD. (2016). Electricity generation (indicator). doi: 10.1787/c6e6caa2-en (Accedido el 6 de septiembre de 2016). Retrieved from <https://data.oecd.org/energy/electricity-generation.htm>

R_Foundation. (2016). The R Project for Statistical Computing. Retrieved from <https://www.r-project.org/>

Rosato, A., Altilio, R., Araneo, R., & Panella, M. (2016, 7-10 June 2016). *Embedding of time series for the prediction in photovoltaic power plants*. Paper presented at the 2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC).

SENER. (2015). Balance Nacional de Energía: Indicadores económicos y energéticos. Consultado en línea en 6 de septiembre de 2016. Retrieved from <http://sie.energia.gob.mx/bdiController.do?action=cuadro&cvecua=IE0C01>

Steeb, W.-H. (2005). *The Nonlinear Workbook*. Chaos, Fractals, Cellular Automata, Genetic Algorithms, Gene Expression Programming, Support Vector Machine, Wavelets, Hidden Markov Models, Fuzzy Logic with C++, Java and

SymbolicC++ Programs (3d ed.). Singapore: World Scientific Publishing Company.

Witten, I., & Frank, E. (2005). *Data mining. Practical machine learning tools and techniques* (2nd ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers.