TECNOLOGÍA QUÍMICA
UNIVERSIDAD DE ORIENTE
Vol.44, No 1, enero-abril 2024
ISSN: 2224-6185

Experiencias en la aplicación de la minería de datos en la industria biofarmacéutica cubana

Experiences in the application of data mining in the Cuban biopharmaceutical industry

Osvaldo Gozá-León[1]* https://orcid.org/0000-0002-1426-5910

Arturo Toledo-Rivero[2] https://orcid.org/0000-0001-7054-7210

[1]Facultad Ingeniería Química, Universidad Tecnológica de La Habana (CUJAE), Cuba

[2]Centro de Inmunología Molecular (CIM), La Habana, Cuba

*Autor para la correspondencia: ogoza@quimica.cujae.edu.cu

RESUMEN

El Centro de Inmunología Molecular es una institución exponente de la biotecnología cubana dedicado a la investigación básica, desarrollo, producción y comercialización de biofármacos, con el objetivo de diagnosticar y tratar el cáncer, así como enfermedades relacionadas con el sistema inmune. En este trabajo se realizó una síntesis bibliográfica de investigaciones que se llevaron a cabo en este centro, en el periodo comprendido del 2012 al 2018, con el propósito de analizar la aplicación de algunas técnicas de Minería de Datos en la evaluación de las etapas de fermentación y purificación del proceso de obtención de biofármacos en tres instalaciones productivas. Se caracterizaron las fases de la Minería de Datos y se presentó la implicación práctica de los resultados. Se aplicó como modelo descriptivo el Análisis de Componentes Principales mediante el software THE UNSCRAMBLER, y como modelo predictivo las Redes Neuronales Artificiales con el empleo de la caja de herramientas

de redes neuronales del MATLAB. La utilización de estos modelos permitió extraer información útil para la toma de decisiones, se logró con ello explicar el comportamiento de los parámetros que influyen en la calidad del producto final, así como estimar variables de importancia como la concentración de la proteína de interés en el sobrenadante de la fermentación y el rendimiento de la etapa de purificación, en función de las variables del proceso que mayor influencia tienen en su comportamiento.

Palabras clave: análisis de componentes principales; fermentación; minería de datos; purificación cromatográfica; redes neuronales.

ABSTRACT

The Center of Molecular Immunology is an exponent of Cuban biotechnology institution dedicated to basic research, development, production and commercialization of biopharmaceuticals, with the aim of diagnosing and treating cancer and diseases related to the immune system. In this work, a bibliographic synthesis of research that was carried out in this center in the period from 2012 to 2018, was performed, with the purpose of analyzing the application of some Data Mining techniques in the evaluation of fermentation and purification stages of the process for obtaining biopharmaceuticals in three production facilities. The phases of Data Mining were characterized and the practical implication of the results was presented. Principal Components Analysis was applied as a descriptive model using THE UNSCRAMBLER software, and Artificial Neural Networks were applied as a predictive model using the MATLAB neural networks toolbox. The use of these models made it possible to extract useful information for decision making, thereby explaining the behavior of the parameters that influence the quality of the final product, as well as estimating important variables such as the concentration of the protein of interest in the fermentation supernatant and the performance of the purification stage, depending on the process variables that have the greatest influence on its behavior.

Keywords: principal component analysis; fermentation; data mining; chromatographic purification;neural networks.

Recibido: 10/05/2023

Aceptado: 18/08/2023

Introducción

El Centro de Inmunología Molecular (CIM) es una institución exponente de la biotecnología cubana que se dedica a la investigación básica, desarrollo, producción y comercialización de biofármacos a partir de plataformas tecnológicas basadas en lo anterior, con el objetivo de diagnosticar y tratar el cáncer y enfermedades relacionadas con el sistema inmune. En este centro se produce una variedad de proteínas de alto valor terapéutico, como la Eritropoyetina Humana Recombinante (EPOhr), hormona glicoproteica que al aumentar la cantidad de glóbulos rojos se utiliza de forma efectiva en el tratamiento de la anemia severa; el hR3 o nimotuzumab, anticuerpo monoclonal humanizado que ha encontrado una efectiva aplicación en el tratamiento de pacientes aquejados de cáncer de cabeza y cuello; y otros productos en desarrollo como el biosimilar RITUXIMAB, anticuerpo monoclonal quimérico que se emplea con demostrada eficacia en el tratamiento de los linfomas no-Hodgkin de células B y enfermedades autoinmunes como el lupus eritematoso sistémico y la artritis reumatoide.

Una vez que la producción de un biofármaco ha sido aprobada, cualquier cambio sustancial en el proceso productivo requiere por lo general de nuevas pruebas clínicas para garantizar la seguridad y eficacia del producto. Como las pruebas clínicas son muy caras, las mejoras del proceso en la industria biofarmacéutica son llevadas a cabo bajo restricciones muy fuertes; de aquí que los procesos de producción sean conducidos normalmente con un comportamiento bien por debajo de su máximo potencial, lo cual limita en gran medida el mejoramiento continuo de los procesos.[(1)]

En el CIM los procesos productivos de obtención de la EPOhr, el hR3 y el RITUXIMAB tienen potencialidades para su mejora. Dichos procesos constan de dos etapas fundamentales, una etapa de fermentación seguida de una etapa de purificación en columnas cromatográficas. En los procesos estudiados en este trabajo ha habido inestabilidad en los rendimientos de ambas etapas y la calidad del producto final. A pesar de que existe una gran cantidad de información registrada, ha faltado conocimiento con relación a la forma en que las variables operacionales impactan en la variabilidad de dichos rendimientos y la calidad del producto final.

A partir de esta situación, se han realizado investigaciones en las etapas de fermentación y purificación que utilizan la técnica de Minería de Datos.

El término Minería de Datos (*Data Mining*) se utiliza mayoritariamente para referirse al proceso genérico correspondiente a las técnicas y herramientas de investigación usadas para extraer información útil de una base de datos. Dentro de estas técnicas se pueden considerar todos aquellos modelos matemáticos y técnicas basadas en aplicaciones de software para el análisis inteligente de los datos y búsqueda de patrones o tendencias en los mismos aplicados de forma iterativa e interactiva.

La Minería de Datos se ha convertido en una herramienta clave en la industria biofarmacéutica para ayudar en la identificación y el análisis de patrones y relaciones en grandes conjuntos de datos. Dicho análisis puede combinarse con otros métodos y técnicas para lograr una comprensión mejor del proceso y ejercer un control de calidad más efectivo. Los procesos de la industria biofarmacéutica típicamente generan conjuntos grandes de datos multivariados los cuales se caracterizan por ser muy heterogéneos, correlacionados y no lineales por naturaleza, así como por tener altos niveles de redundancia y ruido. En este sentido la utilidad de las técnicas de análisis de datos multivariados, como parte esencial de las técnicas de Minería de Datos, ha sido probada en esta área. Su habilidad para reducir dimensionalidad removiendo la redundancia y el ruido permite identificar las características más sobresalientes de los datos. Estas características pueden ser después utilizadas en el monitoreo de los bioprocesos, la detección de fallas y la optimización, lo cual

ha sido descrito extensamente en la literatura.[1] Los detalles internos de la mayoría de los procesos de la industria biofarmacéutica no son todavía bien comprendidos, lo cual dificulta el desarrollo de modelos matemáticos determinísticos, de ahí que para muchos de estos procesos sea necesario confiar sustancialmente en el desarrollo de modelos orientados a datos. Para el desarrollo de este tipo de modelos la disponibilidad cada vez más creciente de datos registrados tanto en línea como fuera de línea (espectroscópicos o de otro tipo) brinda al ingeniero una información sustancial como punto de partida para el análisis multivariante.[2]

Dada la potencialidad que tiene el análisis multivariado y los datos históricos acumulados de las producciones a escala comercial en el CIM, la aplicación de técnicas de Minería de Datos resulta de gran interés para extraer información útil que permita identificar las variables que mayor aporte tienen a la variabilidad de los procesos y con ello explicar el comportamiento de los parámetros que influyen en la calidad del producto final, así como cuantificar el impacto de las variables que se controlan sobre los rendimientos en cada etapa.

En este trabajo se realizó una síntesis bibliográfica de investigaciones que se llevaron a cabo en el Centro de Inmunología Molecular, Cuba, en el periodo comprendido del año 2012 hasta el 2018, con el propósito de analizar la aplicación de los métodos de Minería de Datos: Análisis de Componentes Principales (ACP) y Redes Neuronales Artificiales (RNA), en la evaluación de las etapas fermentación y purificación del proceso de obtención de biofármacos. Se caracterizaron las fases de la Minería de Datos y se presentó la implicación práctica de los resultados.

Fundamentación teórica

Fases típicas de un proceso de Minería de Datos

Las fases en el proceso global de Minería de Datos no están claramente diferenciadas, lo que hace que sea un proceso iterativo e interactivo con el

usuario experto. Las interacciones entre las decisiones tomadas en diferentes fases, así como los parámetros de los métodos utilizados y la forma de representar el problema suelen ser extremadamente complejos. Típicamente el proceso se estructura en seis fases que se ilustran en la figura 1.[3]

En la Minería de Datos se utilizan modelos descriptivos y predictivos. Los modelos descriptivos exploran las propiedades de los datos que se examinan, e identifican patrones que explican, resumen o caracterizan dichos datos. Estos modelos permiten acometer tareas tales como: asociación, correlación y agrupamiento. En las tareas de agrupamiento se obtienen grupos o *clusters* a partir de los datos, de manera que los objetos de un mismo grupo son muy similares entre sí, y muy distintos a los de otros grupos.

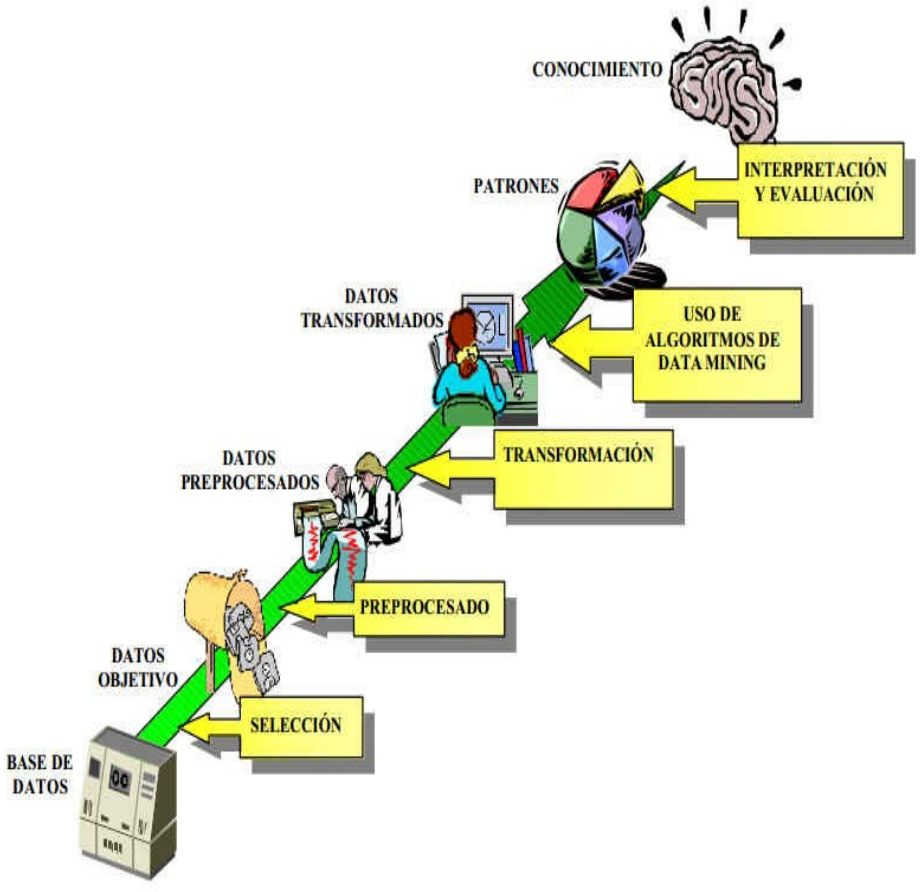

Fig. 1- Fases típicas de un proceso de Minería de Datos

Por su parte, los modelos predictivos estiman o predicen valores futuros de la variable objetivo del análisis, a partir de datos de entrada que se consideran

influyentes en su comportamiento. Estos modelos permiten acometer tareas tales como: clasificación, regresión y predicción. En las tareas de regresión y predicción se examinan los datos, y en función de ellos, se asigna a la variable objetivo (numérica) uno de sus posibles valores.

La preparación de los datos es el paso inicial en el desarrollo del modelo. El objetivo de este paso consiste en adquirir una visión de los datos del proceso y con ello entonces seleccionar los más apropiados para la modelación. Las tareas principales de este paso consisten en extraer el conjunto de datos de las bases de datos históricas, examinar la estructura del conjunto de datos y seleccionar las muestras y variables.[4] Para asegurar la eficiencia de este paso se deben analizar las características de los datos del proceso tales como el alejamiento de la distribución normal, y el nivel de correlación entre las variables. Otro aspecto importante es la selección de las variables y las muestras, lo cual está estrechamente relacionado con el paso de desarrollo del modelo. Dicha selección depende del tipo de modelo que se quiere desarrollar, y de cuál es la tarea principal a acometer con ese modelo.

Una vez que en el paso inicial se conformó el conjunto de datos, es necesario realizar el pre-procesamiento para mejorar la calidad de los datos, y algunas transformaciones apropiadas de los datos pudieran ser necesarias para que el modelado sea más eficiente.[5]

Con el conjunto de datos de entrenamiento, es posible seleccionar un algoritmo de Minería de Datos apropiado para la construcción del modelo. Sobre la base de un análisis detallado de las características de los datos, se valora la complejidad del modelo a obtener. Una vez que se selecciona la estructura del modelo, sus parámetros se pueden determinar implementando un algoritmo de Minería de Datos con el conjunto de datos de entrenamiento. Finalmente, para que el modelo pueda ser utilizado el mismo necesita ser validado, requiriéndose para ello también un conjunto de datos de validación.

Materiales y métodos

En la realización de este trabajo se utilizaron como modelo descriptivo el Análisis de Componentes Principales; y como modelo predictivo, las Redes Neuronales Artificiales.

Análisis de Componentes Principales

El Análisis de Componentes Principales es una de las técnicas estadísticas multivariantes más difundidas en el análisis de datos. Sus principales objetivos son: extraer la información más importante de un conjunto de datos multivariados, comprimir un conjunto de datos multivariados manteniendo solo la información que se considere importante (reducir la dimensionalidad de los datos), simplificar la descripción de un conjunto de datos y analizar la estructura de las observaciones y de las variables.(6)

La idea central del ACP es reducir la dimensionalidad de un conjunto de datos correspondientes a un gran número de variables, y retener tanto como sea posible la variación de los datos originales. Esto se logra transformando las variables originales en un nuevo conjunto de variables, combinación lineal de las primarias, que se denominan componentes principales, los cuales no están correlacionados entre sí y se ordenan de forma tal que el primer componente retiene la mayor parte de la variación presente en las variables originales.(7)

En el presente estudio el Análisis de Componentes Principales se realizó con la versión 8.0 del *software THE UNSCRAMBLER,* el cual está especialmente concebido para análisis multivariado de datos. Este programa permite realizar un análisis de los resultados con la ayuda de cuatro gráficos fundamentales para el entendimiento e interpretación de la información: gráfico de la varianza explicada, gráfico de la influencia, gráfico de las puntuaciones o mapa de las muestras y gráfico de los pesos o mapa de las variables.(8)

Redes Neuronales Artificiales

Las Redes Neuronales Artificiales son modelos que intentan simular la estructura y los aspectos funcionales de las redes neuronales biológicas.

Típicamente las neuronas se agrupan en diferentes capas como capa de entrada, capas de salida y capas ocultas. El uso de las capas ocultas les confiere a las redes neuronales la habilidad de describir sistemas no lineales. Una de los paradigmas de redes neuronales más populares que se han aplicado en la modelación de sistemas no lineales, especialmente en procesos químicos y biológicos, es la red neuronal unidireccional con aprendizaje de propagación hacia atrás (*feed-forward back propagation*), la cual se utilizó en este trabajo. En el presente estudio, el procedimiento de desarrollo de las redes neuronales se llevó a cabo con la caja de herramientas de redes neuronales del MATLAB, versión 14.

Resultados y discusión

En las tres plantas productoras a escala comercial del CIM: EPO/vacunas y anticuerpos monoclonales terapéuticos hR3, RITUXIMAB, en el periodo comprendido entre los años 2012 y 2018, se realizaron investigaciones con aplicación de técnicas de Minería de Datos. Consisten en trabajos de diploma y tesis de maestría, con la tutoría de los autores de este trabajo, en opción a los títulos de Ingeniero Químico, Máster en Análisis y Control de Procesos, y Máster en Ingeniería de los Procesos Biotecnológicos, las cuales se defendieron en la Facultad de Ingeniería Química de la Universidad Tecnológica de La Habana. Presentaron como objeto de estudio el proceso de obtención de biofármacos a partir del cultivo de células de mamíferos, y como campo de acción la etapa de fermentación en biorreactor de tanque agitado o la etapa de purificación en columnas cromatográficas en el proceso en cuestión (tabla 1).

El paso inicial de preparación de los datos constituyó siempre el que mayor tiempo y esfuerzo conllevó, lo cual se corresponde con lo reflejado en la literatura especializada. Fue necesario extraer de las bases de datos históricas los conjuntos de datos para la construcción de los modelos; dichos datos se caracterizaron por ser multivariados con una alta dimensionalidad, variabilidad

y heterogeneidad, así como mostraron estar correlacionados y tener alejamientos de la distribución normal.

Para la selección de las variables a considerar en la modelación se tuvieron en cuenta estudios ya realizados con anterioridad en las plantas, consultas con expertos de las plantas y en varios casos se partió de la determinación de los parámetros críticos de la etapa en cuestión a través de un modelo de riesgo basado en matriz de entradas y salidas.

Tabla 1- Tesis realizadas en el CIM con aplicación de técnicas de Minería de Datos

Título de la tesis (Tipo, año)	Planta, etapa	Modelos
Estimación del rendimiento en la etapa de purificación de EPOhr aplicando técnicas de Minería de Datos (Maestría, 2012)	EPOhr, purificación	RNA
Análisis exploratorio de datos en la etapa de purificación de la planta de EPOhr en el CIM (Diploma, 2013)	EPOhr, purificación	ACP
Análisis exploratorio de datos en la etapa de fermentación de la planta de EPOhr en el CIM (Diploma, 2015)	EPOhr, fermentación	ACP
Análisis del proceso de fermentación en la planta de EPOhr aplicando técnicas de Minería de Datos (Maestría, 2015)	EPOhr, fermentación	ACP, RNA
Análisis exploratorio de datos en el proceso de purificación de un anticuerpo monoclonal (Diploma, 2016)	hR3, purificación	ACP
Determinación de parámetros críticos para la verificación continuada del proceso de purificación durante la producción de un anticuerpo monoclonal (Maestría, 2016)	hR3, purificación	ACP
Análisis exploratorio de datos en el proceso de fermentación de la planta de hR3 (Diploma, 2017)	hR3, purificación	ACP
Análisis exploratorio de datos en el proceso de fermentación de la planta de EPOhr (Diploma, 2017)	EPOhr, fermentación	ACP
Análisis multivariante del proceso de fermentación en la planta de producción del anticuerpo monoclonal hR3 (Maestría, 2017)	hR3, fermentación	ACP, RNA
Análisis exploratorio de datos en el proceso de fermentación de la planta de 1B8 (Diploma, 2018)	1B8, fermentación	ACP
Análisis exploratorio de datos en el proceso de purificación del anticuerpo monoclonal 1B8 (Diploma, 2018)	1B8, purificación	ACP

Aplicación del Análisis de Componentes Principales

En los Análisis de Componentes Principales que se realizaron, la heterogeneidad de los datos, motivada por la presencia de variables de

diferente naturaleza y diferentes magnitudes, condujo a la combinación del autoescalado y la normalización de los datos como parte del preprocesamiento, lo que facilitó la realización del análisis.

Los gráficos de influencia y de las puntuaciones que se obtuvieron por el programa THE UNSCRAMBLER al realizar el ACP, facilitaron grandemente la detección de los puntos discrepantes (*outliers*), lo que permitió definir cuáles eran los de mayor alejamiento de un comportamiento normal y debían ser desechados y cuales a pesar del alejamiento había que considerar por contener información útil del proceso. Todos los puntos discrepantes que no se desecharon se correspondieron con cambios operacionales en el proceso que provocaron afectaciones en las variables estudiadas, lo cual se corroboró con la experiencia práctica de los especialistas en cada planta.

El Análisis de Componentes Principales permitió reducir la dimensionalidad de los datos y definir cuáles son las variables que mayor aporte tienen a la variabilidad del proceso, lo cual resultó de gran utilidad para enriquecer la estrategia de control en cada planta.

En las etapas de fermentación el análisis se centró siempre en el fermentador industrial, donde se obtuvo en unos casos que con un solo componente principal se logra explicar más del 90% de la varianza total y en otros que con dos componentes principales se logra explicar más del 80 % de la varianza total.

En las etapas de purificación el análisis abarcó los distintos pasos cromatográficos que se utilizan. En los casos de las plantas de anticuerpos monoclonales terapéuticos, dada la complejidad de los procesos a lotes, se establecieron dos secciones en serie bien definidas, la primera que se identifica como Purificación I donde se elimina la mayor cantidad de impurezas, y la segunda, llamada Purificación II donde se refina y acondiciona el producto de salida a granel a sus especificaciones de calidad. En todos los casos se obtuvo que con dos componentes principales se logra explicar más del 80% de la varianza total.[9, 10]

A modo de ilustración, en el Análisis de Componentes Principales que se realizó a los datos del proceso de fermentación en la planta de hR3 durante la campaña del 2014, [8] se partió de la determinación de los parámetros críticos de la etapa de fermentación a través de un modelo de riesgo basado en matriz de entradas y salidas, y se definieron diez parámetros críticos en el fermentador industrial: temperatura, tiempo de duración del cultivo, viabilidad celular, equilibrio ácido - base (ph), velocidad específica de crecimiento de la biomasa, velocidad específica de formación del producto de interés, flujo de medio de cultivo, concentración de células vivas, velocidad de agitación y oxígeno disuelto. Como resultado se obtuvo que dos componentes principales logran explicar más del 99 % de la varianza total, y se logró definir cuáles son los parámetros críticos que mayor aporte tienen a la variabilidad del proceso de fermentación. Dichos resultados corroboraron experiencias prácticas de especialistas de la planta y permitieron dar recomendaciones a considerar en el plan de verificación continuada del proceso, como proponer la inclusión en la estrategia de control del proceso a la temperatura, la velocidad de agitación, el oxígeno disuelto y el tiempo de duración del cultivo. En aplicación posterior, el Análisis de Componentes Principales realizado a datos del proceso de fermentación, permitió la determinación de las mejores condiciones nutricionales como base para la optimización del medio de cultivo.[11]

Aplicación de Redes Neuronales Artificiales

Con las Redes Neuronales Artificiales fue posible obtener modelos que permitieron estimar, para la etapa de purificación el rendimiento de EPOhr en bulbos por litro de sobrenadante aplicado y para la etapa de fermentación la concentración de la proteína de interés a la salida del fermentador industrial, en ambos casos en función de las variables del proceso que mayor influencia tienen en su comportamiento.

Las redes utilizadas fueron unidireccionales con una capa oculta y se realizó el entrenamiento con el algoritmo Levenberg-Marquardt de propagación hacia atrás, con el empleo del 70% de los datos para el entrenamiento y el 30% para la validación. Como funciones de transferencia se utilizaron la sigmoidea

(*tansig*) en la capa oculta y la función lineal (*purelin*) para la capa de salida. Los mejores resultados se obtuvieron en la etapa de purificación con errores cuadráticos medios (MSE, *Mean Square Root*) en el orden de las milésimas y coeficientes de correlación (R) mayores que 0,95, mientras que en la etapa de fermentación los errores cuadráticos medios (MSE) estuvieron en el orden de las centésimas y los coeficientes de correlación (R) fueron mayores que 0,9.[12]

A modo de ilustración, en la modelación que se hizo de la etapa de purificación de la planta de EPOhr con datos correspondientes al 2012, [12] se obtuvo una Red Neuronal Artificial para la predicción del rendimiento de dicha etapa que comprende cuatro pasos cromatográficos, medida como el rendimiento de EPOhr en bulbos por litro de sobrenadante aplicado, en función de 11 variables operacionales. La red neuronal que mejor ajustó tenía 11 neuronas en la capa oculta y muy buen desempeño en las fases de entrenamiento y validación con un error cuadrático medio (MSE) de 0,0042 y un coeficiente de correlación (R) de 0,958. Con el método de los pesos de las conexiones se ordenaron las variables de entrada según su contribución a la función objetivo, y se obtuvo que cinco variables predominaban sobre las restantes. Estos resultados suministraron información útil, mostraron que los pasos cromatográficos primero y tercero son decisivos para alcanzar rendimientos altos en la etapa de purificación, lo cual enriqueció la estrategia de control en la planta.

Conclusiones

1. Se realizó una síntesis bibliográfica de investigaciones que se llevaron a cabo en tres plantas productoras a escala comercial del CIM, en el periodo comprendido del año 2012 hasta el 2018, en la que se analizó la aplicación de los métodos de Minería de Datos: Análisis de Componentes Principales (ACP) y Redes Neuronales Artificiales (RNA), en la evaluación de las etapas de fermentación y purificación del proceso de obtención de biofármacos.
2. Se analizaron las peculiaridades de las fases de la Minería de Datos por las que se transitó. Como modelo descriptivo se utilizó el Análisis de Componentes Principales, con el software THE UNSCRAMBLER, en el que se logró explicar

más del 80% de la varianza total en cada caso. Como modelo predictivo se usaron las Redes Neuronales Artificiales, con la caja de herramientas de redes neuronales del MATLAB. Las redes utilizadas fueron unidireccionales con una capa oculta, se realizó el entrenamiento con el algoritmo Levenberg-Marquardt de propagación hacia atrás, y se obtuvieron ajustes con coeficientes de correlación mayores que 0,9.

3. Como implicación práctica de los resultados, el Análisis de Componentes Principales permitió definir cuáles son las variables que mayor aporte tienen a la variabilidad del proceso y con ello explicar el comportamiento de los parámetros que influyen en la calidad del producto final. Las Redes Neuronales Artificiales, permitieron estimar variables de importancia como la concentración de la proteína de interés en el sobrenadante en el caso de la fermentación y el rendimiento de EPOhr en bulbos por litro de sobrenadante aplicado en caso de la etapa de purificación, en función de las variables del proceso que mayor influencia tienen en su comportamiento. Las investigaciones realizadas permitieron profundizar en el entendimiento de los procesos y enriquecer sus estrategias de control.

Referencias bibliográficas

1. TEIXEIRA, A. P., OLIVEIRA, R., ALVES, P. M. AND CARRONDO, M. J. T. Advances in on-line monitoring and control of mammalian cell cultures: Supporting the PAT initiative. *Biotechnology Advances*, [en línea]. 2009, **27**, 726-732. DOI: 10.1016/j.biotechadv.2009.05.003

2. TROUP, G. M., Georgakis, C. Process systems engineering tools in the pharmaceutical industry. *Comput. Chem. Eng.*, [en línea]. 2013, **51**, 157-171.

3. MARTÍNEZ DE PISÓN, F. *Optimización mediante técnicas de minería de datos del ciclo de recocido de una línea de galvanizado*. Tesis de Doctorado. J. B. Ordieres Meré (dir.). Universidad de La Rioja, Logroño, España, 2003.

4. GE, Z.I., SONG, Z., DING, S. X., HUANG, B. Data Mining and Analytics in the Process Industry: The Role of Machine Learning. *IEEE Access*. 2017, 5, 20590-20616. DOI: 10.1109/ACCESS.2017.2756872

5. XU, S., Lu, Bo., Baldea, M., Edgar, T.F., Wojsznis, Willy., Blevins, T., Nixon, M. Data cleaning in the process industries. *Rev. Chem. Eng*, [en línea]. 2015, **31**, 453–490. DOI: 10.1515/revce-2015-0022

6. ABDI, H., WILLIAMS, L.J. Principal Component Analysis. *Rev. Comp. Stat.*, [en línea]. 2010, **2**, 433–459. DOI: 10.1002/wics.101

7. RODIONOVA, O., KUCHERYAVSKIY, S., POMERANTSEV, A. Efficient tools for principal component analysis of complex data - a tutorial. *Chemom. Intell. Lab. Syst.*, [en línea]. 2021, 213. DOI: 10.1016/j.chemolab.2021.104304

8. MESA, L., GOZÁ, O., URANGA, M., TOLEDO, A., GÁLVEZ, Y. Aplicación del Análisis de Componentes Principales en el proceso de fermentación de un anticuerpo monoclonal. *VacciMonitor.*, [en línea]. 2018. **27**, 8-15.

9. GOZÁ, O., Fernández, M., Rodríguez, R.H., Ojito, E. Aplicación del Análisis de Componentes Principales en el proceso de purificación de un biofármaco. *VacciMonitor*, [en línea]. 2020. **29**, 5-13.

10. TOLEDO, A., GOZÁ, O., HERNÁNDEZ, E., LEONARD, I., HIDALGO, G. A continued process verification strategy at first stages of monoclonal antibody purification by integrated risk assessment and multivariate data analysis. *Biotecnol. Apl.,* [en línea]. 2021, **38**,1201-8. ISSN 1027-2852.

11. HERNÁNDEZ, E., CALZADILLA, L., TOLEDO, A., GOZÁ, O., PIETZKE, M., VAZQUEZ, A., RODRÍGUEZ, G., QUINTANA, A., LEON, K., BOGGIANO., T. Determination of Best Nutritional Conditions for a Monoclonal Antibody-Producing Cell Line based on a Multivariate Data Analysis Approach. *Glob. J. Eng. Tech.: J General Engineering,* [en línea]. 2023, **23**(1). ISSN: 2583-3359.

12. RODRÍGUEZ, R.H., Gozá, O., Ojito, E. Preliminary modeling of an industrial recombinant human erythropoietin purification process by artificial neural networks. *Braz. J. Chem. Eng.,* [en línea]. 2015, **32**, 725-734. ISSN 1678-4383. DOI: 10.1590/0104-6632.20150323s00003527

Conflictos de Interés

Los autores declaran que no hay conflictos de intereses.

Contribución de los autores

Osvaldo Gozá León: conceptualización, metodología, investigación, escritura y corrección.

Arturo Toledo Rivero: conceptualización, metodología, investigación y corrección.