SENSORES VIRTUALES MEDIANTE REDES NEURONALES ARTIFICIALES, CASO EN BIOTECNOLOGÍA

Marisol Quispe Aduviri
Simulación de Sistemas
marysol_sa@live.com

RESUMEN

Las redes neuronales artificiales (RNA) se caracterizan por aprender a través de entrenamiento en lugar de descripciones formales, esto las ha hecho la opción preferencial para modelar procesos de variables con interrelaciones complejas. Algunos de estos procesos se encuentran en el área de biotecnología. La biotecnología computacional es un concepto novedoso en el que se combinan dos disciplinas científicas actualmente en fuerte expansión. Por un lado la biotecnología con todas sus implicaciones económicas y sociales. Algunos de estos procesos se encuentran en el área de biotecnología. La estimación robusta de la biomasa en un proceso de fermentación es de particular interés. Una característica deseable para los estimadores es la capacidad que tengan de integrar una variedad de mediciones para producir su estimación y que sean robustos ante fallas. En particular es importante la robustez ante fallas de los sensores que proveen las variables en línea. Por la dificultad de hacer las mediciones de biomasa y la robustez implícita en las RNA se ha escogido implementar sensores virtuales a través de éste método.

Palabras clave

Biotecnología, Análisis de datos, Procesado de señales, Genómica, Proteómica.

1. INTRODUCCIÓN

Cualquier sistema de control o de monitoreo requiere de elementos de interfase con el mundo real, el mundo físico. Así, un proceso industrial requiere de una diversidad de sensores para observar e identificar su estado actual y tomar decisiones de control. Para un robot móvil es esencial tener información de su alrededor para saber hacia donde moverse o desarrollar estrategias para elegir la trayectoria más apropiada. En un proceso de producción es de vital importancia saber si el sistema de producción en realidad está produciendo la calidad para la que fue diseñado, por lo tanto se requieren sensores para verificarlo.

Dado que las áreas de aplicación de los sensores son muy diversas, existen sensores de muchos tipos, que se eligen dependiendo de la naturaleza de la variable a medir y de las condiciones de la aplicación. En el proceso industrial antes mencionado puede haber distintos tipos de sensores midiendo distintas variables que son relevantes en el desarrollo del proceso. El cómo procesar los datos de los sensores en este caso y hacer deducciones del proceso basados en esa información es motivo de muchos estudios y existen ya teorías de control para aplicar la información según la naturaleza del proceso. Para un caso como el descrito se usarían modelos multivariable del proceso y teorías de control adecuadas para esos modelos.

En el caso del robot y el sistema de control de calidad la situación del manejo de los datos de los sensores puede tornarse bastante distinta. En este caso lo más común es que el interés sea una sola variable (como la dirección a overse de tal manera que no haya colisiones ó la brillantez del producto) pero en cambio se tienen varios sensores midiendo la misma variable. En este caso el objetivo sería usar sensores que actúen bajo distintos principios para medir la misma variable, de tal manera que la información sea más confiable. Se busca redundancia para obtener confiabilidad. Ahora el problema de procesamiento de la información es muy distinto ya que se requiere de la fusión de la información de fuentes distintas (distintos tipos de sensor, distintos niveles de las variables, etc.) para obtener el valor de una sola variable.

Existen situaciones en las que el sensor adecuado no existe o es prohibitivamente caro por lo que se desearía una manera de estimar la información de esa variable. En muchos casos la información sobre esa variable inmensurable existe pero en otras formas. Por ejemplo, se sabe que en todo proceso la presión y temperatura están ligadas, por lo que sabiendo una, se podría inferir la otra mediante el procedimiento adecuado. En el caso de procesos de fermentación el monitoreo de la variable de biomasa es esencial, sin embargo, no existe un sensor para medirla. Existen básicamente dos métodos para medir esta variable: a) el método automático a través de un romatógrafo o densidad óptica y b) el método manual a través de muestras tomadas a mano y analizadas por un especialista. El método a) cuesta al menos $8,000.00 dólares por lo tanto se prefiere el método b). Éste método no provee muestras periódicas ni de suficiente frecuencia por lo tanto no es completamente satisfactorio. Basados en esta idea de inferencia de información han nacido los llamados sensores por software o soft-sensors o sensores virtuales. Estos son programas que son los encargados de hacer la inferencia tomando la información existente. Los programas pueden consistir en un modelo matemático de cómo hacer la inferencia, en un modelo heurístico o un 'modelo inteligente'.

Un método usado como modelo inteligente es el de las redes neuronales artificiales (RNA). Las RNA son un conjunto de elementos procesadores adaptables que simulan burdamente el funcionamiento de una neurona animal. En una aplicación como esta, las RNA harían un modelo inferencial de la variable en cuestión tomando como base la información de las variables medidas. En el área de procesos industriales las variables del proceso siempre están ligadas de una u otra manera. En ocasiones la interrelación entre variables es compleja y altamente no lineal, sin embargo, las RNA y han demostrada ser capaces de aprender a base de entrenamiento estas dinámicas.

2. SENSORES VIRTUALES

En muchos tipos de procesos industriales y en otros casos donde se tiene multiplicidad de sensores y monitoreo o control por computadora es posible implementar un sensor virtual. Ya sea porque es difícil medir la variable en cuestión, porque el sensor es muy caro, demasiado lento o inexacto los sensores virtuales se hacen necesarios y ahora son cada vez más populares. Se está desarrollando ya la primera propuesta comercial de un sensor virtual múltiple de este tipo para la industria automotriz. En esta aplicación se usa un modelo neuronal del motor del automóvil para hacer asociaciones altamente complejas, no lineales y multidimensionales

Las asociaciones que hace el modelo neuronal se generan mediante entradas de los sensores estándar de los motores actuales y las salidas deseadas, que en este caso son par del motor, emisiones tóxicas y consumo de combustible en todo el rango de operación. El modelo neuronal se obtiene mediante entrenamiento en una estructura de prueba del motor con un dinamómetro o entrenamiento en línea con el automóvil funcionando. Los microorganismos son siempre muy sensibles al conjunto de pequeñas variaciones en las distintas variables, por lo tanto se requiere de un sistema de estimación altamente robusto y con muy buena capacidad de generalización. Una de las razones por la que los sensores virtuales son muy robustos en su funcionamiento es su característica de redundancia intrínseca sin incurrir en gastos adicionales de hardware o más sensores. El hecho de tener como entradas datos de una multiplicidad de variables y que éstas puedan ser dependientes unas de otras o estar relacionadas de alguna manera, provee esta redundancia intrínseca que evita fallas drásticas aun en el evento de fallas en los sensores.

Presentamos el planteamiento y algunos resultados de dos casos de estudio en los que hemos estado involucrados.

En el primer caso que analizamos se trata de la estimación de biomasa en un proceso de producción de un antibiótico como producto secundario. En el segundo, se produce un pigmento rojo altamente cotizado.

3. CASO I: ESTIMACIÓN DE LA BIOMASA EN LA PRODUCCIÓN DE UN ANTIBIÓTICO.

La mayoría de las aplicaciones de estimación inferencial encontradas en la literatura en esta área son de biomasa y se generan de lecturas de la razón de evolución del dióxido de carbono (‘carbón dióxido evolution rate’, CER) y el tiempo transcurrido en la fermentación. Si el analizador de dióxido de carbono falla, entonces la estimación falla. Se propone, en el primer caso de estudio, tener un sistema para proveer una estimación de la biomasa a partir de varios conjuntos de mediciones diversas, como subgrupos de CER, razón de consumo de oxígeno (‘oxígeno uptake rate’, OUR), pH (o las adiciones ácidas y alcalinas para controlarlo), el tiempo transcurrido en la fermentación, etc. Esta redundancia de información nos lleva a obtener una supervisión y control de la fermentación más robustos y es una característica intrínseca de los sensores virtuales. No sólo eso sino que nos puede llevar a hacer importantes observaciones acerca del proceso mismo, por ejemplo, la influencia relativa de cada variable en una etapa particular de la producción. Varias RNA de alimentación hacia adelante (‘feed-forward’)” se han escogido utilizando el algoritmo “backpropagation” a través del método Levenberg-Marquardt. Se ha observado que, dado el entrenamiento apropiado, la metodología de crear sensores virtuales a través de RNA, es capaz de estimar la biomasa con una precisión comparable a la de la instrumentación, teniendo un proceso no cambiante (es decir, que el método se aplique siempre al mismo proceso). Más aun, los resultados muestran que las estimaciones son casi insensibles a ruido y a fallas en los sensores debido a la integración y redundancia disponibles.

3.1. PREPROCESAMIENTO.

Los datos del proceso fueron normalizados antes de ser presentados a las RNA de tal manera que el rango de todas las variables se redujera a entre 0.1 y 0.9 para obtener una convergencia más rápida en el entrenamiento de las redes. También se aplicó interpolación lineal a los datos históricos de la biomasa para obtener una razón de muestreo periódica y uniforme para todas las variables de interés. Basados en experimentos previos, se decidió aplicar un periodo de muestreo constante de dos horas para todas las variables. Las variables en uso como entradas a las RNA son la razón de evolución del dióxido de carbono (CER, por sus siglas en inglés), oxígeno disuelto, tiempo transcurrido en la fermentación y pH. Se ha llegado a la elección de estas variables, considerando también reportes de otros investigadores.

3.2. LAS REDES NEURONALES ARTIFICIALES.

Estamos usando redes de alimentación hacia adelante con una capa escondida. El número de entradas a cada red es igual al número de variables que se ha escogido para cada una y que se presenta más adelante. Basados en experimentos previos, el número de neuronas en la capa escondida de cada red será de 6. La salida de todas las redes es siempre la biomasa (sólo una salida). Hemos encontrado para esta aplicación que si se quiere mantener una buena capacidad de generalización de las redes, el criterio de convergencia (la figura de error) debe ser holgado. Para este caso elegimos un valor de alrededor de 5% del rango de los datos normalizados. Las funciones de activación de la capa escondida son tipo sigmoide y en la salida se tienen funciones de activación lineales. El algoritmo de entrenamiento es 'retropropagación' (Backpropagation) con regla de ajuste Levenberg-Marquardt. La figura de error que se usa para evaluar el desempeño de las redes es el Error Absoluto Medio

$$EAM \propto \sum_{i}^{n} abs(\hat{Y}_i - T_i) \Big/ n \qquad \text{(EAM) (1),} \qquad (1)$$

donde *Y*$ es el vector de estimación, *T* es el vector de valores deseados, *n* es el número de valores de entrada *i* y abs(•) indica el valor absoluto. El factor de proporcionalidad se debe a que el valor del EAM es un porcentaje del rango de los datos normalizados.

3.3. LA APLICACIÓN.

Como primer paso después de este pre procesamiento se entrenan cuatro RNA para estimar el valor de la biomasa,

teniendo cada una un grupo distinto de entradas. Los grupos de entradas son los siguientes:

1) CER y tiempo transcurrido (CT).
2) DOT y tiempo transcurrido (DT).
3) pH y tiempo transcurrido (PT).
4) CER, DOT y tiempo transcurrido (CDT).

A las estimaciones de biomasa de estas RNA les llamamos resultados de la primera etapa. Existe una segunda etapa de procesamiento que llamamos etapa de fusión, para la que proponemos una de dos opciones, como se explica adelante.
En la primera se usa una RNA a la que le llamamos 'red de fusión'. Para ésta, las estimaciones de las cuatro RNA de la primera etapa, más la señal de tiempo transcurrido se usan como entradas para entrenar la red de fusión para estimar biomasa nuevamente, como se ilustra en la Fig. 2. La arquitectura de la red de fusión es igual a las demás con una capa escondida y 6 neuronas en esa capa.

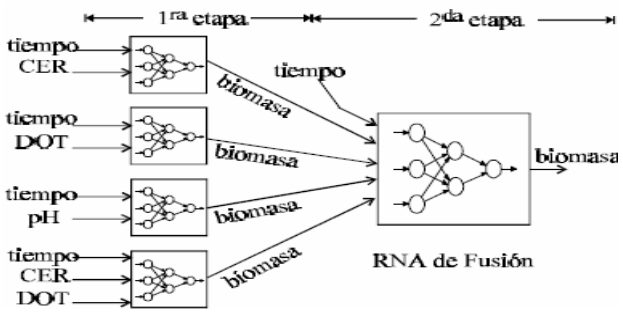

Fig. 1 La RNA de fusión.

Una segunda forma de implementar la fusión se logra mediante un programa que elige las mejores estimaciones de las proporcionadas por la primera etapa. Así, para cada muestra, se elige la mejor estimación con respecto a una referencia. Cómo generar esa referencia es un problema interesante ya que la referencia óptima es precisamente la variable que se está tratando de generar.
La solución que proponemos es tomar esta decisión basados en un pronóstico de un paso hacia adelante. Este es una RNA que, entrenada con datos de biomasa en el tiempo *t* y el tiempo transcurrido (dos datos que se tienen), genera una predicción de la biomasa en el tiempo *t+1*. Esta predicción es bastante buena ya que sólo se predice un paso de tiempo por lo tanto sirve bien para el efecto que se desea. Este método también provee información de qué conjunto de entradas es más importante en cada paso de tiempo. Este segundo método de fusión está ilustrado en la Fig. 3. Se ha observado que errores por fallas de los sensores y ruido son disminuidos en gran medida usando esta técnica. A este procedimiento le hemos llamado Sistema de Análisis Temporal (SAT).

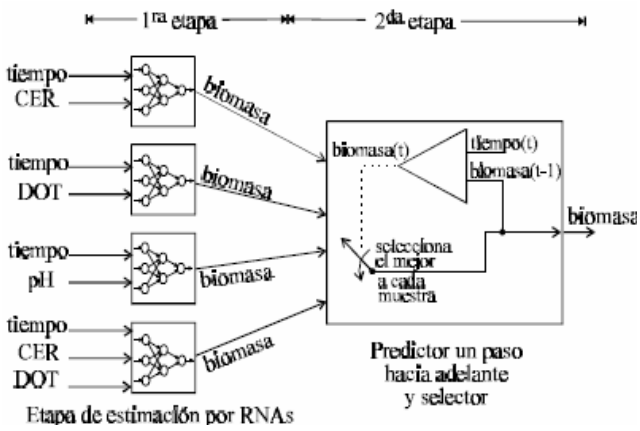

Fig. 2 Nuestro Sistema de Análisis Temporal (SAT).

3.4. PRUEBAS SIN PERTURBACIONES.

Para estos dos métodos de procesamiento, se entrenan las redes con el grupo de datos de una fermentación y se prueba el desempeño con tres grupos distintos de fermentaciones del mismo tipo, repetidas. En la Tabla 1 se muestra un resumen de los resultados de las pruebas.

TABLA 1 RESUMEN DE LOS RESULTADOS DE LAS ESTIMACIONES DE BIOMASA. SE PRUEBAN LOS DOS MÉTODOS DE FUSIÓN (FUS Y SAT) PROPUESTOS PARA LAS CUATRO FERMENTACIONES.

	F181	F187	F253	F256	Media
FUS	4.52%	1.19%	9.62%	5.90%	**5.31%**
SAT	2.65%	2.41%	5.75%	7.49%	**4.58%**

En la tabla de resultados F187 se refiere a los datos de la fermentación que se usó para entrenamiento de las redes, por eso presenta los errores más bajos. En términos de promedios, las estimaciones de los dos métodos están alrededor de 5% lo cual puede no parecer muy impresionante, aunque se está demostrando una capacidad de generalización en realidad muy importante.
Lo que es más impresionante es la demostrada robustez de ambos sistemas. Las figuras Fig. 4 y Fig. 5, muestran gráficas de los resultados para una de las fermentaciones de prueba, F181. En Fig. 5, el símbolo en el recuadro indica el origen de la estimación.

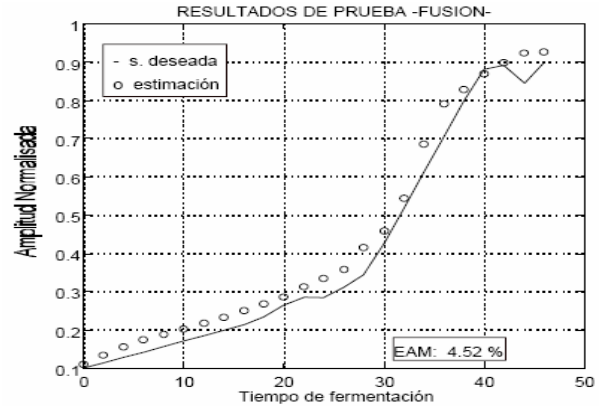

Fig. 3 Resultado de la estimación de biomasa con la red de fusión para la

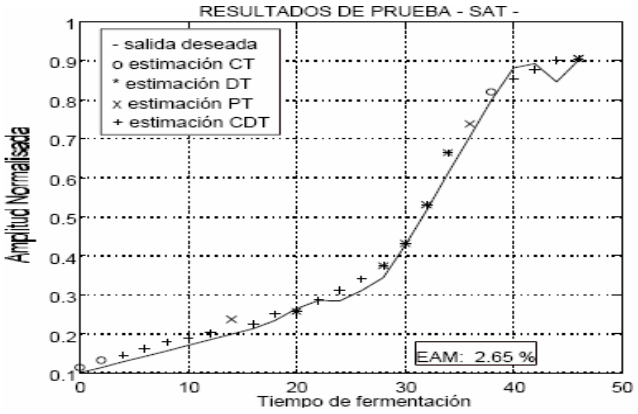

Fig. 4 Estimación de biomasa del Sistema de Análisis Temporal para la fermentación F181.

3.5. PRUEBAS CON PERTURBACIONES.

Con el propósito de probar la robustez de los estimadores, investigamos su comportamiento ante varios tipos de perturbaciones. Un tipo de perturbación que probamos fue ruido aleatorio. Estas pruebas se hicieron añadiendo un vector de ruido a la señal que nos interesaba afectar. La magnitud de la señal de ruido es igual a la de la variable afectada pero el ruido se aplica mediante la siguiente ecuación:

señal_ruidosa (i)=señal(i)+señal(i)*ruido(i) (1)

La razón de ser de Ec. (1) es que creemos que modela bien una situación real donde la magnitud del ruido es proporcional a la magnitud de la señal. Esta perturbación se aplica solamente a la señal CER porque es la más importante. Para las otras variables se usan los datos reales del proceso.

4. CASO II: ESTIMACIÓN DE BIOMASA Y PRODUCTO SECUNDARIO EN LA PRODUCCIÓN DE UN PIGMENTO.

Existen algunos trabajos relacionados con la estimación de variables mediante RNA, donde reportan haber obtenido estimaciones de biomasa y un producto secundario. Leal et al. utilizaron un conjunto de RNA para hacer diferentes estimaciones de biomasa y dejaron que otra red decidiera cual era la mejor estimación de biomasa. En esta aplicación la propuesta de solución es un poco mas sencilla y consta de usar una RNA para la estimación.

Como se mencionó anteriormente el uso de las RNA como sensores virtuales tiene una posible aplicación por demás interesante. Entrenando y probando una red con distintos tipos de entradas se pueden hacer deducciones acerca de la relevancia de el conjunto de entradas con la salida, observando cual conjunto es el que da mejores resultados. Esta metodología se uso para estudiar las variables disponibles en un proceso de producción de un pigmento. También se han hecho estudios extensos sobre la topología de las RNA más adecuada para esta aplicación. La configuración que aparece en la figura 5 es ejemplo de resultados parciales en este trabajo. La variable 'Oxígeno' se refiere al oxígeno consumido por los microorganismos en la aplicación. 'Espuma' es la espuma que se produce en el fermentador también
relacionada con la cantidad de microorganismos. 'Edad_Ferm' es simplemente la edad de la fermentación, es decir, el tiempo transcurrido desde el inicio de ésta.

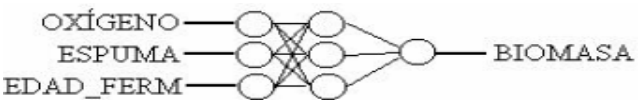

Fig. 5 RNA como estimador de biomasa basado en oxígeno, cantidad de espuma producida y edad de la fermentación.

Los resultados en este estudio son todavía parciales pero abarcan estimaciones de producto secundario y primario.

4. CONCLUSIONES.

Se demostró que una red neuronal artificial es capaz de aprender la interrelación entre la biomasa y algunas de las variables más importantes del proceso en su conceptualización como sensor virtual. Se propone una mejora en las estimaciones a través de un método de dos etapas promoviendo redundancia de datos. En la primera etapa se generan varias estimaciones de biomasa a partir de varios agrupamientos de variables de entrada. En la segunda etapa se fusionan estas estimaciones para obtener una estimación mejorada de biomasa. Para esta fusión se proponen dos métodos. El primero consiste en emplear una RNA de fusión para estimar la biomasa tomando como entradas las estimaciones de la etapa anterior. El segundo método consiste en entrenar un RNA tipo TDNN (Time Delay Neural Network) para hacer un pronóstico de biomasa un paso adelante para elegir la mejor de las estimaciones de la primera etapa. A éste último método le llamamos sistema de análisis temporal (SAT). En ambos casos de estudio se ha observado que, dado el entrenamiento apropiado, la metodología de fusión de datos a través de RNA es capaz de estimar la biomasa con una precisión comparable a errores de instrumentación.

Más aun, los resultados muestran que las estimaciones son casi insensibles a ruido y a fallas en los sensores debido a la integración y redundancia disponibles. Esta será una de las principales ventajas que provean los sensores virtuales.

6. BIBLIOGRAFIA

[1] C. Atkinson, M. Traver, T. Long, E. Hanzevack, "Virtual Sensors. Areal-time neural network-based intelligent diagnostics http://www.cemr.wvu.edu/~virtsens/virtsens.pdf