(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 426 357**

(51) Int. CI.:

G05B 13/02 (2006.01)
G06N 3/04 (2006.01)
F02C 9/00 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **26.08.2008 E 08787465 (7)**

(97) Fecha y número de publicación de la concesión europea: **31.07.2013 EP 2185980**

(54) Título: **Procedimiento para el control y/o regulación apoyados por ordenador con ayuda de redes neuronales**

(30) Prioridad:

06.09.2007 DE 102007042440

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
22.10.2013

(73) Titular/es:

SIEMENS AKTIENGESELLSCHAFT (100.0%)
WITTELSBACHERPLATZ 2
80333 MÜNCHEN, DE

(72) Inventor/es:

SCHNEEGASS, DANIEL y
UDLUFT, STEFFEN

(74) Agente/Representante:

ZUAZO ARALUZE, Alexander

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín europeo de patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre concesión de Patentes Europeas).

DESCRIPCIÓN

Procedimiento para el control y/o regulación apoyados por ordenador con ayuda de redes neuronales.

La invención se refiere a un procedimiento para el control y/o regulación apoyados por ordenador de un sistema técnico y al correspondiente producto de programa de ordenador.

Cuando se controlan sistemas técnicos complejos se desea a menudo que las acciones a realizar en el sistema técnico estén elegidas tal que se obtenga un comportamiento dinámico ventajoso deseado del sistema técnico. No obstante el comportamiento dinámico de sistemas técnicos complejos a menudo no puede predecirse con facilidad, por lo que se necesitan los correspondientes procedimientos de predicción apoyados por ordenador para estimar el comportamiento futuro del sistema técnico y elegir correspondientemente acciones adecuadas para regular o controlar el sistema técnico.

Hoy en día se basa el control de sistemas técnicos a menudo en los conocimientos de expertos, es decir, la regulación automática del sistema se confecciona sobre la base de este conocimiento de expertos. Pero se conocen también planteamientos en los que mediante la ayuda de métodos conocidos del llamado aprendizaje por refuerzo se controlan sistemas técnicos. No obstante los procedimientos conocidos no pueden aplicarse en general a cualesquiera sistemas técnicos y a menudo no aportan resultados suficientemente buenos.

Por BRAM BAKKER “Reinforcement learning by backpropagation through an LSTM model/critic” (Aprendizaje por refuerzo mediante retropropagación en base a un modelo/crítica LSTM), APPROXIMATE DYNAMIC PROGRAMMING AND REINFORCEMENT LEARNING (Programación dinámica aproximada y aprendizaje por refuerzo), 2007. ADPR L 2007. IEEE INTERNATIONAL SYMPOSIUM ON, IEEE, Pl. 1 abril 2007 (2007-04-01), páginas 127-134, XP031095237, ISBN: 978-1-4244-0706-4 se conoce un procedimiento para el control y/o regulación apoyados por ordenador de un sistema técnico en el que el comportamiento dinámico del sistema técnico para varios instantes se caracteriza en cada caso por un estado del sistema técnico y una acción realizada en el sistema técnico, conduciendo la correspondiente acción en un instante determinado a un estado secuencial del sistema técnico en el siguiente instante. En este procedimiento se aprende una regla de elección de acciones con un conjunto de bloques de datos, llevando cada bloque de datos asociada una valoración. El aprendizaje de la regla de elección de acciones incluye la modelización de una función de valor mediante una primera red neuronal, que incluye los estados y acciones del sistema técnico como parámetros. El aprendizaje de las reglas de elección de acciones incluye además el aprendizaje de la primera red neuronal en base a un criterio de optimalidad que depende de las valoraciones de los bloques de datos y de la función de valor, modelizándose una acción óptima con respecto a la función de valor mediante una segunda red neuronal. El sistema técnico se regula y/o controla tal que las acciones a realizar en el sistema técnico se eligen con la regla de elección de acciones aprendida basándose en la segunda red neuronal (π) aprendida.

Por lo tanto es tarea de la invención lograr un procedimiento para el control y/o regulación apoyados por ordenador de un sistema técnico que puede aplicarse con validez general a cualesquiera sistemas técnicos y que aporta buenos resultados.

En el procedimiento correspondiente a la invención se considera el comportamiento dinámico de un sistema técnico en varios instantes, caracterizándose el comportamiento dinámico en cada instante por un estado del sistema técnico y una acción realizada en el sistema técnico, conduciendo la correspondiente acción en el respectivo instante a un estado secuencial del sistema técnico en el siguiente instante.

Para lograr un control y/o regulación óptimos del sistema técnico, se aprende una regla de elección de acciones sobre la base de bloques de datos, incluyendo cada bloque de datos el estado del sistema técnico en el correspondiente instante, la acción realizada en ese instante, así como el estado secuencial, estando asociada a cada bloque de datos una valoración.

El estado del sistema técnico es aquí en particular un vector de estado con una o varias variables, siendo las variables por ejemplo magnitudes de estado observadas en el sistema técnico. Análogamente puede también estar compuesta una acción a realizar en el sistema técnico por el correspondiente vector con un conjunto de variables de acción, representando las variables de acción en particular parámetros ajustables en el sistema técnico.

El procedimiento correspondiente a la invención se caracteriza por una variante especial del aprendizaje de la regla de elección de acciones, que incluye las siguientes etapas:

i) modelización de una función de calidad mediante una primera red neuronal, incluyendo los estados y acciones del sistema técnico como parámetros.

ii) aprendizaje de la primera red neuronal sobre la base de un criterio de optimalidad, que depende de las valoraciones de los bloques de datos y de la función de calidad, modelizándose una acción óptima en cuanto a la función de calidad mediante una segunda red neuronal, que se aprende basándose en la función de calidad.

Mediante un tal procedimiento se determina, mediante un aprendizaje adecuado de la primera y segunda redes neuronales, de manera sencilla y efectiva, una regla de elección de acciones óptima, que depende de las valoraciones de los bloques de datos, estando configurada la regla de elección de acciones tal que en un cierto estado siempre se elige la acción con la mejor valoración posible. Con la regla de elección de acciones aprendida se realiza entonces la regulación o control propiamente dichos del sistema técnico, eligiéndose acciones a realizar en el sistema técnico con la regulación de elección de acciones aprendida basándose en la segunda red neuronal aprendida.

Como criterio de optimalidad se utiliza la minimización de un residuo de Bellman modificado, incluyendo el residuo de Bellman modificado una función auxiliar, que depende del correspondiente estado del sistema técnico y de las acciones a realizar en el correspondiente estado. Una posible configuración mejorada de este residuo de Bellman se describe en la descripción detallada de la solicitud. El residuo de Bellman modificado se denomina allí L_{aux}.

Para utilizar este residuo de Bellman modificado en el procedimiento correspondiente a la invención, se modeliza la función auxiliar mediante una tercera red neuronal, que se aprende sobre la base del criterio de optimalidad, formando la tercera red neuronal una red feed-forward (de propagación hacia adelante) con una capa de entrada que incluye el correspondiente estado del sistema técnico y la acción que puede realizarse en el correspondiente estado, una o varias capas ocultas, así como una capa de salida que incluye la función auxiliar. El aprendizaje de esta tercera red neuronal se realiza en el procedimiento correspondiente a la invención en paralelo al aprendizaje de la primera y segunda redes neuronales.

El procedimiento correspondiente a la invención se probó en base a bloques de datos de prueba y se ha comprobado que con el procedimiento se logran resultados muy buenos.

El procedimiento correspondiente a la invención significa una ampliación del procedimiento descrito en el documento [1], siendo el documento [1] una solicitud de patente alemana, entregada por el mismo solicitante que la presente solicitud. El procedimiento correspondiente a la presente invención tiene frente al procedimiento del documento [1] la ventaja de que se utiliza una segunda red neuronal que aprende la acción óptima basándose en la función de calidad, con lo que la regla de elección de acciones aprendida con el procedimiento viene dada de manera sencilla mediante una segunda red neuronal aprendida, con la que partiendo de un estado del sistema técnico puede calcularse la acción óptima en este estado. Debido a ello el procedimiento no queda limitado a acciones discretas, sino que la segunda red neuronal puede modelizar en particular también acciones continuas. Además, puede incrementarse mediante el procedimiento correspondiente a la invención la eficiencia de datos, es decir, ya con una reducida cantidad de bloques de datos pueden lograrse buenos resultados para el control y/o regulación adecuados del sistema técnico basándose en un criterio de optimalidad.

En una forma de ejecución preferente del procedimiento correspondiente a la invención se modeliza la función de calidad mediante la primera red neuronal tal que se adapta una función de valoración a las valoraciones de los bloques de datos.

En otra configuración mejorada del procedimiento correspondiente a la invención se fija la acción óptima con referencia a la función de calidad, la cual se modeliza mediante la segunda red neuronal tal que la acción óptima maximiza la función de calidad.

En una forma de ejecución preferente del procedimiento correspondiente a la invención, constituye la primera red neuronal una red feed-forward con una capa de entrada que incluye el correspondiente estado del sistema técnico y la acción a realizar en el correspondiente estado, una o varias capas ocultas, así como una capa de salida que incluye la función de calidad. Análogamente está configurada preferiblemente la segunda red neuronal igualmente como una red feed-forward, incluyendo esta red feed-forward las siguientes etapas:

- una capa de entrada que incluye el correspondiente estado secuencial del sistema técnico;
- una o varias capas ocultas con variables ocultas;
- una capa de salida que incluye la acción óptima en el estado secuencial con respecto a la función de calidad.

Las redes feed-forward antes citadas se denominan también perceptrones multicapa y son estructuras de redes neuronales artificiales conocidas desde hace mucho tiempo por el estado de la técnica.

Para aprender la primera y segunda red neuronal en el procedimiento correspondiente a la invención se utiliza preferiblemente el procedimiento de retropropagación (backpropagation) conocido desde hace mucho tiempo por el estado de la técnica.

El criterio de optimalidad puede elegirse en el procedimiento correspondiente a la invención distinto, utilizándose preferiblemente aquel criterio de optimalidad que parametriza un comportamiento dinámico óptimo del sistema técnico. Posibles criterios de optimalidad son por ejemplo la minimización del residuo de Bellman o bien el haber alcanzado el punto fijo de la iteración de Bellman. El residuo de Bellman o bien la iteración de Bellman son

conocidos al especialista en el campo del aprendizaje por refuerzo (en inglés Reinforcement Learning) y por lo tanto no se describirán aquí.

En una forma de ejecución especialmente preferente del procedimiento correspondiente a la invención incluye el criterio de optimalidad un parámetro que puede ajustarse de nuevo, mediante cuya modificación se adapta el criterio de optimalidad. De esta manera se logra una posibilidad flexible de adaptar el procedimiento correspondiente a la invención al criterio de optimalidad más apropiado para el bloque de datos predeterminado.

En otra forma de ejecución del procedimiento correspondiente a la invención puede también tenerse en cuenta de la forma adecuada la historia de estados y acciones precedentes del sistema técnico. Esto se realiza siendo los estados estados del sistema técnico ocultos en los bloques de datos, que se generan mediante una red neuronal recurrente con ayuda de bloques de datos de origen, incluyendo los bloques de datos de origen en cada caso un estado observado del sistema técnico, una acción realizada en el sistema observado, así como el estado secuencial que de ello resulta. Con la red neuronal recurrente se modeliza en particular el comportamiento dinámico del sistema técnico, estando formada la red neuronal recurrente por al menos una capa de entrada que incluye los estados observados del sistema técnico y las acciones realizadas en el sistema técnico, al menos una capa recurrente oculta que incluye los estados ocultos del sistema técnico y al menos una capa de salida que incluye los estados observados del sistema técnico. La red neuronal recurrente se aprende de nuevo con un procedimiento de aprendizaje adecuado, en particular también mediante el procedimiento de retropropagación (back-propagation) conocido.

Con el procedimiento correspondiente a la invención pueden controlarse y regularse cualesquiera sistemas técnicos, siendo un campo de aplicación preferente las turbinas, en particular turbinas de gas. En el control y/o regulación de una turbina de gas son los estados del sistema técnico y/o las acciones que pueden realizarse en los correspondientes estados una o varias de las siguientes magnitudes:

> Potencia total de la turbina de gas; una o varias presiones y/o temperaturas en la turbina de gas o en el entorno de la turbina de gas; aceleraciones en la cámara de combustión de la turbina de gas; uno o varios parámetros de ajuste en la turbina de gas, en particular ajustes de las válvulas y/o relaciones de carburante y/o posiciones de los álabes directores previos.

Junto al procedimiento antes descrito, se refiere la invención además a un producto de programa de ordenador con un código de programa memorizado sobre un soporte legible por máquina para realizar el procedimiento correspondiente a la invención, cuando el programa corre sobre un ordenador.

A continuación se describen en detalle ejemplos de ejecución de la invención en base a las figuras adjuntas.

Se muestra en:

figura 1 una representación esquemática de una primera forma de ejecución del procedimiento correspondiente a la invención; y
figura 2 una representación esquemática de una segunda forma de ejecución del procedimiento correspondiente a la invención.

Las formas de ejecución descritas a continuación del procedimiento correspondiente a la invención se basan en un conjunto de bloques de datos que ha sido observado, es decir, medido y determinado experimentalmente, para cualquier sistema técnico. Un caso de aplicación especialmente preferente de un sistema técnico es al respecto el control de una turbina de gas, para el que existen datos en forma de magnitudes de estado de la turbina, por ejemplo la potencia total de la turbina de gas, una o varias presiones y/o temperaturas en la turbina de gas, aceleraciones en la cámara de combustión y similares. Se encuentran al respecto bloques de datos relativos a múltiples instantes consecutivos, estando caracterizado cada bloque de datos por un estado, que usualmente es un vector de estado con un conjunto de magnitudes de estado, por una acción que representa la variación de magnitudes de estado o por otros parámetros del sistema técnico ajustables, así como mediante un estado secuencial, que reproduce los valores de las magnitudes de estado tras realizarse la acción. Además se dispone para cada bloque de datos de una valoración o premio, que reproduce la calidad de la acción en el correspondiente instante para el control del sistema técnico. La valoración esta configurada aquí preferiblemente tal que se logra el mejor o bien óptimo control del sistema técnico mediante acciones que tienen elevadas valoraciones o premios en los distintos instantes durante el funcionamiento del sistema técnico.

En las formas de ejecución descritas a continuación del procedimiento correspondiente a la invención se aprende directamente en base a los bloques de datos observados del sistema técnico, mediante un procedimiento de aprendizaje por refuerzo (en inglés Reinforcement Learning), una regla de elección de acciones, con la que a continuación puede operarse el sistema técnico de manera adecuada. La regla de elección de acciones indica entonces para un estado del sistema técnico cuál es la mejor acción a realizar en este estado. El sistema técnico se considera aquí como un sistema dinámico estocástico, considerándose el procedimiento de aprendizaje por refuerzo

para averiguar la regla de elección de la acción como una tarea de regresión, en la que se adapta una función de premio a los bloques de datos observados.

En el procedimiento de aprendizaje descrito a continuación se busca la regla de elección de acciones que puede utilizarse óptimamente para controlar el sistema técnico. Matemáticamente se consideran aquí los estados, las acciones y los estados secuenciales como observaciones de un llamado proceso de decisión de Markov. Un proceso de decisión de Markov viene en general dado por un espacio de estados S, un conjunto de acciones A que pueden elegirse en los distintos estados y la dinámica, que se considera como distribución de probabilidades de transición P_T: S X A X S → [0, 1], que depende del estado en ese momento s, de la acción elegida a y del estado secuencial s'. La transición de un estado al estado secuencial se caracteriza mediante los llamados premios R(s,a,s'), que son funciones del estado en ese momento, de la acción y del estado secuencial. Los premios se definen mediante una distribución de probabilidades de premio P_R con el valor de esperanza del premio

$$\mathbf{E}R = \int_{\Re} rP_R(s,a,s',r)dr, s,s' \in S, a \in A$$

Según la forma de ejecución descrita a continuación del procedimiento correspondiente a la invención, se busca el máximo de una función Q de descuento, que corresponde a la función de calidad en el sentido de las reivindicaciones y que está definida mediante la ecuación de Bellman conocida desde hace mucho tiempo por el estado de la técnica como sigue:

$$Q^{\pi}(s,a) = \mathbf{E}_{s'}\left(R(s,a,s') + \gamma Q^{\pi}(s',\pi(s'))\right) \qquad (1).$$

La maximización se realiza aquí en el llamado espacio de regulación Π = (S → A) en todos los estados s y acciones a posibles, siendo 0<γ<1 el factor de descuento, s' el estado secuencial de s y πεΠ la regla de elección de acción utilizada. La maximización se realiza según la invención aquí descrita con un procedimiento de regresión basado en redes neuronales, que utiliza un gradiente basado en la regla de elección de acción óptima (es decir, en la regla de elección que maximiza la función Q) y que se denomina también "Policy Gradient Neural Rewards Regression" o política de regresión de gradientes. Al respecto no se buscan explícitamente acciones discretas - como en el procedimiento correspondiente al documento [1] - que maximizan la función de calidad. En lugar de ello se utiliza la acción ya previamente tomada como óptima como entrada para la función Q, calculándose la acción óptima basándose en una red feed-forward neuronal. La arquitectura del procedimiento utilizado se muestra en la figura 1 y se describirá más abajo.

En las formas de ejecución descritas a continuación del procedimiento correspondiente a la invención se considera un sistema técnico en el que tanto los estados del sistema como también la acciones que pueden realizarse en el correspondiente estado son continuas. La dinámica del sistema es aquí probabilística.

En las formas de ejecución de la figura 1 y figura 2 se describe la función Q antes descrita mediante una única red feed-forward o bien un perceptrón multicapa *N(s,a)=Q(s,a)*, que incluye una capa de entrada I con el estado s y la acción a, una capa oculta H, así como una capa de salida O con la función de calidad Q. Para modelizar la función de premio según la ecuación (1), se utiliza el algoritmo de retropropagación conocido desde hace mucho tiempo por el estado de la técnica, resolviéndose el problema de la optimización minimizando un residuo de Bellman modificado (cuadrado) en todas las 1 transiciones observadas. 1 es así la cantidad de bloques de datos en los datos observados del sistema técnico. El residuo de Bellman modificado utilizado en la forma de ejecución aquí descrita se basa en el documento [2] y es una mejor aproximación del residuo de Bellman auténtico. Se denomina también a continuación residuo auxiliar L_{aux} y es como sigue:

$$L_{aux} = L - \ddot{L} + \Omega(\theta) =$$
$$= \sum_{i=1}^{l} (Q(s_i,a_i) - \gamma Q(s_{i+1},\pi(s_{i+1})) - r_i)^2 - \sum_{i=1}^{l} (h(s_i,a_i) - \gamma Q(s_{i+1},\pi(s_{i+1})) - r_i)^2 + \Omega(\theta)$$

(2)

Aquí representa (θ) los parámetros de la red feed-forward neuronal artificial N(s,a) e incluye en particular las matrices de ponderación entre las distintas capas de neuronas en la red feed-forward. Q es un sistema de regularización adecuado. r_i es el premio o valoración observado en un estado s_i a partir de los bloques de datos y s_{i+1} son estimadores fiables de las magnitudes de estado correspondientes al estado secuencial.

Se sabe que la minimización del residuo de Bellman por un lado tiene la ventaja de que representa un problema de aprendizaje que puede controlarse bien, ya que se aplica con el esquema del aprendizaje vigilado. Por otro lado,

tiende la minimización del residuo de Bellman a minimizar términos de orden superior de la suma descontada de los premios futuros en el caso estocástico, cuando no pueden darse otros bloques de datos no correlados para cada transición. En general están las soluciones para las funciones Q afectadas por prejuicios, que son más suaves para estados secuenciales de las transiciones estocásticas. Cuando S_{i+i} y r_i son estimaciones fiables para los siguientes estados o premios, entonces no es la expresión $(Q(s_i,a_i)-\gamma V(s_{i+1})-r_i)^2$ una estimación fiable del residuo de Bellman cuadrado $(Q(s,a)-(TQ)(s,a))^2$ auténtico, sino de $(Q(s,a)-(TQ)(s,a))^2+(T'Q)(s,a)^2$. T y T' están definidos aquí como sigue:

$$(TQ)(s,a)=\mathbf{E}_{s'}\left(R(s,a,s')+\gamma \max_{a'} Q(s',a')\right)$$

$$(T'Q)(s,a)^2=\mathbf{Var}_{s'}\left(R(s,a,s')+\gamma \max_{a'} Q(s',a')\right)$$

T se denomina aquí también operador Bellman.

Como alternativa a la utilización de dobles trayectorias, se utiliza el residuo de Bellman modificado antes indicado del documento [2] como mejor aproximación al residuo de Bellman auténtico. La tarea de optimización es así la solución

$$\hat{Q}=\arg\min_{Q\in H_Q}\max_{h\in H_h} L_{aux}\ .$$

La idea del residuo de Bellman modificado es el hallazgo de un h que realice la aproximación al operador de Bellman a lo largo de las observaciones.

Se obtiene así:

$$\begin{aligned} Z&=\mathbf{E}_{s'}\left(Q(s,a)-\gamma V(s')-R(s,a,s')\right)^2-\mathbf{E}_{s'}\left(h(s,a)-\gamma V(s')-R(s,a,s')\right)^2\\ &=\left(Q(s,a)-(TQ)(s,a)\right)^2-\mathbf{Err}\left(h(s,a),(TQ)(s,a)\right).\end{aligned}$$

Se trata al respecto de la función de pérdidas auténtica con un término de errores adicional debido a la aproximación sub-óptima de h, en el caso de que H_h no esté en situación de aproximarse al operador de Bellman con la exactitud que se desee.

Esta técnica posibilita limitar el residuo de Bellman auténtico hacia arriba, en el caso de que el error de h con respecto a TQ pueda limitarse. Se observa fácilmente que $\hat{L}\leq L$ rige dentro de un punto de silla de L_{aux}, en el caso de que $H_Q = H_h$. Caso contrario no proporcionaría h el mínimo de $\hat{L}$. Por ello se aportaría un óptimo de L_{aux} mediante cada punto fijo de la iteración de Bellman, en el caso de que este punto exista, ya que sólo en este caso Q puede aproximarse al operador Bellman tan bien como h y es $L_{aux} = 0$. A diferencia de la propuesta del documento [2], se elegiría en la forma de ejecución aquí descrita de la invención H_h bien como una clase de función claramente más potente que H_Q o bien, considerando el conocimiento previo acerca del operador de Bellman auténtico, con lo que $\hat{L}$ no aporta esencialmente una estimación mejor de $T'Q^2$. Puesto que una tal estimación de la varianza sigue sin ser fiable, converge el procedimiento sobre un estimador no fiable del residuo de Bellman auténtico, que minimiza la función $\hat{Q}^*\in H_Q$ sólo dentro del espacio de la función, pero evidentemente aportan una mejor aproximación que los estimadores conocidos por el estado de la técnica.

A partir del antes citado residuo de Bellman L_{aux} según la ecuación (2), resultan los siguientes gradientes $\Delta\theta$, $\Delta\omega$ y $\Delta\psi$, que representan las derivadas del residuo L_{aux} según θ, ω y ψ respectivamente y cuyo punto cero ha de determinarse para solucionar la tarea de optimización:

$$\Delta\theta = \sum_{i=1}^{I}\Big(\big(Q(s_i,a_i)-\gamma Q(s_{i+1},\pi(s_{i+1}))-r_i\big)\cdot \frac{d}{d\theta}\big(Q(s_i,a_i)-\rho\gamma Q(s_{i+1},\pi(s_{i+1}))\big)-\frac{\varepsilon d}{d\theta}Q(s_{i+1},\pi(s_{i+1})) + \beta\rho\gamma\big(h(s_i,a_i)-\gamma Q(s_{i+1},\pi(s_{i+1}))-r_i\big)\frac{d}{d\theta}Q(s_{i+1},\pi(s_{i+1}))\Big)+\frac{d\Omega(\theta)}{d\theta} \qquad (3)$$

$$\Delta\omega = \alpha\beta\sum_{i=1}^{I}\big(h(s_i,a_i)-\gamma Q(s_{i+1},\pi(s_{i+1}))-r_i\big)\frac{d}{d\omega}h(s_i,a_i) \qquad (4)$$

$$\Delta\psi = -\sum_{i=1}^{I}\frac{d}{\varepsilon d\psi}\pi(s_{i+1})\frac{\varepsilon d}{d\pi(s_{i+1})}Q(s_{i+1},\pi(s_{i+1}))=-\sum_{i=1}^{I}\frac{d}{d\psi}Q(s_{i+1},\pi(s_{i+1})) \qquad (5)$$

ω son aquí los correspondientes parámetros que describen la función auxiliar h, que se modeliza como red feed-forward, $0 \le \beta \le 1$ sirve para el control de la influencia de la función auxiliar h y $\alpha \ge 1$ es la magnitud de la optimización de h en comparación con Q. ψ es el parámetro de una red feed-forward π (figura 1), que se utiliza para modelizar la regla de elección óptima. En lugar de este residuo de Bellman modificado L_{aux}, puede utilizarse también el residuo Bellman normal o bien el que se llegue al punto fijo de la ecuación de Bellman para la solución del problema de optimización. En este caso no existe la anterior ecuación (4) y en la ecuación (3) no entra el término que comienza con los factores βργ. El parámetro β permite también una transición deslizante. Se obtiene entonces para ρ = 1 el criterio de optimalidad según la minimización clásica del residuo de Bellman, y por el contrario para ρ = 0 se obtiene la llegada al punto fijo de la iteración de Bellman.

En la forma de ejecución aquí descrita se alcanza la función h en la arquitectura de la figura 1 mediante introducción de la red auxiliar AN junto a la red principal MN. La red auxiliar AN modeliza la función h con una red neuronal, que análogamente a la red N(s,a) presenta una capa de entrada I que incluye el estado s y la acción a, una capa oculta H y una capa de salida O que incluye la función h. La red auxiliar AN cumple dos tareas, maximiza $\hat{L}$ con respecto a θ y minimiza $\hat{L}$ con respecto a ω. La red principal MN, así como la red auxiliar AN, se aprenden aquí simultáneamente.

La arquitectura de la red principal MN representada en la figura 1 reproduce la política de regresión de gradiente utilizada en la invención. La red principal MN esta compuesta por una parte izquierda, que modeliza la función de calidad Q como una única red neuronal N(s,a). Los estados s y las acciones a son aquí continuos. La parte izquierda de la red está acoplada con la parte derecha mediante la función de premio R, controlándose el flujo de gradiente mediante los parámetros -γ y ρ, que se presentan en la ecuación de gradientes (3) anterior. En la parte derecha de la red principal MN se encuentra en los ramales 1 y 2 la regla de elección de acciones óptima P como red neuronal π con los correspondientes parámetros ψ, así como la red neuronal N(s',a') para el estado secuencial. Los parámetros ψ de la red neuronal π se aprenden con referencia a la red de elección óptima P. La regla de elección óptima es aquí la maximización de la función de calidad Q. V' en la fig. 1 significa aquí V(s') = Q(s',π(s')) y refleja la función Q máxima para el estado secuencial s' y la acción secuencial a'. El ramal 2 de la parte derecha de la red principal MN muestra la maximización de la función Q con referencia al término ε de la ecuación (5) anterior. Un ε suficientemente pequeño posibilita que la regla P sólo contribuya a maximizar la función Q.

Según la figura 1 resulta así un procedimiento en el que una única red neuronal N(s,a) se utiliza para estados s y acciones a continuos. Para los estados secuenciales se calcula Q(s',π(s'), suponiéndose que π: S → A con los parámetros ψ tiende a ejecutar la regla P óptima, es decir, a maximizar la función Q. Así se encuentra Q (s',π(s') en las proximidades de $\max_{a'}Q(s',a')$. Esto se logra mediante la maximización de la función Q para los estados secuenciales a la vez que la regresión con respecto a los premios. Así se ejecuta una clase de Batch On-Policy-Iteration o una Batch-Actor-Critic-Iteration, en las que se utiliza la iteración intrínseca entre la función Q y la regla P. Se utiliza entonces la técnica de control de flujo de gradiente conocida por el estado de la técnica junto con Shared Weights (ponderaciones compartidas), siendo esta técnica suficiente para construir la arquitectura adecuada. En la parte de la red para el estado secuencial que se representa en la figura 1 con la referencia 1, corta la red el flujo de gradientes, lo que se representa con el número 0 en la flecha discontinua orientada hacia la acción a'. Con ello no influye la regla P sobre la regresión relativa a los premios. En una parte de red ampliada, que en la figura 1 se designa con la referencia 2, se posibilita mediante un ε suficientemente pequeño que sólo la regla P contribuya a la

maximización de la función Q. Para calcular la regla de elección óptima π se utiliza un procedimiento de retropropagación con shared weights. Además se utiliza la técnica de bloqueo del reflujo de gradientes, desarrollada por los inventores y ya publicada. La técnica de bloqueo del reflujo de gradientes pertenece así al estado de la técnica y es conocida al especialista desde hace mucho tiempo. Por lo tanto no se describirá la misma aquí más en detalle. Tras la convergencia del procedimiento de retropropagación, puede determinarse la regla de elección óptima con ayuda de la red feed-forward π, sin utilizar la función Q como resultado intermedio.

El procedimiento antes descrito no tiene en cuenta la historia de los estados precedentes, lo cual da lugar a que no pueda garantizarse la forma de funcionamiento si no existe un proceso de decisión de Markov. No obstante, en una segunda forma de ejecución del procedimiento correspondiente a la invención puede considerarse también esta historia. Esto sucede debido a que el propio bloque de datos que se utiliza para aprender las redes neuronales fue generado a su vez a partir de un bloque de datos de origen. El bloque de datos de origen es aquí el bloque de datos que en la forma de ejecución de la figura 1 se ha introducido directamente para aprender las redes neuronales. En la segunda forma de ejecución que se muestra en la figura 2 se utiliza este bloque de datos de origen directamente para modelizar el comportamiento dinámico del sistema técnico con una red neuronal recurrente RNN, que incluye una capa de entrada I', una capa oculta H' y una capa de salida O'. Esta red neuronal incluye los estados internos x_t, Z_t, **t** ε **t**-τ,…t+1, así como las matrices de ponderación F, G, H, J. Las transiciones entre los estados están definidas aquí mediante las siguientes ecuaciones:

$$\mathbf{x}_t = \tanh(F\mathbf{s}_t + J\mathbf{z}_{t-1})$$

$$\mathbf{z}_t = G a_t + H\mathbf{x}_t$$

Mediante una matriz M, que reproduce el estado interno sobre el externo, puede lograrse el estado secuencial cumpliendo la siguiente condición:

$$\left\| M\mathbf{z}_t - \mathbf{s}_{t+1} \right\|^2 = \min.$$

Con algoritmos conocidos se determinan según las ecuaciones anteriores los parámetros de la red recurrente neuronal (es decir, las matrices de ponderación de la red) tal que la red neuronal recurrente genera muy bien los bloques de datos observados en el correspondiente instante. La red neuronal recurrente se aprende entonces de nuevo mediante un procedimiento de retropropagación conocido desde hace mucho tiempo por el estado de la técnica. La modelización del comportamiento dinámico mediante la red neuronal recurrente RNN le es conocido al especialista desde hace mucho tiempo y por lo tanto no se describirá en detalle. A diferencia del procedimiento de la figura 1, se introducen ahora los estados ocultos x_t y x_{t+1} como magnitudes de entrada en la capa de entrada I en lugar de los estados s y s' respectivamente. Por lo demás, corresponde el procedimiento de aprendizaje de la figura 2 al procedimiento de aprendizaje de la figura 1, lo cual queda patente en que la parte por encima de la red neuronal recurrente RNN es idéntica a la de la figura 1. Por lo tanto renunciamos a describir de nuevo el aprendizaje de las redes neuronales. En la red neuronal recurrente de la figura 2 se introduce además un parámetro μ, con el que se adapta la velocidad de aprendizaje de la red neuronal recurrente en comparación con la velocidad de aprendizaje de las distintas redes feed-forward. Además se utilizan matrices N, que describiremos posteriormente.

La arquitectura mostrada en la figura 2 posibilita utilizar sólo una red neuronal recurrente tanto para la función Q como también para la función de valores. La red neuronal recurrente está aquí en condiciones de realizar la aproximación a un sistema dinámico determinístico con la exactitud que se desee, lo cual no es posible para un sistema dinámico estocástico. No obstante esto no significa ninguna limitación, ya que la estructura del estado interno oculto puede considerarse como una transformación en un espacio de características adecuado para el procedimiento de aprendizaje por refuerzo estocástico mediante las redes feed-forward. En el caso determinístico este espacio de características es idéntico a la descripción exacta de toda la información que determina el futuro y por el contrario en el caso estocástico general tiene que generarse el estado interno tal que puede predecirse el futuro esperado. Por ello se utiliza en la red neuronal recurrente de la figura 2 una componente de futuro autónoma cuando pueden predecirse estados internos sin observaciones procedentes del proceso de Markov. En la fig. 2 se trata al respecto de los estados $\mathbf{x}_t$ para t>i+1 y rige:

$$\mathbf{x}_t = \tanh(N\mathbf{z}_{t-1}), t > i+1 .$$

Esto es suficiente, ya que la característica de Markov puede ser reconstruida mediante el conocimiento sobre los estados futuros esperados. Mediante la arquitectura recurrente antes descrita de la figura 2 es posible la reconstrucción automática de un proceso de Markov tomado como base, en principio de cualquier orden. Quien

aplica el procedimiento correspondiente a la invención puede así aprovechar más informaciones procedentes de los datos y tiene que proporcionar menos informaciones de expertos.

Las formas de ejecución antes descritas correspondientes a las figuras 1 y 2 del procedimiento correspondiente a la invención se probaron en base a las problemáticas conocidas por el estado de la técnica. Se ha comprobado al respecto que el procedimiento aporta buenos resultados. En particular se comparó el procedimiento también con el procedimiento según el documento [1], pudiendo lograrse mejores resultados con el procedimiento correspondiente a la invención.

El procedimiento antes descrito ofrece un planteamiento de solución eficiente en cuanto a información para problemas de control óptimo generales en cualesquiera sectores técnicos, siendo posible al respecto solucionar con pocos bloques de datos disponibles también problemas de control complejos, que con procedimientos clásicos no pueden solucionarse satisfactoriamente.

Bibliografía

(1) Solicitud de patente alemana 10 2007 017 259.3
(2) Andras Antos, Csaba Szepesvari, Remi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. In Proc. of the Conference on Learning Theory, páginas 574-588, 2006.

REIVINDICACIONES

1. Procedimiento para el control y/o regulación apoyados por ordenador de un sistema técnico, en el que:
   a) el comportamiento dinámico del sistema técnico en varios instantes (t) se caracteriza en cada caso por un estado (s_t, x_t) del sistema técnico y una acción (a_t) realizada en el sistema técnico, conduciendo la correspondiente acción (a_t) en el correspondiente instante (t) a un estado secuencial (s_{t+1}, x_{t+1}) del sistema técnico en el siguiente instante (t+1);
   b) se aprende una regla de elección de acciones con un conjunto de bloques de datos, incluyendo cada bloque de datos el estado (s_t, x_t) del sistema técnico en el correspondiente instante (t), la acción (a_t) realizada en el estado (s_t, x_t), así como el estado secuencial (s_{t+1}, x_{t+1}) y estando asociada a cada bloque de datos una valoración (r_i), incluyendo el aprendizaje de la regla de elección de acciones las siguientes etapas:
      i) modelización de una función de calidad (Q(s, a)) mediante una primera red neuronal (N_a, (s)), incluyendo los estados (s_t, x_t) y acciones (a_t) del sistema técnico como parámetros;
      ii) aprendizaje de la primera red neuronal (N(s,a) sobre la base de un criterio de optimalidad, que depende de las valoraciones (r_i) de los bloques de datos y de la función de calidad (Q(s,a)), modelizándose una acción óptima (a') en cuanto a la función de calidad (Q(s,a)) mediante una segunda red neuronal (π), que se aprende basándose en la función de calidad;
   c) el sistema técnico se regula y/o controla eligiendo las acciones (a_t) a realizar en el sistema técnico con la regla de elección de acciones aprendida basándose en la segunda red neuronal (π) aprendida;

   caracterizado porque

   el criterio de optimalidad es la minimización de un residuo de Bellman modificado, incluyendo el residuo de Bellman modificado una función auxiliar (h(s,a)), que depende del estado (s_t, x_t) del sistema técnico y de la acción (a_t) que puede realizarse en el correspondiente estado (s_t, x_t), modelizándose la función auxiliar mediante una tercera red neuronal, que se aprende sobre la base del criterio de optimalidad, formando la tercera red neuronal una red feed-forward (de propagación hacia adelante) con una capa de entrada (I) que incluye el correspondiente estado (s_t, x_t) del sistema técnico y la acción (a_t) que puede realizarse en el correspondiente estado (s_t, x_t), una o varias capas ocultas (H), así como una capa de salida (O) que incluye la función auxiliar (h(s,a)).

2. Procedimiento según la reivindicación 1,
   en el que la función de calidad (Q(s,a)) se modeliza mediante la primera red neuronal (N(s,a)) tal que una función de valoración (R(s,a,s') se adapta a las valoraciones (r_i) de los bloques de datos.

3. Procedimiento según la reivindicación 1 ó 2,
   en el que la acción óptima (a') con respecto a la función de calidad (Q(s,a)) es aquella acción que maximiza la función de calidad (Q(s,a)).

4. Procedimiento según una de las reivindicaciones precedentes,
   en el que la primera red neuronal (N(s,a) constituye una red feed-forward con una capa de entrada (I) que incluye el correspondiente estado (s_t, x_t) del sistema técnico y la acción (a_t) que puede realizarse en el correspondiente estado (s_t, x_t), una o varias capas ocultas (H), así como una capa de salida (O) que incluye la función de calidad (Q(s,a).

5. Procedimiento según una de las reivindicaciones precedentes,
   en el que la segunda red neuronal (π) constituye una red feed-forward, con una capa de entrada (I) que incluye el correspondiente estado secuencial (s_{t+1}, x_{t+1}) del sistema técnico, una o varias capas ocultas (H) así como una capa de salida (O) que incluye la acción óptima (a') en el estado secuencial (s_{t+1}, x_{t+1}) con respecto a la función de calidad (Q(s,a)).

6. Procedimiento según una de las reivindicaciones precedentes,
   en el que se utiliza un procedimiento de retropropagación para aprender la primera red neuronal (N(s,a), así como la segunda red neuronal (π).

7. Procedimiento según una de las reivindicaciones precedentes,
   en el que el criterio de optimalidad está elegido tal que se parametriza un comportamiento dinámico óptimo del sistema técnico.

8. Procedimiento según una de las reivindicaciones precedentes,
   en el que el criterio de optimalidad es la minimización del residuo de Bellman.

9. Procedimiento según una de las reivindicaciones precedentes,
   en el que el criterio de optimalidad es la llegada al punto fijo de la iteración de Bellman.

10. Procedimiento según una de las reivindicaciones precedentes,

en el que el criterio de optimalidad incluye un parámetro (ρ) ajustable, mediante cuya modificación se adapta el criterio de optimalidad.

11. Procedimiento según una de las reivindicaciones precedentes,
en el que un estado (s_t, x_t) del sistema técnico incluye una o varias variables, en particular magnitudes de estado observadas del sistema técnico y/o una acción (a_t) a realizar en el sistema técnico incluye una o varias variables de acción.

12. Procedimiento según una de las reivindicaciones precedentes,
en el que los estados (s_t, x_t) son estados del sistema técnico ocultos (x_t) en los bloques de datos, que se generan mediante una red neuronal (RNN) recurrente con ayuda de bloques de datos de origen, incluyendo los bloques de datos de origen en cada caso un estado observado (s_t) del sistema técnico, una acción (a_t) realizada en el estado (s_t) observado, así como el estado secuencial(s_{t+1}) que de ello resulta.

13. Procedimiento según la reivindicación 12,
en el que se modeliza el comportamiento dinámico del sistema técnico mediante la red neuronal recurrente (RNN), estando formada la red neuronal recurrente (RNN) por al menos una capa de entrada (I') que incluye los estados observados (s_t) del sistema técnico y las acciones (a_t) realizadas en el sistema técnico, al menos una capa recurrente oculta (H) que incluye los estados ocultos (x_t) y al menos una capa de salida (O') que incluye los estados observados (s_t) del sistema técnico.

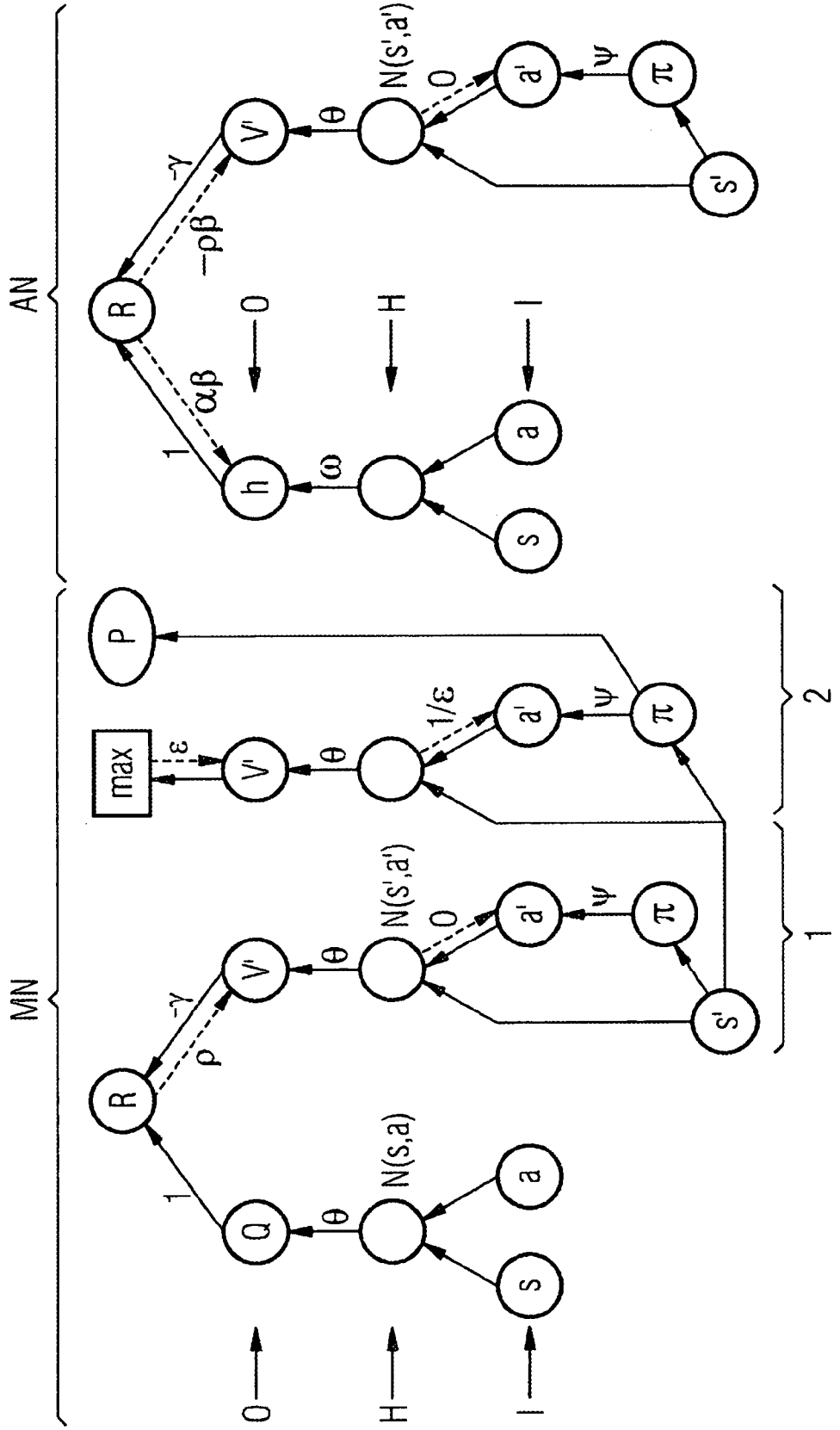
14. Procedimiento según la reivindicación 13,
en el que la red neuronal recurrente (RNN) se aprende con un procedimiento de aprendizaje, en particular un procedimiento de retropropagación (backpropagation).

15. Procedimiento según una de las reivindicaciones precedentes,
en el que el sistema técnico es una turbina, en particular una turbina de gas.

16. Procedimiento según la reivindicación 15,
en el que mediante el procedimiento se regula y/o controla una turbina de gas, incluyendo los estados (s_t, x_t) del sistema técnico y/o las acciones (a_t) que pueden realizarse en los correspondientes estados (s_t, x_t) una o varias de las siguientes magnitudes:
potencia total de la turbina de gas; una o varias presiones y/o temperaturas en la turbina de gas o en el entorno de la turbina de gas; aceleraciones en la cámara de combustión de la turbina de gas; uno o varios parámetros de ajuste en la turbina de gas, en particular ajustes de las válvulas y/o relaciones de carburante y/o posiciones de los álabes directores previos.

17. Programa de ordenador con un código de programa memorizado en un soporte legible por máquina para ejecutar el procedimiento según una de las reivindicaciones precedentes, cuando el programa corre sobre un ordenador.

FIG 1



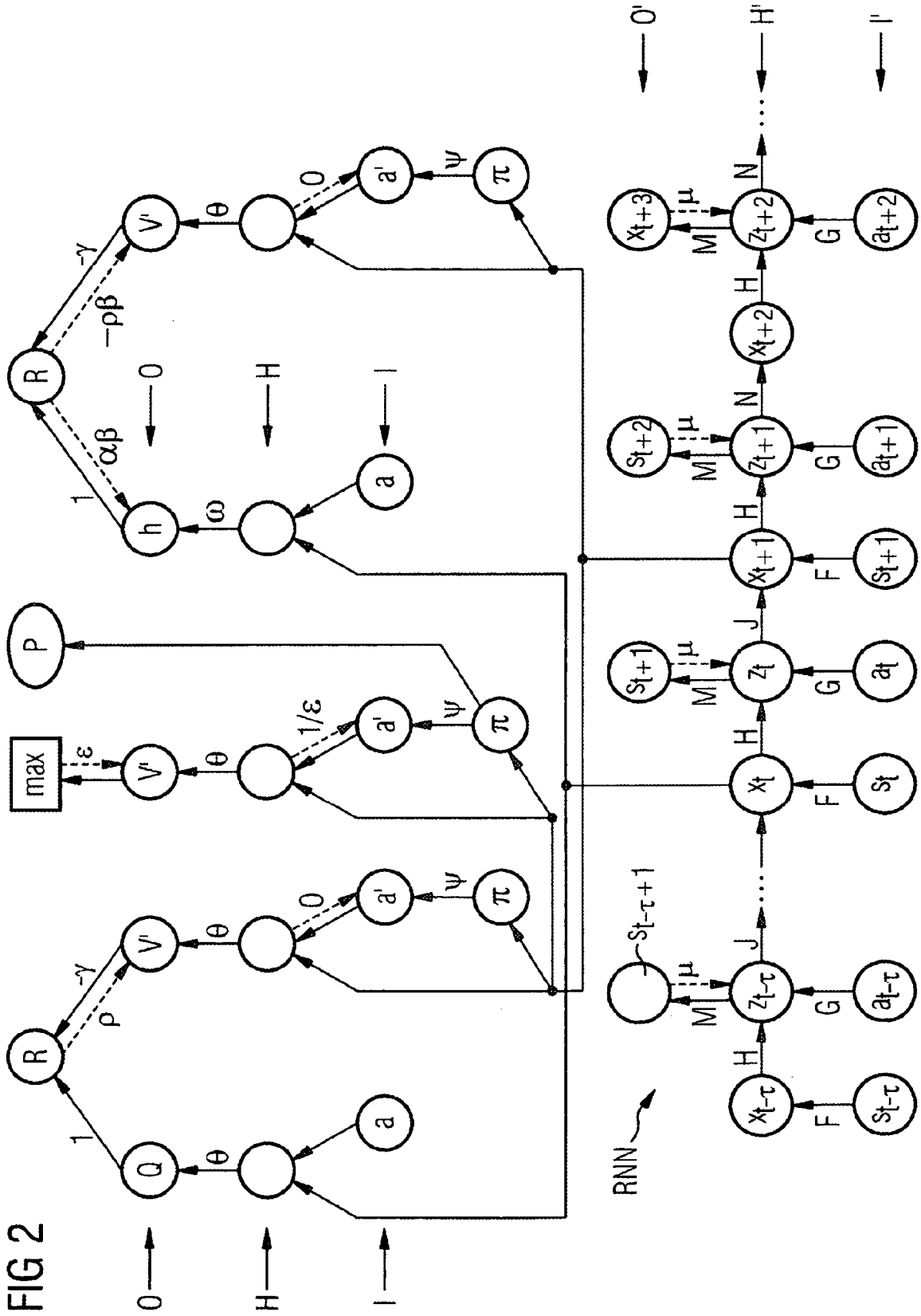


FIG 2