(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 457 891**

(21) Número de solicitud: 201231295

(51) Int. CI.:

G06F 17/10 (2006.01)

(12) PATENTE DE INVENCIÓN B1

(22) Fecha de presentación:

13.08.2012

(43) Fecha de publicación de la solicitud:

29.04.2014

Fecha de la concesión:

29.12.2014

(45) Fecha de publicación de la concesión:

08.01.2015

(56) Se remite a la solicitud internacional:

PCT/ES2013/070590

(73) Titular/es:

UNIVERSITAT DE LES ILLES BALEARS (100.0%)
Campus Universitario. Ctra. de Valldemosa, km. 7,5. Edifici Son Lledo
07071 Palma de Mallorca (Illes Balears) ES

(72) Inventor/es:

ROSELLÓ SANZ, José Luis;
CANALS GUINAND, Vicente J. y
MORRO GOMILA, Antoni

(74) Agente/Representante:

TEMIÑO CENICEROS, Ignacio

(54) Título: **MÉTODO Y SISTEMA DIGITAL PROBABILÍSTICO PARA LA EXPLORACIÓN EFICIENTE DE GRANDES BASES DE DATOS**

(57) Resumen:

Sistema digital probabilístico para la exploración eficiente de grandes bases de datos que comprende una entrada de n+1 vectores, de los cuales 'n' pertenecen a la base de datos y uno es el patrón a identificar vector x_0 (1); donde cada vector consistirá en N números o descriptores definidos por 'm' bits cada uno, y donde cada uno de estos componentes de 'm' bits representarán un descriptor determinado para el objeto que describe cada vector; de tal forma que cada objeto viene determinado por N descriptores distintos codificados cada uno con 'm' bits, y donde las cantidades que describe cada descriptor tendrán que ser normalizadas al rango que abarcan; y donde de la base de datos a escanear se volcarán un total de n vectores (2), cada uno de los cuales se comparará con el vector de referencia x_0 (1) en una pluralidad de comparadores vectoriales (3).

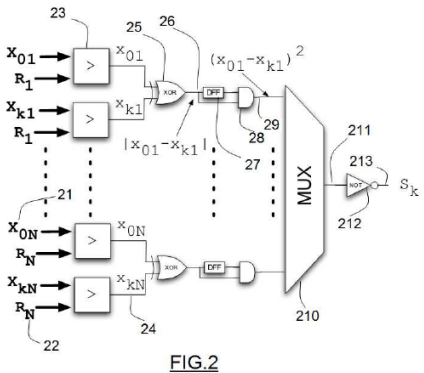

FIG.2

ES 2 457 891 B1

Aviso: Se puede realizar consulta prevista por el art. 37.3.8 LP.

MÉTODO Y SISTEMA DIGITAL PROBABILÍSTICO PARA LA EXPLORACIÓN EFICIENTE DE GRANDES BASES DE DATOS

DESCRIPCIÓN

La presente invención se refiere al diseño de un circuito digital que es capaz de explorar de forma rápida y eficiente grandes bases de datos en la búsqueda de patrones concretos. La presente invención está adaptada para la computación masiva en paralelo y tiene una gran relevancia en muchas áreas de ciencia en donde se han de explorar grandes bases de datos para la extracción de información útil.

Antecedentes de la invención

En [*Gaines, B.R., 1968. "Random pulse machines", IEEE Trans. on Comp. 410*] se describe la idea de aplicar la teoría de probabilidades a los circuitos digitales con el fin de implementar operaciones algebraicas de forma sencilla y altamente inmune al ruido si la comparamos con la lógica digital tradicional. En este documento se utiliza el hecho que, partiendo de señales digitales pulsantes aperiódicas se pueden implementar operaciones aritméticas complejas como la suma o la multiplicación a partir de circuitos digitales sencillos (en estos dos casos dispondríamos de un multiplexor para la suma ponderada y de una puerta AND para la multiplicación). Estas propiedades pueden ser usadas para la implementación de sistemas eficientes de reconocimiento de patrones.

La desventaja de esta forma de utilizar las puertas lógicas se presenta cuando se quiere almacenar o reconvertir la información de cada computación realizada en el sistema al mundo binario. Por cada canal estocástico de información que se desee conocer su actividad (y por tanto conocer el valor numérico que representa) se ha de integrar durante un tiempo suficiente el número de pulsos proporcionados por la señal. Este hecho hace que, en caso de querer suficiente precisión en las mediciones, el tiempo de computación necesario para obtener determinadas operaciones será mucho mayor que el usado por la tecnología digital clásica que se utiliza actualmente. Este hecho hizo abandonar dicha tecnología a mediados de los setenta cuando surgieron los primeros microprocesadores basados en la lógica digital convencional y que proporcionaban altas precisiones en menores tiempos de computación. Otra de las características diferenciadoras de este tipo de lógica es la necesidad de la descorrelación entre las señales que se operan. Esta descorrelación temporal de las señales es necesaria para poder implementar las operaciones aritméticas básicas. Hay que reseñar que, en los principios de la computación, la lógica estocástica se postulaba como una lógica alternativa a la lógica digital tradicional y que por tanto su aplicación es tan generalista como ésta. La principal razón de su entrada en desuso y olvido (durante la década de los 70) se debió a la falta de precisión en las operaciones que se realizaban (debido a su naturaleza probabilística) y debido sobretodo a la necesidad de usar sistemas deterministas de gran precisión para automatizar las actividades económico-financieras tanto en el ámbito doméstico como empresarial, que se correspondía con el principal sector demandante de sistemas de computación.

De esta forma la computación estocástica o probabilística no se desarrollará durante los años siguientes aunque ésta haya sido utilizada últimamente para la implementación de redes neuronales en hardware [*Y. Maeda and Y. Fukuda, "FPGA Implementation of Pulse Density Hopfield Neural Network", In Proc. Int. Joint Conf. on Neural Networks, Florida, USA, 2007; Kondo Y, Sawada, Y, 1992. "Functional Abilities of a Stochastic Logic Neural Network" IEEE Trans. on Neural Networks, 3 (3), 434-443; S. Sato, K. Nemoto, S. Akimoto, M Kinjo and K. Nakajima, "Implementation of a New Neurochip Using Stochastic Logic" IEEE Trans. on Neural Networks, 14 (5), Sept 2003, pp. 1122-1127; S. L. Bade, B. Hutchings. "FPGA-Based Stochastic Neural Networks Implementation" in Proc. IEEE Workshop on FPGAs for Custom Computing Machines, Napa Valley, CA, USA, 1994. pp. 189-198*; *B. Brown, H. Card, "Stochastic Neural Computation I: Computational Elements" IEEE Transactions on Computers, Volume. 50, Issue. 9, pp. 891–905, 2001].*

A medida que la tecnología de integración de los circuitos ha evolucionado desde finales de la década de 1950 hasta la actualidad se han incrementado exponencialmente el número de transistores que se pueden implementar en un solo integrado. Según la conocida ley de Moore, cada dos años la tecnología dobla sus prestaciones. Esto implica por ejemplo que en un período de 20 años (desde principios de los 90 hasta ahora) la tecnología se haya multiplicado por un factor 1000 o más, permitiendo actualmente la integración de miles de millones de transistores en un solo integrado.

Este nuevo escenario abre grandes posibilidades para el uso de lógicas no-deterministas como la lógica estocástica o probabilística en aplicaciones en donde se necesite gran paralelismo y no sea necesario conocer o almacenar los resultados parciales de la computación de cada módulo estocástico que esté operativo. Éste es el caso por ejemplo de los sistemas de reconocimiento de formas en donde a partir de una gran cantidad de información de entrada y de la realización probabilística de muchos procesos en paralelo se determine si determinado estímulo complejo pertenece o no a una categoría concreta. En el caso concreto de la exploración de grandes bases de datos el proceso consiste en identificar el conjunto de vectores de la base de datos que mejor coincidan con el patrón modelo que se esté buscando. De esta forma, y gracias al reducido tamaño de las estructuras lógicas que requieren los elementos computacionales estocásticos o probabilísticos se pueden incrementar considerablemente el número de estos bloques a la vez que la velocidad de proceso de éstos,

posibilitando así la computación masiva en paralelo. Recientes trabajos [*V Canals, A. Morro, J.L. Rosselló, "Stochastic-Based pattern-recognition analysis" Pattern Recognition Letters 20101101 Elsevier*] realzan este hecho y demuestran por ejemplo que para el caso concreto del reconocimiento de patrones la lógica estocástica es ordenes de magnitud más rápida que la lógica tradicional basada en microprocesadores, lo cual puede ser considerado, en este campo concreto, como una gran ventaja de esta tecnología con respecto a la tradicional.

Por tanto, el problema técnico que pretende resolver la presente invención es la identificación comparativa de objetos a partir de una población determinada de objetos definidos por vectores en la base de datos.

En el documento [*"Hardware implementation of stochastic-based Neural Networks", Rossello J L; Canals V; Morro A. Proceedings of the International Joint Conference on Neural Networks - 2010 IEEE World Congress on Computational Intelligence, WCCI 2010 - 2010 International Joint Conference on Neural Networks, IJCNN 2010*] se divulgan elementos hardware para implementar sistemas basados en redes neuronales a partir de señales pulsantes. Se utilizan señales binarias pulsantes descorrelacionadas para la realización de comparaciones rápida y eficiente.

En [*"Practical hardware implementation of self-configuring neural networks", Rossello J L; Canals V; Morro A; De Paul I. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) - Advances in Neural Networks - ISNN 2009 - 6th International Symposium on Neural Networks, ISNN 2009, Proceedings – 2009*] se presenta una revisión sobre elementos HW eficientes para implementar redes neuronales que permiten generar vectores aleatorios y donde se muestra un prototipo de bajo coste que combina una puerta lógica XOR y un element N-bit shift register.

En [*"Using stochastic logic for efficient pattern recognition analysis", Rossello J L; Canals V; De Paul I; Segura J. Proceedings of the International Joint Conference on Neural Networks - 2008 International Joint Conference on Neural Networks, IJCNN 2008 – 2008*] se divulga una metodología probabilística de reconocimiento de patrones basada en comparaciones en paralelo. Se presenta un sistema digital que permite obtener la señal más próxima al patrón utilizando puertas lógicas y bloques comparadores. Las señales pulsantes que se utilizan deben de estar descorrelacionadas entre ellas para poder realizar las operaciones.

El documento US2005108675 "Pattern recognition method for integrated circuit design, involves iteratively comparing rank in tree representation of design instance built around each design unit with related rank in pattern tree for each design unit in list" que se refiere a un sistema digital para un circuito integrado que permite el reconocimiento de patrones combinando puertas lógicas y bloques comparadores. También son de interés para ilustrar el estado del arte los documentos: US5703792 "System for selecting molecules or molecular parts - assigns measure of added diversity for each molecule relative to molecules in base set and directs processor to select molecules on basis of this measure, used in, e.g. pharmaceutical design" que se refiere al reconocimiento de formas, basado en la comparación de vectores para obtener la distancia entre ellos (Euclidean, Manhattan, etc) y realizar cribados de bases de datos utilizando técnicas estadísticas, aunque no se hace referencia un circuito digital para su implementación. En esta línea, el documento US2004117125 "Drug discovery and development for e.g. identifying additional applications and uses of known compounds comprises using databases comprising chemical and biological interaction data and computer-based data analysis program" que se refiere a la aplicación de búsqueda en bases de datos, utilizando la distancia entre los descriptores y la exploración virtual (virtual screening).

Descripción de la invención

El sistema propuesto utiliza un tipo de computación probabilística más amplia que la lógica estocástica tradicional puesto que para realizar las comparaciones utilizamos señales que en determinados bloques guardan una gran correlación mientras que para otros la descorrelación es máxima. Con dicha técnica podemos comparar de forma simple (mediante las señales correlacionadas) y realizar operaciones complejas (mediante las señales descorrelacionadas). En las publicaciones anteriores solamente se utilizan señales descorrelacionadas (característica de la lógica estocástica clásica), lo que obliga a implementar sistemas relativamente complejos como redes neuronales o comparadores bayesianos. Por tanto, el problema técnico que resuelve la presente invención es precisamente la simplificación de los circuitos implicados en la computación probabilística.

El objeto de la presente invención es el de desarrollar un sistema capaz de alcanzar grandes velocidades de exploración debido al alto grado de paralelismo conseguido con la lógica. Dicho paralelismo es debido al uso de técnicas probabilísticas, que son capaces de minimizar el hardware necesario por computación. Esas estructuras hardware minimizadas pueden posteriormente replicarse centenares de veces en un solo circuito integrado, con el objeto de obtener un paralelismo máximo a un coste mínimo.

La presente invención representa un gran paso adelante para la exploración de grandes bases de datos en tiempos razonables y es capaz de presentar velocidades de exploración que son varios órdenes de magnitud más rápidas en comparación con las formas tradicionales basadas en el uso de microprocesadores.

Uno de los campos en los que probablemente tenga más impacto la presente invención es en la búsqueda de nuevos fármacos. En las etapas iniciales de este campo de investigación se realiza lo que se conoce como exploración virtual ("*virtual screening*") el cual consiste en la identificación de moléculas candidatas a posibles fármacos a partir de grandes bases de datos moleculares. Estas bases de datos, que pueden contener miles de millones de componentes, tienen que ser exploradas de forma reiterada en la búsqueda de compuestos que posean ciertas características de forma y composición que los haga óptimos para poder interaccionar con determinados objetivos terapéuticos. Si se pretende realizar dicha exploración con técnicas convencionales en tiempos razonables se necesita del uso de docenas de estaciones de trabajo trabajando en paralelo (clusters y/o superordenadores). Con la presente invención (que es capaz de paralelizar el procesado de la base de datos en cada integrado) se lograrían disminuir costes hardware, energéticos y de mantenimiento en este proceso en la misma proporción en que se aumenta la velocidad de exploración.

Además de la búsqueda de nuevos fármacos, la presente invención es aplicable a cualquier campo científico-tecnológico en donde sea necesaria la exploración de grandes bases de datos en tiempos razonables.

Más concretamente, la invención es capaz de comparar un vector de referencia $\mathbf{x_0}$ con *n* vectores distintos escogidos de una base de datos. De esta forma, una base de datos extensa puede ser explorada en paquetes de *n* vectores. A cada comparación se muestra a la salida cuál de los vectores es más próximo a la referencia (en caso de que lo haya), ya que pasado un tiempo prudencial $\boldsymbol{\tau}$ a determinar por el usuario (y que fija un umbral de similitud), se pasará a cargar un nuevo paquete de *n* vectores procedentes de la base de datos. De esta forma, a la salida del comparador global se puede obtener uno o ningún resultado positivo.

El sistema está pensado para una exploración muy rápida de bases de datos muy extensas. Los resultados de cada comparación de cada grupo de 'n' vectores se van almacenando en una memoria de resultados con la codificación que se desee. Dicho almacenamiento y codificación se puede realizar mediante el uso de técnicas tradicionales (uso de procesadores) que no retardarán el proceso puesto que se realizan a la vez que el sistema propuesto está evaluando el siguiente paquete de información.

Como resultado final se obtienen de la base de datos un conjunto de vectores similares al vector patrón. Uno de los cuales será el más próximo de la base de datos. Aunque no está garantizado que el conjunto de vectores similares obtenidos al final de la exploración de la base de datos sean realmente los más próximos a la referencia (puesto que en el caso en que dos o más vectores próximos se hayan seleccionado en el mismo paquete solamente se selecciona uno de ellos), sí se obtendrá un conjunto de vectores dentro del umbral de similitud elegido por el usuario que puede ser bastante extenso.

En caso de querer un conjunto de vectores más reducido o simplemente obtener cuál es el vector más próximo se tendrán que realizar nuevas exploraciones pero sobre los vectores seleccionados. Esto no implica un retardo mucho mayor puesto que dicho conjunto de vectores que se selecciona ha de ser muchísimo más reducido que el conjunto de vectores totales de la base de datos mediante la selección de un umbral τ adecuado.

La presente invención es aplicable a múltiples campos de investigación en donde tenga que tratarse con extensas bases de datos y de las cuales tenga que obtenerse información útil. Un ejemplo será el proceso de exploración virtual ("virtual screening") que se realiza en las etapas iniciales del desarrollo de nuevos fármacos. Estas bases de datos pueden contener del orden de miles de millones de componentes que vendrán descritos por un vector de descriptores en relación a su forma o composición química. La velocidad de exploración de dichas bases de datos puede ser del orden de 10^7 comparaciones por segundo en el caso de usar un procesador comercial estándar [*P. Ballester, W.G. Richards, "Ultrafast shape recognition for similarity search in molecular databases" Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 463 (2081), pp. 1307-1321*]. Para una base de datos con 10^{12} componentes, si se desean buscar cuáles son los más próximos a 500 patrones concretos (que estarán relacionados con determinadas dianas terapéuticas o que podrán ser fármacos comerciales) se tardaría alrededor de año y medio en hacer la comparación con un solo procesador. Si se deseara realizar una comparación en un tiempo razonable se deberían usar 100 ordenadores en paralelo para realizar el cálculo en alrededor de 5 días con el consiguiente costo en infraestructura y personal de gestión. Con el sistema propuesto se podría usar un dispositivo digital de lógica programable (FPGA de alta gama) para realizar la misma comparación en el mismo período de 5 días.

A lo largo de la descripción y las reivindicaciones, la palabra "comprende" y sus variantes no pretender excluir otras características técnicas, aditivos, componentes o pasos. Para los expertos en la materia, otros objetos, ventajas y características de la invención se desprenderán en parte de la descripción y en parte de la práctica de la invención. Los siguientes ejemplos y dibujos proporcionan a modo de ilustración, y no se pretende que sean limitativos de la presente invención. Además, la presente invención cubre todas las posibles combinaciones de realizaciones particulares y preferidas aquí indicadas.

Breve descripción de las figuras

A continuación se pasa a describir de manera muy breve una serie de dibujos que ayudan a comprender mejor la invención y que se relacionan expresamente con una realización de dicha invención que se presenta como un ejemplo no limitativo de ésta.

La figura 1 muestra el diseño completo de la invención consistente en el uso de comparadores vectoriales (3) cuyas salidas alimentan un comparador global (5). A la salida de comparador WTA "Winner Take All" obtenemos un bus de respuesta (6) indicando cuál es la entrada más parecida al vector de referencia (x_0) en caso de que lo hubiere.

La figura 2 muestra el comportamiento interno de cada comparador vectorial (3). A la salida de este bloque se proporciona una señal de un bit (213) (S_k) que fluctúa con una probabilidad proporcional a la similitud entre los vectores que se comparan.

La figura 3 muestra cómo se generan los números aleatorios R_j que se usan en el comparador vectorial (3) de la figura 1. Dichos bloques pueden ser contadores pseudo-aleatorios LFSRs.

La figura 4 muestra el esquema del comparador probabilístico (5). El esquema está pensado como un "Winner Take All" (WTA) en donde la señal de más actividad acaba desbordando antes su contador y por tanto generando una señal a nivel alto '1' a la salida de su indicador correspondiente.

Realización preferente de la invención

La lógica estocástica es el resultado de la aplicación de las leyes de la probabilidad a la lógica digital en donde las variables son representadas mediante señales pulsantes estocásticas que poseen una probabilidad de activación determinada (p). Dicha probabilidad de activación estará definida entre 0 y 1.

La presente invención utiliza el alto grado de compresión de información que caracteriza dicha computación estocástica o probabilística para conseguir altos grados de paralelismo en un solo integrado. Dicho paralelismo es óptimo para la identificación de patrones concretos en bases de datos extensas.

El sistema global se muestra en la figura 1. En esta figura se parte de una entrada de n+1 vectores, de los cuales 'n' pertenecen a la base de datos y uno es el patrón a identificar (vector $\mathbf{x_0}$ mostrado en (1) de la Fig. 1). Cada vector consistirá en N números (descriptores) definidos por 'm' bits cada uno por lo que cada vector constará de un total de m*N bits. Cada uno de estos componentes de 'm' bits representarán un descriptor determinado para el objeto que describe cada vector. De esta forma cada objeto viene descrito por N descriptores distintos codificados cada uno con 'm' bits. Las cantidades que describe cada descriptor tendrán que ser normalizadas al rango que abarcan, en este caso un número entre 0 y 2^{m-1}.

De la base de datos a escanear se volcarán en el circuito de la Figura 1 un total de *n* vectores (2), cada uno de los cuales se comparará con el vector de referencia $\mathbf{x_0}$ (1) en el comparador vectorial o bloque CP (3). Dicho bloque CP (3) se ilustra en la figura 2 que se explicará más adelante.

Como resultado de dicha comparación se generará para cada vector $\mathbf{x_k}$ que se compare con $\mathbf{x_0}$ una señal de 1 bit (4) que presentará una actividad irregular (estocástica) con probabilidad de activación proporcional a la similitud entre los vectores comparados ($\mathbf{x_0}$ y $\mathbf{x_k}$). Esta señal queda definida como S_k. De esta forma, para cada paquete de 'n' vectores seleccionados de la base de datos se obtienen 'n' bits pulsantes con probabilidad de activación proporcional a la similitud entre cada vector y el de referencia. Dichos bits se conectarán a un bloque de comparación global de pulsos (5), que denominaremos en la presente descripción como bloque WTA. De esta forma, la salida (6) del bloque WTA (5) (señal *category* indicada en la figura 1, con referencia (6)) codificará cuál de los 'n' bits es mayor (y por tanto cuál es el vector más próximo al vector de referencia). El bloque WTA (5) tendrá un tiempo de respuesta determinado τ que estará en función de los contadores internos del bloque WTA (5). Si pasado este tiempo no hay ninguna salida activada (todos los bits del bus category están a 0) querrá decir que ninguno de los 'n' vectores seleccionados tiene una similitud apreciable con respecto al vector de referencia $\mathbf{x_0}$.

En la figura 2 se muestra el contenido del comparador vectorial notado como bloque CP (3) que se mostraba en la figura 1. Dicho bloque consiste en la comparación, descriptor por descriptor, entre el vector de referencia $\mathbf{x_0}$ y el vector k-ésimo $\mathbf{x_k}$. Si los vectores son de N dimensiones tendremos por tanto que comparar un total de N descriptores, par a par.

Para la realización de la comparación se parte de cada par de descriptores (21) y que consiste en un número binario de 'm' bits, $\mathbf{X_{0j}}$ y $\mathbf{X_{kj}}$ donde j es un número entero entre 1 y N definiendo al descriptor j-ésimo y se comparan con un número fluctuante aleatorio de 'm' bits R_j (22) mediante un comparador binario (23).

Este comparador binario (23) proporcionará un '1' ('0') lógico a su salida en caso de que $X_{kj}>R_j$ ($X_{kj}<R_j$). Dicho número aleatorio no estará fijo a un valor determinado sino que fluctuará con una determinada cadencia temporal (el bloque generador de cada número aleatorio R_j se describe más adelante en la figura 3). Como resultado de dichas comparaciones se generarán 'N' señales pulsantes estocásticas (x_{01} a x_{0N} y desde x_{k1} hasta x_{kN}) a la salida (24) de cada comparador (23).

Para la comparación de cada par de descriptores del vector de referencia $\mathbf{x_0}$ y del vector a comparar $\mathbf{x_k}$ se obtendrán las señales pulsantes x_{0j} y x_{kj} que son generadas por los correspondientes comparadores. Dichas señales pulsantes oscilarán de forma estocástica pero correlacionada entre ellas debido a usar la misma comparación con $\mathbf{R_j}$ en los comparadores. De esta forma, puesto que $\mathbf{X_{0j}}$ y $\mathbf{X_{kj}}$ se comparan con el mismo número binario aleatorio oscilante $\mathbf{R_j}$ ambas señales presentarán una correlación máxima. Eso significa que si $\mathbf{X_{0j} > X_{kj}}$ nunca se podrá dar el caso en donde la señal pulsante x_{0j} sea 0 y x_{kj} sea un 1 lógico (y viceversa en el caso $\mathbf{X_{0j} < X_{kj}}$). Esa correlación es óptima para cuantificar la distancia entre descriptores. Puesto que las señales pulsantes x_{0j} y x_{kj} representan probabilidades de activación (entre 0 y 1) y están correlacionadas entre ellas, la probabilidad de que sean ambas distintas en un momento dado es directamente proporcional a la diferencia de probabilidades en valor absoluto $| x_{0j} – x_{kj} |$. Esta es la probabilidad de activación de la señal pulsante (26) que se obtiene a la salida de la puerta XOR (25) que está conectada a ambas señales.

De esta forma, a la salida de la XOR se obtiene una señal pulsante estocástica cuya probabilidad de activación es proporcional a la distancia entre descriptores. Dicha señal estocástica se puede elevar al cuadrado si se la multiplica por ella misma mediante una puerta AND (28). Para realizar dicha multiplicación se han de descorrelacionar las entradas de la puerta AND lo cual se puede obtener mediante un bloque que imponga un determinado retraso como un flip-flop tipo D o un registro de desplazamiento (27). A la salida de la puerta AND (29) tendremos por tanto una señal pulsante estocástica con probabilidad de activación igual a $(x_{0j} – x_{kj})^2$.

Para cada componente vectorial que se compara se replica la estructura para obtener dicha señal (29). Una vez obtenidas las señales de dichos comparadores pulsantes se conectan a la entrada de un multiplexor (210) que se encarga de pasar a su salida (211) todas y cada una de sus N entradas (que tienen que estar descorrelacionadas entre ellas) de forma secuencial. El resultado de dicha multiplexación es una señal pulsante (211) con probabilidad de activación igual al valor medio de las señales de entrada $\Sigma_j(x_{0j}–x_{kj})^2/N$. Esta probabilidad es proporcional al cuadrado de la distancia euclidiana entre los vectores que se comparan. Si se desea obtener la similitud entre vectores se puede invertir la señal con un inversor (212) de forma que se obtiene una señal con probabilidad de activación:

$$S_k=1-\Sigma_j(x_{0j}–x_{kj})^2/N$$

Este parámetro será 1 cuando ambos vectores sean iguales mientras que tenderá a cero a medida que las distancias entre vectores aumente. La salida de este bloque se unirá a otras 'n-1' señales de similitud correspondientes a los 'n' vectores que se estarán comparando en cada instante y serán la entrada del comparador de pulsos WTA (5).

Los números aleatorios R_j serán generados por distintos bloques que tanto podrán ser generadores de números pseudo-aleatorios como las LFSR (Linear-Feedback Shift Register, descrito en [*P.H.R. Scholefield, "Shift Registers Generating Maximum-Length Sequences", Electronic Technology, 10-1960, pp.389-394*]) como generadores de números aleatorios basados en sistemas caóticos (analógicos o digitales). Dichos bloques se muestran en la figura 3 (31) para el caso del uso de LFSRs que serán inicializadas por una señal de inicio global (32) y que proporcionarán a su salida (33) un número aleatorio de 'm' bits cada flanco de la señal de reloj (34). Para el caso del uso de LFSRs, la semilla de cada bloque deberá ser distinto para descorrelacionar entre sí las salidas R_j.

En último lugar está el comportamiento del bloque WTA (5) de la Figura 1. Dicho bloque lo podemos observar en la Figura 4 en donde cada entrada de similitud S_i (i=1..n) de cada comparación realizada se conecta a un contador. De esta forma, las 'n' entradas (41) se conectan a dichos contadores (42) que en caso de desbordar su cuenta (que pueden ser de 16 o más estados) activan su salida (43) y_i.

Dicha salida activará el reset síncrono (45) mediante la puerta OR (44) de todos los contadores que no han llegado a desbordar. A su vez, la señal y_i (43) es capturada por el flip-flop D (46) que proporcionará la codificación de cuál es el vector más parecido a la referencia $\mathbf{x_0}$. De esta forma la salida (47) de cada Flip-Flop (46) se actualizará en el momento que la señal de habilitación (48) se active y que coincide con la señal de clear síncrono (45). De esta forma se obtiene a la salida del WTA (6) un bus de 'n' bits con todos los bits puestos a 0 excepto el bit correspondiente al vector más próximo al de referencia. Externamente se puede esperar un tiempo determinado para obtener una determinada similitud. Una vez pasado este tiempo se puede cargar otro paquete de vectores desde la base de datos y se guardan los datos del resultado del paquete anterior. Si ningún contador ha desbordado en ese tiempo se podrá decir que ningún vector del paquete escogido tiene la similitud requerida.

Esta comparación se ha realizado en paralelo y ha tardado un tiempo relacionado con el tiempo de desbordamiento típico de los contadores internos del WTA (42) que puede ser un valor típico de unas decenas de veces el período de oscilación de la señal de reloj del sistema (señales (49) en Figura 4 y (34) de la Figura 3 que corresponden a la misma señal de reloj). La comparación que ser realiza comprime la información de forma tal que el paralelismo del sistema se maximiza. De esta forma, en una FPGA de tamaño grande se podrán integrar sistemas que comparen cientos de vectores a la vez, lo cual implicará una velocidad considerable que será cientos de veces superior a la velocidad que se puede conseguir con un solo procesador mediante técnicas clásicas.

La principal característica del sistema propuesto y que lo diferencia de otros comparadores estocásticos presentes en la literatura (como los mencionados en el apartado de antecedentes de la invención) es el uso de señales aperiódicas que en determinados bloques están correlacionadas entre sí (para poder obtener una señal de similitud como es el caso de la señal (26)), y que en otros bloques se encuentran descorrelacionadas (como es el caso de las señales de entrada del multiplexor (210) o de las señales de entrada del WTA (5). De esta forma se utilizan las capacidades aritméticas de las señales estocásticas descorrelacionadas (para calcular el cuadrado (29) de una señal (26) o para calcular la media aritmética de N señales (211) con un multiplexor) además de las capacidades de comparación que surgen cuando las señales pulsantes están correlacionadas entre sí (como en la evaluación de la señal (26) usando una puerta XOR (25) puesto que las señales x_{0j} y x_{kj} están correlacionadas entre sí al ser el resultado de comparaciones con números aleatorios iguales R_j). La combinación de ambos tipos de señales (señales correlacionadas para la comparación y señales descorrelacionadas para la realización de operaciones aritméticas) proporciona una gran optimización tanto en términos de bloques digitales como de número de ciclos de reloj necesarios para cada comparación. Este tipo de computación probabilística se diferencia de la estocástica clásica en el sentido que esta última considera que las señales han de estar siempre descorrelacionadas temporalmente para su procesamiento. Los comparadores pulsantes mencionados anteriormente como [*V Canals, A. Morro, J.L. Rosselló, "Stochastic-Based pattern-recognition analysis" Pattern Recognition Letters 20101101 Elsevier*] , [*"Hardware implementation of stochastic-based Neural Networks", Rossello J L; Canals V; Morro A. Proceedings of the International Joint Conference on Neural Networks - 2010 IEEE World Congress on Computational Intelligence, WCCI 2010 - 2010 International Joint Conference on Neural Networks, IJCNN 2010*] , [*"Practical hardware implementation of self-configuring neural networks", Rossello J L; Canals V; Morro A; De Paul I. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) - Advances in Neural Networks - ISNN 2009 - 6th International Symposium on Neural Networks, ISNN 2009, Proceedings – 2009*] ó [*"Using stochastic logic for efficient pattern recognition analysis", Rossello J L; Canals V; De Paul I; Segura J. Proceedings of the International Joint Conference on Neural Networks - 2008 International Joint Conference on Neural Networks, IJCNN 2008 – 2008*] basan su funcionamiento en el uso de señales estocásticas que siempre están descorrelacionadas, construyendo bloques estocásticos con la finalidad de implementar sistemas más complejos como redes neuronales o comparadores bayesianos, lo que hace que internamente sean bastante más complejos que en el diseño propuesto en este documento (basado en el uso de señales tanto correlacionadas como descorrelacionadas según convenga en cada etapa de la comparación y minimizando el hardware necesario).

REIVINDICACIONES

1.- Sistema digital probabilístico para la exploración eficiente de grandes bases de datos del tipo que comprende una entrada de n+1 vectores, de los cuales 'n' pertenecen a la base de datos y uno es el patrón a identificar vector $\mathbf{X_0}$ (1); donde cada vector consistirá en N números o descriptores definidos por 'm' bits cada uno por lo que cada vector constará de un total de m*N bits, y donde cada uno de estos componentes de 'm' bits representarán un descriptor determinado para el objeto que describe cada vector; de tal forma que cada objeto viene determinado por N descriptores distintos codificados cada uno con 'm' bits, y donde las cantidades que describe cada descriptor tendrán que ser normalizadas al rango que abarcan; y donde de la base de datos a escanear se volcarán un total de *n* vectores (2), cada uno de los cuales se comparará con el vector de referencia $\mathbf{X_0}$ (1) en una pluralidad de comparadores vectoriales (3); todo ello **caracterizado porque:**

como resultado de dicha comparación se genera para cada vector $\mathbf{X_k}$ que se compare con $\mathbf{X_0}$, una señal $\mathbf{s_k}$ de 1 bit (4) que presenta una actividad irregular estocástica con probabilidad de activación proporcional a la similitud entre los vectores comparados $\mathbf{X_0}$ y $\mathbf{X_k}$; todo ello de tal forma que para cada vector de los 'n' vectores seleccionados de la base de datos se obtienen 'n' bits pulsantes con probabilidad de activación proporcional a la similitud entre vectores, y donde dichos bits están conectados con un bloque de comparación global de pulsos (5) cuya salida (6) codifica cuál de los 'n' bits es mayor;

donde además, el comparador vectorial (3) comprende medios de comparación (21,22,23) para generar N señales pulsantes estocásticas correlacionadas entre sí y medios para la realización de operaciones aritméticas (27,28,29) con dichas señales descorrelacionadas;

donde además, cada comparador vectorial (3) comprende al menos un comparador binario (23) con dos entradas (21,22), en donde una primera entrada (21) está conectada con los descriptores que consisten en números binarios de 'm' bits $\mathbf{X_{0j}}$ y $\mathbf{X_{kj}}$ donde j es un número entero entre 1 y N definiendo al descriptor j-ésimo y la segunda entrada (22) consiste en un número aleatorio de 'm' bits R_j; todo ello de tal forma que como resultado de dichas comparaciones se generarán 'N' señales pulsantes estocásticas ($\mathbf{x_{0j}}$, $\mathbf{x_{kj}}$) a la salida (24) de cada comparador (23); y donde estas señales de salida (24) están conectadas con un puerta XOR (25) cuya salida (26) es una señal pulsante estocástica cuya probabilidad de activación es proporcional a la distancia entre descriptores; y donde dicha señal estocástica se puede elevar al cuadrado si se la multiplica por ella misma mediante una puerta AND (28), descorrelacionando las señales mediante un registro de desplazamiento (27), obteniendo a la salida de la puerta AND (29) una señal pulsante estocástica con probabilidad de activación igual a $(\mathbf{x_{0j} - x_{kj}})^2$;

y donde además, para cada componente de los vectores que se comparan, se replica la estructura para obtener la señal de salida de la puerta AND (29); que están conectadas con la entrada de un multiplexor (210) que se encarga de pasar a su salida (211) todas y cada una de sus N entradas de forma secuencial, cuyo resultado es una señal pulsante (211) con probabilidad de activación igual al valor medio de las señales de activación $\Sigma_j(x_{0j}-x_{kj})^2/N$, siendo esta probabilidad proporcional al cuadrado de la distancia euclidiana entre los vectores que se comparan.

2.- Sistema de acuerdo con la reivindicación 1 donde el bloque de comparación global de pulsos (5) tiene un tiempo de respuesta determinado τ que estará en función de los contadores internos del comparador global de pulsos (5).

3.- Sistema de acuerdo con la reivindicación 1 donde la señal pulsante (211) de salida del multiplexor (210) se invierte la señal con un inversor (212) de forma que se obtiene una señal con probabilidad de activación $S_k=1-\Sigma_j(x_{0j}-x_{kj})^2/N$ (4); donde este parámetro será 1 cuando ambos vectores sean iguales mientras que tiende a cero a medida que las distancias entre vectores aumenta.

4.- Sistema de acuerdo con la reivindicación 1 donde los números aleatorios R_j serán generados por distintos bloques que tanto podrán ser generadores de números pseudo-aleatorios como las LFSR (31) como generadores de números aleatorios basados en sistemas caóticos ó estocásticos.

5.- Sistema de acuerdo con la reivindicación 4 en donde los LFSR (31) son inicializados por una señal de inicio global (32) y que proporcionarán a su salida (33) un número aleatorio de 'm' bits cada flanco de la señal de reloj (34).

6.- Sistema de acuerdo con cualquiera de las reivindicaciones anteriores en donde el comparador global de pulsos (5) para cada entrada de similitud S_i (i=1..n) de cada comparación realizada se conecta a un contador, de tal forma que las 'n' entradas (41) se conectan a dichos contadores (42) que en caso de desbordar su cuenta activan su salida (43) y_i, que a su vez activa el reset síncrono (45) mediante una puerta OR (44) de todos los contadores que no han llegado a desbordar; y donde su vez, la señal y_i (43) es capturada por un elemento de memoria (flip-flop D) (46) que proporciona a su salida la codificación de cuál es el vector más parecido a la referencia $\mathbf{x_0}$; y donde la salida (47) de cada Flip-Flop (46) se actualiza en el momento que la señal de habilitación (48) se activa y que coincide con la señal de reset síncrono (45), obteniendo a la salida (6) del comparador global de pulsos (5) un bus de 'n' bits con todos los bits puestos a 0 excepto el bit correspondiente al vector más próximo al de referencia.

7.- Método para la exploración eficiente de grandes bases de datos implementado en un sistema según cualquiera de las reivindicaciones anteriores que comprende una entrada de n+1 vectores, de los cuales 'n' pertenecen a la base de datos y uno es el patrón a identificar vector $\mathbf{X_0}$ (1); donde cada vector consistirá en N números o descriptores definidos por 'm' bits cada uno por lo que cada vector constará de un total de m*N bits, y donde cada uno de estos componentes de 'm' bits representarán un descriptor determinado para el objeto que describe cada vector; de tal forma que cada objeto viene determinado por N descriptores distintos codificados cada uno con 'm' bits, y donde las cantidades que describe cada descriptor están normalizadas al rango que abarcan; **caracterizado porque** comprende:

un volcado de la base de datos a escanear de un total de *n* vectores (2), cada uno de los cuales se compara con el vector de referencia $\mathbf{X_0}$ (1) en una pluralidad de comparadores vectoriales (3); y donde como resultado de dicha comparación se genera para cada vector $\mathbf{X_k}$ que se compare con $\mathbf{X_0}$, una señal $\mathbf{s_k}$ de 1 bit (4) que presenta una actividad irregular estocástica con probabilidad de activación proporcional a la similitud entre los vectores comparados $\mathbf{X_0}$ y $\mathbf{X_k}$; todo ello de tal forma que para cada vector de los 'n' vectores seleccionados de la base de datos se obtienen 'n' bits o señales pulsantes con probabilidad de activación proporcional a la similitud entre vectores; y donde finalmente se establece una etapa de codificación de los 'n' bits definiendo cuál es el mayor;

donde el comparador vectorial (3) comprende medios de comparación (21,22,23) para generar N señales pulsantes estocásticas correlacionadas entre sí y medios para la realización de operaciones aritméticas (27,28,29) con dichas señales descorrelacionadas;

donde cada comparador vectorial (3) comprende al menos un comparador binario (23) con dos entradas (21,22), en donde una primera entrada (21) está conectada con los descriptores que consisten en números binarios de 'm' bits $\mathbf{X_{0j}}$ y $\mathbf{X_{kj}}$ donde j es un número entero entre 1 y N definiendo al descriptor j-ésimo y la segunda entrada (22) consiste en un número aleatorio de 'm' bits R_j; todo ello de tal forma que como resultado de dichas comparaciones se generarán 'N' señales pulsantes estocásticas ($\mathbf{x_{0j}}$, $\mathbf{x_{kj}}$) a la salida (24) de cada comparador (23); y donde estas señales de salida (24) están conectadas con un puerta XOR (25) cuya salida (26) es una señal pulsante estocástica cuya probabilidad de activación es proporcional a la distancia entre descriptores; y donde dicha señal estocástica se puede elevar al cuadrado si se la multiplica por ella misma mediante una puerta AND (28), descorrelacionando las señales mediante un registro de desplazamiento (27), obteniendo a la salida de la puerta AND (29) una señal pulsante estocástica con probabilidad de activación igual a $(\mathbf{x_{0j} - x_{kj}})^2$;

y donde, para cada componente de los vectores que se comparan, se replica la estructura para obtener la señal de salida de la puerta AND (29); que están conectadas con la entrada de un multiplexor (210) que se encarga de pasar a su salida (211) todas y cada una de sus N entradas de forma secuencial, cuyo resultado es una señal pulsante (211) con probabilidad de activación igual al valor medio de las señales de activación $\Sigma_j(x_{0j}-x_{kj})^2/N$, siendo esta probabilidad proporcional al cuadrado de la distancia euclidiana entre los vectores que se comparan.

8.- Uso del sistema según cualquiera de las reivindicaciones 1 a 6 en la identificación de moléculas candidatas a posibles fármacos a partir de grandes bases de datos moleculares.

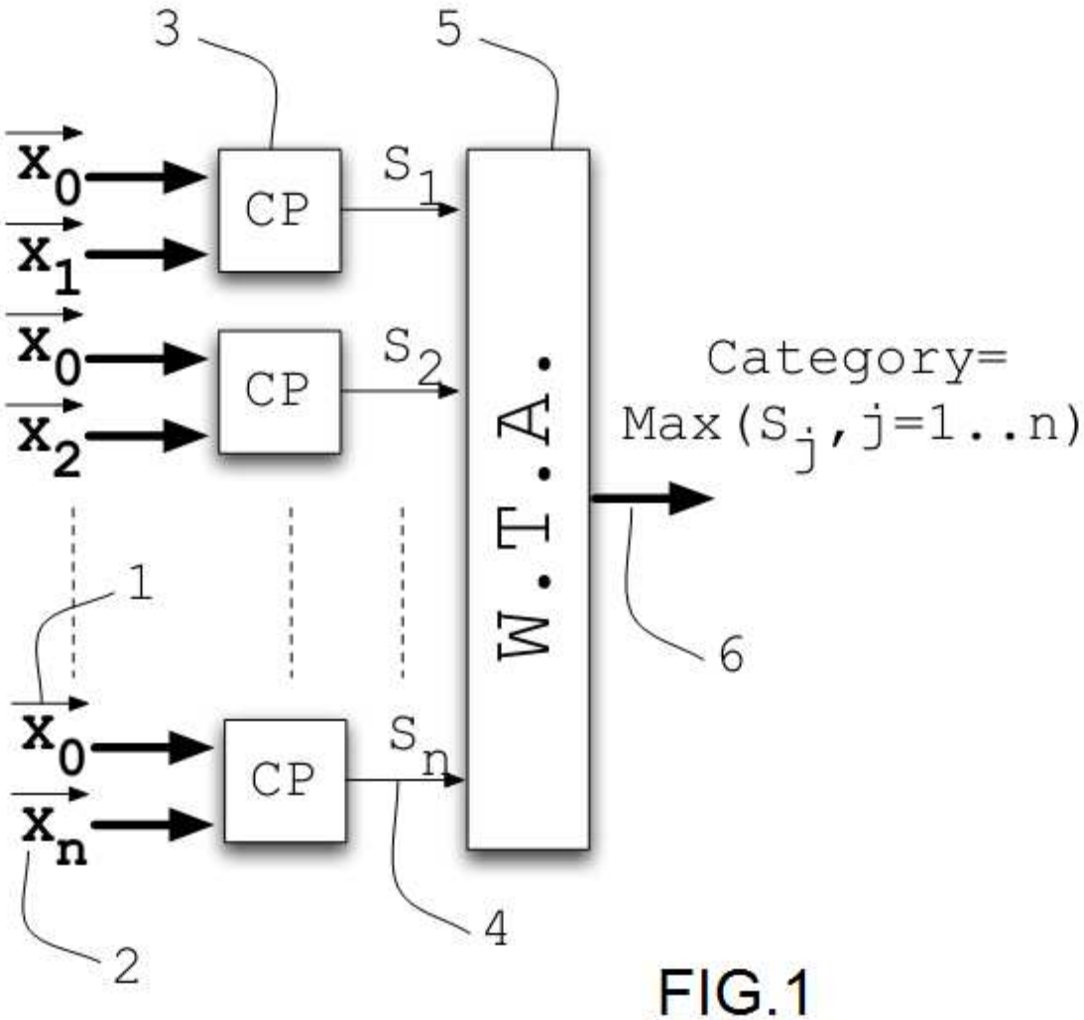

FIG.1

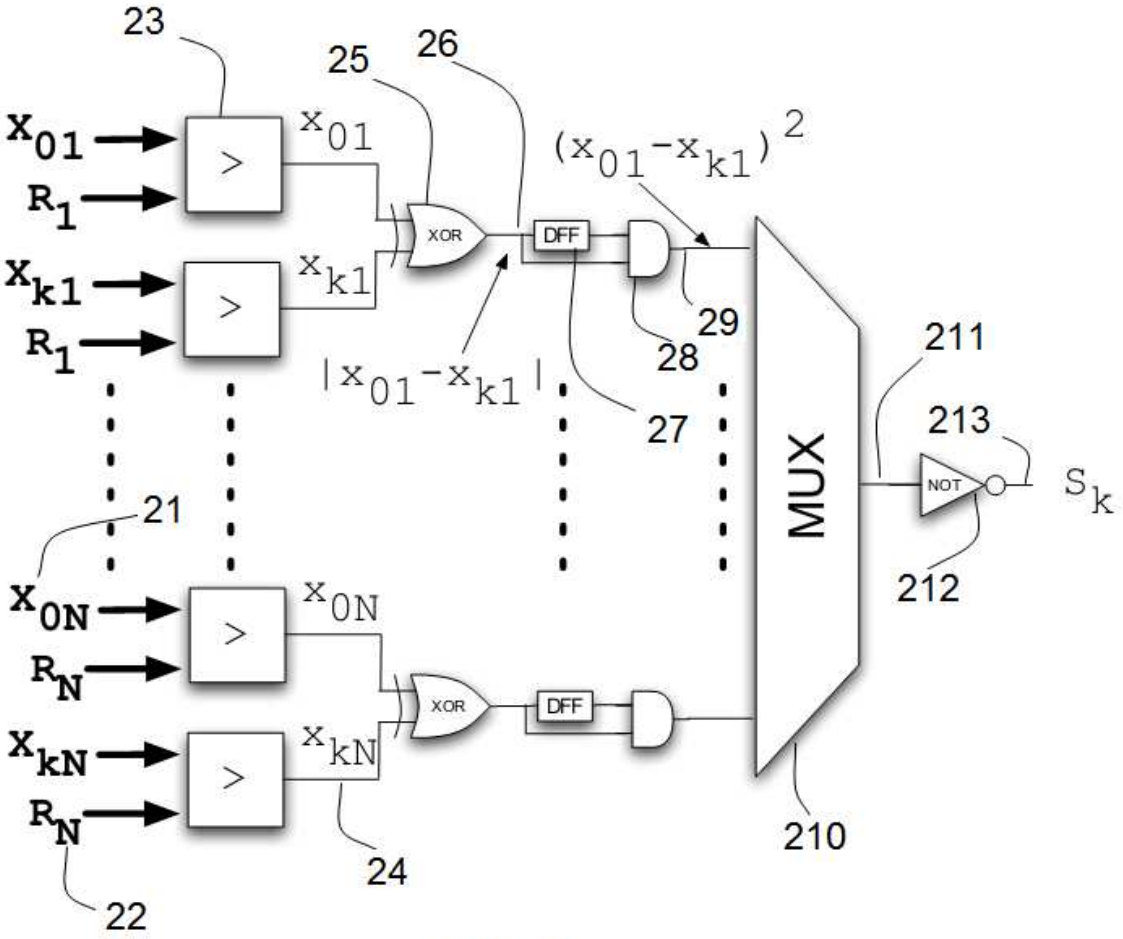
23
25
26
x_{01}
R_1
x_{k1}
R_1
XOR
DFF
$(x_{01}-x_{k1})^2$
29
28
$|x_{01}-x_{k1}|$
27
211
213
MUX
NOT
S_k
212
21
x_{0N}
R_N
x_{kN}
R_N
XOR
DFF
210
24
22

FIG.2

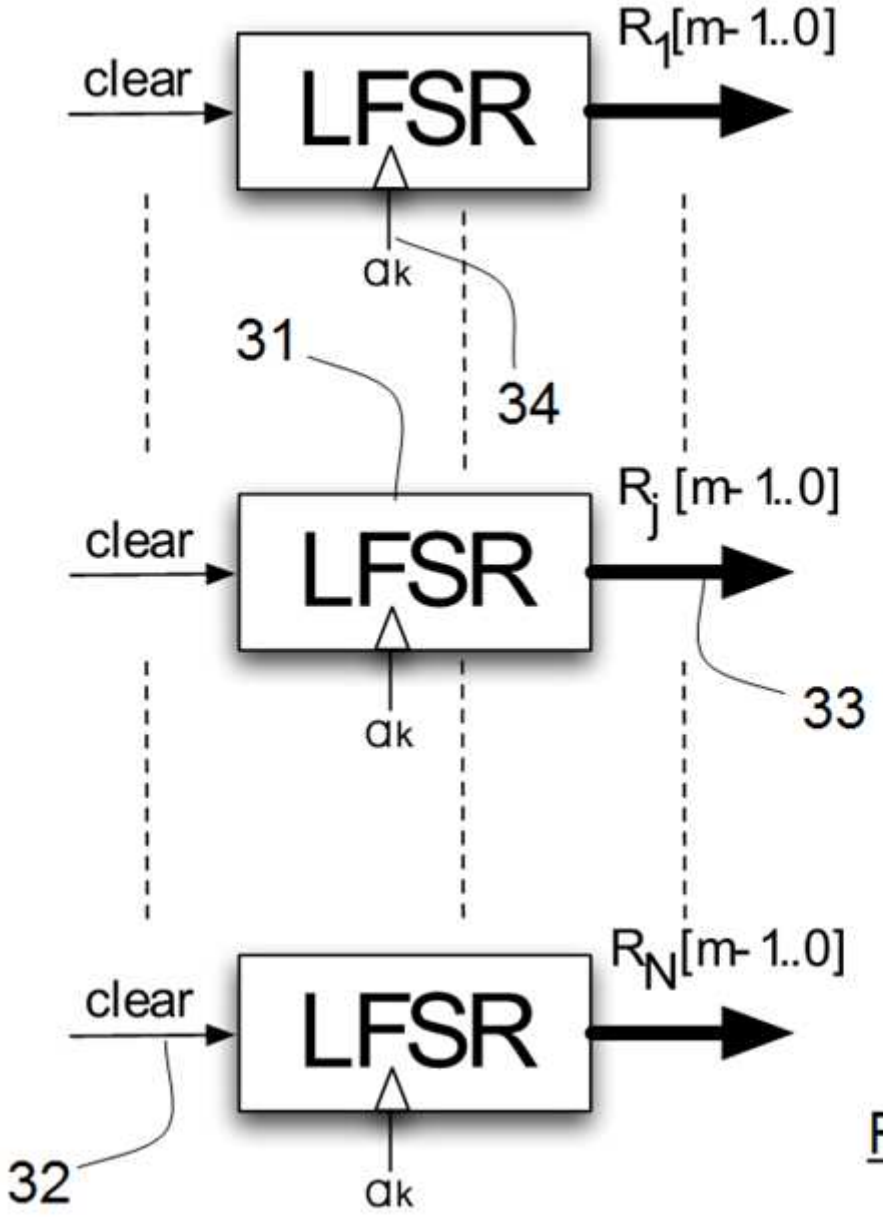
clear
LFSR
$R_1[m-1..0]$
ak
31
34
clear
LFSR
$R_j[m-1..0]$
ak
33
clear
LFSR
$R_N[m-1..0]$
32
ak

FIG.3

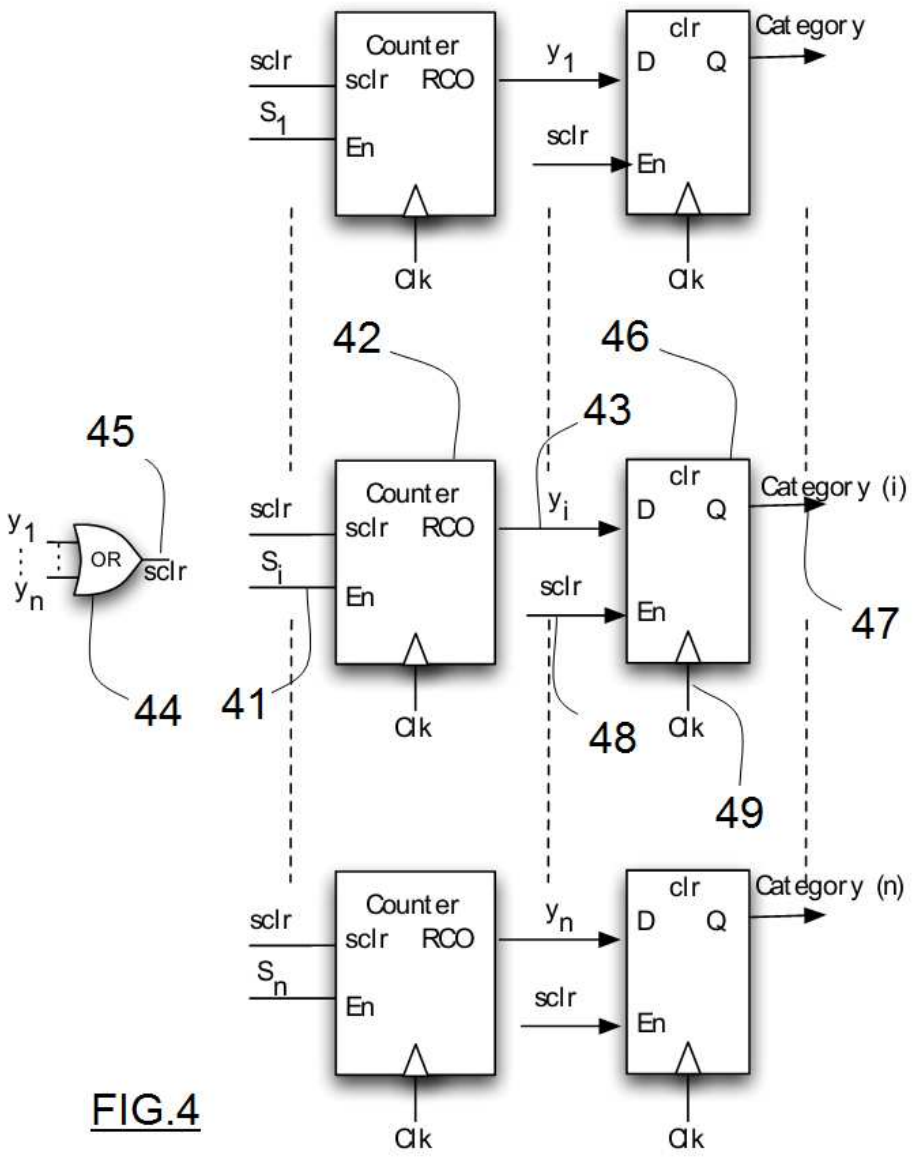
Counter
sclr
RCO
En
Clk
clr
D
Q
Category
sclr
S_1
y_1
42
46
43
45
y_i
Category (i)
47
S_i
41
44
OR
y_n
48
49
S_n
y_n
Category (n)

FIG.4