SEDFE: Un Sistema Experto para el Diagnóstico Fitosanitario del Espárrago usando Redes Bayesianas

Pedro Nelson Shiguihara Juárez
Universidad Nacional de Trujillo, Escuela de Informática,
Trujillo, Perú, 044
p.shiguihara@gmail.com

and

Jorge Carlos Valverde Rebaza
Universidad Nacional de Trujillo, Escuela de Informática,
Trujillo, Perú, 044
jorge.carlos14@gmail.com

Abstract

This paper proposes an expert system based on a probabilistic model of bayesian network for diagnosis of plagues and diseases of asparagus, using a likelihood propagation technique based on the messages passing algorithm of Kim and Pearl, it allows the nodes updating in the diagnosis network by reaching results with a short difference margin with an exact calculus of joint probability distribution using an Enumeration algorithm. In addition, the system is able to establish consistent results according to the conventional patterns of each pathogenic germ and its manifestations.

Keywords: Joint Probability Distribution, Conditional probability, Bayes theorem.

Resumen

Este artículo propone un sistema experto basado en el modelo probabilístico de redes Bayesianas para el diagnóstico de plagas y enfermedades del espárrago, el cual, hace uso de la técnica de propagación de certeza basada en el algoritmo de paso de mensajes de Kim y Pearl para la actualización de nodos dentro de la red de diagnóstico. De esta manera se logra alcanzar resultados con un margen de diferencia de centésimas con respecto al cálculo exacto obtenido con la tabla de distribución conjunta completa usando un algoritmo de Enumeración. Además, el sistema experto logra establecer resultados coherentes de acuerdo a los patrones convencionales de cada germen patógeno del espárrago y sus manifestaciones.

Palabras claves: Distribución de probabilidad conjunta, Probabilidad condicional, Teorema de Bayes.

INTRODUCCIÓN

La producción de cultivos agroindustriales se ha visto siempre diezmada por el ataque de plagas o el padecimiento de enfermedades en buena parte de toda la producción del cultivo agrícola, conllevando esto siempre a grandes pérdidas económicas en las empresas agroindustriales, en donde se pierden grandes áreas de cultivos azotadas por alguna anomalía que en muchos casos pudo haber sido evitada si se detectaba con anterioridad. El diagnóstico de enfermedades o plagas se lleva a cabo a través del monitoreo de las plantaciones de cultivo, en donde el objetivo es detectar anomalías presentes sobre las plantas, como por ejemplo, tallos secos, presencia de pústulas en alguna parte de la planta, presencia de insectos nocivos, entre otros. Así, teniendo un conjunto de síntomas visibles dentro del ambiente o campo de cultivo se puede realizar un diagnóstico que permitirá tomar las medidas necesarias y adecuadas para minimizar el daño sobre los cultivos.

En este contexto el trabajo aborda un problema específico, enfocándose en el diagnóstico de enfermedades y plagas para el producto del espárrago. El espárrago es un producto agroindustrial de gran consumo, y así como otros productos agroindustriales, tienen un tratamiento especial y adecuado para lograr su cosecha. Este producto resulta normalmente afectado por muchas enfermedades como la roya, estemfiliosis, etc.; y plagas como el gusano de alambre, entre otros. Esto depende muchas veces de las condiciones ambientales y tipo de suelo del lugar de cultivo. Dado que las plagas y enfermedades se presentan en función de las características del lugar de cultivo, tenemos que no todos los lugares de cultivos tendrán las mismas enfermedades o plagas, o al menos no con el mismo nivel de presencia.

Por años, el diagnóstico de enfermedades ha venido siendo un tema de interés en el área computacional de la inteligencia artificial. Sistemas expertos como MYCIN en la década de 1970 para el diagnóstico de enfermedades de la sangre, fueron pioneros en esta rama. Desde esa época hasta la actualidad, los sistemas expertos han evolucionado en muchos aspectos tales como: el modo de inferencia, interfaces de usuario, representación de la base de conocimiento, etc.; mencionados en [2]. Muchos modelos han sido presentados para el "*razonamiento*" de los computadores a fin de lograr un diagnóstico efectivo. Uno de los modos de razonamiento que se usaron en un primer momento fue el basado en la lógica de primer orden con una base de conocimiento construido con múltiples reglas lógicas.

Muchos sistemas expertos basados en reglas han sido desarrollados. En [5] se presenta un sistema experto para el diagnóstico de plantas basado en la observación visual de síntomas, el cual puede diagnosticar diferentes agentes infecciosos presentes en las plantas; los resultados logrados son buenos, sin embargo, son muy genéricos. [6] desarrolló un sistema experto basado en reglas para el diagnóstico de plagas, enfermedades y desordenes para el mango, así mismo, [7] propone el diseño e implementación de JAPIEST, un sistema de inteligencia integral para el diagnóstico de enfermedades y pestes del tomate. En [8] se presenta un sistema experto basado en reglas para el diagnóstico de enfermedades y pestes en plantaciones de olivo.

Puesto que los resultados de inferir en base a reglas son infalibles, la desventaja radica en que es necesario tener todas las reglas para poder efectuar la inferencia, algo que no es posible en la realidad en problemas como la detección de anomalías del espárrago, puesto que podemos observar sólo unos cuantos síntomas del total que representa a una enfermedad o plaga. A los síntomas visibles dentro de un ambiente se les denomina evidencia.

Por ello, una de las soluciones a dicho problema fue establecer un modelo probabilístico, en donde todas las enfermedades y síntomas tenían inicialmente un valor

de probabilidad que representaba el nivel de presencia en el entorno, llamada prevalencia. En este otro enfoque, si bien es cierto se establecían reglas lógicas para relacionar los síntomas con enfermedades, éstas no obedecían al modelo de inferencia de la lógica de primer orden, dado que no respetaban las reglas de inferencia sino que las quebrantaban. A pesar de ello, los resultados fueron mayoritariamente satisfactorios dado que las prevalencias permitían adaptar el modelo del conocimiento a un entorno específico.

La red Bayesiana es un modelo probabilístico basado en el teorema de Bayes y que establece una relación entre síntomas y enfermedades mediante la construcción de la base de conocimiento en un grafo, en donde los nodos padres representan a los efectos y los nodos hijos a las causas, teniendo así, un conjunto de reglas implícitamente representando el modo causa-efecto, de nodo hijo a nodo padre (para mayor detalle ver [2] y [4]). Este tipo de grafo se le denomina red de diagnóstico, puesto que posibilita mantener información que es necesaria para lograr efectuar una inferencia sobre el grafo y obtener un diagnóstico resultante. Si bien el modelo probabilístico es la parte principal del sistema experto presentado, es necesario realizar un proceso previo a su construcción, la selección adecuada de un conjunto de síntomas para cada enfermedad y plaga del espárrago. Quizá, parecería lógico establecer como un conjunto de síntomas para una enfermedad determinada, a todos los síntomas que se hayan presentado cuando ocurrió una enfermedad en el pasado. Sin embargo, realizar la selección de este modo no sería lo adecuado, por la existencia de dos factores: la *sensibilidad* y la *especificidad*, que serán explicadas más adelante. Luego de la selección de los síntomas viene la construcción del grafo, considerando la representación de una red de diagnóstico. Posteriormente el motor de inferencia del sistema experto realiza su trabajo de inferir un diagnóstico, para lo cual hace uso del algoritmo de paso de mensajes propuesto por Kim y Pearl en [1] y presentado también en [2].

El resto del documento se organiza de la siguiente manera. En la sección 2, se explica brevemente la teoría del modelo probabilístico presentado, en la sección 3 se presentan los experimentos y resultados obtenidos, finalmente, en la sección 4 mostramos nuestras conclusiones.

FUNDAMENTO TEÓRICO

2.1 Red Bayesiana

La red Bayesiana de tipo causal [2], es el conjunto de síntomas y causas (enfermedades) representadas por un nodo, en donde el sentido de las aristas denotan una relación directa entre efecto-causa. En la figura 1 podemos observar una red de tipo causal, en donde la evidencia proviene normalmente de los nodos síntomas, y en donde se trata de determinar cuales enfermedades representan mejor a los síntomas observados.

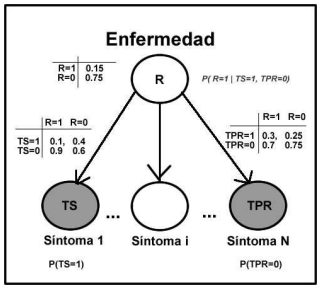
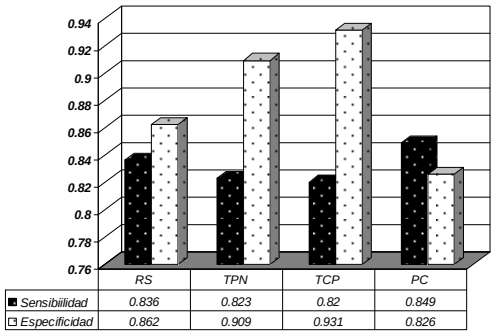


Figura 1: Ejemplo de una Red Bayesiana de diagnóstico, en donde los nodos sombreados representan la evidencia y cada uno tiene una *tabla de probabilidad condicional* conocido como CPT.

En la figura 1 por razones de simplicidad usaremos variables para denotar a alguna enfermedad o síntoma. El nodo "*R*" representa a la enfermedad del espárrago conocida como Roya, mientras que los síntomas Tallo Seco y Tallo con Pústulas Rojizas son representados como "*TS*" y "*TPR*" respectivamente. En la figura 1 se observa que "*TS*" y "*TPR*" han sido evidenciadas, confirmando la presencia de Tallo Seco *P(TS=1)* y la ausencia de Tallo con Pústulas Rojizas *P(TPR=0)*. Ahora podemos estimar la probabilidad de que aparezca Roya dada las dos evidencias anteriores, como se expresa en la consulta: *P(R=1 | TS=1, TPR=0).*
La notación que se presenta en este informe acerca de las variables o nodos, como en la figura 1 y tablas de resultados, se basan en las referencias de [4] y [2].

2.2 Sensibilidad y especificidad

La sensibilidad es la probabilidad que indica una correlación entre la aparición de la enfermedad y la aparición del síntoma. La especificidad en cambio, es la probabilidad de no tener el síntoma, cuando no está la enfermedad presente. Ambos valores son indispensables para poder representar claramente una enfermedad. De esta manera, en la figura 2 se muestra los niveles de especificidad y sensibilidad de lo síntomas de la Estemfiliosis del espárrago que veremos más adelante en la sección de experimentos.



	RS	TPN	TCP	PC
Sensibiilidad	0.836	0.823	0.82	0.849
Especificidad	0.862	0.909	0.931	0.826

Figura 2: Niveles de sensibilidad y especificidad de los síntomas de la Estemfiliosis

Dicho de otro modo, la sensibilidad es un valor cuantitativo, entre cero y uno, que explica el nivel de relación entre aparición-síntoma aparición-enfermedad, es decir, cuanta probabilidad hay de que al aparecer dicho síntoma, aparecerá la enfermedad.
Por otro lado, la especificidad es un valor cuantitativo, entre cero y uno, que explica el nivel de relación entre ausencia-síntoma ausencia-enfermedad, indicando la probabilidad de que al no existir el síntoma, no aparezca la enfermedad.
Para realizar una selección adecuada de síntomas que expliquen acertadamente la aparición de una enfermedad, es necesario seleccionar síntomas cuya relación de sensibilidad y especificidad con respecto a una enfermedad o plaga sea elevada. Tómese en cuenta que mientras más alta sea la relación, hay mayor seguridad de que los resultados serán idóneos. Considerar niveles de sensibilidad y especificidad por debajo de 70% no es muy recomendable.

2.3. Inferencia de la Red Bayesiana

Como hemos mencionado, el proceso de inferencia será basado en el algoritmo de paso de mensajes de Kim y Pearl [1], explicado también en [2], el cual, es similar al de inferencia por enumeración, sin embargo considera la comunicación bidireccional, es decir de padres a hijos y viceversa, lo que permite establecer razonamiento de diagnóstico como de predicción. Por otro lado, el algoritmo almacena valores parciales en base a los mensajes π y λ, recibidos de los nodos padres e hijos respectivamente

[2]. Esto permite realizar un pre-cálculo de probabilidades, lo que permite ahorrar tiempo para la siguiente vez en el que se tenga que realizar esta tarea.
A continuación describiremos los valores usados en el algoritmo de inferencia, teniendo en cuenta que un nodo actual *X* tiene un conjunto de padres *U* y un conjunto de hijos *Y*.

2.3.1 Probabilidad λ

Es el valor que se usa para realizar un diagnóstico. Se obtiene a través de la contribución de cada enlace saliente del nodo actual, es decir por el conjunto de nodos hijos.

$$\lambda_{Y_j}(X) = P(E_{Y_j/X} \mid X) \qquad (5)$$

Donde $E_{Y_j/X}$ es toda la evidencia conectada a Y_j a través de sus padres, excepto *X*.

2.3.2 Probabilidad π

Es el valor que da soporte a la inferencia predictiva, y es la contribución de cada enlace entrante a *X*, es decir, los padres de *X*.

$$\pi_X(U_i) = P(U_i \mid E_{U_i/X}) \qquad (6)$$

Donde $E_{U_i/X}$ son todas las evidencias conectadas a U_i con excepción, a través de *X*.

2.3.3 Mensaje λ

El nodo actual *X* calcula nuevos mensajes para enviarlos a cada uno de sus padres. En este caso la propagación del mensaje se realiza de abajo hacia arriba.

$$\lambda_X(u_i) = \lambda(x_i)\sum_{uk \neq ui} P(x_i \mid u_1, \ldots, u_n)\prod_{k \neq i} \pi_X(u_k) \qquad (7)$$

2.3.4 Mensaje π

El nodo *X* calcula los nuevos mensajes para enviarlos a sus hijos:

$$\pi_{Y_j}(x_i) = \begin{cases} 1 \text{ ,si el valor de evidencia } x_i \text{ es entero} \\ 0 \text{ ,si la evidencia es para cualquier otro valor } x_j \\ \alpha \left[\prod_{k \neq j} \lambda_{Y_k(x_i)}\right] \sum_{u_1,\ldots,u_n} (P(x_i|u_1,\ldots,u_n) \prod_i \pi_X(u_i)) \end{cases} \qquad (8)$$

Esto mantiene actualizado a sus nodos padres con respecto a la propagación de evidencia vía X.

EXPERIMENTOS Y RESULTADOS

Para la realización de pruebas, se emplearon características usuales (síntomas) de las enfermedades del espárrago. Los datos fueron elegidos según las características comunes de cada enfermedad.
De esta manera, para un campo de cultivo denominado *entorno A*, se modeló la red Bayesiana que se muestra en la figura 3.

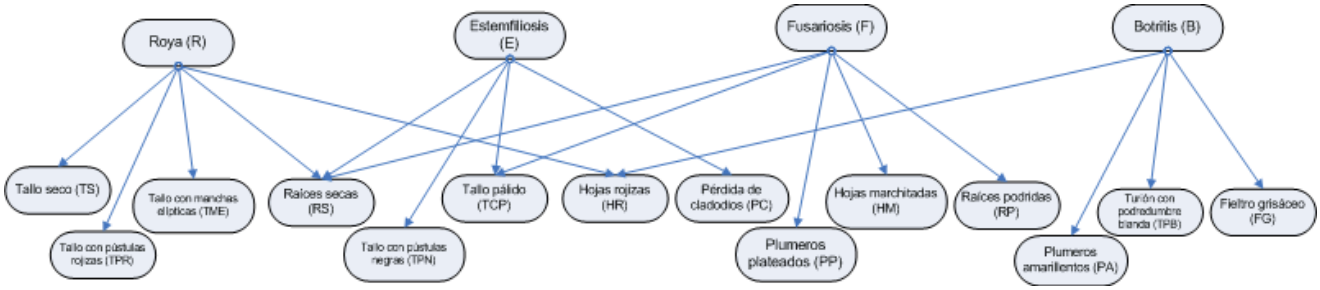


Figura 3: Red Bayesiana de diagnóstico modelada para el *entorno A*

Enfermedad	Consideración en el entorno
Roya	28.5 %
Estemfiliosis	13.6 %
Fusariosis	13.8 %
Botritis	9.44 %

Tabla 1: Se muestra la prevalencia de las enfermedades consideradas para el *entorno A*

Como se muestra en la figura 3, hay ciertos síntomas que influyen en más de una enfermedad. En la tabla 1, se muestra la prevalencia de las enfermedades consideradas para el entorno A.

P(E=1) = 0.136	$\Rightarrow$ P(E=0) = 0.864
P(PC=1 | E=1) = 0.849	$\Rightarrow$ P(PC=0 | E=1) = 0.151
P(PC=1 | E=0) = 0.174	$\Rightarrow$ P(PC=0 | E=0) = 0.826
P(TCP=1 | E=1)=0.802	$\Rightarrow$ P(TCP=0 | E=1)= 0.198
P(TCP=1 | E=0)=0.069	$\Rightarrow$ P(TCP=0 | E=0)= 0.931
P(RS=1 | E=1) = 0.836	$\Rightarrow$ P(RS=0 | E=1) = 0.164
P(RS=1 | E=0) = 0.138	$\Rightarrow$ P(RS=0 | E=0) = 0.862
P(TPN=1 | E=1) = 0.823	$\Rightarrow$ P(TPN=0 | E=1) = 0.177
P(TPN=1 | E=0) = 0.091	$\Rightarrow$ P(TPN=0 | E=0) = 0.909

Tabla 2: Se muestra los niveles de sensibilidad y especificidad de los síntomas para la enfermedad Estemfiliosis (E). Cuando E=0, indica ausencia de la enfermedad Estemfiliosis, E=1 indica presencia de la enfermedad, esto cumple para los demás casos.

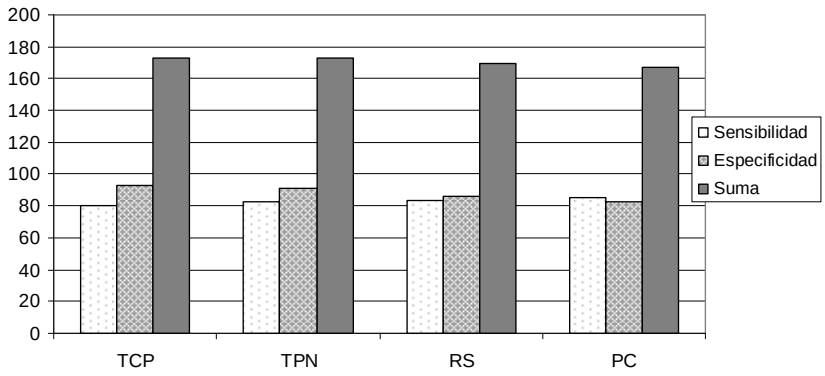


Figura 4: Muestra la sensibilidad y especificidad en porcentaje, de los síntomas típicos de la Estemfiliosis así como la suma de sensibilidad y especificidad para cada síntoma.

Síntoma/ Enfermedad	Roya (%)	Estemfiliosis (%)	Fusariosis (%)	Botritis (%)
Tallo seco	4.78	0.58	1.17	0.33
Tallo con pústulas rojizas	5.11	0.58	1.17	0.33
Tallo con manchas elípticas	2.33	0.58	1.17	0.33
Raíces secas	1.75	2.96	4.86	0.33
Tallo con pústulas negras	0.84	3.91	1.17	0.33
Tallo pálido	0.84	4.42	3.92	0.33
Hojas rojizas	1.62	0.58	1.17	1.73
Pérdida de cladiodos	0.84	2.49	1.17	0.33
Plumeros plateados	0.84	0.58	4.24	0.33
Hojas marchitadas	0.84	0.58	1.81	0.33
Raíces podridas	0.84	0.58	3.04	0.33
Plumeros amarillentos	0.84	0.58	1.17	2.34
Turión con podredumbre blanda	0.84	0.58	1.17	3.08
Fieltro grisáceo	0.84	0.58	1.17	2.93

Tabla 3: Niveles de presencia para cada enfermedad, dada la presencia de un síntoma en el *entorno A.*

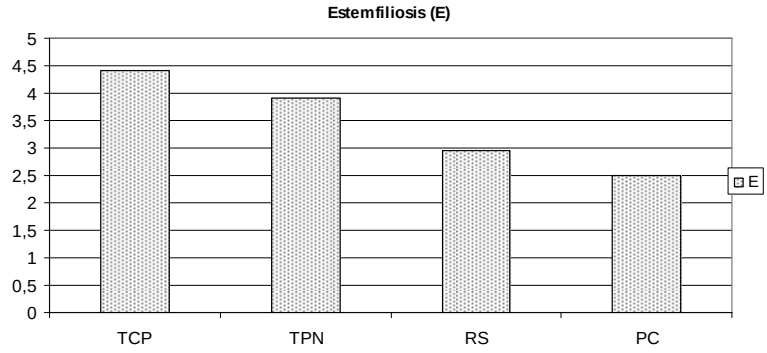


Figura 5: Muestra el nivel de presencia (en porcentaje) de la enfermedad Estemfiliosis cuando aparece un determinado síntoma que está relacionado con la enfermedad.

Síntomas/ Enfermedad	Roya (%)	Estemfiliosis (%)	Fusariosis (%)	Botritis (%)
TS, TPR	23.95	0.58	1.17	0.33
TS, TME	12.82	0.58	1.17	0.33
TPR, TME	13.44	0.58	1.17	0.33
TS, TPR, TME	56.7	0.58	1.17	0.33

Tabla 4: Dada la presencia de síntomas exclusivos de la enfermedad Roya, se muestran los resultados en el diagnóstico general.

En la tabla 2, se muestra la relación de sensibilidad y especificidad de los síntomas relacionados a la Estemfiliosis, la figura 4 aclara la idea presentando la sensibilidad,

especificidad y la suma de ambas; que representa finalmente el grado de interrelación entre el síntoma y la enfermedad.
En la tabla 3, se muestran los resultados de diagnosticar dado algún síntoma presentado, logrando obtener resultados acordes con la relación entre al nivel de sensibilidad y especificidad que tiene un síntoma con la enfermedad. Por ejemplo, puede observarse que la enfermedad Estemfiliosis obtiene un alto porcentaje de presencia cuando se manifiesta el síntoma de tallo pálido, cuya suma de nivel de sensibilidad y especificidad es la más alta entre los otros síntomas de la Estemfiliosis (ver figura 4). La figura 5 muestra el nivel de presencia de la Estemfiliosis dado un síntoma propio de ésta, y pueden observarse las figuras 4 y 5 en donde se aprecia la relación directamente proporcional entre la suma de sensibilidad y especificidad con la probabilidad de aparición de la enfermedad Estemfiliosis.
En la tabla 4, se muestra un diagnóstico, presentando un conjunto de síntomas, todos pertenecientes a la enfermedad de la Roya. Puede observarse que ningún síntoma afecta a otras enfermedades, salvo a la Roya, puesto que son síntomas exclusivos de esta enfermedad; además el nivel de presencia de la Roya varía según la combinación dada de síntomas.

CONCLUSIONES

Los síntomas con baja especificidad y sensibilidad, cuando no son visibles, permiten establecer mucha más duda de que la enfermedad pueda aparecer, aún teniendo síntomas con altas sensibilidades visibles, por ello es recomendable usar siempre síntomas que traten de caracterizar en lo mejor posible a una enfermedad, teniendo altas sensibilidades y especificidades para tener mucha más precisión a la hora de responder una consulta. Ello requiere que los estudios estadísticos acerca del entorno donde se va a implantar el sistema, tengan mucha confiabilidad en los datos, acerca de experiencias pasadas, entre las relaciones de los síntomas y la enfermedad (o enfermedades).

El sistema es totalmente ajustable a cualquier entorno, con el estudio estadístico de síntomas y enfermedades correspondientes. Esto permite adaptar los resultados de diagnóstico al mismo lugar en donde se implantará el sistema, lo cual es muy ventajoso puesto que los resultados llegan a ser más fiables, si el estudio estadístico del entorno es correcto.

Referencias

[1] Kim J. and Pearl J. A computational model for causal and diagnostic reasoning in inference systems. *In Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, (1983), pp. 190-193.

[2] Korb K.B. and Nicholson A.E. *Bayesian Artificial Intelligence*. Chapman & Hall/CRC. 2004.

[3] Pearl J. Fusion, propagation and structuring in belief networks. *Artificial Intelligence*. Vol 29, (1986), pp. 241-288.

[4] Russell S. and Norving P. *Artificial Intelligence: A modern Approach*. Second Edition, Prentice Hall. 2003.

[5] Riley, M.B., M.R. Williamson and O. Maloy, 2002. *Plant disease diagnosis. The Plant Health Instructor*, 10.1094/PHI-I-2002-1021-01

[6] Rajkishore, P., K. Ranjan and A.K. Sinha, 2006. *AMRAPALIKA: An expert system for the diagnosis of pests, diseases and disorders in Indian mango. Knowledge-Based Syst*., 19: 9-21.

[7] Lopez-Morales, V., O. Lopez-Ortega, J. Ramos-Fernandez and L.B. Munoz, 2008. *JAPIEST: An integral intelligent system for the diagnosis and control of tomatoes diseases and pests in hydroponic greenhouses*. Expert Syst. Appl., 35: 1506-1512.

[8] Gonzalez-Andujar, J.L., 2008. *Expert system for pests, diseases and weeds identification in olive crops*. Expert Syst. Appl., 10.1016/j.eswa.2008.01.007