Reducción del coste computacional de la Inferencia de Contexto Bayesiana mediante el uso de rangos de valores dinámicos

Sergio Fortes[(1)], Korbinian Frank[(2)], Raquel Barco[(3)].
sergio.fortes.rodriguez@ieee.org, korbinian.frank@dlr.de, rbm@ic.uma.es.
[(1)]Dpto. de Ingeniería de Comunicaciones. Universidad de Málaga. 29071 Málaga, Spain.
[(2)]Institute of Communications and Navigation. German Aerospace Center (DLR). 82234 Oberpfaffenhofen, Germany.
[(3)]Dpto. de Ingeniería de Comunicaciones. Universidad de Málaga. 29071 Málaga, Spain.

***Abstract*— This paper shows how to reduce evaluation time for context inference. Probabilistic Context Inference has proven to be more powerful and better adapted to the challenges of the physical reality with uncertain or missing information. As the inference complexity is very high, the complexity of the to be evaluated rule (representing a share of the real world) should be reduced as far as possible. Therefore we present an approach to select only relevant values of context types and to adapt this selection during its usage time. In an evaluation we show that with only a few evaluations of the reduced inference rules the reduction costs will have amortized and the system brings significant benefit to context aware computing.**

I. Introducción

A. Motivación

Los sistemas de conciencia de contexto o *context aware* representan el próximo paradigma de computación y de los sistemas de telecomunicación. Los computadores se integrarán usando información de contexto y, además, obteniendo conclusiones a partir de él. Este proceso es denominado *inferencia*. Informaciones de "contexto de alto nivel", cómo por ejemplo la actividad actual del usuario, serán la base de multitud de acciones proactivas, preferencias y procesos de toma de decisiones. Las redes Bayesianas (*Bayesian Networks, BN*) son consideradas como la opción más factible a la hora de modelar la realidad y proveer inferencia con la suficiente precisión y rendimiento [1]. En ellas la aplicación de información conocida (*evidencia*) sobre ciertas *variables* de la red permiten obtener la probabilidad de que el resto de las variables tomen un cierto valor. Nuestro estudio se centra en las BNs discretas, es decir aquellas cuyas variables (o *nodos*) solo pueden tomar valores de naturaleza discreta. El conjunto de *valores*, también denominados *estados*, que puede adoptar cada variable es su *rango de valores* (*value range, VR*). En las BNs la relación entre las variables se expresa mediante el uso de probabilidades condicionales de cada estado dado un valor de las variables relacionadas. Un problema de la redes Bayesianas es el alto coste computacional del proceso de inferencia para redes con gran número de variables y/o estados. Una forma de reducir la complejidad de la evaluación es disminuyendo el número de probabilidades condicionales. La tabla de probabilidades condicionales (*Conditional Probability Table, CPT*) de una variable aleatoria con n variables causas o *padres*, tiene el tamaño $size \leq k^{n+1}$ con k_i siendo el número de *estados*, que el nodo (variable) i puede tomar y siendo $k = \max_{i=1}^{n} k_i$. Cualquier reducción en k implica una mejora sustancial en el tiempo de evaluación, por lo que si reducimos el número de estados obtendremos una importante reducción en los tiempos de cómputo de la evaluación o inferencia del sistema. Sin embargo la relevancia de cada estado depende del contexto: usuarios, servicios, tiempo, localización, ..., por lo que los VRs deberán ser modificados dinamicamente basándonos en los requisitos de cada momento. Este trabajo tiene por tanto el objetivo de desarrollar la metodología y la base matemática necesaria para la reducción de las CPTs, evitando en los posible la pérdida de precisión en la inferencia debida a esta simplificación. Esta reducción se realizará mediante la unión de estados.

B. Trabajos relacionados

La tarea de discretización o *repartición* de los rangos de valores ha sido desarrollada en un gran número de trabajos. Muchos de ellos se centran en la discretización de variables continuas para la clasificación o *classification learning*.

Este es el caso de Fayyad en [3], donde un *Entropy Minimum Description Length Principle* es usado para seleccionar recursivamente los límites en un proceso de discretización *top-down*.

Para aplicaciones Bayesianas, Barco y otros [4], [5] usan parte de estos conceptos para implementar diagnósticos en redes de comunicaciones móviles, analizando además diferentes técnicas y su rendimiento.

En el caso de Clarke y Burton en [6], se muestran métodos para repartición previamente o durante la construcción de la BN. Esta repartición sigue un esquema *bottom-up*, estableciéndose criterios de selección de estados y de punto de parada.

Muy cercano a nuestros objetivos, pero para redes Bayesianas Híbridas, se encuentra el proceso de *discretización dinámica en cualquier instante* desarrollado por Kozlov y

Koller [7]. En éste hacen uso de *la distancia de Kullback-Leibler* y versiones ponderadas de ésta, calculadas entre las diferentes densidades de probabilidad de la BN.

Con todo, el análisis de los trabajos relacionados nos indica la necesidad de estudios posteriores. Así aunque parte de la teoría y los métodos existentes puede ser reutilizados no existe un desarrollo completo adaptable para el uso de rangos de valores dinámicos en redes Bayesianas discretas.

II. Arquitectura

La Fig. 1 muestra la arquitectura desarrollada para permitir variaciones dinámicas (*repartición* o *rediscretización*) durante tiempo de ejecución, para un sistema de inferencia de contexto, en comparación con la arquitectura clásica. Un paso intermedio es introducido con anterioridad a la inferencia. En este paso son aplicables una serie de reglas y criterios para tomar decisiones y establecer como debe ser realizado el procesado o *repartición* de los rangos de valores. De este modo ciertos estados de cada nodo son unidos para reducir las CPTs.

III. Unión de Estados

Nuestro procesado de VRs seguirá una aproximación bottom-up, donde cada variable comienza con su VR completo. A partir de éste nuestro proceso consistirá en unir valores siguiendo métodos y criterios tendientes a reducir la pérdida de información y el coste computacional. En esta sección se establece los fundamentos matemáticos para la unión de estados.

El concepto de unión de estados implica una reducción de las CPTs de los nodos de la red. Este proceso supone la eliminación de las probabilidades condicionales de los estados originales a unir y la adición de las probabilidades condicionales calculadas para el nuevo estado fruto de la unión.

La repartición o rediscretización de los rangos de valores implica modificaciones en las tablas de probabilidades condicionales (*CPTs*) de los nodos. Estos cambios se realizan en dos etapas:

A. Modificaciones internas al nodo

Uniendo un subconjunto de valores, la probabilidad condicional de un nuevo estado $V = \mathbf{v_{stm}}$, fruto de la unión de varios estados originales $\mathbf{v_{stm}} = v_{stm_0} \cup v_{stm_1} \cup \dots$, dada una determinada configuración de sus variables padres, $\mathbf{\Pi_v} = \pi_\mathbf{v}$, es:

$$p(V = \mathbf{v_{stm}} | \mathbf{\Pi_v} = \pi_\mathbf{v}) = \sum_{v \in \mathbf{v_{stm}}} p(v|\pi_\mathbf{v}) \qquad (1)$$

B. Actualización de los Hijos

Después de la modificación de los rangos de valores de un nodo V causa o padre de otros, se hace necesaria la actualización de las CPTs de sus nodos *hijos*. Desarrollando la expresión de la nueva probabilidad condicional para un nodo hijo Y, obtenemos:

$$p(Y = y | \mathbf{\Pi'_y} = \pi'_\mathbf{y}) = \frac{\sum_{v \in \mathbf{v_{stm}}} p(y|v, \widehat{\pi_y})}{\sum_{v \in \mathbf{v_{stm}}} p(v, \widehat{\pi_y})} \qquad (2)$$

Donde esta ecuación provee la nueva probabilidad condicional de $Y = y$ dado el nuevo estado del nodo padre $V = \mathbf{v_{stm}}$, siendo $\mathbf{\Pi'_y} = \pi'_\mathbf{y}$ una cierta configuración de las variables padres, entre las que se incluye el nuevo estado de V. $\widehat{\pi}$ representa el subconjunto de la configuración π' de los padres de Y excluyendo a V. El término $p(v, \widehat{\pi_y})$ es la probabilidad conjunta de una configuración $\widehat{\pi_y}$ con cualquiera de los estados $V = v$ de los unidos en la fase anterior.

IV. Personalización de VRs: Extensión de Protección

Los métodos desarrollados con anterioridad a este trabajo se aplican en fases tempranas del proceso de inferencia (cómo veiamos en la figura Fig. 1) por lo que carecen de una adaptación real y dinámica a las necesidades de los usuarios y/o servicios que obtienen información a partir de la BN. Para incrementar el nivel de adaptación y personalización de los sistemas una metodología completamente nueva ha sido desarrollada.

En nuestro modelo un servicio puede establecer como estados de interés ciertos valores de uno o varios nodos de una red Bayesiana. A estos estados los denominaremos *estados de interés* del servicio. Parece lógico por tanto que debamos evitar que estos estados se vean sometidos a unión con otros en nuestros esfuerzos por reducir las CPTs de la red Bayesiana. Los estados libres de ser sometidos a unión con otros se denominarán *estados protegidos*.

Si se analiza el hecho de la fuerte relación que existe entre las probabilidades de un nodo con sus nodos relacionados nos damos cuenta que la protección de ciertos estados y la aplicación descuidada de métodos de unión en otros nodos supone un fuerte aumento en el error. Para evitarlo, deben ser igualmente protegidos estos estados, de otros nodos, que muestran una fuerte dependencia con respecto a los estados de interés.

Para ello y basándonos en [9], obtuvimos la siguiente expresión para el cálculo de la información mutua como medida de la dependencia entre el subconjunto de estados de interés $\mathbf{y_s}$ de un nodo Y y el estado de un nodo padre $X = x$.

$$I(Y = \mathbf{y_s}, X = x) = \sum_{y \in \mathbf{y_s}} p(x,y) \left| log(\frac{p(x,y)}{p(x)p(y)}) \right| \qquad (3)$$

Donde el término $\left| log(\frac{p(x,y)}{p(x)p(y)}) \right|$ provee información sobre cuán dependientes son los estados entre si. $p(x,y)$ pondera la expresión basándose en la probabilidad conjunta de los dos estados.

Una ecuación equivalente es desarrollada para el estado de un nodo hijo, $Y = y$, cuyo padre contiene un conjunto de estados de interés, $\mathbf{x_s}$:

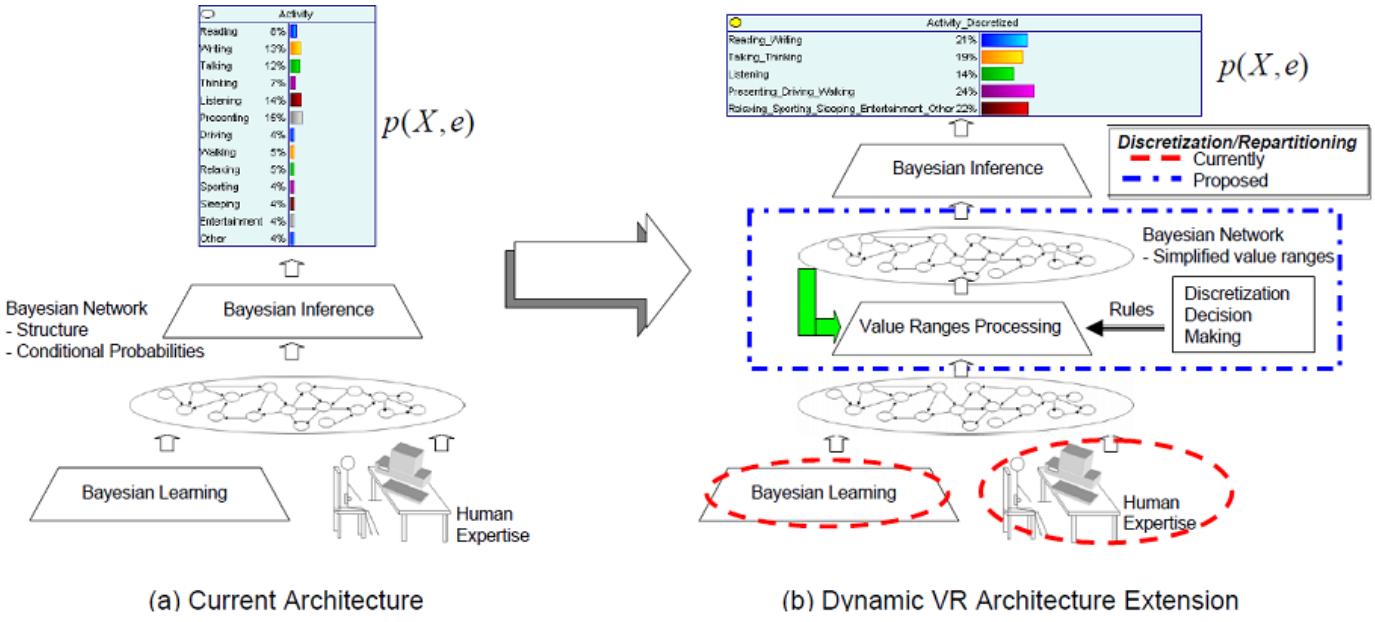

Fig. 1. Arquitecturas para la Inferencia de Contexto Bayesiana

$$I(Y = y, X = \mathbf{x_s}) = \sum_{x \in \mathbf{x_s}} p(x, y) \left| log(\frac{p(x, y)}{p(x)p(y)}) \right| \quad (4)$$

Por consiguiente, estableciendo un conjunto de estados de interés en ciertos nodos, es posible extender la protección hacia los estados de otros nodos que compartan la mayor información mutua y, por tanto, mayor efecto sobre su evaluación. Aplicando este proceso recursivamente, puede extenderse la protección de estados a lo largo de toda una red Bayesiana.

Esta extensión de protección o *protection extension* (PE_M), puede ser ajustada estableciendo un *mínimo valor de la información mutua para ser protegido*, M. A menor valor un mayor número de estados serán protegidos, mejorando la calidad de la inferencia pero reduciendo la simplificación de la BN.

Después de este proceso, sobre el resto de estados no protegidos pueden aplicarse otros criterios de repartición, o bien unir todos ellos en un único estado nuevo, que fue la solución adoptada para la fase de Evaluación. De este modo generamos versiones simplificadas de una BN original y adaptada a las necesidades del servicio que vaya a hacer uso de ella.

V. Evaluación de Resultados

En esta sección evaluaremos las principales características de una aplicación de la metodología descrita.

A. Entorno de Pruebas

Para probar las capacidades de los diferentes métodos se desarrolló un motor de inferencia Bayesiana y una aplicación-servicio básica, ambos en Java. Éste debería adaptar sus VRs a los requisitos de dicho servicio.

La red Bayesiana a usar fue implementada y diseñada mediante experto humano. Esta red representa el entorno de una oficina y tiene las siguientes características: 17 nodos, 7 relacionados con sensores y 133 estados en total para todas las variables de la red.

Con el propósito de evaluar los métodos descritos se desarrollo un servicio de ejemplo consistente en una pantalla de pared, como las que pueden ser encontradas a la entrada de grandes compañías. Ésta ofrecerá a los empleados un mensaje de bienvenida, información sobre el menú de un restaurante o información del tráfico dependiendo de la actividad actual del usuario. Esta información será producida y suministrada automáticamente por el motor de inferencia Bayesiana y sin intervención directa del usuario. Dado el servicio, los estados de interés serán `Llegando`, `YendoAComer` y `YendoACasa`, pertenecientes al nodo `Actividad`. Para reducir el consumo de recursos los VRs los nodos de la red serán reducidos.

La calidad del método desarrollado debe ser medida mediante el análisis dos características esenciales: el error en la inferencia introducido por el uso de rangos de valores simplificados y el consumo de CPU.

B. Error

Para analizar el error se comparan las probabilidades de cada estado obtenidas en la BN original con los resultados para las BNs con procesado de VRs, dada una cierta evidencia inducida en los nodos que representan sensores.

De este modo, considerando las probabilidades proporcionadas por la red sin simplificar como exactos, el error cometido por el uso de VRs simplificados para los estados del nodo `Actividad` se muestra en la Tabla I. Para los tres estados de interés del servicio: `Llegando`, LLG, `YendoAComer`, YCO y `YendoACasa`, YCA, los mejores

resultados se dan con extensión de protección con menor *mínimo valor de la información mutua para ser protegido*. Nuestro conocimiento del servicio nos indica, además, que estos valores de este error son aceptables para el servicio. El resto de los valores de la lista representan los demás estados del nodo de `Actividad` sobre los cuáles no nos extenderemos al carecer de valor para el servicio. Observamos además cómo el error provocado por el uso de los métodos PE crece en estos estados fuera del rango de interés.

	$PE_{0.03}$	$PE_{0.01}$
LLG	**1.03**	**1.99**
YCO	**17.24**	**2.34**
YCA	**17.24**	**2.34**
PRG	83.28	62.71
RDG	21.05	39.67
WRT	125.88	77.66
OTH	51.03	30.66

TABLE I
ERROR RELATIVO (EN %) DE LA INFERENCIA CON RESPECTO A SU VALOR EN LA RED SIN SIMPLIFICAR.

C. Coste Computacional

El análisis presentado en Fig. 2 muestra el número de inferencias ejecutadas en el tiempo para la red original no reparticionada, $NoRep$, y las BNs simplificadas mediante $PE_0.03$ y $PE_0.01$. Se observa cómo la primera evaluación se prodice con anterioridad en la red sin simplificar, debido a que no se ejecuta sobre ella el paso intermedio necesario para la unión de estados. Sin embargo pasada la segunda inferencia se evidencia como las redes simplificadas se evalúan mucho más rápidamente, proporcionando un mayor número de inferencias. De la gráfica se denota también que el menor error en la inferencia mediante $PE_{0.01}$, observado en el apartado anterior, se consigue a cambio de un mayor coste computacional con respecto a $PE_{0.03}$. Esto es así debido a que un mayor *mínimo valor de la información mutua para ser protegido* supone un menor número de estados a proteger, lo cuál proporciona unos VRs finales con un menor número de estados y por tanto menor coste computacional de la inferencia, pero a cambio de un mayor error.

VI. CONCLUSIÓN

Este documento muestra como la modificación dinámica de los rangos de valores es un factor de importancia para la inferencia de contexto en aplicaciones reales. Se propone una arquitectura capaz de realizar modificaciones en los VRs y diferentes métodos para determinar los estados relevantes.

En el capítulo de Evaluación se muestran los resultados del análisis de los sistemas desarrollados. Así se observa que el error introducido es aceptable, en los valores de interés, cuando se usa el método de *extensión de protección* desarrollado en este proyecto. Además el costo de generar los rangos de valores reducidos es ampliamente compensado por la reducción en la evaluación de la BN.

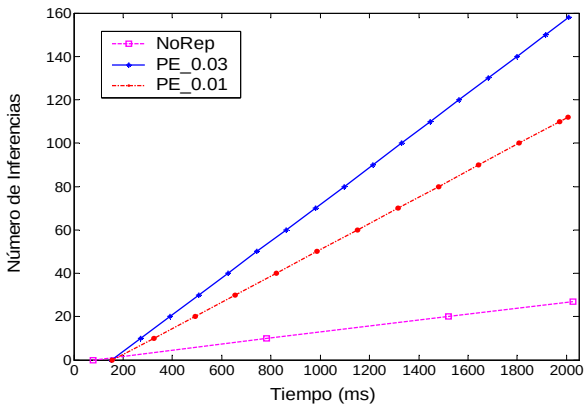

Fig. 2. Número de inferencias de la BN original y sus versiones simplificadas, ejecutadas en un PC de gama media actual.

Se pretende que estas metodologías se apliquen en el desarrollo de servicios en entornos completos de conciencia de contexto, conectando diferentes servicios y usuarios mediante sistemas de posicionamiento en interiores. Esperamos de este modo obtener información realista sobre frecuencias de actualización y modificación de VRs.

REFERENCES

[1] K. Frank, M. Röckl, P. Gallego Hermann, and M. T. Morillas Vera, "Knowledge representation and inference in context-aware computing environments," in *The Second International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM08)*, N. C. K.-C. P. M. D. A. Lloret Mauri, Jaime; Cardona, Ed. IEEE Computer Society Conference Publishing Services (CPS), 09 2008, pp. 89 – 95. [Online]. Available: http://elib.dlr.de/54514

[2] K. Frank, M. Röckl, and P. Robertson, "The Bayeslet concept for modular context inference," in *Proceedings of UBICOMM08*. Valencia, Spain: IEEE Computer Society, 2008.

[3] U. Fayyad and K. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proceedings of the International Joint Conference on Uncertainty in AI*, 1993, pp. 1022–1027.

[4] R. Barco, V. Wille, L. Díez, and M. Toril, "Learning of model parameters for fault diagnosis in wireless networks," *Wireless Networks*, 2008. [Online]. Available: http://dx.doi.org/10.1007/s11276-008-0128-z

[5] R. Barco, P. Lázaro, L. Díez, and V. Wille, "Continuous versus discrete model in autodiagnosis systems for wireless networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 6, pp. 673–681, 2008.

[6] E. J. Clarke and B. A. Barton, "Entropy and mdl discretization of continuous variables for bayesian belief networks," *International Journal of Intelligent Systems*, vol. 15, no. 1, pp. 61–92, 2000.

[7] A. V. Kozlov and D. Koller, "Nonuniform dynamic discretization in hybrid networks," in *In Proc. UAI*. Morgan Kaufmann, 1997, pp. 314–325.

[8] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *International Conference on Machine Learning*, 1995, pp. 194–202.

[9] N. Friedman and M. Goldszmidt, "Discretizing continuous attributes while learning Bayesian neworks," in *Proc. 13th International Conference on Machine Learning*. Morgan Kaufmann, 1996, pp. 157–165.