Probability Dynamics in Causal Networks: A Probabilistic Prediction Engine

Silvano Mussi
Member of AICA – Associazione Italiana per l'Informatica ed il Calcolo Automatico
Piazza Rodolfo Morandi 2, 20121 Milano, Italy
Email: mussisilvano [AT] tiscali.it

Abstract—**The proposal presented in the paper considers the following scenario. There is a population of subjects and an event E that can occur to each subject of the population. The population is constituted of subjects that, each with his/her specific past, begin, at a certain point of their life, a period of aging (observed-aging) under certain conditions (possibly changing in time). The probability of E occurrence to a subject during his/her observed-aging depends on the particular combinations of conditions under which the subject spends such a period of time. The basic question underlying the proposal is: if, given his/her specific past, a subject S had to spend his/her next future time ΔT under certain (simulated) combinations of conditions, what would be the occurrence probability of E to S at the end of the future time interval ΔT ?**

The paper presents both a method for step by step modelling the physical world and the mathematical model of a probabilistic prediction engine that provides answer to the basic question.

Moreover the basic general laws of probability dynamics in causal networks are also illustrated. This is motivated by the fact that these laws underlie the probabilistic reasoning that produces the prediction algorithm. The prediction engine, presented in this paper at a theoretical level, has also been implemented in a software prototype in turn embedded in a general software environment (www.cheerup.it) for creating and administering specific application oriented predictive tools in heterogeneous fields .

The proposal can be useful in all the problems in which there is the need to take suitable and personalized measures in advance. For example, in the field of medicine, the proposal might be useful to Predictive, Preventive and Personalized Medicine. In fact the prediction engine, producing probabilistic predictions might be useful to Predictive Medicine, allowing to take suitable measures in advance might be useful to Preventive Medicine, producing specific predictions for each specific subject might be useful to Personalized Medicine.

Keywords—*Computer applications, Knowledge engineering, Decision support systems, Predictive monitoring.*

I. Introduction

In order to allow the reader to briefly get an intuitive and synthetic view of the proposal let us organize this section as an explanation of the title of the paper. Precisely, let us first explain the terms used in the first part of the title, then let us pass to examine the terms used in the second part.

A. Probability dynamics in causal networks : what are we talking about ?

"I consider the world probability as meaning the state of mind with respect to an assertion, a coming event, or any other matter on which absolute knowledge does not exists" – August de Morgan, 1838. Such citation, that appears at the beginning of the book [1], points out the natural concept of probability. Probability is not a property of events of the physical world, but a mental state, a psychological state (belief, trust, expectation, etc.) of who asks, with respect to his/her own information state, about the occurrence of future events. The term "probability degree" has therefore the meaning of "degree of belief, trust, expectation, …" about the occurrence of a future event, in synthesis: how much one believes that a certain event E will occur.

As uncertainty is pervasive in all fields (especially in medicine), the probability calculus about the occurrence of an event supports the rational decision (that cannot be the winner one) about which measure should be taken in advance in order to prevent/favour the occurrence of an undesired/desired event E. For decades hundreds of researchers have faced the problem of how to represent and manage uncertainty [2] and a great number of papers have been published on this topic. Let us avoid facing such a crucial subject, it would be beyond the scope of the paper. The proposal presented in this paper follows the probabilistic approach. Uncertainty is represented in probabilistic terms and probability theory is adopted as the theoretic and mathematical foundation of the mental state called probability.

In many kinds of problems pervaded by uncertainty the probability concept is effective if it is intended as conditioned probability playing the role of strength of causal relation. More precisely, the world is modelled in two steps. The first step consists in modelling the world in terms of cause-effect relations building this way a causal network whose nodes are multistate variables. The second step consists in putting uncertainty into the causal network, that is assigning conditioned probabilities to the causal relations of the network. A causal network with conditional probabilities is called probabilistic network. The probabilistic network is a basic conceptual paradigm for modelling real world under uncertainty [1], [3]. In each node of the network, probability is distributed (as a sort of fluid) on the states of the node and the total sum of the single quantities is 1. When an evidence enters the network, that is when a node is instantiated to one of its states, the consequences of this fact propagate through the network. Such consequences consist in redistributing the probabilities on the states of the remaining nodes. Probability calculus in probabilistic networks consists in calculating the probability distributions on the states of the end nodes given the probability distributions on the states of the start nodes. End nodes are the ones that, representing future events, cannot be directly observed and yet are relevant for taking measures in advance. The proposal concerns both a method for building a type of probabilistic network suitable to certain categories of problems (see the scenario illustrated in the next sub-section), and the definition of an algorithm for probability calculus in such type of network: the Probabilistic Prediction Engine.

After this general premise, necessary for the sake of conceptual clarity, let us pass to examine the second part of the title.

B. *Probabilistic Prediction Engine: what is it ?*

Let us start this section by examining the general scenario considered by the proposal. Let us consider a population of subjects and an event E that can occur to each subject of the population. The population is constituted of subjects that, each with his/her specific past, begins, at a certain point of their life, a period of aging under certain conditions (possibly changing in time), period that we will call *observed-aging*. What is interesting is the fact that the probability of E occurrence to a subject during his/her observed-aging depends on the particular combinations of conditions under which the subject spends such a period of time. For the sake of clarity let us consider the following example concerning the occupational medicine. A subject S, with his/her own chronological age and specific past (familial anamnesis, operations and/or illness undergone, etc.), at a certain point of his/her life begins a certain working activity that requires to spend a long period of time under certain working conditions harmful to health (for example, unhealthy environmental conditions, stress conditions, etc.). If E represents a professional illness typical of that working activity, it can be stated that such working conditions, lasting in time, affect the occurrence probability of E to S.

Turning back to the general scenario, let us consider the following basic question: if, given his/her specific past, a subject S had to spend his/her next future ΔT time under certain (simulated) combinations of conditions, what would be the occurrence probability of E to S at the end of the future time interval ΔT ? Answering this question is the goal of the proposal.

In order to allow the reader to better understand the basic question and what the proposal consists in, it is necessary to informally introduce some basic concepts.

At the moment of beginning of the observed-aging on behalf of a subject S, an *initial pro*file of S is built. The initial profile is an informative record that describes the status of S at the beginning of his/her observed-aging: the age range (in the following called *initial age-range*) to which the current chronological age (expressed with a proper time unit) of S belongs, and a set of *initial information* concerning the past and present status of S (anamnesis information, familial predispositions to E occurrence, current health state, operations undergone in the past, illness undergone (e.g. hepatitis, cardiac infarcts), or still present (e.g. diabetes), etc.).

After a subject S has passed a certain period of observed-aging, the information that specifies for how long (in the observed-aging period) S has been under the condition C1, for how long under the condition C2, etc. constitutes the *profile of observed aging* of S. Let us call *whole profile* of S the informative set: initial profile (of S) AND profile of observed aging (of S). Finally let us take into account if E has occurred or not to S. The period of observed-aging of S ends when such piece of information is: E has occurred to S.

The Probabilistic Prediction Engine presented in the proposal and implemented in a software prototype [4] provides the possibility to simulate that a subject S will spend a future time interval ΔT under certain combinations of conditions defined by the user (of the Engine). In other words, given a subject S, the Engine allows the user to define (for S), for a future time interval ΔT, a hypothetical profile of observed aging, profile that we will call *future profile of observed aging*. In this case the whole profile of S extends even to include the simulated future: whole profile = initial profile AND extended profile of observed aging, where *extended profile of observed aging* = “profile of observed aging” & “future profile of observed aging” (& is a merging operator). The user, after having defined a simulated future, can ask the Engine to calculate the answer to the basic question (section 4 will illustrate how such a calculus is accomplished). Let us notice that the possibility to compare different predictions resulting from different simulated cases may help decision making in trade-off problems.

Let us also notice that the calculated probability values are sufficiently reliable only if the Engine has at its disposal a great number of subject stories. How to collect such stories? As it is well known, there are two methods for filling a database: either by means of batch loading of data or by means of collecting every single story in real time. This second method is carried out by monitoring single subjects, monitoring that, being equipped with prediction features, is called: predictive monitoring.

There are many fields (not necessarily the medicine field) in which the proposal could be applied. In general the proposal could be a useful support to Predictive, Preventive and Personalized Medicine. In fact probabilistic predictions may be used in Predictive Medicine. Predictions may be used in Preventive Medicine for taking suitable measures in advance. Predictions, being specific for each specific subject, may be used in Personalized Medicine.

Let us also notice that the event E might also be a desired event. For example, in the fields of fitness or sport what might be interesting is calculating the probability that a subject, given his/her initial profile, has to reach a certain target (event E) if he/she spends his/her next years under certain conditions of physical exercise, diet, etc.

C. Paper organization

Section 2 "Step by step modelling the physical world" presents the general conceptual tools needed to build a real world model suitable to the proposal. Section 3 "Modelling the dynamics of probability in a causal network: a qualitative view" presents the concept of probability dynamics in causal networks. Section 4 "Computational probability dynamics: Probabilistic Predictive Engine" presents the Probabilistic Predictive Engine. Sections 5 "Discussion and related work" and 6 "Conclusions" terminate the paper.

II. STEP BY STEP MODELLING THE PHYSICAL WORLD

This section, that concerns the construction of a model that will be used by the Probabilistic Prediction Engine, is organized in three sub-sections. The first sub-section presents the formal definitions of the basic concepts underlying the theoretical model of the proposal and informally illustrated in section 1. The second sub-section introduces the concepts of time-slice, concept necessary for modelling the real time. Finally the third sub-section illustrates how the concepts presented in the preceding sub-sections are organized in order to build a suitable causal network: the basic conceptual platform for the Probabilistic Prediction Engine.

A. Concept of observed-aging

- Let us represent the event E by means of a <u>variable</u> with two states: "occurred", "not-occurred". For short the notation E=n stands for E = not-occurred, E=y for E = occurred.
- Let us define the concept of "*state of a condition* C1".

A condition C1 is represented by a variable with a set of possible states s1, s2, … . The set of possible states of C1 is denoted by C1={s1, s2, ...}, and C1=s1 means: "the condition C1 is in state s1" or, equivalently, "the condition C1 is instantiated to the state s1". For example, the condition "Cigarette smoke" might have three states: "yes under 10 cigarettes a day" (s1), "yes 10 or more cigarettes a day" (s2), "no" (s3).

- Let us define the concept of "*observed-aging*, of a subject S1, related to a condition C1". Let us consider a condition C1. Let $\Delta T = [0 , z)$ be the period of observed-aging of a subject S1. It is reasonably to consider the possibility that S1 elapses ΔT by spending n time units with C1=s1, then m time units with C1=s2, etc. Let us adopt the following dot notation: <condition state>.<number of time units> to represent the fact that a subject spends a certain number of time units under a certain condition state. As a consequence, the present concept is formally defined by the sequence $S1_{C1}$ = (st1.n, st2.m, ...), where $n + m + \ldots = \Delta T$.
- Let us define the concept of "<u>variable</u> of observed-aging related to a condition C1". Let us consider a set of subjects {S1, S2, S3, …}. The present concept is represented by the variable HC1 whose set of values is $\{S1_{C1}$, $S2_{C1}$, $S3_{C1}$ …}. HC1 is a *history variable*, it stands for "History of C1".

In this sub-section we have so far formalized the concept of observed-aging and its consequences, but, since such concept involves time, we are prompted to face the problem of defining a proper model of time.

B. Concept of time slice

The concept of observed-aging of a subject implies that we are interested in time elapsing only after the beginning of the observed-aging (before the beginning of the observed-aging the past of a subject is summarized in his/her initial profile). In other terms, the beginning of the observed-aging can be viewed as a sort of "birth" for a subject, i.e. the birth of his/her observed-aging. Let *tu* be the time unit suitable for measuring the observed-aging in a given application. Let us subdivide the time axis into segments, each with length equal to 1 *tu* . More precisely, let us subdivide the time axis into the intervals: [0 , 1) , [1 , 2), …, where 0 is the point on the time-axis that corresponds to the birth of the observed-aging.

In real life, time elapses in a continuous way, but in a computational model we must get time to elapse in a discrete way, that is as a sequence of time-slices: time-slice 1, time-slice 2, etc. (for short, time-slice 1 is denoted by ts1, time slice 2 by ts2, etc.). Precisely, the time interval [0 , 1) concentrates in ts1, the time interval [1 , 2) concentrates in ts2, etc. Each time slice collects both the history-variables and the E-variable. Let us refer, for example, to the condition C1. In ts1 there is $HC1_1$ that means HC1 with ΔT = [0 , 1), in ts2 there is $HC1_2$ that means HC1 with ΔT = [0 , 2), etc. So, in general, given a set of conditions {C1, C2, …, CN}, in a time slice tsm (m>0) there are: $HC1_m$, $HC2_m$, …, HCN_m , with ΔT = [0 , m). As for E, in tsm there is also the variable E_m . If in ΔT = [0 , m) E has not occurred, tsm has E_m=n, if E has occurred, tsm has E_m=y. Let us real world E occurs inside the interval [2 , 3) and the sub-interval before E occurrence elapses with C1=s2. Such a situation is represented in the model by: {HC_1=(s1.1), E_1=n} in ts1; {HC_2=(s1.1,s2.1), E_2=n} in ts2; {HC_3=(s1.1,s2.2), E_3=y} in ts3. The period of observed-aging of the current subject terminates with ts3. In point 0 we have the special time slice ts0 with only E_0=n. After having formalized the concept of time slice, let us consider again the concept of observed-aging by examining its consequences in terms of time slices. To this end let us face the

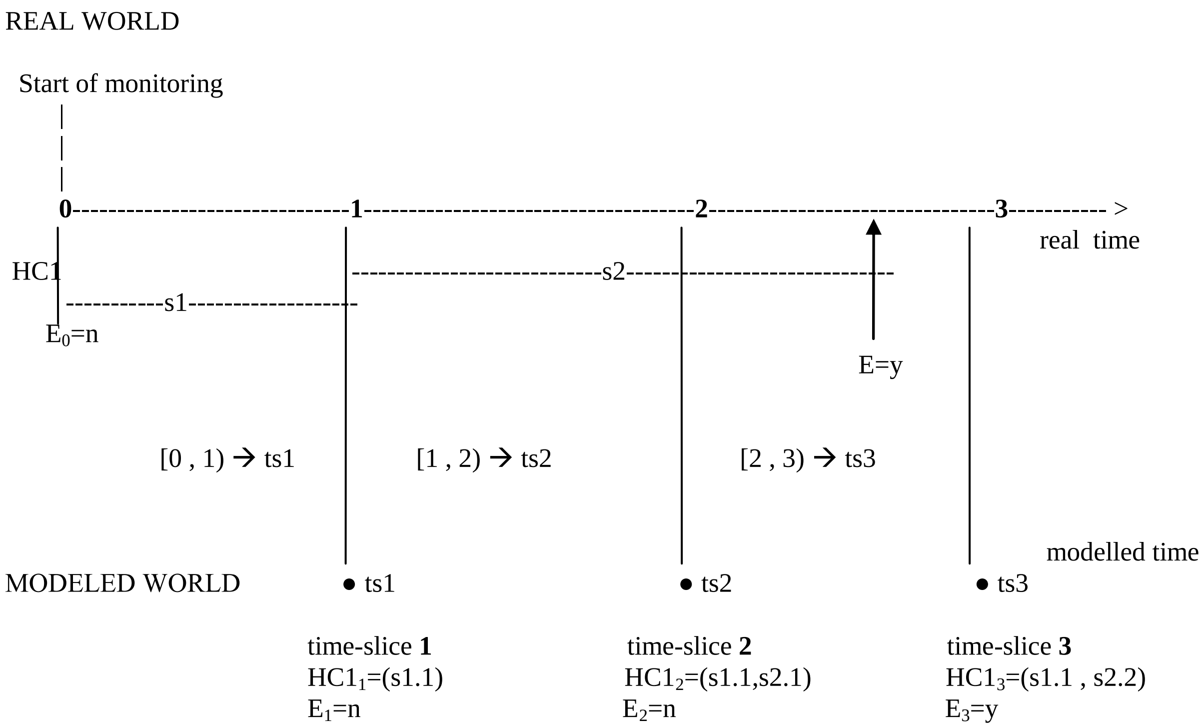


Fig. 1 An application example of the time-slice concept.

suppose that b is the point of E occurrence in [m-1 , m). If in the sub-interval [m-1 , b) a condition C is in state st, then let us make an approximation by assuming that such state lasts even during the remaining sub-interval [b , m).

The concept of time slice is synthetically illustrated in fig. 1. Let us briefly comment it. For the sake of simplicity the figure considers a hypothetical period of observed-aging constituted by only 3 time slices. Moreover C1 is the only condition being considered. The interval time defined, in the real world, by [0 , 1), is concentrated, in the model, in a single point: the time-slice 1 (i.e. ts1), the interval [1 , 2) is concentrated in ts2, etc. In the definition of the profile concept.

C. *Concept of profile*

Let us define, in more formal terms, the concept of initial profile already presented in section 1 in informal terms. Let {IF1, IF2, …} be the set of initial information.

- Let us define the concept of "*state of a piece of initial information* IF1". A piece of initial information IF1 is represented by a variable with a set of possible states: s1, s2, … and IF1=s1 means "the piece of initial information IF1 is in state s1" or, equivalently, "the piece of initial information IF1 is instantiated to the state s1". For example, the piece of initial information "Genetic predisposition" might have two states: "yes" (s1), "no" (s2).

● Let us define the concept of "*initial information* of a subject S1". Let us consider all the single pieces of initial information: IF1, IF2, IF3, … On the basis of the real history of a subject S1 each single piece of initial information has a proper instantiation to one of its possible states. For example IF1=s2, IF2=s1, IF3=s1, … The present concept is represented by such a set of instantiations of the initial information.

● Let us define the concept of "*initial profile* of a subject S1". Let us define the present concept as: "initial age-range of S1" AND "initial information of S1". The concept is denoted with $iprof_{S1}$.

● Let us define the concept of "variable of initial profile". This concept is represented by the variable IP whose set of values is defined by the initial profiles of the single subjects.

● Let us define the concept of "*profile of observed-aging* of a subject S1 as long as a time slice k". Let us consider the set of conditions {C1, C2, …, CN} and a subject S1. Let us define the present concept as the set: $\{S1_{C1}, S1_{C2}, \ldots, S1_{CN}\}$ with $\Delta T = [0 , k)$, and let us denote it with $prof_{S1,k}$.

As already stated in section 1, the Prediction Engine provides the possibility to simulate that a subject S1 spends a future time interval Δt under certain combinations of condition states. In other words, it is possible to define, for a future Δt, different possible simulated profiles of observed-aging. We are therefore prompted to formulate two new definitions.

● Let us define the concept of "*future profile of observed-aging* of a subject S1 referred to the time interval between the present time slice q and the future time slice k". Let us consider the set of conditions {C1, C2, …, CN} and a subject S1. Let us define the present concept as the set: $\{S1_{C1}, S1_{C2}, \ldots, S1_{CN}\}$ referred to the time interval $\Delta t = [q , k)$, where q is the present time slice and k is the future time slice ($q < k$). Let such profile be denoted with $fprof_{S1,q,k}$.

● Let us define the concept of "*profile of observed-aging* of a subject S1 *extended* to a future time slice k". The present concept is represented by the profile resulting by extending the profile of observed-aging of a subject S1 $prof_{S1,q}$ so to also include the future profile $fprof_{S1,q,k}$. In formal terms, let us state: $prof_{S1,k} = prof_{S1,q} \,\&\, fprof_{S1,q,k}$, where the symbol & stands for "merging". That is, the sequences <condition state>.<number of time units> of the two profiles are to be merged. Let us explain what it means. Let us refer, for the sake of simplicity, to the condition C1 only. Let us suppose, for example, that in the segment $\Delta T = [0 , q)$ the sequence $S1_{C1}$ is (s1.a, s2.b, s1.c) where $a + b + c = \Delta T$. Let us suppose that in the segment $\Delta t = [q , n)$ the sequence $S1_{C1}$ is (s1.d, s2.e). Then in the segment $\Delta T = [0 , n)$ the sequence $S1_{C1}$ is the given by: (s1.a, s2.b, s1.c) & (s1.d, s2.e) = (s1.a, s2.b, s1.(c+d), s2.e).

Time slice	**Initial age-range**	**IF1**	**IF2**	**HC1**	**HC2**	**HC3**	**$E_Y$prof**	**$E_N$prof**
….								
3	30-35	s1	s2	(s3.2, s1.1)	(s1.3)	(s2.3)	2	113
3	35-40	s2	s2	(s3.3)	(s2.3)	(s2.1, s1.2)	3	124
3	35-40	s2	s1	(s3.3)	(s2.3)	(s2.1, s1.2)	1	57
3	40-45	s2	s1	(s3.3)	(s2.3)	(s2.1, s1.2)	5	98
…								
…								
4	30-35	s1	s2	(s3.2, s1.2)	(s1.4)	(s2.4)	2	113
4	30-35	s1	s2	(s3.2, s1.2)	(s1.4)	(s2.4)	3	113
…								
5	30-35	s1	s2	(s3.2, s1.3)	(s1.5)	(s2.4, s1.1)	2	125
…								

Fig. 2 An example of how profiles are stored in model memory. In the example, that represents a cross-section of the database, the initial profile is based on initial age range and two variables (IF1, IF2) whereas the observed-aging profile is based on three history-variables (HC1, HC2, HC3).

● Let us define the concept of "*whole profile* of a subject S1 as long as a time slice k". This concept is defined as: $wprof_{S1,k}$ = $iprof_{S1}$ AND $prof_{S1,k}$.

In general, let us use the symbols $wprof_k$, $prof_k$ to respectively indicate a whole profile and a profile related to a time slice k without making explicit the subject the profiles belong to. Similarly, let us use the symbol iprof to indicate an initial profile without making explicit the subject the initial profile belongs to.

● Let us define the concept of "counter variable related to a whole profile $wprof_k$". During the period of observed-aging, subjects give their contributions by adding, for each time slice k, their whole profiles $wprof_k$ to the database. Let us define, for each $wprof_k$, two counter variables: $E_N wprof_k$ and $E_Y wprof_k$. If the $wprof_k$ of the current subject is already present in the database, then one of the two variables will be incremented. Precisely, if E has not occurred, $E_N wprof_k$ is incremented. If E has occurred, $E_Y wprof_k$ is incremented. As a consequence the variable $E_N wprof_k$ contains the number of cases (having profile $wprof_k$) with E_k=n whereas the variable $E_Y wprof_k$ contains the number cases (having profile $wprof_k$) with E_k=y. Let us notice that the fact that $wprof_k$ is added to the database (or, equivalently, the fact that the related counter variables are updated) implicitly implies the fact that the state of E in the preceding time slice k-1 is necessarily in state n (in fact, as stated in section 1.2, the period of observed-aging of a subject S ends when E occurs to S). Formally, for each $wprof_k$ belonging to the time slice k, we should take into account the fact E_{k-1}=n. Finally, let us notice this important fact: the value of $[E_Y wprof_k / (E_N wprof_k + E_Y wprof_k)]$ represents the frequency of occurrence of E given the profile $wprof_k$ and E_{k-1}=n. For short let us denote such value with the symbol $L(wprof_k)$. Formally: $L(wprof_k)$ = frequency-of-occurrence(E_k =y | E_{k-1}=n, $wprof_k$). As a consequence, assuming that the value of ($E_N wprof_k + E_Y wprof_k$) is enough great, if we adopt the frequentist definition of probability, we can state that, with a certain approximation, $L(wprof_k) = P(E_k = y \mid E_{k-1}=n, wprof_k)$.

Figure 2 shows how profiles are represented in the database. Let us briefly comment it. Let us consider for example the time slice 3. There are four rows referring to it. Each row represents the whole profile of a subject in the time slice 3. Let us consider, for example, the first row. The initial profile is represented by {init. age-range = 30-35 , IF1 = s1 , IF2 = s2}. The profile of observed-aging is represented by { HC1 = (s3.2,s1.1) , HC2 = (s1.3) , HC3 = (s2.3) }. Finally let us consider the counter variables: $E_Y wprof_3$ = 2, $E_N wprof_3$ = 113. As a consequence $L(wprof_3)$ = 2 / 115 = 0.017.

In this sub-section we have formalized the concept of profile and its consequences. We have now at our disposal the concept of time slice and the concept of profile and we are therefore ready to organize them in a basic structure: a temporal causal network. Let us define such concept in the next sub-section.

D. Concept of temporal causal network

In general, for many real world domains, we are not sure that the probability distribution on the states of E is constant in time. It may vary due to the only fact that time elapses. In general, given two time-slices: a present one (tsa) and a future one (tsb), we can state that it might be that $P(E_b=y) \neq P(E_a=y)$ for the only reason that the real time interval corresponding to tsb – tsa has elapsed. Let us represent such a situation in the following way. Let us consider, for example, a time interval [(i-1) , m) where the point (i-1) is included whereas the point m is excluded and i ≥ 1. Such interval is subdivided into sub-intervals: [(i-1) , i), concentrated in tsi ; [i , (i+1)), concentrated in ts(i+1) ; …; [(m-1) , m), concentrated in tsm. Let us build a causal chain where the E variables E_i , E_{i+1} ,…, E_m (that are present in the respective time-slices tsi, tsi+1, …, tsm) become nodes of the chain (i.e. E nodes). Such E nodes are connected by causal links: $E_i \rightarrow E_{i+1} \rightarrow \ldots \rightarrow E_m$. Such links connecting E nodes represent time elapsing. For this reason they are called temporal links. Taking into account that the occurrence probability of E for a subject may be affected by both the mere aging of the subject and the conditions C1, C2, …, CN under which the subject ages, we are prompted to enrich the causal chain by adding the history-variables of observed-aging present in the time slices. Precisely, such variables become nodes of the network (i.e. HC nodes) and in each time slice HC nodes are connected to the E node of the related time slice by causal links. For example, in tsi we have: $HC1_i \rightarrow E_i$, $HC2_i \rightarrow E_i$, … , $HCN_i \rightarrow E_i$. And so forth for the other time slices. Considering that also the initial profile of a subject contributes to affect the probability of E occurrence during the observed-aging period, let us also add the IP variable (i.e. IP node). Such node is linked through causal links to the E nodes of all the time slices: $IP \rightarrow E_i$, $IP \rightarrow E_{i+1}$, …, $IP \rightarrow E_m$. Finally, let us add the causal relation $E_{i-1} \rightarrow E_i$. If we suppose that the time slice (i-1) (if i=1, we have to do with the special time slice ts0) represents the time slice of the present time and all the others represent future time slices, we can conclude that E_{i-1} is necessarily in state n. Putting all together we get the temporal causal network illustrated in Figure 3. Let us comment it.

The network will be used by the Prediction Engine to produce predictions. If S1 is the subject for which predictions are to be produced, in each future time slice k the IP variable and the related history-variables are respectively instantiated with $iprof_{S1}$ and $prof_{S1,k}$, where $prof_{S1,k}$ is the profile extended as long as tsk. Formally: $prof_{S1,k} = prof_{S1,(i-1)}$ & $fprof_{S1,(i-1),k}$.

This sub-section has defined the basic structure of the temporal causal network used to produce predictions. However there is still an important concept missing from such network: probability. Let us examine, in the next section, how to obtain a probabilistic temporal causal network.

E. Concept of probabilistic temporal causal network

Let us consider the temporal causal network of Figure 3 and and let us enrich it with the concept of conditioned probability in the following way. Given two events: A, B, where B causes A, the probability of A occurrence conditioned to B occurrence, for short the probability of A given B, denoted by P(A|B), is defined as

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

From this definition we get the so-called Product Rule

$$P(A \mid B) \cdot P(B) = P(A, B)$$

Let us establish that a causal relation like B → A (B causes A) has associated a quantitative aspect: the value of the conditioned probability P(A|B), value that represents the strength of the causal relation. For example, for the time slice i we have (let A be E_i , let B be {E_{i-1} , IP, HC_i ,…, HC_i }):

$$P(E_i \mid E_{i-1}, IP, HC1_i, \ldots HCN_i)$$

As a consequence we obtain a probabilistic temporal causal network (for short: probabilistic network). Such network is dynamically created by the prediction engine when prediction calculus is required for a subject S1. Let us examine what an instantiation of the network of fig. 3 consists in. If (i-1) (where i ≥ 1) is the time-slice of the present time, probabilistic predictions consist in calculating, for each future time-slice k, where i ≤ k ≤ m, the value of

$$P(E_k = y \mid E_{i-1} = n, IP,$$
$$HC1_i, \ldots HCN_i, \ldots, HC1_k, \ldots, HCN_k)$$

where IP is instantiated to the initial profile of S1 (i.e. $iprof_{S1}$), whereas nodes $HC1_i$, …, HCN_i are instantiated to $prof_{S1,i} = prof_{S1,(i-1)}$ & $fprof_{S1,(i-1),i}$, … , nodes $HC1_k$, …, HCN_k are instantiated to $prof_{S1,k} = prof_{S1,(i-1)}$ & $fprof_{S1,(i-1),k}$. Once instantiated, the network of fig. 3 is ready for being used by the prediction engine. The target probability $P(E_k\text{=}y \mid E_{i-1}\text{=}n, IP, HC1_i, \ldots, HCN_k)$ has this meaning: probability that in time slice k E is in state y (that is E_k = y).

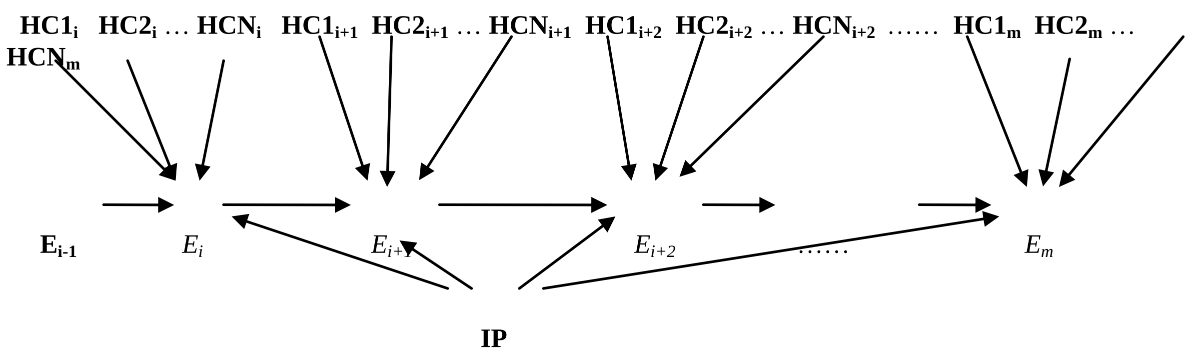


Fig. 3. The structure of the causal network used to produce predictions for a subject S1. The time-slice i (i ≥ 1) represents the first future time slice. The time-slice m represents the last future time slice. The arrows connecting E nodes represent temporal links. The nodes in bold are instantiated: the set of nodes $HC1_i$, …, HCN_i are instantiated by the observed-aging profile $prof_{S1,i}$, the set $HC1_{i+1}$, …, HCN_{i+1} are instantiated by $prof_{S1,i+1}$, etc. The IP node is instantiated by the initial profile $iprof_{S1}$. The node E_{i-1} is instantiated to the state n. The probability values of the E nodes in italic are to be calculated.

Let us conclude the present section with the following considerations. By using the basic concepts of observed-aging, time slice, temporal causal network and probabilistic network, we are able to step-by-step define a real world model that can be used to produce probabilistic predictions. The probabilistic network plays the role of a platform on which computational probability dynamics has to be accomplished by following well defined laws, laws that underlie the algorithm of the Prediction Engine. The next section faces this aspect.

III. MODELLING THE DYNAMICS OF PROBABILITY : A QUALITATIVE VIEW

Let us examine, in general and qualitative terms, the laws of probability dynamics. Let us think of probability as a sort of fluid located in the nodes of a probabilistic network. In each node, probability is distributed on the states of it. In each node, the sum of the probability quantities on the states equals 1. As a consequence, even if nodes are connected through causal relations, probability does not flow from a node to another through such connections. In fact, dynamics of probability has the following meaning. When we gather a certain fact from the real world we instantiate the related node. Instantiating a node means to get the probability distribution on its states to become the following: on one state the level of probability is 1 (that is 100%), on each of the remaining states the level of probability is 0. After having instantiated a node we have to propagate the consequences of such an instantiation throughout the network. Such propagation consists in properly varying the probability distribution on the states of the other nodes of the network (dynamics of probability). Such variations are accomplished by both calculating probability values on the basis of theorems (e.g. Bayes theorem, etc.) and definitions, and taking into account the types of structures of the causal relations that connect nodes.

Basically there are three types of such structures: serial structure, diverging structure, converging structure. Let us examine for each of them what it consists in and how it works taking into account that serial and converging structures are present in the network of fig. 3 and the knowledge of their behaviours will be relevant during the probabilistic reasoning presented in section 4. The diverging structure is not present in the network but, for the sake of completeness, it will be examined all the same.

In order to show that the laws of these structures are coherent with probability dynamics in human mind, let us use a well known example present in [5] : the "wet grass example" (fig. 4). The scenario the network of the figure refers to is the following. Mr. Holmes lives in a house with a garden that is near to the garden of his neighbour: Mr. Watson. Moreover, Mr. Holmes in the evening is used to water his garden with his sprinkler but then sometimes he forgets to switch it off. The wife of Mr. Holmes is used to get up early in the morning and go to the garden. The figure represents the causal relations of the physical world.

In the following of this section let us use this graphic representation: a node in bold means that it is instantiated, a node in italic means that the distribution of probability on its states is consequence of the fact that one or more nodes have been instantiated.

A. Serial structure

A serial structure (with 3 nodes) is represented by:
B → A → C .

Let us suppose that the starting situation is: no node is instantiated. If we instantiate B, then the instantiation consequences propagate descending along the causal path: from B to A to C (from the cause to the effect): **B** → *A* → *C* affecting this way the distribution of probability on the states of A and C. Vice versa, if we instantiate C, then instantiation consequences propagate going up the causal path in the opposite direction: from C to A to B (from the effect to the cause): *B* → *A* → **C** affecting this way the probability distributions on the states of A and B.

Let us now suppose that we start from a situation in which A has already been instantiated: *B* → **A** → *C*. In such a case the propagation channel between B and C is closed. The instantiation of B has no influence on C, that is **B** → **A** → *C* is equivalent to **A** → *C* and so B can be neglected (such a situation will be encountered in the course of probabilistic reasoning in section 4). Similarly, the instantiation of C has no influence on B.

Referring to fig. 4, let us consider the sub-network {R , S} → H → wshoes. If Mr. Holmes knows with certainty that **H**=y (his grass is wet) then he concludes that it is very probable that the shoes of his wife are wet, no matter if that is due to the fact that it has rained or the sprinkler has been forgotten on, that is no matter if it is {R=n, S=y} or {R=y, S=n}.

B. Diverging structure

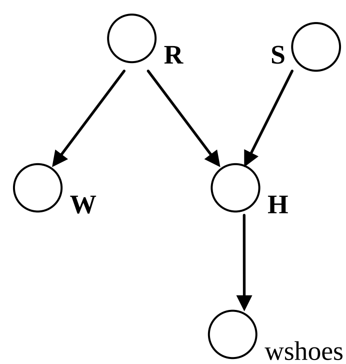


Fig. 4. The causal network of the "wet grass example". Legenda:
R = rain;
S = sprinkler of Mr. Holmes;
H = grass of Mr. Holmes;
W = grass of Mr. Watson;
wshoes = shoes of wife of Mr. Holmes.
Each node has two states: y, n. (that is: yes, no).

As already above stated, the diverging structure is not present in the network of fig. 3 but, for the sake of completeness, let us examine it all the same. A diverging structure (with 3 nodes) is represented by: B ← A → C . Let us suppose that the starting situation is: no node is instantiated. If we instantiate B, then the instantiation consequences propagate going up the causal path in the opposite direction (from B to A): **B** ← *A* and then descending along the causal path (from A to C): *A* → *C* , affecting this way the distribution of probability on the states of A and C. Similarly, if we instantiate C, the instantiation consequences propagate to A and B. Again, if we suppose that we start from a situation in which A has already been instantiated: *B* ← **A** → *C*, then the propagation channel between B and C is closed. Referring to fig. 4, let us consider the diverging structure W ← R → H with no node instantiated.

● Case a: R is not instantiated.

If in the morning Mr. Holmes sees that the grass of Mr. Watson is wet (**W**=y), he concludes that it is very probable that it has rained during the night, that is: R=y is very probable (he is not certain because the grass of Mr Watson might be wet because Mr. Watson has watered the garden). As a consequence Mr. Holmes retains that it is very probable that his grass too is wet, that is: H=y is very probable (and therefore it is also very probable that the shoes of his wife are wet).

● Case b: R is instantiated.

Let us suppose that Mr. Holmes, as soon as gets up in the morning, receives the piece of news that it has not rained in the night (that is he knows with certainty that **R**=n). As a

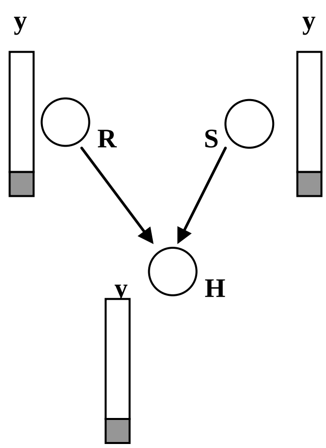


Fig. 5. The converging structure of the "wet grass example". Initial situation: no node is instantiated. Only the state y is illustrated. The state n is complementary.

consequence he retains that it is little probable that the grass of Mr. Watson is wet (he is not sure because Mr. Watson sometimes waters his garden). If later Mr. Holmes sees that his grass is wet (**H**=y), then this fact does not change his retaining that it is little probable that W=y (in fact the grass of Mr. Holmes might be wet because of his sprinkler that had been forgotten in state on the night before).

C. Converging structure

Let us consider the converging structure. Such structure, beside being present in the network of fig. 3, presents interesting aspects of probability dynamics that will have an important role in the probabilistic reasoning of section 4 (see Theorem 1). As a consequence let us deeply examine it. The converging structure (with 3 nodes) is represented by:

B→ A ← C , where A is called "convergence node".

Let us suppose that the starting situation is: no node is instantiated. If we instantiate C, then the instantiation

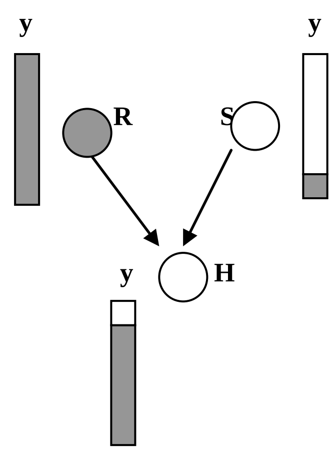


Fig. 6. R is instantiated but since H is not instantiated, the instantiation of R does not affect S.

consequences propagate only to A, the propagation channel from A to B is closed: B→ *A* ← **C** . Similarly, if we instantiate B, the propagation channel from A to C is closed.

Again let us start from the starting situation in which no node is instantiated. If we instantiate A, then the instantiation consequences propagate to B and C: *B*→ **A** ← *C* . If now we instantiate one of the two remaining nodes, say C, then the channel from A to B is open and the instantiation consequences can propagate to the other node, i.e. B. Precisely: the probability distribution on the states of B goes back to the initial values.

Referring to fig. 4, let us consider the converging structure R→ H ← S . The starting situation is: no node is instantiated.

There are two initial probability distributions: one on the states of R, the other on the states of S. The distribution concerning R reflects the percentage of night with rain in the current season. The distribution concerning S reflects the percentage of times that Mr. Holmes forgets to turn his sprinkler off. The starting situation is represented in fig. 5. Probability is represented as a fluid that is distributed on the states of a node.

● Case a: H is not instantiated.

Let us suppose that Mr. Holmes, as soon as he gets up, receives the piece of news that it has rained during the night. He

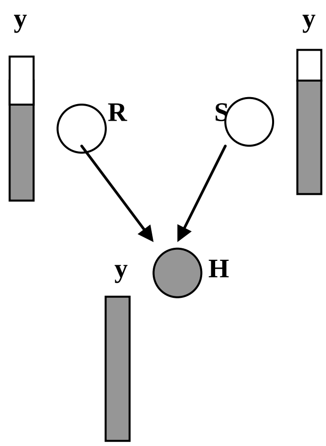
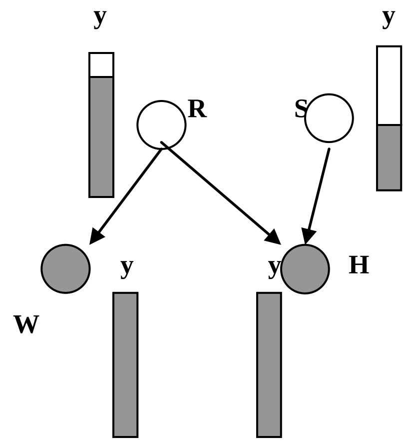


Fig. 7. H is instantiated to state y. The effects of such an instantiation propagate to R and S.

instantiates the node R, that is **R**=y, and propagates the consequences: he retains that it is very probable that his/her grass is wet, that is H=y (and similarly as for W=y and the shoes of his wife). He though does not vary the probability distribution on the states of S because the fact that it has rained has nothing to do with the fact that the sprinkler might has been forgotten in state on. In other words: as for the propagation of the consequences of the instantiation of R, the channel from H to S is closed. The situation here described is represented in fig. 6.

● Case b: H is instantiated.

Starting from the initial situation (fig. 5) let us now suppose that Mr. Holmes, as soon as he gets up, sees his grass wet. He instantiates the node H, that is **H**=y, and propagates the consequences: he retains more probable both that he has forgotten to turn his sprinkler off and that it has rained (and as a consequence it is also more probable that the grass of Mr Watson is wet). In more formal terms, Mr. Holmes increases the probability level associated to the state S=y and R=y (and as a consequence, W=y). The situation here described is represented in fig. 7 (of course he also believes wshoes=y).



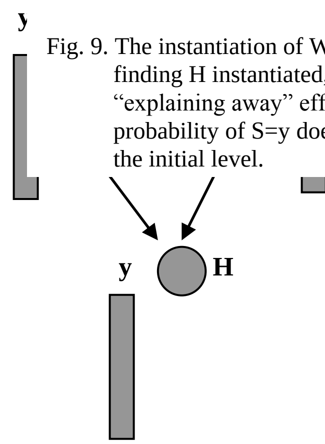


Fig. 9. The instantiation of W to the state y, finding H instantiated, produces an “explaining away” effect, but the level of probability of S=y does not go back to the initial level.

Fig. 8. The instantiation of R to the state y, finding H instantiated, produces the “explaining away” effect: the level of probability of S=y goes back to the initial level.

If then Mr. Holmes receives the piece of news that it has rained during the night, he instantiates the node R, that is **R**=y, and propagates the consequences of such an instantiation. Of course the probability level associated to the state W=y increases, but what is more interesting is that the probability level associated to the state S=y decreases to the initial level, that is the level that corresponds to the fact that some evenings Mr. Holmes forgets to turn his sprinkler off: the level illustrated in fig. 5. In fact, the fact that it has rained during the night (R=y) perfectly explains why the grass of Mr. Holmes is wet and as a consequence there are no reasons for Mr. Holmes to believe in S=y more or less than how much he normally believes. This phenomenon of probability dynamics that occurs in human mind is called “explaining away”. The situation here described is represented in fig. 8.

Finally, for the sake of completeness let us consider the following case of "explaining away". Let us suppose that, starting form the situation described in fig. 7, Mr. Holmes does not know if it has rained during the night, but sees that the grass of Mr. Watson is wet. Then Mr. Holmes instantiates the node W (i.e. **W**=y), instead of R, and propagates the consequences of this instantiation. The probability level associated to the state R=y increases and the probability level associated to the state S=y decreases, carrying out the "explaining away" mechanism, but this time the probability level associated to the state S=y does not decrease to the initial level. This is due to the fact that Mr. Holmes is not certain that it has rained during the night. In fact the fact that the grass of Mr. Watson is wet might be due to the fact that Mr. Watson has watered his garden the evening before and therefore the fact that the grass of Mr. Holmes is wet, might be due to the fact that Mr. Holmes has forgotten to switch off his sprinkler the evening before. The situation here described is represented in fig. 9. (A sort of imperfect explanation of H=y)

Let us notice that the "explaining away" mechanism occurs even if the convergence node is not instantiated, in fact it is sufficient that an effect node (a son node) of the convergence node is instantiated. Referring to fig. 5 it can be stated that if, instead of instantiating H, we instantiate wshoes, the "explaining away" occurs all the same.

We are now at the end of this section. We have at our disposal all the background knowledge necessary for facing the prediction problem considered in the proposal. Let us therefore pass to examine, in the next section, how to build the probabilistic prediction engine.

IV. Computational Probability Dynamics : Probabilistic Prediction Engine

Let us consider the network of fig. 3. Let S1 be the subject for which predictions have to be produced for the future time slices: i , …, m. Let $fprof_{S1,(i-1),m}$ be the future simulated profile for S1. For each time slice k ($i \leq k \leq m$) let us calculate the related extended profile of S1: $prof_{S1,k} = prof_{S1,(i-1)}$ & $fprof_{S1,(i-1),k}$, and then, with such extended profile, let us instantiate the nodes $HC1_k$ … HCN_k . Finally let us instantiate the node IP with $iprof_{S1}$. The network of fig. 3 is now completely instantiated.

A. *Prediction for the first future time slice*

Let us calculate the prediction for the first future time slice: the time slice i. This means that we have to calculate the probability of E_i=y in the network of fig. 3. Let us notice that the node E_{i+1} is the converging node of a converging structure. Let us map such converging structure in fig. 3 into the converging structure R → H ← S in fig. 5. More precisely, the node E_{i+1} corresponds to the node H, the node E_i corresponds to

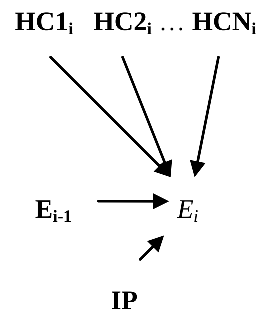


Fig. 10. The structure of the causal network used to produce prediction for the first future time slice. The node E_{i-1} is instantiated to the state n. The probability value of the E_i node has to be calculated.

the node S, the set of nodes {$HC1_{i+1}$,…, HCN_{i+1} , IP} corresponds to the node R. Since the node E_{i+1} is not instantiated, E_i is not instantiated, the nodes $HC1_{i+1}$,…, HCN_{i+1} and IP are instantiated, we are in the situation described in section 3.3 case a fig. 6. As a consequence, the instantiations of

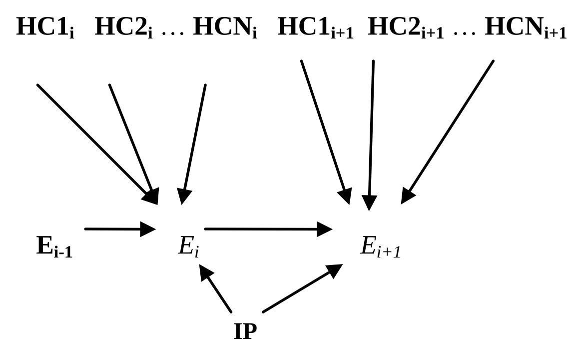


Fig. 11. The structure of the causal network used to produce predictions for the second future time slice. The node E_{i-1} is instantiated to the state n.

the nodes {$HC1_{i+1}$,…, HCN_{i+1} , IP} do not affect the distribution probability on the states of f E_i . The conclusion is that the nodes $HC1_{i+1}$,…, HCN_{i+1} can be neglected along with the link from IP to node E_{i+1} and the node E_{i+1} itself. The instantiated network of fig. 3 therefore reduces to the one of fig. 10.

Let us make the following considerations on the network of fig. 10. Let us suppose that a subject S1 at the time slice i has the whole extended profile $wprof_i$ = iprof AND $prof_i$. As a consequence the extended profile $prof_i$ instantiates the variables

$HC1_i$,…, HCN_i and the initial profile iprof instantiates the variable IP. So, considering the node E_i (the only node that is not instantiated), it can be stated that the probability that E_i is in state y is formally expressed by:
$P(E_i=y \mid E_{i-1}=n$, iprof , $prof_i$). On the basis of the considerations made in section 2.3, precisely in the paragraph concerning the definition of the counter variables, we can state that $P(E_i=y \mid E_{i-1}=n$, iprof , $prof_i$) = $L(wprof_i)$.

We have so reached the first target. We know the prediction for the first future time slice, the time slice i. But what about the prediction fort the next time slice, i.e. the time slice i+1 ?

B. Prediction for the second future time slice: common sense based reasoning

Let us calculate the prediction for the second future time slice: the time slice i+1. This means that we have to calculate the probability of E_{i+1}=y in the network of fig. 3. Let us notice that the node E_{i+2} is the converging node of a converging structure. By applying to the present case the same kind of reasoning illustrated for the first future time slice, we can conclude that the network of fig. 3 becomes the one of fig. 11.

Let us make the following considerations on the network of fig. 11. Let us notice that E_i is a node inside a serial structure: it is a node along the causal paths starting from E_{i-1} , $HC1_i$,…, HCN_i , IP and ending to E_{i+1} . Let us suppose that E_i=n . In such a case, on the basis of the considerations made in section 3.1, the nodes $HC1_i$,…, HCN_i can be neglected along with the node E_{i-1} and the link from IP to node E_i , and as a consequence the network of fig. 11 reduces to the one of fig. 12.

But this network is similar to the one of fig. 10 and so, by applying the same considerations made for fig. 10, we can conclude that for a subject S1, that at the time slice i+1 has the whole extended profile $wprof_{i+1}$ = iprof AND $prof_{i+1}$, it can be stated: $P(E_{i+1}=y \mid E_i=n$, iprof , $prof_{i+1}$) = $L(wprof_{i+1})$.
But, unfortunately, we cannot be sure that E_i=n . So, considering fig. 11, the basic question arises: if we believe $L(wprof_i)$ that E_i=y (given E_{i-1}=n and $wprof_i$), and $L(wprof_{i+1})$ that E_{i+1}=y (given E_i=n and $wprof_{i+1}$), then how much should we *coherently* believe that E_{i+1}=y (given E_{i-1}=n and iprof and the extended profiles $prof_i$ and $prof_{i+1}$) ?

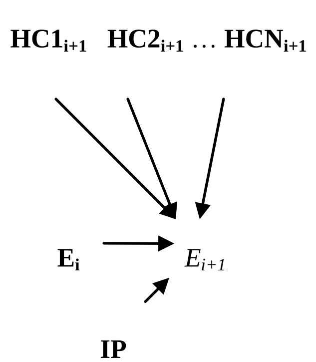


Fig. 12. The node E_i is instantiated to the state n. The probability value of the state y of the node E_{i+1} is given by $L(wprof_{i+1})$.

C. Predictions for all the future time slices: formal and general approach

Let us go back to a formal level. For the sake of shortness and formal clarity let us use, in the following formal reasoning, the shorter denotation L_k instead of $L(wprof_k)$. Referring again to the network of fig. 3 let us consider a time slice k (with i≤k≤m) and let us formalize and generalize the common sense based considerations made in section 4.2. Let us start by reformulating the basic question: if we believe L_i that E_i=y (given E_{i-1}=n and $wprof_i$), …, L_k that E_k=y (given E_{k-1}=n and $wprof_k$), then how much should we *coherently* believe that E_k=y (given E_{i-1}=n , iprof , $prof_i$, …, $prof_k$) ? In formal terms the question becomes: what is the value of the following conditioned probability ?

$$P(E_k = y \mid E_{i-1} = n, IP, HC1_i,...HCN_i,...,HC1_k,...,HCN_k)$$

.

The prediction engine is a machine built for providing answer to this question. The answer is given by the following formula

$$\begin{cases} L_k & \text{if } k=i \\ L_k \cdot (1 - X_{k-1}) + X_{k-1} & \text{if } i < k \le m \end{cases} \qquad (0)$$

where X_{k-1} stands for the prediction value calculated for the preceding time slice k-1, that is formally:

$$X_{k-1} = P(E_{k-1} = y \mid E_{i-1} = n, IP, HC1_i,...,HCN_{k-1})$$

The proof is given by Theorem 1, theorem that is presented in the following sub-section.

D. Probabilistic reasoning: approach based on probability dynamics

This section presents probabilistic reasoning based on the laws of probability dynamics examined in section 3.

Theorem 1 [reasoning based on probability dynamics]
Let us consider the network in fig. 3 and a time slice k (with i≤k≤m)

$$P(E_k = y \mid E_{i-1} = n, IP, HC1_i,...HCN_i,...,HC1_k,...,HCN_k)$$

$$=$$

$$L_k \qquad \text{if } k=i$$

$$L_k \cdot (1 - X_{k-1}) + X_{k-1} \qquad \text{if } i < k \le m$$

where

$$L_k = \frac{E_Y wprof_k}{E_Y wprof_k + E_N wprof_k}$$

where $wprof_k$ represents the whole profile instantiating both the variable IP (with iprof) and the set of history variables $HC1_k, \dots, HCN_k$ related to time-slice k (with $prof_k$), and where X_{k-1} stands for the prediction value calculated for the preceding time slice k-1.

Proof

Let us premise the following considerations.

For each whole profile $wprof_k$ (where $i \le k \le m$, and $i \ge 1$) present in the database it can be stated the implicit fact $E_{k-1}=n$

As a consequence, by adopting the frequentist probability definition, it can be stated that

$$P(E_k = y \mid E_{k-1} = n, \\ IP, HC1_k, \dots HCN_k) = L_k$$

Obviously it has to be intended as an empirical probability value approximating the theoretical probability value, approximation that is as smaller as greater the value of ($E_N wprof_k + E_Y wprof_k$) is. Given this premise, let us enter the theorem proof.

- Let us consider first the case of k = i. If k=i we have

$$P(E_i = y \mid E_{i-1} = n, IP, HC1_i, \dots HCN_i)$$

but this is just L_i

- Let us now consider the case $i < k \le m$.

For short let us use the symbol A to denote the sequence:

$$E_{i-1} = n, IP, \\ HC1_i, \dots, HCN_i, \dots, HC1_k, \dots, HCN_k$$

It can be stated that

$$P(E_k = y \mid A) = \\ P(E_k = y \mid E_{k-1} = n, A) \cdot P(E_{k-1} = n \mid A) + \\ P(E_k = y \mid E_{k-1} = y, A) \cdot P(E_{k-1} = y \mid A) \qquad (1)$$

In fact:

1) by applying the product rule we have

$$P(E_k = y \mid E_{k-1} = n, A) \cdot P(E_{k-1} = n \mid A) = \\ P(E_k = y, E_{k-1} = n \mid A)$$

and similarly

$$P(E_k = y \mid E_{k-1} = y, A) \cdot P(E_{k-1} = y \mid A) = \\ P(E_k = y, E_{k-1} = y \mid A)$$

2) since the two joint events (E_k=y, E_{k-1}=n) and (E_k=y, E_{k-1}=y) are mutually exclusive, on the basis of the addition axiom we have:

$$P(E_k = y, E_{k-1} = n \mid A) + P(E_k = y, E_{k-1} = y \mid A) = \\ P((E_k = y, E_{k-1} = n \mid A) OR (E_k = y, E_{k-1} = y \mid A))$$

3) since the set of states {E_{k-1}=n , E_{k-1}=y} is exhaustive, we have:

$$P((E_k = y, E_{k-1} = n \mid A) OR (E_k = y, E_{k-1} = y \mid A)) = \\ P(E_k = y \mid A)$$

On the basis of these considerations let us rewrite the (1) as follows (for short the sequence $HC1_i$, …, HCN_i, …, $HC1_k$, …, HCN_k is represented by $HC1_i$, …, HCN_k):

$$P(E_k = y \mid E_{i-1} = n, IP, HC1_i, \dots, HCN_k)$$

$$=$$

$$P(E_k = y \mid E_{k-1} = n, \\ E_{i-1} = n, IP, HC1_i, \dots, HCN_k) \cdot \qquad (2)$$

$$P(E_{k-1} = n \mid E_{i-1} = n, IP, HC1_i, \dots, HCN_k) + \qquad (3)$$

$$P(E_k = y \mid E_{k-1} = y, \\ E_{i-1} = n, IP, HC1_i, \dots, HCN_k) \cdot \qquad (4)$$

$$P(E_{k-1} = y \mid E_{i-1} = n, IP, HC1_i, \dots, HCN_k) \qquad (5)$$

● Let us consider the (2). Every causal path connecting the nodes E_{i-1} , $HC1_i$,…, HCN_i ,…, $HC1_{k-1}$,…, HCN_{k-1} to the node E_k is a serial structure in which E_{k-1} is the last but one node. Since E_{k-1} is instantiated to a state (i.e. the state n), its antecedents can be neglected along with the link IP → E_{k-1} Referring to section 3.1 and fig. 4 we can notice that the node E_k plays the role of the node wshoes, the node E_{k-1} plays the role of the node H, the set of the remaining nodes plays the role of the set {R , S}. In conclusion the (2) is equivalent to

$$P(E_k = y \mid E_{k-1} = n, IP, HC1_k, \ldots HCN_k)$$

which is L_k

● Let us consider the (3). The value of the (3) is complementary to the value of the (5).

● Let us consider the (4). The value of the (4) is 1. In fact if E_{k-1}=y, then E_k=y independently of the combination of context states in session k: if when we are in time-slice k-1 we know that E has occurred, then the knowledge of that fact does not change for all the subsequent time-slices.

● Finally let us consider the (5). The nodes E_{k-1} , IP, $HC1_k$,…, HCN_k are all direct causes of the node E_k : there is a causal structure converging to E_k and since E_k is not instantiated, its causes are all independent. We have a case that instantiates the case a in section 3.3. In fact, referring to fig. 6 we can notice that the converging node E_k plays the role of H, the node E_{k-1} plays the role of S, the (instantiated) nodes $HC1_k$,…, HCN_k and IP play the role of R and as a consequence do not affect the distribution probability on the states of E_{k-1} . The conclusion is that the nodes $HC1_k$,…, HCN_k can be neglected along with the link from IP to node E_k and the node E_k itself. The ultimate consequence is that the (5) is equivalent to:

$$P(E_{k-1} = y \mid E_{i-1} = n, IP, HC1_i, \ldots, HCN_{k-1}) \qquad (6)$$

But the value of the (6) is the prediction value calculated for the time slice k-1. Putting all together, it can be stated that:

$$P(E_k = y \mid E_{i-1} = n, IP, HC1_i, \ldots, HCN_k) = L_k \cdot (1 - X_{k-1}) + X_{k-1}$$

where X_{k-1} stands for the prediction value calculated for session k-1. In conclusion:

$$P(E_k = y \mid E_{i-1} = n, IP, HC1_i, \ldots HCN_i, \ldots, HC1_k, \ldots, HCN_k)$$

$$= \begin{cases} L_k & \text{if } k=i \\ L_k \cdot (1 - X_{k-1}) + X_{k-1} & \text{if } i < k \le m \end{cases}$$

End of proof [Theorem 1]

It is interesting to notice that we can get to an equivalent conclusion by reasoning with the global joint probability table of the network of fig. 3. The next sub-section faces this problem.

E. Probabilistic reasoning: global joint probability-table based approach

We know that from the global joint probability table of a network we can calculate the probability values of all the nodes of the network. Such a joint probability table can be built by applying the so-called Chain Rule. The Chain Rule, which is an application of the Product Rule, is defined as:

$$P(A_1, \ldots, A_n) = \prod_{i=1}^{n} P(A_i \mid pa(A_i))$$

where $pa(A_i)$ stands for “direct parents of A_i”.
Let us look at Figure 3 again and let us build the following theorem (Theorem 2) for calculating the value of

$$P(E_k = y \mid E_{i-1} = n, IP, HC1_i, \ldots HCN_i, \ldots, HC1_k, \ldots, HCN_k)$$

Theorem 2, based on Chain Rule application, defines a general algorithm, that is equivalent to the one defined in Theorem 1.

Theorem 2 [reasoning based on global joint probability-table]

$$P(E_k = y \mid E_{i-1} = n, IP,$$
$$HC1_i ,...HCN_i ,..., HC1_k ,..., HCN_k) =$$
$$L_k \cdot (1 - L_{k-1}) \cdot (1 - L_{k-2}) \cdot ... \cdot (1 - L_{i+1}) \cdot (1 - L_i) +$$
$$L_{k-1} \cdot (1 - L_{k-2}) \cdot ... \cdot (1 - L_{i+1}) \cdot (1 - L_i) +$$
$$...$$
$$L_{i+1} \cdot (1 - L_i) +$$
$$L_i$$

where L_j (where j is such that: $i \leq j \leq k$), is given by

$$L_j = \frac{E_Y wprof_j}{E_Y wprof_j + E_N wprof_j}$$

where $wprof_j$ represents the whole profile instantiating both the variable IP and the set of the variables $HC1_j, \ldots, HCN_j$.

Proof

For short, the sequence $HC1_i$, …, HCN_i, …, $HC1_k$, …, HCN_k is represented by $HC1_i$, …, HCN_k . Before entering the reasoning of the proof let us define the following five general rules:

● *RULE 1*) Given that $P(E_{i-1}=n)=1$, and the symbols IP, $HC1_i$,…, HCN_k denote condition instantiations (that is IP means IP=st, $HC1_i$ means $HC1_i$ = st, etc.) we have that: P(IP)=1, $P(HC1_i)=1$,…, $P(HCN_k)=1$. As a consequence:

$$P(E_{i-1} = n) \cdot P(IP) \cdot$$
$$P(HC1_i) \cdot ... \cdot P(HCN_k) = 1$$

● *RULE 2*) Given a time-slice j ($j \geq 1$) it can be stated that:

$$P(E_j = y \mid E_{j-1} = n, IP, HC1_j ,..., HCN_j) = L_j$$

● *RULE 3*) On the basis of Rule 2 it can be stated that:

$$P(E_j = n \mid E_{j-1} = n, IP,$$
$$HC1_j ,..., HCN_j) = (1 - L_j)$$

● *RULE 4*) As already noticed, if E_{j-1}=y, then E_j=y independently of the combination of context states. As a consequence:

$$P(E_j = y \mid E_{j-1} = y, IP,$$
$$HC1_j ,..., HCN_j) = 1$$

● *RULE 5*) On the basis of Rule 4 it can be stated that:

$$P(E_j = n \mid E_{j-1} = y, IP, HC1_j ,..., HCN_j) = 0$$

Let us now enter the proof. The proof is defined by an algorithm structured in three basic sequential steps.

● *STEP 1)* Let us consider the global joint probability table of the network in Figure 3. The value of

$$P(E_k = y \mid E_{i-1} = n, IP, HC1_i ,..., HCN_k)$$

is calculated by adding the values of all the table rows containing E_k=y and E_{i-1}=n, IP, $HC1_i$, …, HCN_k .
In more formal terms:

$$P(E_k = y \mid E_{i-1} = n, IP, HC1_i ,..., HCN_k) =$$
$$P(E_k = y,$$
$$E_{k-1} = n, E_{k-2} = n, E_{k-3} = n,..., E_i = n,$$
$$E_{i-1} = n, IP, HC1_i ,..., HCN_k) +$$
$$P(E_k = y,$$
$$E_{k-1} = y, E_{k-2} = n, E_{k-3} = n,..., E_i = n,$$
$$E_{i-1} = n, IP, HC1_i ,..., HCN_k) +$$
$$....$$

Let us notice that the number of addenda is 2^{k-i} . In fact between E_k and E_{i-1} there are the k-i E nodes: E_{k-1} , E_{k-2} ,…, E_{i+1} , E_i . Such nodes are not instantiated and since each of them has two states {n, y}, the number of possible combinations for the set of states of these nodes is 2^{k-i} . As a consequence we have 2^{k-i} addenda.

● *STEP 2)* Let us apply the Chain Rule to each addendum. Let us consider the first addendum. We get:

$$P(E_k = y, E_{k-1} = n, \ldots, E_i = n, E_{i-1} = n,$$
$$IP, HC1_i, \ldots, HCN_k) =$$
$$P(E_k = y \mid E_{k-1} = n, IP, HC1_k, \ldots, HCN_k) \cdot$$
$$P(E_{k-1} = n \mid E_{k-2} = n, IP, HC1_{k-1}, \ldots, HCN_{k-1}) \cdot$$
$$P(E_{k-2} = n \mid E_{k-3} = n, IP, HC1_{k-2}, \ldots, HCN_{k-2}) \cdot$$
$$\ldots$$
$$P(E_{i+1} = n \mid E_i = n, IP, HC1_{i+1}, \ldots, HCN_{i+1}) \cdot$$
$$P(E_i = n \mid E_{i-1} = n, IP, HC1_i, \ldots, HCN_i) \cdot$$
$$P(E_{i-1} = n) \cdot P(IP) \cdot P(HC1_i) \cdot \ldots \cdot P(HCN_k)$$

By applying rules 1, 2, 3 the first addendum becomes:

$$P(E_k = y, E_{k-1} = n, E_{k-2} = n, \ldots, E_i = n$$
$$E_{i-1} = n, IP, HC1_i, \ldots, HCN_k) =$$
$$L_k \cdot (1 - L_{k-1}) \cdot (1 - L_{k-2}) \cdot \ldots \cdot (1 - L_{i+1}) \cdot (1 - L_i)$$

Let us now consider the second addendum. By applying the Chain Rule and then rules 1, 2, 3, 4, we get:

$$P(E_k = y, E_{k-1} = y, E_{k-2} = n, \ldots,$$
$$E_{i-1} = n, IP, HC1_i, \ldots, HCN_k) =$$
$$L_{k-1} \cdot (1 - L_{k-2}) \cdot \ldots \cdot (1 - L_{i+1}) \cdot (1 - L_i)$$

And so forth for the remaining addenda.

● *STEP 3)* Let us sum the results obtained by applying the chain Rule and then the above five rules. At the end we get:

$$P(E_k = y \mid E_{i-1} = n, IP,$$
$$HC1_i, \ldots HCN_i, \ldots, HC1_k, \ldots, HCN_k) =$$
$$L_k \cdot (1 - L_{k-1}) \cdot (1 - L_{k-2}) \cdot \ldots \cdot (1 - L_{i+1}) \cdot (1 - L_i) +$$
$$L_{k-1} \cdot (1 - L_{k-2}) \cdot \ldots \cdot (1 - L_{i+1}) \cdot (1 - L_i) +$$
$$\ldots$$
$$L_{i+1} \cdot (1 - L_i) +$$
$$L_i$$

End of proof [Theorem 2]

Let us examine, in the next sub-section, the equivalence between the conclusions of the two theorems.

F. Equivalence of the two approaches

For the sake of simplicity let us avoid facing a formal general proof of equivalence of the conclusions of the two theorems. Let us limit to show that there is equivalence when considering specific time slices.
For example, let us consider the time slice k where k = i+2. By applying the algorithm of Theorem 2 we have:

$$P(E_{i+2} = y \mid E_{i-1} = n, IP, HC1_i, \ldots, HCN_{i+2}) =$$
$$L_{i+2}(1 - L_{i+1})(1 - L_i) +$$
$$L_{i+1}(1 - L_i) +$$
$$L_i$$

By applying the algorithm of Theorem 1 with k=i+2 we have:

$$P(E_{i+2} = y \mid E_{i-1} = n, IP, HC1_i, \ldots, HCN_{i+2}) =$$
$$L_{i+2} \cdot (1 - X_{i+1}) + X_{i+1}$$

where

$$X_{i+1} = L_{i+1} \cdot (1 - X_i) + X_i$$

where

$$X_i = L_i$$

We can notice (by a simple algebraic calculus) that even if the two results appear to be formally different, they express the same conclusion.

Let us conclude this section by noticing that the fact that by means of Theorem 2 (that uses only probability axioms) we have reached the same conclusions reached by means of Theorem 1 (that uses the laws of probability dynamics in the structures presented in section 3) can be seen as a mathematical confirmation of the validity of the common sense based laws of the dynamics of that sort of “fluid” called probability.

V. Related work and discussion

The Prediction Engine has been implemented as a software tool embedded in a general software environment for predictive monitoring. The great number of works concerning predictive monitoring, published in scientific journals and conferences both in past and in recent years, gives evidence of both the modernity of the topic and the remarkable effort so far accomplished by the researchers community with respect to such a theme.

Industry is a typical world in which predictive monitoring, mostly intended as preventive monitoring, has find numerous applications with a variety of approaches. Twenty years ago already, preventive monitoring was a crucial theme for manufacturing processes (typically, for example, in the world of the large car manufacturing companies [6]). In manufacturing industries there is a considerable attention to reduce costly and unexpected breakdowns. As a consequence preventive maintenance is becoming more and more important. Maintenance should abandon the traditional “fail and fix” approach to pass to the more modern “predict and prevent” one [7]. As a consequence the fundamental need is monitoring degradation instead of detecting faults. A predictive performance and degradation monitoring is what is needed for an effective proactive maintenance to prevent machines from breakdown. The theme of degradation monitoring for failure prevention applied to vehicle electronics and sensor systems is faced in [8] where the authors propose a unified monitoring and prognostics approach that prevents failures by analyzing degradation features, driven by physics-of-failure. The need, for manufacturers of complex systems, to optimize equipment performance and reduce costs and unscheduled downtime, gives rise to system health monitoring. System states monitoring is augmented with prediction of future system health states and predictive diagnosis of possible future failure states [9]. Predictive monitoring has been also applied to flexible manufacturing systems. In [10], the main objective is to manage progressive failures in order to avoid breakdown state for the flexible manufacturing system. The approach to predictive monitoring proposed in [11] uses predictions from a dynamic model to predict whether process variables will violate an emergency limit in the future (predictions are based on a Kalman filter and disturbance estimation). Predictive monitoring has also been applied in many specific industry worlds like, for example, cold extrusion and forging processes [12] and chemical plants [13], [14]. In many industrial applications predictive monitoring assumes the meaning of preventive monitoring and aims to enhance the effectiveness of preventive maintenance by making it proactive. In some cases though, predictive monitoring is finalized to early intervening to maintain a system at a high level of performance. It is the case of a predicting monitoring application for wireless sensor networks: “...by monitoring and subsequently predicting trends on network load or sensor nodes energy levels, the wireless sensor network can proactively initiate self-reconfiguration…” [15]. In most industry applications the acquisition of monitoring data is carried out through sensors [16].

Predictive monitoring has found many applications in medicine too. Applications concern both clinical trials [17] and several specific fields. For example, interesting applications have been carried out in the field of diabetes therapy. In [18] and [19], continuous glucose monitoring devices provide data that are processed by mathematical forecasting models to predict future glucose levels in order to prevent hypo-/hyperglycemic events. Many other specific applications of preventive monitoring may be found in medicine. For example, in [20] the authors present the experience of predictive monitoring applied to some patients exposed to gentamicin (a commonly used antibiotic medication) ototoxicity: the most common single known cause of bilateral vestibulopathy. Patients undergoing exercise rehabilitation therapy were tested repeatedly during follow-up visits to monitor changes in their vestibulo-ocular reflex. Predictive monitoring turned out to be useful for continuing or modifying the course of vestibular rehabilitation therapy.
Very recently predictive monitoring has found many applications in the field of environment pollution [21], [22], [23].

Literature shows that, in general, prediction has been intended in the sense of prevention, that is as a means for preventing undesired events. Actually the possibility of getting early warnings before an undesired event may occur has always been very appealing. Let us think, for example, of prevention of high risk events for health, or serious faults or anomalies of costly and strategic industrial equipments or plants.

The proposal presented in the paper has the ultimate purpose that is in common with all the cited applications, but, at the same time, it has many aspects that distinguish it from them. The proposal, is neither a predictive monitoring application nor a general prognostics tool for preventing undesired events in some fields like, for example, manufacturing industries, medicine, etc. In fact the proposal is general, it presents both a general method for modelling real problems and a general algorithm for producing predictions. Moreover the proposal concerns prediction applied to both preventing undesired events and favouring desired events. Since the prediction engine has been implemented in a software tool oriented to predictive monitoring, it is the domain expert (i.e. a human agent) that carries out monitoring sessions and enters data about the current subject situation (context states, etc.). Again, it is the domain expert that defines the starting conditions for simulating the future (which condition states are supposed to be present in the future), and it is the domain expert that reads the simulation results and takes suitable measures. Let us notice though that the algorithm of the prediction engine might be embedded in a software program. In such a case it is an automatic agent that plays the role of a human user of a specific predictive monitoring tool. For example, a software agent might periodically gather data about a subject (e.g. a machine) by means of sensors or other software programs interfacing a database. It is the software agent that activates simulations and then examines the results and as a consequence takes suitable measures. Finally, let us consider that, with respect to other approaches to predictive monitoring, the predictive engine considered in the proposal is probabilistic. This means that the tool becomes predictive only after having collected a number of cases that is statistically significant, i.e. sufficient to be able to produce probabilistic inferences. Before reaching that condition the tool works like a mere monitoring tool (subject data acquisition and data entry into the database).

VI. Conclusion

The paper has presented both a method for modelling real world and a Probabilistic Prediction Engine at a theoretical level. However the author has also developed a software tool, in a prototype version, that implements the algorithm of the prediction engine. Moreover, such tool has been embedded into a general software environment for creating and administering specific application oriented predictive tools (in heterogeneous fields). Such environment, that has been illustrated in [4] and can be found at the web-page www.cheerup.it , is oriented to predictive monitoring. It provides numerous and effective facilities for probabilistic predictive monitoring: facilities for creating new application oriented tools (equipped with specific domain knowledge), monitoring subjects and simulating possible future probabilistic scenarios, administering tools and subjects and regulating co-operation among working groups. Such environment is ready to be used for facing probabilistic prediction applications in real world fields.

Acknowledgment

I thank Prof. Frank Jensen for his clarifying answers (via e-mail) to some scientific questions.

References

[1] D. A. Clark, “Numerical and symbolic approaches to uncertainty management in AI”, *Artificial Intelligence Review,* vol. 4, no. 2, pp. 109-146, 1990.

[2] J. Pearl, “Probabilistic reasoning in intelligent systems: Networks of Plausible Inference”, Morgan Kaufmann Publishers Inc, san Mateo, California, 1988.

[3] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, D. J. Spiegelhalter, “Probabilistic Networks and Expert systems”, Springer –Verlag New York inc, 1999.

[4] S. Mussi, “General environment for probabilistic predictive monitoring”, *International Journal of Computers*, vol. 7, no. 2, pp. 31-49, 2013.

[5] F. V. Jensen, “An introduction to Bayesian networks”, UCL Press, 1996.

[6] S. Spiewak and M. Szafarczyk, “A Predictive Monitoring and Diagnosis System for Manufacturing”, *CIRP Annals - Manufacturing Technology*, vol. 40, no. 1, pp. 400-403, 1991.

[7] J. Lee, J. Ni, D. Djurdjanovic, H. Qiu and H. Liao, “Intelligent prognostics tools and e-maintenance”, *Computers in Industry*, vol.57, no. 6, pp. 476-489, 2006. [Online]. Available: http://dx.doi.org/10.1016/j.compind.2006.02.014

[8] H. Liao and J. Lee, “Predictive Monitoring and Failure Prevention of Vehicle Electronic Components and Sensor Systems”, *SAE 2006 World Congress & Exhibition, April 2006*, Detroit, MI, USA, Session: Automobile Electronics and Systems Reliability (Part 1 of 2).

[9] R. Kothamasu, S. H. Huang and William H. VerDuin, “System health monitoring and prognostics—a review of current paradigms and practices”, *The International Journal of Advanced Manufacturing Technology*, vol. 28, no. 9-10, pp. 1012-1024, 2006. [Online]. Available: http://dx.doi.org/10.1007/s00170-004-2131-6

[10] F. Ly, A. K. A. Toguyeni and E. Craye, “Indirect predictive monitoring in flexible manufacturing systems”, *Robotics and Computer-Integrated Manufacturing*, vol. 16, no. 5, pp. 321-338, 2000. [Online]. Available: http://dx.doi.org/10.1016/S0736-5845(00)00015-6

[11] B. C. Juricek, D. E. Seborg and W. E. Larimore, “Predictive monitoring for abnormal situation management”, *Journal of Process Control*, vol. 11, no. 2, pp. 111-128, 2001. [Online]. Available: http://dx.doi.org/10.1016/S0959-1524(00)00043-3

[12] S. A. Spiewak, R. Duggirala and K. Barnett, “Predictive Monitoring and Control of the Cold Extrusion Process”, *CIRP Annals - Manufacturing Technology*, vol. 49, no. 1, pp. 383-386, 2000.

[13] J Jeng, C Li, H Huang, “Dynamic Processes Monitoring Using Predictive PCA”, *Journal of the Chinese Institute of Engineers*, vol. 29, no. 2, pp. 311-318, 2006.

[14] G. Chen, T. J. McAvov, “Predictive on-line monitoring of continuous processess”, *Journal of Process Control*, vol. 8, no. 5-6, pp. 409-420, 1998.

[15] A. Ali, A. Khelil, F. K. Shaikh and N. Suri, “MPM: Map based Predictive Monitoring for Wireless Sensor Networks”, presented at Autonomic Computing and Communications Systems, Third Int. ICST Conf. Autonomics 2009, Limassol, Cyprus, September 9-11, 2009.

[16] J. Chen, T.-Y. Hsu, C.-C. Chen, and Y.-C. Cheng, “Online Predictive Monitoring Using Dynamic Imaging of Furnaces with the Combinational Method of Multiway Principal Component Analysis and Hidden Markov Model”, *Industrial & Engineering Chemistry Research*, vol. 50, no 5, pp. 2946-2958, 2011. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ie100671j

[17] S. C. Choi and P. A. Pepple, “Monitoring Clinical Trials Based on Predictive Probability of Significance”, *Biometrics*, vol. 45, no. 1, pp. 317-323, 1989. [Online]. Available: http://www.jstor.org/stable/2532056

[18] J. Reifman, S. Rajaraman, A. Gribok and W. K. Ward, “Predictive Monitoring for Improved Management of Glucose Levels”, *Diabetes Science Technology*, vol. 1, no. 4, pp. 478-486, 2007.

[19] C. Pérez-Gandía, A. Facchinetti, G. Sparacino, C. Cobelli, E.J. Gómez, M. Rigla, A. de Leiva and M.E. Hernando, “Artificial Neural Network Algorithm for Online Glucose Prediction from Continuous Glucose Monitoring”, *Diabetes Technology & Therapeutics*, vol. 12, no. 1, pp. 81-88, 2010. [Online]. Available: http://dx.doi.org/10.1089/dia.2009.0076

[20] D. P. O'leary, L. L. Davis and S. Li, “Predictive Monitoring of High-frequency Vestibulo-ocular Reflex Rehabilitation Following Gentamicin

Ototoxicity", *Acta Oto-Laryngologica*, vol. 115, no. S520, pp. 202-204, 1995.

[21] Y. Kim, M. Kim, J. Lim, J. T. Kim, C. Yoo, "Predictive monitoring and diagnosis of periodic air pollution in a subway station", *Journal of Hazardous Materials*, vol. 183, no. 1-3, pp. 448-459, 2010.

[22] M. M. Zamzam, V. B. Reddy, K Al Bisher, "Reducing Energy and Emissions through Predictive Performance Monitoring System ",*Abu Dhabi International Petroleum Conference and Exhibition*, 11-14 November 2012, Abu Dhabi, UAE.

[23] E. Pathmanathan, R. Ibrahim, V. S. Arsivadm, "CO2 emission model development employing particle swarm optimized — Least squared SVR (PSO-LSSVR) hybrid algorithm", *International Conference on Intelligent and Advanced Systems (ICIAS*), 12-14 June 2012.

Silvano Mussi received the degree in Physics in 1975 from the University of Milan, Italy. He has been with a telecommunication company (ITALTEL) for over 10 years, working in the Research and Development department, in the fields of Software Engineering and Functional Discrete Simulation of Real-time Systems. Then, for over 20 years he has been with an interuniversity consortium for information technology (CILEA) where for over 10 years he has played the role of marketing manager. For over 10 years he has done research in the fields of Knowledge Engineering and Expert Systems. For some years he has been contract-professor of Artificial Intelligence at the University of Brescia. His current research interests are in the fields of probabilistic expert systems, predictive monitoring and decision support systems. His scientific publications can be found at the Web-page: www.cheerup.it