Appl Intell (2016) 45:1034–1046
DOI 10.1007/s10489-016-0807-x

Quality-optimized predictive analytics

Christos Anagnostopoulos[1]

Published online: 27 June 2016

Abstract On-line statistical and machine learning analytic tasks over large-scale contextual data streams coming from e.g., wireless sensor networks, Internet of Things environments, have gained high popularity nowadays due to their significance in knowledge extraction, regression and classification tasks, and, more generally, in making sense from large-scale streaming data. The quality of the received contextual information, however, impacts predictive analytics tasks especially when dealing with uncertain data, outliers data, and data containing missing values. Low quality of received contextual data significantly spoils the progressive inference and on-line statistical reasoning tasks, thus, bias is introduced in the induced knowledge, e.g., classification and decision making. To alleviate such situation, which is not so rare in real time contextual information processing systems, we propose a progressive time-optimized data quality-aware mechanism, which attempts to deliver contextual information of high quality to predictive analytics engines by progressively introducing a certain controlled delay. Such a mechanism progressively delivers high quality data as much as possible, thus eliminating possible biases in knowledge extraction and predictive analysis tasks. We propose an analytical model for this mechanism and show the benefits stem from this approach through comprehensive experimental evaluation and comparative assessment with quality-unaware methods over real sensory multivariate contextual data.

Keywords Predictive analytics · Stochastic decision making · Optimal stopping theory · Contextual data streams · Quality of information

1 Introduction

In real-life scenarios, wireless sensor networks in Internet of Things (IoT) environments have been widely utilized in contextual information monitoring and on-line large-scale predictive analytics, including environmental monitoring, forest/marine environmental monitoring, and smart cities intelligence applications. IoT predictive intelligence applications process contextual information captured from a number of dedicated sensor (stationary and/or mobile) nodes (sources of contextual information) with advanced sensing and computing capabilities. Sources sense and monitor, e.g., physical contextual parameters (*context*) and transmit the collected pieces of context to a central predictive analytics and information processing system (hereinafter referred to as *System*) using wireless communication technologies, e.g., multi-hop communication. However, the sensory field of the sources, e.g., IoT wireless devices within a city area, has a number of inherent characteristics including uncontrollable environments and topological constraints. Sources are typically powered by batteries and thus having limited energy resources. Moreover, environmental monitoring, IoT smart applications, and on-line statistical analytics applications require efficient, accurate and timely data analysis in order to facilitate (near) real-time critical decision-making, and situation- and context- awareness.

Accurate predictive analytics relies on the *quality* of context and *quality* of context inference expressed by meta-information [1], e.g., contextual value validity thresholds, outliers, expiration thresholds, contextual information with

✉ Christos Anagnostopoulos
christos.anagnostopoulos@glasgow.ac.uk

[1] School of Computing Science, University of Glasgow, G12 8QQ, Glasgow, UK

enhanced semantics. Raw contextual observations collected from sources, however, may have low quality and reliability due to limited energy and computational resources and harsh deployment environments. Predictive analytic tasks like outliers detection, multivariate regression and classification, information fusion (e.g., aggregation), and situational context inference and reasoning, are in need of high quality of sensed context. Inaccurate observations resulting from sources malfunction need to be corrected or removed [8]. This however yields bias in the extracted knowledge and analytics tasks, e.g., false alarms for fire detection, high prediction error in regression models, incompatible context inference, high misclassification errors, inconsistent reasoning. Machine and Statistical Learning (MSL) methods are adopted for (i) identifying and (ii) (ideally) correcting *problematic* context (e.g., missing values, obsolete data, and outliers). Such MSL methods are of high importance for knowledge extraction, inference, and decision making over incomplete underlying data [6]. Most MSL techniques, such as neural networks and support vector machines, fail if one or more inputs contains missing values and thus cannot be used for predictive analytics and decision-making purposes [7].

In the state of the art, it is possible to find quite a few IoT monitoring and predictive analytics solutions such as forest monitoring [2], fire-event prediction and classification [3], agriculture monitoring [4], marine environment states prediction [5], watershed prediction systems [20], health states prediction in rivers [21], or energy management solutions to reduce both the amount of resources needed and the atmospheric emissions [22]. The reader could also refer to the survey [23] and the references therein. Sensor networks as the pillars of the contextual information sources promise to revolutionize sensing in a wide range of intelligent application domains because of their reliability, accuracy, flexibility, cost effectiveness [24] and ease of deployment. However, contextual data streams pose a challenge to large-scale predictive analytics because, traditional approaches to quality control cannot efficiently (i) handle large-scale observations and (ii) deal with the demands of real-time processing. There is an increasing need for **predictive intelligence methods** to check and correct (sensed) context to ensure that is delivered in near real time and is of the highest quality. Time-optimized context quality control expedites post-processing and analytics (e.g., missing values substitutions, concept drift correction) so that the final delivered context is of high quality for further processing regression/classification tasks. This motivated us to introduce an *optimally scheduled context quality aware mechanism* which improves the quality of the delivered context to the System for near real time predictive analytics and knowledge extraction. The proposed mechanism materializes quality assessment prior to delivery of the context to the System by minimizing the induced bias in statistical inference and/or estimation processes due to problematic sensed context. As it will be shown in the experimental evaluation section, our mechanism delivers contextual information to the System of high quality (e.g., as much non-problematic and accurate data as possible) inducing a relatively small delay compared to solutions that either immediately deliver context or decide on context delivery upon threshold-based rules that do not take into account the quality dynamics of the contextual data.

2 Rationale

The rationale behind the proposed mechanism is to deliver high quality context to the predictive analytics System through a stochastically, optimally controlled (delivery) delay. Within such delay tolerant delivery, the mechanism optimally decides when to deliver context with the highest possible quality, thus, improving predictive analytics tasks. The mechanism delivers context (represented by a row vector) $\mathbf{x} = [x_1, \ldots, x_n]$ of n measurements (values), where each x_i corresponding to the i-th source, with the least possible *problematic* pieces of data. We require that System receive *good* context $\mathbf{x}$ in the sense that it consists with as many non-problematic values as possible. This is mandatory since the quality of $\mathbf{x}$ affects the predictive analytics tasks for monitoring the state of nature in the receptive field and/or MSL methods for knowledge extraction. We abstract such methods/tasks through a function $f(\mathbf{x})$ over sensed context $\mathbf{x}$, which formulates a MSL/predictive analytics process. For instance, $f(\mathbf{x})$ refers to a statistical metric like mean value, or to a multivariate regression model, e.g., linear regression model $f(\mathbf{x}) = \mathbf{w}^{\top}\mathbf{x} + b, b > 0$ with $\mathbf{x}$ being the predictor vector and $\mathbf{w}$ the learned parameter, or to a classification model, e.g., $f(\mathbf{x}) = sign(\mathbf{w}^{\top}\mathbf{x} + b)$. Inevitably, the more non-problematic values the System receives, the more accurate an analytics process in terms of $f(\mathbf{x})$ can be achieved. Our mechanism attempts to deliver as good context as possible to achieve a high quality of the invoked analytics process. In that sense, we delay the delivery of context to the System in hope of observing a relatively good one to deliver, however at the expense of a certain delay. Figure 1 shows the rationale of the proposed mechanism. The baseline solution is to immediately deliver the current context $\mathbf{x}$ to the System not taking into account the context quality semantics.

2.1 Motivation

We report on four real-life cases / scenarios in order to further exemplify our motivation on the application of quality-optimized predictive analytics.

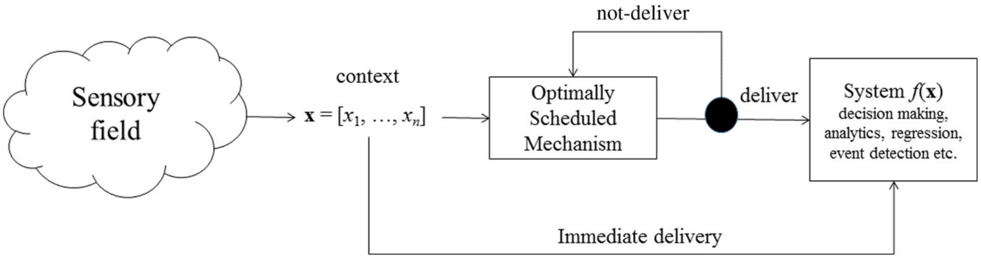
Sensory field
context
$\mathbf{x} = [x_1, \ldots, x_n]$
not-deliver
Optimally Scheduled Mechanism
deliver
System $f(\mathbf{x})$
decision making, analytics, regression, event detection etc.
Immediate delivery

Fig. 1 Overall idea: context **x** from the sensory field is fed to the optimally quality-driven scheduled mechanism, which either delivers **x** to the System or waits for possibly high quality context

Case 1 [Incomplete Contextual Data] If, at a given time instance, a portion of the received values to System are problematic, say $x_1, \ldots, x_m$ with $m < n$ missing values, then there might be a bias in further processing of **x**. For instance, consider the deviation on estimating the mean value of $n-m$ observed values i.e., $f(\mathbf{x}) = \frac{1}{n-m}\sum_{i=1}^{n-m} x_i$ instead of n values $\frac{1}{n}\sum_{i=1}^{n} x_i$, or on estimating the order statistics, e.g., $f(\mathbf{x}) = \min(x_1, \ldots, x_{n-m})$ instead of $\min(x_1, \ldots, x_n)$; recall the 'effect size' problem in statistics [25] where the statistical error is proportional to $1/\sqrt{n-m}$. Moreover, a missing value substitution algorithm (MVA) [26] running on the System, which is able to predict the most plausible values for the m missing values of **x**, results in higher accuracy when m is relatively smaller compared to n. Hence, a delay in avoiding the delivery of *bad* **x** (with a high number of missing values) could be of high importance in terms of accuracy of prediction and, more interestingly, avoiding the MVA invocation each time a bad vector is available, thus eliminating redundant waste of resources [9].

Case 2 [Validity of Contextual Data] Consider that an analytic task like concept drift detection or novelty detection task that requires its input **x** to contain a high number of non-expired values. Here we deal with the fact that the validity of each value x_i is characterized by an expiration window. That is, for each value x_i there is an expiration indicator $I_i(x_i) = 1$ if x_i is a valid value; 0 otherwise (i.e., expired value). The mechanism has to 'delay' the delivery of **x** to the detection algorithm by attempting to find a *better* vector of n values at some unknown time in the future, which maximizes the $f(\mathbf{x}) = \sum_{i=1}^{n} I_i(x_i)$, i.e., context that contains a high number of valid values.

Case 3 [Contextualized Inference] Contextual data fusion processing has gained significant importance [10]. Contextual data fusion refers to the problem of combining diverse and conflicting contextual information provided by sources, in a consistent and coherent manner [11]. The objective of the contextualized inference is to infer a subtaxonomy of situations (from the very abstract to the very specific) of a system that is being observed or taxonomy of activities being performed [13, 14]. Specifically, contextualized inference methods [14] are generally applied in situation- and context-aware systems [16, 17], where a more specific situation (positioned at the lower levels of the situational taxonomy) is represented by the logical conjunction of situational components [12, 15]. Let us adopt the by far popular IF-THEN situational knowledge representation inference rule, i.e., the logical conjunction $f(\mathbf{x}) = \bigwedge_{i=1}^{n}(f_i(x_i)) \in \{\texttt{TRUE}, \texttt{FALSE}\}$ of n logical operators $f_i(x_i)$ over aggregated (or not) values x_i, e.g., the situational component $f_i(x_i) = \texttt{TRUE}$ if $x_i \in [x_i^{low}, x_i^{high}]$; $\texttt{FLASE}$, otherwise. That is, $f(\mathbf{x})$ is envisaged as an IF-THEN situational rule for evaluating the current situational context given the current context **x**. A predictive analytics system caters for inferring the most specific situation within a situational taxonomy. That is, situation $f(\mathbf{x})$ conveys more information to the system than situation $f'(\mathbf{x})$ iff one can deduct $f'(\mathbf{x})$ from $f(\mathbf{x})$, i.e., $f(\mathbf{x})$ contains more TRUE situational components than the $f'(\mathbf{x})$. Such a situation-aware system has to 'delay' its situational inference by observing as much true facts, i.e., components with $\{x_i = \texttt{TRUE}\}$, as possible to reason about more specific situations, which further activates more specific actuation rules and decisions, compared with the 'trivial' abstract situations, i.e., those containing a high number of $\{x_i = \texttt{FALSE}\}$ components.

Case 4 [Progressive and Maintenance Analytics] The author would like to mention the prior work [18] and [19] on dealing with the optimal maintenance of the top-k list of objects over incomplete multivariate data streams and intelligent progressive Big Data analytics. The work [18] refers to an intelligent scheduling of top-k list maintenance with the purpose of increasing the quality of the delivered list to a analytics back-end system. Generally speaking, in this case the $f(\mathbf{x})$ abstracts the degree of updates of sequential partial results **x** from merged top-k lists. Hence, a predictive analytics system 'delays' its final top-k list maintenance based on the up-to-now seen quality of partial results. The work in [19] deals with continuous queries over a distributed federation of data nodes and returns the final outcome to users or analytics applications. The system based on the current quality of the up-to-now retrieved partial results (abstracted by a non-trivial $f(\mathbf{x})$ over partial results **x**) engages a subset of query processors to further execute the issued queries.

In both analytics systems, one has to define an optimally scheduled mechanism over queries to provide optimal decisions on when to invoke a maintenance process [18] or further analyzing data given analytics queries [19].

In all these real-life cases, the predictive analytics system requires *more* information or *quality* information in order to proceed with an analytics task, e.g., either situational inference, aggregation, or classification tasks. However, a delay in the delivery of vectors **x** to the System incurs some *penalty*, especially when dealing with real time predictive analytics as in the above mentioned cases. On the one hand, we require immediate consumption of the observed pieces of context **x** by the predictive analytic tasks. On the other hand, we require a high quality of the analytics / prediction / classification results, which fundamentally relies on the quality of the received pieces of context, i.e., the input to the System. We attempt to reduce the redundant invocations of predictive analytics tasks with inputs of low quality, which inevitably lead to 'biased' inference and statistical reasoning results. Evidently, there is a trade-off between delaying the consumption of the observed context (thus feeding the System with high quality of context) and the *near* real time processing associated with a delay-tolerant predictive analytics process. The problem here is to determine *when* to deliver high quality context balancing between quality of analytics results and near real time predictive analytics.

2.2 Contribution & organization

The contribution of this paper is an analytical stochastic optimization mechanism, which monitors streams of pieces of context and optimally determines when to deliver context of high quality to the System for predictive analytics. Such mechanism is based on the principles of the theory of optimal stopping [27] through which we derive an optimal decision time to 'stop' observing the contextual data stream and to 'deliver' context such that the expected predictive analytics quality is maximized given a certain cost per observation. The theory of optimal stopping [27] is proved to be very efficient in cases where we try to find the appropriate decision time instance to stop the observation of a stochastic process with the objective of maximizing our payoff or reward. Naturally, we build our mechanism on the principles of the optimal stopping theory to maximize the quality of predictive analytics results by inducing a controlled delay. Through this delay we attempt to balance between immediate and delayed predictive analytics in hopes of observing higher quality pieces of contextual information as illustrated in Cases 1–3. The outcome of the mechanism indicates whether we should stop observing the quality of the context streams and activate a predictive analytics and/or MSL method, or to continue. This delay-tolerant activation supports intelligent analytics applications that can tolerate some delay in hopes of obtaining high quality results, like: (i) progressive query analytics applications in large-scale distributed systems [19], (ii) results maintenance of rank-based queries over data streams [18], (iii) efficient networking analytics applications for location-based services [34], (iv) efficient and progressive recommendations of recommendation systems and applications [35], (v) efficient user's mobility and trajectory patterns extraction in mobile computing environments [36], (vi) quality information forwarding and dissemination in mobile applications over IoT environments [37–39], and (vii) security analytics for location-privacy [40].

As it will be shown in the performance assessment, our mechanism provides a wide range of quality results, ranging between medium quality results with almost zero delay and high quality results with an acceptable delay. Through this delay (in terms of the application tolerance), the System saves computational resources and eliminates redundant activations of MSL methods/analytics tasks.

The contribution of this work is summarized as follows:

– A novel stochastic optimization mechanism which decides when a predictive analytics task should be activated over large-scale contextual data streams by guaranteeing the highest possible quality results.
– An analytical model under the principles of the optimal stopping theory that derives the optimal time for activating the predictive analytics tasks.
– Comprehensive experimental results showcasing the benefits of our mechanism to real life intelligent predictive analytics applications over real contextual data involving widely applied aggregation analytics vis-à-

Table 1 Nomenclature

Concept	Description
t, T	discrete time instance, optimal stopping time instance
n	number of contextual data streams
$\mathbf{x}$	context vector
$f(\mathbf{x})$	predictive analytics function over **x** (abstraction)
β	probability of 'good' value
X_t^i	quality indicator of the i-th measurement at time t
Y	quality reward
M	quantity of 'good' values
$\mathcal{F}_t$	filtration up to time t
V^*	maximum expected quality reward
y	scalar value/estimation of V^*
c	delay cost per observation
$\mathcal{L}_t$	log-likelihood up to t

vis the threshold-based and immediate context delivery approaches.

The paper is organized as follows: Section 3 introduces the concept of context quality for data streams of (possibly problematic) contextual data and some preliminaries in the theory of optimal stopping. Section 4 formulates and provides a solution to the quality-optimized mechanism for the considered stochastic optimization problem. Section 5 reports on the experimental results of our mechanism through a sensitivity analysis of the basic parameters and provides a comparative assessment with threshold-based and immediate context delivery rules over real sensors contextual data. Finally, Section 6 concludes the paper and discusses future research on that topic.

3 Definitions

Table 1 refers to the nomenclature.

3.1 Quality of contextual information

Consider a discrete time domain $\mathbb{T} = \{1, 2, \ldots\}$ such that $\mathbf{x} = [x_1, \ldots, x_n]$ contains real values $x_i \in \mathbb{R}$ at time $t \in \mathbb{T}$ for each dimension $i \in 1, \ldots, n$ (or in a compact notation $i \in [n]$). We assume that x_i at time t refers to the measurement of source i or the aggregation result over K measurements $x_{i1}, \ldots, x_{iK}$ launched on source i, $K > 0$. (The value x_{ij} could refer to a measurement of the j-th neighboring node in the spatial neighborhood of source i, $j \in [K]$.) Each measurement x_i is received instantly and that a new possible value might be received from the same source i only at the next time slot $t + 1$, i.e., in the interval $[t, t + 1)$ source i reports only once or not at all.

We proceed with a generic model representation to capture the idea of a good piece of context $\mathbf{x}$. Specifically, the characterization of $\mathbf{x}$ as a 'good' piece of context intuitively indicates that $\mathbf{x}$ contains a relatively high number of good values, e.g., a percentage of 75 % of the n values of context $\mathbf{x}$ refers to non-missing values. A 'good' value x_i at time t means, for instance, that x_i is a valid value, a non-incomplete value, or a TRUE fact/situation, i.e., $I_i(x_i) = 1$ as discussed in Cases 1 and 2 or $I_i(x_i) = \texttt{TRUE}$ in Case 3, while $I_i(x_i) = 0$ indicates a bad value, or a missing datum (Cases 1,2) or a situation does not hold true ($I_i(x_i) = \texttt{FALSE}$ in Case 3). Or, if x_i is observed at time t thus not being missed as discussed in Case 1, then x_i is called a good value, otherwise it is called a bad value, i.e., a missing value. Based on all these interpretations, we provide the following definitions:

Definition 1 The quality indicator of the i-th measurement (i.e., from the i-th source) is define as the random variable (r.v.) X_t^i such that:

$$X_t^i = \begin{cases} 1(\texttt{TRUE}) & \text{with probability } \beta_i \\ 0(\texttt{FALSE}) & \text{with probability } 1 - \beta_i, \end{cases} \tag{1}$$

where a zero value, i.e., $X_t^i = 0$, indicates a bad value of dimension i at time t while a value $X_t^i = 1$ refers to a good value x_i at t.

The r.v. $X_1^i, X_2^i, \ldots$ are independent and identically distributed (i.i.d.). with expectation $E[X^i] = 1 \cdot P(X^i = 1) + 0 \cdot P(X^i = 0) = \beta_i > 0$ given that $\beta_i \in (0, 1), i \in [n]$. The value of β_i can be estimated by historical data and/or combined with information provided by the manufacturer of source i, e.g., quantifying sensor node degree of reliability of measurement. (Remark 2 provides an estimation of the β parameter.) Each time t the mechanism observes context $\mathbf{x}$ and does not immediately deliver it to the System, we encounter fixed a (delay) cost of observation $c > 0$.

Definition 2 We define as *quality reward* of context $\mathbf{x}$ at time t the r.v. Y_t, which refers to the quantity of the good values $M_t = \sum_{i=1}^n X_t^i$ minus the total observation cost up to time t, i.e.,

$$Y_t = \sum_{i=1}^{n} X_t^i - t \cdot c = M_t - t \cdot c. \tag{2}$$

3.2 Preliminaries on the optimal stopping theory

The theory of optimal stopping [27, 28] is concerned with the problem of choosing a time instance to take a certain action, in order to minimize an expected loss (or maximize an expected payoff). A stopping rule problem is associated with:

- a sequence of random variables (r.v.) $M_1, M_2, \ldots$, whose joint distribution is assumed to be known and
- a sequence of payoff (reward) functions $(Y_t(M_1, \ldots, M_t))_{1 \leq t}$ which depend only on the observed values of the corresponding r.v.s $M_1, \ldots, M_t$.

The available information up to t is a sequence $\mathcal{F}_t$ of values of the r.v.s $M_1, \ldots, M_t$ (a.k.a. filtration). The optimal stopping rule problem is defined as follows: We are observing the sequence of the r.v.s $(M_t)_{1 \leq t}$, and at each time instance t, we can choose to either stop observing or continue. If we stop observing at time instance t, we get reward Y_t. We desire to choose a stopping rule or stopping time to maximize our expected reward.

Definition 3 An optimal stopping rule problem is to find the stopping time T which maximizes the expected reward, i.e., $E[Y_T] = \sup_{0 \leq t \leq \mathcal{T}} E[Y_t]$. Note, $\mathcal{T}$ might be ∞.

4 Time-optimized quality-driven mechanism

The mechanism observes the sequence of r.v. $M_1, M_2, \ldots, M_t$ without delivering the corresponding pieces of context $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$ to the System. Our aim is to find the best strategy in the sense of having the highest expected quality reward $E[Y]$ at the lowest cumulative cost of delay. At each time t we only need to decide:

- whether to deliver $\mathbf{x}_t$ to the System, thus, proceeding with a predictive analytic task over $f(\mathbf{x}_t)$ or
- to continue with the next observation $\mathbf{x}_{t+1}$ without delivering $\mathbf{x}_t$ to System, thus, delaying the predictive analytic task.

Hence, a strategy is a function which assigns to each sequence $M_1, M_2, \ldots$ a *stopping time*. Furthermore, since we cannot see the future, a decision to stop observation at time t can only depend upon $M_1, M_2, \ldots$ Formally we have to solve the following problem:

Problem 1 Given the sequence of sums of quality indicators $M_1, \ldots, M_t$, find the optimal stopping time T which maximizes $E[Y_T] = \sup_{0 \leq t < \infty} E[Y_t]$.

The idea is to find a criterion at time instance t such that given the current value of M_t, denoting the current quality of context observed at the mechanism, the latter immediately decides whether to deliver $\mathbf{x}_t$ to the System or to continue to the next observation. We require an immediate decision making over the contextual data streams, thus, avoiding any redundant computations. As it will be shown in the remainder, the mechanism at time instance t proceeds with a time-optimized decision in $O(n)$ time involving simply the counting of quality indicators X_t^i from all n sources, $i \in [n]$.

In order to solve Problem 1, we rest on the principle of optimality. Specifically, let T be the optimal stopping time where the supremum in our Problem 1 is attained, i.e., $E[Y_T] = V^*$ with $V^* = \sup_t E[Y_t]$. We can now provide the optimality equation given the filtration $\mathcal{F}_t$, i.e., after observing $M_1, \ldots, M_t$, as follows:

Theorem 1 *Let T be an arbitrary stopping time and $V_t^* = \sup_{T \geq t} E[Y_t|\mathcal{F}_t]$. Then, $V_t^* = \max(Y_t, E[V_{t+1}^*|\mathcal{F}_t])$*

Proof See [27] □ □

The optimal stopping time T given by the principle of optimality from Theorem 1 is represented by the rule:

$$T = \min\{t \geq 0 | Y_t = V_t^*\}. \tag{3}$$

Let us put the reward $Y_0 = -\infty$ to force our mechanism to take at least one observation. Also, we put $Y_\infty = -\infty$ as naturally the cost of an infinite number of observation is infinite. Consider now the V^* the expected quality reward for the System based on an optimal stopping rule in (3). Suppose that the mechanism induces cost c and observe the M_1. Note that if the mechanism continues from this point then quality M_1 is 'lost' and the cost c is already paid. Hence, it is just like starting the problem over again. That is, if the mechanism continues from this point, the System can obtain an expected quality reward of V^* but no more. Therefore, from the principle of optimality in Theorem 1 we derive that if $M_1 < V^*$ then the mechanism should continue; if $M_1 > V^*$, then the mechanism should stop and deliver context to the System. For $M_1 = V^*$ both decisions are optimal; we adopt here a stopping decision. This argument is made at any stage t by the mechanism, thus, in our case we provide the optimal stopping rule, which is adopted by the mechanism, as follows:

Theorem 2 *Given the sequence $M_1, \ldots, M_t$ there is a real number $y = V^*$ such that the optimal stopping time T is given by $T = \min\{t \geq 1 | M_t > y\}$ with $E[Y_T] = y$.*

Proof The r.v. $M = \sum_{i=1}^n X^i$ takes realization discrete values from $\{0, 1, \ldots, n\}$. Now, at the optimal stopping time $t = T$, i.e., the first time at which $M_t > y$, we obtain $E[Y_T] = E[M_T] - E[T]c = \sum_{i=1}^n E[X_T^i] - E[T]c$. Moreover, let $\gamma = P(M > y)$ and $\delta = 1 - \gamma = P(M \leq y)$. Then, we obtain

$$\begin{aligned} E[M_T] &= \sum_{k=1}^{\infty} E[M_k | M_k > y, M_1 \leq y, \ldots M_{k-1} \leq y] \\ &= \sum_{k=1}^{\infty} E[M_k | M_k > y] \delta^{k-1} = E[M_1 | M_1 > y] \frac{1}{\gamma}. \end{aligned}$$

The quantity $E[M_T] = \frac{1}{\gamma} E[M_1 | M_1 > y]$ indicates that at the optimal stopping time T, the expected context quality equals to the expected context quality given that the latter is above the criterion threshold $y = V^*$. In addition, for the optimal stopping time T we obtain

$$\begin{aligned} E[T] &= \sum_{k=1}^{\infty} k P(M_k, M_k > y, M_1 \leq y, \ldots M_{k-1} \leq y) \\ &= \gamma \sum_{k=1}^{\infty} k \delta^{k-1} = \frac{1}{\gamma}. \end{aligned}$$

The problem now is to compute $y = V^*$. This is done through the optimality equation in Theorem 1 and the above mentioned argument, i.e.,

$$\begin{aligned} V^* &= E[\max(M_1, V^*)] - c \Leftrightarrow \\ c &= E[(M_1 - V^*)^+] \end{aligned}$$

That is a quality reward $E[Y_T]$ is obtained at the optimal stopping time T with quality reward greater than y and y is the solution of the $E[(M_1 - y)^+] = c$, with $(x - y)^+ = \max(0, x - y)$

Hence, having an y such that $c = E[\max(0, M_1 - y)] = E[(\sum_{i=1}^{n} X_1^i - y)^+]$, we obtain

$$\begin{aligned} E[Y_T] &= E[M_T] - \frac{1}{\gamma} c \\ &= \frac{1}{\gamma} (E[M_1 | M_1 > y] - E[M_1 - y | M_1 > y]) \\ &= \frac{1}{\gamma} E[y] P(M_1 > y) = y. \end{aligned}$$

Hence, the optimal stopping time T achieves the maximal expected quality reward $E[Y_T] = y$. □

Remark 1 The optimal rule in Theorem 2 is optimal for our problem since $E[(M - y)^+] - c$ is monotonically non-decreasing with M for $M > y$ almost surely and $E[(M - y)^+]$ is continuous in y and decreasing from $+\infty$ to zero. Hence there is a unique solution for y for any $c > 0$.

The mechanism stops the observation process of pieces of context and delivers context $\mathbf{x}_t$ at the first time instance t at which the quantity of the good values M_t is above a threshold $y \in \mathbb{R}$, which refers to the highest quality of reward that can be obtained. The problem now reduces on the evaluation of the y value such that $E[(M_1 - y)^+] = c$. The algorithm of our mechanism is shown in Fig. 2. The input of the algorithm is the stopping criterion y. At each received context $\mathbf{x}_t$, the mechanism calculates M_t and decides whether to deliver $\mathbf{x}_t$ to the System or not. In the former case, the mechanism start-off with the next sequence of (M_t). Evidently, the computational time for evaluating the criterion $M_t > y$ is $O(n)$.

```
Require: stopping quality criterion $y$
  $t \leftarrow 1$
  STOP ← FALSE
  while STOP = FALSE do
    RECEIVE context $\mathbf{x}_t = [x_1, \ldots, x_n]$
    CALCULATE context quality $M_t = \sum_{i=1}^{n} X_t^i$
    if $M_t > y$ then
      STOP ← TRUE
    else {take next observation}
      $t \leftarrow t + 1$
    end if
  end while
  DELIVER $\mathbf{x}_t$ to the System for further processing/predictive analytics.
```

Fig. 2 Algorithm of the quality-optimized mechanism

We proceed our analysis with the case where $\beta_i = \beta$ for all sources, $i \in [n]$. If we notate $Z = \max(M - y, 0)$ and $F_M(y) = P(M \leq y)$ be the cumulative distribution function of M then y is the solution of $E[Z] = c$. We have that $E[Z] = E[M - y | M > y](1 - P(M \leq y)) = (E[M | M > y] - y(1 - F_M(y)))(1 - F_M(y))$. In this case, $M = \sum_{i=1}^{n} X^i$ is a Binomial random variable with parameters (n, β). Hence, we obtain $F_M(y) = \sum_{j=0}^{\lfloor y \rfloor} \binom{n}{j} \beta^j (1 - \beta)^{n-j}$. Moreover, we have that $E[M | M > y] = \sum_{m=0}^{n} m P(M = m | M > y)$ or

$$E[M | M > y] = \frac{1}{1 - F_M(y)} \sum_{m=y+1}^{n} m \binom{n}{m} \beta^m (1 - \beta)^{n-m}.$$

Hence, the expectation of Z is:

$$E[Z] = \sum_{m=y+1}^{n} m \binom{n}{m} \beta^m (1 - \beta)^{n-m} - y(1 - F_M(y))^2 \quad (4)$$

Based on the criterion $E[Z] = c$ and on (4), we can find analytically the value of y. However, the assumption $\beta_i = \beta, \forall i$ does not spoil the theoretical results and is adopted for eliminating the computations of $F_M(y)$ for solving $E[(M_1 - y)^+] = c$. Obviously, when $\beta_i \neq \beta_j, i, j \in [n]$ then $F_M(y)$ is provided in [29] (a.k.a. Poisson-Binomial distribution) thus, we can obtain the corresponding value for y.

Remark 2 The probability β of a non-problematic piece of contextual value X^i can be incrementally estimated by the maximum likelihood estimation of β of the Binomial distribution with parameters (n, β) after observing a series of m pieces of context $(\mathbf{x}_t)_{t=1}^{m}$, $m > 1$. Specifically, recall that the probability density function for the Binomial is $\binom{n}{M} \beta^M (1 - \beta)^{n-M}$ with $M = 0, \ldots, n$. Hence, the log-likelihood $\mathcal{L}_m(\beta)$ of a series of m samples of $M_1, \ldots, M_m$ is

$$\begin{aligned} \mathcal{L}_m(\beta) = & \sum_{i=1}^{m} \ln \binom{n}{M_i} + \ln \beta \sum_{i=1}^{m} M_i \\ & + \left(nm - \sum_{i=1}^{m} M_i \right) \ln(1 - \beta). \end{aligned}$$

Since $\mathcal{L}_m(\beta)$ is a continuous function of β given m observations, i.e., $\beta = \beta_m$, its maximum value derives from the derivative of $\mathcal{L}_m(\beta)$ with respect to β_m by setting it equal to zero, i.e., $\frac{\partial \mathcal{L}}{\partial \beta_m} = 0$. After this calculation, we obtain that up to the m-th observation, the probability β_m is: $\beta_m = \frac{1}{nm} \sum_{i=1}^{m} M_i$. Hence, we can incrementally estimate the β_m value by the previous β_{m-1} and the current value of M_m by using the recursion $\beta_m = \frac{m-1}{m}\beta_{m-1} + \frac{1}{nm}M_m$, with $\beta_1 = \frac{1}{n}M_1$. After a series of m observations, we can learn the $\beta = \beta_m$ and then initiate our mechanism.

5 Experimental evaluation

5.1 Sensitivity analysis

5.1.1 Simulation setup

We study the performance of the proposed Optimal Delivery Approach (ODA) on both analytical model and simulations with respect to the basic parameters, i.e., probability of a good value β, number of sources n, and cost per observation c. We also provide a comparative assessment with a Threshold-based Delivery Approach (TDA) on deciding when to deliver context to System for further processing. Specifically, TDA choses a threshold $\theta \in \{1, n\}$ and delivers context $\mathbf{x}$ at the first time t at which $M_t \geq \theta$. That is, when context $\mathbf{x}$ has at least θ (out of n) non-problematic values, then TDA immediately delivers $\mathbf{x}$ to System.

We define as ‘epoch’ the number of pieces of context an approach (ODA, TDA) has observed until it decides to deliver the current context to the System. Each time t context $\mathbf{x}_t$ is delivered to System, then a new epoch for the approach starts-off. TDA at the beginning of each epoch choses a threshold θ uniformly at random from $\{1, n\}$, while ODA for every epoch applies the threshold y as estimated using (4). In the j-th epoch we measure the quality reward Y_{t_j} when an approach (ODA, TDA) delivers context $\mathbf{x}_{t_j}$ at stopping time t_j. We run experiments for $N = 10^4$ epochs, thus obtaining the average value of Y, i.e., $E[Y] \sim \frac{1}{N}\sum_{j=1}^{N} Y_{t_j}$ for both approaches.

5.1.2 Performance assessment

Figure 3(left) shows the impact of probability β on the average quality reward $E[Y]$ with different cost values c for the analytical model and the simulation results using $n = 30$; we obtain similar results for other n values. It is worth mentioning how accurately the simulation curves fit with the analytical model curves for all parameter values, denoting the capability of the proposed model for predicting the average quality reward given β and c values. Moreover, we observe that as β increases then we obtain higher quality rewards, as expected, since we deal with *less problematic* pieces of data. With the term problematic piece of data, here, we denote that the context vector $\mathbf{x}$ contains more non-missing values than missing values. Statistically, for $\beta > 0.5$, context $\mathbf{x}$ is less problematic than a piece of context $\mathbf{x}'$, with $\beta' < 0.5$, since the former contains, at least, more non-missing values than the latter one. That is, in context $\mathbf{x}$, over 50% of the n values are non-missing given that each value is non-missing with probability over 0.5 by expectation of the Binomial distribution $\sim B(n, \beta)$. Also, the impact of the delay cost on $E[Y]$ is low compared to the impact of β especially when $c > 0.5$.

In Fig. 3(right) we plot the average delay $E[T]$ against the cost c for different values of n with $\beta = 0.8$. $E[T]$ indicates the average number of observations that the mechanism neglects in each epoch before stopping and then delivering context to the System for predictive analytics. As shown in Fig. 3(right), a relatively small delay is tolerated in order to proceed with delivering context of high quality. This indicates the applicability of the proposed ODA to near real-time predictive analytics. Moreover, as the cost per observation decreases then a relatively higher delay is encountered, since low cost c gives the ‘opportunity’ to the mechanism to observe more pieces of context before stopping at a good one, thus, increasing the likelihood of receiving context of high quality. On the other hand, a high cost value reinforces the mechanism to stop (and thus deliver context) at an early stage of each epoch. For instance, for $c = 0.8$ the mechanism, on average, delivers the second received context to System. By tuning the cost we can control the degree of tolerance of the statistical analytics process, with $c \to 1$ indicating a very

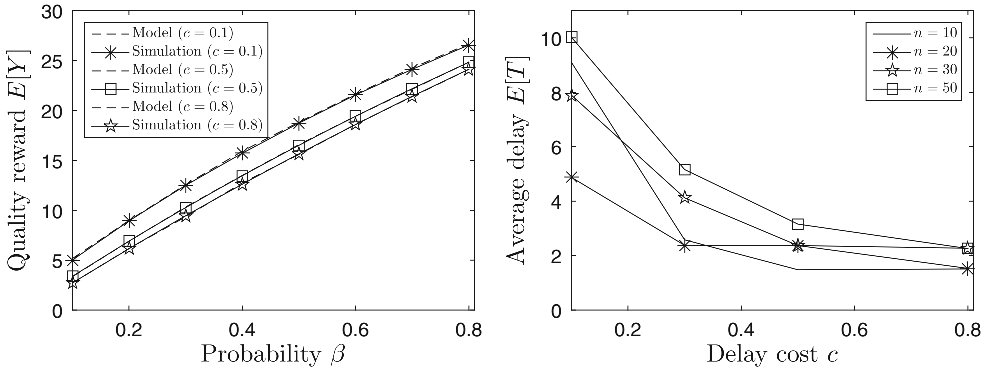

Fig. 3 (*Left*) Quality reward $E[Y]$ against probability β for analytical model and simulations with different cost c and $n = 30$; (*right*) average delivery delay of the proposed approach, i.e., $E[T]$, against cost c for different n values with $\beta = 0.8$

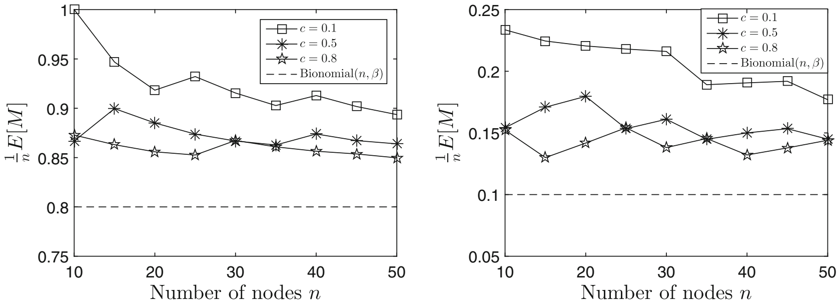

Fig. 4 The NQI and BNQI against number of sources n for different cost c with (*left*) $\beta = 0.8$ and (*right*) $\beta = 0.1$

conservative system, while $c \to 0$ indicating high tolerance to information processing.

Let us define the Normalized Quality Indicator (NQI) $\frac{1}{n}M_t$ of an approach which evaluates the quality of the delivered context $\mathbf{x}_t$ when stopping at time t within an epoch. Recall that M_t indicates the number of non-problematic values that context $\mathbf{x}_t$ contains with $0 \leq M_t \leq n$. Hence a high NQI value close to unity denotes delivered context of high quality. Figure 4 illustrates the average NQI for the ODA (for all epochs, i.e., $\frac{1}{n}E[M] \sim \frac{1}{nN}\sum_{j=1}^{N} M_{t_j}$) against number of sources n for different cost values c and $\beta \in \{0.1, 0.8\}$. It is worth noting that the NQI of an approach that stops the observation process at an arbitrary time and, then, delivers context at that time to the System is $\frac{1}{n}E[M] = \frac{1}{n}\beta n = \beta$, where $E[M] = \beta n$ is the expectation of the Binomial distribution $\sim B(n, \beta)$; we notate this value as BNQI. This approach does not take into consideration the sequence of the r.v. $M_1, \ldots, M_{t-1}$ in order to proceed with a decision at stopping at time t. On the other hand, ODA takes into account the sequence $(M_t)_{t=1}^{T}$ thus exploiting the *knowledge* up to T and then obtaining always higher values than BNQI, even for high cost values as shown in Fig. 4. In addition, NQI for relatively medium/high cost values does not depend on the number of sources n, which means that $E[M]$ increases linearly with n. Note also that the higher the β value, the higher NQI gets since the received context is of high quality, while as $\beta \to 0$ then NQI comes with lower values. However, in that case, NQI is always higher than BNQI indicating the applicability of ODA in cases where the received context contains a high portion of problematic values. Indicatively, for $\beta = 0.01$ we obtain NQI = 1.16 and BNQI = 0.01, i.e., our approach delivers two orders of magnitude more quality context with $n = 30, c = 0.1$. Nonetheless, we have to evaluate the performance of the ODA including also the incurred delay, i.e., $E[T]$, required to proceed with context delivery of high quality. We compare the expected quality reward $E[Y]$ for both approaches (ODA / TDA) for certain values of c, β and n.

Tables 2 and 3 show the average reward $E[Y]$ for both approaches against cost per observation c and probability β with $n \in \{30, 50\}$, respectively. $E[Y]$ quantifies the quality of context delivered when an approach stops at a stopping time t accounting also the cumulative cost for observing t pieces of context. ODA achieves always higher $E[Y]$ value than TDA for all parameters. More interestingly, ODA is deemed appropriate for adopting for delay-tolerant predictive analytics when context contains a high portion of problematic values, i.e., low β values, compared with the performance of TDA. We can observe that for $\beta = 0.1$ and, especially, when the cost of observation is relatively high, i.e., $c = 0.8$, ODA delivers context of (112,129) % more quality compared to TDA in terms of quality reward with $n = (30, 50)$. Moreover, as β increases then ODA and TDA proceed with relatively high $E[Y]$. This is due to the fact that high β values refer to received context of high quality, thus, evidently both approaches would deliver high quality context. However even in this case, ODA outperforms TDA. When the cost of observation is relatively high and the received context contains a low portion of problematic values, ODA is 84 % and 48 % more efficient than TDA in terms of quality reward for $n = 30$ and $n = 50$, respectively; see Tables 2 and 3.

Overall, ODA delivers high quality context to the System, thus, improving the quality of predictive analytics, even when context contains, with a high probability, problematic values and the cost per observation is not negligible.

Table 2 Average quality reward $E[Y]$ for ODA and TDA with $n = 30$

	ODA	TDA	ODA	TDA	ODA	TDA
β	$c = 0.1$		$c = 0.5$		$c = 0.8$	
0.1	5.15	-0.13	3.33	-7.94	2.82	-22.80
0.5	18.82	12.16	16.50	2.62	15.65	-8.69
0.8	26.59	22.37	24.84	19.79	24.21	13.81

Table 3 Average quality reward $E[Y]$ for ODA and TDA with $n = 50$

	ODA	TDA	ODA	TDA	ODA	TDA
β	$c = 0.1$		$c = 0.5$		$c = 0.8$	
0.1	7.98	0.23	5.76	-4.08	5.11	-17.15
0.5	30.36	17.85	27.57	13.89	26.57	5.94
0.8	43.76	36.05	41.55	33.36	40.69	27.84

This is attributed to the fact that ODA exploits the history of the observed sequence of M_t and then decides on the optimal stopping time to deliver context at the expense of a controlled (relatively low) delay.

5.2 Comparative assessment

5.2.1 Experiment setup

We experiment with real contextual data from $K = 16$ chemical sensors exposed to three gases of three chemical compounds at a certain concentration level [32, 33]. Each sensor detects three specific environmental contextual parameters corresponding to Ethylene, Ammonia, and Toluene, respectively. Each sensor $k \in [K]$ measures a triplet $\mathbf{s}_k = [x_{k1}, x_{k2}, x_{k3}]$, where each dimension of $\mathbf{s}_k$ corresponds to the three contextual parameters. The context is then a n-dimensional vector with $n = 3K = 48$ dimensions at time instance t, i.e., $\mathbf{x}_t = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K)$ and the dataset contains 13,910 48-dimensional contextual vectors. We focus in the case where there are missing values for each dimension of the context vector at time instance t. For experimentation, we set the probability of a missing (problematic) value in a dimension with $p = 1 - \beta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, i.e., the probability of being a parameter non-problematic at time instance t is $\beta = 1 - p$.

We consider two scenarios. In the first scenario (Scenario 1), the System processes the delivered context vector $\mathbf{x}_t$, which might include missing values. The process of the System refers to a fusion operator over the contextual values of the vector (described later). In the second scenario (Scenario 2), the System before processing the context vector $\mathbf{x}_t$ invokes a Missing Value substitution Algorithm (MVA) for handling the missing values in $\mathbf{x}_t$. After the invocation of the MVA, the System calls for a fusion operator over the 'imputed' contextual values. The process of the System over context $\mathbf{x}$ refers to two fusion operators over the contextual data. For demonstration, we define two vectorial fusion operators: $f_{avg}(\mathbf{x})$ is associated with the mean value of each chemical compound over all K sensors, and $f_{\min}(\mathbf{x})$ is associated with the minimum value of each chemical compound over all K sensors, as follows:

$$f_{avg}(\mathbf{x}) = \left[\frac{1}{K}\sum_{k=1}^{K} x_{kj}\right], j = 1, 2, 3 \tag{5}$$

$$f_{\min}(\mathbf{x}) = \left[\min_{k\in[K]}\{x_{kj}\}\right], j = 1, 2, 3 \tag{6}$$

Scenario 1 In this scenario, when a dimension x_{kj} is missing, $k \in [K]$, $j = 1, 2, 3$ then, evidently, the operators f_{avg} and $f_{\min}$ do not take into account that dimension in the calculation of the mean or the minimum, respectively; note, there is not MVA invocation in this scenario. We experiment with three approaches (we repeat the ODA and TDA for convenience):

- The Optimal Delivery Approach (ODA), which observes M_t and delivers $\mathbf{x}_t$ when $M_t > y$. Then the System invokes the vectorial operators $f_{avg}(\mathbf{x}_t)$ (and $f_{\min}(\mathbf{x}_t)$). Otherwise, the System takes the next observation, i.e., the next incoming context vector.
- The Immediate Delivery Approach (IDA), which delivers context $\mathbf{x}_t$ at each time instance t to the System. Then, the System invokes at each t the vectorial operators $f_{avg}(\mathbf{x}_t)$ (and $f_{\min}(\mathbf{x}_t)$).
- The Threshold-based Delivery Approach (TDA) with threshold parameter $\theta \in (0, n)$, which observes M_t and delivers $\mathbf{x}_t$ when $M_t > \theta$. Then the System invokes the vectorial operators $f_{avg}(\mathbf{x}_t)$ (and $f_{\min}(\mathbf{x}_t)$). Otherwise, the System waits for the next time instance to process the incoming vector.

The comparative assessment in Scenario 1 is to examine whether the ODA compared with the delay of the TDA and the non-delay of the IDA results to *accurate* fusion results. Specifically, if $\mathbf{x}'_{t_i}$ is the delivered context to the System by an approach at some time instance t_i within the i-th epoch, $i = 1, \ldots, N$, and $\mathbf{x}_t$ is the ground truth (actual) context at that time instance (i.e., without missing values), then we define as mean *fusion error* for the $f_{avg}(\cdot)$ operator as the root mean squared error of the vectorial fused vector, i.e.,

$$e_{avg} = \left(\frac{1}{N}\sum_{i=1}^{N} \parallel f_{avg}(\mathbf{x}_{t_i}) - f_{avg}(\mathbf{x}'_{t_i}) \parallel^2\right)^{1/2} \tag{7}$$

The fusion error $e_{\min}$ for the $f_{\min}(\cdot)$ operator is similarly defined and N is the total number of epochs for each approach. Moreover, we have to include the corresponding expected delay ω_{avg} (and $\omega_{\min}$) of context delivery to the System by an approach (ODA,TDA,IDA) to obtain a certain fusion error. Evidently, the delay for the IDA is zero, since it immediately delivers context to the System for fusion. The expected delay for both ODA and TDA is defined as:

$$\omega = \frac{1}{N}\sum_{i=1}^{N} t_i^*, \tag{8}$$

where t_i^* refers (i) to the optimal stopping time T for the i-th epoch in the ODA, i.e., the first time instance at which $M_{t_i} > y$ and (ii) to the threshold-based stopping time for the i-th epoch in the TDA, i.e., the first time instance at which $M_{t_i} \geq \theta$ for a specific θ.

Scenario 2 In this scenario, when a dimension x_{kj} is missing then its value is filled-in (a.k.a. imputed) by the Exponential Smoothing MVA (ES-MVA) [30] with smoothing factor $a \in (0, 1)$, which is used in time-series contextual

Table 4 Scenario 1: Fusion error e_{avg} and delay ω (in parenthesis)

	$\beta = 0.5$			$\beta = 0.7$		
$\eta = \frac{\theta}{n}$	ODA	IDA	TDA	ODA	IDA	TDA
0.1	80.88 (3.4)	161.96	161 (0)	52.8 (2)	109.5	109.5 (0)
0.3			160 (10)			106.3 (7)
0.5			95.7 (15)			94.5 (12)
0.7			12 (504)			14.3 (443)
0.9			10 (988)			9.5 (1099)

data. Specifically, if the dimension $x_{kj,t}$ at time instance t is missing, which corresponds to sensor $k \in [K]$ and to the chemical compound $j \in \{1, 2, 3\}$, then the ES-MVA replaces it with an estimate $u_{kj,t}$ based on $x_{kj,t-1}$ and the trajectory of this dimension up to $t - 1$, that is:

$$u_{kj,t} = ax_{kj,t-1} + (1 - a)u_{kj,t-1}, \qquad (9)$$

with $u_{kj,1} = x_{kj,0}$. The smoothed statistical estimate $u_{kj,t}$ for the corresponding missing value $x_{kj,t}$ is a weighted average of the previous observation $x_{kj,t-1}$ and the previous smoothed statistical estimate $u_{kj,t-1}$. In this scenario, the three approaches ODA, TDA and IDA deliver the context $\mathbf{x}_t$ to the System as described in Scenario 1. Nonetheless, the System upon receiving the $\mathbf{x}_t$ vector it firstly involves the ES-MVA for imputation and then invoking the fusion operators $f_{avg}(\mathbf{x}_t)$ and $f_{\min}(\mathbf{x}_t)$ of the imputed vector $\mathbf{x}_t$. Moreover, the fusion errors e_{avg} and $e_{\min}$ in this scenario is defined as in Scenario 1 by simply involving the imputed contextual values.

5.2.2 Comparison evaluation

Tables 4 and 5 show the fusion errors e_{avg} and $e_{\min}$ for the f_{avg} and $f_{\min}$ operators, respectively, and the corresponding delay ω (shown within parenthesis) with $\beta \in \{0.5, 0.7\}$ using the approaches ODA, IDA, and TDA repeated for $N = 10^4$ epochs. The results are produced with observation cost $c = 1$; similar results are obtain with other c values. The ODA achieves the lowest error compared to IDA for all cases with a relatively small delay, i.e., number of observations until the mechanism delivers context to the System. This indicates the applicability of our approach for near real-time predictive analytics, by achieving low fusion error compared with the IDA, which achieves 100% higher fusion error by immediately delivering context. Moreover, we experiment with different threshold values for the TDA, i.e., $\theta = \eta n$, with different $\eta \in \{0.1, \ldots, 0.9\}$ percentage. Evidently, the lower the threshold, i.e., the TDA stops at the first time instance the percentage of non-problematic values out of n is over η, the sooner that mechanism delivers context to the System. As shown in Tables 4 and 5, TDA achieves higher fusion error than ODA with relatively higher delay. Specifically, with $\eta \leq 0.5$, ODA outperforms TDA in both error and delay. On the other hand, for $\eta > 0.5$, i.e., TDA considers stopping when at least more than 50 % of the contextual values are non-problematic, it achieves lower fusion error compared to ODA. However, this comes at the expense of a significantly high delay (indicatively % for $\eta = 0.7$). This high delay is prohibitive for (near) real-time statistics analytics, especially in the environmental monitoring, since significant events cannot be captured at the early stages of a monitoring process, e.g., fire or flood detection. Evidently, as β increases all approaches obtain relatively lower fusion error, since less problematic pieces of context are observed. Nonetheless, in this case, TDA achieves extremely high delay for obtaining a low error. In both cases for all β values, the proposed mechanism with significantly low delay achieves low fusion error (in both types of fusion operators). The IDA approach never outperforms ODA in each case, while TDA for $\eta > 0.5$ attempts lower fusion error with one or two orders of magnitude higher delay than that of ODA, thus, yielding it inappropriate for real-time monitoring. It is worth noting that similar behavior will be obtained with other fusion operators that take into consideration the number of current contextual values, since the more non-problematic values we receive the better the accuracy of the event detection. For instance fusion operators over the current context $\mathbf{x}_t$ could be higher order statistics over the n current measurements, the top-K

Table 5 Scenario 1: Fusion error $e_{\min}$ and delay ω (in parenthesis)

	$\beta = 0.5$			$\beta = 0.7$		
$\eta = \frac{\theta}{n}$	ODA	IDA	TDA	ODA	IDA	TDA
0.1	621.2 (3.4)	1240	1240 (0)	439.7	806.9	806 (0)
0.3			1202 (10)			784 (7)
0.5			765 (15)			702 (15)
0.7			50 (504)			46 (443)
0.9			16 (988)			2 (1099)

Table 6 Scenario 2: Fusion error e_{avg} and delay ω (in parenthesis)

	$\beta = 0.5$			$\beta = 0.7$		
$\eta = \frac{\theta}{n}$	ODA	IDA	TDA	ODA	IDA	TDA
0.1	61.02 (3.4)	111.55	111 (0)	32.8 (2)	99.6	99 (0)
0.3			102 (10)			92.4 (7)
0.5			88.3 (15)			74.4 (12)
0.7			9 (504)			11.9 (443)
0.9			7 (988)			8.1 (1099)

Table 7 Scenario 2: Fusion error $e_{\min}$ and delay ω (in parenthesis)

	$\beta = 0.5$			$\beta = 0.7$		
$\eta = \frac{\theta}{n}$	ODA	IDA	TDA	ODA	IDA	TDA
0.1	580 (3.4)	1224	1224 (0)	377.5	838.6	838 (0)
0.3			1212 (10)			824 (7)
0.5			715 (15)			587.6 (15)
0.7			46.7 (504)			47.6 (443)
0.9			17.8 (988)			2.3 (1099)

sources with respect to score functions over their measurements, the outliers of $\mathbf{x}_t$ using the median absolute deviation about the median [31], or a weighted sum over the current contextual values.

In the case we adopt a MVA for missing values imputation before delivering context to the System, we obtain analogous performance of all mechanisms. Tables 6 and 7 show the impact of the adoption of the ES-MVA on the fusion errors for both fusion operators using all approaches. Obviously, by adopting a MVA, we obtain lower fusion errors since the missing values are replaced with the most plausible enough thus, statistically reducing the error. Even in this case, ODA outperforms IDA significantly. This is due to the fact that the ODA takes into account all information (i.e., the series M_t) before proceeding with an optimal decision whether to stop at time t or continue and take the next observation. Recall that the highest possible expected context quality reward is obtained by the stopping rule stated in Theorem 2. This justifies the capability of our mechanism to deliver context of high quality with relatively low delay. The TDA assumes low fusion error but with very high delay compared with the ODA and, obviously, IDA. Overall, in both scenarios (by either adopting MVA algorithms or not) the ODA is deemed as an appropriate mechanism for near real-time analytics assuring high quality of delivered context, thus, improving the quality of MVAs inducing a tolerable delay.

6 Conclusions

We introduce a quality-optimized mechanism for delaying context delivery to predictive analytics engines in hope of receiving context of higher quality in data streams, thus eliminating possible biases in knowledge extraction and in decision making. The idea behind this mechanism is to avoid immediately delivering context by introducing a certain controlled delay. The proposed mechanism, based on the principles of optimal stopping theory, proceeds with an optimal stopping rule for delivering context taking into consideration the observation cost and the statistics of the quality indicators seen so far. An analytical stochastic optimization model is proposed and, through experimental evaluation and comparative assessment with a threshold-based and immediate delivery approach, our mechanism is deemed appropriate for adoption especially when the received context is (stochastically) of low quality and the observation cost is not negligible. In our future agenda we study the analysis and development of a mechanism in which the decision time for context delivery is contained within a finite time interval which is application specific.

References

1. Abbott D (2014) Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst, (1 ed.). Wiley Publishing
2. Awang A et al. (2007) RIMBAMON: A forest monitoring system using wireless sensor networks. In: ICIAS, pp 1101–1106
3. Zervas E et al. (2011) Multisensor data fusion for fire detection. Inform Fusion, Elsevier 12(3):1566–2535
4. Nittel S (2009) A Survey of geosensor networks: Advances in dynamic environmental monitoring. Sensors 9:5664–5678
5. Xu G et al. (2014) Applications Of wireless sensor networks in marine environment monitoring: a survey. Sensors 14(9):16932–16954
6. Su X et al. (2011) Using classifier-based nominal imputation to improve machine learning. In: 15th PAKDD, Part I, LNAI 6634, pp 124–135
7. Farhangfar A et al. (2008) Impact of imputation of missing values on classification error for discrete data. Pattern Recogn 41(12):3692–3705
8. Enders CK (2010) Applied Missing data analysis. Guilford Press, NY
9. Anagnostopoulos C, Triantafillou P (2014) Scaling out big data missing value imputations: pythia vs. godzilla. In: 20th ACM SIGKDD (KDD '14), pp 651–660
10. Hall DL, McMullen SAH (2004) Mathematical techniques in multisensor data fusion, Second. Artech House, Norwood
11. Das S (2008) High-Level Data fusion. Artech House Publishers, Norwood
12. Bettini C, Brdiczka O, Henricksen K, Indulska J, Nicklas D, Ranganathan A, Riboni D (2010) A survey of context modelling and reasoning techniques. Pervasive Mob Comput 6(2):161–180
13. Jong-yi H, Eui-ho S, Sung-Jin K (2009) Context-aware systems: A literature review and classification. Expert Syst Appl 36(4):8509–8522
14. Henricksen K, Indulska J (2006) Developing context-aware pervasive computing applications: Models and approach. Pervasive Mob Comput 2(1):37–64
15. Ye J, Dobson S, McKeever S (2012) Situation identification techniques in pervasive computing: A review. Pervasive Mob Comput 8(1):36–66
16. Anagnostopoulos C, Ntarladimas Y, Hadjiefthymiades S (2007) Situational computing: An innovative architecture with imprecise reasoning. J Syst Softw 80(12):1993–2014
17. Anagnostopoulos C, Hadjiefthymiades S (2008) Enhancing situation-aware systems through imprecise reasoning. IEEE Trans Mob Comput 7(10):1153–1168
18. Kolomvatsos K, Anagnostopoulos C, Hadjiefthymiades S (2015) A Time optimized scheme for top-k list maintenance over incomplete data streams. Inform Sci 311, C:59–73

19. Kolomvatsos K, Anagnostopoulos C, Hadjiefthymiades S (2015) An efficient time optimized scheme for progressive analytics in big data. Big Data Res 2(4):155–165
20. Eidson GW et al. (2009) The South carolina digital Watershed: end-to-end support for realtime management of water resources, Proc. 4th Intl. Symposium on Innovations and Real-time Applications of Distributed Sensor Networks (IRADSN 09), 2010, USA
21. Xia HB et al. (2009) Design of water environment data monitoring node based on ZigBee technology. Proc. Intl. Conference on Computational Intelligence and Software Engineering (CiSE 09), 1–4
22. Nguyen N et al. (2010) A Real-time control using wireless sensor network for intelligent energy management system in buildings. Proc. IEEE Worskshop on Environmental Energy and Structural Monitoring Systems (EESMS 10), 87–92
23. Oliveira LM, Rodrigues JJ (2011) Wireless Sensor networks: a survey on environmental monitoring. J Commun 6(2):143–151
24. Kim J.-J. et al. (2010) Wireless monitoring of indoor air quality by a sensor network. Indoor Built Environ 19(1):145–150
25. Kelley K et al. (2012) On effect size. Psychol Methods 17(2):137–152
26. Little R, Rubin D (2002) Statistical Analysis with Missing Data, Wiley Series in Probability and Statistics
27. Peskir G, Shiryaev A (2006) Optimal Stopping and Free-Boundary problems, Ed. 1, Lectures in Mathematics, ETH Zuerich, Birkhauser Basel
28. Shiryaev A (2007) Optimal stopping rules, series: Stochastic modelling and applied probability, vol. 8 springer
29. Daskalakis C et al. (2012) Learning poisson binomial distributions. In: 44th ACM STOC '12, pp 709–728
30. Tomas C (2006) Exponential smoothing for irregular data. Appl Math 51(6):597–604
31. Rousseeuw PJ, Croux C (1993) Alternatives to the Median Absolute Deviation. J Am Stat Assoc 88(424)
32. Vergara A, Vembu S, Ayhan T, Ryan MA, Homer ML, Huerta R (2012) Chemical gas sensor drift compensation using classifier ensembles. Sensors Actuators B Chem 166:320–329
33. Rodriguez-Lujan I, Fonollosa J, Vergara A, Homer M, Huerta R (2014) On the calibration of sensor arrays for pattern recognition using the minimal number of experiments. Chemometr Intell Lab Syst 130:123–134
34. Anagnostopoulos C, Kolomvatsos K, Hadjiefthymiades S (2015) Time-optimised user grouping in location based services. Comput Netw, Elsevier 81:220–244
35. Kolomvatsos K, Anagnostopoulos C, Hadjiefthymiades S (2014) An efficient recommendation system based on the optimal stopping theory. Expert Syst Appl, Elsevier 41(15):6796–6806
36. Anagnostopoulos C, Hadjiefthymiades S (2014) Intelligent trajectory classification for improved movement prediction. IEEE Trans Syst Man Cybern Syst Hum 44(10):1301–1314
37. Anagnostopoulos C (2014) Time-optimized contextual information forwarding in mobile sensor networks. J Parallel Distrib Comput, Elsevier 74(5):2317–2332
38. Anagnostopoulos C, Hadjiefthymiades S (2013) Multivariate context collection in mobile sensor networks. Comput Netw, Elsevier 57(6):1394–1407
39. Anagnostopoulos C, Hadjiefthymiades S, Zervas E (2013) Optimal stopping of the context collection process in mobile sensor networks. In: IEEE 24Th international symposium on personal, indoor and mobile radio communications (PIMRC), london, UK, pp 8–11
40. Delakouridis C, Anagnostopoulos C (2013) On enhancement of 'share the secret' scheme for location privacy. In: 9th International Workshop on Security and Trust Management (STM 2013), England, UK, pp 09–13