Proceedings of the Federated Conference on Computer Science and Information Systems pp. 825–834
DOI: 10.15439/2017F121
ISSN 2300-5963 ACSIS, Vol. 11

An Incremental Evidential Conflict Resolution Method for Data stream Fusion In IoT

Walid Cherifi
Faculty of Cybernetics,
Military University of Technology,
Warsaw, Poland.
Email: walid.cherifi@wat.edu.pl

Bolesław Szafrański
Faculty of Cybernetics,
Military University of Technology,
Warsaw, Poland.
Email: b.szafranski@milstar.pl

***Abstract*—During the last decade, several Internet of Things (IoT) applications has been developed to facilitate machine-to-human and machine-to-machine communication with the physical world by integrating both digital and physical entities through the internet. However, multiple important challenges need to be addressed in order to take the full advantage of these applications. One of the most important of these challenges concerns the management of IoT data, practically the data generated in dynamic and volatile environments and then provided in the form of streaming datasets. To enable reliable IoT applications in such scenario, it is crucial to develop methods that are able to automatically resolve any possible data conflict between diverse information sources in the case where the data is coming in a streaming fashion. In this paper, an incremental evidential conflict resolution method (I-ECRM) that is able to overcome this problem is introduced. The efficiency and effectiveness of the proposed method have been tested and evaluated through extensive experiments on synthetic datasets. The obtained results have shown that our method achieves a nice performance over different tradeoffs dimensions.**

I. Introduction

ACTIVE research, industry and standardization efforts in the field of next-generation networking are pushing towards a smart connected world where everyday objects will be dotted with the ability to sense, act and exchange information about their surroundings[1, 2]. Combined with today's Internet infrastructure, these objects can make a huge difference in our way of life. Thus, the number of expected applications is only bounded by imagination with applications on smart grid, smart cities, smart homes, smart health, industrial automation, and connected cars to name a few [3].

This new trend is commonly referred to as the future Internet architecture or simply the Internet of Things (IoT) [3]. This vision is starting to gain widespread adoption in today's world by building upon advances in a multitude of fields including micro-electromechanical systems, advances in wired and wireless communication technologies, networking, machine learning and big data. Many challenges are however being tackled and/or need to be addressed in order to unleash the full potential of the IoT applications. One of the most fundamental of these challenges is related to the quality of the generated data by the information sources (also known as things). In fact, due to the variety of the reliability level of the information sources, different sources can provide different contradictory information about the same real-world object, and thus a conflict between the sources' provided information may occur. In this situation, the collected information pieces about the same real world object need to be corrected according to their corresponding source reliability level and then fused in order to reduce uncertainty and obtain a more coherent, integrated information.

In general, the value of sources reliability degrees can be either obtained from external sources such as human experts, learned by using training datasets or constructed as a function of general agreement and corroboration between various sources. In this paper, we consider the case when no training dataset is available to assess the quality of the information sources. Thus, it is quite challenging to ascertain the reliability of each information source from the massive amounts of unlabeled data without knowing whether their provided information pieces are correct or wrong. Therefore, one of the main questions exposed in this paper is how to develop an efficient and effective unsupervised method that can both learn the sources reliability degree and determine the credibility degree of each provided information pieces without relying on manual user interaction, master data, or training dataset.

With the great evolution of computers technology, low-cost wireless sensor devices, Web technologies, and their recent multiple applications provide access to new types of data, which were not taken into account by the traditional processing applications. Two particularly interesting features of such data sets include their large volume and high velocity [4]. In several IoT applications, the amount of everyday generated data has grown exponentially during the last few years. This means that it is impossible to store and manipulate all that data since even a large scale algorithms exceed the processing capacity of the current single computing systems. Furthermore, the batch unsupervised data processing methods cannot handle its complex structure and size, fulfill very strict constraints as even simple computational operations are too costly. On the other hand, this could be seen as an opportunity to try to design and develop new methods that are able to deal with this new types of data complexity.

The above-mentioned requirements and challenges are particularly noticeable in emerging online data-intensive IoT

applications, on which data are being continuously generated at a high speed and/or large volume in a streaming format [5]. Sensor networks, weather forecast, traffic management, stock price prediction, or social media information analysis are just a few representatives of such applications where multiple data sources, such as sensors, human crowd, as well as web services, working in dynamic environments generate high rate data stream. Compared to static environments, streaming data sets arrive at a great speed and their processing algorithms have to meet tight computational requirements including limited memory usage, short processing time, and an online scan of incoming sets. Thus, the batch unsupervised evidential conflict resolution method cannot be used in such a scenario, as this technique is based on cost-effective iterative updates of the sources reliability degrees and information credibility values, which requires the totality of the data for the processing. Therefore, it is vital to develop efficient and effective techniques for data conflict resolution in the data streams scenario.

In this paper, we tackle this challenging scenario of conflict resolution problem. This scenario concerns the situation where the collected information pieces from the different sources arrive in a streaming fashion, i.e. the sources' provided information pieces are continuously collected by the fusion system in sequential chunks over a long period of time. In the light of this challenge, we propose an I-ECRM. This incremental method is able to resolve any probable conflict among the information sources and it can update the estimated evidential source reliability mass functions simultaneously in the case where the collected information is arriving in a streaming way.

The proposed method is based on the belief functions theory [6, 7]. This mathematical theory has been recently recognized to be one of the most effective tools to encode and manage information imperfection that is abundant in IoT applications [8]. This is due to its remarkable ability to represent and manipulate various types of imperfection (incomplete, imprecise, uncertain, or a combination of them). In particular, the belief function theory has an appealing tool that is able to combine multiple imperfect information pieces, and thus aiming at reducing uncertainty and obtaining more coherent, integrated information.

The rest of this paper is organized as follows. First, we briefly introduce the basic notions of the belief functions theory. After that, we motivate in Section 3 the need for a new incremental method for the evidential conflict resolution problem in the context of data streams. We then formalize the problem in Section 4. In Section 5, we introduce our I-ECRM. Next, we provide in Section 6 preliminary simulation results about the effectiveness and efficiency of the proposed incremental method via experimental evaluation over synthetic datasets. Finally, we conclude the paper Section 7.

II. Belief functions theory

The belief functions theory, also known as Dempster Shafer theory or evidence theory, is considered as one of the most widespread mathematical frameworks for data fusion. It was first introduced by Dempster in the 1960s [6] and later developed and improved by Shafer in the 1970s [7]. Some more recent advances in this theory were introduced later in the Transferable Belief Model (TBM) proposed by Smets [9]. The belief functions theory is also considered as a generalization of probability theory [10]. It provides an attractive, powerful and efficient mathematical framework to encode and aggregate a wide spectrum of imperfect information.

A. Basic notations

In the framework of belief functions, a problem domain is represented by a finite non empty set $\Theta = \{H_1, H_2, ..., H_N\}$ of N mutually exclusive and exhaustive hypotheses (events) representing the possible solutions of the considered task that we attend to determine its real value H. 2^Θ represents the power set composed of all the possible subsets of Θ. The basic belief assignment (*bba*), also known as mass function, is a function m mapping from 2^Θ to $[0, 1]$ and verifies the following conditions:

$$\begin{cases} m(\emptyset) = 0 \\ m(A) \geq 0 \\ \sum_{A \in 2^\Theta}^{m(A)=1} \end{cases} , \forall A \in 2^\Theta \tag{1}$$

$m(A)$ is the support degree that is assigned exactly to a proposition A and to no smaller subset. The mass functions m assigned to all the subsets of Θ are summed to unity and there is no belief left to the empty set. A mass function assigned exactly to Θ is referred to as the degree of global ignorance, denoted by $m(\Theta)$, and a mass function assigned exactly to a smaller subset of Θ except for any singleton proposition or Θ is referred to as the degree of local ignorance. If there is no local or global ignorance, a mass function will reduce to a classical probability function.

Besides the mass function, there are two other important functions to encode pieces of evidence: the belief function Bel and the plausibility function Pl [7]. These functions represent differently the same piece of information as the mass function. They are especially used to facilitate the manipulation and reasoning within the framework of belief functions. They are formally defined as follows:

$$\begin{cases} Bel : 2^\Theta \rightarrow [0, 1] \\ \quad A \mapsto \sum_{\substack{B \in 2^\Theta \\ B \subseteq A}} m(B) \end{cases} \tag{2}$$

$$\begin{cases} Pl : 2^\Theta \rightarrow [0, 1] \\ \quad A \mapsto \sum_{\substack{B \in 2^\Theta \\ A \cap B \neq \emptyset}} m(B) \end{cases} \tag{3}$$

$Bel(A)$ represents all masses assigned exactly to A and its smaller subsets, and $Pl(A)$ represents all possible masses that could be assigned to A and its smaller subsets. Note that $Bel(A)$ and $Pl(A)$ can be interpreted as the lower and upper bounds of the real probability $P(A)$.

B. Discounting operation

In several real-world situations, the information sources are not considered equally fully reliable. In this case, it is reasonable to discount each unreliable source s by a reliability factor $\alpha \in [0,1]$. Following the classical discounting method [7], a new discounted mass function m^α is obtained from the initial mass function m provided by the partially reliable source s as follows:

$$\begin{cases} m^\alpha(A) = \alpha \times m(A) & for A \neq \Theta \\ m^\alpha(\Theta) = (1-\alpha) + \alpha \times m(\Theta) \end{cases} \quad (4)$$

The discounting operation is mostly applied to model a situation where a source s delivers a mass function m, and the reliability of s is quantified by α. If the information source s is totally reliable (i.e. $\alpha = 1$), then m is left unchanged and it is considered as an acceptable piece of evidence. On the other hand, if the source s is completely unreliable, the mass function m is converted into the vacuous mass function (i.e. $m^\alpha(\Theta) = 1$), and thus this piece of evidence cannot be taken into consideration. In practice, the discounting operation can be used efficiently if one has an accurate estimation of the reliability value of the considered information source.

C. Combination of mass functions

The kernel of belief functions theory is Dempster's rule of combination that was originally adopted as the sole . this rule is the normalized conjunctive operation which aims to aggregate various mass functions from multiple independent information sources defined within the same frame of discernment. Given two mass functions m_1 and m_2 derived from two independent information sources s_1 and s_2, the combined mass function by Dempster' rule, denoted by $m_{1\oplus 2}(A) = m_1(A) \oplus m_2(A)$, is defined by the following equation:

$$m_{1\oplus 2}(A) = \begin{cases} \frac{\sum_{B,C\in 2^\Theta \; B\cap C=A} m_1(B)*m_2(C)}{1-\sum_{B,C\in 2^\Theta \; B\cap C=\emptyset} m_1(B)*m_2(C)} & A \in 2^\Theta, A \neq \emptyset \\ 0 & A = \emptyset \end{cases} \quad (5)$$

where the denominator represents the conflict coefficient, reflecting the degree of conflict between the two mass functions m_1 and m_2.

It is worth noting that this rule is widely used by the belief functions' community. This is due to its interesting mathematical properties. Indeed, Dempster's rule of combination is inherently commutative and associative, meaning that it can be used to aggregate several pieces of information in any order without changing the final results. This fact makes Dempster's rule very attractive from an engineering implementation perspective. In addition to these two properties, Dempster's rule is Non-idempotent i.e. the combination of two similar independent mass functions gives generally another more precise combined mass function. This is due to the fact that aggregating these two independent mass function may increase the total amount of information. Moreover, the vacuous mass function, that support the total ignorance, can be easily proved to be the neutral element for Dempster's rule for any mass function m defined over a frame of discernment Θ. This property is reasonable since the total ignorant evidence should not affect the fusion outcome since it doesn't provide any useful information that can be valuable to make a difference between the components of the power set 2^Θ.

D. Decision making

In addition to the combination operation, one of the main goal of using the belief functions theory is to make preeminent decisions by selecting the hypothesis that best fits the solution of the fusion problem under consideration. Therefore, the ultimate step in this framework is to make a decision about the studied task based on the reasoning results.

In order to make a reasonable decision, it is usually preferable to use a well-defined probability function. Probabilistic transformation is a great tool to map mass functions to probabilities. A classical transformation is the pignistic transformation [9], defined formally as follows:

$$BetP(A) = \sum_{B\subseteq\Theta, A\cap B\neq\emptyset} \frac{|A\cap B|}{|B|} m(B) \quad (6)$$

where $|B|$ is the number of elements in subset B. $BetP$ transfers uniformly the positive mass of each nonspecific element onto the singletons involved in that element according to the cardinal number of the proposition. Once the pignistic probability $BetP$ is computed, the decision can be made based on selecting the hypothesis $\hat{H}_j$ with the largest pignistic probability.

III. Data stream fusion in IoT

A streaming datasets can be considered as a set of potentially unlimited, ordered sequence of information pieces that are continuously coming at a fast speed, in such a way that it is impossible to permanently store and keep the entire information in memory or an external data repository [11]. In general, data streams have the following important properties:

- Data streams are sequences of information pieces, ordered by arrival time or another ordered property which can be, for instance, the generation time. This fact makes information pieces in data streams arrive for processing over time instead of being available a priori.
- Since data streams are produced continually and have unlimited or at least unknown length. Thus, their volume is considered as extremely huge.
- The arriving rate of data streams is very high with respect to the processing power of the fusion system.
- The qualitative behavior of the sources providing streaming datasets are susceptibility to change, and hence the quality of the provided information pieces may change over time.

Due to the above properties, processing methods that deal with streaming dataset should differ from the batch methods that need to process the whole complete dataset at once.

Table I
DIFFERENCES BETWEEN BATCH AND STREAM DATA PROCESSING METHODS [12].

	Batch	Stream
Number of passes	Multiple	Single
Processing time	Unlimited	Restricted
Memory usage	Unlimited	Restricted
Type of result	Accurate	Approximate
Distributed	No	Yes

The main dissimilarities include the sequential nature of the arriving information pieces, immense volumes, processing speed constraints, and the fact that the information pieces in the streaming dataset can generally be accessed only one time compared with the batch methods, where multiple access to the complete static dataset is possible. A summary of the differences between batch and stream data processing is presented in Table I.

In [13], the Unsupervised Evidential Conflict Resolution Method (U-ECRM) was developed to resolve the evidential conflict among the diverse sources by simultaneously estimating the evidential source reliability mass functions and determining the correct value of each considered objects. Unfortunately, this method cannot be directly applied to data streams due to the fact that this iterative method was specially designed to deal with static datasets. This fact makes the unsupervised evidential conflict resolution method do multiple passes through the entire dataset in order to resolve the conflict. Thus, this batch evidential method is impractical in the case where the dataset is in the form of a continuous flow of data streams. More importantly, the behavior of the information source can change over time. Thus, this evolving behavior needs to be captured and the evidential source reliability values have to be adjusted according to these changes. Furthermore, the method needs to take into account the problem of resource allocation when dealing with unbounded streaming datasets, which is mainly due to the massive volume and rapid speed of data streams. Accordingly, how to achieve greatest results under different resource constraints becomes a challenging task. The principal goal of this task is to decrease the resource allocation as compared to the batch iterative method and maximize the effectiveness of the method's outputs.

As a consequence, the applications that need to process data streams require a novel method that can do intelligent data processing and real-time analysis of the massive quantity of the generated streaming datasets in reasonable processing time and restricted memory space.

IV. PROBLEM FORMULATION

Suppose we have a set of N sources $S = \{s_1, s_2, s_3, \ \dots \ , s_N\}$ where the reliability level of each information source s_i is encoded as an evidential source reliability mass function m_i^{Θ} defined over the frame of discernment $\Theta = \{T, D, R\}$. Here T means that the source is trustworthy, D means that the source is defective and R means that the source is Random.

Each information source s_i provides information pieces in the form of mass functions $m_{i,j}^{\Omega,T=t}$. These pieces of information are continuously delivered in a streaming way i.e. the information pieces arrive in a sequential sets of information $D = \{D^{T=0}, D^{T=1}, ..., D^{T=t}, ...\}$, where each set $D^{T=t}$ contains a number of the sources' delivered information pieces $D^{T=t} = \left\{m_{i\in\{1,2,...,N\},j\in\{1,2,...,M^t\}}^{\Omega,T=t}\right\}$ about the actual values of a specific set of objects $O^{T=t} = \{o_1^t, o_2^t, o_3^t, \ \dots \ , o_{M^t}^t\}$ where each variable $o_j^t \in O^t$ can takes its unique true value $\hat{H}_j^t$ from the exhaustive and mutually exclusive frame of discernment $\Omega_j^t = \left\{H_{1,j}^t, H_{2,j}^t, H_{3,j}^t, ..., H_{K_j,j}^t\right\}$. It is worth noting that different sets of objects in different time t can contain different objects. In other words, the two objects o_1^{t-1} and o_1^t may not represent the same object. This meaning can be the same as if we consider the same object o_1 but its correct value change over time. For instance, if we suppose that the object o_1 represents the weather prediction of a specific city, the actual value of this object is different and independent from one day to another.

Due to several reasons, the information pieces that are provided by different sources about the same object can be conflicting. As a consequence, the main objective in this paper is to find a robust solution to this problem. This can be done by designing a method that is able to resolve the probable conflict between the information sources by determining the correct value of each object o_j^t in a specific time t. Moreover, this method should resolve the conflict and determine the correct value of each object with a single scan of the streaming dataset, short processing time, and use a limited memory space. Furthermore, the method should capture any changes in the behavior of the information sources, and thus adjusting the evidential source reliability mass function of each source according to its new state.

To achieve this objective, we adjust, in this paper, our proposed U-ECRM [13] so that the evidential source reliability mass functions and the correct values determination can be learned incrementally. This incremental method can also be used in the case of datasets with a gigantic volume that can only permit one single sequential pass through the whole datasets.

In fact, incremental methods have been used by several researcher to deal with computational problems that need to process streaming datasets [11, 14]. This kind of methods aims at analyzing and processing the newly arriving information pieces sequentially in such a way that the obtained results are as accurate, or approximately as accurate, as a traditional batch method that uses the entire dataset at once. A well-developed incremental method that is able to deal with data streaming scenarios should have the following important practical merits [15]

- **Use an incremental data access:** The method should process chunks of information pieces at a time, rather

than require the entire set of information pieces at the beginning of the processing.

- **Consider a single pass nature:** The method needs to handle and to process the newly arrived information pieces at once in the arriving order. This is due to the fact that the incoming information pieces cannot be kept permanently in memory, and thus the method should make only one pass through the available dataset.
- **Proceed in real time fashion:** the method should treat each information piece that belongs to the streaming datasets in real time fashion i.e. the newly arrived information pieces should be processed in an approximately short time once they are arrived. This processing time should be shorter than or at least equal to the data stream incoming rate, otherwise some important information pieces may be lost without treating and analyzing them.
- **Use bounded storage space:** since the streaming datasets is considered as an unlimited set of information pieces that are continuously arriving, it is impossible to store the entire streaming datasets in memory. As a consequence, the incremental method that deals with data streams should exploit a limited memory space to store a summary of the predicted model as well as the recently arrived information pieces.
- **Be ready to predict at any time:** the method should produce the best possible result at any point of time regardless the number of the past information pieces that are used to predict the model's parameters. Particularly, the results obtained from the incremental method should be as accurate as possible compared with the results achieved by the traditional batch methods that use the entire dataset up to a specific time t.

An incremental method with the above-stated capabilities can effectively process and deal with large streaming dataset without the need of re-executing the method from scratch after the arrival of a new set of information pieces. Such incremental methods can be built by scaling up traditional batch methods. This can be achieved by modifying the batch methods and tailored them to fit the data stream setting. In the next section, we introduce our proposed I-ECRM that is designed specifically to handle data streams or a static dataset with a massive volume.

V. The Proposed incremental evidential method

The key idea behind the proposed I-ECRM is to determine the correct value $\hat{H}_j^t$ of each considered object o_j^t in the time-stamp t based on the evidential source reliability mass functions $m_i^{\Theta,t-1}$ that are learned from the past interactions of the sources. Once done, the evidential source reliability mass functions $m_i^{\Theta,t}$ at time t should be updated according to the newly determined correct values without the need to re-execute the method on the complete dataset from scratch every time a new chunk of the streaming dataset is collected by the fusion system. Applying this idea in the U-ECRM [13], we modify the evidential source reliability mass functions update and the correct value mass functions determination steps to conduct I-ECRM.

Figure 1 presents the main concepts and the key idea in the architecture of the proposed I-ECRM. Specifically, a set of information sources continuously generate and provide chunks of streaming datasets to the fusion system. For each new arrived chunk of the streaming datasets at the time-stamp $T = t$, the incremental method first uses the evidential source reliability mass functions $m_{i\in\{1,2,\ldots,N\}}^{\Theta,T=t-1}$ learned from the previously processed chunks of information to correct the sources' provided information pieces. After that, the incremental method combines the sources' corrected information pieces by using Dempster's combination rule in order to obtain the correct value mass function $m_j^{\Omega,T=t}$ of each object o_j^t. Once done, the evidential source reliability mass functions $m_{i\in\{1,2,\ldots,N\}}^{\Theta,T=t}$ can be updated based on the difference between the computed correct value mass functions $m_{j\in\{1,2,\ldots,M^t\}}^{\Omega,T=t}$ and the sources' provided information pieces $m_{i\in\{1,2,\ldots,N\},j\in\{1,2,\ldots,M^t\}}^{\Omega,T=t}$.

The detailed description of the I-ECRM is summarized in Algorithm 1. This algorithm starts with an initialization step where it first uses Murphy aggregation method [16] to combine and fuse the information pieces of the first chunk $D^{T=0}$ of the streaming dataset. This can be done by computing the pignistic probability $BetP_j^{T=0}$ for each object $o_j^{T=0}$ and then selecting the hypothesis $\hat{H}_j^{T=0}$ that has the maximum pignistic probability.

$$\hat{H}_j^{T=0} = \underset{H_{l,j}^{T=0}\in\Omega_j^{T=0}}{arg\max} \left(BetP_j^{T=0}\left(H_{l,j}^{T=0}\right)\right) \tag{7}$$

Next, the initial evidential source reliability mass function $m_i^{\Theta,T=0}$ of each source can be estimated. To do so, the algorithm begins by evaluating the correctness degree of each of the mass functions that is provided by this source with regard to available information about the correct values. This evaluation step produces a set of evidence correctness mass functions $m_{i,j}^{\Psi,T=0}$, which encode how correct and relevant the source's information pieces are. The evidence correctness mass function $m_{i,j}^{\Psi,T=0}$ is defined over the frame of discernment $\Psi_{i,j} = \left\{C, \bar{C}\right\}$ where C encodes the hypothesis that the provided information $m_{i,j}^{\Omega,T=0}$ is correct, whereas $\bar{C}$ represents the hypothesis that the provided information $m_{i,j}^{\Omega}$ is incorrect. In order to compute $m_{i,j}^{\Psi,T=0}$, we use equation 8.

$$\begin{cases} m_{i,j}^{\Psi,T=0}(C) = \sum\limits_{B\in 2^{\Omega,T=0}} m_j^{\Omega,T=0}(B)\left(\sum\limits_{B\cap A=B} f(|A|)m_{i,j}^{\Omega,T=0}(A)\right) \\ m_{i,j}^{\Psi,T=0}(\bar{C}) = \sum\limits_{B\in 2^{\Omega,T=0}} m_j^{\Omega,T=0}(B)\left(\sum\limits_{B\cap A=\varnothing} m_{i,j}^{\Omega,T=0}(A)\right) \\ m_{i,j}^{\Psi,T=0}(C,\bar{C}) = 1 - \left(m_{i,j}^{\Psi,T=0}(C) + m_{i,j}^{\Psi,T=0}(\bar{C})\right) \end{cases} \tag{8}$$

where f is a function which distributes the imprecision of the source s_i between the support degree that that the given evidence is correct $m_{i,j}^{\Psi,T=0}(C)$ and the support degree that

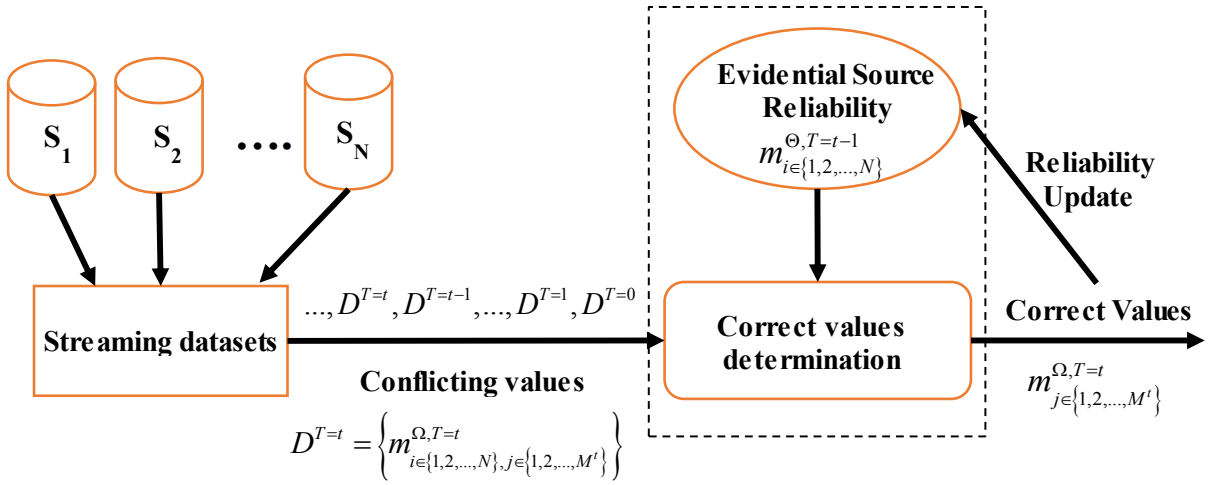

Figure 1. Conceptual view of the incremental evidential conflict resolution method for streaming datasets.

the provided information is irrelevant $m_{i,j}^{\Psi,T=0}(C,\bar{C})$. This function can be defined as follows:

$$f(|A|) = \frac{|\Omega_j| - |A|}{|\Omega_j| - 1} \tag{9}$$

Function f is, in reality, based on the uniform distribution of the correct identification of wrong hypotheses. In fact, one can reason about the wrong hypothesis that was correctly mentioned. Indeed, if source s_i supports proposition A i.e. the actual value $\hat{H}_{l,j}$ belongs to subset A, this can also mean that source s_i claims that the complement set of A does not contain the correct value. In other words, the piece of evidence provided by source s_i was (in somehow) correct concerning the identification of some wrong hypotheses. Therefore, one can give a proportion of $m_{i,j}^{\Omega,T=0}(A)$ to the mass function supporting the correctness of the provided information i.e. $m_{i,j}^{\Psi,T=0}(C)$. Whereas the rest of the proportion should be allocated to proposition $\{C,\bar{C}\}$, where the meaning is that the provided piece of evidence is irrelevant and does not contain any useful information.

After obtaining the evidence correctness mass functions of all objects, the total true positive $TP_i^{T=0}$ and the total false negative $FN_i^{T=0}$ of the source s_i are calculated by means of equation 10 and equation 11 respectively.

$$TP_i^{T=0} = \sum_{j=1}^{M} m_{i,j}^{\Psi,T=0}(C) \tag{10}$$

$$FN_i^{T=0} = \sum_{j=1}^{M} m_{i,j}^{\Psi,T=0}(\bar{C}) \tag{11}$$

After that, the algorithm uses the $TP_i^{T=0}$ and $FN_i^{T=0}$ along with an application-specific user-specified cautious parameter $C_{cautious}$ to estimate the reliability of the source by using equation 12 or equation 13 depending on the difference between $TP_i^{T=0}$ and $FN_i^{T=0}$.

- Case 1: $TP_i^{T=0} \geq FN_i^{T=0}$**:**

$$\begin{cases} m_i^{\Theta,T=0}(T) &= \frac{TP_i^{T=0} - FN_i^{T=0}}{TP_i^{T=0} + FN_i^{T=0} + C_{cautious}} \\ m_i^{\Theta,T=0}(D) &= 0 \\ m_i^{\Theta,T=0}(R) &= \frac{2FN_i^{T=0}}{TP_i^{T=0} + FN_i^{T=0} + C_{cautious}} \\ m_i^{\Theta,T=0}(T,D,R) &= \frac{C_{cautious}}{TP_i^{T=0} + FN_i^{T=0} + C_{cautious}} \end{cases} \tag{12}$$

- Case 2: $TP_i^{T=0} \leq FN_i^{T=0}$**:**

$$\begin{cases} m_i^{\Theta,T=0}(T) &= 0 \\ m_i^{\Theta,T=0}(D) &= \frac{FN_i^{T=0} - TP_i^{T=0}}{TP_i^{T=0} + FN_i^{T=0} + C_{cautious}} \\ m_i^{\Theta,T=0}(R) &= \frac{2TP_i^{T=0}}{TP_i^{T=0} + FN_i^{T=0} + C_{cautious}} \\ m_i^{\Theta,T=0}(T,D,R) &= \frac{C_{cautious}}{TP_i^{T=0} + FN_i^{T=0} + C_{cautious}} \end{cases} \tag{13}$$

At this point, the incremental method is ready to incrementally process the newly arriving streaming chunks. For each newly arrived chunk $D^{T=t}$ of the streaming dataset at time t, the algorithm uses the previously learned evidential source reliability mass functions $m_{i\in\{1,2,\ldots,N\}}^{\Theta,T=t-1}$ to compute the correct values mass function $m_j^{\Omega,T=t}$ for each object o_j of the chunk $D^{T=t}$. For each object $o_j^{T=t}$, the algorithm starts by correcting the provided mass functions $m_{i\in\{1,2,\ldots,N\},j}^{\Omega,T=t}$ according to their appropriate sources' evidential source reliability mass functions $m_{i\in\{1,2,\ldots,N\}}^{\Theta,T=t-1}$ by means of the evidential correction mechanism. This mechanism can be formally defined in equation 14.

Once correcting all the provided information pieces about the actual value of the considered object o_j, these corrected mass functions $m_{i\in\{1,2,\ldots,N\},j}^{\Omega*,T=t}$ can be aggregated by Dempster's combination rule so as to produce the combined correct

$$\begin{cases} m_{i,j}^{\Omega*,T=t}(A) = m_i^{\Theta,T=t-1}(T) m_{i,j}^{\Omega,T=t}(A) + m_i^{\Theta,T=t-1}(D) m_{i,j}^{\Omega,T=t}(\bar{A}) & \forall \mathrm{A} \in 2^{\Omega,T=t}/\Omega \\ m_{i,j}^{\Omega*,T=t}(\Omega) = m_{i,j}^{\Omega,T=t}(\Omega) + \left[m_i^{\Theta,T=t-1}(R) + m_i^{\Theta,T=t-1}(T,D,R)\right] \sum\limits_{\mathrm{A} \in 2^{\Omega}/\Omega} m_{i,j}^{\Omega,T=t}(A) \end{cases} \tag{14}$$

value mass function $m_j^{\Omega*,T=t}$. Immediately after that, the algorithm selects the correct values $\hat{H}_{j\in\{1,2,\ldots,M^t\}}$ by choosing the hypothesis $\hat{H}_j^{T=t}$ that has the maximum pignistic probability.

$$\hat{H}_j^{T=t} = \underset{H_{l,j}^{T=t} \in \Omega_j^{T=t}}{arg\max} \left(BetP_j^{T=t}\left(H_{l,j}^{T=t}\right)\right) \tag{15}$$

After that, the values of the evidential source reliability mass functions can be updated according to the estimated correct values $\hat{H}_{j\in\{1,2,\ldots,M^t\}}^{T=t}$ of the current chunk $D^{T=t}$. To do so, the algorithm begins by computing the true positive value $TP_i^{\Delta t}$ and the false negative value $FN_i^{\Delta t}$ of each source over the current streaming chunk $D^{T=t}$. These two important values can be obtained by means of equation 16 and equation 17 respectively.

$$TP_i{}^{\Delta t} = \sum_{j=1}^{M} m_{i,j}^{\Psi,\Delta t}(C) \tag{16}$$

$$FN_i{}^{\Delta t} = \sum_{j=1}^{M} m_{i,j}^{\Psi,\Delta t}\left(\bar{C}\right) \tag{17}$$

where $m_{i,j}^{\Psi,\Delta t}$ is the evidence correctness mass function of each provided information piece $m_{i,j}^{\Omega,T=t}$ with regard to the obtained correct value $\hat{H}_j^{T=t}$.

In order to control the effect of possible changing behaviors of the information sources, the I-ECRM uses a decay parameter $\lambda \in [0,1]$ that determines the impact of historical interaction on the current evidential source reliability mass function $m_i^{\Theta,T=t}$. Intuitively, the recent interactions of the sources $m_{i\in\{1,2,\ldots,N\},j\in\{1,2,\ldots,M^t\}}^{\Omega,T=t}$ should play a more important role in the estimation of $m_{i\in\{1,2,\ldots,N\}}^{\Theta,T=t}$ than the historical interaction when $T < t$. In other words, the key idea of the use of the decay parameter is to scale the past information about the behaviors of the sources by a constant factor λ, i.e. each time a new chunk of the streaming dataset is arrived, the past learned total true positive values $TP_{i\in\{1,2,\ldots,N\}}^{T=t-1}$ and the total false negative values $FN_{i\in\{1,2,\ldots,N\}}^{T=t-1}$ are scaled down by the factor λ. Qualitatively, this means that the smaller the decay parameter λ is, the less impact from historical interactions in the estimation of the current evidential reliability values and hence it will make the model respond quickly to any behavioral changes. As a result, the newly computed $TP_{i\in\{1,2,\ldots,N\}}^{T=t}$ and $FN_{i\in\{1,2,\ldots,N\}}^{T=t}$ can be obtained as follows:

$$\begin{cases} TP_{i\in\{1,2,\ldots,N\}}^{T=t} = \lambda \cdot TP_{i\in\{1,2,\ldots,N\}}^{T=t-1} + TP_{i\in\{1,2,\ldots,N\}}^{\Delta t} \\ FN_{i\in\{1,2,\ldots,N\}}^{T=t} = \lambda \cdot FN_{i\in\{1,2,\ldots,N\}}^{T=t-1} + FN_{i\in\{1,2,\ldots,N\}}^{\Delta t} \end{cases} \tag{18}$$

Once the $TP_i^{T=t}$ and $FN_i^{T=t}$ are computed, the method can estimate the $m_{i\in\{1,2,\ldots,N\}}^{\Theta,T=t}$ by means of equation 19 or equation 20 depending on the difference between the values of $TP_i^{T=t}$ and $FN_i^{T=t}$. These newly estimated evidential source reliability mass function $m_{i\in\{1,2,\ldots,N\}}^{\Theta,T=t}$ can be further used to resolve the conflict of the newly arriving chunk $D^{T=t+1}$ at time $T = t+1$.

- Case 1: $TP_i^{T=t} \geq FN_i^{T=t}$**:**

$$\begin{cases} m_i^{\Theta,T=t}(T) &= \frac{TP_i^{T=t} - FN_i^{T=t}}{TP_i^{T=t} + FN_i^{T=t} + C_{cautious}} \\ m_i^{\Theta,T=t}(D) &= 0 \\ m_i^{\Theta,T=t}(R) &= \frac{2FN_i^{T=t}}{TP_i^{T=t} + FN_i^{T=t} + C_{cautious}} \\ m_i^{\Theta,T=t}(T,D,R) &= \frac{C_{cautious}}{TP_i^{T=t} + FN_i^{T=t} + C_{cautious}} \end{cases} \tag{19}$$

- Case 2: $TP_i^{T=t} \leq FN_i^{T=t}$**:**

$$\begin{cases} m_i^{\Theta,T=t}(T) &= 0 \\ m_i^{\Theta,T=t}(D) &= \frac{FN_i^{T=t} - TP_i^{T=t}}{TP_i^{T=t} + FN_i^{T=t} + C_{cautious}} \\ m_i^{\Theta,T=t}(R) &= \frac{2TP_i^{T=t}}{TP_i^{T=t} + FN_i^{T=t} + C_{cautious}} \\ m_i^{\Theta,T=t}(T,D,R) &= \frac{C_{cautious}}{TP_i^{T=t} + FN_i^{T=t} + C_{cautious}} \end{cases} \tag{20}$$

We now show how the I-ECRM can effectively address the computational requirements of processing and dealing with data streams introduced in Section 6.2. First, the I-ECRM makes a single scan (one-pass) through the streaming datasets since it is obvious that the proposed incremental method process the provided information pieces only once. Second, the I-ECRM uses a limited memory space to process the whole data streams because it only exploits a size of memory space equivalent to the size of the evidential source reliability mass functions as well as only one chunk of the provided information pieces at any time t in the stream. Third, the I-ECRM processes the streaming datasets in short time since the algorithm computes and then reports the objects' correct values online, which is in effect much shorter than the computation time of the batch unsupervised evidential conflict resolution method. Finally, the proposed incremental method can capture and handle any changes in the behavior of the sources. This is ensured by the decay parameter which allows the fusion system to gradually forget about the sources' old interactions and mainly focus focus on the current interactions.

VI. Experimental evaluation

In this section, we report and analyze the initial experimental results of the proposed I-ECRM on some instances of synthetic datasets. The obtained experimental results demonstrate that our proposed incremental evidential method can achieve a good efficiency-effectiveness trade-off. We first introduce the overall experiment settings in subsection VI-A, and then we present and discuss the experimental results in subsection VI-B.

Algorithm 1: Incremental Evidential Conflict Resolution Method I-ECRM

Input : Streaming dataset $\left\{D^{T=0}, D^{T=1}, ..., D^{T=t}, ...\right\}$ where: $D^{T=t} = \left\{m^{\Omega,T=t}_{i\in\{1,2,...,N\},j\in\{1,2,...,M^t\}}\right\}$.
A cautious parameter $C_{cautious}$.
A decay parameter λ.

Output: The set of all $\hat{H}^{T=t}_{j\in\{1,2,...,M^t\}}$ representing the correct values of objects $o^{T=t}_{j\in\{1,2,...,M^t\}}$.

begin
  `// Init of parameter using` $D^{T=0}$**:**
  Compute $m^{\Omega,T=0}_{j\in\{1,2,...,M^0\}}$ *by means of Murphy method [16].*
  Find $\hat{H}^{T=0}_{j\in\{1,2,...,M^0\}}$ *by means eq 7.*
  Compute $m^{\Theta,T=0}_{i\in\{1,2,...,N\}}$ *by means of eq 12 or eq 13.*
  while *new streaming dataset* $D^{T=t>0}$ *is arriving* **do**
    `// Correct value mass function computation:`
    Compute $m^{\Omega,T=t}_{j\in\{1,2,...,M^t\}}$
    `// Correct values decision making:`
    Find $\hat{H}^{T=t}_{j\in\{1,2,...,M^t\}}$ *by means of eq 15.*
    `// Evidential source reliability updating`
    foreach *source* s_i *in the set of all sources* S **do**
      `// 1. Compute` $TP_i^{\Delta t}$ `and` $FN_i^{\Delta t}$**:**
      foreach *object* o_j^t *in the set of all objects* O^t **do**
        Compute $m^{\Psi,\Delta t}_{i,j}$ of $m^{\Omega,T=t}_{i,j}$ with regard to the categorical mass function $m_j^{\Omega,T=t}(\hat{H}_j) = 1$ by means of equation 8.
      end
      $TP_i^{\Delta t} = \sum_{j=1}^{M} m^{\Psi}_{i,j}(C)$
      $FN_i^{\Delta t} = \sum_{j=1}^{M} m^{\Psi}_{i,j}(\bar{C})$
      `// 2. Compute` $TP_i^{T=t}$ `and` $FN_i^{T=t}$**:**
      $TP_i^{T=t} = \lambda \cdot TP_i^{T=t-1} + TP_i^{\Delta t}$
      $FN_i^{T=t} = \lambda \cdot FN_i^{T=t-1} + FN_i^{\Delta t}$
      `// 3. Compute the reliability` $m_i^{\Theta,T=t}$**:**
      if $TP_i^{T=t} \geq FN_i^{T=t}$ **then**
        Estimate $m_i^{\Theta,T=t}$ using equation 19.
      else
        Estimate $m_i^{\Theta,T=t}$ using equation 20.
      end
    end
  end
end

A. Experimental setting

1) Datasets: In order to show the benefit of the I-ECRM over the unsupervised conflict resolution method, we use the synthetic dataset generator developed by Waguih et al. [17] to produce some instances of synthetic datasets. This dataset generator was developed in order to generate and simulate a wide range of real-world situations where the behaviors of the information sources can be controlled and configured in terms of a set of parameters such as coverage, reliability level, conflicting information, to name a few.

2) Methods in comparison: We evaluate the performance of the I-ECRM with regard to the batch unsupervised evidential conflict resolution method (U-ECRM) and the native voting method where the correct value is the one which is supported by the majority of the sources.

3) Evaluation metric: we use the following metrics to evaluate the performance of the proposed methods:

Precision rate: We use the precision rate to evaluate the effectiveness of the proposed methods. A highest precision rate implies a better and a more effective method.

CPU time: We use the CPU time to evaluate the time efficiency of the proposed evidential conflict resolution method. A shorter CPU time implies a faster and a more efficient method.

Space usage: We use the memory space occupation of the proposed methods to evaluate the space efficiency. A smaller memory space occupation implies a more space efficient method.

4) Environment: To ensure the implementation of our method, we have developed our incremental evidential conflict resolution Matlab R2010a. We have further conducted our experiments on PC 8GB RAM, Intel(R) core (TM) i2CPU 2.30GHz, and windows 10 installed.

B. Experimental results

We begin by setting the scale parameters of our considered scenario as follows: we set the number of sources to 60, and the number of possible values for each object to 4. We also select the uniform distribution for the distribution of the distinct values per object. In addition, we configure the source coverage to follow the exponential distributions. Furthermore, we select 80-pessimistic distributions for the ground truth distribution. As for the number of objects, we change this parameter from 1,000 to 10,000 objects with increments of 1000 objects. The key idea behind varying this parameter is to evaluate the effect of changing the number of object in the effectiveness and the efficiency of the proposed conflict resolution methods.

Based on the above setting, we generate 20 synthetic datasets for each experiment of a specific number of sources. In order to reduce the randomness of the dataset generation process, the evaluation metrics of each considered conflict resolution method is computed as the average of these 20 generated datasets included in the dataset of the same number of objects.

To simulate the scenario of streaming dataset, we consider that every time t a chunk containing the information pieces

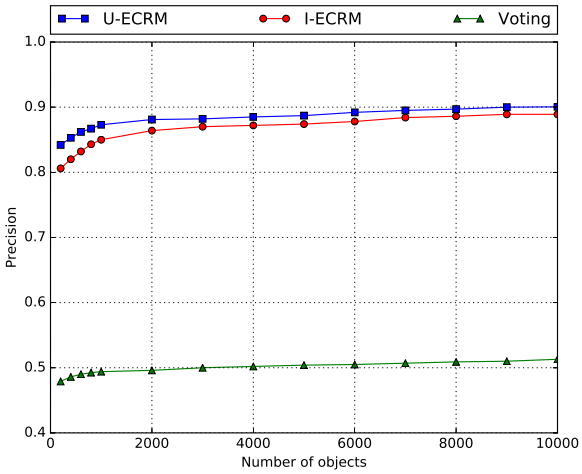

Figure 2. The evaluation of the precision of the considered method with regard to the number of objects.

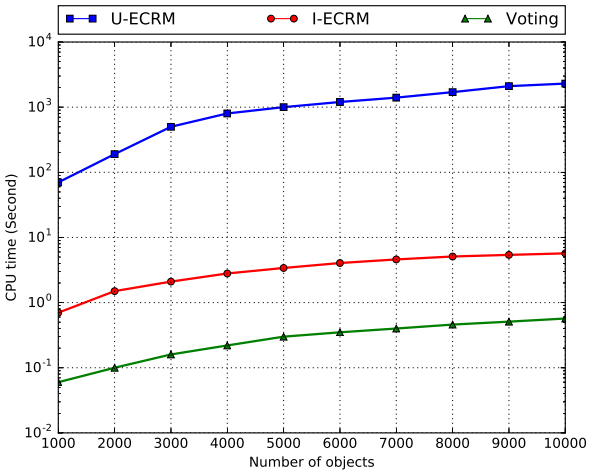

Figure 3. The evaluation of the processing CPU time efficiency of the considered conflict resolution methods with regard to the number of objects.

about 50 objects are arriving to the fusion system. In this case, we should process each time 50 objects by the considered conflict resolution methods.

We first start by comparing the effectiveness of the I-ECRM with regard to the batch unsupervised evidential conflict resolution method and the trivial voting method. Then, we provide the time and space efficiency analysis of the considered methods.

Effectiveness results: Figure 2 plots the precision of the considered conflict resolution methods on the synthetic dataset. As can be seen in Figure 2, the precision of the considered methods quickly increase in the beginning when more objects are involved. However, these precision values become on average approximately constant after the number of objects is greater than 1000. Also, it can be observed that the unsupervised evidential conflict resolution method is the most effective method, followed by the I-ECRM. This latter method performs only slightly worse than the former one. In the opposite, the voting method is the less effective method. This is due to the fact that this trivial method does not consider the reliability of the source while determining the correct value of each object.

Time efficiency results: Figure 3 plots the CPU time of the considered conflict resolution methods on the considered synthetic dataset. The results obtained from Figure 3 show that the voting method is the most time efficient, followed by the I-ECRM. When processing 10,000 objects, the voting and incremental methods take around 0.7 seconds and 5.5 seconds respectively. The unsupervised evidential conflict resolution method is the less time efficient as it needs to make several iterations over the entire datasets. Its CPU time increases quickly as more objects are involved, which exceeds 1,000 seconds when processing 6000 objects. Accordingly, the unsupervised evidential conflict resolution method is not appropriate for processing and analyzing streaming datasets or datasets with massive volumes.

It is worth mentioning that when new chunks of information pieces arrive over time, the voting and I-ECRM need only to process these new coming chunks. Therefore, Their CPU time relies only on the size of the chunk to be processed. On the other hand, the supervised evidential conflict resolution method needs to process the entire dataset each time a new chunk arrives. Thus, its CPU time depends on the size whole dataset.

Space efficiency results: Figure 4 plots the memory space used by the considered conflict resolution methods to process the synthetic dataset. As can be seen from Figure 4, the voting method has the lowest memory consumption, as it is a method that processes each time only one object and its corresponding provided information pieces. Thus the voting method is considered as the most space efficient. The second most space efficient method is the I-ECRM. This incremental method cache only the newly arrived chunk of information pieces each time. Moreover, it needs to cache additional information concerning the evidential source reliability mass functions (the model parameters). Finally, the worst space efficient method is the supervised evidential conflict resolution method which is the most space consuming. This is due to the fact that this method needs to cache the complete streaming dataset in memory (the old as well as the newly arrived streaming chunks).

VII. Conclusion

In this paper, we addressed the challenging problem of resolving information conflict in the case where the sources' provided information pieces are continuously arriving at the fusion system in the form of streaming datasets. This problem is very important because recent years have witnessed a huge range of online IoT applications that need to process data streams. To deal with this problem, we proposed and developed an incremental evidential conflict resolution method that is able to resolve the evidential conflict among sources by jointly and incrementally estimating the evidential source

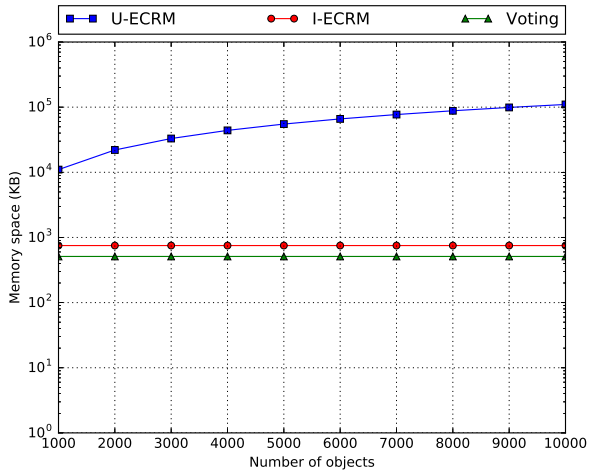

Figure 4. The evaluation of the processing memory space efficiency of the considered conflict resolution methods with regards to the number of the objects.

reliability mass function of each information source and discovering the correct value of each object among the set of all possible values. This incremental method works under the constraints of a single scan of the streaming data, real-time processing fashion, and a limited memory space usage. The proposed method was empirically evaluated by using synthetic datasets in order to verify its efficiency and effectiveness. The obtained results show that the proposed incremental evidential method has a nice efficiency-effectiveness trade-off.

BIBLIOGRAPHY

[1] L. Da Xu, W. He, and S. Li, "Internet of things in industries: A survey," *IEEE Transactions on industrial informatics*, vol. 10, no. 4, pp. 2233–2243, 2014. [Online]. Available: http://dx.doi.org/10.1109/TII.2014.2300753

[2] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (iot): A vision, architectural elements, and future directions," *Future generation computer systems*, vol. 29, no. 7, pp. 1645–1660, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.future.2013.01.010

[3] H. Sundmaeker, P. Guillemin, P. Friess, and S. Woelfflé, Eds., *Vision and Challenges for Realising the Internet of Things*. Luxembourg: Publications Office of the European Union, 2010. [Online]. Available: http://dx.doi.org/10.2759/26127

[4] M. Wang, C. Perera, P. P. Jayaraman, M. Zhang, P. Strazdins, R. Shyamsundar, and R. Ranjan, "City data fusion: Sensor data fusion in the internet of things," *International Journal of Distributed Systems and Technologies (IJDST)*, vol. 7, no. 1, pp. 15–36, 2016. [Online]. Available: http://dx.doi.org/10.4018/IJDST.2016010102

[5] Y. Qin, Q. Z. Sheng, N. J. Falkner, S. Dustdar, H. Wang, and A. V. Vasilakos, "When things matter: A survey on data-centric internet of things," *Journal of Network and Computer Applications*, vol. 64, pp. 137–153, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.jnca.2015.12.016

[6] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *The annals of mathematical statistics*, pp. 325–339, 1967.

[7] G. Shafer *et al.*, *A mathematical theory of evidence*. Princeton University Press, 1976, vol. 1.

[8] A. Bossae and B. Solaiman, *Information Fusion and Analytics for Big Data and Iot*. Norwood, MA, USA: Artech House, Inc., 2016.

[9] P. Smets, "Decision making in the tbm: the necessity of the pignistic transformation," *International Journal of Approximate Reasoning*, vol. 38, no. 2, pp. 133–147, 2005. [Online]. Available: https://doi.org/10.1016/j.ijar.2004.05.003

[10] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion*, vol. 14, no. 1, pp. 28–44, 2013. [Online]. Available: https://doi.org/10.1016/j.inffus.2011.08.001

[11] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams: A review," *SIGMOD Rec.*, vol. 34, no. 2, pp. 18–26, Jun. 2005. [Online]. Available: http://doi.acm.org/10.1145/1083784.1083789

[12] J. Gama, *Knowledge discovery from data streams*. CRC Press, 2010.

[13] W. Cherifi and B. Szafrański, "An unsupervised evidential conflict resolution method for data fusion in iot," *Submitted to IoT-ECAW'17*.

[14] M. A. Maloof and R. S. Michalski, "Incremental learning with partial instance memory," *Artificial intelligence*, vol. 154, no. 1-2, pp. 95–126, 2004. [Online]. Available: http://dx.doi.org/10.1016/j.artint.2003.04.001

[15] A. Bifet and R. Kirkby, "Data stream mining a practical approach."

[16] C. K. Murphy, "Combining belief functions when evidence conflicts," *Decision support systems*, vol. 29, no. 1, pp. 1–9, 2000. [Online]. Available: https://doi.org/10.1016/S0167-9236(99)00084-6

[17] D. A. Waguih and L. Berti-Equille, "Truth discovery algorithms: An experimental evaluation," *CoRR*, vol. abs/1409.6428, 2014. [Online]. Available: http://arxiv.org/abs/1409.6428