A semantic-based approach for Machine Learning data analysis

Agnese Pinto, Floriano Scioscia, Giuseppe Loseto, Michele Ruta, Eliana Bove, Eugenio Di Sciascio
Politecnico di Bari, Bari, Italy
Email: name.surname@poliba.it

***Abstract*—Pervasive applications and services are increasingly based on the intelligent interpretation of data gathered via heterogeneous sensors dipped in the environment. Classical Machine Learning (ML) techniques often do not go beyond a basic classification, lacking a meaningful representation of the detected events. This paper introduces a early proposal for a semantic-enhanced machine learning analysis on data of sensors streams, performing better even on resource-constrained pervasive smart objects. The framework merges an ontology-driven characterization of statistical data distributions with non-standard matchmaking services, enabling a fine-grained event detection by treating the typical classification problem of ML as a resource discovery.**

I. Introduction

The increasingly large amounts of data available through widespread diffusion of sensing and capturing devices dipped in everyday environments, strongly evidence the need for effective applications able to process them with the final goal to give a meaningful interpretation of retrieved information. In the last few years, along with the deep penetration of pervasive technologies, several data analysis techniques have been devised and used for that, producing undeniable benefits, but also revealing some weakness. Among such techniques, classical Machine Learning (ML) is one of the more largely adopted strategy. In spite of computational manageability and simple nature of ML approaches, is even more evident they lack of a meaningful, machine-understandable interpretation of detected information. This reveals a not-negligible limit in a possible exploitation of such theoretical frameworks in fully autonomic case studies and application scenarios.

This paper aims to early propose an enhancement to ML analysis on data streams, attempting to go beyond a trivial classification and associating semantically rich (compact) descriptions to real-world retrieved data. The proposed approach also exploits non-standard matchmaking services in [1] to obtain a logic-based characterization of statistical data distributions so enabling a fine-grained event detection: a typical Machine Learning classification problem is treated as an ontology-driven resource discovery. The proposal grounds on both ideas and technologies of distributed semantic-based systems [2], whose individuals (assertional knowledge) are physically tied to objects disseminated in a given environment, without centralized coordination. Each annotation refers to an ontology providing the conceptualization for the particular domain. This model is supported by an advanced matchmaking carried out using metadata stored in sensing and capturing devices dipped in a context, lacking fixed knowledge bases, and inference tasks distributed among computing devices which provide minimal computational capabilities. Stream reasoning techniques then provide the means to harness the flow of semantically annotated updates inferred from low-level data, in order to enable adaptive context-aware behaviors in a large array of applications.

The experimental campaign of the proposed framework is still ongoing and it is basically devoted to preliminary assess both feasibility and sustainability of ideas and algorithms, but also to better understand strategies for surmounting challenges deriving from any practical deployment of proposed ideas.

The remainder of the paper is as follows. In Section II, we provide some motivation for the proposal before analyzing literature in Section III. Section IV discusses the devised framework in detail, while the subsequent section reports on main challenges the proposal contains. Conclusion finally terminates the paper.

II. Motivation

Main motivation for this paper stems from our studies in pervasive computing field, and particularly in the Semantic Web of Things (SWoT) one [3], [2]. In such scenarios, information is gathered through micro-devices enriching everyday items or deployed in given environments and interconnected wirelessly. Basically, due to their small size, such *objects* have minimum processing capabilities, a small storage and low-throughput communication capabilities. They continuously produce raw data to be processed by advanced remote applications which leverage an intelligent interpretation of retrieved information. Classical ML techniques have been largely used for that, but their main weakness is in the lack of a meaningful representation of detected events.

Hence the main contribution of the proposed approach can be summarized as follows:

1. Semantically rich and compact descriptions are associated to real-world objects, retrieved data and monitored environments. Semantic-based automatic inferences could be exploited to infer implicit information starting from an explicit event and context detection. This is a peculiar feature of the proposed approach and enables novel classes of smart applications.
2. A general and scalable framework for ontology-driven data fusion allows each WSN node collecting raw data and

IEEE ICSC 2015, February 7-9, 2015, Anaheim, California, USA
978-1-4799-7935-6/15/$31.00 ©2015 IEEE

aggregating annotated information according to its storage and computational resources. Other data fusion approaches are based on statistics or control theory, not allowing explicit, machine-understandable knowledge representation.

3. Each micro-device can process locally and automatically retrieved information in order to describe itself and the context where it is located toward a range of applications. This enables pervasive knowledge-based systems with high degrees of autonomic capability not already allowed by typical WoT paradigms.

Thanks to this vision, innovative analysis methods could be applied to data extracted by inexpensive off-the-shelf pervasive micro-devices so providing useful results in event monitoring and recognition without requiring large computational resources: studies in machine learning techniques, algorithms and tools have enabled novel classes of data interpretation approaches, but the exploitation of logic-based and approximate discovery strategies leverage non-exact matching results to compensate possible faults in capturing activities and the unreliability of wireless communications.

III. Related Work

Many real world events show patterns, which can be identified and predicted from data continuously gathered *e.g.*, by Wireless Sensor Networks (WSNs) or smart objects [4]. Several machine learning methods are exploited for WSNs scenarios and real-world applications, including *decision tree classifiers* [5], [6], *bayesian algorithms* [7] and *support vector machines* [8].

The main limitation of the above-cited works is that classification yields trivial labels, lacking any structured information about the characteristics of the output. *Multi-label* variants of the above basic methods were devised to increase richness and accuracy of representation. In [9] five ML techniques were tested in a multi-label classification task for data fault detection in WSNs, whereas [10] proposed a multi-label classification algorithm applicable when label dependencies are structured as either a tree or a directed acyclic graph. In [11] a distributed classification approach was proposed to monitor physical environments via inexpensive wireless sensor devices forming a spanning tree. Similarly, an incremental online learning algorithm was proposed in [12] to determine the root causes of problems in IT services. The above multi-label approaches use trees or graphs to model class dependencies, while in the approach proposed here sensor data populate ontological knowledge bases, whose classes can have articulated relationships and high-level descriptions suitable for further high-level processing.

Other semantic-enhanced machine learning methods exist in literature. In [13], the authors proposed an approach using wireless sensor network data and ontologies to represent and infer knowledge about traffic conditions. In [14], an unsupervised model was used to classify Web Service data types in a large number of ontology classes by adopting an extended neural network. In both works, however, mining was exploited only to map data to a single class. Conversely, [15] proposed a *Semantic Decision Tree* to derive high-level annotations for ontology classes. In spite of that, no inference algorithm was exploited during learning steps.

IV. Proposed approach

The proposed approach keeps the typical workflow of data mining and machine learning: data collection and cleansing, model training, system predictions exploitation and evaluation. Nevertheless, semantic-based enhancements change the way each step is performed.

The workflow starts with raw **data** gathered *e.g.*, by sensors dipped in a given environment for m different measuring parameters, generally named *features*. In order to support semantic-based data annotation and interpretation, an **ontology** $\mathcal{T}$ models the domain conceptualization along properly defined patterns, and $\mathcal{T}$ is supposed acyclic and expressed in the moderately expressive $\mathcal{ALN}$ (Attributive Language with unqualified Number restrictions) language of the DL (Description Logic) family. This is required by the further adoption of non-standard inference services for semantic matchmaking. For each measuring parameter $\mathcal{T}$ will include a hierarchy of concepts (each one with its own properties), forming a partonomy of the top-most concept. In other words, each parameter will be represented via a classes/subclasses taxonomy featuring all significant configurations it can assume in the domain of interest. The depth of the hierarchy and the breadth of each level will be proportional to both resolution and range of sensing/capturing equipment, as well as to the needed degree of detail in data representation.

The goal of the following **training** step is to build a semantic annotation for each possible output class, connoting the observed event/phenomenon according to input data. The description is obtained by joining the above logical concepts in a conjunctive expression. The training phase works on a set S of n training samples, each with m features. The samples can be grouped from the streaming data by defining proper time windows. Let us suppose that w distinct outputs exist in the training set. Each feature value is mapped to the most specific corresponding concept in the reference ontology $\mathcal{T}$. Therefore the i-th sample $\forall\ i = 1, \ldots, n$ is composed by: (a) m concept components $C_{i,1}, \ldots, C_{i,m}$ (defined as in [16]) annotating its features; (b) an observed output O_i labeled with a class of the ontology. Samples are processed sequentially by Algorithm 1 in order to build the so-called *Training Matrix* $\mathcal{M}$ (the pseudocode uses a MATLAB-like notation for matrix access). $\mathcal{M}$ is a $(w+1) \times (k+1)$ matrix having all the different outputs on the first column, all the distinct concept components on the first row and, in each element, the number of occurrences of the column header concept component in the samples having the row header output. Initially, $\mathcal{M}$ is empty. For each sample, the `findConceptIndex` function finds the row index related to the observed output. If it does not exist, a new one will be appended to $\mathcal{M}$ with all concepts columns set to zero (lines 4–10). Subsequently, a similar function is used (line 13) to find the column indexes corresponding to the concept components of the sample S_i.

Algorithm 1 Training Matrix generation

Require:
– $\mathcal{L}$ Description Logic;
– acyclic TBox $\mathcal{T}$;
– training set $S = \{S_1, S_2, \ldots S_n\}$, with $S_i = (C_{i,1}, \ldots, C_{i,m}, O_i)\ \forall\ i = 1, \ldots, n$, where all $C_{i,j}$ and O_i are expressed in $\mathcal{L}$ and satisfiable in $\mathcal{T}$.
Ensure:
– $\mathcal{M} : (w+1) \times (k+1)$ matrix of occurrences of the concepts for each observed output, where k is the total number of distinct concepts appearing in S

$\mathcal{M} := 0$ // (1×1) matrix
$r := 1, c := 1$
for $i := 1$ to $|S|$ **do**
  $u_r :=$ *findConceptIndex*$(O_i, \mathcal{M}(:, 1))$
  if $u_r = null$ **then**
    append a row to $\mathcal{M}$
    $r := r + 1$
    $u_r := r$
    $\mathcal{M}(u_r, 1) := O_i$
    initialize $\mathcal{M}(u_r, 2 : c)$ to zeros
  end if
  for $j := 1$ to m **do**
    $u_c :=$ *findConceptIndex*$(C_{i,j}, \mathcal{M}(1, :))$
    if $u_c = null$ **then**
      append a column to $\mathcal{M}$
      $c := c + 1$
      $u_c := c$
      $\mathcal{M}(1, u_c) := C_{i,j}$
      initialize $\mathcal{M}(2 : r, u_c)$ to zeros
    end if
    $\mathcal{M}(u_r, u_c) = \mathcal{M}(u_r, u_c) + 1$ // update occurrences
  end for
end for
return $\mathcal{M}$

Also in this case, if a concept column is not defined in $\mathcal{M}$ a new one is created.

$\mathcal{M}$ gives a complete picture of the training set. Each output class can now be defined as conjunction of the concepts having greater-than-zero occurrences in the corresponding row. By doing so, however, even very rare concepts are included, which may be not significant in defining the class. Therefore it is useful to define a *significance threshold* T_s as the minimum number of samples where a particular concept must appear to be considered significant for the occurrence of a given output. $\mathcal{M}$'s structure suggests the possibility to define different thresholds for each output and for each feature, as

$$T_{s(i,j)} = \theta_{(i,j)} |S|$$

with $0 < \theta_{(i,j)} \leq 1 \quad \forall i, j$ being adaptive ratios computed through a cross-validation process on the input dataset. Customized thresholds allow to focus sensitivity on the features with highest variance and/or the outputs most difficult to predict. This training approach produces a Knowledge Base with conceptual knowledge (TBox) modeled by human experts and factual knowledge (ABox) created automatically from the available data stream, with instances representing the events of interest the system should be able to recognize.

The subsequent **classification** task exploits a semantic matchmaking process based on *Concept Contraction* and *Concept Abduction* non-standard inference services [1]. Given an ontology $\mathcal{T}$ and two concept expressions A and B, if they have conflicting characteristics Concept Contraction determines a concept expression G (Give up) which is an explanation about what in A is not compatible with B and returns a value $penalty^{(c)}$ representing the semantic distance associated to it. Otherwise, if A is compatible with B but does not cover it fully, Concept Abduction calculates a concept expression H (Hypothesis) representing what should be hypothesized (*i.e.*, is underspecified) in B in order to completely satisfy A, and it provides a related $penalty^{(a)}$ value. Concept Contraction and Concept Abduction can be considered as extensions respectively to *Satisfiability* and *Subsumption* standard inference services, which can only provide "yes/no" answers.

In the proposed approach, data of the instance to be classified are first labeled w.r.t. the reference ontology as in the first step of training, then their conjunction is taken as annotation of the instance itself. Penalty values obtained from matchmaking are used to compute the *semantic distance* between the input instance and the event descriptions generated during training. The predicted/recognized event will be the one with the lowest distance. Semantic matchmaking produces ranked similarity measures, associated with a logic-based explanation. Therefore the prediction outcome has a formally grounded and understandable confidence value. This is a clear benefit w.r.t. many standard ML techniques which produce opaque predictions. Furthermore, the approach does not take the instance annotation directly as the output, because the inherent data volatility could lead to inconsistent assertions, which would be impossible to reason on.

System **evaluation** works with a validation set, consisting of several classified instances represented w.r.t. the same ontology used for building the training set. The goal is to check how much the predicted event class corresponds to the actual event associated to each instance of the validation set. Beyond classical tools such as confusion matrix and statistical performance metrics as accuracy, precision and recall, the graded nature of the predictions can be exploited to evaluate, *e.g.*, the average semantic distance of the predicted class from the actual one, analogously to error measure in regression analysis.

As said above, cross-validation allows to tune system parameters if performance is not satisfactory. If computing resources permit it, incoming test data can also be used to update the training matrix on-the-fly allowing evolve the model as new data is observed. A proper extension of the baseline Algorithm 1 requires a fading mechanism to allow the system to "forget" the oldest training samples.

V. Challenges

The proposed semantic-based framework for data stream analysis is a general-purpose approach for advanced event identification. It is intended specifically for mobile and pervasive computing. Those contexts are characterized by severe resource limitations affecting processing, memory, storage and energy consumption. Therefore, hardware and software limitations should be taken into account during systems and applications design. First of all, storage and analysis of a large data set can be impractical on devices equipped with very small memory. The proposed approach allows a streaming approach where samples are discarded as soon as the Training Matrix is updated. The matrix can be orders of magnitude

smaller than the original data. Moreover, straightforward extensions of Algorithm 1 can allow distributed processing by more than one node with a final merging step, which could reduce the communication overhead within a sensor network if intermediate nodes have enough storage capacity. Nevertheless, experimental evaluation campaigns must be carried out to assess and optimize the proposed algorithms. The goal is to identify parameters values ensuring the best trade-off between memory consumption and CPU time.

Efficiency of the semantic matchmaking for on-line classification on data streams is a further concern. When processing logic-based information to infer implicit knowledge, careful optimization is needed to achieve acceptable performance for adequately expressive languages. In [17] an early approach was proposed to adapt non-standard logic-based inferences to pervasive computing contexts. By limiting expressiveness to acyclic TBoxes in the $\mathcal{ALN}$ DL and exploiting performance-optimized data structures, structural algorithms could be adopted with polynomial complexity for both standard and non-standard inferences.

Finally, the association of training labels is a critical task: it is quite difficult to get accurate labels at a finer semantic granularity. Moreover, in problems with a large number of possible classes, many inference procedures are needed for each event recognition, with impact on efficiency of on-line data stream processing. Therefore the ontology modeler should define the appropriate level of class taxonomy according to available training data, application requirements and time constraints compared with node computational resources. This can be achieved by starting with fine-grained class descriptions and then exploiting their semantics to identify the *least common subsumers* for sets of similar classes.

VI. Conclusion and future work

This paper introduced am early proposal for a semantic-enhanced machine learning on heterogeneous data streams. Mapping raw data to ontology-based concept labels provides a low-level semantic interpretation of the statistical distribution of information, while the conjunctive aggregation of concept components allows building automatically a rich and meaningful representation of events during the training phase. Finally, the exploitation of non-standard matchmaking inferences enables a fine-grained event detection by treating the ML classification problem as a resource discovery.

The proposed approach is currently under prototypical implementation. Experimental evaluation on proper testbeds and datasets is ongoing and will allow to assess effectiveness w.r.t. the state of the art in (pervasive) machine learning, as well as feasibility with reference to a possible exploitation on resource-constrained platforms. Besides the already mentioned algorithmic refinements and variants, further investigation is in progress, *e.g.*, about the use of weighted concept expressions [1] to grade matchmaking outcomes in order to reflect the relative relevance of different concept components within output class characterizations.

Acknowledgments

The work was supported by the Italian PON project Puglia@Service and E.T.C.P. Greece-Italy ARGES (pAssengeRs and loGistics information Exchange System) project.

References

[1] M. Ruta, E. Di Sciascio, and F. Scioscia, "Concept Abduction and Contraction in Semantic-based P2P Environments," *Web Intelligence and Agent Systems*, vol. 9, no. 3, pp. 179–207, 2011.

[2] M. Ruta, F. Scioscia, and E. Di Sciascio, "Enabling the Semantic Web of Things: framework and architecture," in *Sixth IEEE International Conference on Semantic Computing (ICSC 2012)*, IEEE. IEEE, sep 2012, pp. 345–347.

[3] M. Ruta, F. Scioscia, A. Pinto, E. Di Sciascio, F. Gramegna, S. Ieva, and G. Loseto, "Resource annotation, dissemination and discovery in the Semantic Web of Things: a CoAP-based framework," in *Internet of Things and Cyber, Physical and Social Computing (iThings/CPSCom), IEEE International Conference on*. IEEE, 2013, pp. 527–534.

[4] M. Abu Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," *IEEE Communications Surveys Tutorials*, vol. 16, no. 4, pp. 1996–2018, 2014.

[5] M. Bahrepour, N. Meratnia, M. Poel, Z. Taghikhaki, and P. J. Havinga, "Distributed event detection in wireless sensor networks for disaster management," in *2nd International Conference on Intelligent Networking and Collaborative Systems (INCOS)*. IEEE, 2010, pp. 507–512.

[6] S. Kher, V. Nutt, D. Dasgupta, H. Ali, and P. Mixon, "A prediction model for anomalies in smart grid with sensor network," in *Proceedings of the Eighth Annual Cyber Security and Information Intelligence Research Workshop*. ACM, 2013, p. 60.

[7] R. Tan, G. Xing, J. Chen, W.-Z. Song, and R. Huang, "Fusion-based volcanic earthquake detection and timing in wireless sensor networks," *ACM Transactions on Sensor Networks*, vol. 9, no. 2, p. 17, 2013.

[8] R. Haber, A. Peter, C. Otero, I. Kostanic, and A. Ejnioui, "A support vector machine for terrain classification in on-demand deployments of wireless sensor networks," in *Systems Conference (SysCon), 2013 IEEE International*, April 2013, pp. 841–846.

[9] Z. Zhang, S. Li, and Z. Li, "Data fault detection using multi-label classification in sensor network," in *Practical Applications of Intelligent Systems*. Springer, 2014, pp. 101–111.

[10] W. Bi and J. T. Kwok, "Multi-label classification on tree-and dag-structured hierarchies," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 17–24.

[11] X. Cheng, J. Xu, J. Pei, and J. Liu, "Hierarchical distributed data classification in wireless sensor networks," *Computer Communications*, vol. 33, no. 12, pp. 1404–1413, 2010.

[12] Y. Song, A. Sailer, and H. Shaikh, "Hierarchical online problem classification for it support services," *Services Computing, IEEE Transactions on*, vol. 5, no. 3, pp. 345–357, 2012.

[13] M. Stocker, M. Ronkko, and M. Kolehmainen, "Situational Knowledge Representation for Traffic Observed by a Pavement Vibration Sensor Network," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 15, no. 4, pp. 1441–1450, 2014.

[14] E. S. Chifu and I. A. Letia, "Unsupervised semantic annotation of Web service datatypes," in *Intelligent Computer Communication and Processing (ICCP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 43–50.

[15] D. Jeon and W. Kim, "Development of Semantic Decision Tree," in *3rd International Conference on Data Mining and Intelligent Information Technology Applications (ICMiA 2011)*, Oct 2011, pp. 28–34.

[16] M. Ruta, S. Colucci, F. Scioscia, E. Di Sciascio, and F. M. Donini, "Finding commonalities in RFID semantic streams," *Procedia Computer Science*, vol. 5, pp. 857–864, 2011.

[17] M. Ruta, F. Scioscia, E. Di Sciascio, F. Gramegna, and G. Loseto, "Mini-ME: the Mini Matchmaking Engine," in *OWL Reasoner Evaluation Workshop (ORE 2012)*, ser. CEUR Workshop Proceedings, vol. 858, 2012, pp. 52–63.