Towards Automatic Ontology Alignment for Enriching Sensor Data Analysis

Marjan Alirezaie and Amy Loutfi

Applied Autonomous Sensor Systems, Dept of Technology, Örebro University,
SE-701 82, Örebro, Sweden
{marjan.alirezaie,amy.loutfi}@oru.se
http://www.oru.se/aass

Abstract. In this work ontology alignment is used to align an ontology comprising high level knowledge to a structure representing the results of low-level sensor data classification. To resolve inherent uncertainties from the data driven classifier, an ontology about application domain is aligned to the classifier output and the result is recommendation system able to suggest a course of action that will resolve the uncertainty. This work is instantiated in a medical application domain where signals from an electronic nose are classified into different bacteria types. In case of misclassification resulting from the data driven classifier, the alignment to an ontology representing traditional microbiology tests suggests a subset of tests most relevant to use. The result is a hybrid classification system (electronic nose and traditional testing) that automatically exploits domain knowledge in the identification process.

Keywords: Ontology Alignment, Sensor Data, Classification, Semantic Gap

1 Introduction

We are surrounded by billions of sensors generating huge amount of data (1800 exabytes in 2011) in our daily lives [8]. There is a large potential to use this sensor data to provide a deeper and better understanding of the world around us. For this to be possible, data must be interpreted and represented in a manner that is compatible for humans. Typically, this interpretation is automated by using complex data driven analysis methods. The output of such methods can still contain inaccuracies due to the fact that low level sensor data is subject to shortcomings due to selectivity, uncertainties and errors. For specific domains e.g. medical domains, the high inaccuracy can hinder the uptake of using automatic data analysis, and thus new sensor technologies.

Infusing more knowledge into the domain generally helps to improve data interpretation. Nonetheless, data driven processes manipulating sensor readings are not able to automatically consider the wealth of high level domain-related knowledge for the integration. In other words, what is needed is a method that is able to automatically fuse the high level knowledge to low level data which are inherently unintuitive and difficult to interpret.

Table 1. Bacteria Species

Bacteria Name	Short Name	Bacteria Name	Short Name
Escherichia coli	EColi	Entercoccus faecalis	ENTFL
Pseudomonas aeruginosa	PSAER	Staphylococcus lugdunensis	STLUG
Staphylococcus aureus	STA	Pasteurella multocida	PASMU
Klebsiella oxytoca	KLOXY	Steptococcus pyogenes	STRPY
Proteus mirabilis	PRMIR	Hemophilus influenzae	HINFL

In this work, we extend the ontological alignment approach explained in [1] for improving signal level classification results so that the method is no longer dependent to the structure of the classifier. We show how to use ontology alignment methods to find high level annotations for the results of signal data analysis in order to resolve ambiguities. Instantiating the methods using the electronic nose, we particularly examine the task of blood bacteria identification (10 blood bacteria species listed in Table 1). The alignment method presented here is applied over classification results of raw sensor data coming from the electronic nose and information pertaining to distinguishing bacteria laboratory tests which are modelled in an ontology. The ontology is aligned to the sensor data which is represented as a tree structure. Obtaining the tree structure depends on the classification method. A hierarchical classifier such as a decision tree, produces a tree structure that can be directly aligned with the ontology. A non-hierarchical classifier, on the other hand, requires further processing to generate a tree structure that can be used in the proposed alignment method. Using this method, if the classifier does not have a hierarchical structure, we find cases causing uncertainties in the classification and make tree-shaped structures with them to feed the alignment process. This alignment technique also uses a heuristic to decrease the number of required laboratory tests that should be applied on samples of bacteria. The result is an enriched classification method which is able to use information represented at a high-level (symbolic) to provide recommendations based on the automatised interpretation of the sensor data.

To familiarize the reader with the sensors used in this work we begin with a brief introduction of electronic noses and their applications to bacteria identification. The paper then proceeds to outline the related works in Sect.3. After that, the next section concentrates on details of the methodology. In Sect.5, our data set structure along with a short description about the sampling process is discussed. Then, Sect.6 represents results of each step of the methodology. The paper ends with the discussion and the conclusion.

2 Electronic Noses (E-Noses)

Comprising a set of chemical sensors, an electronic nose is a machine with the ability of detecting and discriminating odors or gases. It consists of an array of sensors each of with partial selectivity and a pattern recognition algorithm. Each odor based on its chemical characteristics makes a "fingerprint" from the

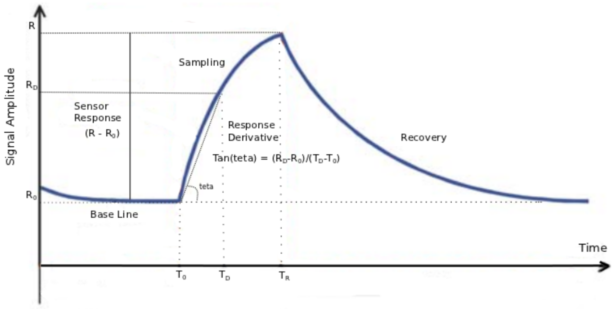


Fig. 1. A signal with three phases (Baseline, Sampling, Recovery).

sensor array which are converted to time series signals [24]. Figure 1 illustrates the response from one sensor in an e-nose when first exposed to the clean air (baseline phase: $t < T_0$), then to a target gas (sampling phase: $T_0 \leq t < T_R$) and finally again to the clean air (recovery phase: $t \geq T_R$).

Electronic noses have been applied to a number of different application domains, ranging from food process monitoring to environmental monitoring [24]. In this work an e-nose is used to provide a quick identification of bacteria existing in blood samples. This is an alternative method to traditional laboratory analysis where samples are first incubated and cultured requiring several days before an identification result is possible. With an electronic nose, it is possible to reduce this time significantly as the sampling of an odour requires anywhere from 3-6 minutes. More details on the use of electronic noses for the detection of bacteria can be found in [9, 11].

The construction of an electronic nose instrument can vary depending on the number and types of sensors used. As described in Sect.5, two different types of e-noses are used in this work, resulting in two data sets, each of which is associated with a classification method.

3 Related Works

To clarify our alignment approach it is worth mentioning that it concerns fusion of information at different levels of abstraction in order to bridge the semantic gap which occurs between these levels [23, 1].

Data fusion is known as one of the most popular solutions in data processing efforts. The common part of works related to integration of data is keeping synchronised the different types of data that come from various sensors measuring same environment [25]. Typically, in the area of sensor fusion, fusion methods consider homogeneous raw data sources i.e. numeric data, whereas the integration methodology in this work is applied on both numeric and symbolic data.

Works which consider fusing symbolic knowledge to numeric data have gained attention in AI fields related to robotics and physically embedded systems[17, 18]. The symbol grounding problem [16] in general and the anchoring problem [10] in particular concentrate on the process of creating and maintaining the relation between a symbol chosen to label an object in the world and those data coming from sensors observing the same physical object in the environment. In most of works such as [15] the challenge is how to perform the anchoring in an artificial system and how to find relevant concepts related to symbols to improve the recognition process. In these efforts the association is done in two ways: grounding well-defined concepts in data (top-down process) and conceptualizing data that exemplifies the concept (bottom-up process). In these kinds of mapping we need to have the information about the objects in the environment which is manually (not automatically) modelled and labelled with symbols. This work leverages from these two approaches. First, the bottom-up knowledge acquisition is acquired solely based on categorical information obtained by physical sensors e.g. class labels. Secondly, a top-down approach is achieved by using the created ontology to find similarity with the result of data analysis methods.

Likewise, ontology grounding is by definition the process of associating abstract concepts to low level data [6, 19]. Since ontologies provide the possibility of expressing entities in conceptual categories which are shareable, they are suitable structures whereby high level knowledge can be modelled. In other words, the interleaved structures of ontologies make it possible to enrich the symbolic representation of a concept by retrieving its different features reified in an ontology for provisioning a more accurate grounding [31]. Our reason to use ontologies, however, is not only for the purpose of grounding. In this work, the reusability of relevant concepts which are needed to make annotation and description for our final recommender system is basically considered.

For some situations where there are intelligible meanings for the measured features of data, it is certainly possible to extract feature-related concepts from ontologies and re-structure the data set according to them for the sake of improvement in data interpretation processes [3, 4]. In the case of electronic noses, because of speciality of sensors, there was no chance to find some extra meanings related to features measured by sensors to be intermixed with raw data.

From another point of view, it is worth to consider the automatic ontology development part of our work as well. According to the survey [7], the ontology generation tasks are categorized into 4 groups: conversion or translation, mining based, external knowledge based and frameworks; and our method belongs to the third one since we are utilizing a set of available ontologies [28] related to the domain. The first step of our ontology building uses the same approach applied by most of works such as [5] and [30] where they query the WordNet database to retrieve the initial synonyms or a simple definition for the concept. Instead of WordNet we use another repository that contains most updated biomedical ontologies being full of subsumption relations [28].

Finally, the alignment which is defined as the process of determining correspondences between concepts [26], is mostly used when two sides of the process are ontologies. However, in this work, we map an ontology with the result of a data driven method according to the names of bacteria.

4 Methodology

The methodology used in this work applies the following general steps which are discussed in next sections. First, we classify pre-processed sensor data. Second, a tree structure that contains the class labels (bacteria names), participating in misclassification situations, is built. Third, the resulted tree is aligned with the ontology containing relevant knowledge. Finally, the method replaces candidate parts of the ontology with their counterparts in the tree.

4.1 Classification of Sensor Data

As mentioned before, in this work we make the alignment process independent of the classification method. At this first step, to evaluate the resilience of the methodology both hierarchical and non-hierarchical classification techniques are used. For the hierarchical group the candidate method is the C4.5 decision tree and for non-hierarchical the Dynamic Time Warping algorithm is evaluated.

The Decision Tree (D-Tree) A decision tree classification has the advantage that it provides transparency in the representation of the outputs [20] and has a suitable (hierarchical) structure for our alignment process [1]. The C4.5 algorithm is used and finds a feature of the training set providing the maximum degree of discrimination between different classes of bacteria. The algorithm iterates, each time splitting instances of the training set according to the most informative selected feature. Each feature value creates a decision node for the tree [20]. Using the confusion matrix[1] from the classification result, a second process finds misclassification positions among leaf nodes of the tree and assigns them all bacteria names sharing these nodes. During this process leaf nodes are divided into two groups A (containing all leaves without misclassification) and B (holding the rest).

Starting with group B nodes, a third process checks if the sibling node also belongs to the group B. In these cases where two sibling leaves belong to the group B, the process labels their parent by all bacteria names shared by them; otherwise, the process keeps the candidate B leaf node by its own bacteria names. Two procedures of our previous work, *RelabelDTree* and *CheckParent*, described in [1], show the details of the decision tree relabelling process. Once all nodes in group B or in the parents of group B are labelled, the tree is ready to be transferred to the alignment process.

[1] Confusion matrix is a table that visualizes the performance of a supervised learning algorithm so that rows and columns are labelled with actual and predicted classes, respectively [21].

Dynamic Time Warping (DTW) The accuracy of the d-tree is highly dependent on the selected features of the signal. Alternative approaches to classification such as Dynamic Time Warping are able to consider the entire signal data rather than extracting specific features from data [20]. The Dynamic Time Warping algorithm [29] is an extensively used technique to measure the similarity of two sequences of data over the time. Taking any two signals as input, the DTW calculates the (Euclidean) distance between them at each time point to see how dissimilar they are. As our sensor signals are the same size, they are suitable for the DTW method.

In order to apply our alignment method, a confusion matrix is needed. Following is a brief explanation on how we build the confusion matrix for the DTW signal processing results:

After preparing the training and test set, we build the $n \times m$ similarity matrix S, where n and m are the length of the test and the training set, respectively. The element $a_{i,j}$ of this matrix is the distance (dissimilarity) value between the ith test case and the jth training case. The confusion matrix C of dimension $l \times l$ is initialized to the zero matrix. Considering each row of the matrix S as r, we update the matrix C so that its element $b_{i,j}$ receives 1-unit increase if i and j are r's test case label index and r's element label index that has the minimum distance value, respectively. To resolve the uncertainities in C, a binary tree is built from candidate labels[2].

To develop the binary tree, we start from the root node having been labelled with a set containing l bacteria names. These l names are those "actual" labels assigned to non-zero elements of a column of C that belongs to a "predicted" label. Creating two branch nodes of the root, the process has to label them so that each child node's label set is a subset of the parent's label set and the complementary set of its sibling as well. In order to keep the completeness[3] of the algorithm, in the process of creating children nodes we have to consider the power set[4] of the parent's label set. Consequently, we have $2^l/2$ different options to split the root node. However, for the sake of reducing the number of required microbiology tests in total (increasing the information gain), we prioritize those branches that are almost in balance in terms of the number of labels. In other words, we divide the number of labels l by 2 to find a laboratory test being able to differentiate between the two biggest subsets (among all elements of the power set) of bacteria names. The number of different branches holding $\lceil \frac{l}{2} \rceil$ labels is hence as follows:

$$\binom{l}{\lceil \frac{l}{2} \rceil} = \frac{l!}{(l - \lceil \frac{l}{2} \rceil)! \times \lceil \frac{l}{2} \rceil!} \tag{1}$$

The first subset of labels with length $\lceil \frac{l}{2} \rceil$ has a sibling set (a complementary label set) signed by the rest of input class labels. Once a branching-labelling step is finished, the alignment process (Sect.4.3) is called to check if there is a

[2] There is an uncertainty if at least one non-diagonal element of a confusion matrix is non-zero.

[3] An algorithm is complete when it finds the solution if there is one.

[4] The power set is a set of all subsets of a set.

counterpart node for the root in the ontology. If the alignment returns a result, we proceed until a complete tree whose each leaf node is labelled by a single bacteria name is built; otherwise we switch to the next candidate for the parent node. In this way, since we divide the list by 2 at each step, the depth of the created tree is m if: $2^{m-1} < l \leq 2^m$.

Algorithm 1 Building tree from candidate class labels

```
function BUILDDTREE($labels$)
    $root \leftarrow makeNode(labels)$
    if $length(labels) = 1$ then return $root$
    end if
    $subSetList \leftarrow getSubSets(labels, length(labels)/2)$
    for all $s$ in $subSetList$ do
        $p \leftarrow getComplementarySet(s)$
        if $ontologyAlign(root, s, p) \neq null$ then
            $rightNode \leftarrow BuildTree(s)$
            $leftNode \leftarrow BuildTree(p)$
            $attachToTree(root, rightNode, leftNode)$
        else return $null$
        end if
    end for return $root$
end function
```

Algorithm 1 represents the process of the building tree from the selected candidate class labels. Provided a confusion matrix, this method can indeed be applied on classification results of any classification method and make them ready for the alignment process with an ontology.

4.2 Microbiology Tests Ontology

The resulting tree from the classification process is aligned to an ontology that contains the high level knowledge related to the domain of bacteria laboratory tests. This ontology has been created in a semi-automatic way so that only the initial concepts such as bacteria represented in Table 1 and microbiology tests related to general categories of bacteria[5] with the positive/negative results have been manually modelled in the ontology.

The automated phase of building the ontology regards knowledge existing in *BioPortal* [28]. *BioPortal* is the repository of biomedical ontologies among which *SNOMED CT* contains key terminologies in biomedicine (about 40,000 classes [12]). This information[6], which is useful in finding the expected results of some microbiological tests for specific bacterial species, are certainly considered during traditional bacteria identification processes in laboratories [27].

[5] http://www.atsu.edu/faculty/chamberlain/Website/lab/idlab/flowchp.htm

[6] The chemical features of bacteria such as cell morphology and gram stain [22].

Having the bacteria name as the input, the Java interface using the Jena API with the ARQ query engine runs a query in a loop to retrieve a hierarchical list that contains the whole indirect super classes of this bacteria. This loop is run until it meets the class named *Bacteria* in the ontology. The returned list contains different (more general) categories of bacteria for which the microbiology tests with the positive/negative results are available in the ontology.

For example, the following SPARQL code returns the name of the first super class of the class which is labeled by the first bacteria name in the Table 1, *Escherichia Coli*.

```
SELECT DISTINCT ?ss0 ?nstep ?label
{
 GRAPH <SNOMEDCT URI>{
 snterm:112283007 rdfs:subClassOf ?ss0.
 ?ss0 skos:prefLabel ?label.
 BIND ('1' AS ?nstep).}
}
```

In an iterative process, other classes are returned such as *GramNegative* or *Bacillus*, which state the gram stain and wall shape type of the bacteria, along with the laboratory tests such as the *Lactose* test which is positive. This extracted information is populated into our ontology after adding the object properties such as *hasTest* and *hasValue* for the *Escherichia Coli* class. Continuing this way of knowledge acquisition, we obtain an ontology that contains results of 7 different microbiology tests for all 10 types of blood bacteria.

4.3 Ontology Alignment

The main task of the aligning step is comparing two types of entities, a tree node holding a set of class labels (bacteria name), and the ontology class(es) containing information related to the bacteria species. As the sensor data lacks semantics, we apply terminological and structural alignment methods [23].

The alignment method calculates the similarities between the two mentioned type of entities by using the Jaro-Winkler algorithm [13]. The algorithm is based on Jaro-Winkler distance (2) and counts the number of matching characters in two strings to measure the distance between them. The lower the JaroWinkler value, the more similar are the strings (bacteria names) [13].

$$distance = \frac{1}{3} \times (\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}) \tag{2}$$

Where:

m: number of matching characters.
t: half the number of transpositions.
s_i: length of i^{th} string.

Because of having no isomorphism relation between the tree and the ontology [14], the graph inexact matching as the structural alignment method is used to determine if there is any similar structure between labelled nodes in the tree and the ontology.

Given the similarities between two structures, a replacement process copies the labels of the classes found in the ontology to relabel the counterpart nodes in the tree. In this way, the tree structure is an annotated structure that contains two types of information, sensor values and microbiology tests, that can act as a recommender in case of uncertainties caused by misclassification.

5 DataSet

Ten types of bacteria species listed in Table 1, sub-cultured on blood and agar plates, are clinical samples in this scenario. In order to discriminate between them, we consider two sampling methods using two different electronic noses. The following section concisely describes the sampling configurations for these two e-noses:

5.1 Sampling Process (Bacteria in Blood and Agar)

The sampling system used for "sniffing" the bacteria in blood is a NST 3220 Emission Analyser from Applied Sensors, Linköping, Sweden. This machine is composed of 22 chemical sensors[7]. In this experiment the baseline acquisition lasts for 10 seconds. Next, the headspace[8] gases are injected into the sensor chambers and the sensors are exposed for 30 seconds in the sampling phase which is followed by the recovery phase lasting for 260 seconds (Fig.1). Each of the 10 bacteria has been sampled 60 times with the 600 "sniffs". Further details of the sampling process and preparation are given in [11].

To make a suitable structured training set for the classification, we pass the continuous time series data generated by 22 sensors through a pre processing phase that includes two steps: Baseline manipulation normalizing the sensor data and compression extracting informative descriptors of signals to make feature vectors [20]. In this way, we replace each sensor signal with its two descriptors indicated in Fig.1. The static response calculating the difference between the end point of the sampling phase and the baseline, gives one single parameter; and the response derivative which is equal to the slope of the line contiguous to the third second of the the sampling phase. A total of 44 feature values are produced for the dataset of 600 samples accompanied by a label list containing bacteria species names listed in the Table 1.

The second data set was collected using a Cyranose 320, which contains a sensor array of 32 conducting polymer sensors. This electronic nose samples bacteria which has been cultured on agar plates. The sampling and recovery phases are 20 and 80 seconds long, respectively. Each of the 10 bacteria has been sampled 40 times resulting in 800 "sniffs". Additional details of this sampling method are given in [9]. Due to the similarities of these signals, we keep the whole signals without any feature extraction phase.

[7] 10 MOS and 12 MOSFET sensors [24].

[8] The headspace is the space just above the liquid sample in a bottle [24].

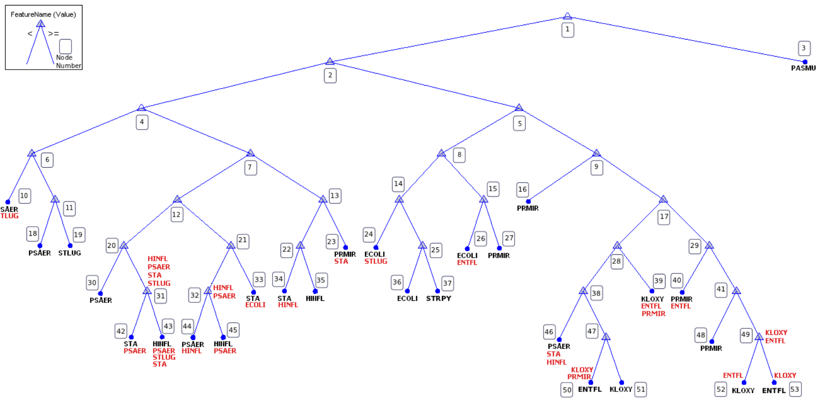


Fig. 2. The Decision Tree Showing Misclassification Cases with Red Labels

6 Results

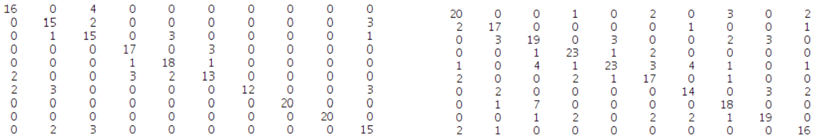
Based on the aforementioned assumption about dividing the classification methods into hierarchical and non-hierarchical groups, we first discuss on results of the two methods separately and then concentrate on the alignment technique results for both.

6.1 Hierarchical Classifier Results

The C4.5 decision tree algorithm was applied for the first data set read by the NST 3220 Emission Analyser electronic nose. This data set as described in Sect.5.1 contains 600 instances with 44 features. After applying a 10-fold cross validation on this data set in order to generalize the error estimation of the classification [21], two thirds (400 cases) of the samples were chosen to form the training set and the rest were considered as test cases. Figure 2 shows the result of the classification fed by the training set. Decision nodes of the tree have been labelled by feature names and criteria values. Leaf nodes of the tree have also been marked by bacteria species names. The confusion matrix of this classification is depicted in Fig.3(a). According to this matrix and (3), among the 200 test cases, there are 39 misclassification corresponding to an accuracy of 80%.

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} = \frac{\sum_{i=1}^{l} a_{ii}}{sum_{i=1}^{l} sum_{j=1}^{l} a_{ij}} \tag{3}$$

Table 2 shows the details of group B nodes which hold misclassified cases. For example, as we can see in the table, the node number 10 is shared with two types of bacteria, the predicted type 2 (according to the training set) and the actual type 7 (according to the test set). In the same way, the details about node number 43 which is shared by bacteria type 10, 2, 3 and 7 are listed shown by 11th, 12th



16	0	4	0	0	0	0	0	0	0
0	15	2	0	0	0	0	0	0	3
0	1	15	0	3	0	0	0	0	1
0	0	0	17	0	3	0	0	0	0
0	0	0	1	18	1	0	0	0	0
2	0	0	3	2	13	0	0	0	0
2	3	0	0	0	0	12	0	0	3
0	0	0	0	0	0	0	20	0	0
0	0	0	0	0	0	0	0	20	0
0	2	3	0	0	0	0	0	0	15

(a) DTree Confusion Matrix.

20	0	0	1	0	2	0	3	0	2
2	17	0	0	0	0	1	0	0	1
0	3	19	0	3	0	0	2	3	0
0	0	1	23	1	2	0	0	0	0
1	0	4	1	23	3	4	1	0	1
2	0	0	2	1	17	0	1	0	0
0	2	0	0	0	0	14	0	3	2
0	1	7	0	0	0	0	18	0	0
0	0	1	2	0	2	2	1	19	0
2	1	0	0	0	0	0	0	0	16

(b) DTW Confusion Matrix.

Fig. 3. Resulted Confusion Matrices.

and 13th items in Table 2. The details of this table are visually depicted in Fig.2. For example, the subtree containing node 49 as the parent and nodes 52 and 53 as children belonging to group B and are sharing bacteria number 4 (*Klebsiella Oxytoca* or *KLOXY*) and 6 (*Entercoccus faecalis* or *ENTFL*).

The misclassification caused by the inconsistencies between the predicted and actual classes is resolved by utilizing the ontology information which is about the discriminating laboratory tests. The alignment results is explained in Sect.6.3.

6.2 Non-Hierarchical Classifier Results

We chose 260 test cases and 540 samples for the training set. The confusion matrix from the DTW method is shown in Fig.3(b). It shows that there are 74 misclassification among 260 test cases that lead to an accuracy of 72%.

Table 3 indicates how incorrect predictions correspond to the actual classifications. For example, for some situations where the test case actual class is bacteria number 1 (*EColi*), 3 (*STA*), 5 (*PRMIR*), 6 (*ENTFL*) or 9 (*STRPY*) the classifier has predicted this test case as the bacteria number 8 (*PASMU*) (column 8 in Figure 3(b)). Therefore, the label set *EColi*, *STA*, *PRMIR*, *ENTFL*, *PASMU*, *STRPY* will be the input for the mentioned building tree process. One of the built tree hierarchical structure returned for this example is represented in Figure 5(a). The size of the label set is $l = 6$ and the depth of the tree is $m = 3$ ($2^2 < 6 \leq 2^3$). Having this tree, the alignment process can follow the similarity checking process explained in the following section.

6.3 Alignment Results

By the string matching method, the alignment process finds all bacteria names in the ontology that are similar to the candidates. Table 4 demonstrates some parts of Jaro-Winkler distances between bacteria names in the classifier and in the ontology. As we can see, the minimum value of each column is located in the diagonal position which proves the correctness of the mapping of the bacteria names. Considering these values, the graph matching method then extracts the most similar parts of the ontology to the tree in terms of the bacteria names labelling the nodes.

The alignment result related to the decision tree classifier is depicted in Fig.4(a). The laboratory test candidate for the parent of nodes 52 and 53 (Table 2) sharing *KLOXY* and *ENTFL* is the *Catalas* test which discriminates

Table 2. Decision Tree, B-leaf Nodes (Prctd = Predicted, Actl = Actual)

Item#	Node	Prctd	Actl	Number	Item#	Node	Prctd	Actl	Number
1	10	2	7	3	11	43	10	2	2
2	23	5	3	3	12	43	10	3	1
3	24	1	7	2	13	43	10	7	3
4	26	1	6	2	14	44	2	10	1
5	33	3	1	4	15	45	10	2	1
6	34	3	10	3	16	46	2	3	1
7	39	4	5	1	17	50	6	4	2
8	39	4	6	1	18	50	6	5	1
9	40	5	6	2	19	52	4	6	2
10	42	3	2	2	20	53	6	4	1

Table 3. DTW Misclassification Cases (Prctd = Predicted, Actl = Actual)

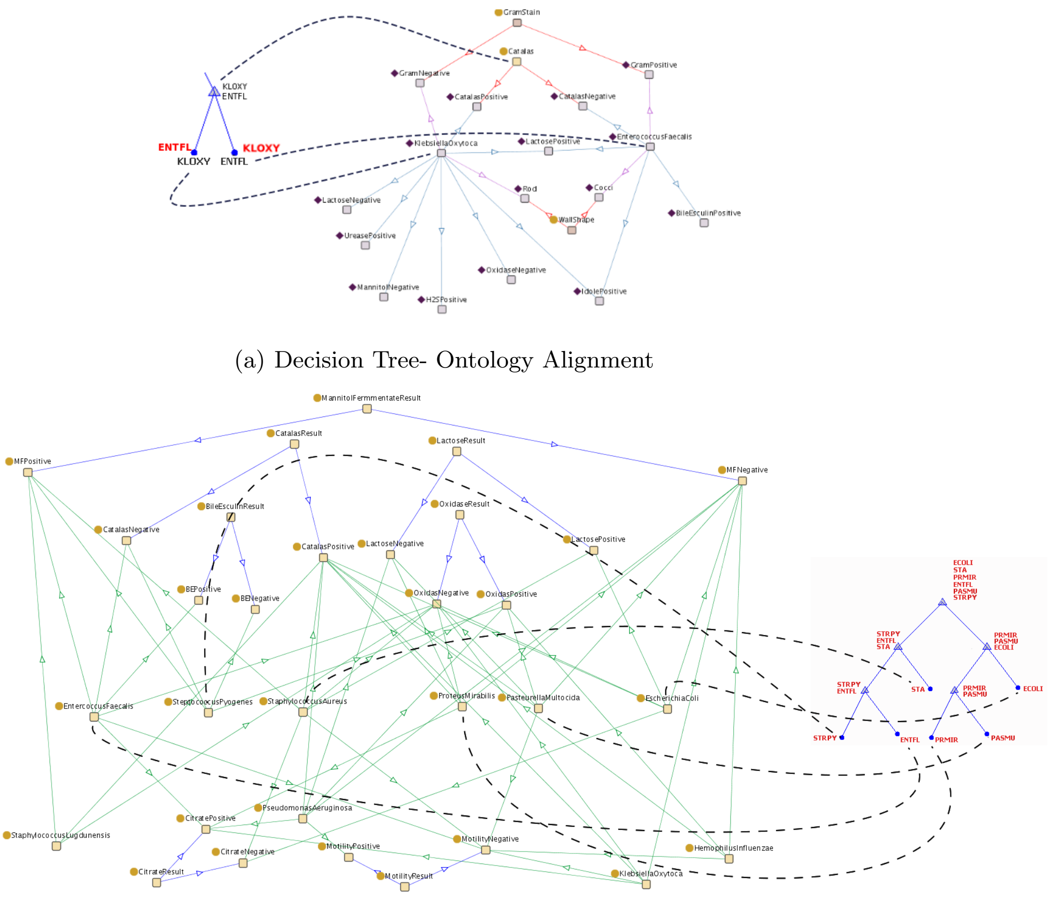
Prctd	Actl	Actl	Actl	Actl	Actl	Prctd	Actl	Actl	Actl	Actl	Actl
1	2	5	6	10	-	6	1	4	5	9	-
2	3	7	8	10	-	7	2	5	9	-	-
3	4	5	8	9	-	8	1	3	5	6	9
4	1	5	6	9	-	9	3	7	-	-	-
5	3	4	6	-	-	10	1	2	5	7	-

between $ENTFL$ and $KLOXY$ with its negative and positive response. Therefore, the ontology suggestion can annotate the subtree holding information about these leaf nodes. By applying the alignment process on the whole nodes in group B, we will finally have an annotated decision tree demonstrated in Fig.6.

Similarly, the mapping of laboratory tests on nodes of the binary trees built based on the Algorithm 1 is represented in Fig.4(b). In this figure, we can see the *MannitolFermentation* is the best microbiology test candidate to divide the 6 bacteria labeling the leaf nodes of the tree (Fig.5(a)). *EColi*, $PASMU$ and $PRMIR$ have the negative reaction to the *Mannitol* test in contrast to STA, $STRPY$ and $ENTFL$ which have positive reaction to it. It means that the *Mannitol* test is chosen to annotate the root of the tree which is pointing to the 8^{th} column of the confusion matrix. In other words, whenever the result of the classifier is bacteria number 8 ($PASMU$) the *MannitolFermentation* test is recommended to be applied first. In the same way, we find the second best laboratory test for each branch of the tree. *Lactose* and *Catalase* tests are suggested for annotating the right and left child of the root, in order. *EColi* has the positive reaction to the Lactose test whereas the $PASMU$ and $PRMIR$ result to the negative one. By follwoing all the object properties named *hasReaction* (green arrows in the Fig.4(b)) we obtain the annotated tree demonstrated in Fig.5(b).

Table 4. Jaro-Winkler distances of names between the classifier and the ontology.

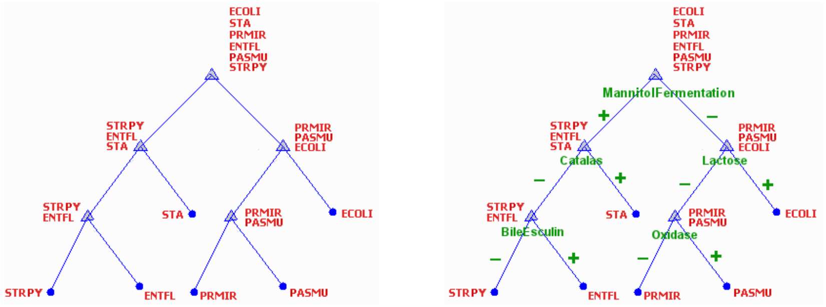
D-Tree Ontology	ECOli	PSAER	STA	KLOXY	PRMIR	ENTFL	STLUG	PASMU	STRPY	HINFL
Escherichia coli	**0.364**	0.515	0.535	1	0.515	0.521	0.579	0.579	0.569	0.492
Pseudomonas ae...	0.503	**0.259**	0.414	0.585	0.379	0.503	0.503	0.352	0.414	0.585
Staphylococcus a...	0.641	0.530	**0.200**	0.530	0.584	0.502	0.279	0.503	0.397	0.502
Klebsiella oxytoca	0.522	0.522	0.407	**0.397**	0.581	0.663	0.663	0.663	0.407	0.663
Proteus mira...	0.519	0.439	0.572	0.580	**0.242**	0.661	0.519	0.364	0.536	1
Entercoccus fa...	0.477	0.502	0.540	0.584	0.502	**0.293**	0.502	0.668	1	0.584
Staphylococcus lu...	0.650	0.539	0.206	0.539	0.587	0.508	**0.269**	0.515	0.406	0.508
Pasteurella mul...	0.502	0.289	0.397	0.584	0.377	0.530	0.420	**0.224**	0.579	0.584
Steptococcus py...	0.532	0.506	0.331	0.585	0.532	0.670	0.337	0.532	**0.238**	1
Hemophilus infl...	0.423	0.423	0.581	0.670	0.532	0.503	0.532	0.391	0.359	**0.379**



(a) Decision Tree- Ontology Alignment

(b) Tree (from DTW)- Ontology Alignment

Fig. 4. Alignment Process (Between Candidate SubTree and Matched Sub-Ontology).



(a) The Misclassification Tree (b) Annotated Tree of DTW results

Fig. 5. The Tree Built from Misclassification of DTW (Predicted Bacteria PASMU)

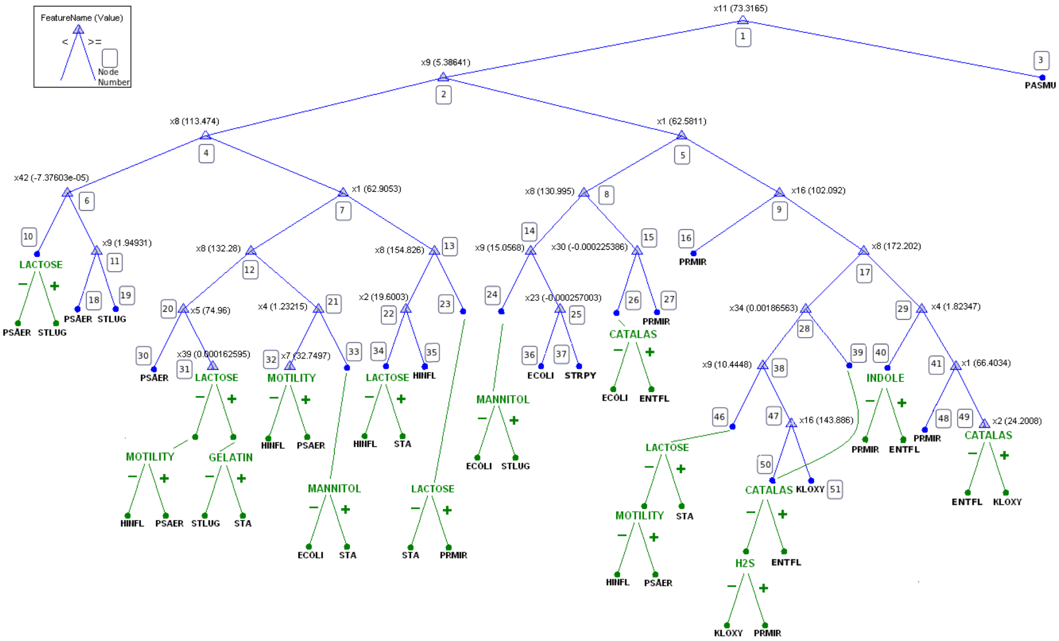


Fig. 6. Annotated Decision Tree by Laboratory Test Information.

7 Conclusions

This work is as the extension of our previous work [1] where we implemented an ontology alignment method to improve classification results of electronic nose sensors readings. We developed the work in terms of two aspects. First, regardless of the classification method and its structure, we can use the generic alignment methodology introduced in Sect.4.3 and generalize the work to cover all applications using different kinds of data coming from various sensor types. Secondly, an extension that automates the ontology development has been considered in this work. The efficiency of this development technique can be further improved in future works. Specifically, there is a limitation due to the amount of acquired knowledge which is dependent on public knowledge repositories.

Considering this generic methodology, future work can also examine the result of the alignment process for a different scenario where a more intelligible meaning of the measured features can be obtained. In this case, it could be possible to include the feature set into the alignment process as well. We will need to appraise the string matching methods if we confront more labels and names than the current label set containing only 10 names of bacteria.

Acknowledgments. This work has been supported by the Swedish National Research Council, Vetenskapsrådet, project nr 2010-4769, on cognitive olfaction.

References

1. Alirezaie, M., Loutfi, A. : Ontology Alignment for Classification of Low Level Sensor Data. Proceedings of 4th KEOD International Conference on Knowledge Engineering and Ontology Development, pp.89-97, Springer-Verlag, (2012)
2. Salvadores, M., Horridge, M., Alexander, P.R., Fergerson, R.W., Musen, M.A., Noy, N.F.: Using SPARQL to Query BioPortal Ontologies and Metadata. Proceedings of International Semantic Web Conference (2), pp.180-195, (2012)
3. Zhang, J., Silvescu, A., Honavar, V.: Ontology-Driven Induction of Decision Trees at Multiple Levels of Abstraction. Proceedings of Symposium on Abstraction, Reformulation, and Approximation, Springer-Verlag, (2002)
4. Bouza, A., Reif, G., Bernstein, A., Gall, H.: SemTree: Ontology-Based Decision Tree Algorithm for Recommender Systems. International Semantic Web Conference (Posters & Demos), (2008)
5. Kong, H., Hwang, M., Kim, P.: Design of the automatic ontology building system about the specific domain knowledge. 8th ICACT International Conference on Advanced Communication Technology, International Symposium on High Performance Distributed Computing, (2006)
6. Jakulin, A., Mladenić, D.: Ontology Grounding. Proceedings of 8th International multi-conference Information Society, pp.170–173, (2005)
7. Bedini, I., Nguyen, B.: Automatic Ontology Generation: State of the Art. Technical report, University of Versailles, (2007)
8. Gantz, J.F., Chute, C., Manfrediz, A., Minton, S., Reinsel, D., Schlichting, W., Toncheva, A.: The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011. Technical report, emc, IDC, (2008)

9. Längkvist, M., Loutfi, A.: Unsupervised feature learning for electronic nose data applied to bacteria identification in blood. NIPS Workshop on Deep Learning and Unsupervised Feature Learning, (2011)
10. Loutfi, A., Coradeschi, S., Saffiotti, A.: Maintaining Coherent Perceptual Information using Anchoring. The 19th International Joint Conference on Artificial Intelligence (IJCAI). pp.1477-1482, (2005)
11. Trincavelli, M., Coradeschi, S., Loutfi, A., Söderquist, B., Thunberg, P.: Direct identication of bacteria in blood culture samples using an electronic nose. IEEE Trans Biomedical Engineering. 57, (2010)
12. Price, C., Spackman, K.: SNOMED clinical terms. British Journal of Healthcare Computing & Information Management. 17(3), pp.27-31, (2000)
13. Jaro, M.: Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Society. 84, pp.414-420, (1989)
14. Hlaoui, A.: A new algorithm for inexact graph matching. Object recognition supported by user interaction for service robots. 4, pp.180-183, (2002)
15. Melchert, J., Coradeschi, S., Loutfi, A.: Knowledge Representation and Reasoning for Perceptual Anchoring. Tools with Artificial Intelligence. 1, pp.129-136, (2007)
16. Harnad, S.: The Symbol Grounding Problem. Physica D: Nonlinear Phenomena. 42, pp.335–346, (1990)
17. Sossai, C., Bison, P., Chemello, G.: Fusion of symbolic knowledge and uncertain information in robotics. Int. J. Intell. Syst. 16, pp.1299-1320, (2001)
18. Chella, A., Frixione, M., Gaglio, S.: Anchoring symbols to conceptual spaces: the case of dynamic scenarios. Robotics and Autonomous Systems. 43, pp.175188, (2003).
19. Fiorini, S.R., Abel, M., Scherer, C.M.S.: An approach for grounding ontologies in raw data using foundational ontology. Information Systems, Elsevier, (2012).
20. Quinlan, R.: C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning), 1st ed. Morgan Kaufmann, (1992)
21. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, (2006)
22. Seltmann, G., Holst, O.: The Bacterial Cell Wall. Springer-Verlag, (2002)
23. Ehrig, M.: Ontology Alignment: Bridging the Semantic Gap. Springer, (2007)
24. Pearce, T.C., Schiffman, S.S., Nagle, H.T., Gardner, J.W.: Handbook of machine olfaction: electronic nose technology. Wiley-VCH, (2003)
25. Joshi, R., Sanderson, A.C.: Multisensor Fusion: A Minimal Representation Framework. World Scientific, Series in Intelligent Control and Intelligent Automation, (1999)
26. Euzenat, J., Shvaiko, P.: Ontology matching. Springer-Verlag, (2007)
27. A national clinical and anatomic pathology reference laboratory, 2006, `http://www.aruplab.com`
28. The BioPortal Metadata Ontology, 2012, `http://www.aruplab.com`
29. Ratanamahatana, C.A., Lin, J., Gunopulos, D., Keogh, E.J., Vlachos, M., Das, G.: Mining Time Series Data. Data Mining and Knowledge Discovery Handbook. pp.1049-1077, (2010)
30. Moldovan, D., Girju, R.: Domain-Specific Knowledge Acquisition and Classification using WordNet. (2000)
31. Jakulin, A., Mladenić, D.: Ontology Grounding. In SIKDD at Multiconference IS, (2005)