Ontology Alignment for Classification of Low Level Sensor Data

Marjan Alirezaie and Amy Loutfi
Applied Autonomous Sensor Systems, School of Science and Technology, Örebro University, SE-701 82, Örebro, Sweden

Keywords: Ontology Alignment, Decision Tree, Classification, Semantic Gap.

Abstract: In this work we show how alignment techniques can be used to align an ontology to a decision tree representing the features used in classification of sensor signals. The sensor data represents time-series data from an electronic nose when measuring bacteria in blood samples. The objective is to provide from the classification of these signals an estimate of the type of bacteria present in the sample. As these classification are inherently uncertain, knowledge about standard laboratory tests are used together with the classification result in order to determine a subset of tests to conduct that should result in a fast identification of the bacteria. The information about the laboratory tests are contained in an ontology. The result from the alignment is new classifier where recommendations are given to a user (expert) based on the interpretation of the sensor data that is done automatically.

1 INTRODUCTION

The uptake of automatic analysis of sensor data in certain applications can be hindered by the difficulty for end users to understand the data-driven processes done by the computer. This is particularly true where the liability of human error can be high, e.g. medical diagnosis. In this work, we examine such an example where a new sensor technology based on chemical sensors is applied to the identification of bacteria in blood. As the presence of bacteria in blood can be life threatening to a patient, it is important to identify the bacteria strain and apply an appropriate antibiotic as quickly as possible. Using the sensor the identification process could be reduced by several days, however, current identification accuracy is approximately 80% using state of the art machine learning methods. This is due to the fact that the low level sensor data is dependent to properties such as sensor type and selectivity of the sensors and as such results in misclassifications and an inaccuracy that is not accpetable for medical domains.

In this paper, an ontological approach is used for improving signal level classification results. On one hand, we rely on the sensor data from the electronic nose to make the identification of bacteria and on the other hand we use information about traditional laboratory testing to resolve ambiguities in the sensor data classification. In this way, the uncertainties about the sensor data are resolved using traditional techniques with the added benefit that only a subset of traditional techniques need to be applied.

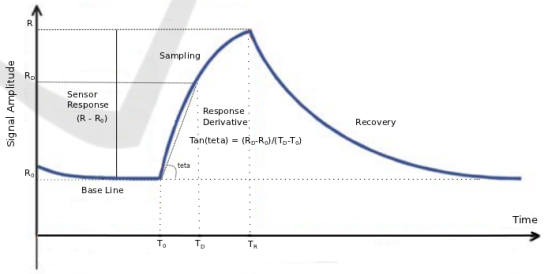

Figure 1: A signal with three phases (Baseline acquisition, Sampling, Recovery).

Using alignment techniques, we show how it is possible to align the ontology with the decision tree representing the features used in classification of the sensor signals. Our method replaces nodes in the decision tree (the classifier) that are particularly uncertain with information from the ontology. The resulting classifier therefore provides a recommendation of which laboratory tests should be conducted taking into account both the ontology and the sensor data. This method is implemented for a task of identification of 10 blood bacteria species listed in Table 1. Each sample contains a time-series response from each sensor in the electronic nose, depicted in Figure 1. However, the method proposed in this paper is generic and could be extended to other ontologies as well as to other types of sensor data.

This paper is structured as follows: In section 2 we address related works having ontological solutions

Alirezaie M. and Loutfi A..
Ontology Alignment for Classification of Low Level Sensor Data.
DOI: 10.5220/0004137400890097
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2012), pages 89-97
ISBN: 978-989-8565-30-3

Table 1: Bacteria Species.

Code	Bacteria Species Name	Short Name
1	Escherichia coli	EColi
2	Pseudomonas aeruginosa	PSAER
3	Staphylococcus aureus	STA
4	Klebsiella oxytoca	KLOXY
5	Proteus mirabilis	PRMIR
6	Entercoccus faecalis	ENTFL
7	Staphylococcus lugdunensis	STLUG
8	Pasteurella multocida	PASMU
9	Steptococcus pyogenes	STRPY
10	Hemophilus influenzae	HINFL

for different semantic gaps problems. The next section concentartes on details of the methodology. After that, in section 4, our data set structure along with a short description about sampling process will be discussed. Then, section 5 represents results of each step of the methodology. The paper ends with discussion and conclusion.

2 RELATED WORKS

In order to empower results of signal level data analysis, several works with data integration approaches have been used. Multisensor data fusion is known as one of the most important effort in low level data processing. The main point of these works is keeping the synchronization among low level data that comes from different sources observing same or related phenomena (Joshi and Sanderson, 1999). In this paper, our approach concerns fusion of information at different levels of abstraction rather than from different sources. In particular, we are concerned with bridging a semantic gap which occurs between these levels (Ehrig, 2007).

Integrating knowledge bases into architectures of multi sensor fusion systems is known as a further step in low level sensor data processing. Some works such as (Yuguang et al., 2008) tried to find common concepts related to an object expected to be recognized by sensors for a better object identification and processing. In some other works similar to (Melchert et al., 2007), knowledge representation for reasoning on data fusion is considered to improve results of anchoring defined as symbol-perception connections for physical objects observed by sensors. While these methods work well for sensor data representing information about objects, they have yet to be extended to cope with time series sensor data.

In works which utilize concepts in the form of high level knowledge for sensor level data annotation, some focus on ontologies as their knowledge representation and reasoning framework (Chen, 2010). Ontologies make it possible to reuse existing knowledge available about measuring data in order to achieve an annotated data set which is essential for a more meaningful processing result. For example (Zhang et al., 2002) tried to induce a new decision tree as a classifier from an updated data set by including new related concepts to the feature set from ontologies. Likewise, in (Bouza et al., 2008) by restructuring data based on concepts extracted from ontologies of the features of data, a recommender system equiped with decision rules in different levels of abstraction has been developed. In these works, features measured by sensors have intelligible meanings with themselves so that their integration with other kinds of data or high level concepts can provide some outstanding improvements in outputs. Alignment, defined as the process of determining correspondences between concepts (Euzenat and Shvaiko, 2007), is mostly used when two sides of the process are ontologies. However, in this work, we map an ontology with the decision tree according to the names of bacteria assigned to different categories in these structures.

3 METHODOLOGY

The methodology used in this work applies the following steps:

- Classifying pre-processed sensor data using the C4.5 algorithm
- Localizing misclassified cases in the output of the classifier
- Aligning the classifier and the ontology to find similar parts between the two structures
- Replacing candidate parts of the ontology with their counterparts in the classifier

3.1 Classification of Sensor Data

A decision tree classifier is used to classify the output from the electronic nose. The decision tree has the advantage that it provides transparency in the representation of the outputs (Quinlan, 1993) and has a suitable structure for the alignment process.

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 (Quinlan, 1993).

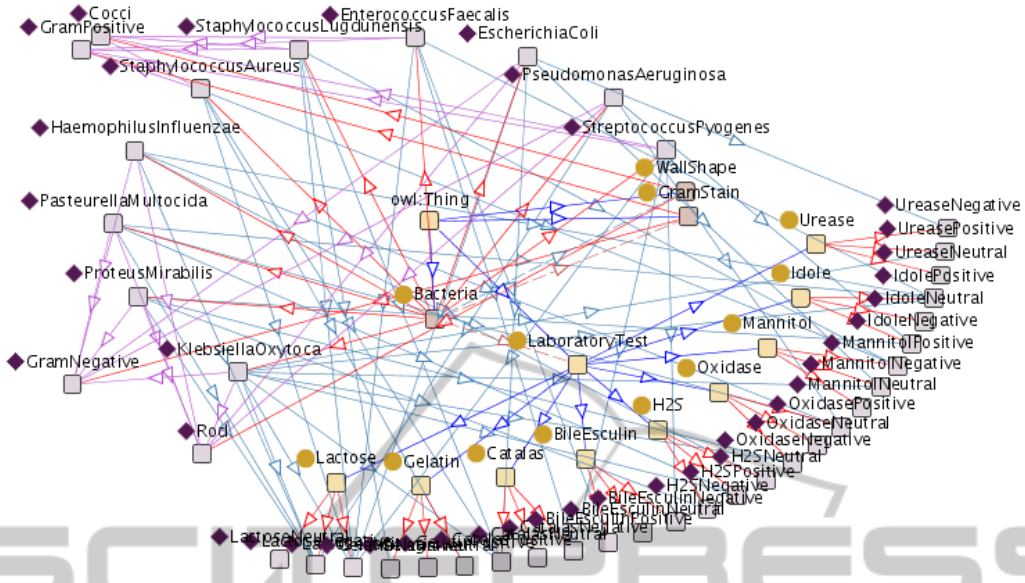

Figure 2: The Bacteria Laboratory Test Ontology.

Using the confusion matrix from the classification result, another process finds misclassification positions among leaf nodes of the tree and assigns them all bacteria names sharing these nodes. This process divides leaves of the tree into two groups A and B where group A contains nodes without misclassification; and the rest of leaf nodes belong to the group B. After this division, a java class runs a sibling checking process for each leaf node of group B. If the sibling also belongs to group B, the process labels the common parent node by all bacteria names shared by its children. If however, the sibling is a member of the group A, the process relabels the candidate leaf node by all bacteria sharing this node (true positive and false positive cases). Algorithm 1 and 2 show details of decision tree relabeling process. Once all nodes in group B or in the parents of group B are labeled, the alignment process begins and searches through the ontology in order to propose a laboratory test which is discriminatory among bacteria sharing the node. Eventually, if the process encounters a leaf node belonging to the group A, it leaves it without any replacement as these nodes are well classified.

3.2 Bacteria Laboratory Test Ontology

The ontology depicted in Figure 2 totally includes 27 classes among which 8 classes, such as *Bacteria*, *GramStain*, *LaboratoryTest* and *WallShape*, are directly subsumed by the *thing* super class. It also contains information about results (positive or negative) of 15 laboratory tests related to bacteria species. Moreover, this ontology provides information about the physical and chemical properties of bacteria cell walls (Gram Positive and Gram Negative) as well as their cell wall shapes (Cocci and Rods shapes) (Seltmann and Holst, 2002). Furthermore, 9 properties undergo relationships through this class hierarchy. For example, the domain of the *hasLaboratoryTest* property is *Bacteria* and its range is *GramStain* class; likewise, the property *hasWallShape* makes a relation between a sub class of *Bacteria* and the *WallShape* class. This information was collected from (ARUP, 2006) and then modeled in a RDF ontology via Protege 3.4.4 framework. Since we aim to launch some parts of this ontology to the classifier implemented in Java, a Java interface using Jena API converts the RDF file into a Java class which is able to be queried by ARQ- the Java RDF query engine.

3.3 Alignment Step

In this work, we align the classifier to the ontology. One could consider an alternative approach in which first the classifier is converted to an ontology and two ontologies are aligned to each other. However, our work considers a classifier which is intended to be online and incremental and therefore conversion of the classifier to an ontology would require an additional step each time the classifier is re-trained.

To perform the alignment between the decision tree and the ontology, we concentrate on finding similarities between two entities: the different categories of bacteria directly assigned to each node in the decision tree, and the information within class nodes in the ontology. To do this, mixing terminological

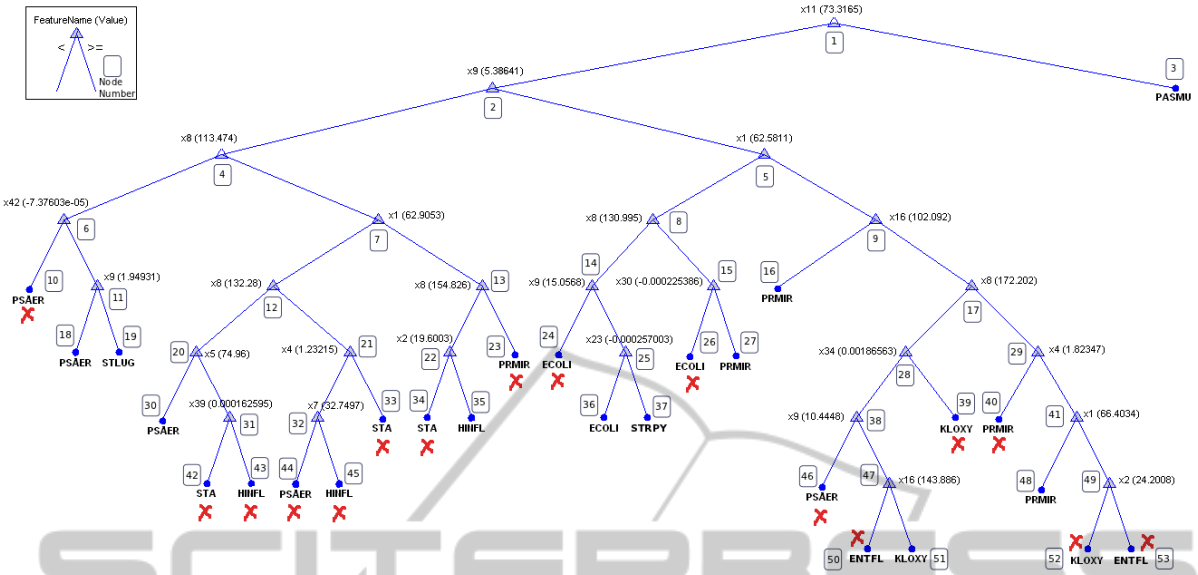

Figure 3: The Decision Tree Classifier-Group B nodes are labeled by red signs.

Algorithm 1: Relabeling DecisionTree.

```
1:  procedure RELABELDTREE(tree)
2:      groupB ← GetMisclasifiedLeaves(tree)
3:      for all l in groupB do
4:          l.Labels ← GetLabels(l)
5:      end for
6:      Sort(groupB)          ▷ Descending based on nodeID
7:      for all l in groupB do
8:          if ¬l.Checked then
9:              l.Checked ← true
10:             s ← GetSibling(l)
11:             if isLeaf(s) then
12:                 if member(s, groupB) then
13:                     if ¬CheckParent(l, s) then
14:                         s.Checked ← true
15:                         s.Replace ← true
16:                         l.Replace ← true
17:                     end if
18:                 else
19:                     l.Replace ← true
20:                 end if
21:             else
22:                 if s.Checked then
23:                     if ¬CheckParent(l, s) then
24:                         l.Replace ← true
25:                     end if
26:                 else
27:                     l.Replace ← true
28:                 end if
29:             end if
30:         end if
31:     end for
32: end procedure
```

Algorithm 2: Check a Parent Node.

```
1:  function CHECKPARENT(l, s)
2:      if hasCommon(l.Labels, s.Labels) then
3:          common ← true
4:          p ← GetParent(l, s)
5:          p.Labels ← GetLabels(l, s)
6:          s.Checked ← true
7:          p.Replace ← true
8:      else
9:          common ← false
10:     end if
11:     return common
12: end function
```

and structural alignment methods is used (Ehrig, 2007)[1]. The Jaro-Winkler algorithm (Jaro, 1989) finds the most similar name for a selected bacteria in the decision tree from the ontology. This algorithm works based on Jaro-Winkler distance (Formula 1) and counts the number of same characters in two strings by considering their positions to measure the distance between them. The higher the JaroWinkler value is, the more similar the strings (bacteria names) are (Jaro, 1989).

$$distance = \frac{1}{3} \times \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) \qquad (1)$$

Where:

m: number of matching characters.

[1]If the data set is rich enough, semantics should also be considered in the alignment process (Ehrig, 2007). However, our alignment methods are verifiable via the classification.

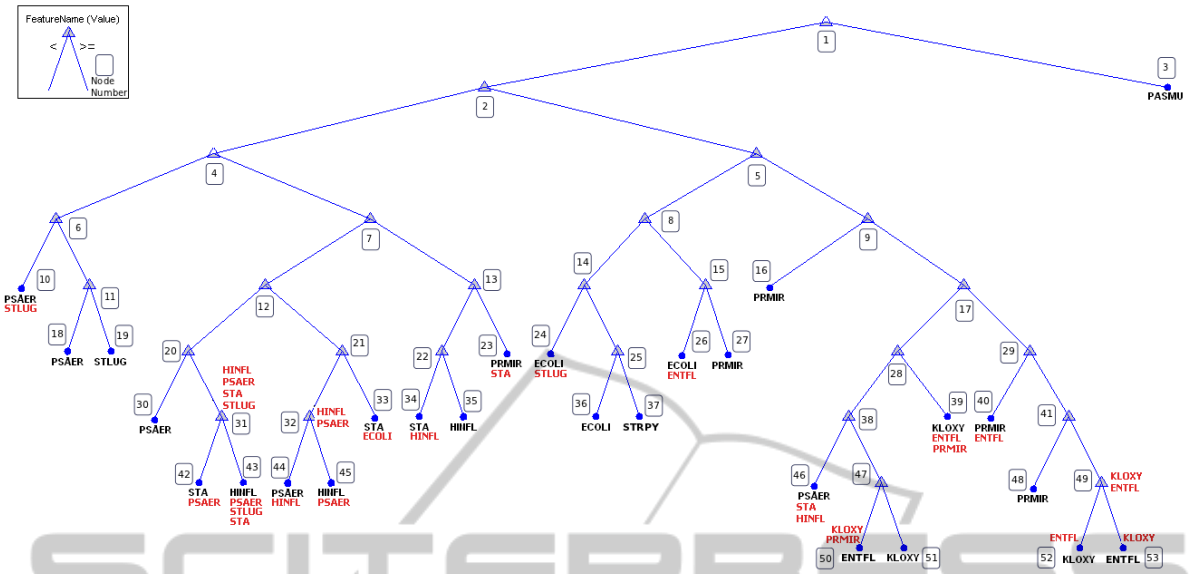

Figure 4: Relabeling Decision Tree By Bacteria Names. (Algorithm 1).

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

The graph inexact matching is used for the structural alignment as there is no isomorphism between the decision tree and the ontology (Hlaoui, 2002). Indeed, by using this kind of graph matching, the alignment process focuses on existing relations between labeled nodes in the decision tree to find similar subsumption relations in the ontology.

After finding the similarities, a replacement process transfers information from the ontology to the classifier in order to provide an annotated decision tree that contains two kinds of information, sensor values and laboratory tests. The algorithmic details of this process are represented in Algorithm 3 and 4. The resulting decision tree is a classfier for those cases mapped to leaf nodes of group A and as a recommender for group B nodes.

Algorithm 3: Alignment (D-Tree and Ontology)

```
1: procedure ALIGNMENT(tree, ontology)
2:     Sort(tree)                          ▷ Ascending based on nodeID
3:     for all tn in tree do
4:         if tn.Replace then
5:             c ← GetSimilar(tn, ontology)
6:             Replace(tn, c, tree)
7:         end if
8:     end for
9: end procedure
```

Algorithm 4: Finding Similar Parts in the Ontology.

```
1: function GETSIMILAR(tn, ontology)
2:     min ← ABigNumber
3:     n ← GetLabelsNumber(tn.Labels)
4:     ontoNodes ← GetNodes(ontology, n)        ▷ same number of labels
5:     for all on in ontoNodes do
6:         for i ← 1, n do
7:             for j ← 1, n do
8:                 d[i][j] ← JaroWinkler(tn.Labels[i], on.Labels[j])
9:             end for
10:        end for
11:        on.distance ← GetSumBestMinimumSet(d)
12:        if on.distance ≤ min then
13:            min ← on.distance
14:            candidate ← on
15:        end if
16:    end for
17:    return candidate
18: end function
```

4 DATASET

The clinical samples in this scenario are 10 types of bacteria species listed in Table 1, sub-cultured on blood agar plates and a bacterial suspension solution. Further details of the sampling process and prepration are given in (Trincavelli et al., 2010). Each sampling cycle is 5 minutes and contains three phases (Figure 1). The first phase is called the baseline acquisition and lasts for 10 seconds. In this phase sensors are exposed to a reference gas which is air in this experiment. Next, the headspace[2] gases are injected into

[2]The headspace is the space just above the liquid sample in a bottle (Pearce et al., 2003)

the sensor chambers and sensors are exposed for 30 seconds. The last phase is a recovery phase of 260 seconds to recover sensors for the next round of testing by flushing the sensors with the reference gas. Each of the 10 bacteria has been sampled 60 times.

To make a more suitable structured training set for the classification, we need to pass sensor readings which are continuous time series data generated by 22 sensors through a pre processing phase that includes two steps. Baseline manipulation and compression normalizes the sensor data according to the baseline phase (Pearce et al., 2003) and extracts informative descriptors of signals to make feature vectors (Pearce et al., 2003), respectively.

We use two descriptors indicated in Figure 1 for each signal: The static response calculating the difference between end of the sampling phase and baseline gives one single parameter; and the response derivative which is equal to the slope of the line contiguous to that segment of the signal related to the first three seconds of the sampling phase. Eventually, we produce 44 feature values for the dataset of 600 samples accompanied by a label list containing bacteria species names listed in the third column of Table 1.

5 RESULTS

A 10-fold cross validation is applied on the data set to generalize the error estimation of the classification (Bishop, 2006). In this process, two thirds (400 cases) of samples in the data set were assignd to the training set and the remainder were used as testing set.

Figure 3 shows the result of the classification fed by the training set. Decision nodes of the tree are labeled by feature names and criteria values. Leaf nodes of the tree are also marked by bacteria species names. The confusion matrix of this classification is depicted in Figure 5. According to this matrix and Formula 2, among the 200 test cases there are 39 misclassifications corresponding to an accuracy of 80%.

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

Figure 5: Classififcation Confusion Matrix.

$$accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (2)$$

The misclassified nodes are shown in Figure 3 and these form the B group nodes. Table 2 also shows more information about leaf nodes of group B. For example in the first row of the table two types of bacteria are sharing node number 10, predicted type 2 (according to the training set) and actual type 7 (according to the test set). Likewise, the 11th, 12th and 13th rows illustrate the details of node number 43 which is shared by 4 kinds of bacteria, bacteria type 10, 2, 3 and 7. To resolve these inconsistencies between predicted and actual bacteria types, we utilize the ontology suggestions related to the laboratory tests and update our decision tree based on the ontology offers to make the classifier to a recommender system. As mentioned above, the alignment process uses terminological and structural methods to find similarities between two structures. To visually make more sense about the structural matching process, Figure 4 depicts the information of Table 2 directly on the decision tree.

Table 2: B-leaf Nodes.

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

The alignment process finds bacteria names sharing a leaf node belonging to group B. For example, the sub tree containing node 49 as the parent and nodes 52 and 53 as children belonging to group B (Figure 4), are sharing bacteria number 4 (Klebsiella Oxytoca or KLOXY) and 6 (Entercoccus faecalis or ENTFL). By the string matching method, the alignment process finds all bacteria names in the ontology that are similar to the candidates. Table 3 demonstrates some parts of Jaro-Winkler distances between bacteria names in the decision tree 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 bacteria names mapping. The graph matching method

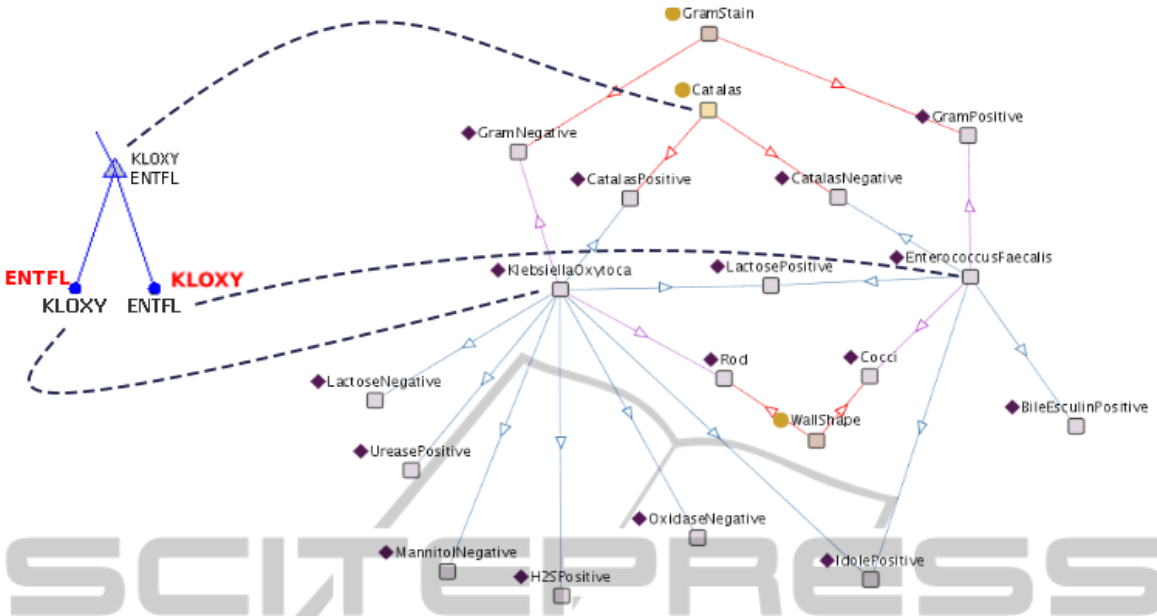

Figure 6: Alignment Process (Between Candidate Sub-Tree and Matched Sub-Ontology).

Table 3: Jaro-Winkler distances of bacteria names between the decision tree and the ontology. (minimum value of each column is in red).

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

then extracts the most similar structure to this part of the sub tree depicted in Figure 6. The laboratory tests candidates for parent of node 52 and 53 sharing KLOXY and ENTFL are Catalas, Mannitol Fermentation, Urease and Methyl Red. However, the cost and duration issues of laboratory tests considered in the designing phase of the ontology cause the ontology to suggest Catalas test which has negative response for ENTFL and positive answer for KLOXY. Therefore, now the ontology suggestion can be replaced by the sub tree holding information about these leaf nodes that contain some uncertainties about the bacteria types. By applying the alignment process on the whole nodes in group B, we will finally have an annotated decision tree demonstrated in Figure 7.

6 CONCLUSIONS

In this work, we implemented an ontological methodology to improve classification results of electronic nose sensors readings. High level information coming from the ontology facilitate decision making and help to compensate the ambiguity existing in some responses of the decision tree as the classifier of bacteria types.

Indeed, using the bacteria laboratory tests alone, the ontology may suggest about 6 different laboratory tests for identification of these 10 types of bacteria (ARUP, 2006). On the other hand, the classification from the electronic nose does not have a sufficently precise response for medical staffs who may offer different kinds of treatment based on the bacteria type detected in a blood sample. By mixing a low level

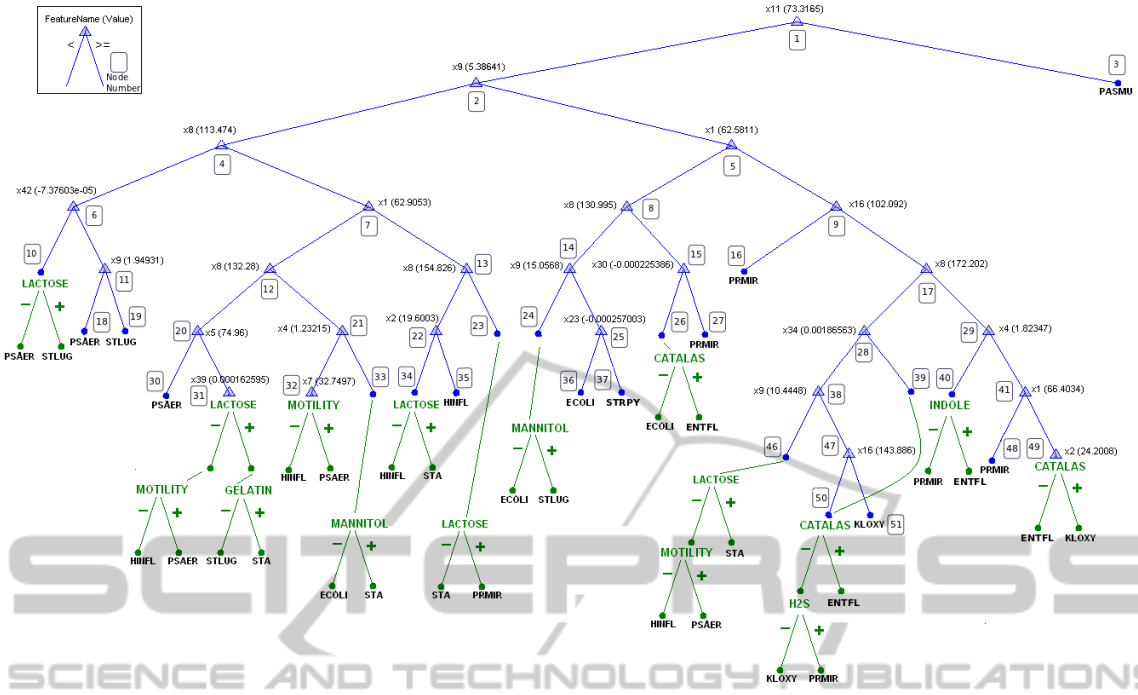

Figure 7: Annotated Decision Tree by Laboratory Test Information.

sensor data structured in a decision tree and high level knowledge in the form of an ontology we could bridge the gap existing between these two kinds of data and reach to a system that has a classifier-recommender role in the identification process. Instead of offering many numbers of laboratory tests, it offers 2 tests to resolve the ambiguities in worst cases (Figure 7).

As the details of the methodology show, there is yet a quantitative evaluation for the system and the future work will focus on obtaining the feedback from experts of the domain where they encounter some suggestions about laboratory tests given by the system. From a technical point of view, the system could be further generalized in the many regards such as the classification and/or the automatic acquisition of knowledge. For this reason, we have created a method using ontologies to allow for this generalization and future work will make full use of the ontological structure. Regarding the classification, decision trees were used due to their hierarchical structures making more sense in the process of alignment. However, a more ideal alignment model is one that is independent from the classifier structure. Moreover, modeling available knowledge into ontologies was accomplished manually. We will address the possibility to use automatic knowledge acquisition in order to find knowledge form disparate and perhaps unstructured contexts by more powerful matching methods to which we can apply some OAEI[3] algorithms in order to assess the whole alignment process.

REFERENCES

ARUP (2006). A national clinical and anatomic pathology reference laboratory. www.aruplab.com.

Bishop, C. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition.

Bouza, A., Reif, G., Bernstein, A., and Gall, H. (2008). Semtree: Ontology-based decision tree algorithm for recommender systems. In *International Semantic Web Conference (Posters & Demos)*.

Chen, Y. (2010). Development of a method for ontology-based empirical knowledge representation and reasoning. *Decision Support Systems*, 50(1):1–20.

Ehrig, M. (2007). *Ontology Alignment: Bridging the Semantic Gap*, volume 4 of *Semantic Web And Beyond Computing for Human Experience*. Springer.

Euzenat, J. and Shvaiko, P. (2007). *Ontology matching*. Springer-Verlag, Heidelberg (DE).

Hlaoui, A. (2002). A new algorithm for inexact graph matching. *Object recognition supported by user interaction for service robots*, 4(c):180–183.

Jaro, M. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Society*.

[3]Ontology Alignment Evaluation Initiative (http://oaei.ontologymatching.org/)

Joshi, R. and Sanderson, A. (1999). *Multisensor Fusion: A Minimal Representation Framework*. Series in Intelligent Control and Intelligent Automation. World Scientific.

Melchert, J., Coradeschi, S., and Loutfi, A. (2007). Knowledge representation and reasoning for perceptual anchoring. *Tools with Artificial Intelligence*.

Pearce, T., Schiffman, S., Nagle, H., and Gardner, J. (2003). *Handbook of machine olfaction: electronic nose technology*. Wiley-VCH.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, C.

Seltmann, G. and Holst, O. (2002). *The Bacterial Cell Wall*. Springer-Verlag.

Trincavelli, M., Coradeschi, S., Lout, A., Sderquist, B., and Thunberg, P. (2010). Direct identication of bacteria in blood culture samples using an electronic nose. *IEEE Trans Biomedical Engineering*.

Yuguang, N., Gaowei, Y., Gang, X., Zehua, C., and Keming, X. (2008). Multi-sensor fusion using knowledge-based mind evolutionary algorithm. *Convergence and Hybrid Information Technology*.

Zhang, J., Silvescu, A., and Honavar, V. (2002). Ontology-driven induction of decision trees at multiple levels of abstraction. In *In Proceedings of Symposium on Abstraction, Reformulation, and Approximation 2002*. Springer-Verlag.