A Bounded Neural Network for Open Set Recognition

Douglas O. Cardoso and Felipe França
Universidade Federal do Rio de Janeiro
PESC-COPPE
Rio de Janeiro, RJ, Brazil

João Gama
Universidade do Porto
LIAAD-INESC
Oporto, Portugal

***Abstract*—Open set recognition is, more than an interesting research subject, a component of various machine learning applications which is sometimes neglected: it is not unusual the existence of learning systems developed on the top of closed-set assumptions, ignoring the error risk involved in a prediction. This risk is strictly related to the location in feature space where the prediction has to be made, compared to the location of the training data: the more distant the training observations are, less is known, higher is the risk. Proper handling of this risk can be necessary in various situation where classification and its variants are employed. This paper presents an approach to open set recognition based on an elaborate distance-like computation provided by a weightless neural network model. The results obtained in the proposed test scenarios are quite interesting, placing the proposed method among the current best ones.**

I. Introduction

Classification is one of the most basic activity of machine learning. Generally, this activity is based on knowledge extracted of a data sample which ideally should reflect most characteristics of a population. Therefore, it can be expected that all observations classes have representatives in this sample. However, this is not required by all classification-like tasks.

For example, fault and anomaly detection [1], [2] can be done relying solely on modelling the default behavior, that considered normal. This is a practical situation where one-class classifiers [3] can be successfully used. An interesting feature of this approach is the capacity to robustly identify unprecedented abnormal behavior just considering how different it is of a known standard: the learning system stays up-to-date as far as the regular behavior remains unchanged.

Also in some multi-class classification tasks the ability to label observations as elements of an 'unknown' class, instead of any of those previously seen, can be very useful. That is the case of human activity recognition [4], [5], where the number of targeted activities is usually smaller than the number of elements in the universe of possibilities [6]. A trend in this area is semi-supervised learning [7], [8]: knowledge is obtained from a set of labelled data combined with a larger unlabelled sample. Depending on how data was collected, it is unreasonable to assume that each unlabelled observation is a member of one the classes represented in the labelled sample.

From the area of computer vision we bring an interesting concept which formalizes the subject addressed by this paper. Open set recognition [9] can be viewed as more challenging version of classification: "incomplete knowledge of the world is present at training time, and unknown classes can be submitted to an algorithm during testing". It is not difficult to realize that similar conditions occurs in machine learning subareas other than computer vision. The examples described in the preceding paragraphs can be seen as some of these cases.

In this paper a methodology to open set recognition is presented. Using a classifier which rates the similarity between unknown data and the stored knowledge, an estimation of this rate to observations not belonging to any known class is computed from the training sample. This estimate is used to define the boundaries between observations belonging to the known classes and the probable outliers. This similarity rating resembles the posterior probability, but do not rely on assumptions about the prior probability distribution of the classes, which is usually unavailable in open-set tasks. The performance of the proposed approach was evaluated in a diverse collection of experiments and it looks promising: its results, when not superior, were at least as good as the compared alternatives in all tested scenarios, what supports its general use.

This is how the remainder of this paper is organized. First, section II presents an introductory overview of open set recognition. Section III introduces the methodology developed in this research. This methodology was evaluated through a group of tasks from distinct domains. The description of these tests and the results obtained are presented in section IV. Section V brings concluding remarks and future works ideas.

II. Background

Considering the default notion of classification, the target of tasks of this type is to identify to which class, among a collection of previously modelled ones, an observation of an unknown class best fits. The idea of dividing the feature space using hypersurfaces defined from training data, so that each region is associated to a class and an observation is classified according to the region where it lies, is undoubtfully valid and useful: discriminative models for learning, as the Support Vector Machine (SVM) and the Multilayer Perceptron, work based on this principle.

However, it is remarkable that these regions are potentially infinite (as the half-spaces defined by a linear-kernel SVM, for example), despite being established according to observations in a finite range of the feature space. To assign a class to an observation that could be considered far away from the training data has its own risk, but it is necessary for the sake

978-1-4799-1959-8/15/$31.00

of classification: a decision about the best fitting class has to be made; therefore, an "educated guess" is to use the already defined regions of the feature space for this purpose.

An interesting question regarding this risk is the following: can it be estimated? In other words, is it possible to assign a confidence level to a prediction provided by one of these discriminative models? Consider an hypothetical binary classification task, wherein the feature space is $\mathbb{R}^n$. During training, these models define a function $f : \mathbb{R}^n \to \mathbb{R}$, which aims to map an observation to a value which represents its class: the most common practice is to relate the values -1 and +1 to the classes and consider an unlabelled observation x an element of the class $sgn(f(x))$. The space-dividing hypersurface, commonly called *decision boundary*, is defined by $f(x) = 0$. This function f indeed computes a signed distance between x the decision boundary. Therefore, to use $f(x)$ seems a straightforward way to compute confidence level of a prediction $sgn(f(x))$.

But a weakness in this idea can be noticed by the following reasoning: let x and x' be unlabeled observations; let the function $\delta : \mathbb{R}^n \to \mathbb{R}$ return the average distance between a given observation and the observations in the training set; if $f(x) = f(x')$ but $\delta(x) \gg \delta(x')$, is it sensible to be equally sure about the classification of both points? The decision about x appears to be more difficult, as it concerns a region of the space which is less known, but nonetheless is divided by the decision boundary defined according to the available training data. Thus, the function f alone may not provide a good estimation of the aforementioned risk.

Generative learning models, opposed to the discriminative ones, attempt not to define the boundaries between classes, but to model the distribution of the classes in the feature space. In probabilistic terms, for any observation x and class y, models of the first type target to define the joint probability $P(x, y)$; the ones of the last type try to approximate $\text{argmax}_y P(y|x)$ using $f(x)$. The confidence level of a prediction made by a generative classifier can consistently depend solely on $P(x, y)$. In this aspect, generative models can be considered superior to discriminative ones: trying to establish a complete probabilistic model of the data, these models can natively provide an estimation of the risk involved with every classification.

Open set recognition is, in a rough description, risk-aware classification: any observation is labeled as the class it best fits if the uncertainty of this prediction is considered acceptable; otherwise, the observation is labeled as belonging to an *unknown* class. During training, the learner should not only model the classes represented in the training set but also discover how much uncertainty should be considered acceptable: if too much tolerant, the system behaves as a classical, closed-set classifier, considering any observation a member of one of the known classes; but if excessively cautious, it has a greater chance of misidentifying observations from the modelled classes as elements of an unknown one.

Some approaches were previously considered for tasks of this kind. A large part of them is based on discriminative classification principles: variants of SVM capable of rejecting (i.e. ruling as outliers) observations [10]–[14] and ensembles of one-class classifiers based on support vectors [15]–[17] are arguably the most common descriptions of the methods recently tested in this regard. This can be considered a natural consequence of the unquestionable competence and flexibility of SVM for closed-set classification, verified in huge variety of applications. However, for open set recognition, a solution with a generative background could fit in more naturally thanks to its embedded confidence estimation, what contrasts with the adaptations made on discriminative models. In some sense, a solution of this kind bounds its decisions to its area of expertise.

Despite this, the use of generative method also prompts some challenges [13], [18]: a precise probability density estimation requires a large and noise-free training set, what is not always available; the prior probability distribution of the classes is generally unknown in open set tasks, what disallow the ordinary use of Bayes' theorem to pick the most probable class of a given observation. A proposed countermeasure to these issues is the use of a model based on the distance between observations to be recognized and class-related prototypes as the centers resulting from k-means clustering, for example.

The WiSARD [19], a n-tuple weightless neural network model, classifies observations rating how well they fit to the known classes. That is, fitness scores regarding each class are assigned to a given observation, and the predicted class is the one with the highest score. The knowledge of each class is kept apart in structures called *discriminators*, which provide these scores. A discriminator works as a complex "distance" computing unit: during training it stores binary features extracted from observation of the class it represents; the similarity between an observation and the knowledge maintained by the discriminator is rated according to the number of binary features extracted from this observation which match those features previously stored.

Although using structures named discriminators, WiSARD has generative capabilities. In one of its most used setups, instead of simply storing features obtained during training, it counts the frequency of occurrence of each feature. These counts can be used to increase the resolution of the fitness scores, allowing more insightful comparisons of them, but also to extract prototypes of the stored knowledge through a procedure called DRASiW [20], a reference to the reversal of the ordinary model operation. Thus, instead of fitting a mathematical model to define boundaries between modelled classes, this neural network model learns explicitly collecting characteristics of the training data for later use.

The work presented in this paper investigated the development a rejection-capable WiSARD. This could be achieved by thresholding the similarity rates provided by the discriminators. A scheme using individual threshold for each class allows greater flexibility to handle unbalanced and noisy data sets [21]. The proposed procedure to define these thresholds is presented next.

III. Estimation of similarity rates

This work presents a modification of the original WiSARD, here identified as tWiSARD. Before training, tWiSARD computes score thresholds for each class. These thresholds are used to rule if an observation being classified belongs to the highest scoring class or if this score is not high enough to

conclude this. In other words, if the best score is below the established threshold this observation is considered an outlier to all known classes. The procedure used to calculate the thresholds is detailed next.

1: Let T be the training sample
2: Let T_y be the observations of class y in T
3: Let FITSCORE(x, Δ) be the fitness score of an observation x to the knowledge a discriminator Δ possess
4: Let PREDICTION$(A, B, f) = \forall_{x \in A}(x \in B \Leftrightarrow f(x))$ be a conditional statement of the pertinence to B of the elements of A, according to f, a boolean function
5: Let EVAL(Z) be a numerical evaluation of a prediction Z, as accuracy, precision or F_1 score, for example
6: **procedure** CLASSTHRESHOLD(class y, # folds k)
7: Let $P = \{p_1, \ldots, p_k\}$ be a partition of T_y in k sets
8: **for all** $p_i \in P$ **do**
9: Train a discriminator Δ with $T_y \setminus p_i$
10: **for all** observation $x \in p_i$ **do**
11: $s_x \leftarrow$ FITSCORE(x, Δ)
12: **end for**
13: **end for**
14: Train a discriminator Δ with T_y
15: **for all** observation $x \in T \setminus T_y$ **do**
16: $s_x \leftarrow$ FITSCORE(x, Δ)
17: **end for**
18: Let $S = \{v : \exists_{x \in T}(s_x = v)\}$ be the set of score values
19: **return** $\underset{v \in S}{\operatorname{argmax}}$ EVAL(PREDICTION$(T, T_y, s_x \geq v)$)
20: **end procedure**

This procedure results from the natural expectation that a discriminator trained with observations of a class y provides higher similarity scores to other observations of y than to observations of other classes. Ideally, there would be a score value which establishes a dichotomy: the scores of all observations of y are greater than this value, and the observations of other classes have scores lower than this value. As this optimal value is usually nonexistent, find a suboptimal value to get as close as possible to this dichotomy is desired.

The estimation of optimal thresholds for rejection was previously assessed [21]. But the procedure herein introduced deals with an aspect not fully analyzed before: to consider the use of different measures to evaluate the possible thresholds. That is, instead of relying on accuracy, any method to rate the quality of a prediction as precision, F_1 score and others can be employed. This is useful, as an example, for tasks in which the distribution of the classes is unbalanced: accuracy, the usual default measure, is not reliable in this kind of situation; a popular alternative is the F_1 score. Still regarding the use of F_1 score, it could be substituted by a F_β score [22], so that instead of considering precision and recall equally important, this could be tuned according to task-specific targets: for example, reducing the rejection rate by the increase of recall weight in the calculation of F_β score.

The use of k-fold cross-validation (the loop starting at line 8) intends not only to allow a more robust estimation of the fitness scores compared to that obtained using a fixed validation set. System speed is also a critical in some applications. Therefore, the number of folds to be used can be adjusted also considering this matter. During the experiments the influence of the variation this parameter in the overall performance of tWiSARD was evaluated.

IV. EXPERIMENTAL EVALUATION

Through this section a collection of tests with an open-set orientation is detailed together with their results. As an indication of the existence of open-set tasks in various knowledge domains, these tests also have different backgrounds, but share some similar requirements. The experiments were developed in Python, and used the Scikit-learn [23] module which provides implementations of SVM and performance metrics employed. SVM default parameter values were used. WiSARD and its derivatives were implemented by the authors, and shared the same parameter configuration. Their parameters were not defined with extreme rigour, but relying on previous experience regarding their setup. The configurations used are detailed in the descriptions of the tasks.

A. Anomaly detection

The most simple tasks which allow an open-set interpretation are anomaly detection problems: binary classification tasks having a default class, in which data needs to be deemed as belonging to the standard or not. The challenge provided by the 'DGA' data set [1] allows this modelling: power transformers should be classified as 'faulty', failure-prone; or 'normal', the default behavior. Observations consists of concentrations of 7 gases dissolved in transformers insulation oil.

For WiSARD, each attribute value was quantized and converted to a 120-bit unary representation. Thus, each observation was converted to a 840-bit string. The length of the addresses used to access WiSARD RAM nodes was set to $\lfloor \log_2 840 \rfloor = 9$.

This is a small data set, composed of 50 'normal' observations and 117 'faulty' ones. To estimate robust rejection thresholds in this situation is challenging, as some characteristics of the population may not be represented in the training sample. On the other hand, it is possible to use methods which are more computationally expensive to process this sample because of its size. Therefore, a leave-one-out cross-validation procedure was performed by CLASSTHRESHOLD by appropriate setup of the parameter k.

The following experiment was performed on this data set. First, the 'faulty' observations were clustered into 10 groups. Assuming the existence of different kinds of faults, this step aims to separate data accordingly. Now consider all proper, non-empty subsets of these 10 clusters. The number of such subsets is $2^{10} - 1$. Each of these subsets was used to define a train-test split of the data set: the training sample was composed of the observations in the aforementioned subset plus 80% of the observations of the 'normal' class, randomly picked; the remaining observations composed the test sample. Contrasted to a default k-fold cross-validation setup, this train-test layout brings a different real-world challenge: there is no

guarantee that all kinds of faults would be covered in the training sample.

The classification performance of SVM and WiSARD as binary classifiers was compared to that of tWiSARD as a one-class classifier. The performances should improve as more information about the faults is provided, what also enables a clearer definition of the standard behavior. Thus, the results regarding each train-test split were grouped and averaged according to the respective number of fault clusters in the training set. For greater statistical soundness, the full procedure was repeated with different clusterings of faulty observations. The performances of the classifiers in a default 5-fold cross-validation task were also reported, for the sake of comparison.

Figure 1 shows results of this experiment. SVM and WiSARD had the superior performances in the 5-fold cross-validation test, but poorer results as the number of fault clusters in the training sample diminished. Of the three tested options, tWiSARD had the best results: 8192 train-test runs were realized; the average F_1 score of SVM, WiSARD and tWiSARD was respectively 80%, 84% and 90%; according to Wilcoxon signed-ranks tests [24], [25] regarding each pair of classifiers, the null-hypothesis that two of them perform equally well is rejected with a significance level (α) of 0.001. Therefore, according to some prior domain knowledge as the probability of emergence of unprecedented anomalies, a one-class learner should be preferred over the binary classifiers.

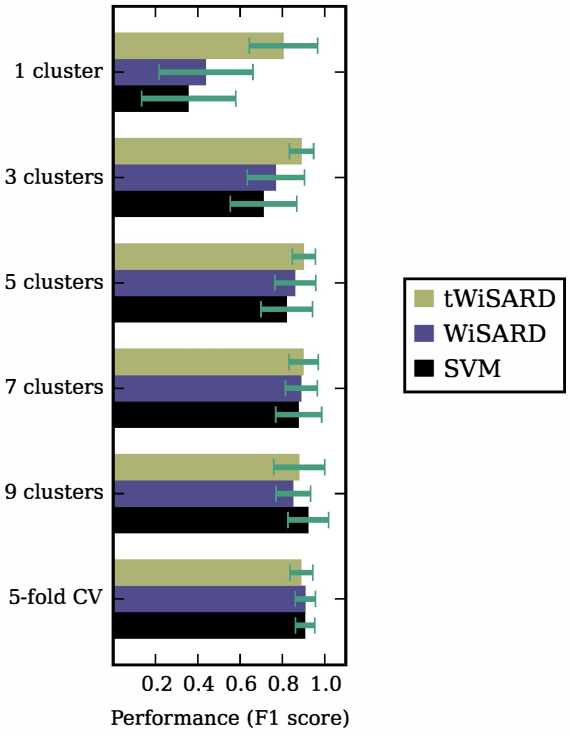

Figure 1. Results of the experiment with the 'DGA' data set. The first five bar groups regard the tests with the clustered 'faulty' observations.

As the other classifiers, tWiSARD also benefits from the availability of more diverse data, which allows a better definition of the rejection thresholds: its worst performance was in the most challenging 1-cluster setup; its biggest performance variation was an increase of around 9% from the 1-cluster to the 3-cluster case. However, it is clear that the alternatives were more disadvantaged by the lack of data variety than tWiSARD.

B. *Human activity recognition*

A drawback in the use of the score thresholds is the chance to reject an observation which would be correctly classified if this methodology was not applied. Therefore, a decrease in the recall of the known classes could be expected when using this mechanism. However, this side effect is acceptable if the overall performance is good enough: the observations correctly rejected would have to make up for the recall decrease. This rejection-recall equilibrium was analyzed though the following situation: the performances of WiSARD, SVM and tWiSARD were compared when one of the data classes was left out of the training sample. Of the three options, only tWiSARD could rule observations as outliers, while the others classified all observations as items of the known classes.

The 'UCI-HAR' data set [4] is, quoting its authors, "an Activity Recognition database, built from the recordings of 30 subjects doing Activities of Daily Living (ADL) while carrying a waist-mounted smartphone with embedded inertial sensors". Each observation is a collection of 561 statistics of the sensor readings. However, in this work just a subset of 46 attributes was used: those related to the mean of the readings. The original training and test sets were used, so that each of the six classes of 'UCI-HAR' was omitted at a time from the training set. This scheme mimics a realistic human activity recognition task, where it is impossible to consider that all possible activities are known and modelled.

For the desired performance in this task, a greater number of bits was required to represent the attributes and, consequently, the observations. Each observation was represented by a total of 20000 bits, which were divided as evenly as possible between the attributes: approximately 435 bits were used to represent the quantized version of the value of each feature. The length of the RAM nodes addresses was set to 40.

For the 'DGA' experiment, all observations in the training sample were used to define the score threshold of the 'normal' class. A similar strategy was not optimal for the 'UCI-HAR' data set: when defining the threshold of a class y, using trivially distinguishable observations of other classes appeared to be harmful: the fitness score of these observations to class y is too low, what pushed the class threshold towards smallish, useless values. As a matter of fact, these trivially distinguishable observations of other classes would hardly be classified as elements of y, not requiring the use of the score threshold of y to avoid their misclassification.

To solve this question the following strategy was adopted: a 3-fold cross-validation test was performed on training sample; the threshold of a class y was calculated using observations which were classified, correctly or not, as members of y in this cross-validation test. The number of folds, k, of the ClassThreshold procedure also differed from the previous experiment: 'UCI-HAR' used $k = 3$, while 'DGA', which is much smaller than the former, set a leave-one-out scheme through this parameter.

The results regarding the 'UCI-HAR' data set are shown in Figure 2. Again, tWiSARD has the best results on average. The bad results for classes 'Sitting' and 'Standing' can be explained: these two classes are closely related. Thus, when one of them was omitted, most observations of the class left out were classified as belonging to the other. In the opposite direction, class 'Laying' is the most isolated of all classes. Consequently, when it was omitted, tWiSARD had its best performance while the performance of the other classifiers was the worst.

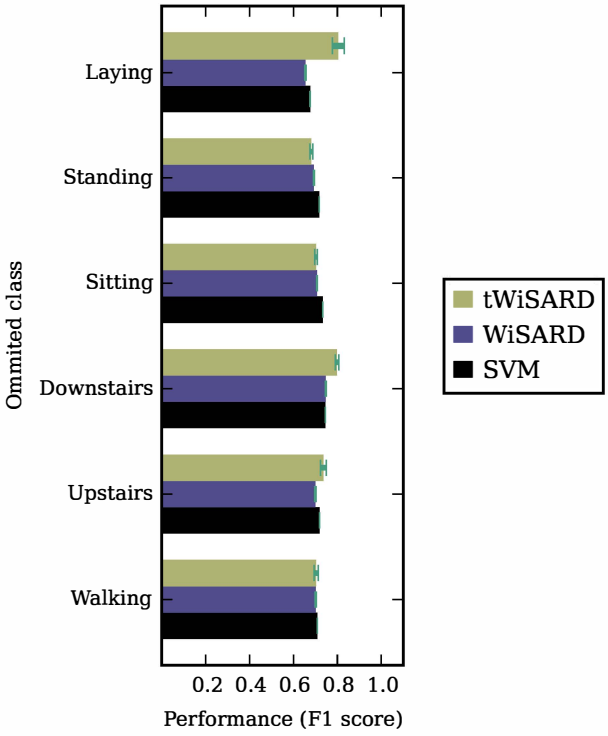

Figure 2. Performance on the 'UCI-HAR' data set, omitting each class at a time. The error bars are negligible as the average standard deviation of the measurements is below 0.005.

C. Handling higher openness

The concept of openness [9] was defined to provide a quantitative degree of complexity of open-set problems, according to the number of classes represented in the training sample compared to those to be handled during the effective use of the consolidated knowledge. Considering C_t, C_r and C_e, respectively, the number of classes with observations in the training sample, the number of classes which should be later recognized ($C_r \leq C_t$, as the training sample may contain "strictly" negative examples, not prompting the modelling of its class) and the number of classes to deal with during evaluation, this is how openness is defined:

$$\text{Openness} = 1 - \sqrt{\frac{2 \times C_t}{C_r + C_e}}$$

According to this definition, the openness of the 'DGA' task, considering each of the 10 fault clusters a different class and that just one of these clusters is in the training sample, is $1 - \sqrt{(2 \times 2)/(1 + 10)} \approx 40\%$. Although still challenging, 'UCI-HAR' task is almost closed: its openness is below 5%. This last task is an interesting benchmark, designed specifically for open set recognition, and openness over 63%.

The data[1] used in this experiment comes from two different image sets, Caltech 256 [26] and ImageNet [27]. The first was used to provide the training sample, while the test set was composed of positive observations of the first source and negative ones from the last, a cross-data set design which requires the proper rejection of observations from classes not targeted, independently of its origin. In each of 5 test rounds, 88 classes were randomly selected. Each of these 88 classes was used once as the one to be recognized, being represented in the training and test samples by 70 and 30 observations, respectively. The remainder of the training set were 70 (5×14) observations of 5 classes randomly chosen from the 87 negative classes. Still in this regard, the test set also had 5 observations from each of the 87 classes not targeted. Adding up, the training and test samples had 140 and 465 observations, respectively.

Each observation of this data set is described by 59 real-valued attributes. Each of those was represented using 512 bits, being converted to an unary value after quantization, similarly to what was done in the previous tasks. The RAM nodes were addressed using bit strings of length $\lfloor \log_2 59 \times 512 \rfloor = 14$.

In this test a larger number of alternatives was compared: a default SVM binary classifier; a rejection-capable SVM, which associated a probability to each prediction according to Platt scaling [28]; a one-class SVM; a default WiSARD binary classifier; a one-class WiSARD, relying on the threshold computed for the positive class only; a two-class tWiSARD; and a one-class P_I SVM [14], which represents the state of the art. The rejection-capable alternatives could opt not to label an observation when in doubt about its class, what is preferred over an uncontested mistake. The Platt-scaling-based SVM used this alternative whenever the respective prediction probability was under 0.5, while tWiSARD did the same when the top similarity score was lower than the threshold previously computed for the best fitting class. The one-class P_I SVM was chosen as the best performing alternative found. Its configuration was optimized, leading to the use its default setup except for $\gamma = 35$ and the rejection threshold $\delta = 30\%$.

Figure 3 presents the results of this test. It can be noticed that each rejection-capable approach performed at the same level or better than its default counterparts. In the case of WiSARD and tWiSARD, the superiority of the later is clear. It is also remarkable that even the one-class WiSARD had a good performance, the second best overall.

V. Conclusion

Dealing with unprecedented non-targeted classes is an important aspect of classification, indispensable in some real-world applications. This fact is sometimes ignored, leading to poor results when a learning system is deployed out of its testbed. The truth is that a great variety of computational applications has a degree of openness which needs proper

[1] http://www.metarecognition.com/openset/ (accessed 2015/04/21), LBP-like Features, Open Universe of 88 Classes

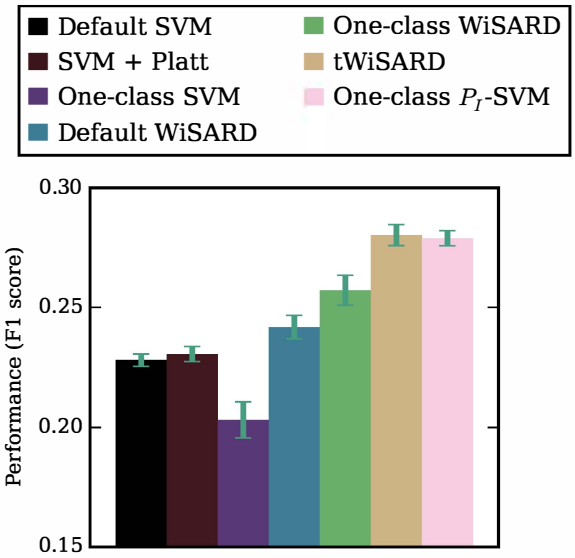

Figure 3. Performance on the cross-data set task.

handling. This paper discusses these points, in the context of examples of current machine learning challenges.

The developed methodology, although simple, is promising: its results in the proposed tests are insightful, highlighting some interesting characteristics of the data which did not emerge during the exclusive use of the classifiers to which the proposed approach was compared. The conception of this technique on the top of WiSARD boosts the use of this well-established learner in situations where it is necessary to define more strictly the boundaries inside which it is possible to make conscious decisions.

A possible continuation of this work is a study on the influence of parameters as the number of folds used to estimate the thresholds and also those of the WiSARD model: during this work, it was noticed some kind of relation between these items, so that it was possible to reduce the number of folds at the expense of defining a more complex neural network. It is possible that this equilibrium can be optimized, or tunned according to a desired system behavior: speed increase, or a slower training procedure targeting higher accuracy.

Another interesting follow-up would be to evaluate the application of the proposed technique to semi-supervised learning tasks. The unlabelled data in the training sample could include observations which are outliers for all targeted classes. A strategy to filter these outliers could use the score thresholds as one of its main building blocks.

Acknowledgments

Douglas O. Cardoso thanks CAPES process 99999.005992/2014-01, CNPq and GE for financial support; Daniel Alves and Diego Souza for the valuable discussions.

Felipe França thanks project SCDP, FINEP project number 01.12.0214.01 (1954/10).

João Gama thanks to European Commission through the project MAESTRA (Grant number ICT-2013-612944).

References

[1] P. Mirowski and Y. LeCun, "Statistical machine learning and dissolved gas analysis: a review," *Power Delivery, IEEE Transactions on*, vol. 27, no. 4, pp. 1791–1799, 2012. 1, 3

[2] D. de O. Cardoso, D. S. Carvalho, D. S. F. Alves, D. F. P. de Souza, H. C. C. Carneiro, C. E. Pedreira, P. M. V. Lima, and F. M. G. França, "Credit analysis with a clustering RAM-based neural classifier," in *22th European Symposium on Artificial Neural Networks, ESANN 2014, Bruges, Belgium, April 23-25, 2014*, 2014. 1

[3] S. S. Khan and M. G. Madden, "A survey of recent trends in one class classification," in *Artificial Intelligence and Cognitive Science - 20th Irish Conference, AICS 2009, Dublin, Ireland, August 19-21, 2009, Revised Selected Papers*, 2009, pp. 188–197. 1

[4] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *21st European Symposium on Artificial Neural Networks, ESANN 2013, Bruges, Belgium, April 24-26, 2013*, 2013. 1, 4

[5] D. O. Cardoso, M. D. Gregorio, P. M. V. Lima, J. Gama, and F. M. G. França, "A weightless neural network-based approach for stream data clustering," in *Intelligent Data Engineering and Automated Learning - IDEAL 2012 - 13th International Conference, Natal, Brazil, August 29-31, 2012. Proceedings*. Springer, 2012, pp. 328–335. 1

[6] B. Hu, Y. Chen, and E. J. Keogh, "Time series classification under more realistic assumptions," in *Proceedings of the 13th SIAM International Conference on Data Mining, May 2-4, 2013. Austin, Texas, USA.*, 2013, pp. 578–586. 1

[7] Y. Lee and S. Cho, "Activity recognition with Android phone using mixture-of-experts co-trained with labeled and unlabeled data," *Neurocomputing*, vol. 126, pp. 106–115, 2014. 1

[8] Z. S. Abdallah, M. M. Gaber, B. Srinivasan, and S. Krishnaswamy, "Adaptive mobile activity recognition system with evolving data streams," *Neurocomputing*, vol. 150, pp. 304–317, 2015. 1

[9] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1757–1772, 2013. 1, 5

[10] G. Fumera and F. Roli, "Support vector machines with embedded reject option," in *Pattern Recognition with Support Vector Machines, First International Workshop, SVM 2002, Niagara Falls, Canada, August 10, 2002, Proceedings*, 2002, pp. 68–82. 2

[11] R. Zhang and D. N. Metaxas, "RO-SVM: support vector machine with reject option for image categorization," in *Proceedings of the British Machine Vision Conference 2006, Edinburgh, UK, September 4-7, 2006*, 2006, pp. 1209–1218. 2

[12] Y. Grandvalet, A. Rakotomamonjy, J. Keshet, and S. Canu, "Support vector machines with a reject option," in *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, 2008, pp. 537–544. 2

[13] W. J. Scheirer, L. P. Jain, and T. E. Boult, "Probability models for open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2317–2324, 2014. 2

[14] L. P. Jain, W. J. Scheirer, and T. E. Boult, "Multi-class open set recognition using probability of inclusion," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III*, 2014, pp. 393–409. 2, 5

[15] C. Chen, Y. Zhan, and C. Wen, "Hierarchical face recognition based on SVDD and SVM," in *2009 International Conference on Environmental Science and Information Application Technology, ESIAT 2009, Wuhan, China, 4-5 July 2009, 3 Volumes*, 2009, pp. 692–695. 2

[16] B. Hanczar and M. Sebag, "Combination of one-class support vector machines for classification with reject option," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I*, 2014, pp. 547–562. 2

[17] W. Homenda, M. Luckner, and W. Pedrycz, "Classification with rejection based on various SVM techniques," in *2014 International Joint Conference on Neural Networks, IJCNN 2014, Beijing, China, July 6-11, 2014*, 2014, pp. 3480–3487. 2

[18] D. M. J. Tax and R. P. W. Duin, "Growing a multi-class classifier with a reject option," *Pattern Recognition Letters*, vol. 29, no. 10, pp. 1565–1570, 2008. 2

[19] I. Aleksander, M. D. Gregorio, F. M. G. França, P. M. V. Lima, and H. Morton, "A brief introduction to weightless neural systems," in *ESANN 2009, 17th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 22-24, 2009, Proceedings*, 2009. 2

[20] B. P. A. Grieco, P. M. V. Lima, M. D. Gregorio, and F. M. G. França, "Producing pattern examples from "mental" images," *Neurocomputing*, vol. 73, no. 7-9, pp. 1057–1064, 2010. 2

[21] G. Fumera, F. Roli, and G. Giacinto, "Reject option with multiple thresholds," *Pattern Recognition*, vol. 33, no. 12, pp. 2099–2101, 2000. 2, 3

[22] C. Goutte and É. Gaussier, "A probabilistic interpretation of precision, recall and *F*-score, with implication for evaluation," in *Advances in Information Retrieval, 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005, Proceedings*, 2005, pp. 345–359. 3

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. 3

[24] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 12 1945. 4

[25] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006. 4

[26] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Institute of Technology, Tech. Rep. 7694, 2007. 5

[27] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 2009, pp. 248–255. 5

[28] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. MIT Press, 1999, pp. 61–74. 5