2011 10th International Conference on Machine Learning and Applications

Max-Coupled Learning: Application To Breast Cancer

Jaime S. Cardoso and Inês Domingues
jaime.cardoso@inescporto.pt inesdomingues@gmail.com
INESC Porto, Faculdade de Engenharia, Universidade do Porto
Campus da FEUP, Rua Dr. Roberto Frias, 378. 4200 - 465 Porto, Portugal
Telephone: +351 222 094 000
Fax: +351 222 094 050

***Abstract*—In the predictive modeling tasks, a clear distinction is often made between learning problems that are supervised or unsupervised, the first involving only labeled data (training patterns with known category labels) while the latter involving only unlabeled data. There is a growing interest in a hybrid setting, called semi-supervised learning; in semi-supervised classification, the labels of only a small portion of the training data set are available. The unlabeled data, instead of being discarded, are also used in the learning process.**

Motivated by a breast cancer application, in this work we address a new learning task, in-between classification and semi-supervised classification. Each example is described using two different feature sets, not necessarily both observed for a given example. If a single view is observed, then the class is only due to that feature set; if both views are present the observed class label is the maximum of the two values corresponding to the individual views.

We propose new learning methodologies adapted to this learning paradigm and experimentally compare them with baseline methods from the conventional supervised and unsupervised settings. The experimental results verify the usefulness of the proposed approaches.

***Keywords*-Ordinal learning, Semi-supervised learning, Support vector machines, Bi-RADS, Decision support systems.**

I. Introduction

According to the World Health Organization, breast cancer was responsible for approximately 519 000 deaths in 2004 comprising 16% of all cancer incidence among women. In 2008, it was the most common form of cancer and cancer related death in women worldwide [1]. For this reason, breast cancer early detection and diagnosis is essential to decrease its associated mortality rate. Therefore, screening is recommended by all medical community [2], [3]. X-ray mammography is currently considered the best imaging method for breast cancer screening and the most effective tool for early detection of this disease [4]. Screening mammographic examinations are performed annually on asymptomatic women to detect early, clinically unsuspected lesions.

When radiologists examine mammograms, they look for specific abnormalities [5]. The most common findings that can be seen on mammography are masses and calcifications. In order to standardize the terminology of the mammographic report, the assessment of findings and the recommendation of action to be taken, the American College of Radiology (ACR) developed the Breast Imaging Reporting and Data System (BI-RADS) scale [6]. Based on level of suspicion, lesions can be placed into one of six BI-RADS scores: score 0 when the exam is not conclusive, score 1 for no findings, score 2 for benign findings, score 3 for probably benign findings, score 4 for suspicious findings, score 5 when there is a big probability of malignancy, and score 6 for proved cancer. **When more than one finding is present in the mammogram, the overall BI-RADS in the medical report corresponds to the finding with highest BI-RADS.** This is the key observation that will motivate our work.

Computer-Aided Detection and Diagnosis (CAD) systems have been developed in the past two decades to assist the radiologists in the interpretation of medical images [7], [8]. Recently there is a surge of interest in systems to predict the final BI-RADS assessment. The design of a CAD system to predict the BI-RADS classification demands a dataset of mammograms with known BI-RADS classification.

An approach based on standard classification techniques would extract features from calcifications and masses, when present, and design a classifier in the joint space. One disadvantage of this approach is that it is not clear how to use the cases with masses only or calcifications only in the design of the CAD nor how to use the CAD is such cases. Moreover, the classifier would have to learn automatically from the data that the final classification is the maximum of the values obtained from the two 'views', masses and calcifications; it would be better to incorporate this knowledge in the learning process.

A second standard option is to train a classifier to make the prediction for one type of findings (e.g. masses) and a second classifier for the other type of findings (e.g. calcifications); the final classification would be the maximum of the two predicted values. To train each individual classifier, one could use the cases with that finding only, for which one knows the true class. The training could be improved by using semi-supervised learning techniques: the cases with both findings would be used as unlabeled data to improve the performance of the individual classifier. The disadvantage of this approach is that, by ignoring the classification when both findings are present, one is not using all the information available during training: although when both findings are present one does not know which one is responsible by the score, one does know

978-0-7695-4607-0/11 $26.00 © 2011 IEEE
DOI 10.1109/ICMLA.2011.93

that at least one of them motivated that score.

Motivated by the described application, we formalize a new learning paradigm and propose new learning methodologies to make efficient use of all the available information.

In Section II, the learning taxonomy is briefly presented; Section III gives the problem formulation; in Section IV, the three different methodologies are proposed to address the learning problem previously formulated; Section V details and presents the experiments done; the paper concludes in Section VI with some final remarks and future directions.

II. Related work

The goal of unsupervised learning or "learning without a teacher" is to directly infer the properties of the data probability density without the help of a supervisor or teacher providing correct answers or degree-of-error for each observation. Clustering is arguably the most often proposed goal for unsupervised learning, having k-means as the simplest and most popular scheme.

Supervised classification is concerned with predicting the values of one or more outputs or response variables for a given set of input or predictor variables. The predictions are based on the training sample of previously solved cases, where the joint values of all of the variables are known. Two of the state of the art supervised algorithms are Neural Networks (NNs) and Support Vector Machines (SVMs).

Traditional classifiers use only labeled data (input / output pairs) to train. Labeled instances however are often difficult, expensive, or time consuming to obtain, as they require the efforts of experienced human annotators. In the medical field, for instance, it requires an expert and consequently time and money in order to have reliable labels. Meanwhile unlabeled data may be relatively easy to collect, but there has been few ways to use them. Semi-supervised learning addresses this problem by using large amount of unlabeled data, together with the labeled data, to build better classifiers. Intuitively, unlabeled data may be useful in order to gain some knowledge on the data structure. Because semi-supervised learning requires less human effort and gives higher accuracy, it is of great interest both in theory and in practice.

According to [9], some of the most representative methods for semi-supervised learning are: (1) generative models (assume the form of a joint probability); (2) semi-supervised SVMs (assume that the decision boundary is situated in a low-density region of unlabeled data between the classes); (3) graph-based models (assume that there is a graph such that the vertices are the labeled and unlabeled training instances, and the undirected edges connect the instances with a given weight); (4) co-training and multi-view models (assume that there are multiple, different learners trained on the same labeled data, and these learners agree on the unlabeled data); and (5) self-training (a classifier is first trained with a small amount of labeled data and then used to classify the unlabeled data; the most confident unlabeled points, together with their predicted labels, are added to the training set).

The learning setting that we address next is neither classification (we do not have complete information about the labels in the full training set) nor semi-supervised (we do have some information about the label for each training example), sitting in-between these two standard scenarios. This is a learning problem with incomplete label knowledge. We argue that this setting is of interest in scenarios other than BI-RADS classification, where based on information from different sources, the decision is made on a worst case analysis of the 'suggestions' from the multiple sources.

Probably, the most similar learning methodology to the one being addressed here is the Multiple Instance Learning (MIL) [10], [11]. The basic idea of MIL is that during training examples are presented in sets (often call 'bags'), and labels are provided for the bags rather than for individual instances. If a bag is labeled positive, it is assumed to contain at least one positive instance, otherwise the bag is negative. Note that this paradigm is for binary settings only and that all the 'views' in the bag come from the same feature set.

III. Max-coupling Semi-Supervised Learning

Consider a training set comprising three different type of observations:

1) $S_1 = \{\mathbf{x}_i, y_i = f(\mathbf{x}_i)\}$, where $i = 1, \ldots, N_1$ and $\mathbf{x}_i \in \Re^{d_1}$, with d_1 the dimension of the feature space (in the breast cancer application, these observations correspond to the cases where only mass was detected in the mammogram) and $f(.)$ is unknown.
2) $S_2 = \{\mathbf{z}_i, y_i = g(\mathbf{z}_i)\}$, where $i = 1, \ldots, N_2$ and $\mathbf{z}_i \in \Re^{d_2}$, with d_2 the dimension of the feature space (in the breast cancer application, these observations correspond to the cases where only calcifications were detected in the mammogram) and $g(.)$ is unknown.
3) $S_{12} = \{\mathbf{x}_i, \mathbf{z}_i, y_i\}$, where $i = 1, \ldots, N_{12}$, $\mathbf{x}_i \in \Re^{d_1}$ and $\mathbf{z}_i \in \Re^{d_2}$ (in the breast cancer application, these observations correspond to the cases where both mass and calcifications were detected in the mammogram). It is known that $y_i = \max(f(\mathbf{x}_i), g(\mathbf{z}_i))$ but $f(\mathbf{x}_i)$ and $g(\mathbf{z}_i)$ are both unobserved (in the breast cancer application, $f(\mathbf{x}_i)$ would correspond to the BI-RADS classification due to the presence of the mass only; similarly to $g(\mathbf{z}_i)$).

For every observation, y_i corresponds to a known classification in one of K ordinal classes (in the breast cancer application, this label corresponds to the overall BI-RADS label present in the medical report). An illustration is given in Fig. 1.

IV. Learning the Max-coupling dependencies

In this section, three new proposals, each with different characteristics and motivations, are presented, all aiming to make an efficient use of the available data in the design of the models. In Section IV-A the proposed method is inspired by Semi-Supervised methods; the algorithm presented in Section IV-B is based on on-line classification approaches; finally, in Section IV-C, a batch approach is given.

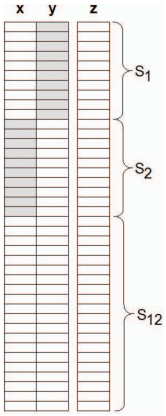


Fig. 1: Training set illustration. White represents observed and Gray unobserved features.

A. A Modified semi-supervised approach

As described in the introduction, a standard solution, with the disadvantage of discarding the information of labels when both features are present, is to train, using semi-supervised techniques, a first classifier to approximate $f(\mathbf{x})$, using the $\mathbf{x}$ features from the complete set and the output label only in the $\mathcal{S}_1$ subset, and a second classifier to approximate $g(\mathbf{z})$, using the $\mathbf{z}$ features from the complete set and the output label only in the $\mathcal{S}_2$ subset. The final classification is the maximum of the two predictions.

As previously mentioned, self-training is one of the most representative approaches. Self-training first learns a separate classifier for each view ($\mathbf{x}$ and $\mathbf{z}$) using any labeled examples. The most confident predictions of each classifier on the unlabeled data are then used to iteratively construct additional labeled training data. Our first proposal to make use of all the available information is inspired on self-training. Two classifiers are initially trained with the samples $\{\mathbf{x}_i, f(\mathbf{x}_i)\}$, where $i = 1, \ldots, N_1$ and $\{\mathbf{z}_i, g(\mathbf{z}_i)\}$, where $i = 1, \ldots, N_2$, respectively. The two classifiers are used to make prediction in the subset $\mathcal{S}_{12}$. If the maximum of the two predictions agree with the known label, the labeled training point is added only to the classifier predicting the maximum value (in case of a tie, both models receive the new training data). Intuitively, instead of selecting the new training points based on the estimated confidence, the points are chosen if the final predictions agrees with the known label.

B. A Modified on-line supervised approach

We have already explained in the introduction why the design of a single classifier in the joint space is not desirable. The use of two independent classifiers also does not seem advantageous.

However, that does not preclude the use of standard supervised learners to address the max-coupling learning problem. Due to its simplicity, let's start by considering on-line classifiers.

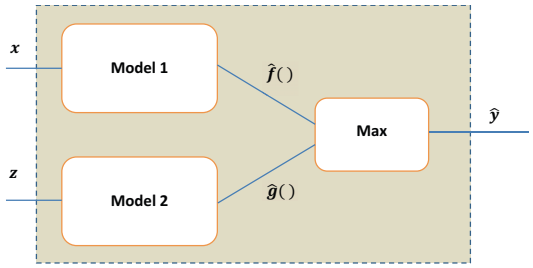


Fig. 2: Proposed supervised architecture for the max-coupling learning.

On-line supervised learning algorithms incrementally build the model and dynamically refine it over time using the most recent observation. Denoting by $\mathbf{w}$ the set of parameters of the model, a typical update rule follows the format

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{w}}|_t,$$

where η is the learning rate and $\mathcal{L}$ some loss function. We are assuming an update rule based on the gradient but similar rules exist for other rationales.

We propose to incorporate in the architecture of the model the knowledge about the output, namely that it is the maximum of two independent values; adapting the architecture to the learning problem, we then update the parameters using the aforementioned update rule. We propose to design the global model as the parallel two individual models coupled by a max computation in the end, see Fig. 2. Each of the individual models is parameterized by its own set of parameters, $\mathbf{w}_1$ and $\mathbf{w}_2$. Consider now that we receive the current observation and we want to update the join model.

When both $\mathbf{x}_i$ and $\mathbf{z}_i$ are present, three different cases should be considered:

1) $\hat{\mathbf{y}}_i = \hat{f}(\mathbf{x}_i) \quad \wedge \quad \hat{\mathbf{y}}_i > \hat{g}(\mathbf{z}_i)$
   Assuming that both models are continuous functions of the parameters (and the magnitude of the gradient is bounded), then 'small' changes in the parameters of the second model will not affect the output of the joint model nor the loss function. Therefore, the derivative of the loss in respect to the parameters of the second model is zero and one only needs to update the first model. Since the loss function at the output of the first model equals the loss at the output of the joint model, the update follows the conventional rule, as if observing the output $\hat{y}_i$ in the output of the first model.
2) $\hat{\mathbf{y}}_i = \hat{g}(\mathbf{z}_i) \quad \wedge \quad \hat{\mathbf{y}}_i > \hat{f}(\mathbf{x}_i)$
   In this case the roles of the first and second models are reversed, and one only needs to conventionally update the second model.
3) $\hat{\mathbf{y}}_i = \hat{g}(\mathbf{z}_i) = \hat{f}(\mathbf{x}_i)$ In this case both models should be updated, as if observing the output $\hat{y}_i$ in the output of each of the models. This follows from the observation that

$$\frac{\partial \mathcal{L}(y_i, \hat{y}_i)}{\partial \mathbf{w}_1} = \frac{\partial \mathcal{L}(y_i, \hat{f}(\mathbf{x}_i))}{\partial \mathbf{w}_1}$$

and

$$\frac{\partial \mathcal{L}(y_i, \hat{y}_i)}{\partial \mathbf{w}_2} = \frac{\partial \mathcal{L}(y_i, \hat{g}(\mathbf{z}_i))}{\partial \mathbf{w}_2},$$

where $\mathbf{w}_1$ and $\mathbf{w}_2$ represent the vector of parameters for the first and second models, respectively.

When only one of $\mathbf{x}_i$ and $\mathbf{z}_i$ is present only the corresponding model is updated.

Some observations are in order:

- Let's start by stressing that the prior knowledge about the problem was incorporated in the proposed architecture; from there we just proceeded with standard update rules based on the gradient.
- Note that this setting is easily generalized for more than two views or sources of data, where the final observed decision is the maximum of $M \geq 2$ individual results.
- We are not constrained in the choice of the individual learning schemes. We can adopt any preferred family of on-line/incremental learning schemes and, in generality, we can adopt different schemes for different individual models, tuning the scheme to the specificities of the individual problem.

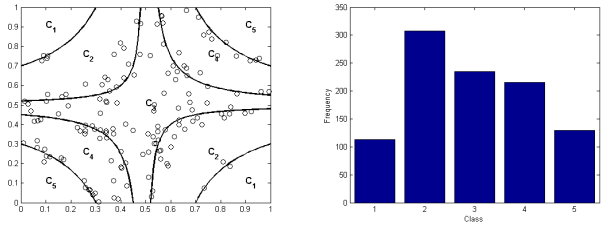
C. A Modified batch supervised learning approach

The online proposed model suggests that the most informative observations are the ones currently being misclassified, since these are the ones leading to an update of the models. This paves the way to a new batch mode supervised learning: instead of adding to the current training set the points where both classifiers agree (as in the standard co-training) or the set of points where the maximum of the two outputs agrees with the observed global classification, every point will be used, either in the training set of model 1 or in the training set of model 2 (or in both), Algorithm 1. This algorithm follows closely the on-line learning ideas from the previous section.

Algorithm 1 Pseudo-Code of the batch-mode proposed approach

$TrainingSet_1 = \mathcal{S}_1$
$TrainingSet_2 = \mathcal{S}_2$
for $epoch = 1 : max\_epoch$ **do**
  Train $Model_1$ in $TrainingSet_1$
  Train $Model_2$ in $TrainingSet_2$
  $P_1 \leftarrow$ predict($Model_1$, $\mathcal{S}_{12}$)
  $P_2 \leftarrow$ predict($Model_2$, $\mathcal{S}_{12}$)
  $idx_1 = find(P_1 \geq P_2)$
  $idx_2 = find(P_2 > P_1)$
  $TrainingSet_1 = \mathcal{S}_1 \cup \mathcal{S}_{12}(idx_1)$
  $TrainingSet_2 = \mathcal{S}_2 \cup \mathcal{S}_{12}(idx_2)$
end for

It is important to highlight a key difference to the proposed model based in semi-supervised learning techniques. With the semi-supervised based model, only examples where the final prediction agrees with the known label are incorporated in the training set of one or both models; moreover when an example is added in one of the training sets it is no longer removed. With the now proposed approach, all examples are selected to integrate one or both training sets; between two epochs examples can be moved from one training set to the other.



(a) (b)

Fig. 3: Artificial dataset corresponding to one of the views of dataset 1. (a) Scatter plot of the examples wrongly classified (due to the noise added) in the artificial dataset. Also shown are the class boundaries. (b) Class distribution in the artificial dataset.

V. Experimental Validation

Three types of synthetic datasets (described in Section V-A) were used in order to validate the proposed methods. Results are given in Section V-C.

A. Synthetic Data

Dataset 1: Example points $\mathbf{x} = (x_1, x_2)^t$ were randomly generated in the unit square $[0, 1] \times [0, 1] \in \Re^2$ according to the uniform distribution [12]. To each point was assigned a class y from the set $\{1, 2, 3, 4, 5\}$, according to

$$y = \min_r \left\{ r : b_{r-1} < 10 \prod_{i=1}^{2} (x_i - 0.5) + \epsilon < b_r \right\}$$

$(b_0, b_1, b_2, b_3, b_4, b_5) = (-\infty, -1, -0.1, 0.25, 1, +\infty)$

where $\epsilon \sim N(0, 0.125^2)$ simulates the possible existence of error in the assignment of the true class on $\mathbf{x}$.

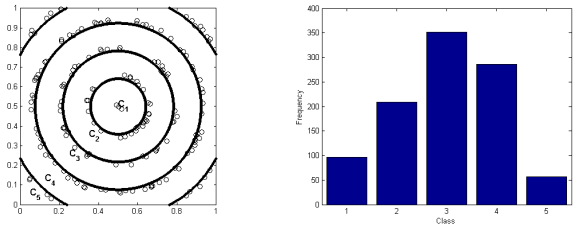
The $\mathbf{z}$-view of the data was similarly generated. We generated N_1 samples for $\mathbf{x}$-view only, N_2 samples for $\mathbf{z}$-view only, and N_{12} samples for both views, where examples from both views were concatenated in a $\Re^4$-feature, keeping only the maximum of the corresponding labels for the observed output of the example. An illustration of this dataset is given in Figure 3.

Dataset 2: The only difference on this dataset to the previous dataset is on the $\mathbf{z}$-view, which was generated similarly to the previous dataset but now in a 3-dimensional space according to:

$$y = \min_r \left\{ r : b_{r-1} < 100 \prod_{i=1}^{3} (x_i - 0.5) + \epsilon < b_r \right\},$$

with the same b_i's and noise distribution.

Dataset 3: The only difference on this dataset to the Dataset 2 is on the $\mathbf{x}$-view. For the $\mathbf{x}$-view, example points



(a) (b)

Fig. 4: Artificial dataset corresponding to the first view of dataset 3. (a) Scatter plot of the examples wrongly classified (due to the noise added) in the artificial dataset. Also shown are the class boundaries. (b) Class distribution in the artificial dataset.

$\mathbf{x} = (x_1, x_2)^t$ were once again randomly generated in the unit square $[0,1]\times[0,1] \in \Re^2$ according to the uniform distribution. However, to each point, the class y was assigned according to

$$y = \min_r \left\{ r : b_{r-1} < \sqrt{\sum_{i=1}^{2}(x_i - 0.5)^2} + \epsilon < b_r \right\}$$

$(b_0, b_1, b_2, b_3, b_4, b_5) = \left(0, \frac{1}{5}\frac{\sqrt{2}}{2}, \frac{2}{5}\frac{\sqrt{2}}{2}, \frac{3}{5}\frac{\sqrt{2}}{2}, \frac{4}{5}\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$ where ϵ is as before. An illustration of this dataset is given in Figure 4.

B. Methodology

We randomly split the generated datasets into training and test sets. The total number of generated points for each dataset was 1000. In order to study the effect of varying the size of the training set, we considered two possibilities: 10% and 50% of all the data used for training. In order to study the effect of varying the proportion in the data with a single-view, we considered two possibilities: 10% and 40% of all the data had information from a single-view (within this, the cases were equally divided by the two views). In each of the four combinations, the splitting of the data was repeated twenty times in order to obtain more stable results for performance estimation. Since the data is ordinal, we adopted the mean absolute error (MAE) as a measure of performance. Two kernel types, degree three polynomial and radial basis function, together with their parameterization were optimized by cross-validation inside the training set.

All models were instantiated with support vector machines (SVMs) by using the LIBSVM package [13]. For multi-class, LIBSVM uses the one-against-one method [14].

The standard two classifiers model trains a classifier for each of the views, ignoring the subset S_{12} for training. The standard one classifier model trains a single classifier in the complete dataset, previously replacing unknown feature-values by the corresponding average value. As standard semi-supervised technique, Tri-training [15] was selected. In this method, the labeled data are split in three sets and a classifier is trained in each set. If two of them agree on the classification of an unlabeled point, the classification is used to "teach" the other classifier.

C. Results

Overall results are presented in Table I. Besides the MAE in the complete test set, we also show the performance for the test subset with information only for a single view and for the subset with the two views. Since the results for 10% and 50% of the data for training were consistent in terms of relative performance, we only show results for the last percentage.

A first conclusion is that the tri-training model was the worst performing model. Since the training data had to be split in three different subsets, it seems the labeled data was not enough to bootstrap the learning process.

The single classifier approach has also difficulties learning from the data, both when the percentage of single-view data is low and moderate in the complete set. This single class model is, in the overall, the second worst-performing model. It is also possible to observe that the conventional two classifier approach improves the performance with the increase of the proportion of single-view data (the single classifier exhibits an opposite behavior). For 40% of single-view data in the dataset, this standard learning becomes quite competitive. However, like in standard semi-supervised learning, the proportion of single-view data in practice is much lower; and for low proportions, the results for this method are not robust.

Analyzing the performance in the complete test set and in the subset comprising only single-view examples, in general, all models perform worst on the single view examples. Nevertheless, the proposed batch model continues to be the best-performing model in this analysis per subset.

Among the proposed methods, the method motivated by semi-supervised ideas exhibits a performance just slightly better that the standard two-classifier method. It seems that in this case the technique to make use of the data form the S_{12} subset is not effective. A different analysis can be made for the batch model, performing clearly above the others. It is interesting to note that the performance for this model is quite good, for low and moderate proportions of single-view data.

So far, we are still collecting our database of mammograms and manually annotating them and therefore we are unable to present results for this specific application. Nevertheless, the results obtained for synthetic data are very promising and potentially interesting for different scenarios.

VI. Conclusion

The typical learning settings, already dissected in the literature, are not necessarily the most interesting for practical applications. The more or less recent surge of interest in semi-supervised techniques is also an attempt to build methods that can better answer the information that is typically available when building the systems.

In this paper we propose a new learning paradigm, in between classification and semi-supervised classification. With classification, one has perfect knowledge of the label for each training example. With semi-supervised classification, for some examples one continues to have perfect knowledge about the label, while for the others the label is completely unknown. In proposed max-coupled setting, for every observation, we do

Single view percentage	Synthetic Dataset	Standard 2 classifiers	Standard 1 classifier	Standard tri-training	Proposed semi-supervised	Proposed batch
10%	Synthetic 1	39(41/39)%	36(45/35)%	67(69/67)%	37(36/37)%	20(25/20)%
	Synthetic 2	70(69/70)%	51(54/51)%	90(90/90)%	65(67/64)%	26(31/25)%
	Synthetic 3	64(76/62)%	43(50/42)%	88(104/87)%	55(66/54)%	24(36/22)%
40%	Synthetic 1	25(25/25)%	36(28/41)%	30(30/30)%	21(22/21)%	19(19/19)%
	Synthetic 2	37(35/38)%	48(35/58)%	61(63/60)%	33(33/32)%	24(26/22)%
	Synthetic 3	31(34/29)%	41(36/44)%	55(58/53)%	27(28/26)%	22(25/20)%

TABLE I: Overall MAE results with 50% data for training. Results for are present in the format 'complete test set (single view/two views)'.

possess some information about the label; however, in a subset of the examples, the knowledge is incomplete, corresponding to the worst-case classification of the individual views of the example.

After motivating and formalizing this new learning paradigm, we proposed modifications to existing learning schemes, suggested both from semi-supervised and supervised ideas. The quality of our results motivate us to further pursue the problem. An interesting question, which we plan to research in the future, is if the design 'from scratch' of models with embedded knowledge of the max-coupled setting can further improve the results. We will also conduct a larger validation of these ideas on real datasets.

ACKNOWLEDGMENTS

This work was partially supported by Fundação para a Ciência e Tecnologia (FCT) - Portugal through projects PTDC/SAU-ENB/114951/2009 and SFRH/BD/70713/2010.

REFERENCES

[1] P. Boyle and B. Levin, "World cancer report," International Agency for research and cancer, Lyon, France, Tech. Rep., 2008.

[2] "Health statistics: atlas on mortality in the european union," *Eurostat*, 2009.

[3] "Fact sheet no. 297: Cancer," *World Health Organization*, 2009.

[4] S. Misra, N. L. Solomon, F. L. Moffat, and L. G. Koniaris, "Screening criteria for breast cancer," *Advances in surgery*, vol. 44, pp. 87–100, 2010.

[5] M. P. Sampat, A. C. Bovik, G. J. Whitman, and M. K. Markey, "A model-based framework for the detection of speculated masses on mammography," *Medical Physics*, vol. 35, no. 5, pp. 2110–2123, 2008.

[6] C. J. Orsi, "The american college of radiology mammography lexicon: an initial attempt to standardize terminology," *American journal of roentgenology (AJR)*, vol. 166, no. 4, pp. 779–780, 1996.

[7] B. Zheng, L. A. Hardesty, W. R. Polle, J. H. Sumkin, and S. Golla, "Mammography with Computer-Aided detection: Reproducibility assessment - initial experience," *Radiology*, vol. 228, pp. 58–62, 2003.

[8] R. A. Castellino, "Computer aided detection (CAD): an overview," *Cancer Imaging*, vol. 5, pp. 17–19, 2005.

[9] X. Zhu, "Semi-Supervised learning," in *Encyclopedia entry in Claude Sammut and Geoffrey Webb, editors, Encyclopedia of Machine Learning*. Springer, 2009.

[10] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems 15*. MIT Press, 2003, pp. 561–568.

[11] T. G. Dietterich and R. H. Lathrop, "Solving the multiple-instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, pp. 31–71, 1997.

[12] J. S. Cardoso and J. F. P. da Costa, "Learning to classify ordinal data: the data replication method," *Journal of Machine Learning Research*, vol. 8, pp. 1393–1429, 2007.

[13] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[14] C. W. Hsu and C. J. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, pp. 415–425, 2002.

[15] Z. Zhou and M. Li, "Tri-Training: exploiting unlabeled data using three classifiers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 11, pp. 1529 – 1541, 2005.