# Max-Coupled Ordinal Classification

Jaime Cardoso
http://www.inescporto.pt/~jsc/
Inês Domingues
http://vcmi.inescporto.pt/

INESC Porto
Universidade do Porto
Porto, Portugal

## Abstract

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.

## 1 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. 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.

In order to standardize the terminology of the mammographic report, mammography findings are classified into the BI-RADS scale. 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 motivates this work.

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 classifier nor how to use the classifier 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 thus 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 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.

## 2 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 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 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 $\mathbf{z}_i \in \Re^{d_2}$. 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.

For every observation, $y_i$ corresponds to a known classification in one of $K$ ordinal classes.

## 3 Learning the Max-coupling dependencies

### 3.1 A Modified semi-supervised approach

Self-training first learns a separate classifier for each view (**x** and **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.

### 3.2 A Modified on-line supervised approach

On-line supervised learning algorithms incrementally build the model and dynamically refine it over time using the most recent observation. Denoting by **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 $\eta$ 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 of two individual models coupled by a max computation in the end. 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 only the first models needs to be updated. 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.

| single view percentage | Dataset | Standard Two Classifiers | Standard One Classifier | Standard tri-Training | proposed semi-supervised | proposed batch |
|---|---|---|---|---|---|---|
| 40% | ESL vs ESL | 47(49/45)% | 28(23/34)% | 14(15/11)% | 10(11/9)% | 8(8/7)% |
| | ESL vs bcct | 39(40/18)% | 21(20/50)% | 15(16/11)% | 9(9/5)% | 7(8/5)% |
| | ESL vs Pasture | 18(11/41)% | 20(13/45)% | 11(10/13)% | 5(3/10)% | 3(2/6)% |
| | bcct vs bcct | 32(36/27)% | 24(21/27)% | 13(14/11)% | 15(18/11)% | 9(10/7)% |
| | bcct vs Pasture | 4(4/27)% | 2(2/18)% | 8(7/17)% | 1(1/5)% | 1(1/2)% |
| | Pasture vs Pasture | 18(17/18)% | 7(5/10)% | 11(11/12)% | 3(4/3)% | 3(4/3)% |
| Time elapsed (sec) | | 6.52 | 12.47 | 714.18 | 780.82 | 884.22 |

Table 1: Overall MAE results for real datasets with 50% data for training. Results for are present in the format 'complete test set (single view/two views)'.

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.

### 3.3 A Modified batch supervised learning approach

The on-line 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).

The algorithm consists thus in first training two models, one with $S_1$ and another with $S_2$. Each model will be used to predict the class of the samples in $S_{12}$. Each model is then iteratively retrained with the original samples plus all the samples from $S_{12}$ for which the model prediction was higher. For example, if for the first sample in $S_{12}$, model A predicts a higher class than model B, that sample will be added to the training set of model A.

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.

## 4 Experimental Validation

### 4.1 Datasets

In our previous work [2], experiments with synthetic datasets were presented. Here, three types of real data were used and combined. The first dataset, ESL, contains 488 profiles of applicants for certain industrial jobs. The class assigned to each applicant was an overall score corresponding to the degree of fitness for the type of job. bcct dataset, encompasses on 960 observation taken from [1] and expresses the aesthetic evaluation of Breast Cancer Conservative Treatment (BCCT). The aesthetic outcome of the treatment for each and every patient was classified in one of the four categories: Excellent, Good, Fair and Poor. The last dataset, Pasture, contains information on the pasture production from a variety of biophysical factors. The target feature has been categorized in three classes, (Low, Medium, High), evenly distributed in the dataset of 36 instances. Each one of the above three datasets was considered as one of the views, yielding a total of 6 datasets.

### 4.2 Methodology

We randomly split the datasets into training and test sets. 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. Since the data is ordinal, we adopted the MAE as a measure of performance. Each model parameterization was optimized by cross-validation inside the training set.

All models were instantiated with SVMs. 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 [3] 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.

### 4.3 Results

Overall results are presented in Table 1. 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 standard two classifiers model was the worst performing model. The single classifier approach has also difficulties learning from the data. The proposed batch model is the best-performing model both in the complete test set and in the subset comprising only single-view examples.

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 the presented data are very promising and potentially interesting for different scenarios.

## 5 Conclusion

In this paper we propose a new learning paradigm, in between classification and semi-supervised classification. In proposed max-coupled setting, for every observation, we do 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. 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.

## References


[1] J. S. Cardoso and M. J. Cardoso. Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. *Art Int in Med*, 40(2):115–126, 2007.

[2] J. S. Cardoso and I. Domingues. Max-coupled learning: application to breast cancer. In *ICMLA*, Hawaii, 2011.

[3] Z. H. Zhou and M. Li. Tri-training: exploiting unlabeled data using three classifiers. *TKDE*, 17(11):1529–1541, 2005.