Transfer Learning with Partial Observability Applied to Cervical Cancer Screening

Kelwin Fernandes[1,2], Jaime S. Cardoso[1,2], and Jessica Fernandes[3]

[1] INESC TEC, Portugal
{kafc, jaime.cardoso}@inesctec.pt,
[2] Universidade do Porto, Portugal
[3] Universidad Central de Venezuela

Abstract. Cervical cancer remains a significant cause of mortality in low-income countries. As in many other diseases, the existence of several screening/diagnosis methods and subjective physician preferences creates a complex ecosystem for automated methods. In order to diminish the amount of labeled data from each modality/expert we propose a regularization-based transfer learning strategy that encourages source and target models to share the same coefficient signs. We instantiated the proposed framework to predict cross-modality individual risk and cross-expert subjective quality assessment of colposcopic images for different modalities. Thus, we are able to transfer knowledge gained from one expert/modality to another.

Keywords: transfer learning, regularization, cervical cancer, digital colposcopy

1 Introduction

Despite the possibility of prevention with regular cytological screening, cervical cancer remains a significant cause of mortality in low-income countries. This being the cause of more than half a million cases per year, and killing more than a quarter of a million in the same period [1]. As in many other diseases, the existence of several screening and diagnosis methods creates a complex ecosystem from a Computed Aided Diagnosis (CAD) system point of view. For instance, in the detection of pre-cancerous cervical lesions, screening strategies include cytology, colposcopy (covering its several modalities [1]) and the gold-standard biopsy. In developing countries resources are very limited and patients usually have poor adherence to routine screening due to low problem awareness. Consequently, the prediction of the individual patient's risk and the best screening strategy during her diagnosis becomes a fundamental problem. Most of these screening methods highly depend on the physician expertise and subjective comfort on the decision process, being a key aspect to improve data acquisition using the physician preferences.

Thereby, from a technical point of view, all these predictive tasks are immersed in a multi-modal and multi-expert setting. Traditionally, supervised

learning techniques would require to collect a vast amount of data from each source (i.e. modalities and experts) and to build predictive models separately for each task. Transfer learning (TL) aims to extract knowledge from at least one source task and use it when learning a predictive model for a new target task [2]. The intuition behind this idea is that learning a new task from related tasks should be easier (faster, with better solutions or with less amount of labeled data) than learning the target task in isolation. In this work, we focus on inductive TL, where both domains are represented by the same feature space and where the source and target tasks are different but related [2]. A main trend in inductive transfer consists on transferring data, namely, strategically including data from the source task in the target dataset [3]. Another approach consists on finding a shared source-target low-dimensional feature representation that is suitable for learning the target task [4]. We group these two approaches under the umbrella of data-driven transfer, where source data is re-used to train the target task. Although these approaches may seem appealing, the vast amount of training data in the source task turns the process prohibitively expensive.

TL techniques (and its community) should be focused on adapting knowledge instead of data. This idea is handled by parameter transfer approaches, which rely on the idea that individual models for related tasks should share some structure (parameters or hyper-parameters) [2]. In this sense, the knowledge generated from a source task is understood as the parameters (and hyperparameters) that define a given model: the coefficients of a regression, the weights of a neural network, the feature hierarchy of a decision tree. Previous works [5–8] explored transferring knowledge from/to linear models by means of regularizing the coefficient difference between different tasks. In this work, we extend this idea by including the notion of partial transfer where high-level properties of the source model are transferred instead of the whole model structure. Partial transfer can be understood as improving the model performance on the target task by using a partially observable source model. This capability is specially important in some scenarios, where unlimited access to the model parameters is not possible due to privacy and security concerns (e.g. health and biometrics applications). In these cases just high-level properties of the model are available. Also, regularizing high-level properties of the models allows transfer between less similar tasks. Therefore, even when the source model is fully observable, it can be interesting to study partial transfer mechanisms.

In this work we focus on transferring the coefficient sign by proposing a new regularization scheme that encourages coefficients to share the same contribution type (i.e. positive, negative) instead of the coefficient impact (i.e. actual value). In order to prove its adequacy to different problems, we instantiated this idea to two different problems: cross-modal individual risk prediction and cross-modal and cross-expert quality assessment of digital colposcopies.

2 Proposed Method

We consider the following scenario in this work. We have two learning tasks (source and target) denoted by *src* and *tgt*. We assume that both tasks share the same feature space $X \subset \mathbb{R}^d$ and output type $Y \subset \mathbb{T}$ (e.g. regression, classification). For a given task $T \in \{src, tgt\}$, we have labelled training data $D^T = X^T \times Y^T$. In order to induce similar models, a TL objective can be understood as finding the best model that balances the tradeoff between model performance on the target data and its similarity with the source model:

$$\arg\min_{M} \left(dataLoss(M, X^{tgt}) + \lambda \; dissimilarity(M, M^{src}) \right), \; \lambda \geq 0 \tag{1}$$

Since a predictive model is a succinct representation of the data, this framework is an efficient way to introduce knowledge obtained from the source task without resorting to the source data. Therefore, it is also useful in scenarios where source data is unavailable at transfer time or in online learning settings.

2.1 Partial Model Transfer: Sign Regularization

Using the proposed framework we can selectively transfer knowledge. This can be done by considering regularization schemes that explore high-level properties of the model instead of its actual state (i.e. assumed values). This can be understood as having partial observability of the model structure.

In this work we focus on linear predictive models for regression (e.g. Linear Regression) and classification (e.g. Logistic Regression, Support Vector Machines). Thereby, we assume that our model can be defined by a vector of coefficients $\omega \in \mathbb{R}^{d+1}$, which includes the bias term ω_0. Here, we are interested in transferring the contribution direction of each feature (i.e. coefficient sign) instead of its importance in the source task (i.e. coefficient magnitude). Eq. (2) defines a dissimilarity regularizer that encourages sign relateness, where ω^{src} and ω^{tgt} denote the source and target coefficients respectively.

$$\delta_p(\omega^{tgt}, \omega^{src}) = \sum_{i=1}^{d} \max(0, -\omega_i^{tgt} \cdot \text{sign}(\omega_i^{src}))^p, \; p > 0 \tag{2}$$

Although this regularizer is able to control the sign change between source and target task, it does not establish any type of control on models with large coefficients with the same sign. Thereby, we introduce the classical L_p-norm regularizer (see Eq. (3)). Figure 1 illustrates the behavior of two particular instances of the proposed regularizer with $p = 1$ and $p = 2$.

$$\Delta_{p,\alpha}(\omega_i^{tgt}, \omega_i^{src}) = \alpha \delta_p(\omega_i^{tgt}, \omega_i^{src}) + (1 - \alpha) \parallel \omega^{tgt} \parallel_p^p, \; 0 \leq \alpha \leq 1 \tag{3}$$

The proposed regularizer is based on the Hinge loss traditionally used in the optimization of Support Vector Machines. In this sense, the particular case when

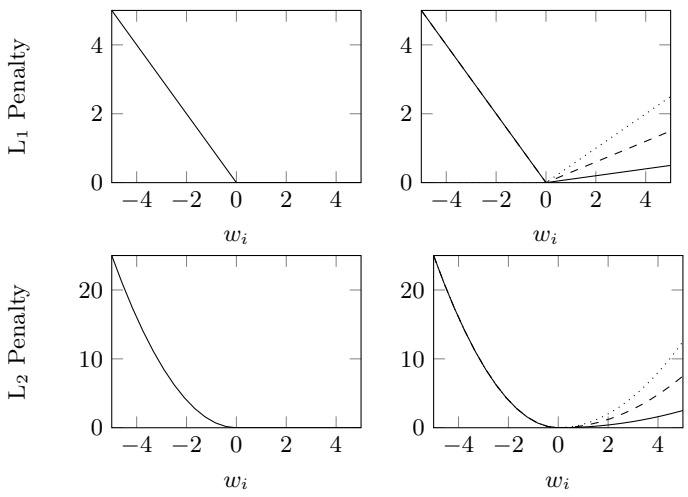

Fig. 1. Regularization factors assuming $w_i^{src} > 0$. First row illustrates the penalization using L_1 regularizers ($p = 1$) with same-sign uncontrolled penalty on the left and with different α values on the right (0.9 - solid, 0.7 - dashed, 0.5 - dotted). Second row is analogous to the first row but using L_2 penalty ($p = 2$).

$p = 2$ is a smooth version that allows gradient computation on its entire domain (see Eq. (4)). Thereby, it can be easily included in gradient descent optimization strategies.

$$\frac{\partial}{\partial \omega_i^{tgt}} \Delta_{2,\alpha} = (1-\alpha)\omega_i^{tgt} + \alpha \begin{cases} 0 & , \text{sign}(\omega_i^{src}) = \text{sign}(\omega_i^{tgt}) \\ -|\omega_i^{tgt}| \, \text{sign}(\omega_i^{src}) & , \text{otherwise} \end{cases} \quad (4)$$

On the other hand, when $p = 1$, the derivative at $\omega_i = 0$ is non-deterministic. However, the subgradient at $\omega_i = 0$ can be computed, inducing a subgradient descent optimization strategy. Due to space limitations we only present results for the smooth version of the proposed regularizer.

3 Experiments

Data was split using a stratified training-test partition (80-20). Then, in order to validate the model performance on different stages of the data acquisition process, the training set was randomly subsampled in 10 nested subsets with several sizes $(10\%, 20\%, 30\%, \ldots, 100\%)$. Each experiment was repeated 30 times varying the test partition. The regularization factor (λ) and all the remaining intrinsic hyper-parameters were learned using Stratified K-fold cross-validation (K = 3) over the training set. For reproducibility purposes, the datasets are made available[4].

[4] http://vcmi.inescporto.pt/reproducible_research/ibpria2017/CervicalCancer/

For each method, the normalized signed Area Under the gain Curve (sAUC) is measured when compared with training the model using target data only, where gain is measured in terms of percentage relative gain. Thus, positive gain reflects positive transfer and, analogously, negative gain reflects negative transfer.

We instantiate the proposed sign-transfer method to two linear models: linear regression for the risk prediction task and Support Vector Machines for the quality assessment task. In each case, we validate the proposed method with fixed sign importance ($\alpha = 1$) - denoted as Sign - and with varying tradeoff between sign agreement and coefficient magnitude ($0 \leq \alpha \leq 1$) - denoted as α-Sign. The proposed regularizers are compared to the state-of-the-art approach, hereafter referred as Diff, where the model is learned using full-observability transfer by regularizing coefficients to be similar to the source-model coefficients [5–8].

3.1 Risk Factors

In this section we instantiate the proposed partial transfer technique to predict the individual patient's risk when multiple screening strategies are available (i.e. colposcopy using acetic acid - Hinselmann, colposcopy using Lugol iodine - Schiller, cytology and biopsy). For this purpose a database with 858 patients including demographic information, habits and historic medical records was collected (see Table 1). Several patients decided not to answer some of the questions due to privacy concerns. Hence, the features denoted by bool $\times$ T, T $\in \{\text{bool}, \text{int}\}$, were encoded as two independent values: whether or not the patient answered the question and the reported value. Missing values were filled using the sample mean. Categorical features were encoded using the one-of-K scheme.

Table 1. Features acquired in the risk factors dataset.

Feature	Type	Feature	Type
Age	int	IUD (years)	int
# sexual partners	bool $\times$ int	STDs	bool $\times$ bool
Age of 1st sexual intercourse	bool $\times$ int	STDs (how many?)	int
# of pregnancies	bool $\times$ int	Diagnosed STDs	categorical
Smokes?	bool $\times$ bool	STDs (years since first diag.)	int
Smokes? (years & packs)	int $\times$ int	STDs (years last diag.)	int
Hormonal Contraceptives?	bool	Has previous cervical diag.?	bool
Horm. Contr.? (years)	int	Prev. cervical diag. (years)	int
Intrauterine device? (IUD)	bool	Prev. cervical diagnosis	categorical

Table 2 shows the results for this task using a regularized linear regression. It was validated that gains achieved by the proposed partial transfer framework were higher than the obtained by the fully observable transfer recently used in the literature. In most cases, the best results were obtained by the α-controlled sign regularization approach.

Table 2. sAUC obtained by the TL approaches on the risk prediction task with multiple screening strategies: Hinselmann (H), Schiller (S), Cytology (C) and Biopsy (B). Performance is measured in terms of Rooted Mean Squared Error (RMSE).

Source	**Target**	**Diff**	**Sign**	**α-Sign**	**Source**	**Target**	**Diff**	**Sign**	**α-Sign**
H	S	66.09	66.02	**68.96**	C	H	35.05	34.51	**35.11**
H	C	19.51	24.67	**37.12**	C	S	55.45	53.97	**55.81**
H	B	54.70	52.39	**54.96**	C	B	47.37	47.40	**47.54**
S	H	38.72	36.44	**38.74**	B	H	47.99	47.39	**48.80**
S	C	33.55	34.21	**39.90**	B	S	64.10	61.89	**66.66**
S	B	**45.48**	42.19	45.34	B	C	28.18	34.14	**43.69**

3.2 Quality Assessment

Choosing frames with good quality to perform the screening is an important step on improving physician's effectiveness. However, several challenges arise when defining the quality in this context. Thus, quality becomes a subjective concept subject to human preferences. In this work we consider a binary annotation scheme (e.g. good and bad quality) to simplify the presentation of the proposed framework. However, in the future we will consider ordinal scales (e.g. poor, fair, good, excellent) and pairwise relative preferences (e.g. the image A is better than the image B). The following semantic medical features were considered:

- Image area occupied by each anatomical body part (cervix, external os and vaginal walls) and occluding objects (speculum and other artifacts).
- The area of each region occluded by artifacts or by specular reflections.
- The maximum area difference between the four cervix quadrants.
- Fitness goodness of the cervix to a given geometric model: convex hull, bounding box, circle and ellipse.
- Distance between the image center and the cervix centroid/external os.
- Mean and standard deviation of each RGB and HSV channel in the cervix area and in the entire image.

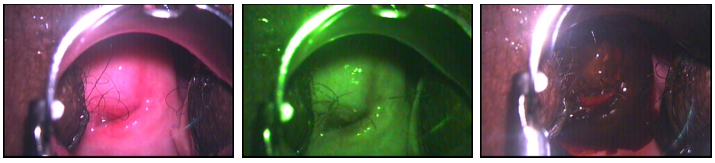

Fig. 2. Colposcopy modalities. From left to right: Hinselmann, Green light and Schiller.

In a joint collaboration with *Hospital Universitario de Caracas*, a dataset with annotations from 6 experts on about 100 cervigrams per modality (see Figure 2) was collected [1]. In the experimental evaluation, each region of interest

Table 3. sAUC obtained by the TL approaches on the quality prediction task with several colposcopic modalities: Hinselmann (H), Green (G) and Schiller (S). Performance is measured in terms of accuracy.

Source	**Target**	**Diff**	**Sign**	**α-Sign**	**Source**	**Target**	**Diff**	**Sign**	**α-Sign**
H	G	53.31	**54.14**	53.83	H	S	**47.82**	46.58	45.73
G	H	64.13	68.05	**68.30**	G	S	47.07	47.98	**48.15**
S	H	**63.73**	62.67	61.02	S	G	47.16	**49.28**	48.54

was manually segmented by an expert to simplify the comparison of the transfer learning approaches.

Table 3 shows the results for the binary classification of the subjective image quality using SVM. The target labels are assigned using the mode of the annotations given by the physicians. Contrary to the linear regression case, the version with $\alpha = 1$ obtained better results than the α-Sign approach. This can be explained by the fact that each modality has a few annotated instances per expert (about 100), turning it difficult to correctly estimate the α parameter.

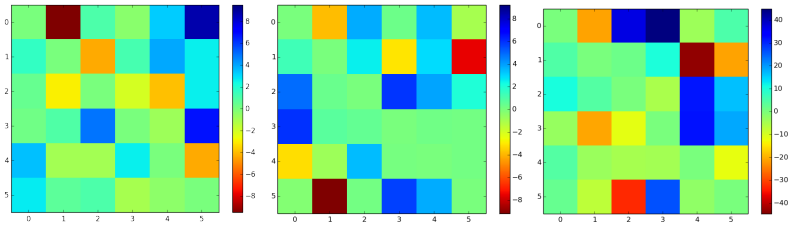

Fig. 3. Heatmap of the transfer gain obtained by the α-Sign regularizer when compared to the state-of-the-art regularizer. Transfer is done from a given expert's preferences (row) to another expert's preferences (column) between the same modality. The modalities are, from left to right: Hinselmann, Green light and Schiller.

Fig. 3 shows the gains obtained by the α-Sign version of the regularizer when compared with the state-of-the-art approach on a multi-expert setting. Here, source and target tasks represent different annotators' preferences (i.e. transferring from the i-th expert in the row to the j-th expert in the column). Analogously to previous experiments, the proposed transfer with partial observability obtained the best results in most cases. Schiller was the modality with highest gains. However, it was also the most unstable, being also the one with lowest gains in some cases. Using partial transfer schemes, some experts reflected poor performance as source in some modalities (e.g. expert 2 in Hinselmann) while behave as good sources in other modalities (e.g. expert 2 in Green). Moreover, since the partial model observability is a weak prior over the model space, the set of models that achieves an optimal regularization value is infinite, inducing a non-symmetric gain matrix.

4 Conclusions

In this work we presented a regularization-based TL approach to transfer the contribution type for each feature on linear models. In order to show its adequacy to different contexts, the proposed model-relatedness regularizer was instantiated to several learning tasks related to cervical cancer screening. Positive results were obtained in most experiments, being competitive with other methods in the literature. This work suggests that the analysis of how models encode high-level properties of the domain may improve transfer performance. Future research lines will tackle this type of transfer in multi-class and ordinal classification settings. Also, we will study how to synthesize high-level transferable knowledge in other non-linear models.

Acknowledgements

This work was funded by the Project "NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/-NORTE-01-0145-FEDER-000016" financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF), and also by Fundação para a Ciência e a Tecnologia (FCT) within PhD grant number SFRH/BD/93012/2013. The authors would like to thank the Gynecology Service of the *Hospital Universitario de Caracas*. In particular, we would like to recognize the efforts of Drs. Geramel Montero, Dulce Almeida, Jose Valentin, Leonardo Amado and Leticia Parpacen.

References

1. Fernandes, K., Cardoso, J.S., Fernandes, J.: Temporal segmentation of digital colposcopies. In: Pattern Recognition and Image Analysis. Springer (2015) 262–271
2. Pan, S.J., Yang, Q.: A survey on transfer learning. Knowledge and Data Engineering, IEEE Transactions on **22**(10) (2010) 1345–1359
3. Garcke, J., Vanck, T.: Importance weighted inductive transfer learning for regression. In: Machine Learning and Knowledge Discovery in Databases. Springer (2014) 466–481
4. Rückert, U., Kramer, S.: Kernel-based inductive transfer. In: Machine Learning and Knowledge Discovery in Databases. Springer (2008) 220–233
5. Evgeniou, T., Pontil, M.: Regularized multi–task learning. In: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM (2004) 109–117
6. Lee, C., Jang, M.G.: A prior model of structural SVMs for domain adaptation. ETRI Journal **33**(5) (2011) 712–719
7. Kuzborskij, I., Orabona, F.: Stability and hypothesis transfer learning. In: ICML (3). (2013) 942–950
8. Perrot, M., Habrard, A.: A theoretical analysis of metric hypothesis transfer learning. In: Proceedings of the 32nd International Conference on Machine Learning (ICML-15). (2015) 1708–1717