Transfer Learning in Astronomy:  
A New Machine-Learning Paradigm

Ricardo Vilalta  
Department of Computer Science, University of Houston.  
Philip G. Hoffman Hall, 3551 Cullen Blvd., Houston, TX 77204-3010, USA  
E-mail: rvilalta@uh.edu

Abstract. The widespread dissemination of machine learning tools in science, particularly in  
astronomy, has revealed the limitation of working with simple single-task scenarios in which  
any task in need of a predictive model is looked in isolation, and ignores the existence of other  
similar tasks. In contrast, a new generation of techniques is emerging where predictive models  
can take advantage of previous experience to leverage information from similar tasks. The new  
emerging area is referred to as Transfer Learning. In this paper, I briefly describe the motivation  
behind the use of transfer learning techniques, and explain how such techniques can be used  
to solve popular problems in astronomy. As an example, a prevalent problem in astronomy  
is to estimate the class of an object (e.g., supernova Ia) using a generation of photometric  
light-curve datasets where data abounds, but class labels are scarce; such analysis can benefit  
from spectroscopic data where class labels are known with high confidence, but data is of small  
size. Transfer learning provides a robust and practical solution to leverage information from one  
domain to improve the accuracy of a model built on a different domain. In the example above,  
transfer learning would look to overcome the difficulty in the compatibility of models between  
spectroscopic data and photometric data, since data properties such as size, class priors, and  
underlying distributions, are all expected to be significantly different.

1. Introduction

The abundance of large datasets generated for scientific research through sophisticated sensors  
(e.g., modern telescopes) or complex simulations, has led to a widespread interest for automated  
mechanisms that can analyze the data and generate models that classify events in an accurate  
and precise manner. A popular approach uses machine learning [8, 1] to train a computer  
to recognize different events using case examples that belong to different categories. In most  
applications, a common assumption pervading most traditional work in machine learning is that  
the probability distribution from which a training sample is drawn is static; future samples must  
follow the same distribution for any model to remain valid. While such assumption is sensible,  
and has found a plethora of successful real-world applications, recent work in machine learning  
has shown an equally large number of applications that do not follow such assumption.

An example of such discrepancy in sample distributions lies in light curve classification  
from star samples obtained from different galaxies [13]. An original source task may consist  
of identifying certain types of stars on a nearby galaxy. But if we move to galaxies lying farther  
away, and we try to repeat the same identification task, we will find that the distribution of  
stars has now changed (sometimes drastically). A major reason for such change is that at
Figure 1. Transfer Learning leverages information from previous knowledge.

greater distances, less luminous stars fall below the detection threshold and more luminous stars are preferentially detected. The corresponding change in distribution precludes the direct utilization of one single model across galaxies; it calls for a form of model adaptation to compensate for the change in the data distribution. Many other domains exist where changes in the underlying probability distribution are primarily caused by measurements obtained under different circumstances (e.g., observing galaxies that lie at different distances), by changing the orientation or position of the same sensing device, or by utilizing a similar but more powerful devices. In particle physics, for example, a model built to identify a certain particle is rendered unapplicable when we collect samples obtained using more powerful particle accelerators; this is because the range of parameter values shifts as we reach out to higher energies.

The abundance of examples exhibiting a shift in distribution as circumstances change over time calls for automated methods that adapt as a response to the dynamic nature of many learning tasks. The type of problems mentioned above have recently led to the development of a new area of study called transfer learning, where the mechanism leverages previous experience to increase the accuracy of predictive models. In this paper, I briefly introduce central ideas in transfer learning and describe a practical application in astronomy that captures the essence of this new paradigm.

1.1. Rationale for Transfer Learning

Before the advent of transfer learning techniques, practical solutions to the problem of shifts in probability distributions led to a continuous re-training of a predictive model over time. Each new instance of a classification or regression problem was considered in isolation, lacking the capacity to exploit information from previous experience. Figure 1(left) shows this operational mode. A training set serves as input to the process of modelling data using machine learning tools. Other datasets are treated similarly but in isolation. No information is exploited that can serve to learn experience across tasks. The traditional approach is particularly detrimental in situations where similar datasets share information that is useful across all tasks. This is clearly the case in the astronomical or particle-physics scenarios mentioned above, where a change in the sensor device or the event generator does not render any previous form of data analysis useless; instead one should be able to exploit the presence of similar patterns across tasks.

The new approach to the construction of adaptive learning models is to gather experience from previous tasks to improve on the current target task. This is also known as the transfer learning problem. Figure 1(right) depicts the new approach. Previous experience is now retained for future analysis (source domains). The repetitive use of machine learning is stored in a knowledge database that varies in nature according to the problem or task under analysis. The main difference with the previous approach becomes evident when a new target domain arrives (i.e., new task). The data modelling process is now reinforced with previous experience to
strengthen the predictive model, and gain leverage from patterns found in previous tasks. A brief description of these adaptive techniques is given in the following section.

2. Different Modalities in Transfer Learning

Before describing different approaches to transfer learning, I introduce some notation. It is common to assume a source dataset \( T_s = \{ (x, y) \} \) made of training examples where \( x \) is a feature vector, and \( y \) is a class or category label. Dataset \( T_s \), the source dataset, corresponds to a previous application of machine learning that produced a predictive model \( f_s(x) \); the model may need some form of adaptation because of changes in the original data distribution. The second dataset, \( T_t \), the target dataset, corresponds to a new application that is similar to the source task, but not identical. Now, rather than building \( f_t(x) \) from scratch, transfer learning can be invoked to exploit previous experience [3, 4].

Many techniques have been proposed in the area of transfer learning [9, 14]. The central idea is to exploit information from a source domain for which class labels are known with high confidence, on a target domain where class labels are scarce and the underlying distribution between the two domains is different.

and the target dataset \( T_t = \{ x_j \}_{j=1}^p \). The source dataset is drawn from a source distribution \( P_s(x, y) \). The target dataset is drawn from a marginal distribution \( P_t(x) \) according to a (different) target distribution, \( P_t(x, y) \), over \( X \times Y \). We will consider the case where both source and target domains differ in the marginal distributions \( P_s(x) \neq P_t(x) \), and class posteriors (i.e., conditional distributions) \( P_s(y|x) \neq P_t(y|x) \). Note that \( T_t \) omits all class labels, a problem known as unsupervised domain adaptation (as opposed to supervised domain adaptation where \( T_t \) contains an additional small set of labeled examples). The absence of labels poses a special challenge during classification. The central question is as follows: what can be exploited or transferred from the source domain to construct an accurate model on the target domain?

Many approaches have been proposed in domain adaptation. Broadly speaking, existing techniques can be placed along three different groups: instance-based, feature-based, and iterative based methods. The strategy behind instance-based methods is to increase the weight of source instances populating regions of high density in the target domain. A piece of work along these lines is known as covariate shift [2, 3]. Under the covariance-shift assumption, the model built on the new weighted source distribution can be directly applied to the target domain [2, 3]. A strong assumption here requires close proximity between source and target distributions.

Feature-based domain adaptation methods project both the source and the target datasets into a common feature space where the covariate-shift assumption holds. The new model built on the transformed space acts as the classifier on the target. Instances of this family include structural corresponding learning [2], and subspace alignment methods [2], among others [2].

Iterative-based methods build a model on the target domain by gradually shifting from a complete reliance on the source domain, to a state of sole reliance on the target domain. The central idea is to exploit source examples to initialize the target model; subsequent iterations continuously refine the model while pushing aside the influence provided by source examples [2].

Theoretical work in domain adaptation includes studies to estimate the distance between source and target distributions [2, 3, 2]; and using regularization terms to learn models that perform well on both source and target domains [2].
An Application of Transfer Learning in Astronomy

An application of transfer learning in astronomy can be found in the automated identification of Supernova Ia [7]. The problem is of great importance to astronomy, because SNe Ia are standard candles instrumental in probing large cosmological distances; the correlation between their luminosity and distance independent quantities has led to multiple discoveries, including the accelerating expansion of the Universe [10]. The discrepancy in distribution that makes this a problem amenable to transfer learning is due to the difference between spectroscopic and photometric observations. Spectroscopy provides a high-resolution description of electromagnetic radiation and is crucial to estimate the chemical composition (through spectral lines) and distances (redshift) with high precision. But spectroscopic observations are costly and time-consuming; in practice, it is more convenient to perform photometric observations that summarize radiation in a set of broad wavelength windows or filters, albeit by forfeiting information from spectral lines.

Many modern surveys including the Sloan Digital Sky Survey, the Dark Energy Survey, and the upcoming Large Synoptic Survey (Telescope) capture pure photometric observations for SNe Ia. A big challenge in modern astronomy is to infer spectroscopic properties from purely photometric data. In our current discussion, the source domain corresponds to spectroscopic data where supernovae are confidently classified (as type Ia or different); the target domain corresponds to the new generation of abundant photometric light-curve datasets where class labels are scarce. Our goal is to take advantage of the source domain to attain high predictive performance on the target domain. The use of transfer learning here is crucial, since photometric observations lack a precise class label (e.g. type of SNe) to conform a reliable training set. At the same time, the naive approach of training a model on the spectroscopic data and applying it directly on the photometric data is prone to failure because of the distributional shift observed between both types of observations.

Methodology

I now describe a study aimed at building accurate SNe Ia classifiers from photometric data (target) while exploiting information from spectroscopic data (source) [7]. The first step is almost always invariably to reduce the dimensionality of the (source and target) data, since the original samples contain a large number of features. A recommended approach is to use Kernel Principal Component Analysis (Kernel PCA) [12]; the technique is a generalization of Principal Component Analysis (PCA) by using non-linear components. Recently, new approaches have emerged that compute equally robust features, such as dimensionality reduction techniques using deep learning architectures [11].

The second step is to use transfer learning to leverage experience from the source domain (spectroscopic dataset) to attain an accurate classifier on the target domain (photometric dataset). The study reported here considered two methods used: Kernel Mean Matching KMM, and Subspace Alignment SA. The first method, KMM, works through a re-weighting scheme that gives importance to those examples on the source dataset that appear close to the target dataset [6]. Specifically, the method projects the data into a new space (reproducing Hilbert kernel space) where it minimizes the (maximum) distance between the means on each distribution (source and target). The projection helps to identify source examples that can be incorporated when building a classifier on the target domain.

The second method, Subspace Alignment, is a common domain adaptation method based on feature subset selection [5]. The idea is to apply PCA on source $T_s$ and target $T_t$ datasets separately by choosing a common space. It then attempts to align the projected source dataset with the projected target dataset in this common subspace using a subspace alignment matrix. Once source and target are aligned, a classifier is built on the transformed source dataset $T^\alpha_s$ and subsequently applied to the transformed target dataset $T^\alpha_t$.
Empirical Results

Experimental results are shown in Figure 2. The horizontal axis

Active Learning. Our experimental work shows how the direct application of domain adaptation techniques does improve generalization performance, albeit to a small degree. To boost the performance of the predictive model on the target dataset, it is necessary to invoke active-learning techniques. Active learning points to those few instances on the target set where knowing the class label with confidence suffices to attain an accurate model. We conclude by showing how the automated classification of Supernovae Ia is best achieved when domain adaptation and active learning are combined synergistically.

We now describe our approach to combine domain adaptation with active learning. As mentioned before, active learning selectively chooses (unlabeled) examples deemed relevant for classification. In our study, we begin by randomly choosing a set of instances from the target dataset and querying them (each class label incurs a fixed cost). A model is then built on this labeled target sample, followed by an iterative process that queries the next most informative instance from the target dataset, and builds a new model after the last queried instance is added to the sample. The algorithm stops when it reaches a maximum cost i.e., when it runs out of budget.

In this paper we use pool-based active learning with margin sampling as the uncertainty-sampling technique. Here, the most informative instance to query, $x^*$, is selected based on the concept of margin as a measure of class-label uncertainty. While many interpretations exist for the margin, here we define it as the difference between the first $P(y' | x)$ and second $P(y'' | x)$ highest class posterior probabilities conditioned on $x$. Instance $x^*$ is then selected as follows:

![Figure 2. Accuracy on Supernova Target Data with and without Domain Adaptation.](image)

![Figure 3. Accuracy and Precision on Target Data using Domain Adaptation and Active Learning.](image)

Our approach to combine domain adaptation with active learning is described in
Algorithm ??, We use the model $M_s$ created by the domain adaptation algorithm as the initial model for active learning. This step obviates randomly querying instances from the target dataset to create an initial labeled sample, and can be seen as a form of transfer of knowledge between source and target. The final goal is to build an accurate model $M_t$ from the target domain.

Active learning [2]

4. Conclusions
The traditional approach to machine learning where

4.1. Acknowledgments
This work was partly supported by the Center for Advanced Computing and Data Systems (CACDS), and by the Texas Institute for Measurement, Evaluation, and Statistics (TIMES) at the University of Houston.

References
[1] Abu-Mostafa, Y. S.; Magdon-Ismail, M.; Lin, H.-T. 2012. Learning from Data, volume 4. AMLBook Singapore.
[2] Balcan, M.-F.; Beygelzimer, A.; Langford, J. 2009 Agnostic active learning. Journal of Computer and System Sciences 75(1):78 – 89.
[3] Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; Vaughan, J. 2010 A theory of learning from different domains. Machine Learning, vol. 79, pp. 151175, 2010.
[4] Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; Wortman, J. 2008 Learning Bounds for Domain Adaptation. Advances in Neural Information Processing Systems 20. Platt, J. C.; Koller, D.; Singer, Y.; Roweis, S. T.; Eds. Curran Associates, Inc., 2008, pp. 129136.
[5] Fernando, B.; Habrard, A.; Sebban, M.; Tuytelaars, T. Subspace Alignment for Domain Adaptation. CoRR, vol. abs/1409.5241.
[6] Gretton, A.; Smola, A.; Schmitfull, M.; Borgwardt, K.; Scholkopf, B. 2009 Covariate shift by kernel mean matching. Dataset shift in Machine Learning, vol. 3, no. 4.
[7] Dhar Gupta, K.; Pampana, R.; Vilalta, R.; Ishida, E. E. O.; de Souza, R. S. 2016 Automated Supernova Ia Classification Using Adaptive Learning Techniques. IEEE Symposium on Computational Intelligence and Data Mining (CIDM-16), Athens, Greece.
[8] Hastie, T.; Tibshirani, R.; Friedman, J. 2009 The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, 2nd edition.
[9] Pan, S. J.; Yang, Q. 2010 A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering 22(10):1345–1359.
[10] Riess, A. G., et al. 1998 Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. The Astronomical Journal, AJ, V. 116, No. 3, 1009.
[11] Sasdelli, M., et al. 2016 Exploring the Spectroscopic Diversity of Type Ia Supernovae with DRACULA: A Machine Learning Approach. Monthly Notices of the Royal Astronomical Society (MNRAS), 461, no.2, pp. 2044-2059, Oxford University Press.
[12] Scholkopf, B.; Smola, A.; Muller, K.-R. 1997 Kernel Principal Component Analysis. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 583588.
[13] Vilalta, R.; Dhar Gupta K.; Macri, L. 2013 A Machine Learning Approach to Cepheid Variable Star Classification Using Data Alignment and Maximum Likelihood. Astronomy and Computing Journal, Vol. 2, pp. 46-53. Elsevier. DOI: 10.1016/j.ascom.2013.07.002.
[14] Weiss, K.; Khoshgoftar, T.; Wang, D. 2016 A Survey of Transfer Learning Journal of Big Data 3:9 DOI 10.1186/s40537-016-0043-6.