What's in a Domain? Multi-Domain Learning for Multi-Attribute Data

Mahesh Joshi[⋆] **Mark Dredze**[†] **William W. Cohen**[⋆] **Carolyn P. Rosé**[⋆]
[⋆] School of Computer Science, Carnegie Mellon University
Pittsburgh, PA, 15213, USA
[†] Human Language Technology Center of Excellence, Johns Hopkins University
Baltimore, MD, 21211, USA
maheshj@cs.cmu.edu,mdredze@cs.jhu.edu
wcohen@cs.cmu.edu,cprose@cs.cmu.edu

Abstract

Multi-Domain learning assumes that a single metadata attribute is used in order to divide the data into so-called domains. However, real-world datasets often have multiple metadata attributes that can divide the data into domains. It is not always apparent which single attribute will lead to the best domains, and more than one attribute might impact classification. We propose extensions to two multi-domain learning techniques for our *multi-attribute* setting, enabling them to simultaneously learn from several metadata attributes. Experimentally, they outperform the multi-domain learning baseline, even when it selects the single "best" attribute.

1 Introduction

Multi-Domain Learning (Evgeniou and Pontil, 2004; Daumé III, 2007; Dredze and Crammer, 2008; Finkel and Manning, 2009; Zhang and Yeung, 2010; Saha et al., 2011) algorithms learn when training instances are spread across many domains, which impact model parameters. These algorithms use examples from each domain to learn a general model that is also sensitive to individual domain differences.

However, many data sets include a host of metadata attributes, many of which can potentially define the domains to use. Consider the case of restaurant reviews, which can be categorized into domains corresponding to the cuisine, location, price range, or several other factors. For multi-domain learning, we should use the metadata attribute most likely to characterize a domain: a change in vocabulary (i.e. features) that most impacts the classification decision (Ben-David et al., 2009). This choice is not easy. First, we may not know which metadata attribute is most likely to fit this role. Perhaps the location most impacts the review language, but it could easily be the price of the meal. Second, multiple metadata attributes could impact the classification decision, and picking a single one might reduce classification accuracy. Therefore, we seek multi-domain learning algorithms which can simultaneously learn from many types of domains (metadata attributes).

We introduce the *multi-attribute multi-domain* (MAMD) learning problem, in which each learning instance is associated with multiple metadata attributes, each of which may impact feature behavior. We present extensions to two popular multi-domain learning algorithms, FEDA (Daumé III, 2007) and MDR (Dredze et al., 2009). Rather than selecting a single domain division, our algorithms consider all attributes as possible distinctions and discover changes in features across attributes. We evaluate our algorithms using two different data sets – a data set of restaurant reviews (Chahuneau et al., 2012), and a dataset of transcribed speech segments from floor debates in the United States Congress (Thomas et al., 2006). We demonstrate that multi-attribute algorithms improve over their multi-domain counterparts, which can learn distinctions from only a single attribute.

2 MAMD Learning

In multi-domain learning, each instance $\mathbf{x}$ is drawn from a domain d with distribution $\mathbf{x} \sim \mathcal{D}_d$ over a vectors space $\mathbb{R}^D$ and labeled with a domain specific function f_d with label $y \in \{-1, +1\}$ (for binary classification). In multi-attribute multi-domain

Proceedings of NAACL-HLT 2013, pages 685–690,
Atlanta, Georgia, 9–14 June 2013.

(MAMD) learning, we have M metadata attributes in a data set, where the mth metadata attribute has K_m possible unique values which represent the domains induced by that metadata attribute. Each instance $\mathbf{x}_i$ is drawn from a distribution $\mathbf{x}_i \sim \mathcal{D}_a$ specific to a *set* of attribute values $\mathcal{A}_i$ associated with each instance. Additionally, each unique set of attributes indexes a function $f_{\mathcal{A}}$.[1] $\mathcal{A}_i$ could contain a value for each attribute, or no values for any attribute (which would index a domain-agnostic "background" distribution and labeling function). Just as a domain can change a feature's probability and behavior, so can each metadata attribute.

Examples of data for MAMD learning abound. The commonly used Amazon product reviews data set (Blitzer et al., 2007) only includes product types, but the original reviews can be attributed with author, product price, brand, and so on. Additional examples include congressional floor debate records (e.g. political party, speaker, bill) (Joshi et al., 2012). In this paper, we use restaurant reviews (Chahuneau et al., 2012), which have upto 20 metadata attributes that define domains, and congressional floor debates, with two attributes that define domains.

It is difficult to apply multi-domain learning algorithms when it is unclear which metadata attribute to choose for defining the "domains". It is possible that there is a single "best" attribute to use for defining domains, one that when used in multi-domain learning will yield the best classifier. To find this attribute, one must rely on one's intuition about the problem,[2] or perform an exhaustive empirical search over all attributes using some validation set. Both these strategies can be brittle, because as the nature of data changes over time so may the "best" domain distinction. Additionally, multi-domain learning was not designed to benefit from multiple helpful attributes.

We note here that Eisenstein et al. (2011), as well as Wang et al. (2012), worked with a "multifaceted topic model" using the framework of sparse additive generative models (SAGE). Both those models capture interactions between topics and multiple aspects, and can be adapted to the case of MAMD. While our problem formulation has significant conceptual overlap with the SAGE–like multifaceted topic models framework, our proposed methods are motivated from a fast online learning perspective.

A naive approach for MAMD would be to treat every unique set of attributes as a domain, including unique proper subsets of different attributes to account for the case of missing attributes in some instances.[3] However, introducing an exponential number of domains requires a similar increase in training data, clearly an infeasible requirement. Instead, we develop multi-attribute extensions for two multi-domain learning algorithms, such that the increase in parameters is linear in the number of metadata attributes, and no special handling is required for the case where some metadata attributes might be missing from an instance.

Multi-Attribute FEDA The key idea behind FEDA (Daumé III, 2007) is to encode each domain using its own parameters, one per feature. FEDA maps a feature vector $\mathbf{x}$ in $\mathbb{R}^D$ to $\mathbb{R}^{D(K+1)}$. This provides a separate parameter sub-space for every domain $k \in 1 \ldots K$, and also maintains a domain-agnostic shared sub-space. Essentially, each feature is duplicated for every instance in the appropriate sub-space of $\mathbb{R}^{D(K+1)}$ that corresponds to the instance's domain. We extend this idea to the MAMD setting by using one parameter per attribute value. The original instance $\mathbf{x} \in \mathbb{R}^D$ is now mapped into $\mathbb{R}^{D(1+\sum_m K_m)}$; a separate parameter for each attribute value and a shared set of parameters. In effect, for every metadata attribute $a \in \mathcal{A}_i$, the original features are copied into the appropriate sub-space. This grows linearly with the number of metadata attribute values, as opposed to exponentially in our naive solution. While this is still substantial growth, each instance retains the same feature sparsity as in the original input space. In this new setup, FEDA allows an instance to contribute towards learning the shared parameters, and the attribute-specific parameters for all the attributes present on an instance. Just like multi-domain FEDA, any supervised learning algorithm can be applied to the transformed representation.

[1] Distributions and functions that share attributes could share parameters.

[2] Intuition is often critical for learning and in some cases can help, such as in the Amazon product reviews data set, where product type clearly corresponds to domain. However, for other data sets the choice may be less clear.

[3] While we used a similar setup for formulating our problem, we did not rule out the potential for factoring the distributions.

Multi-Attribute MDR We make a similar change to MDR (Dredze et al., 2009) to extend it for the MAMD setting. In the original formulation, Dredze et al. used confidence-weighted (CW) learning (Dredze et al., 2008) for learning shared and domain-specific classifiers, which are combined based on the confidence scores associated with the feature weights. For training the MDR approaches in a multi-domain learning setup, they found that computing updates for the combined classifier and then equally distributing them to the shared and domain-specific classifiers was the best strategy, although it approximated the true objective that they aimed to optimize. In our multi-attribute setup confidence-weighted (CW) classifiers are learned for each of the $\sum_m K_m$ attribute values in addition to a shared CW classifier. At classification time, a combined classifier is computed for every instance. However, instead of combining the shared classifier and a *single* domain-specific classifier, we combine the shared CW classifier and $|\mathcal{A}_i|$ different attribute value-specific CW classifiers associated with $\mathbf{x}_i$. The combined classifier is found by minimizing the KL-divergence of the combined classifier with respect to each of the underlying classifiers.[4]

When learning the shared and domain-specific classifiers, we follow the best result in Dredze et al. and use the "averaged update" strategy (§7.3 in Dredze et al.), where updates are computed for the combined classifier, and are then distributed to the shared and domain-specific classifiers. MDR-U will indicate that the updates to the combined classifiers are *uniformly* distributed to the underlying shared and domain-specific classifiers.

Dredze et al. also used another scheme called "variance" to distribute the combined update to the underlying classifiers (§4, last paragraph in Dredze et al.) Their idea was to give a lower portion of the update to the underlying classifier that has higher variance (or in their terminology, "less confidence") since it contributed less to the combined classifier. We refer to this as MDR-V. However, this conflicts with the original CW intuition that features with higher variance (lower confidence) should receive higher updates; since they are more in need of change. Therefore, we implemented a modified "variance" scheme, where the updates are distributed to the underlying classifiers such that higher variance features receive the larger updates. We refer to this as MDR-NV. We observed significant improvements with this modified scheme.

3 Experiments

To evaluate our multi-attribute algorithms we consider two datasets. First, we use two subsets of the restaurant reviews dataset (1,180,308 reviews) introduced by Chahuneau et al. (2012) with the goal of labeling reviews as positive or negative. The first subset (50K-RND) randomly selects 50,000 reviews while the second (50K-BAL) is a class-balanced sample. Following the approach of Blitzer et al. (2007), scores above and below 3-stars indicated positive and negative reviews, while 3-star reviews were discarded. Second, we use the transcribed segments of speech from the United States Congress floor debates (Convote), introduced by Thomas et al. (2006). The binary classification task on this dataset is that of predicting whether a given speech segment supports or opposes a bill under discussion in the floor debate.

In the WordSalad datasets, each restaurant review can have many metadata attributes, including a unique identifier, name (which may not be unique), address (we extract the zipcode), and type (Italian, Chinese, etc.). We select the 20 most common metadata attributes (excluding latitude, longitude, and the average rating). [5] In the Convote dataset, each speech segment is associated with the political party affiliation of the speaker (democrat, independent, or republican) and the speaker identifier (we use bill identifiers for creating folds in our 10-fold cross-validation setup).

In addition to our new algorithms, we evaluate several baselines. All methods use confidence-weighted (CW) learning (Crammer et al., 2012).

BASE A single classifier trained on all the data, and which ignores metadata attributes and uses unigram features. For CW, we use the best-performing setting from Dredze et al. (2008) — the "variance" algorithm, which computes approximate but closed–form updates, which also lead to faster learning. Parameters are tuned over a validation set within each training fold.

[4] We also tried the l_2 distance method of Dredze et al. (2009) but it gave consistently worse results.

[5] Our method requires categorical metadata attributes, although real-valued attributes can be discretized.

	metadata	1-META	FEDA	MDR-U	MDR-V	MDR-NV
50K-RND	NONE (BASE)	92.29 (±0.14)				
	ALL (META)	† 92.69 (±0.10)				
	CATEGORY	† 92.48 (±0.11)	92.47 (±0.10)	†‡ 92.99 (±0.12)	91.16 (±0.16)	†‡ **93.24 (±0.13)**
	ZIPCODE	92.40 (±0.09)	† 92.73 (±0.09)	†‡ 92.99 (±0.12)	91.19 (±0.20)	†‡ **93.22 (±0.11)**
	NEIGHBORHOOD	92.42 (±0.11)	† 92.65 (±0.13)	†‡ 93.02 (±0.13)	91.17 (±0.21)	†‡ **93.21 (±0.12)**
50K-BAL	NONE (BASE)	89.95 (±0.10)				
	ALL (META)	† 90.39 (±0.09)				
	CATEGORY	90.09 (±0.11)	† 90.50 (±0.11)	† 90.60 (±0.11)	87.89 (±0.13)	†‡ **91.33 (±0.08)**
	ZIPCODE	89.97 (±0.12)	† 90.42 (±0.13)	† 90.56 (±0.09)	87.78 (±0.16)	†‡ **91.30 (±0.10)**
	ID	† 90.42 (±0.11)	†‡ 90.64 (±0.11)	† 90.50 (±0.11)	87.78 (±0.25)	†‡ **91.27 (±0.09)**

Table 1: Average accuracy (± standard error) for the best three metadata attributes, when using a single attribute at a time. Results that are *numerically the best* within a row are in **bold**. Results significantly better than BASE are marked with †, and better than META are marked with ‡. Significance is measured using a two-tailed paired t-test with $\alpha = 0.05$.

	#attributes	FEDA	MDR-U	MDR-V	MDR-NV
50K-RND	MAMD	†‡ 93.07 (±0.19)	†‡ 93.12 (±0.11)	87.08 (±1.72)	†‡ **93.19 (±0.12)**
	1-ORCL	†‡ 93.06 (±0.11)	†‡ 93.17 (±0.11)	92.37 (±0.11)	†‡ **93.39 (±0.12)**
	1-TUNE	† 92.64 (±0.12)	† 92.81 (±0.16)	92.15 (±0.17)	†‡ **93.07 (±0.14)**
	1-MEAN	† **92.61 (±0.09)**	† 92.59 (±0.10)	91.41 (±0.12)	† 92.58 (±0.10)
50K-BAL	MAMD	†‡ **91.42 (±0.09)**	†‡ 91.06 (±0.04)	81.43 (±2.79)	†‡ 91.40 (±0.08)
	1-ORCL	†‡ 90.89 (±0.10)	†‡ 90.87 (±0.11)	89.33 (±0.13)	†‡ **91.45 (±0.07)**
	1-TUNE	† 90.33 (±0.10)	†‡ 90.70 (±0.14)	89.13 (±0.16)	†‡ **91.26 (±0.08)**
	1-MEAN	† **90.30 (±0.06)**	89.92 (±0.07)	88.25 (±0.07)	90.06 (±0.08)

Table 2: Average accuracy (± standard error) using 10-fold cross-validation for methods that use all attributes, either directly (our proposed methods) or for selecting the "best" single attribute using one of the strategies described earlier. Formatting and significance symbols are the same as in Table 1.

META Identical to BASE with a unique bias feature added for each attribute value (Joshi et al., 2012).

1-META A special case of META where a unique bias feature is added *only for a single attribute*.

To use multi-domain learning directly, we could select a single attribute as the domain. We consider several strategies for picking this attribute and evaluate both FEDA and MDR in this setting.

1-MEAN Choose an attribute randomly, equivalent to the expected (mean) error over all attributes.

1-TUNE Select the best performing attribute on a validation set.

1-ORCL Select the best performing attribute on the *test* set. Though impossible in practice, this gives the oracle upper bound on multi-domain learning.

All experiments use ten-fold cross-validation. We report the mean accuracy, along with standard error.

4 Results

Table 1 shows the results of single-attribute multi-domain learning methods for the WordSalad datasets. The table shows the three best-performing metadata attributes (as decided by the highest accuracy among all the methods across all 20 metadata attributes). Clearly, several of the attributes can provide meaningful domains, which demonstrates that methods that can select multiple attributes at once are desirable. We also see that our modification to MDR (MDR-NV) works the best.

Table 3 shows the results of single-attribute multi-domain learning methods for the Convote dataset. The first observation to be made on this dataset is that neither the PARTY, nor the SPEAKER attribute individually achieve significant improvement over the META baseline, which uses both these attributes as features. This is in contrast with the results on the WordSalad dataset, where some attributes by themselves showed an improvement over the META baseline. Thus, this dataset represents a more challenging setup for our multi–attribute multi–domain learning methods — they need to exploit the two weak attributes simultaneously.

We next demonstrate multi-attribute improvements over the multi-domain baselines (Tables 2 and 4). For WordSalad datasets, our extensions that can use all metadata attributes simultaneously are consistently better than both the 1-MEAN and the 1-TUNE strategies (except for the case of the old variance scheme used by (Dredze et al., 2009)). For the skewed subset

metadata	1-META	FEDA	MDR-U	MDR-V	MDR-NV
NONE (BASE)	67.08 (±1.74)				
ALL (META)	† 82.60 (±1.95)				
PARTY	† 78.81 (±1.47)	† **84.19 (±2.44)**	† 83.23 (±2.48)	† 81.38 (±2.22)	† 83.92 (±2.31)
SPEAKER	† 77.49 (±1.75)	† **82.88 (±2.43)**	† 78.32 (±1.91)	62.43 (±2.20)	† 72.26 (±1.37)

Table 3: Convote: Average accuracy (± standard error) when using a single attribute at a time. Results that are *numerically the best* within a row are in **bold**. Results significantly better than BASE are marked with †, and better than META are marked with ‡. Significance is measured using a two-tailed paired t-test with $\alpha = 0.05$.

#attributes	FEDA	MDR-U	MDR-V	MDR-NV
MAMD	†‡ 85.71 (±2.74)	† 84.12 (±2.56)	50.44 (±1.78)	†‡ **86.19 (±2.49)**
1-ORCL	† **84.77 (±2.47)**	† 83.88 (±2.27)	† 81.38 (±2.22)	† 83.92 (±2.31)
1-TUNE	† **84.19 (±2.44)**	† 83.23 (±2.48)	† 81.38 (±2.22)	† 83.92 (±2.31)
1-MEAN	† **83.53 (±2.40)**	† 80.77 (±1.92)	† 71.91 (±1.82)	† 78.09 (±1.69)

Table 4: Convote: Average accuracy (± standard error) using 10-fold cross-validation for methods that use all attributes, either directly (our proposed methods) or for selecting the "best" single attribute using one of the strategies described earlier. Formatting and significance symbols are the same as in Table 3.

50K-RND, MAMD+FEDA is significantly better than 1-TUNE+FEDA; MAMD+MDR-U is significantly better than 1-TUNE+MDR-U; MAMD+MDR-NV is not significantly different from 1-TUNE+MDR-U. For the balanced subset 50K-BAL, a similar pattern holds, except that MAMD+MDR-NV is significantly better than 1-TUNE+MDR-NV. Clearly, our multi-attribute algorithms provide a benefit over existing approaches. Even with oracle knowledge of the test performance using multi-domain learning, we can still obtain improvements (FEDA and MDR-U in the 50K-BAL set, and all the Convote results, except MDR-V).

Although MAMD+MDR-NV is not significantly better than 1-TUNE+MDR-NV on the 50K-RND set, we found that in every single fold in our ten-fold cross-validation experiments, the "best" single metadata attribute decided using a validation set did not match the best-performing single metadata attribute on the corresponding test set. This shows the potential instability of choosing a single best attribute. Also, note that MDR-NV is a variant that we have proposed in the current work, and in fact for the earlier variant of MDR (MDR-U), as well as for FEDA, we do see significant improvements when using all metadata attributes. Furthermore, the computational cost of evaluating every metadata attribute independently to tune the single best metadata attribute can be high and often impractical. Our approach requires no such tuning. Finally, observe that for FEDA, the 1-TUNE strategy is not significantly different from 1-MEAN, which just randomly picks a single best metadata attribute. For MDR-U, 1-TUNE is significantly better than 1-MEAN on the balanced subset 50K-BAL, but not on the skewed subset 50K-RND.

As mentioned earlier, the Convote dataset is a challenging setting for our methods due to the fact that no single attribute is strong enough to yield improvements over the META baseline. In this setting, both MAMD+FEDA and MAMD+MDR-NV achieve a significant improvement over the META baseline, with MDR-NV being the best (though not significantly better than FEDA). Additionally, both of them are significantly better than their corresponding 1-TUNE strategies. This result further supports our claim that using multiple attributes in combination for defining domains (even when any single one of them is not particularly beneficial for multi–domain learning) is important.

5 Conclusions

We propose multi-attribute multi-domain learning methods that can utilize multiple metadata attributes simultaneously for defining domains. Using these methods, the definition of "domains" does not have to be restricted to a single metadata attribute. Our methods achieve a better performance on two multi-attribute datasets as compared to traditional multi-domain learning methods that are tuned to use a single "best" attribute.

Acknowledgments

This research is supported by the Office of Naval Research grant number N000141110221.

References

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2009. A theory of learning from different domains. *Machine Learning*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447. Association for Computational Linguistics.

Victor Chahuneau, Kevin Gimpel, Bryan R. Routledge, Lily Scherlis, and Noah A. Smith. 2012. Word Salad: Relating Food Prices and Descriptions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP 2012)*.

Koby Crammer, Mark Dredze, and Fernando Pereira. 2012. Confidence-weighted linear classification for text categorization. *Journal of Machine Learning Research (JMLR)*.

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263. Association for Computational Linguistics.

Mark Dredze and Koby Crammer. 2008. Online methods for multi-domain learning and adaptation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*.

Mark Dredze, Koby Crammer, and Fernando Pereira. 2008. Confidence-weighted linear classification. *Proceedings of the 25th international conference on Machine learning - ICML '08*.

Mark Dredze, Alex Kulesza, and Koby Crammer. 2009. Multi-domain learning by confidence-weighted parameter combination. *Machine Learning*, 79(1–2):123–149.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse Additive Generative Models of Text. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*.

Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi–task learning. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*.

Jenny R Finkel and Christopher D Manning. 2009. Hierarchical Bayesian Domain Adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–610. Association for Computational Linguistics.

Mahesh Joshi, Mark Dredze, William W. Cohen, and Carolyn P. Rosé. 2012. Multi-domain learning: When do domains matter? In *Proceedings of EMNLP-CoNLL 2012*, pages 1302–1312.

Avishek Saha, Piyush Rai, Hal Daumé III, and Suresh Venkatasubramanian. 2011. Online learning of multiple tasks and their relationships. In *Proceedings of AISTATS 2011*.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, pages 327–335.

William Yang Wang, Elijah Mayfield, Suresh Naidu, and Jeremiah Dittmar. 2012. Historical Analysis of Legal Opinions with a Sparse Mixed-Effects Latent Variable Model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*.

Yu Zhang and Dit-Yan Yeung. 2010. A Convex Formulation for Learning Task Relationships in Multi-Task Learning. In *Proceedings of the Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*.