Active Zero-Shot Learning

Sihong Xie[¶] Shaoxiong Wang[§] Philip S. Yu[†]
[†]Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA
[§]Department of Computer Science, Tsinghua University, Beijing, China
[¶]Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, USA

ABSTRACT

In multi-label classification in the big data age, the number of classes can be in thousands, and obtaining sufficient training data for each class is infeasible. Zero-shot learning aims at predicting a large number of unseen classes using only labeled data from a small set of classes and external knowledge about class relations. However, previous zero-shot learning models passively accept labeled data collected beforehand, relinquishing the opportunity to select the proper set of classes to inquire labeled data and optimize the performance of unseen class prediction. To resolve this issue, we propose an active class selection strategy to intelligently query labeled data for a parsimonious set of informative classes. We demonstrate two desirable probabilistic properties of the proposed method that can facilitate unseen classes prediction. Experiments on 4 text datasets demonstrate that the active zero-shot learning algorithm is superior to a wide spectrum of baselines. We indicate promising future directions at the end of this paper.

INTRODUCTION

In some real world applications such as image, text and video classification, it is not uncommon to have tens of thousands of classes to predict. It becomes extremely difficult for traditional classification methods to learn the mapping from data and each and every class, as collecting training data for all classes is simply infeasible. For example, when one wants to identify many potentially useful tags for online contents like blogs, videos and images, it is markedly laborious to exhaust *all* possible tags and assign each piece of content to its relevant classes. Therefore, many classes will not have labeled data and traditional classification models will fail. Zero-shot learning utilizes external knowledge bases describing class relations to predict classes that do not have any training data (the "unseen classes") [5, 7, 9]. Usually, zero-shot learning algorithms first map instances to intermediate attributes, which can be seen classes (those with labeled data), human-specified or data-dependent attributes. Then the predicted attributes are mapped to a large number of unseen classes through the knowledge bases. In this way, prediction of unseen classes become possible and no training data is necessary for those classes.

Zero-shot learning has been studied for more than a decade [1] from various aspects. Two fundamental zero-shot paradigms, direct attribute prediction (DAP) and indirect attribute prediction (IAP) are given in [5]. More sophisticated zero-shot models are proposed, such as max-margin semi-supervised learning to exploit the unlabeled data [6], and multi-view zero-shot learning utilizing multiple data sources [3]. Multiple knowledge bases such as WordNet [7, 12], Wikipedia [2, 9], web search logs [8] and human-annotated images [5] are compared. The authors in [2, 9, 13] propose to learn the intermediate attributes using deep learning.

Although the zero-shot learning literature has addressed some of the crucial issues, most existing works assume that a zero-shot model can only passively learn from labeled data collected for a pre-determined and fixed subset of classes [6, 9]. However, given the ever-increasing amount of online contents, potentially hundreds of thousands of classes can be identified. It is infeasible to tag large amoutn of relevant contents for each class, while which classes have higher priority to be tagged is a question. Instead, one has to actively decide which classes are the most useful to collect labeled data to train a zero-shot model to predict the remaining classes. The key is that complex dependencies exist among classes such that some classes provide more global information about the other classes, and unseen class predictions can benefit from the "properly" selected seen classes. The work [10] proposed a representative class selection strategy. Our experiments shows that representativeness is not a proper metric for seen class selection.

To solve the above challenges, we propose to actively select a parsimonious set of informative classes to collect labeled data, and keep the large number of remaining classes unseen to save labeling efforts. Traditional active learning strategies query labels of instances for *all* instead of *some* classes [11], and the effectiveness of the learning methods is not related to the seen-unseen class connections, which we study here. We formulate zero-shot learning as a two-phase procedure, corresponding to which the informativeness of a seen class can be characterized by its discriminative power (accuracy) and the information it provides for the unseen classes (connectivity) via the external knowledge base. The proposed strategy exploits the entropy of inter-class similarities to measure the above two aspects of a class. We discover that the inter-class similarity follows a beta distribution, based on which we further reveal the relationship between the entropy and the probability that an unseen class is sufficiently connected to the seen ones. Extensive experiments on 4 text classification datasets with up to thousands of classes show that by obtaining labeled data for a small number of classes, we are able to significantly improve the unseen class prediction, compared with other active class selection strategies.

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983866

THE PROPOSED APPROACH

Let the set of d seen classes be denoted by $\mathcal{S}$ and the set of k unseen ones be denoted by $\mathcal{U}$. Without loss of generality, assume that the seen classes are indexed by $\{1, \ldots, d\}$, and the remaining unseen classes are indexed by $\{d+1, \ldots, d+k\}$. Training data is given by $\mathcal{D}_0 = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$, where $\mathbf{x}_i \in \mathcal{X} = \mathbb{R}^p$ is the feature vector of the i-th instance and $\mathbf{y}_i \in \mathcal{Y} = \{0,1\}^d$ is the corresponding seen class vector. Zero-shot learning aims at predicting the ground truth of the k unseen classes $\mathbf{z} \in \mathcal{Z} = \{0,1\}^k$ for any test data $\mathbf{x} \in \mathcal{X}$. The above procedure can be captured by two mappings: $f : \mathcal{X} \to \mathcal{Y}$ and $g : \mathcal{Y} \to \mathcal{Z}$ such that the composed predictive model $g \circ f : \mathcal{X} \to \mathcal{Z}$ has good performance.

We focus on the intelligent selection of d classes to form the attribute space $\mathcal{Y}$, such that d is small to minimize labeling efforts, and the prediction in the unseen class space $\mathcal{Z}$ is optimized. Since we focus on the effects brought by class split, we fix the other components: the mapping f:$\mathcal{X} \mapsto \mathcal{Y}$ consists of multiple logistic regression models, each of which predicts one and only one seen class; a class similarity matrix K is derived from a related corpus as the knowledge base. Given two index sets $\mathcal{I}$ and $\mathcal{J}$, let $K^{\mathcal{I}\mathcal{J}}$ be the sub-matrix of K that consists of the rows indexed by $\mathcal{I}$ and columns indexed by $\mathcal{J}$. Then $K^{\mathcal{U}\mathcal{U}}$ is the similarity matrix for the unseen classes, and $K^{\mathcal{U}\mathcal{S}}_{ij}$ being the similarity between the i-th unseen class and the j-th seen class. One can view K as the adjacent matrix of the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of all classes and the edge weights are the class similarities. Given the above notations, with $\hat{\mathbf{y}} \in \mathbb{R}^d$ being the predicted seen classes, the mapping g is defined by $g : \hat{\mathbf{y}} \mapsto K^{\mathcal{U}\mathcal{S}}\hat{\mathbf{y}}$.

Methodology

Intuitively, we want to select an unseen class that can convery more information about the remaining unseen ones, and make it a seen class. The connectivity between the i-th unseen class and the remaining ones can be a measurement of the amount of information conveyed. Such connectivity can be calculated by various centrality metrics of the i-th node on the sub-graph of $\mathcal{G}$ consisting of all the curent unseen classes. The degree centrality of the i-th unseen class is one such measures and can be calculated as $\sum_{j=1}^{k} K^{\mathcal{U}\mathcal{U}}_{ij}$, where k is the current number of unseen classes. We then select the unseen class with the largest degree centrality as the next seen class (this selection strategy is called "*max-deg-uu*"). One possible drawback is that, the i-th class can be connected to only a small number of unseen classes with high weights $K^{\mathcal{U}\mathcal{U}}_{ij}$, but not connected to others. Selecting this class will not add much information about the unseen classes that are not well connected to this class.

We take a probabilistic perspective and use entropy to characterize the distribution of the significant connections between a class and the others. First, the similarities between each unseen class and the other unseen ones are normalized to a probability distribution:

$$P^{\mathcal{U}\mathcal{U}} = \text{diag}(\mathbb{1}^\top K^{\mathcal{U}\mathcal{U}})^{-1} K^{\mathcal{U}\mathcal{U}}, \tag{1}$$

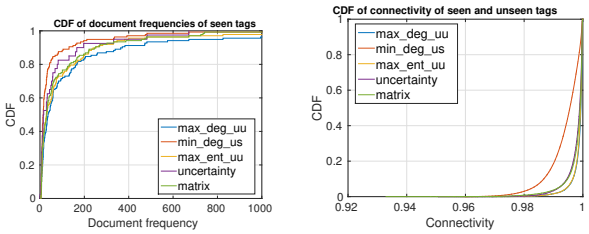
where $\text{diag}(\mathbf{v})$ denotes the diagonal matrix with diagonal elements being the entries of $\mathbf{v} \in \mathbb{R}^u$, and $\mathbb{1}$ is the all-one vector. Then we calculate the entropy of the connections between the i-th unseen class and others, $i = 1, \ldots, k$:

$$H(i) = -\sum_{j=1}^{u} P^{\mathcal{U}\mathcal{U}}_{ij} \log P^{\mathcal{U}\mathcal{U}}_{ij}. \tag{2}$$

We select the class that have the highest entropy, acquire labeled data for that class, and move it from $\mathcal{U}$ to $\mathcal{S}$. The labeled instances for the selected classes are used to train a prediction model, which is added to the mapping f. This selection procedure is repeated until the labeling budget runs out. Lastly, we obtain the zero-shot model consisting of f and g. The algorithm is shown in Algorithm 1. Intuitively, the larger the $H(i)$, the more classes in the pool of unseen classes are connected to the i-th unseen class, which shall have the highest utility to be added to the pool of seen classes. An extreme case is that there is only one unseen class connected to the i-th unseen class (not counting self connection), then the i-th unseen class is not providing information to the remaining unseen classes and shall not be selected. Formally, we want to minimize $\mathsf{Pr}(j \notin C_j(s), \forall s \in \mathcal{S}))$, the probability that j-th unseen class that is not well connected to any seen class. Let $C_j(s)$ be the event that the j-th unseen class is well connected to the s-th seen class.

$$\mathsf{Pr}(j \notin C_j(s), \forall s \in \mathcal{S}) = \prod_{s \in \mathcal{S}} \mathsf{Pr}(j \notin C_j(s)), \tag{3}$$

by assuming that the entries in $P^{\mathcal{S}\mathcal{U}}_{sj}, s \in \mathcal{S}$ are independently distributed conditioned on j. With a higher $H(s)$, $P^{\mathcal{S}\mathcal{U}}_{sj}, j = 1, \ldots, u$ are more evenly distributed (over the unseen classes) and more of the probabilities $\mathsf{Pr}(j \notin C_j(s)), s \in \mathcal{S}$ will be small, leading to a smaller $\mathsf{Pr}(j \notin C_j(s), \forall s \in \mathcal{S}))$.



(a) CDF of document frequency of selected classes (b) CDF of the seen-unseen classes connectivities

Figure 1: Analysis of seen-unseen tag split resulting from the proposed selection method

Algorithm 1 Active Zero-shot Learning

Input: Unlabeled training data $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, class similarity matrix K, budget b.
Output: Zero-shot prediction model $g \circ f$.
Randomly select seeding classes to form $\mathcal{S}$, and the remaining classes go to $\mathcal{U} = \{1, \ldots, d+k\} \setminus \mathcal{S}$.
while $b > 0$ **do**
  Normalize the sub-matrix $K^{\mathcal{U}\mathcal{S}}$ by Eq. (1).
  Calculate the entropies using Eq. (2).
  Select the class (denoted by t) that have the largest entropies.
  Query the labels of the training instances for the class t.
  $\mathcal{S} = \mathcal{S} \cup \{t\}, \mathcal{U} = \mathcal{U} \setminus \{t\}, b = b - 1$.
end while

EXPERIMENTS

StackExchange is a QA (questions and answers) system where members can ask and answer questions. The predictions of zero-shot learning can be used to prompt users to associate their questions with tags that have not used by any user (unseen tags), and thus facilitate question organization and retrieval. We adopt 3 sites from StackExchange: askubuntu, dba and unix (see Table 1). Bag-of-words representation with TF-IDF transformation is used as the feature vectors of the questions. Each tag is treated as a class, and a question can have multiple tags, so the tasks can be formulated as multi-label classification problems. Only those tags that appear in at least 10 questions are kept. Tags provided by the users for

Table 1: Datasets

	askubuntu	dba	unix
# Training	55684	12070	23069
# Test	55883	12211	23025
# Tags	1003	345	775

the questions are used as ground truth. Each selection strategy is tested on 20 randomly picked seeding seen classes, and we report the averaged performances over 20 runs. Questions are randomly split into disjoint training and test sets. The training data is used to train classification models (Liblinear with default settings), each of which maps from features to a seen tag. Then we map the predicted seen classes on the test data to unseen classes via a knowledge base, which is an embedding of the tags in a low dimensional space via restricted Boltzmann machine trained on the text corpus of questions. Tag similarity is calculated through a kernel function: $K(\mathbf{t}_1, \mathbf{t}_2) = \exp(\|\mathbf{t}_1 - \mathbf{t}_2\|^2/\sigma^2)$ where $\mathbf{t}_1$ and $\mathbf{t}_2$ are the low dimensional representations of two tags, and $\sigma = 10$ throughout the experiments.

We adopt 3 common metrics (precision@5, NDCG@5 and micro-AUC) to evaluate the performance of tag retrieval. Since there is no previous study on active zero-shot learning, we compare *max-ent-uu* with the following baselines that capture different aspects of seen-unseen class splits.

- *max-deg-uu*: as mentioned in the methodology section, this method labels data for the classes that have the highest degree centrality.
- *min-deg-us*: we take the row sums of the matrix $K^{\mathcal{US}}$, which captures the total similarity between unseen tags and seen tags. The unseen class with smallest row sum is picked. The intuition is that unseen tags that are farthest away from the current seen classes can provide complementary information to the current seen ones.
- *uncertainty*: uncertainty-based sampling queries the labels of the top unseen classes that has the highest entropy in their predictions on the training data, according to the current class split. This baseline runs in an incremental manner as *max-ent-uu* and *min-deg-us*.
- *matrix*: in [4] the author proposed a matrix partition algorithm to split a set of instances into two such that the mutual information between the distributions of the two sets is maximized. This method is considered to be a representativeness-based active learning method. We adapt their model and treat classes as instances. This algorithm runs in batch-mode and we only report its performance when 100 additional classes are selected.

We set the number of unseen classes selected in each iteration to 2 ($c = 2$) in Algorithm 1 and the other iterative baselines. We test other values for this parameter ($c = 5$ and $c = 10$), and find out that $c = 2$ gives the best results.

Results In Figure 2, we show how the zero-shot prediction performances of 4 iterative algorithms evolve as more seen classes are added, plus the batch-mode method *matrix*. Each row in Figure 2 consists of 3 sub-figures showing the performance in precision@5, ndcg@5 and micro AUC, respectively. In each sub-figure, the performance of *max-ent-uu* (shown in green solid lines), is compared with those of the 4 baselines. From the figures, we can see that across all datasets and all metrics, except that in Figures 2(c), *max-ent-uu* consistently outperforms all the baselines. In some cases, *max-ent-uu* ends up with performance two times better than the runner-up (Figures 2(g) and 2(b)). The baselines *min-deg-us, uncertainty* and *matrix* consistently have performance between those of *max-ent-uu* and *max-deg-uu* in 8 out of 9 cases.

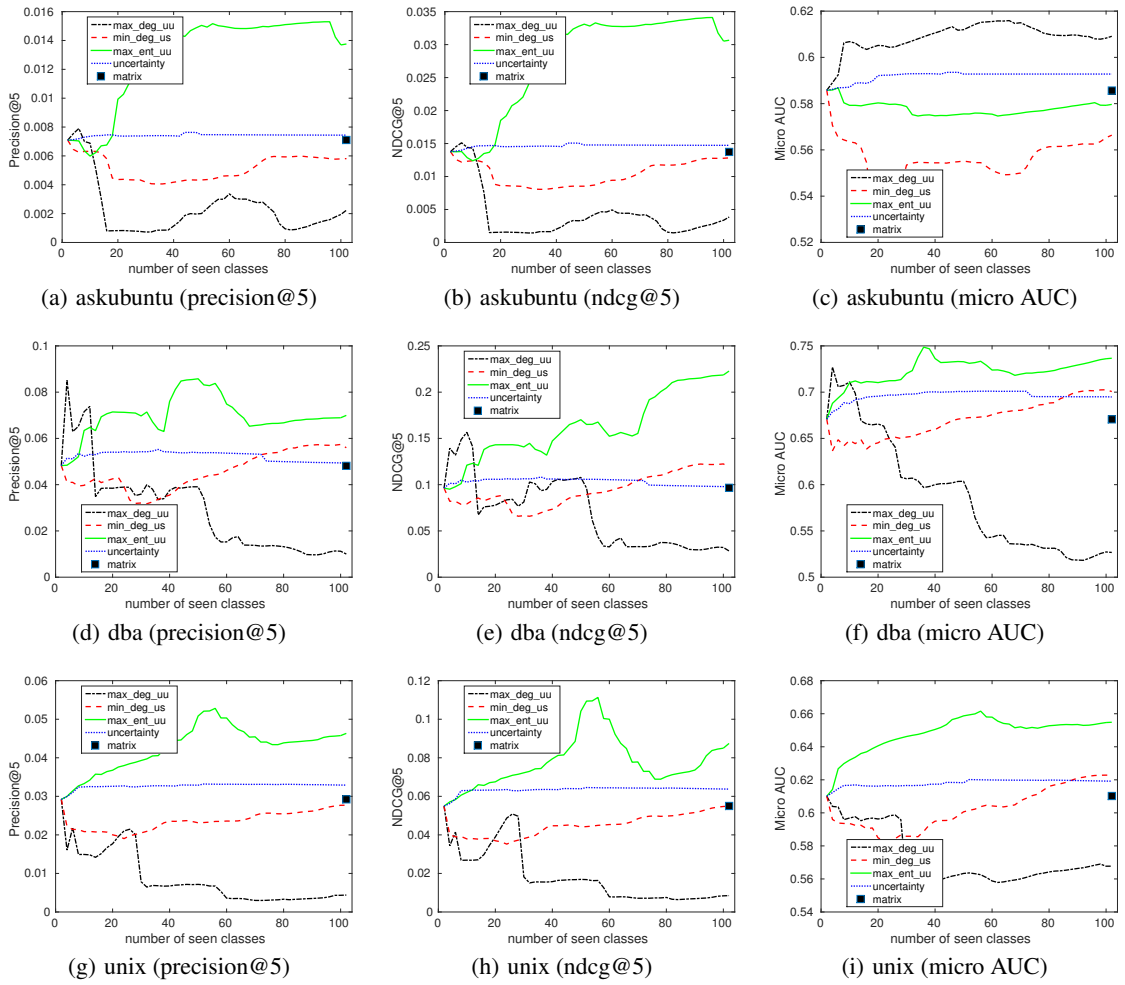
Surprisingly, the seemingly naive method *min-deg-us* can gradually pick up its performance and ends up with similar or better performance with the more sophisticated methods *matrix* in the dba and unix datasets, although its performance at the beginning is not very impressive. Our explanation is that by selecting the classes that are least similar to the already picked ones, more information can be revealed. However, this baseline fails to consider unseen class coverage information, and the selected classes may not be well-connected to the large clusters of unseen classes (as we will see next), leading to less effective seen-to-unseen class mapping. Furthermore, the performance of *matrix* is quite close to *uncertainty* in all cases. Our conjecture is that by picking the current unseen classes that do not have confident predictions, *uncertainty* is able to explore the class space that has not been explored before, and ends up with a seen class space that represent the whole class space quite well, which is what *matrix* aims for.

Analyzing the seen-unseen classes relations Here we empirically shows why the proposed method works via the analysis of the resulting seen-unseen connection matrix $K^{\mathcal{US}}$. In Figure 1(a), we plot the CDFs of the frequencies of the selected classes that appear in the training instances (namely document frequencies) on one dataset (best viewed in color). We can see that among the 5 strategies, the *max-deg-uu* tends to select classes that have higher document frequencies than those selected by *max-ent-uu*, as the CDF of *max-deg-uu* is more shifted to the right. It has been shown in text classification that, the more frequent a word appears in the corpus, the less informative it is, as evidenced by the commonly used tf-idf transformation. A frequent seen class is likely to be predicted more often by predictive models that takes into account of the class prior distribution. Such seen class become a less discriminative feature when used as features for the mapping g. This partly explains why the baseline *max-deg-uu* has the worst performance in 10 out of 12 cases among all methods.

From Figure 1(a), we see that the baseline *min-deg-us* also tends to select classes that are less frequent than those selected by *max-ent-uu*, then why hasn't *min-deg-us* outperformed *max-ent-uu*? In Figure 1(b), we plot the CDF of the seen-unseen class connectivities on the same dataset, where connectivities are the row sums of $K^{\mathcal{SU}}$. We see that *max-ent-uu* produces connectivities as strong as those produced by *max-deg-uu*. The seen classes selected by *min-deg-us* tend to have low connectivities with unseen classes, and the baselines *uncertainty* and *matrix* produces medium connectivities. If the connectivities are strong, then the seen classes can provide significant information about the unseen classes. We have similar observations on the other datasets, and we conclude that *max-ent-uu* is more likely to find seen classes that simultaneously possess the following properties: 1) discriminative about the test data (low document frequency) and 2) informative about the unseen classes (high coverage). The unique combination of these two properties helps *max-ent-uu* outperform the baselines.

CONCLUSIONS AND FUTURE WORK

We study active learning in the zero-shot prediction setting for the purpose of finding a small number of informative seen classes to facilitate unseen class predictions. We propose an entropy-based selection method, which is demonstrated to be able to capture the desirable distribution of seen-unseen similarity. Experiments show that the proposed method outperform both representativeness and uncertainty based active learning methods. In the future, we plan to



(a) askubuntu (precision@5) (b) askubuntu (ndcg@5) (c) askubuntu (micro AUC)

(d) dba (precision@5) (e) dba (ndcg@5) (f) dba (micro AUC)

(g) unix (precision@5) (h) unix (ndcg@5) (i) unix (micro AUC)

Figure 2: Comparisons of the proposed method and the baselines

explore more selection strategies based on other zero-shot learning properties. Further reducing the labeling efforts by instance selecting is also a promising direction.

Acknowledgment

This work is supported in part by NSF Award III-1526499, and NVIDIA Corporation with the donation of the Titan X GPU.

REFERENCES

[1] T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *J. Artif. Int. Res.*, 1995.

[2] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. DeViSE: A Deep Visual-Semantic Embedding Model. In *NIPS*. 2013.

[3] Y. Fu, T. Hospedales, T. Xiang, Z. Fu, and S. Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *ECCV*. 2014.

[4] Y. Guo. Active instance sampling via matrix partition. In *NIPS*. 2010.

[5] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[6] X. Li and Y. Guo. Max-margin zero-shot learning for multi-class classification. In *AISTAT*, 2015.

[7] X. Li, S. Liao, W. Lan, X. Du, and G. Yang. Zero-shot Image Tagging by Hierarchical Semantic Embedding. In *SIGIR*, 2015.

[8] T. Mensink, E. Gavves, and C. G. M. Snoek. COSTA: Co-Occurrence Statistics for Zero-Shot Classification. In *CVPR*, 2014.

[9] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean. Zero-Shot Learning by Convex Combination of Semantic Embeddings. *CoRR*, 2013.

[10] A. Pentina and C. H. Lampert. Active task selection for multi-task learning. 2016.

[11] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang. Two-dimensional active learning for image classification. In *CVPR*, 2008.

[12] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, 2011.

[13] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*. 2013.