LEARNING MULTIPLE CONCEPTS WITH INCREMENTAL DIVERSE DENSITY

James Gibson and Shrikanth Narayanan

Signal Analysis and Interpretation Lab
University of Southern California
Los Angeles, CA 90089

jjgibson@usc.edu, shri@sipi.usc.edu

ABSTRACT

We present a novel method of learning multiple disjunct concepts with diverse density using an incremental approach. We demonstrate that by maximizing the diverse density over individual target concept points and minimizing the probability of their intersection, concepts can be learned incrementally. This method reduces the complexity of the algorithm from factorial, with respect to the number of targets, to exponential order. We demonstrate that this greedy approach successfully learns disjunctive target concepts with competitive classification accuracy on a benchmark multiple instance learning dataset in comparison to other common diverse density approaches. We also introduce a novel application of the multiple instance learning framework to an emotion recognition task using prosodic and spectral speech features.

1. INTRODUCTION

Multiple instance learning (MIL) has become popular for learning problems using corpora with summatively, and often ambiguously, labeled data. In the MIL framework, datasets consist of labeled prototypes, called bags, which are comprised of several example feature vectors rather than every feature vector having an explicit label as in conventional supervised learning problems. The example feature vectors, called instances, may vary in the extent to which they contribute to the label given to the bag. Correctly identifying representative examples in a bag should ultimately lead to higher classification accuracy and may lend itself to interpretive insights about the problem domain based on these representative instances. The regions in feature space near representative instances are considered concepts and the goal of many MIL algorithms is to identify these concepts and distinguish bags of different labels by the proximity of their instances to these concepts.

The notion of diverse density (DD) was introduced by Maron and Lozano-Pérez in [1] to address this problem. The main intuition behind diverse density is that target concepts will lie in areas where many bags of the same label intersect or are very close. One benefit of this model is that points in feature space that are close to points from negatively labeled bags are given lower diverse density. This allows bag labels to be taken into account when searching for target concepts which leads to learned concepts with greater discriminability. The diverse density is one of the most popular formulations of MIL. It has been the subject of numerous advances including an expectation maximization formulation [2]. Additionally, it has been formulated as an instance selection task that is subsequently paired with a support vector machine classifier [3]. Chen et al. further improved this work by embedding the instance selection task in the optimization of the support vector machine [4].

In the original algorithm, the diverse density is calculated over the instance feature space and a threshold is determined to classify unseen bags according to where their instances lie with respect to the learned threshold [5]. Fould and Frank presented an algorithm for speeding up the diverse density learning process by choosing the target concept to be the single instance with highest diverse density from all the bags [6]. Maron (and Ratan) described how the diverse density algorithm could be extended to a search for multiple disjunctive target concepts in [5] and [7]. Maron also discussed considerations that must be made when searching for learning multiple concepts. One important consideration is how to choose d, the number of disjuncts to be learned. It was indicated that adding more disjuncts will always increase the diverse density but will lead to less generalization of the model when classifying unseen bags. How to address this issue remains an open problem. Because of the factorial complexity of learning multiple disjunct target concepts, experiments have not been conducted to explore learning more than two concepts.

This paper introduces an efficient method for learning multiple disjunct concepts with a variant of the diverse density algorithm that reduces the complexity of maximizing the diverse density over multiple disjuncts from factorial with the number of concepts to exponential order. We introduce a new estimator for the diverse density, called the Incremental Diverse Density (eq. 7), which allows for instances to be selected in a greedy manner. We empirically demonstrate the usefulness of the method with competitive classification accuracy on a canonical MIL dataset. Subsequently, we demonstrate a novel application of multiple instance learning to an emotion recognition task using speech features.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. METHODOLOGY

2.1. Diverse Density Learning of Multiple Concepts

The diverse density algorithm was originally presented for learning single point concepts in ambiguously labeled data. Subsequently, it was extended to learn multiple concepts by treating the single points as a disjunctitve set. The diverse density of a set of disjunct points in the observed instance space is defined as:

$$DD(\mathcal{D}) \equiv P(\mathcal{D}|\mathbf{B}), \tag{1}$$

where, $\mathbf{B} = \{B_1^+, ..., B_l^+, B_1^-, ..., B_m^-\}$ is the set of n labeled bags ($n = l + m$) and $\mathcal{D} = \{c_1 \vee c_2 \vee \cdots \vee c_d\}$ is a disjunction of d concepts. Then by invoking Bayes Rule the set of disjunct concepts with maximum diverse density can be determined with maximum likelihood estimation,

$$\mathcal{D}^* = \underset{\mathcal{D} \in \mathbf{I}^{(d)}}{\operatorname{argmax}}[P(\mathcal{D}|\mathbf{B})] = \underset{\mathcal{D} \in \mathbf{I}^{(d)}}{\operatorname{argmax}} \left[\frac{P(\mathbf{B}|\mathcal{D})P(\mathcal{D})}{P(\mathbf{B})} \right], \tag{2}$$

where $\mathbf{I}^{(d)}$ is the set of all d-element subsets of disjunct instances in $\mathbf{I}$, the set of the L observed instances from the l positive bags in $\mathbf{B}$. Then assuming a uniform prior, $P(\mathcal{D})$, conditional independence of observed bags given the target concepts, and invoking Bayes Rule once more, this becomes,

$$\begin{aligned} \mathcal{D}^* =& \underset{\mathcal{D} \in \mathbf{I}^{(d)}}{\operatorname{argmax}} \left[\prod_{i=1}^{l} P(B_i^+|\mathcal{D}) \prod_{i'=1}^{m} P(B_{i'}^-|\mathcal{D}) \right] \\ =& \underset{\mathcal{D} \in \mathbf{I}^{(d)}}{\operatorname{argmax}} \left[\prod_{i=1}^{l} P(\mathcal{D}|B_i^+) \prod_{i'=1}^{m} P(\mathcal{D}|B_{i'}^-) \right]. \end{aligned} \tag{3}$$

In [5], Maron compares common density estimators (e.g., most-likely-cause and noisy-or) to estimate $P(\mathcal{D}|B_i)$. For this work we use the most-likely-cause estimator because it is the most consistent with the assumption that bags can be represented by their single most representative instance. Thus, the conditional probability becomes,

$$P(\mathcal{D}|B_i) \propto 1 - \left| \frac{1 + y_i}{2} - \max_{1 \le j \le N_i} [P(B_{ij} \in \mathcal{D})] \right|, \tag{4}$$

where, $y_i \in \{-1, 1\}$ is the label of bag i, N_i is the number of instances in bag i, and $P(B_{ij} \in \mathcal{D})$ is the probability that the jth instance from the ith bag is in the set of hypothesized concepts $\mathcal{D}$. Note this value is not a proper probability because it is left unnormalized (hence, the $\propto$ relation). By taking the hypothesized concept which is closest to the instance B_{ij}, it is assumed that each instance is generated by one concept. Subsequently, the probability that B_{ij} is in any one of the target concepts is estimated by,

$$P(B_{ij} \in \mathcal{D}) \propto \max_{1 \le k \le d} \left(e^{-||B_{ij} - c_k||^2} \right). \tag{5}$$

We use the max operator as an estimator for the logical 'or' to be consistent with the previous literature.

2.2. Incremental Learning of Multiple Concepts

In order to maximize the diverse density over d disjunct instances with an exhaustive search it is necessary to search $\binom{L}{d}$ possible combinations of instances in $\mathbf{I}^{(d)}$. By making the assumption that each disjunct concept can be learned incrementally (in a greedy manner), target concepts can be learned one at a time reducing the search to $L(2^d - 1)$ iterations. The intuition behind this assumption is that when finding the maximum diverse density of the disjunction of the globally optimum point (the instance with highest diverse density) at the first step with all other points will yield a disjunct concept of high diverse density. We begin the incremental maximization over single instances from positive bags. Clearly this reduces exactly to the single concept (point-wise) diverse density learning when $\mathcal{D}$ consists of a single target concept. Once a single concept is discovered, it is fixed and paired with all other $(L-1)$ instances and two disjunct concepts are learned. This is then repeated until d total concepts are learned.

The existing algorithm for maximizing the diverse density over disjunct concepts is not sufficient for incremental learning because the strongest hypothesized point is often in an area of much higher diverse density and it will overwhelm other areas in the space that would have been considered in a global optimization compared to an incremental one. Ultimately, attempting to incrementally learn concepts in this way will lead to redundant concepts. Since it is of interest to learn a diverse set of strong concepts, it is necessary to minimize the similarity (or probability of intersection) of the hypothesized concepts. Because of the estimators used to approximate the diverse density (see equations 4 and 5), the general addition theorem does not hold. However by making the approximation,

$$DD(c_1 \vee c_2) \approx DD(c_1) + DD(c_2) - DD(c_1 \wedge c_2), \tag{6}$$

we have the desired result of maximizing the diverse density over the concepts while minimizing their intersection. We use this approximation to define a new quantity for incremental learning of the diverse density of multiple concepts which we will refer to as the *Incremental Diverse Density* (IDD). More explicitly, we define the incremental diverse density of d disjunct concepts as,

$$\begin{aligned} IDD(\mathcal{D}) \equiv& \sum_{i=1}^{d} DD(c_i) - \sum_{i,j:1 \le i < j \le d} DD(c_i \wedge c_j) \\ &+ \sum_{i,j,k:1 \le i < j < k \le d} DD(c_i \wedge c_j \wedge c_k) - \cdots \\ &+ (-1)^{d-1} DD(c_1 \wedge \cdots \wedge c_d). \end{aligned} \tag{7}$$

Or, equivalently (and more compactly),

$$IDD(\mathcal{D}) \equiv \sum_{\forall \mathcal{S} \in 2^{\mathcal{C}}} (-1)^{|\mathcal{S}|-1} DD(\mathcal{S}) \tag{8}$$

where $\mathcal{C}$ is the conjunction of hypothesized concepts, $\mathcal{C} = \{c_1 \wedge c_2 \wedge \cdots \wedge c_d\}$.

While the diverse density over disjunct instances has been previously defined, there is no existing definition of the diverse density of conjunctive instances. Thus, we define the probability that a given instance is in all the target concepts as:

$$P(B_{ij} \in \mathcal{C}) \propto \min_{1 \leq k \leq d} \left(e^{-||B_{ij}-c_k||^2} \right), \tag{9}$$

This definition uses the min function as an estimator for a logical 'and' to be consistent with the estimator for logical 'or' defined in equation 5. The total number of disjuncts d, is a parameter that can be tuned with cross-validation or set by the user with prior knowledge of the number of target concepts to be learned in a particular dataset.

2.3. Nearest Concept Features

In order to evaluate the efficacy of IDD for accurately classify bags a simple nearest concept classification rule is used. First, concepts are learned from positive bags. Once the positive concepts are learned, each bag is classified according to the minimum distance of any instance in that bag to the concepts, i.e., bags with an instance very close to a positive concept are labeled positive. This results in a single dimensional feature that is used to train a simple linear SVM classifier [8] to determine the decision threshold. Thus, each bag is represented by the feature:

$$\phi(\mathbf{B_i}) = \min_{1 \leq k \leq d} \left(\min_{1 \leq j \leq N_i} ||B_{ij} - c_k||^2 \right). \tag{10}$$

This can be easily extended to a nearest concept task in which the distances from a particular concept are taken from all points in the bag and then bags would be classified according to either a majority voting or minimum total distance rule or the multi-dimensional feature could be used to train a subsequent classifier.

2.4. Point-and-scaling Concepts

Thus far, we have only discussed concepts in terms of their points in feature space. One of the main strengths of the diverse density is that a concept point and the corresponding feature scaling can be learned simultaneously to learn concepts that emphasize important features and de-emphasize less important features. This adds the ability to learn better concepts but at the cost of increased complexity. Point-and-scaling concepts are learned by maximizing the diverse density with respect to both the concept point c and corresponding scaling s. Thus the probability that a particular instance B_{ij} resulted from a point-and-scaling concept is estimated by,

$$P(B_{ij} \in \{c, s\}) = \exp\left(-\sum_{q=1}^{r} s_q (B_{ijq} - c_q)^2 \right), \tag{11}$$

where r is the dimensionality of the feature space. This maximization is done using gradient based optimization. For implementing this task we use the `dpfmin` routine [9]. In particular we use an adaptation from code written by Chen et al. [3]. For incremental learning there are two possible approaches. The first is to maximize the diverse density over single point-and-scaling instances from positive bags and to continue using those scaling parameters when choosing subsequent disjunct concepts. Another solution is to re-learn the scaling parameters of each concept point as additional disjuncts are learned. In [6] it is demonstrated that scaling once, iteratively, or after concepts points were determined did not have a significant impact on classification accuracy in the single concept case. In this paper, we learn feature scalings for the single target points and use those same scalings when maximizing the incremental diverse density with respect to disjunct concepts. This scaling is subsequently used to compute the nearest concept features. Thus, the bag features become:

$$\phi(\mathbf{B_i}) = \min_{1 \leq k \leq d} \left(\min_{1 \leq j \leq N_i} \left[\sum_{q=1}^{r} s_{k_q} (B_{ijq} - c_{k_q})^2 \right] \right). \tag{12}$$

3. EXPERIMENTAL RESULTS

3.1. MUSK Experiments

The MUSK 1 and MUSK 2 data sets are benchmark tests used for evaluating MIL algorithms. Both corpora are available online in the University of California, Irvine Machine Learning Repository [10]. The data are comprised of molecules which have been labeled as either 'musk' or 'non-musk' based on whether they smell musky. Each molecule has many feature vectors to describe different shapes that molecules can take. For this reason, molecules are considered bags and the feature vectors describing them are considered instances. A molecule is considered a 'musk' if any of its feature vector descriptions are a musk and a 'non-musk' only if all of its descriptions are not musks. Hence, we attempt to learn the musk concept(s) in the data to perform classification.

Table 1. Classification accuracies (%) on the MUSK data sets. Average of 10 runs of 10-fold cross validation results are shown (with 95% confidence intervals in brackets for IDD).

ALGORITHM	MUSK 1	MUSK 2
IDD	88.04 [86.88, 89.20]	86.37 [84.95, 87.79]
DD [1]	88.9	82.5
EM-DD [11]	84.8	84.9
QUICKDD [6]	86.4	87.2
DD-SVM [3]	85.8	91.3
MILES [4]	86.3	87.7

We report classification accuracies[1] for the MUSK data in Table 1. To choose the number of target concepts for IDD, we perform 2-fold cross validation on the training set, varying the number of disjuncts from 1 to 5. The average accuracy is on par with previously reported results on the MUSK 1 data. In fact, IDD gave the highest average accuracy for diverse density or multiple instance variants on the MUSK 1

[1] We report results for the QuickDD *Scaling Only* method because it is the most directly comparable to the scaling method used here [6]. The results for the diverse density (DD) were not conducted for 10 runs of 10-fold cross validation so they are reported here for reference not direct comparison [1].

Table 2. Single versus multiple concept IDD unweighted and weighted classification accuracy, precision, recall, and F1-measure (%) for the sadness detection task.

ALGORITHM	d	U.W. ACCURACY	W. ACCURACY	PRECISION	RECALL	F1-MEASURE
IDD	1	82.31	81.79	**78.69**	78.69	78.69
IDD	1-5	**83.67**	**84.72**	77.94	**86.89**	**82.17**

dataset (with respect to 10 runs of 10 fold cross validation). The lower bound of the 95% confidence interval for IDD is above the average accuracy of all other methods.

The IDD algorithm was also competitive on the MUSK 2 data. Only the average accuracy for DD-SVM was above the higher bound of the 95% confidence interval of IDD. It is important to note that IDD, in contrast to DD-SVM, only estimates the diverse density for the positive class, there is no need to find the optimum parameters for a radial basis function kernel SVM, and final classification is performed with a single dimensional feature vector.

3.2. Sadness Detection in Spoken Dialogs

Label ambiguity is a major challenge in human emotional and affective state recognition [12]. Moreover, such labels are often provided at a summative level such as over an entire spoken dialog. This makes the multiple instance learning paradigm an attractive approach as it was developed in order to address label ambiguity. The IEMOCAP database consists of conversational dialogs in which pairs of actors express a number of target emotions [13], [14]. There are ten speakers in the corpus who participated in both scripted and improvisational dialogs. There were multiple sessions recorded for each pair (scripted and improvisational each approximately five minutes long). In each session they were asked to act emotions from the target set of happiness, anger, sadness, frustration and neutral. In total there are 147 sessions (bags) consisting of 4947 utterance turns (instances) which were each subsequently given an emotion label.

We adapt this emotion recognition task to the multiple instance learning framework by choosing an emotion that is both well represented in the data and seemed that it would be well modeled by a salient instance representation. We chose sadness for this task as it is represented in a large proportion of the sessions (42.11%) and it can be emoted in ambiguous ways (e.g., apathy versus grief). We label any session with at least one sad utterance (labeled as sad) as a 'sad' session, reserving the label 'not sad' for sessions without any sad utterances. This is consistent with the original MIL formulation in which it only takes one positive instance to label a bag as positive, while negative bags only contain negative instances. This results in a challenging classification task because many of the sessions contain a variety of different emotions (e.g., some sessions contain both sad and happy utterances) that are non prototypically expressed or represent blended emotions.

We model these sessions by extracting speech prosodic and spectral features from each utterance. Pitch, intensity and 12 Mel Filter Bank (MFB) coefficients were extracted over 25 ms frames with a 10 ms shift. These signal level feature dimensions were z-normalized with the mean and standard deviation of the training data by fold. Eleven functionals of these frame-level acoustic features were taken across each speaker utterance to produce instance feature representations. The computed functionals were mean, variance, median, inter-quartile range, 1st percentile, 25th percentile, 75th percentile, 99th percentile, range, skewness, and kurtosis. This resulted in a 154 dimensional feature vector representation of each instance (14 features by 11 functionals). The classification experiments were conducted using leave-one-speaker-out cross validation to ensure speaker independent modeling. With ten speakers this yielded a 10-fold cross validation. As Maron mentioned in [5] there is an inherent tradeoff between the number of concepts estimated and generalization of the model. For this task we use 2-fold cross-validation on the training data to determine the optimum number of concepts.

As shown in table 2, using the IDD algorithm to estimate multiple concepts improved the unweighted and weighted accuracy, recall, and F1-measure for the sadness detection task. Only precision decreased by estimating multiple concepts. Recall had the greatest increase indicating that multiple concepts helped retrieve sessions with sad instances that were missed by a single concept. However, this came at the cost of increasing false positives. This result is intuitive because by allowing more regions to be considered salient to the classification task, it is more likely that a similar instance may appear in a negative bag.

4. CONCLUSION AND FUTURE WORK

This paper presented a new variant of the diverse density for incrementally estimating multiple disjunct concepts in multiple instance data. This method allows for more complex concepts to be learned efficiently. More complex concepts lead to better modeling of some multiple instance data as verified by competitive accuracy on benchmark data. In the future, we will investigate the problem of estimating the ideal number of concepts to be learned during training. In this work, we did this estimation by performing cross validation on the training data. Unfortunately, this is a very computationally costly procedure. We plan to investigate more efficient ways of doing this estimation such as using clustering or information measures. In addition to improving this work, we are interested in applying the multiple instance framework to more classification and concept learning tasks such as utterance level analysis of conversational dialogs.

5. REFERENCES

[1] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in neural information processing systems*, 1998, pp. 570–576.

[2] Q. Zhang and S. Goldman, "EM-DD: An improved multiple-instance learning technique," in *Advances in neural information processing systems*, vol. 2, 2002, pp. 1073–1080.

[3] Y. Chen and J. Wang, "Image categorization by learning and reasoning with regions," *The Journal of Machine Learning Research*, vol. 5, pp. 913–939, 2004.

[4] Y. Chen, J. Bi, and J. Wang, "Miles: Multiple-instance learning via embedded instance selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1931–1947, 2006.

[5] O. Maron, "Learning from ambiguity," Ph.D. dissertation, Massachusetts Institute of Technology, 1998.

[6] J. Foulds and E. Frank, "Speeding up and boosting diverse density learning," *Discovery Science*, pp. 102–116, 2010.

[7] O. Maron and A. Ratan, "Multiple-instance learning for natural scene classification," in *Proceedings of the Fifteenth International Conference on Machine Learning*, vol. 15, 1998, pp. 341–349.

[8] C. Chang and C. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[9] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical recipes in C: the art of scientific programming*. Cambridge University Press, New York, 1992.

[10] K. Bache and M. Lichman, "UCI machine learning repository," University of California, Irvine, School of Information and Computer Sciences, 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[11] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in neural information processing systems*, vol. 15, 2002, pp. 561–568.

[12] E. Mower, A. Metallinou, C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *Affective Computing and Intelligent Interaction and Workshops*, 2009.

[13] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[14] ——, "The interactive emotional dyadic motion capture (iemocap) database," [Online]. Available: http://sail.usc.edu/iemocap/iemocap_release.htm, 2008.