

Logistic Regression and Boosting for Labeled Bags of Instances

Xin Xu and Eibe Frank

Department of Computer Science
University of Waikato
Hamilton, New Zealand
{xx5, eibe}@cs.waikato.ac.nz

Abstract. In this paper we upgrade linear logistic regression and boosting to multi-instance data, where each example consists of a labeled bag of instances. This is done by connecting predictions for individual instances to a bag-level probability estimate by simple averaging and maximizing the likelihood at the bag level—in other words, by assuming that all instances contribute equally and independently to a bag’s label. We present empirical results for artificial data generated according to the underlying generative model that we assume, and also show that the two algorithms produce competitive results on the Musk benchmark datasets.

1 Introduction

Multi-instance (MI) learning differs from standard supervised learning in that each example is not just a single instance: examples are collections of instances, called “bags”, containing an arbitrary number of instances. However, as in standard supervised learning there is only one label for each example (i.e. bag). This makes MI learning inherently more difficult because it is not clear how the bag label relates to the individual instances in a bag. The standard way to approach the problem is to assume that there is one “key” instance in a bag that triggers whether the bag’s class label will be positive or negative.¹ This approach has been proposed for predicting the activity of a molecule, where the instances correspond to different shapes of the molecule, and the assumption is that the presence of a particular shape determines whether a molecule is chemically active or not [1]. The task is then to identify these key instances. A different approach is to assume that all instances contribute equally and independently to a bag’s class label. In the above context this corresponds to assuming that the different shapes of a molecule have a certain probability of being active and the average of these probabilities determines the probability that the molecule will be active. In this paper, we use the latter approach to upgrade logistic regression and boosting to MI problems.

The paper is organized as follows. In Section 2 we explain the underlying generative model that we assume and illustrate it using artificial MI data based on

¹ Throughout this paper we assume a classification task with two possible classes.

this model. In Section 3 we describe how we use the generative model to upgrade linear logistic regression and boosting to MI learning. In Section 4 we evaluate the resulting algorithms on both the artificial data and the Musk benchmark problems [1]. Section 5 summarizes our results.

2 A Generative Model

We assume that the class label of a bag is generated in a two-stage process. The first stage determines class probabilities for the individual instances in a bag, and the second stage combines these probability estimates in some fashion to assign a class label to the bag. The difficulty lies in the fact that we cannot observe class labels for individual instances and therefore cannot estimate the instance-level class probability function directly.

This general two-stage framework has previously been used in the Diverse Density (DD) algorithm [2] to learn a classifier for MI problems. In the first step, DD assumes a radial (or “Gaussian-like”) formulation for the instance-level class probability function $Pr(y|x)$ (where y is a class label—either 0 or 1—and x an instance). In the second step, it assumes that the values of $Pr(y|x)$ for the instances in a bag are combined using either a multi-stage (based on the noisy-or model) or a one-stage (based on the most-likely-cause model) Bernoulli process to determine the bag’s class label. In both cases DD essentially attempts to identify ellipsoid areas of the instance space for which the probability of observing a positive instance is high, i.e, it attempts to identify areas that have a high probability of containing “key” instances. The parameters involved are estimated by maximizing the bag-level likelihood.

In this paper we use the same two-stage framework to upgrade linear logistic regression and boosting to MI data. However, we assume a different process for combining the instance-level class probability estimates, namely that all instances contribute equally to a bag’s class probability. Consider an instance-level model that estimates $Pr(y|x)$ or the log-odds function $\log \frac{Pr(y=1|x)}{Pr(y=0|x)}$, and a bag b that corresponds to a certain area in the instance space. Given this bag b with n instances $x_i \in b$, we assume that the bag-level class probability is either given by

$$Pr(Y|b) = \frac{1}{n} \sum_{i=1}^n Pr(y|x_i) \tag{1}$$

or by

$$\begin{aligned} \log \frac{Pr(y = 1|b)}{Pr(y = 0|b)} &= \frac{1}{n} \sum_{i=1}^n \log \frac{Pr(y = 1|x_i)}{Pr(y = 0|x_i)} \\ &\Rightarrow \begin{cases} Pr(y = 1|b) = \frac{[\prod_i^n Pr(y=1|x_i)]^{1/n}}{[\prod_i^n Pr(y=1|x_i)]^{1/n} + [\prod_i^n Pr(y=0|x_i)]^{1/n}} \\ Pr(y = 0|b) = \frac{[\prod_i^n Pr(y=0|x_i)]^{1/n}}{[\prod_i^n Pr(y=1|x_i)]^{1/n} + [\prod_i^n Pr(y=0|x_i)]^{1/n}} \end{cases} \end{aligned} \tag{2}$$

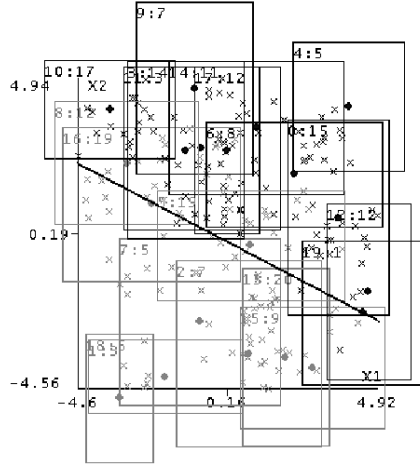


Fig. 1. An artificial dataset with 20 bags.

Equation 1 is based on the arithmetic mean of the instance-level class probability estimates while Equation 2 involves the geometric mean.

The above generative models are best illustrated based on an artificial dataset. Consider an artificial domain with two attributes. We create bags of instances by defining rectangular regions in this two-dimensional instance space and sampling instances from within each region. First, we generate coordinates for the centroids of the rectangles according to a uniform distribution with a range of $[-5, 5]$ for each of the two dimensions. The size of a rectangle in each dimension is chosen from 2 to 6 with equal probability. Each rectangle is used to create a bag of instances. For sampling the instances, we assume a symmetric triangle distribution with ranges $[-5, 5]$ in each dimension (and density function $f(x) = 0.2 - 0.04|x|$). From this distribution we sample n instances from within a rectangle.

The number of instances n for each bag is chosen from 1 to 20 with equal probability. We define the instance-level class probability by the linear logistic model $Pr(y = 1|x_1, x_2) = \frac{1}{1 + e^{-x_1 - 2x_2}}$, where x_1 and x_2 are the two attribute values of an instance. Thus the log-odds function is a simple linear function, which means the instance-level decision boundary is a line. Then we take Equation 2 to calculate $Pr(y|b)$. Finally we label each bag by flipping a coin according to this probability.

Figure 1 shows a dataset with 20 bags that was generated according to this model. The black line in the middle is the instance-level decision boundary (i.e. where $Pr(y = 1|x_1, x_2) = 0.5$) and the sub-space on the right side has instances with higher probability to be positive. A rectangle indicates the region used to sample points for the corresponding bag (and a dot indicates its centroid). The top-left corner of each rectangle shows the bag index, followed by the number of instances in the bag. Bags in gray belong to class “negative” and bags in black to

class “positive”. Note that bags can be on the “wrong” side of the instance-level decision boundary because each bag was labeled by flipping a coin based on its class probability.

3 Upgrading Linear Logistic Regression and Boosting

In this section we first show how to implement linear logistic regression based on the above generative model. Then we do the same for additive logistic regression (i.e. boosting)—more specifically, AdaBoost.M1 [3].

3.1 Linear Logistic Regression

The standard logistic regression model does not apply to multi-instance data because the instances’ class labels are masked by the “collective” class label of a bag. However, suppose we know how the instance-level class probabilities are combined to form a bag-level probability, say, by Equation 2. Then we can apply a standard optimization method (e.g. gradient descent) to search for a maximum of the (bag-level) binomial likelihood based on the training data. This gives us an indirect estimate of the instance-level logistic model.

The instance-level class probabilities are given by $Pr(y = 1|x) = 1/(1 + \exp(-\beta x))$ and $Pr(y = 0|x) = 1/(1 + \exp(\beta x))$ respectively, where β is the parameter vector to be estimated. If we assume bag-level probabilities are given by Equation 2, this results in the bag-level model

$$\begin{aligned} Pr(y = 1|b) &= \frac{[\prod_i^n Pr(y=1|x_i)]^{1/n}}{[\prod_i^n Pr(y=1|x_i)]^{1/n} + [\prod_i^n Pr(y=0|x_i)]^{1/n}} = \frac{\exp(\frac{1}{n}\beta \sum_i \mathbf{x}_i)}{1 + \exp(\frac{1}{n}\beta \sum_i \mathbf{x}_i)} \\ Pr(y = 0|b) &= \frac{[\prod_i^n Pr(y=0|x_i)]^{1/n}}{[\prod_i^n Pr(y=1|x_i)]^{1/n} + [\prod_i^n Pr(y=0|x_i)]^{1/n}} = \frac{1}{1 + \exp(\frac{1}{n}\beta \sum_i \mathbf{x}_i)} \end{aligned} \quad (3)$$

Based on this we can estimate the parameter vector β by maximizing the bag-level binomial log-likelihood function:

$$LL = \sum_i^N [y_i \log Pr(y = 1|b) + (1 - y_i) \log Pr(y = 0|b)] \quad (4)$$

where N is the number of bags.

Note that this problem can actually be reduced to standard single-instance logistic regression by converting each bag of instances into a single instance representing the bag’s mean (because the instances only enter Equation 3 in the sum $\sum_i x_i$). The reason for this is the simple linear structure of the model and the fact that the bag-level probabilities are generated based on the geometric mean (Equation 2). In practice, it is usually not possible to say whether the geometric mean is more appropriate than the arithmetic one, so we may want to use Equation 1 as an alternative. In that case, we change the formulation of $Pr(y|b)$ based on Equation 1 and use this in conjunction with the same log-likelihood function (Equation 4). This problem can no longer be solved by

1. Initialize weight of each bag to $W_i = 1/N$, $i = 1, 2, \dots, N$.
2. Repeat for $m = 1, 2, \dots, M$:
 - (a) Set $W_{ij} \leftarrow W_i/n_i$, assign the bag’s class label to each of its instances, and build an instance-level model $h_m(x_{ij}) \in \{-1, 1\}$.
 - (b) Within the i^{th} bag (with n_i instances), compute the error rate $e_i \in [0, 1]$ by counting the number of misclassified instances within that bag, i.e. $e_i = \sum_j 1_{(h_m(x_{ij}) \neq y_i)}/n_i$.
 - (c) If $e_i < 0.5$ for all i ’s, go to Step 3.
 - (d) Compute $c_m = \operatorname{argmin} \sum_i W_i \exp[(2e_i - 1)c_m]$ using numeric optimization.
 - (e) If $(c_m \leq 0)$, go to Step 3.
 - (f) Set $W_i \leftarrow W_i \exp[(2e_i - 1)c_m]$ and renormalize so that $\sum_i W_i = 1$.
3. return $\operatorname{sign}[\sum_j \sum_m c_m h_m(x_j)]$.

Fig. 2. The MIBoosting Algorithm.

applying the above transformation in conjunction with standard single-instance logistic regression. In the remainder of this paper, we will call the former method MILogisticRegressionGEOM and the latter one MILogisticRegressionARITH.

As usual, the maximization of the log-likelihood function is carried out via numeric optimization because there is no direct analytical solution. The optimization problem can be solved very efficiently because we are working with a linear model. (Note that the radial formulation in the DD [2] algorithm, for example, poses a difficult global optimization problem that is expensive to solve.) In our implementation we use a quasi-Newton optimization procedure with BFGS updates suggested in [4], searching for parameter values around zero.

3.2 Boosting

Both linear logistic regression and the model used by the DD algorithm assume a limited family of patterns. Boosting is a popular algorithm for modeling more complex relationships. It constructs an ensemble of so-called “weak” classifiers. In the following we explain how to upgrade the AdaBoost.M1 algorithm to MI problems, assuming the same generative model as in the case of linear logistic regression. Note that AdaBoost.M1 can be directly applied to MI problems (without any modifications) if the “weak” learner is a full-blown MI learner, in the same way as ensemble classifiers for MI learning have been built using bagging in [5]. In the following we consider the case where the weak learner is a standard single-instance learner (e.g. C4.5 [6]).

Boosting, more specifically, AdaBoost.M1, originates from work in computational learning theory [3], but received a statistical explanation as additive logistic regression in [7]. It can be shown that it minimizes an exponential loss function in a forward stage-wise manner, ultimately estimating the log-odds function $\frac{1}{2} \log \frac{\Pr(y=1|X)}{\Pr(y=0|X)}$ based on an additive model [7]. To upgrade AdaBoost.M1 we use the generative model described in Section 2 in conjunction with Equation 2 (i.e. the geometric average). The pseudo code for the algorithm, called “MIBoosting” is shown in Figure 2. Here, N is the number of bags, and we

use the subscript i to denote the i^{th} bag, where $i = 1, 2, \dots, N$. There are n_i instances in the i^{th} bag, and we use the subscript j to refer to the j^{th} instance, where $j = 1, 2, \dots, n_i$. The j^{th} instance in the i^{th} bag is x_{ij} . Note that, for convenience, we assume that the class label of a bag is either 1 or -1 (i.e. $y \in \{1, -1\}$), rather than 1 or 0.

The derivation of the algorithm is analogous to that in [7]. We regard the expectation sign E as the sample average instead of the population expectation. We are looking for a function (i.e. classifier) $F(b)$ that minimizes the exponential loss $E_B E_{Y|B}[\exp(-yF(b))]$. In each iteration of boosting, the aim is to expand $F(b)$ into $F(b) + cf(b)$ (i.e. adding a new “weak” classifier) so that the exponential loss is minimized. In the following we will just write $E[\cdot]$ to denote $E_B E_{Y|B}[\cdot]$, and use $E_w[\cdot]$ to denote the weighted expectation, as in [7].

In each iteration of the algorithm, we search for the best $f(b)$ to add to the model. Second order expansion of $\exp(-ycf(b))$ about $f(b) = 0$ shows that we can achieve this by searching for the maximum of $E_w[yf(b)]$, given bag-level weights $W_B = \exp(-yF(b))$ [7]. If we had an MI-capable weak learner at hand that could deal with bag weights, we could estimate $f(b)$ directly. However we are interested in wrapping our boosting algorithm around a single-instance weak learner. Thus we expand $f(b)$ into $f(b) = \sum_n h(x_j)/n$, where $h(x_j) \in \{-1, 1\}$ is the prediction of the weak classifier $h(\cdot)$ for the j^{th} instance in b . We are seeking a weak classifier $h(\cdot)$ that maximizes

$$E_w[yh(x_b)/n] = \sum_{i=1}^N \sum_{j=1}^{n_i} \left[\frac{1}{n_i} W_i y_i h(x_{ij}) \right].$$

This is maximized if $h(x_{ij}) = y_i$, which means that we are seeking a classifier $h(\cdot)$ that attempts to correctly predict the label of the bag that each instance pertains to. More precisely, $h(\cdot)$ should minimize the weighted instance-level classification error given instance weights $\frac{W_i}{n_i}$ (assuming we give every instance its bag’s label). This is exactly what standard single-instance learners are designed to do (provided they can deal with instance weights). Hence we can use any (weak) single-instance learner to generate the function $h(\cdot)$ by assigning each instance the label of its bag and the corresponding weight $\frac{W_i}{n_i}$. This constitutes Step 2a of the algorithm in Figure 2.

Now that we have found $f(b)$, we have to look for the best multiplier $c > 0$. To do this we can directly optimize the objective function

$$\begin{aligned} E_B E_{Y|B}[\exp(-yF(b) + c(-yf(b)))] &= \sum_i W_i \exp\left[c_m \left(-\frac{y \sum_j h(x_{ij})}{n_i}\right)\right] \\ &= \sum_i W_i \exp[(2e_i - 1)c_m], \end{aligned}$$

where $e_i = \sum_j 1_{(h_m(x_{ij}) \neq y_i)} / n_i$ (computed in Step 2b).

Minimization of this expectation constitutes Step 2d. Note that this function will not have a global minimum if all $e_i < 0.5$. In that case all bags will be

correctly classified by $f(b)$, and no further boosting iterations can be performed. Therefore this is checked in Step 2c. This is analogous to what happens in standard AdaBoost.M1 if the current weak learner has zero error on the training data.

Note that the solution of the optimization problem involving c may not have an analytical form. Therefore we simply search for c using a quasi-Newton method in Step 2d. The computational cost for this is negligible compared to the time it takes to learn the weak classifier. The resulting value for c is not necessarily positive. If it is negative we can simply reverse the prediction of the weak classifier and get a positive c . However, we use the AdaBoost.M1 convention and stop the learning process if c becomes negative. This is checked in Step 2e.

Finally, we update the bag-level weights in Step 2f according to the additive structure of $F(b)$, in the same way as in standard AdaBoost.M1. Note that, the more misclassified instances a bag has, the greater its weight in the next iteration. This is analogous to what happens in standard AdaBoost.M1 at the instance level.

To classify a test bag, we can simply regard $F(b)$ as the bag-level log-odds function and take Equation 2 to make a prediction (Step 3). An appealing property of this algorithm is that, if there is only one instance per bag, i.e. for single-instance data, the algorithm naturally degenerates to normal AdaBoost.M1. To see why, note that the solution for c in Step 2d will be exactly $\frac{1}{2} \log \frac{1 - err_W}{err_W}$, where err_W is the weighted error. Hence the weight update will be the same as in standard AdaBoost.M1.

4 Experimental Results

In this section we first discuss empirical results for the artificial data from Section 2. Then we evaluate our techniques on the Musk benchmark datasets.

4.1 Results for the Artificial Data

Because we know the artificial data from Section 2 is generated using a linear logistic model based on the geometric mean formulation in Equation 2, MILogisticRegressionGEOM is the natural candidate to test on this specific data. Figure 3 shows the parameters estimated by this method when the number of bags increases. As expected, the estimated parameters converge to the true parameters asymptotically. However, more than 1000 bags are necessary to produce accurate estimates.

To evaluate the performance of MIBoosting we created an independent test set with 10,000 bags based on the same generative model. As the “weak” learner we used decision stumps, performing 30 iterations of the boosting algorithm. The test error for different numbers of training bags is plotted in Figure 4, and compared to that of MI linear logistic regression. Not surprisingly, the latter outperforms boosting because it matches the underlying generative model exactly. However, both methods approach the optimum error rate on the test data.

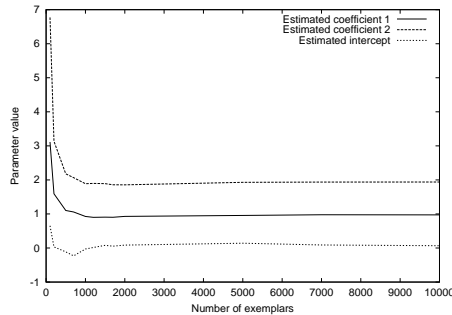


Fig. 3. Parameters estimated by MI Logistic Regression on the artificial data.

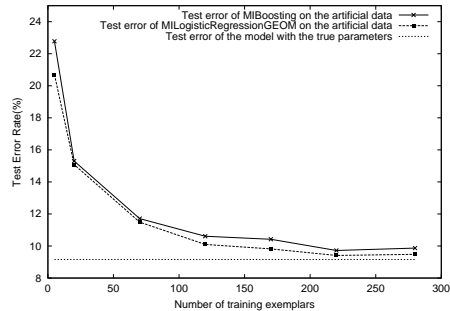


Fig. 4. Test error rates for boosting and logistic regression on the artificial data.

4.2 Results for the Musk Problems

In this section we present results for the Musk drug activity datasets [1]. The Musk 1 data has 92 bags and 476 instances, the Musk 2 data 102 bags and 6598 instances. Both are two-class problems with 166 attributes. Because of the large number of attributes, some regularization proved necessary to achieve good results with logistic regression. To this end we added an L_2 penalty term $\lambda\|\beta\|^2$ to the likelihood function in Equation 4, where λ is the ridge parameter.² In our experiments we set $\lambda = 2$. For boosting, we used C4.5 trees as the “weak” classifiers and 50 boosting iterations. We introduced some regularization by setting the minimum number of instances per leaf to twice the average bag size (10 instances for Musk 1, and 120 instances for Musk 2). The post-pruning mechanism in C4.5 was turned off.

The upper part of Table 1 shows error estimates for the two variants of MLogisticRegression and for MIBoosting. These were obtained by 10 runs of stratified 10-fold cross-validation (CV) (at the bag level). The standard deviation of the 10 estimates from the 10 runs is also shown. The lower part summarizes some results for closely related statistical methods that can be found in the literature. A summary of results for other multi-instance learning algorithms can be found in [8]. Note also that the performance of MI learning algorithms can be improved further using bagging, as shown in [5].

The first result in the lower part of Table 1 is for the Diverse Density algorithm applied in conjunction with the noisy-or model [2]. This result was also obtained by 10-fold cross-validation. The second result was produced by a neural network. The MI Neural Network is a multi-layer perceptron adapted to multi-instance problems by using a soft-max function to hone in on the “key” instances in a bag [9].³ The last entry in Table 1 refers to a support vector machine (SVM) used in conjunction with a Gaussian multi-instance kernel [8]. This method re-

² We standardized the training data using the weighted mean and variance, each instance being weighted by the inverse of the number of instances in its bag.

³ It is unclear how the error estimates for the neural network were generated.

Method	Musk 1	Musk 2
MILogisticRegressionGEOM	14.13±2.23	17.74±1.17
MILogisticRegressionARITH	13.26±1.83	15.88±1.29
MIBoosting with 50 iterations	12.07±1.95	15.98±1.31
Diverse Density	11.1	17.5
MI Neural Network	12.0	18.0
SVM with Gaussian MI kernel	13.6±1.1	12.0±1.0

Table 1. Error rate estimates for the Musk datasets and standard deviations (if available).

places the dot product between two instances (used in standard SVMs) by the sum of the dot products over all possible pairs of instances from two bags. This method resembles ours in the sense that it also implicitly assumes all instances in a bag are equally relevant to the classification of the bag. The error-estimates for the SVM were generated using leave-10-out, which produces similar results as 10-fold cross-validation on these datasets.

The results for logistic regression indicate that the arithmetic mean (Equation 1) is more appropriate than the geometric one (Equation 2) on the Musk data. Furthermore, boosting improves only slightly on logistic regression. For both methods the estimated error rates are slightly higher than for DD and the neural network on Musk 1, and slightly lower on Musk 2. Compared to the SVM, the results are similar for Musk 1, but the SVM appears to have an edge on the Musk 2 data. However, given that logistic regression finds a simple linear model, it is quite surprising that it is so competitive.

Figure 5 shows the effect of different numbers of iterations in the boosting algorithm. When applying boosting to single-instance data, it is often observed that the classification error on the test data continues to drop after the error on the training data has reached zero. Figure 5 shows that this effect also occurs in the multi-instance version of boosting—in this case applied to the Musk 1 data and using decision stumps as the weak classifiers.

Figure 5 also plots the Root Relative Squared Error (RRSE) of the probability estimates generated by boosting. It shows that the classification error on the training data is reduced to zero after about 100 iterations but the RRSE keeps decreasing until around 800 iterations. The generalization error, estimated using 10 runs of 10-fold CV, also reaches a minimum around 800 iterations, at 10.44% (standard deviation: 2.62%). Note that this error rate is lower than the one for boosted decision trees (Table 1). However, boosting decision stumps is too slow for the Musk 2 data, where even 8000 iterations were not sufficient to minimize the RRSE.

5 Conclusions

We have introduced multi-instance versions of logistic regression and boosting, and shown that they produce results comparable to other statistical multi-

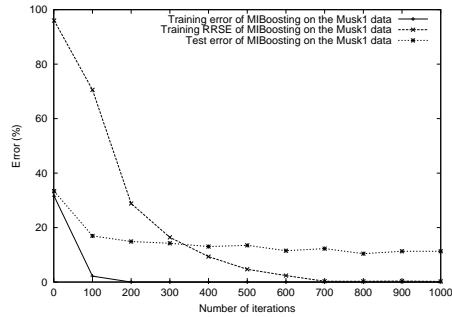


Fig. 5. The effect of varying the number of boosting iterations on the Musk1 data.

instance techniques on the Musk benchmark datasets. Our multi-instance algorithms were derived by assuming that every instance in a bag contributes equally to the bag’s class label—a departure from the standard multi-instance assumption that relates the likelihood of a particular class label to the presence (or absence) of a certain “key” instance in the bag.

Acknowledgments

This research was supported by Marsden Grant 01-UOW-019.

References

1. T.G. Dietterich, R.H. Lathrop, and T. Lozano-Pérez. Solving the multiple-instance problem with the axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.
2. O. Maron. *Learning from Ambiguity*. PhD thesis, MIT, United States, 1998.
3. Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *Proc of the 13th Int Conf on Machine Learning*, pages 148–156. Morgan Kaufman, 1996.
4. P.E. Gill, W. Murray, and M.H. Wright. *Practical Optimization*. Academic Press, 1981.
5. Z.-H. Zhou and M.-L. Zhang. Ensembles of multi-instance learners. In *Proc of the 14th European Conf on Machine Learning*, pages 492–501. Springer, 2003.
6. J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
7. J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression, a statistical view of boosting (with discussion). *Annals of Statistics*, 28:307–337, 2000.
8. T. Gärtner, P.A. Flach, A. Kowalczyk, and A.J. Smola. Multi-instance kernels. In *Proc of the 19th Int Conf on Machine Learning*, pages 179–186. Morgan Kaufmann, 2002.
9. J. Ramon and L. De Raedt. Multi instance neural networks. In *Attribute-Value and Relational Learning: Crossing the Boundaries*, 2000. Workshop at the 17th Int Conf on Machine Learning.