Robust Fuzzy Clustering for Multiple Instance Regression

Mohamed Trabelsi, Hichem Frigui*

Multimedia Research Lab, CECS dept.
University of Louisville, KY, USA

Abstract

Multiple instance regression (MIR) operates on a collection of bags, where each bag contains many instances sharing the same real-valued label. Only few instances, called primary instances, contribute to the bag labels. The remaining ones are noisy observations. The goal in MIR is to identify the primary instances within each bag and learn a regression model that can predict the label of a previously unseen bag. In this paper, we show that regression models can be identified as clusters when appropriate features and distances are used. We introduce an algorithm, called Robust Fuzzy Clustering for Multiple Instance Regression (RFC-MIR), that can learn multiple linear models simultaneously. First, RFC-MIR uses constrained fuzzy memberships to obtain an initial partition where instances can belong to multiple models with various degrees. Then, it uses unconstrained possibilistic memberships to allow the initial local models to expand and converge to the global model. These memberships are also used to identify the primary instances within each bag. After clustering, the possibilistic memberships are used to identify the optimal number of regression models. We evaluate our approach on synthetic data sets generated by varying the dimensionality of the feature space, the number of instances per bag, and the noise level. We also validate the RFC-MIR using two real applications: prediction of the yearly average yield of a crop using remote sensing data; and drug activity prediction. These applications have been used consistently to validate

*Corresponding author

existing MIR algorithms. We show that our approach achieves higher accuracy than existing methods.

Keywords: Multiple instance regression, Fuzzy clustering, Possibilistic clustering, Multiple model regression.

---

1. Introduction

In standard supervised learning, each object is represented by a single feature vector and a label. This label is categorical for classification problems and real-valued for regression problems. However, some learning applications cannot provide a label to each observation, and thus could not be solved with this traditional learning paradigm. An alternative framework of learning that tackles the inherent labeling ambiguity better than supervised learning is the multiple instance learning (MIL) paradigm [1, 2, 3]. In MIL, an object is represented by a collection of feature vectors, or instances, called a bag. Each bag can contain a different number of instances. Labels are available at the bag level, however, labels of individual instances within a bag are unknown. This many-to-one relationship between instances and data labels produces an inherent ambiguity in determining which instances in a given bag are responsible for its associated label. MIL was formalized in 1997 by Dietterich et al. providing a solution to drug activity prediction [1]. Ever since, MIL has increasingly been applied to a wide variety of tasks including drug discovery [4], image analysis [5, 6, 7, 8], content-based information retrieval [9], time series prediction [2], landmine detection [10], information fusion [11, 12], and remote sensing [13].

Most of the existing work in MIL has focused on multiple instance classification (MIC). In MIC, a bag is labeled negative if all of its instances are negative, and positive if at least one of its instances is positive. Given a training set of labeled bags, the goal of MIC is to learn a concept that predicts the labels of training data at the instance level and generalizes to predict the labels of testing bags and their instances [1]. In addition to the above approach that is based on the standard MIL assumption, multiple MIL paradigms have been proposed

[14].

Multiple instance regression (MIR) has received much less attention. In MIR, bags have real-valued labels and the goal is to learn a regression model that can predict the label of a new bag from the features of its instances. MIR is a challenging learning task since we have no prior knowledge of the primary instances, i.e., instances within each bag that are relevant to its label. In fact, for the general MIR setting, the unknown number of relevant instances can vary from one bag to another. Predicting the label of a new test bag using a learned MIR model is even more challenging.

In this paper, we introduce a novel MIR framework, called Robust Fuzzy Clustering for MIR (RFC-MIR). In RFC-MIR, we show that regression models can be identified as clusters when appropriate features and distances are used. We also show that fuzzy memberships are useful in obtaining an initial partition where instances can belong to multiple models with various degrees, and possibilistic memberships can be used to identify non-primary instances as noise and outliers and reduce their influence on the learned regression parameters.

2. Related work

Most existing work in MIL has focused on multiple instance classification (MIC). MIC algorithms can be categorized into three main paradigms: instance space, bag space, and embedded instance space. Instance space-based algorithms rely on the standard multiple instance assumption, which states that a positive bag must contain at least one positive instance [1]; the labels of remaining instances are irrelevant. These algorithms seek points in the instance feature space with strong correlation to instances from positive bags and no or low correlation to instances from negative bags. These points, called target concepts (TC), serve as loci for instance-level class labeling. Examples of instance space-based algorithms include the Axis-Parallel Rectangles (APR) [1] which constructs a set of boundaries in the problem feature space to capture the TC. Other instance space-based approaches include the Diverse Density (DD) [4]

and EM-DD [15], which use optimization techniques to learn the TC. In [16], clustering methods were used to generalize the DD algorithm to learn multiple target concepts simultaneously.

In the bag space MIC, each bag is mapped to an N-dimensional feature vector based on a bag-to-bag comparator metric with respect to all N bags within the training data. A key advantage of the bag space paradigm is that the mapped bag representation removes the instance-level ambiguity from the problem. Examples of such methods include the Citation-kNN classifier [17] and Multiple Instance Dissimilarity (MInD) [18].

Embedded instance space methods also map each bag to a single feature vector. The difference is that target concepts in the instance space are used for this mapping, rather than bags. Examples include methods based on learning a dictionary [19, 20] and other methods based on learning target concepts such as DD-SVM [21] and MILES [6]. DD-SVM locates candidate TCs across multiple runs of the DD algorithm with distinct starting points. MILES [6] considers each instance from both positive and negative bags as a potential TC and uses a sparse SVM to select an optimal subset of instances. MIRSVM [8] is another algorithm that extends the SVM classifier to multiple instance data. Other classifiers that have been extended to handle multiple instance data include kNN [17] and Neural Networks [22].

In contrast to MIC, in MIR there is no notion of positive/negative bags and target concepts. MIR aims to learn a regression model that maps each bag to a real-valued output.

The two simplest approaches to MIR, that are commonly used as baselines (e.g., in [23, 24, 13]), are the Aggregated-MIR and Instance-MIR. The Aggregated-MIR represents each bag by a single meta-instance, typically the mean of all the bag's instances. Then, a model is learned by applying traditional regression techniques on the meta-instances. Instance-MIR propagates the bag label to all of its instances and then uses all instances and traditional regression techniques to learn the model.

Primary instance regression (PIR) [25] is one of the earliest MIR that main-

tains the bag structure. PIR assumes that the label of each bag is determined by a single instance, called primary instance (i.e "true instance"), and that the rest of the instances in the bag are noisy observations. PIR is an iterative algorithm that uses an EM-based approach to alternate between selecting the most likely primary instances and fitting a linear regression to these instances.

EM-MIR [13] is another multiple instance regression algorithm that assumes that each bag contains a prime instance which determines the bag label. EM-MIR treats the bag label as a random variable described with a mixture model. The contribution of each instance to its bag label is proportional to its probability of being the prime instance. The EM algorithm is then used to learn the prior function and the prediction function parameters.

MI-ClusterRegress [23] is a different approach to MIR that uses clustering to reduce MIR to a standard regression problem. It is motivated by the fact that bags can contain instances drawn from a number of distinct underlying data distributions. MI-ClusterRegress uses a clustering step to group instances of all bags into a predefined number of clusters. Instances that are relevant to each cluster, called exemplars, are identified and used to build a local model for each cluster using traditional regression techniques. The cluster with the best fitting error is identified as the "prime" cluster and is responsible for the bags' labels. The potential drawback of MI-ClusterRegress is that clustering is performed in an unsupervised manner, without considering the bag labels. Moreover, it assumes that all primary instances will be grouped into one cluster, which is usually not the case especially in high dimensional feature spaces. In fact, if the primary instances of all bags are split among multiple clusters, then the cluster with the best fitting error may not necessarily correspond to the prime cluster (as will be illustrated in Section 3.1). For instance, a small cluster that has few primary instances will have a better chance at being selected (lower error of fit) as the prime cluster than a larger one that has the bulk of primary instances. Since both clusters may include other non-primary instances, regression models learned from exemplars of both clusters may deviate from the correct model. The deviation of the small cluster can be more severe due to reduced number

of samples.

Even though many MIR algorithms have been proposed in the past few years, predicting the label of a new test bag using the learned MIR model remains very challenging and no existing methods have proved to accomplish this task without making restrictive assumptions. Early methods assume that a single instance within each bag determines its label. These methods can identify primary instances for labeled bags during learning, but cannot make prediction for a new unlabeled bag unless its primary instance is known *a priori*. Other methods such as Instance-MIR, PIR, and Aggregated-MIR assume all instances within a bag are relevant and are noisy versions of the primary instance. Instance-MIR and PIR first predict the label of each instance within the test bag using the learned regression model. Then, the labels of all instances are aggregated using the *mean* or *median* to predict the label of the bag. Aggregated-MIR first computes the test bag's meta-instance (e.g. mean of all its instances) and uses it as input to the learned model. The above two approaches are reliable only when all instances within each bag represent the "true instance" with small deviations.

To label a new test bag, MI-ClusterRegress constructs the bag's meta-instance as the average of the bag's instances weighted by their relevance to the prime cluster identified during learning. The predicted label of this exemplar is then treated as the bag's label. This labeling approach is based on the assumption that only the primary instances of the test bag will be assigned to the prime cluster, which is not necessarily true.

3. Robust clustering to learn multiple regression models

Let $D = \{B_j, j = 1 \dots N_B\}$ be a collection of N_B bags, where $B_j = \{(\mathbf{b}_{ij}, y_j), i = 1 \dots n_j\}$, $\mathbf{b}_{ij} \in \mathbb{R}^d$ is the attribute vector representing the i^{th}th instance from the jth bag, y_j is the real-valued target value of the j_{th} bag and n_j is the number of instances in the j^{th} bag. The instances b_{ij} that determine the label y_j, called primary instances, are unknown. The objective of MIR is to identify the primary instances within each bag, learn the regression model, and

be able to predict the label of previously unseen bags.

3.1. Motivating Example

First, we motivate our clustering-based approach by using a simple 1-D data to illustrate the MI-ClusterRegress algorithm and its shortcomings. The data include 100 bags and each bag has 5 instances and are displayed in figure 1(a) where the x-axis represents the 1-D feature of the instances and the y-axis represents the label of each bag (all instances in one bag have the same y value as they share the same label). All primary instances are displayed as green dots and the remaining instances are displayed as red dots. Recall that in MIR this information is not available, and that we use it here for illustrative purposes only. The first step in MI-ClusterRegress is to partition all instances into k clusters. We use the K-Means algorithm [26] for this example and we let $k = 4$. The resulting partition is shown in figure 1(b). Since the K-Means is applied to the instance features only, it simply partitions the x-axis into 4 intervals. The next step in MI-ClusterRegress is to identify the closest instance from each bag to each cluster center. These instances are referred to as exemplars. In figure 1(c), we show exemplars of each cluster. The basic assumption in MI-ClusterRegress is that the exemplars of one of the clusters will correspond to the primary instances of all bags. However, comparing the primary instances in figure 1(a) to the exemplars in figure 1(c), we notice that this is not the case. The next step in MI-ClusterRegress is to fit a linear regression model to the exemplars of each cluster (as shown in figure 1(d)) and identify the cluster that has the smallest error fit. For this example, cluster 4 was selected. However, as illustrated in figure 1(e), the learned regression model is quite different from the true model used to generate the data.

In the above example, MI-ClusterRegress failed to learn the correct regression model because the assumption that most exemplars of one of the clusters will correspond to the true primary instances did not hold. This assumption will be harder to maintain as the dimensionality of the feature space increases.

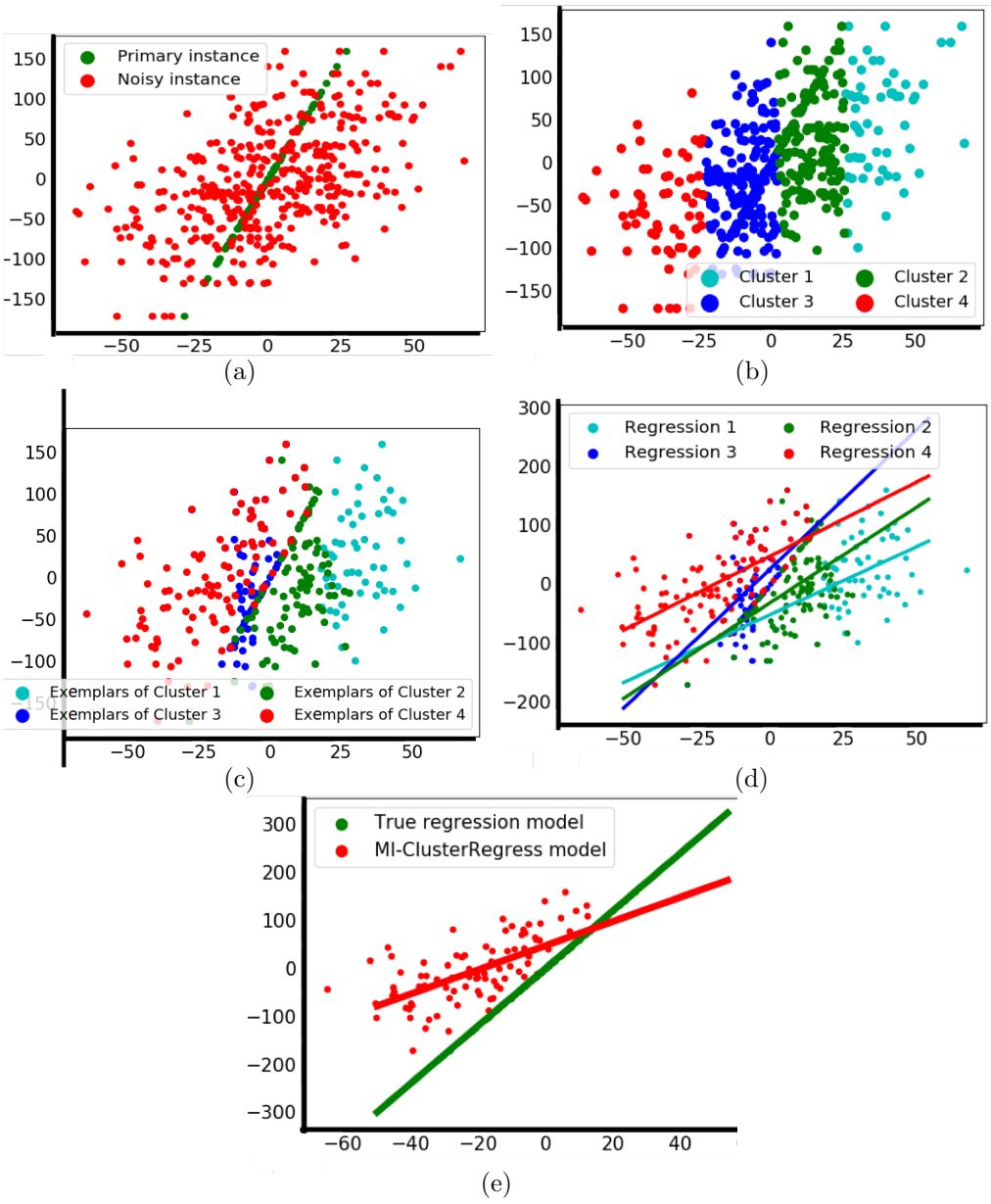

Figure 1: Illustration of the MI-ClusterRegress algorithm [23] to learn a regression model from multiple instance data. (a) Multiple instance data. Each bag has one primary instance (green dots) and 4 noisy instances (red dots). (b) The 4 clusters obtained after partitioning all instances. (c) Exemplars of the 4 clusters. (d) Regression models learned using exemplars of every cluster. (e) True regression model vs model learned using MI-ClusterRegress.

3.2. Robust clustering for MIR

MI-ClusterRegress offers an interesting approach to solve the MIR problem. It uses unsupervised learning (clustering) to separate the primary instances from the noisy ones. However, it relies on a simple clustering algorithm that seeks spherical clusters in the feature space, that are not necessarily relevant to the regression model, and cannot identify noise and outliers. In this paper, inspired by MI-ClusterRegress, we propose a new approach, called Robust Fuzzy Clustering for Multiple Instance Regression (RFC-MIR). RFC-MIR performs clustering and multiple model fitting simultaneously. Compared to MI-ClusterRegress, it has four additional properties. First, instead of using clustering to partition the instances in the feature space regardless of the labels of their bags, we combine the features and labels and use clustering, with an appropriate distance, to identify multiple local regression models. Second, we use a robust clustering approach so that non-primary instances (that incorrectly inherit the label of the bag they belong to) can be treated as noise and outliers to minimize their influence on the learned regression parameters. Third, we use fuzzy clustering so that each instance can contribute to each local regression with a fuzzy membership degree. Finally, we use properties of the possibilistic memberships to find the optimal number of regression models.

Let $\mathbf{x}_{ji} = [\mathbf{b}_{ji}, y_i] \in \mathbb{R}^{d+1}$ represents the concatenation of the j^{th} instance from the i^{th} bag and the label of its bag. Recall that labels are not available at the instance level and that y_i is valid only for the primary instances of bag i. Thus, many of the $\mathbf{x}_{ji}$'s can have an irrelevant y_i. We combine $\mathbf{x}_{ji}$ from all training bags into $D = \{\mathbf{x}_{ji}, i = 1 \dots N_B, j = 1 \dots n_i\}$. To simplify notation, we assume that all bags have the same number of instances $n_i = n$ for $i = 1 \dots N_B$, and we rewrite $D = \{\mathbf{x}_i, i = 1 \dots N\}$, where $N = n \times N_B$. Next, we show how clustering could be used to identify the primary instances from all of the N instances and learn the MIR models simultaneously.

The fuzzy c-means (FCM) [27] algorithm minimizes

$$J_F = \sum_{i=1}^{C} \sum_{j=1}^{N} (u_{ij}^F)^m dist_{ij}^2 \tag{1}$$

In (1), C is the number of clusters, $dist_{ij}$ is the distance from $\mathbf{x}_j$ to cluster i, $m > 1$ is a weighting exponent called the fuzzifier, and u_{ij}^F is the fuzzy membership of $\mathbf{x}_j$ in cluster i and satisfies the constraint:

$$u_{ij}^F \in [0,1] \text{ for all i,j; and } \sum_{i=1}^{C} u_{ij}^F = 1 \text{ for all j .} \tag{2}$$

The distance $dist_{ij}$ used in (1) controls the type and shape of clusters that will be identified. Various distances have been proposed to identify ellipsoidal, linear, and shell clusters such as lines, circles, ellipses, and general quadratics [28, 29]. In this paper, we assume that the underlying regression model is linear and we use (1) to identify multiple linear models. In particular, we use a generalization of the distance in [30, 31] and let:

$$dist_{ij}^2 = \sum_{k=1}^{d+1} v_{ik}((\mathbf{x}_j - \mathbf{c}_i) \cdot \mathbf{e}_{ik})^2 \tag{3}$$

where $\mathbf{c}_i$ is the center of cluster i, $\mathbf{e}_{ik}$ is the k^{th} unit eigenvector of the covariance matrix Σ_i of cluster i. The eigenvectors are assumed to be arranged in ascending order of the corresponding eigenvalues λ_{ik}. In (3), we let

$$v_{ik} = \frac{\left[\prod_{j=1}^{d+1} \lambda_{ij}\right]^{\frac{1}{d+1}}}{\lambda_{ik}}, \tag{4}$$

that is, more importance will be given to distances projected on the eigenvectors associated with the smaller eigenvalues.

Optimization of (1) with $dist_{ij}$ in (3) subject to (2), using alternate optimization, results in an iterative algorithm that alternates between updating the fuzzy memberships using

$$u_{ij}^F = \left[\sum_{k=1}^{C} \left(\frac{dist_{ij}^2}{dist_{kj}^2}\right)^{\frac{1}{m-1}}\right]^{-1} \tag{5}$$

and the center $\mathbf{c}_i$ and covariance Σ_i of cluster i using

$$\mathbf{c}_i = \frac{\sum_{j=1}^{N} (u_{ij}^F)^m \mathbf{x}_j}{\sum_{j=1}^{N} (u_{ij}^F)^m}, \tag{6}$$

and

$$\Sigma_i = \frac{\sum_{j=1}^{N} (u_{ij}^F)^m (\mathbf{x}_j - \mathbf{c}_i)(\mathbf{x_j} - \mathbf{c_i})^{\mathbf{T}}}{\sum_{j=1}^{N} (u_{ij}^F)^m}. \tag{7}$$

The objective function of the FCM in (1) is known to be sensitive to noise and outliers, and thus, is not suitable for the considered MIR application where we know a priori that the data is very noisy as non-primary instances and their labels should be treated as noise. Instead, we use the possibilistic c means (PCM) [27], which relaxes the constraint in (2) and minimizes

$$J_P = \sum_{i=1}^{C} \sum_{j=1}^{N} (u_{ij}^P)^m dist_{ij}^2 + \sum_{i=1}^{C} \eta_i \sum_{j=1}^{N} (1 - u_{ij}^P)^m \tag{8}$$

where $u_{ij}^P \in [0, 1]$ is a possibilistic membership degree that is not constrained to sum to 1 across all clusters. It is close to 0 for samples that are considered outliers, and close to 1 for inliers. In (8), η_i is a cluster resolution parameter that could be fixed a priori or estimated using the distribution of the data within each cluster [27].

Optimization of (8) also results in an iterative algorithm that alternates between updating u_{ij}^P using

$$u_{ij}^P = \frac{1}{1 + (\frac{dist_{ij}^2}{\eta_i})^{\frac{1}{m-1}}} \tag{9}$$

and the center $\mathbf{c}_i$ and covariance Σ_i as in (6) and (7) respectively.

Since the PCM does not constraint the memberships u_{ij}^P to sum to 1, it can result in several identical clusters. We use this feature to identify the optimal number of regression models [32]. We simply start with an over-specified number

of clusters, then identify and merge similar ones. Two clusters are considered similar and merged if

$$\frac{\sum_{k=1}^{N} |u_{ik}^P - u_{jk}^P|}{\sum_{k=1}^{N} |u_{ik}^P| + \sum_{k=1}^{N} |u_{jk}^P|} < \theta_M \tag{10}$$

where θ_M is a threshold constant.

Currently, we assume that the underlying regression model is linear and thus, it can be captured by a single linear cluster. Consequently, if the algorithm identifies more than one cluster, say $c' > 1$, we need to select the "optimal" one, p. Two possible criteria can be used to select this cluster. The first one is based on minimizing the fitting errors, i.e.,

$$p = arg \min_{i=1,\ldots c'} \left\{ \varepsilon_i = \sum_{j=1}^{N} (u_{ij}^P)^m dist_{ij}^2 \right\} \tag{11}$$

An alternative approach is to select the cluster that covers the maximum number of bags. Let

$$\mathcal{P}^i = \{\mathbf{x}_j, j = 1 \ldots N \mid u_{ij}^P > \theta_P\} \tag{12}$$

be the set of inliers (i.e primary instances) assigned to cluster i, and

$$\mathcal{B}^i = \{B_k \mid \mathbf{x}_j \in \mathcal{P}^i \text{ and } \mathbf{x}_j \text{ is an instance of } B_k\}$$

be the set of bags that contribute to cluster i. In (12), $\theta_P \in [0, 1]$ is a constant threshold. The "optimal" cluster, p, can be identified as the one that has the largest number of unique bags in $\mathcal{B}^i$. In this paper, we report results using the latter approach.

After identifying the optimal cluster, p, we let the primary instances of the data D be the primary instances of cluster p, i.e., $\mathcal{P} = \mathcal{P}^p$ as it will be described in Section 3.3.

The linear regression model parameters can be identified from the cluster center $\mathbf{c}_p$ and covariance matrix Σ_p. Let $\mathbf{e}_{min} = [e_{min}^1, \ldots, e_{min}^{d+1}]$ be the eigenvector associated with the smallest eigenvalue λ_{min} of Σ_p and let $\mathbf{x} = [x_1, \ldots, x_d, y] \in \mathcal{P}$ be a primary instance. The fact that $\mathbf{x}$ and $\mathbf{c}_p$ belong to the

regression model leads to

$$\mathbf{e}_{min} \cdot (\mathbf{x} - \mathbf{c}_p) = 0,$$

or

$$\mathbf{e}_{min} \cdot \mathbf{x} = \mathbf{e}_{min} \cdot \mathbf{c}_p.$$

Decomposing $\mathbf{x}$ into the instance feature vector $[x_1, \ldots, x_d]$ and its label y, we obtain

$$e_{min}^{d+1} y + \sum_{k=1}^{d} e_{min}^{k} x^k = \mathbf{e}_{min} \cdot \mathbf{c}_p$$

Solving for y, we obtain the regression model:

$$y = f(x) = \frac{\mathbf{e}_{min} \cdot \mathbf{c}_p}{e_{min}^{d+1}} - \sum_{k=1}^{d} \frac{e_{min}^{k}}{e_{min}^{d+1}} x^k \tag{13}$$

The resulting RFC-MIR algorithm is summarized in Algorithm 1. The objective of the first part (lines 5 – 11) is to partition the feature space and obtain an initial set of local models that approximate the global model. Accuracy is not needed at this level as long as all dense regions in the feature space are covered by some clusters. This is needed because possibilistic clustering can potentially ignore dense regions (treats them as outliers) if they are not represented by the initial clusters [33]. Different clustering algorithms could be used for this initialization step including those that are based on crisp sets (e.g. K-Means algorithm [26]). In fact, even a standard clustering algorithm, with the Euclidean distance, could be used to first partition the feature space into spherical clusters. Then, a local linear model could be estimated for each cluster. In this paper, we use the distance in (3) and fuzzy memberships to make the transition to the possibilistic part of the RFC-MIR (lines 13 – 17) smoother. This choice tends to reduce the total number of iterations needed for the possibilistic clustering to converge. The clusters' parameters are updated for I_{init} iterations, where the default value of I_{init} is set to 10.

The second component of RFC-MIR (lines 13 – 17) uses unconstrained memberships to allow local models to expand by considering neighboring points (even if they belong to different clusters). Consequently, if the primary instances can

be fit by one linear model, all initial local models will converge to the same global model. As a final step, possibilistic memberships are used to identify similar models (using (10)) and merge them. Crisp methods could not be used for this component since they assign each data point exclusively to the best model. This constraint prevents clusters from absorbing nearby points if they are slightly closer to other clusters. Thus, local models cannot expand and evolve to the global model. In fact, even fuzzy methods where memberships in all clusters are constrained to sum to one could not be used for this component.

3.3. Prediction Algorithm for RFC-MIR

Primary instances in the training data can be identified using (12) as the inliers, i.e., points that have high possibilistic membership. For testing, this process is not as trivial since labels are needed to assign new memberships. Thus, as all existing MIR methods, the proposed RFC-MIR needs to make assumptions to predict the label of a new test bag. For instance, we could use the simple approach used in Instance-MIR and average the predicted labels of all instances. Similarly, we could compute the test bag's meta-instance as in the Aggregated-MIR and predict its label. In this paper, we report results using an approach similar to the one used in MI-ClusterRegress that takes advantage of the data structure identified during training. We assume that the primary instance(s) of the test bag will be assigned to the prime cluster. Unlike MI-ClusterRegress, RFC-MIR uses both the instances' features and the bags' labels to learn the prime cluster. It also uses possibilistic memberships to identify and ignore the effect of noisy instances. This imposes additional constraints while learning the structure of the data and while identifying the primary instance of a test bag. Thus, the risk of violating the assumption is minimized.

Let $B^t = \{\mathbf{x}_1^t \dots \mathbf{x}_n^t\}$ be a test bag with n instances. First, for each $\mathbf{x}_i^t \in B^t$, we identify the closest primary instance (from training data) $\mathbf{x}_i^P \in \mathcal{P}$. Then, we assume that y_i^P, the label of $\mathbf{x}_i^P$, is a good initial estimate of the label of $\mathbf{x}_i^t$ and use $[\mathbf{x}_i^t, y_i^P]$ to estimate the possibilistic membership u_i^P of $\mathbf{x}_i^t$ in the regression model f. The primary instance of test bag B^t is identified as the instance that

Algorithm 1 The RFC-MIR Algorithm

procedure RFC-MIR(D,C,m)
  Inputs:
Training data D, an overestimated number of clusters C, fuzzifier m
  Outputs: learned regression model f, set of primary instances $\mathcal{P}$

  Run FCM[30] for I_{iter} iterations to get initial partition
  % get initial C distinct regression models
  for I_{iter} iterations **do**
    update centers using (6)
    update covariance matrices using (7)
    update fuzzy memberships using (5)
  end
  % Refine C models by ignoring noise and outliers
  repeat
    update centers using (6)
    update covariance matrices using (7)
    update possibilistic memberships using (9)
  until All possibilistic memberships do not change significantly
  Merge similar clusters using (10)
  if number of remaining clusters $c' > 1$ **then**
    select "optimal" cluster using (11)
  Identify $\mathcal{P}$, the set of primary instances using (12)
  Identify regression model using (13)

has the highest possibilistic membership, i.e.

$$\mathbf{x}^t_{prim} = \{\mathbf{x}^t_k \mid u^P_k = \max_{i=1\ldots n} \{u^P_i\}\} \tag{14}$$

Finally, test bag B^t is labeled using

$$\hat{y}(B^t) = f\left(\mathbf{x}^t_{prim}\right) \tag{15}$$

Algorithm 2 The RFC-MIR-Predict Algorithm

procedure RFC-MIR-PREDICT(B^t,f,$\mathcal{P}$)
  Inputs: New test bag B^t, primary instances $\mathcal{P}$ from training data, learned regression model f.
  Outputs: Primary instance of B^t: $\mathbf{x}^t_{prim}$, Prediction for B^t: $\hat{y}(B^t)$
  for each $\mathbf{x}^t_i \in B^t$ **do**
    Find closest primary instance in $\mathcal{P}$, $\mathbf{x}^P_i$, to $\mathbf{x}^t_i$
    Approximate the label of $\mathbf{x}^t_i$ with the label of $\mathbf{x}^P_i$, y^P_i
    Estimate $u^P_i(\mathbf{x}^t_i)$ using $[\mathbf{x}^t_i, y^P_i]$ in (9)
  end for
  Identify primary instance of B^t,$\mathbf{x}^t_{prim}$, using (14)
  Label B^t using (15)

We should note here that it is possible to select multiple primary instances for each test bag (e.g all instances with possibilistic membership above a threshold). In this case, the label of B^t can be taken as the average of the labels of all primary instances. The proposed labeling algorithm is summarized in Algorithm 2.

4. Experimental Results

4.1. Synthetic datasets

To validate the proposed MIR and compare its performance to existing MIR approaches, we generate a series of synthetic multiple instance data sets with linear models. We vary the dimensionality of the feature space, the number of instances per bag, and the noise level added to the instances' features and bags' labels.

First, we generate the instances features, $\mathbf{x}_{ij} \in \mathbb{R}^d$, using

$$\mathbf{x}_{ij} = \mathbf{t}_i + \epsilon^F_{ij}, \quad for\ i = 1, \cdots, N_B, \quad and \quad j = 1, \cdots, n_i, \tag{16}$$

where $\mathbf{t}_i$ is the primary instance of bag, B_i, generated from a d-dimensional Gaussian distribution with zero mean and covariance $=10I^{d\times d}$. In (16), ϵ^F_{ij} is

a noise term added to the features. It is generated using a normal distribution $\mathcal{N}^F(\mu^F{=}0, \sum^F{=}\sigma^F I^{d\times d})$ for different values of σ^F. As the noise level increases, $\mathbf{x}_{ij}$ will divert from being a primary instance to an irrelevant one.

The label of each bag, B_i, is generated using

$$y_i = h(\mathbf{t}_i) + \epsilon_i^L, \tag{17}$$

where

$$h(\mathbf{x}) = \sum_{k=1}^{d} a_k x_k \tag{18}$$

is a linear d-dimensional function. In (18), a_k are constant coefficients that will be generated randomly. In (17), ϵ_i^L is a noise term, added to the true label. It is generated from a normal distribution $\mathcal{N}^L(\mu^L{=}0, \sigma^L)$ for different values of σ^L.

Using the above strategy, we generate multiple data sets by varying:

1. The dimensionality of the feature space, d from 1 to 10.
2. The noise level added to the features in (16). We let

$$\sigma^F = k_1 \times \sigma_0^F, \tag{19}$$

   with $\sigma_0^F{=}0.1$ and k_1 varies from 1 to 100.
3. The noise level added to the bags' labels in (17). We let

$$\sigma^L = k_2 \times \sigma_0^L, \tag{20}$$

   with $\sigma_0^L{=}0.05$ and k_2 varies from 1 to 25.
4. The number of instances per bag, n_i, from 5 to 100.

For each set of parameters, we create 10 linear models by generating random coefficients a_k (used in (18)). For each model, we generate one data collection that includes 100 bags, i.e. $N_B{=}100$.

First, we use a simple 1-D data to illustrate the different steps of the proposed RFC-MIR. The true model of this data is $h(x) = 6x$, and each bag has 5 instances. For the noise levels, we use $k_1{=}100$ and $k_2{=}2$.

This data is displayed in figure 2(a) where the x-axis represents the 1-D feature of the instances and the y-axis represents the label of each bag (all instances

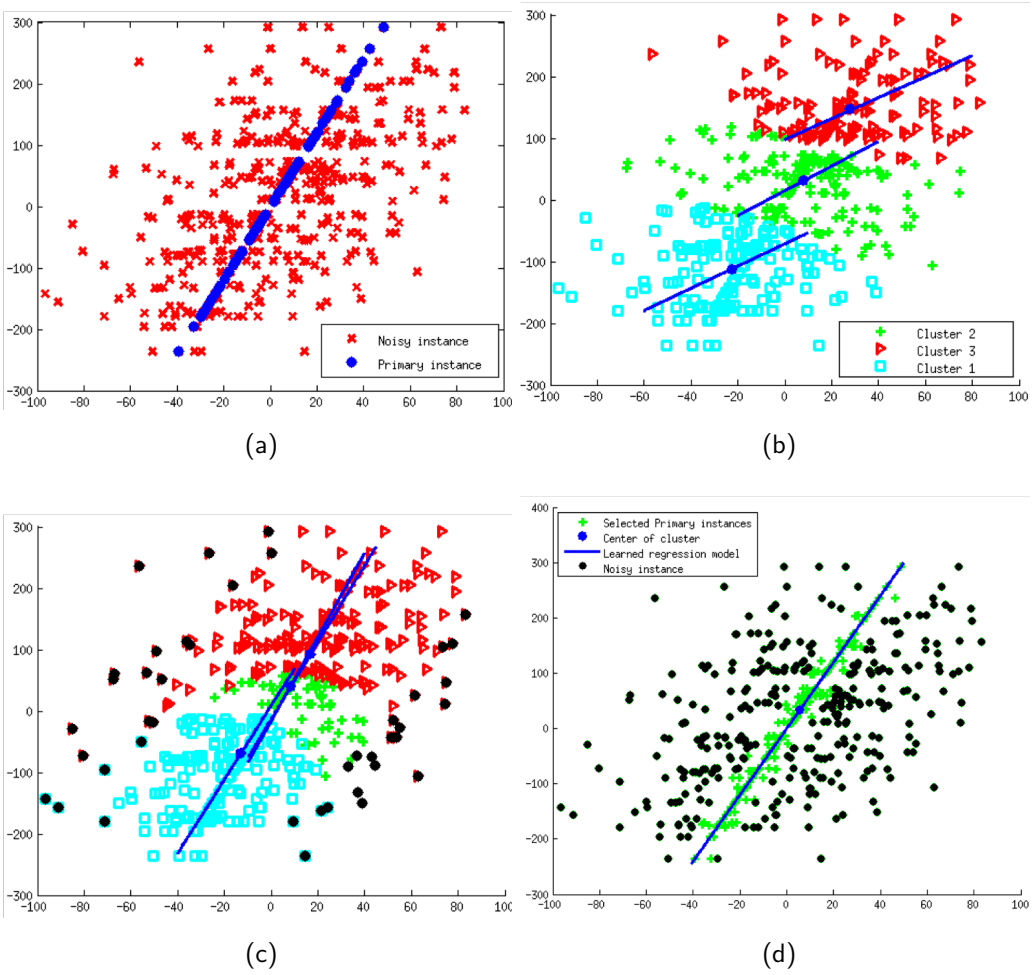

Figure 2: Illustrations of the steps of RFC-MIR: (a) Example of MIR data, (b) initial 3 clusters obtained using fuzzy memberships, (c) Result after 3 iterations with possibilistic memberships where the 3 clusters started converging to the same model and RFC-MIR starts identifying noisy instances (displayed as black dots), (d) Result after convergence and merging similar clusters.

in one bag have the same y value as they share the same label). All primary instances are displayed as filled blue circles and the remaining ones are displayed as red 'x'. Recall that in MIR this information is not available, and that we use it here for illustrative purposes only. Using C=3, figure 2(b) displays the 3 initial clusters obtained after running the RFC-MIR for few iterations with fuzzy memberships. Points that belong to different clusters are displayed with different symbols and colors. Figure 2(c) displays the results after switching from fuzzy to possibilistic memberships and running the algorithm for 3 iterations. As it can be seen, RFC-MIR started identifying noisy instances (displayed as black circles) and the 3 linear clusters started converging to the same true model. Figure 2(d) displays the final results after the clusters became identical and got merged into one using (10). Points with high possibilistic memberships (> 0.75) are located along the linear model. These points will be considered the primary instances. All others, will be treated as irrelevant ones.

Next, we compare the results of RFC-MIR with 4 MIR algorithms that were outlined in Section 2. These are the MI-Cluster Regress [23], the Instance-MIR and Aggregated-MIR [23, 24], and the Primary-MIR [25]. For each data set, we compute the mean square error (MSE) using:

$$MSE = \frac{1}{N_B} \sum_{i=1}^{N_B} (y_i - \hat{y}(B_i))^2, \tag{21}$$

where y_i is the true label of bag B_i and $\hat{y}(B_i)$ is the label estimated using the MIR algorithm being evaluated.

For all of the remaining experiments, we set the initial number of models C to 10, θ_M in (10) to 0.1, θ_P in (12) to 0.75, and the fuzzifier m to 2. The value of η_i in (9) is estimated using the average fuzzy intra-cluster distance of cluster i as recommended in [27]. In Algorithm 1, RFC-MIR converges (line 17) when the possibilitic memberships of all bags in all clusters do not change by more than 0.01 between 2 consecutive iterations.

In the following experiments, unless stated otherwise, we fix k_1, used to control the level of noise added to the instances in (19) and k_2, used to control the

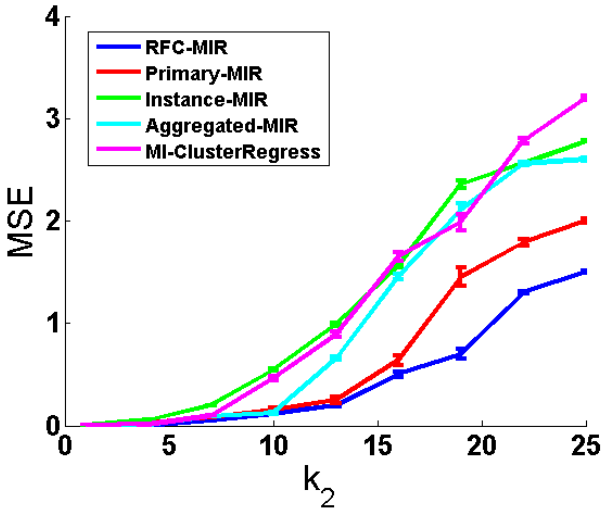

Figure 3: Comparison of RFC-MIR with 4 other MIR algorithms as the level of noise added to the true labels is increased.

noise added to bags' labels in (20) to 10. We also set the number of instances per bag, n_i, and the dimensionality of the instance space, d, to 5 and 1 respectively.

In the first experiment, we vary the noise level added to the bags' labels by increasing k_2 from 1 to 25. For each value of k_2, we generate 10 data sets using 10 linear models that use random coefficients (a_k's in (18)). The results of this experiment are displayed in Figure 3 where for each value of k_2, we display the mean MSE averaged over the 10 random models. We also display the variance of the MSE as a vertical error bar. As it can be seen, RFC-MIR has the lowest error. Moreover, the results of the 10 random models are consistent as indicated by the low MSE variations across the random models.

In a second experiment, we vary k_1 from 1 to 100. The results are displayed in Figure 4 where the proposed RFC-MIR has the lowest MSE average and variation. It is interesting to note that MI-ClusterRegress has almost the worst performance in Figure 3 and very competitive results in Figure 4. This suggests that MI-ClusterRegress is robust to noisy features since they are included in the clustering step. In fact, noisy features can be assigned to the non-prime cluster

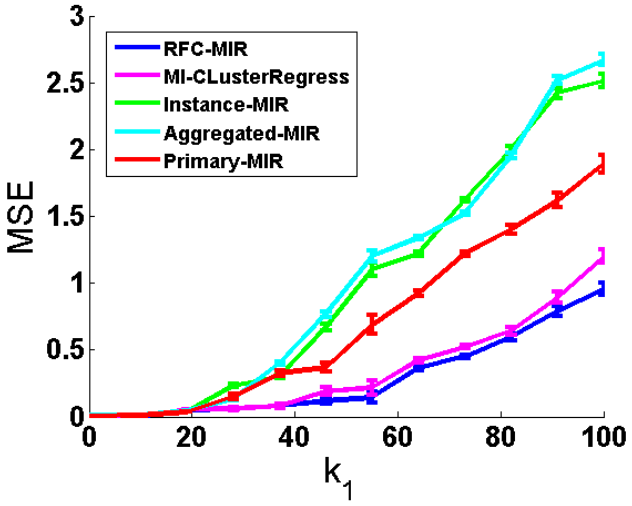

Figure 4: Comparison of RFC-MIR with 4 other MIR algorithms as the level of noise added to the features is increased.

and will not affect the learned model. On the other hand, MI-ClusterRegress is sensitive to noise present in the bag's labels. This is expected since these labels are not used in the clustering step but used later in learning the model of the prime cluster.

In a third experiment, we vary the number of instances per bag, n_i from 5 to 100. In general, adding more instances increases the number of irrelevant ones and makes the MIR problem more challenging. The results of this experiment are displayed in Figure 5. As it can be seen, the proposed RFC-MIR algorithm is very robust even in the presence of a large number of irrelevant instances. On the other hand, for all other 4 algorithms the average MSE increases at a much higher rate as more irrelevant instances are added to each bag.

In a fourth experiment, we vary the dimensionality of the instances, d, from 1 to 10. The results are displayed in Figure 6. As for the previous experiments, the proposed RFC-MIR maintains the lowest MSE values.

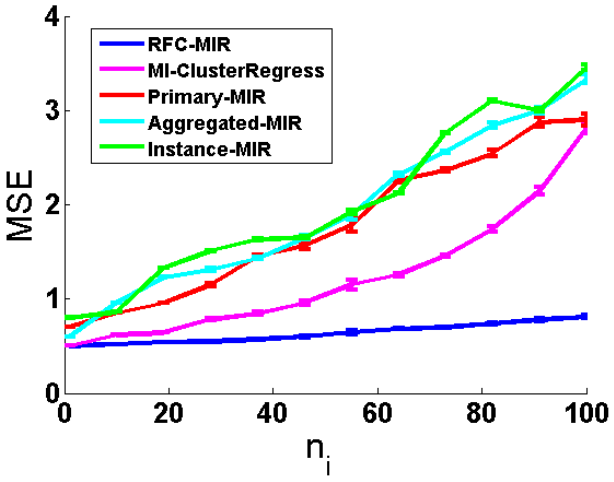

Figure 5: Comparison of RFC-MIR with 4 other MIR algorithms as the number of instances per bag is increased.

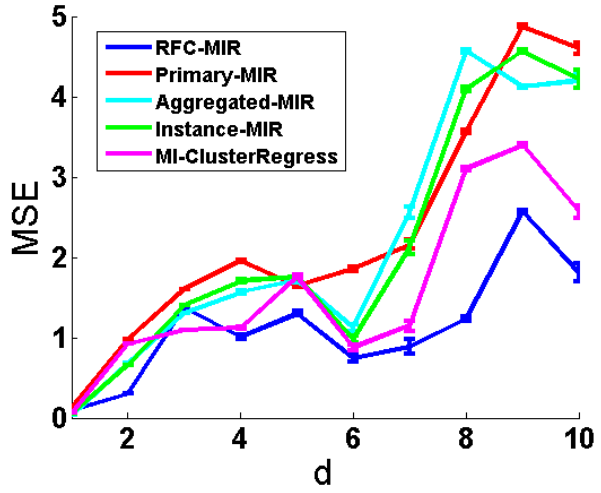

Figure 6: Comparison of RFC-MIR with 4 other MIR algorithms as we increase the dimensionality of the instances.

4.2. Application to remote sensing

A common application that has been used to validate most existing MIR algorithms is the prediction of crop yield based on remote sensing observations [13, 34, 23, 35]. Predicting the yearly average yield of a crop per acre for a given

region, especially when done early in the growing seasons, can be very beneficial. We use data collected by the MODIS instruments onboard satellites that provide cover of the entire US every 1-2 days [35]. In particular, we use the 8-day aggregate product which provides observations, in the red and near infrared (NIR), of each pixel location (250m×250m on the surface of the earth) every 8 days. The RED and NIR values are combined to generate the Normalized Difference Vegetation Index (NDVI) using:

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{22}$$

NDVI provides good indication of vegetation abundance and is good for identifying pixels that contain crops. Consequently, each pixel is represented by a time series where the i^{th} observation corresponds to the pixel's NDVI after $8 \times i$ days.

Multiple instance representation and learning is used for this application because it involves uncertainties at multiple levels. First, labels (average yield per acre) are available at the county level but are almost impossible to report at the pixel level. Also, the pixel-level NDVI feature can be used to discriminate between vegetation and other categories. However, it cannot distinguish between pixels that correspond to different types of crop. Thus, a group of pixels (e.g. within a county) should be considered collectively.

Considering all pixels within a bag as a group makes it possible to apply multiple instance learning methods. However, as with most MIR applications, additional assumptions are needed to predict the label of a new test bag. On one hand, predicting the output of each instance and combining them may be intuitive but it ignores the fact that a large number of pixels may belong to different regions such as cities, forest, water, etc. On the other hand, selecting one primary instance from each bag ignores the fact that multiple instances may be needed to predict the average yield of a county and that an average obtained from a single sample may not be accurate. In our experiments, we compare the prediction error of few methods that use different assumptions.

We use data from the California region over a period of 5 years (2001-2005)

to predict the yield of corn and wheat in each county. These are the same data sets used in [23] to validate MI-ClusterRegress. This application is challenging because each county contains thousands of pixels and we do not know which pixels (or even how many) contain the crop of interest. We use a randomly sub-sampled data such that 100 pixels are selected for each county. Thus, each county is represented by one bag of 100 instances[1] . Observations from the first 4 years (2001 – 2004) are used for training. The learned regression models are then used to predict the yield for 2005. Let f_D, for $D = 8, 16, \cdots, 360$ be the regression model to predict the yield at day D. f_D is trained with the sequence of NDVI observations taken every 8 days from the beginning of the year until day D. Thus, f_D will involve D/8-dimensional instance vectors. For each data, we run the five MIR algorithms 10 times and report the mean MSE and standard deviation of all runs.

Figure 7 compares the MSE of the five algorithms to predict corn yield and Figure 8 compares the results to predict wheat yield. We only consider the days of the growing season (days 140-280 for corn and days 0-180 for wheat). As it can be seen for both crops, RFC-MIR provides more accurate and consistent prediction.

To gain more insights and investigate the validity of the assumptions made by the different MIR algorithms to label test bags, in Figure 9 we plot the prediction error, $(y_i - \hat{y}(B_i))^2$, for each test bag, B_i, versus the fraction of instances within B_i that have been assigned high possibilistic membership values ($u^P \geq \theta^P = 0.75$) by RFC-MIR. We assume that these learned memberships can provide good estimates of the number of primary instances within each test bag. For this experiment, we use the corn data and we consider the values predicted at day 176 for each algorithm.

First, we note that the number of primary instances per bag can vary significantly: from 5% to almost 95%. For RFC-MIR, the prediction errors vary

[1]The pre-processed and sub-sampled data is publicly available at http://harvist.jpl.nasa.gov/papers.shtml

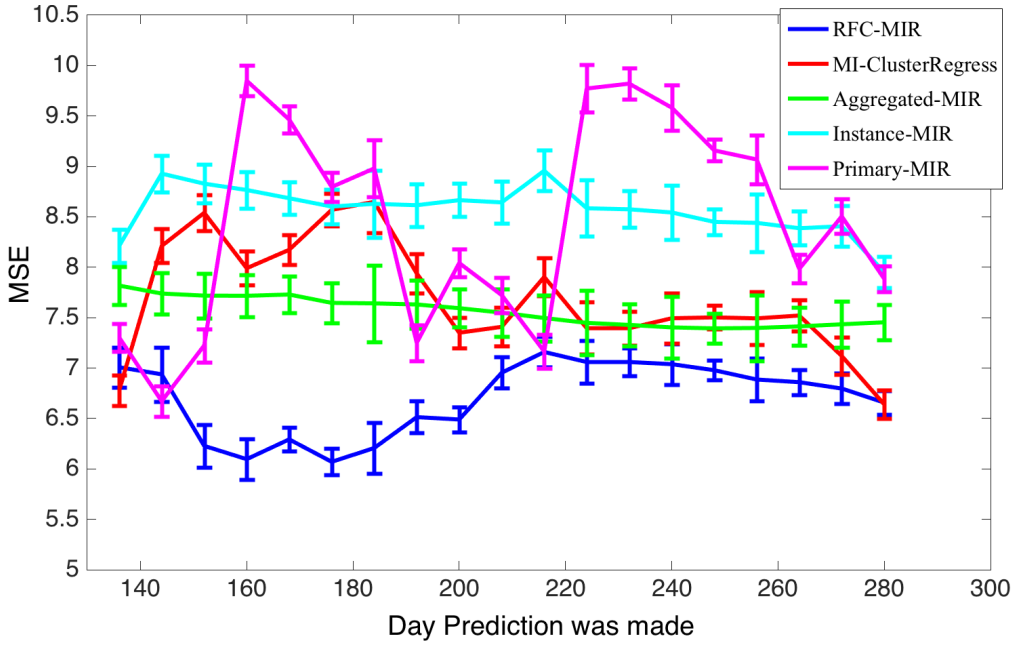

Figure 7: Comparison of the MSE of MI-ClusterRegress, Instance-MIR, Aggregated-MIR, Primary-MIR and RFC-MIR for predicting the yearly average yield of corn at different days of the growing season.

between 0 and 15 with no correlation to the number of primary instances per bag. This is not the case for the other algorithms. For example, in Figure 9(b), for Instance-MIR the prediction error is around 20 when the fraction of primary instances is less than 0.75, then it drops around 5 when more than 80% of the instances in a bag are primary. These results are expected since this method predicts the bag's output as the average output of all of its instances. Aggregated-MIR, Primary-MIR, and MI-ClusterRegress have similar behavior where the prediction error drops for most bags with more primary instances. For Aggregated-MIR and Primary-MIR, the prediction error remains high for few bags even if they have more than 80% primary instances. One possible explanation for Aggregated-MIR is that this method predicts the bag's output as the output of its meta-instance (mean of all instances), and this exemplar can be affected even by very few noisy instances. For Primary-MIR, one possible explanation is that the optimization (EM) used to identify the primary instance

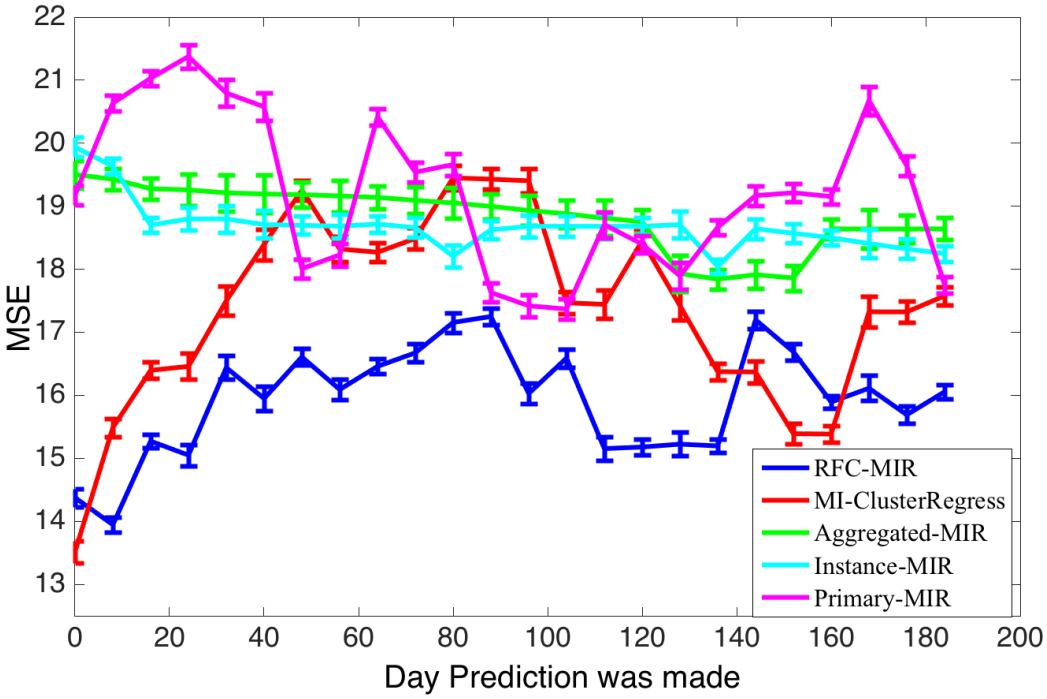

Figure 8: Comparison of the MSE of MI-ClusterRegress, Instance-MIR, Aggregated-MIR, Primary-MIR and RFC-MIR for predicting the yearly average yield of wheat at different days of the growing season.

of each bag can lead to sub-optimal solution for some bags.

4.3. Applications to Drug Activity Prediction

A well-known application, in pharmaceutical industry, that has been used to validate MIR algorithms is the drug activity prediction [36]. This application, known as Quantitative Structure-Activity Relationships (QSAR) [37], is based on the concept that a biological effect of a given drug is a function of its chemical structure.

Molecules can adopt multiple shapes by rotating some of their internal bonds. These rotations result in different conformations. Each conformation is characterized by potential energy that is determined by the interactions between the molecule's atoms. Conformers, that have the lowest energy, determine the chemical and biological properties of a molecule. Thus, only conformations that

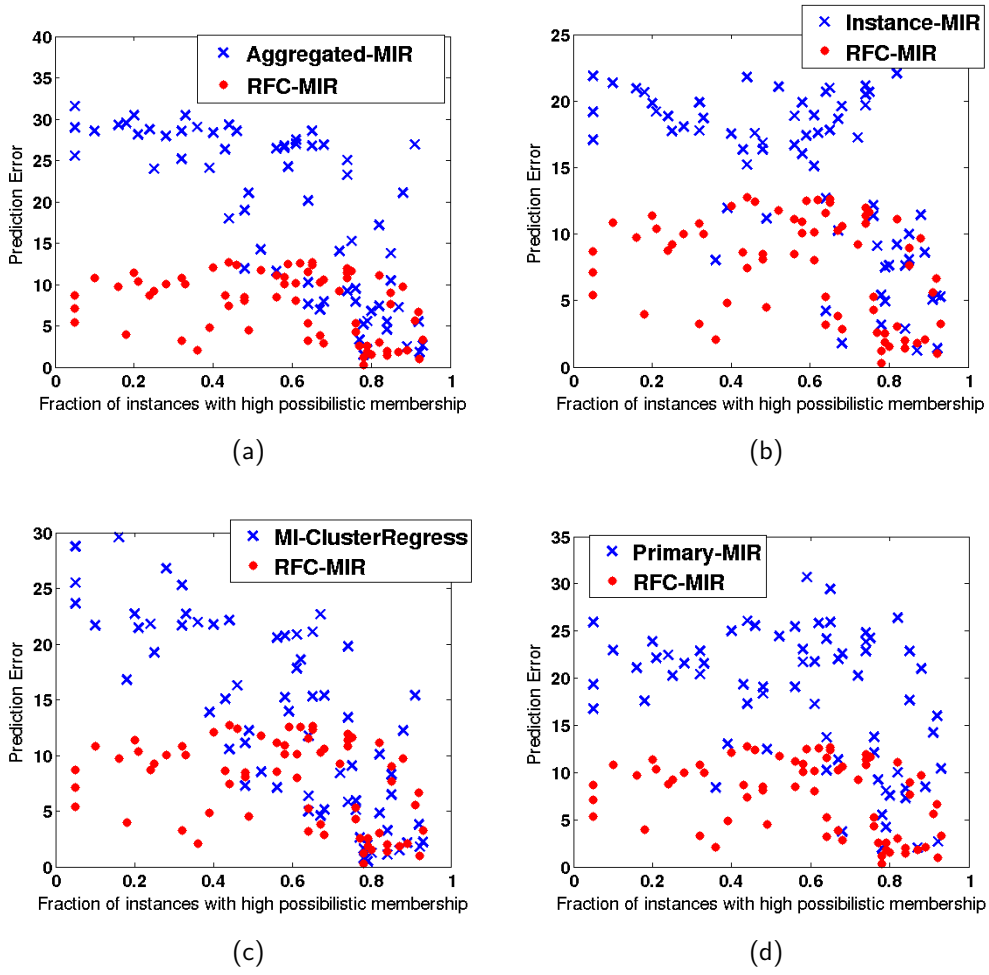

Figure 9: Prediction error for each bag vs. the fraction of instances within the bag that were identified by RFC-MIR as primary instances. The yearly average yield of corn was predicted at day 176 for each algorithm. (a) RFC-MIR vs. Aggregated-MIR; (b) RFC-MIR vs. Instance-MIR; (c) RFC-MIR vs. MI-ClusterRegress; and (d) RFC-MIR vs. Primary-MIR

correspond to local energy minima are possible candidates for binding. These low energy conformations can be computed using several methods including Monte Carlo search of bond-angle space [38], systematic bond-angle search [39], simulated annealing [40] and genetic algorithms [41, 42, 43].

Recently, some nonlinear approaches based on machine learning techniques, such as artificial neural network, have been proposed to predict drug-target interaction [44, 45]. Other approaches, such as Bayesian ranking prediction [46], are based on predetermined interactions between known molecules and targets. In [47], the authors propose an invariant representation of the molecule, using an inductive logic programming (ILP). The above approaches use a single feature vector representation as input to traditional regression algorithms. Using this encoding, information about individual conformations is lost. Consequently, the conformation that is responsible for the interaction with a given target cannot be recovered.

The multiple instance learning approach is suitable for this application since it maintains a representation of the multiple low energy conformations and any one can potentially be a binding candidate to the target protein. Within MIL, some approaches treat the drug-target interactions as a classification problem [48, 49]. These methods use a binary label for drug-target interactions, and predict the presence or absence of interaction between the pair. However, It may be more valuable and challenging to determine the binding affinity as a real value that represents the strength of the interaction between drug and target protein. In this case, the problem is treated as a multiple instance regression. Using MIR to predict drug-target interaction offers two main benefits: first, there is no need to have a global representation of the molecule and information about the individual conformations can be maintained. Second, the strength of the binding can be predicted. In an MIR setting, a drug is represented by a bag that contains all possible structures of this molecule and features extracted from each structure represent an instance within the bag. Labels at the instance level are not available and not needed, and the bag's label is the affinity of the drug to a given target protein. Given a set of drugs with their possible structures and

Table 1: Comparison of the accuracy of the proposed RFC-MIR with 4 MIR methods on the Thrombin Inhibitors dataset .

MI-Cluster Regress	Instance-MIR	Aggregated-MIR	Primary-MIR	RFC-MIR
3.89 ± 0.87	4.83 ± 0.97	4.25 ± 0.77	3.92 ± 0.89	$\mathbf{3.74 \pm 0.76}$

known affinities to a target protein, the objective of MIR is to learn a model that can predict the affinity to the target of a new drug.

To validate the applicability of our RFC-MIR appraoch to the drug-target interaction problem, we use a publicly available dataset that consists of Thrombin inhibitors [50] that can be used as anti-coagulant. This dataset consists of 40 thrombin inhibitors. Each drug or inhibitor contains between 3 and 334 structures and is assigned a real valued affinity to a target protein. Each instance or structure is a 6 dimensional feature vector. It corresponds to a 4-point pharmacophore representation [51]. In this representation, the Euclidean distances between 4 different chemical groups are calculated. This leads to a $\binom{4}{2}$=6 dimensional feature vector.

We divide the data into training and testing using 5 fold cross validation to evaluate RFC-MIR and compare it to other MIR algorithms. As in the previous experiments, we run each MIR algorithm 10 times and report the mean and standard deviation of the MSE across all runs.

The results are summarized in Table 1. As it can be seen, RFC-MIR outperforms the other methods in term of MSE and standard deviation. As in the previous experiments, we can attribute the better performance of the RFC-MIR to the possibilistic membership function that can ignore noisy instances, identify the primary instances and use them to learn the regression model. For this application, Instance-MIR and Aggregated-MIR have the worst performance because some bags can contain a very large number of instances (up to 334 conformations). In these cases, many of these instances are irrelevant. Thus, computing one meta-instance as the average of all instances or averaging the

output of all instances can lead to large prediction errors.

5. Conclusions

We proposed a new approach to multiple instance regression based on robust multi-model fitting. By combining the bags' instances and labels, and using an appropriate distance that measures the deviation of a point from a linear model, we showed that a possibilistic clustering algorithm can be used to estimate the regression model in a MIR setting. More importantly, we showed that the possibilistic memberships can be used to identify the primary instances and the irrelevant instances within each bag. Using several synthetic data sets with known structure and different levels of noise and difficulty, we showed that our approach achieves higher accuracy than state of the art methods. We have also validated our approach using real applications in remote sensing and drug activity prediction. Using a multiple instance data representation, we showed that RFC-MIR can be used to predict the bag's output without the need to label the training data at the instance level. We also showed that RFC-MIR can provide more accurate and consistent predictions than state of the art methods.

Currently, we assume that the regression model is linear and after clustering, we identify a single model that has instances from the maximum number of distinct bags and/or minimizes the fitting error. We are currently investigating two strategies to generalize our approach to non-linear regression. The first one is based on the assumption that a non-linear model can be approximated by multiple piecewise linear models. In this case, after convergence, instead of selecting the best model, we need to identify the multiple valid models and their domains. The second approach modifies the distance measure used within the clustering objective function to represent the fitting error with respect to a non-linear model.

Acknowledgment

This work was supported in part by U.S. Army Research Office Grants Number W911NF-13-1-0066 and W911NF-14-1-0589. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office, or the U.S. Government.

References

[1] T. Dietterich, R. Lathrop, T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, Artificial Intelligence 89 (1997) 31–71.

[2] O. Maron, Learning from ambiguity, Ph.D. thesis, Massachusetts Institute of Technology (1998).

[3] M.-A. Carbonneau, V. Cheplygina, E. Granger, G. Gagnon, Multiple instance learning: A survey of problem characteristics and applications, Pattern Recognition 77 (2018) 329 – 353.

[4] O. Maron, T. Lozano-Pérez, A framework for multiple-instance learning, Advances in Neural Information Processing Systems 10 (1) (1998) 570–576.

[5] R. Rahmani, S. A. Goldman, MISSL: Multiple-instance semi-supervised learning, in: Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 705–712.

[6] Y. Chen, J. Bi, J. Z. Wang, MILES: Multiple-instance learning via embedded instance selection, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (12) (2006) 1931–1947.

[7] C. Yang, M. Dong, F. Fotouhi, Region based image annotation through multiple-instance learning, in: Proceedings of the 13th annual ACM international conference on Multimedia, ACM, 2005, pp. 435–438.

[8] G. Melki, A. Cano, S. Ventura, Mirsvm: Multi-instance support vector machine with bag representatives, Pattern Recognition 79 (2018) 228 – 241.

[9] C. Zhang, X. Chen, W. B. Chen, An online multiple instance learning system for semantic image retrieval, in: Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007), 2007, pp. 83–84.

[10] A. Karem, H. Frigui, A multiple instance learning approach for landmine detection using ground penetrating radar, in: 2011 IEEE International Geoscience and Remote Sensing Symposium, 2011, pp. 878–881.

[11] A. Khalifa, H. Frigui, Fusion of multiple algorithms for detecting buried objects using fuzzy inference, in: Proc. SPIE, Vol. 9072, 2014, pp. 90720V–90720V–10.

[12] A. B. Khalifa, H. Frigui, A multiple instance neuro-fuzzy inference system for fusion of multiple landmine detection algorithms, in: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2015, pp. 4312–4315.

[13] Z. Wang, L. Lan, S. Vucetic, Mixture model for multiple instance regression and applications in remote sensing, IEEE Transactions on Geoscience and Remote Sensing 50 (6) (2012) 2226–2237.

[14] E. Alpaydın, V. Cheplygina, M. Loog, D. M. Tax, Single- vs. multiple-instance classification, Pattern Recognition 48 (9) (2015) 2831 – 2838.

[15] Q. Zhang, S. A. Goldman, EM-DD: An improved multiple-instance learning technique, in: In Advances in Neural Information Processing Systems, MIT Press, 2001, pp. 1073–1080.

[16] A. Karem, H. Frigui, Fuzzy clustering of multiple instance data, in: 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2015, pp. 1–7.

[17] J. Wang, J.-D. Zucker, Solving the multiple-instance problem: A lazy learning approach, in: Proceedings of the Seventeenth International Conference on Machine Learning, ICML '00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000, pp. 1119–1126.

[18] V. Cheplygina, D. M. Tax, M. Loog, Multiple instance learning with bag dissimilarities, Pattern Recognition 48 (1) (2015) 264 – 275.

[19] M. Qiao, L. Liu, J. Yu, C. Xu, D. Tao, Diversified dictionaries for multi-instance learning, Pattern Recognition 64 (2017) 407 – 416.

[20] J. J.-Y. Wang, I. W.-H. Tsang, X. Cui, Z. Lu, X. Gao, Multi-instance dictionary learning via multivariate performance measure optimization, Pattern Recognition 66 (2017) 448 – 459.

[21] Y. Chen, J. Z. Wang, D. Geman, Image categorization by learning and reasoning with regions, Journal of Machine Learning Research 5 (2004) 913–939.

[22] X. Wang, Y. Yan, P. Tang, X. Bai, W. Liu, Revisiting multiple instance neural networks, Pattern Recognition 74 (2018) 15 – 24.

[23] K. L. Wagstaff, T. Lane, A. Roper, Multiple-instance regression with structured data, in: 2008 IEEE International Conference on Data Mining Workshops, 2008, pp. 291–300.

[24] Z. Wang, V. Radosavljevic, B. Han, Z. Obradovic, S. Vucetic, Aerosol optical depth prediction from satellite observations by multiple instance regression, in: Proceedings of the 2008 SIAM International Conference on Data Mining, 2008, pp. 165–176.

[25] S. Ray, D. Page, Multiple instance regression, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Morgan Kaufmann Publishers Inc., 2001, pp. 425–432.

[26] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, University of California Press, 1967, pp. 281–297.

[27] R. Krishnapuram, J. Keller, A possibilistic approach to clustering, Fuzzy Systems, IEEE Transactions on 1 (2) (1993) 98–110.

[28] H. Frigui, R. Krishnapuram, A comparison of fuzzy shell-clustering methods for the detection of ellipses, IEEE Transactions on Fuzzy Systems 4 (2) (1996) 193–199.

[29] F. Höppner, F. Klawonn, R. Kruse, T. Runkler, Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition, John Wiley and Sons, England, 1999.

[30] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer Academic Publishers, Norwell, MA, USA, 1981.

[31] R. N. Dave, Use of the adaptive fuzzy clustering algorithm to detect lines in digital images, in: Proc. SPIE, Vol. 1192, 1990, pp. 600–611.

[32] K. R. Frigui H., A robust algorithm for automatic extraction of an unknown number of clusters from noisy data, Pattern Recognition Letters 17 (12) (1996) 1223 – 1232.

[33] R. Krishnapuram, J. Keller, The possibilistic c-means algorithm: Insights and recommendations, IEEE Transactions on Fuzzy Systems 4 (3) (1996) 385–393.

[34] J. You, X. Li, M. Low, D. Lobell, S. Ermon, Deep gaussian process for crop yield prediction based on remote sensing data, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA., 2017, pp. 4559–4566.

[35] K. L. Wagstaff, T. Lane, Salience assignment for multiple-instance regression, in: ICML '2007 Workshop on Constrained Optimization and Structured Output Spaces, Corvallis, OR, 2007, 2007.

[36] J. Davis, S. Ray, Tightly integrating relational learning and multiple-instance regression for real-valued drug activity prediction, International Conference on Machine Learning.

[37] X.-F. Zhou, Q. Shao, R. A. Coburn, M. E. Morris, Quantitative structure activity relationship and quantitative structure-pharmacokinetics relationship of 1,4-dihydropyridines and pyridines as multidrug resistance modulators, Pharmaceutical Research 22 (12).

[38] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, D. Baker, Protein structure prediction using rosetta methods in enzymology, Numerical Computer Methods, Part D 383 (2004) 66 – 93.

[39] D. D.Beusen, E. Shands, Systematic search strategies in conformational analysis, Drug Discovery Today. 1 (1996) 429 – 437.

[40] S. R. Wilson, W. Cui, Conformation Searching Using Simulated Annealing, Birkhäuser Boston, Boston, MA, 1994, pp. 43–70.

[41] N. Nair, J. M. Goodman, Genetic algorithms in conformational analysis, Journal of Chemical Information and Computer Sciences 38 (2) (1998) 317–320.

[42] Y. Sakae, T. Hiroyasu, M. Miki, K. Ishii, Y. Okamoto, A conformational search method for protein systems using genetic crossover and metropolis criterion, Journal of Physics: Conference Series 487 (1) (2014) 012003.

[43] A. Supady, V. Blum, C. Baldauf, First-principles molecular structure search with a genetic algorithm, Journal of Chemical Information and Modeling 55 (11) (2015) 2338–2348.

[44] A. N. J. Thomas, G. D. Richard, H. L. David, C. R. E., C. J. Barr, E. B. Teresa, A. W. T. Lozano-Perez, Compass: A shape-based machine learning tool for drug design, Journal of Computer-Aided Molecular Design 8 (6) (1994) 635–652.

[45] T. Andrea, H. Kalayeh, Applications of neural networks in quantitative structure-activity-relationships of dihydrofolate-reductase inhibitors, Journal of Medicinal Chemistry 34 (1991) 2824–2836.

[46] L. Peska, K. Buza, J. Koller, Drug-target interaction prediction: A bayesian ranking approach, Computer Methods and Programs in Biomedicine 152 (2017) 15 – 21.

[47] N. Marchand-Geneste, K. A. Watson, B. K. Alsberg, R. D. King, New approach to pharmacophore mapping and qsar analysis using inductive logic programming. application to thermolysin inhibitors and glycogen phosphorylase b inhibitors, Journal of Medicinal Chemistry 45 (2) (2002) 399–409.

[48] Z. Zhao, G. Fu, S. Liu, K. M. Elokely, R. J. Doerksen, Y. Chen, D. E. Wilkins, Drug activity prediction using multiple-instance learning via joint instance and feature selection, BMC Bioinformatics 14 (14) (2013) S16.

[49] G. Fu, X. Nan, H. Liu, R. Y. Patel, P. R. Daga, Y. Chen, D. E. Wilkins, R. J. Doerksen, Implementation of multiple-instance learning in drug activity prediction, BMC Bioinformatics 13 (15) (2012) S3.

[50] J. Cheng, C. Hatzis, H. Hayashi, M.-A. Krogel, S. Morishita, D. Page, J. Sese, Kdd cup 2001 report, SIGKDD Exploration 3 47–64.

[51] P. Finn, S. Muggleton, D. Page, A. Srinivasan, Pharmacophore discovery using the inductive logic programming system progol, Machine Learning 30 (1998) 241–270.