Chapter 6
Multi-instance Regression

Abstract Regression is a popular machine learning task that aims to predict a numerical outcome. In multi-instance regression (MIR), each observation can be described by several instances. After a brief introduction to this topic in Sect. 6.1, we present a formal definition of MIR and its appropriate evaluation measures in Sect. 6.2. We organize the MIR methods in two main categories. Algorithms that focus on individual instances of each bag in their construction of a regression model are examined in Sect. 6.3, while Sect. 6.4 discusses methods that treat bags as single entities to create a regression model operating at the bag level. Section 6.5 lists some summarizing remarks.

6.1 Introduction

The multi-instance regression task (MIR) is the natural extension of traditional (single-instance) regression to the multi-instance setting. MIR models the data in the same way as MIC, with the important difference that each bag is associated with a real-valued outcome and not a class. The MIR objective is to approximate, based on the training bags, a function that can predict the outcome of future bags as accurately as possible. In Sect. 6.2, we present a more formal description of MIR. Compared to the traditional regression task, the ambiguity introduced by the multiple descriptions for every bag as well as the lack of information on how these descriptions relate to the bag label make MIR intrinsically more challenging.

MIR has been studied much less than the multi-instance classification task. Nevertheless, it is of great importance for two main reasons. On the one hand, regression provides a theoretical basis to understand many classification methods and can generate useful ideas to the design of more effective classifiers. On the other hand, an important motivation for the development of new algorithms is that many real-life applications can be successfully modeled as MIR problems. These include drug activity prediction, landmark recognition, remote sensing systems, age estimation, and sentiment analysis (see Sect. 2.4.7).

We have already shown that MIC methods can be grouped into two major categories, namely instance-based and bag-based methods. MIR methods can be divided

F. Herrera et al., *Multiple Instance Learning*, DOI 10.1007/978-3-319-47759-6_6

into these two categories as well. Instance-based MIR methods try to determine a regression function for one prime instance, a subgroup of instances or all instances inside the bag. The bag label is a function of the prime instance label or an aggregation of several instance labels. This group of methods is discussed in Sect. 6.3. Bag-based MIR methods treat each bag as a whole entity. These methods can rely on bag-wise distance or kernel functions or can be based on mapping functions that represent bags as single vectors on which single-instance regression models can be learned. Section 6.4 is devoted to bag-based MIR methods.

6.2 MIR Formulation

We begin our discussion on MIR methods with a brief description of the setting and objective of this learning task. Evaluation measures to assess the quality of MIR models are presented as well.

6.2.1 Problem Description

In a MIR problem, the training set $D = (\mathbf{X}, \mathbf{Y})$ consists of m bags $\mathbf{X} = \langle X_1, \ldots X_m \rangle$ and their corresponding real-valued labels $\mathbf{Y} = \langle y_1, \ldots y_m \rangle$ with $(\forall i = 1 \ldots m)(y_i \in \mathbb{R})$. Each bag X_i has n_i instances $\left\{x_{i1}, \ldots, x_{in_i}\right\}$ and each instance x_{ij} is described by d features. In the simplest case all features take on real values, but in general real-valued features can be mixed with categorical ones. Instance labels are not available. The goal of MIR is to determine a function f over the bag space $\mathbb{N}^{\mathbb{X}}$ which can make predictions $\hat{y}_i = f(X_i)$ of the label y_i of new bags X_i as accurately as possible.

A geometrical view of MIR is presented in Fig. 6.1. If we project the bag space in a Cartesian plane, each bag can be viewed as a region in that plane. For simplicity, the bag space in Fig. 6.1 has only two features $\mathscr{X}_1$ and $\mathscr{X}_2$. Bag labels are represented as scores on the $\mathscr{Y}$ axis, orthogonal to the $\mathscr{X}_1$–$\mathscr{X}_2$ plane. When there is an infinite number of bags covering the whole bag space, their label ordinates form a surface over the plane. MIR aims to find a function f that is the best possible approximation to that surface. As a consequence, f is sometimes referred to as the regression surface.

6.2.2 Evaluation Measures

Evaluation measures are needed to assess how well the regression function f approximates the real process. Any validation scheme (Sect. 1.4.1) can be used as appropriate to the problem at hand. Despite the difference in data representation, MIR is similar to traditional regression in the type of variable to be predicted. The same evaluation metrics can therefore be used. The most common evaluation measures used for

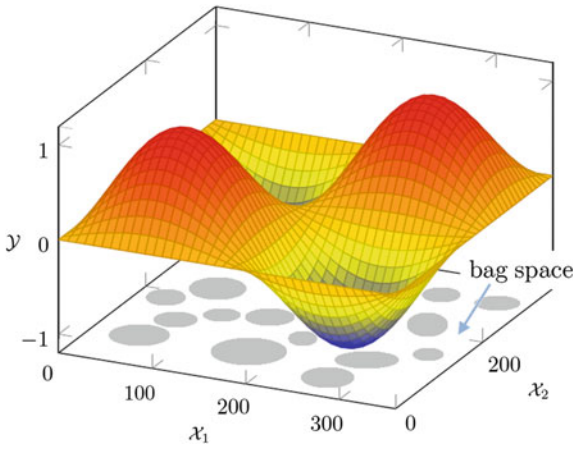

Fig. 6.1 Geometrical view of MIR

regression are the mean absolute deviation (MAD) computed as

$$MAD = \frac{1}{n}\sum_{i=1}^{n} |y_i - f(X_i)|$$

and the mean squared error (MSE) given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - f(X_i))^2 .$$

In both cases, the actual label y_i is compared with the predicted outcome $f(X_i)$ and their differences are averaged over n test bags. In MSE, these differences are squared, such that larger errors are penalized more.

6.3 Instance-Based Regression Methods

Chapter 4 showed that instance-based classifiers perform their main learning process at instance level. The same goes for instance-based regression methods. Attention is payed to the individual instances within the bags. In general, one representative instance is selected or generated for each bag and a regression model is constructed over these instances. Any traditional single-instance regression learner can be used, since it is fitted with just one instance from each bag. The most important design option of instance-based regression methods is the way to model the relation of instances with the bag label. Two assumptions dominate the MIR literature

- **The prime instance assumption (Sect. 6.3.1)**: there is a single instance in every bag which is responsible for the bag label. Algorithms based on the prime instance assumption strive to select this "correct" instance in the bag.
- **The collective assumption (Sect. 6.3.2)**: each instance in the bag makes a (possibly different) contribution to the bag label. Methods based on this assumption try to determine the weight that each instance has in the prediction of the bag label.

A second source of differences between instance-based methods lies with the selection of the regression model and solution procedure to be used. Commonly, one assumes a class of regression functions f on X depending on a set of parameters Θ, i.e., $f \equiv f(X, \Theta)$. As an example, consider the class of linear regression functions consisting of a hyperplane determined by a normal vector W. To guarantee that the obtained regression model $f(X, \Theta)$ is the best possible approximation according to the training data, an optimization approach

$$\Theta^* = \arg\min_{\Theta} \sum_{i=1}^{m} L(y_i, f(X_i, \Theta)) + \lambda R(\Theta) \tag{6.1}$$

must be adopted. In expression (6.1), L is a loss function that indicates how well the prediction model performs and $R(\Theta)$ is a regularization term favoring simple models and thereby avoiding data overfitting. The parameter $\lambda > 0$ is a trade-off between prediction accuracy and model complexity. The optimization solution Θ^* is the best set of parameters defining the optimal regression model. Several loss functions and regularization forms can be used in (6.1). For example, when squared loss is combined with squared norm regularization we can obtain a closed form solution. In other cases, gradient descent optimization methods need to be applied to find an approximation. As this is a typical formulation of regression models, abundant explanations can be found in any modern machine-learning book.

In the following sections, we provide an in-depth analysis of the most relevant assumptions involved in instance-based regression methods. Representative algorithms are briefly described in each case.

6.3.1 Prime Instance Assumption

This approach assumes that the bag label is determined by only one instance in the bag, namely the primary or prime instance. The remaining instances are considered noisy observations of the prime instance. This assumption was proposed in the seminal work of Ray and Page [12] and has had a great impact on MIR works that followed. Inspiration was drawn from the standard MIL assumption for classification, which states that a single positive instance in a bag suffices for a positive bag label (Sect. 3.4.1).

6.3.1.1 Prime-MIR Algorithm

Formally, the prime-instance algorithm looks for the optimal regression model $f(x, \Theta)$ in (6.1) based on the set of prime instances $\{x_{1p}, \ldots, x_{mp}\}$, where x_{ip} is the prime instance in bag X_i. The general prime-instance regression model is defined as

$$\Theta^* = \arg\min_{\Theta} \sum_{i=1}^{m} L\left(y_i, f\left(x_{ip}, \Theta\right)\right) + \lambda R\left(\Theta\right). \tag{6.2}$$

In [10], the authors prove that the exhaustive search of the set of prime instances satisfying (6.2) has an NP-complete computational complexity. An approximating solution method is used in the form of an expectation-maximization (EM) algorithm. First, prime instances are selected at random from each bag. This initial guess is subsequently refined by iterating between expectation and maximization steps. In the maximization step, a new regression model is trained by using the current hypothesis of prime instances. In the expectation step, new candidates of prime instances are found by selecting the instance from each bag that has the lowest prediction error according to the current regression model. These steps are repeated until convergence.

The regression model $f(x, \Theta)$ obtained by means of (6.2) represents the hyperplane Θ that best approximates the prime instance outcomes, which in turn are supposed to be the best approximation of the bag labels. When a new bag is presented to the model, its prime instance needs to be located in order to evaluate the model and obtain the bag label. However, the model of Ray and Page [12] does not provide any information on which element is the prime instance. Although Cheung and Kwok [4] and Ray [10] identified problem domains in which it is possible to assume that the prime instance is the one with the largest output value, it is not possible to generalize this heuristic to other domains. Wang et al. [20] suggest a statistical solution by using the mean of the predictions of all instances in the new bag X_i, namely

$$\hat{y}_i = mean\left(f\left(x_{i1}\right), \ldots, f\left(x_{in_i}\right)\right). \tag{6.3}$$

To increase the prediction robustness against outliers, the median of the instance predictions can be more appropriate in some applications

$$\hat{y}_i = median\left(f\left(x_{i1}\right), \ldots, f\left(x_{in_i}\right)\right). \tag{6.4}$$

While the original proposal [10] was very simple and uses a non-regularized linear regression model, the general prime-instance model is more sophisticated. In particular, nonlinear models can also be applied, as for example neural networks [20]. Another improvement is proposed by Wang et al. [20]. Instead of initializing prime instances at random, the first selection can be based on predictions made by a simple MIR model constructed on the training data. Prime-instance regression algorithms have been used as benchmark in several studies [9, 19, 20].

6.3.1.2 Two Levels Regularization Framework

In Sect. 4.5, we described the MI-SVM method [2], an instance-based classification SVM based on the standard MIL assumption. It selects a primary instance from each bag and trains a standard SVM with these instances. In Sect. 5.2.2, we described the MI kernel [7], a kernel defined over whole bags. It allows for an SVM to be used at bag level in the classification of multi-instance data. Cheung and Kwok [4] translate both ideas to the regression setting and connect them in a unified framework. A general loss function that depends on both the training bags and training instances is defined. This function is split into two parts. The first part considers the loss between each bag label y_i and its prediction $f(X_i)$, using the hinge loss function $\max(0, 1 - y_i f(X_i))$. The second part considers the loss between the prediction of each bag $f(X_i)$ and those of its constituent instances $\left\{f\left(x_{ij}\right) | j = 1, \ldots, n_i\right\}$. It can be defined in various ways. Cheung and Kwok present margin formulations for the L1 loss $\ell(\upsilon_1, \upsilon_2) = |\upsilon_1 - \upsilon_2|$, the L2 loss $\ell(\upsilon_1, \upsilon_2) = (\upsilon_1 - \upsilon_2)^2$ and the ε-insensitive loss $\ell(\upsilon_1, \upsilon_2; \varepsilon) = \max(0, |\upsilon_1 - \upsilon_2| - \varepsilon)$. The complete loss function is defined as

$$V\left(\{X_i, y_i, f(X_i)\}_i, \left\{f\left(x_{ij}\right)\right\}_{ij}\right) = \frac{1}{m}\sum_{i=1}^{m} \max(0, 1 - y_i f(X_i)) + \frac{\lambda}{m}\sum_{i=1}^{m} \ell\left(f(X_i), \max_j f\left(x_{ij}\right)\right), \quad (6.5)$$

where λ is a parameter that trades off the two components. Based on the fact that an instance can also be considered as a bag of size one, they use the representer theorem and the Constrained Concave–Convex Procedure (CCCP) to solve the problem as a quadratic programming problem, which guarantees the convergence to a local optimum. As the max operator is not a smooth function, the gradient is replaced by a convenient sub-gradient in each iteration of the CCCP procedure.

In regression problems, the loss functions has two parts as well. The first part considers the loss between the value of each bag and its corresponding prediction. As in ν-support vector regression [13], they use the ε-insensitive loss and an extra $\nu\varepsilon$ term (where ν is a user-defined parameter) to penalize the value of ε. The second part considers the loss between the prediction of each bag and those of its constituent instances. Following Ray and Page [12], they assume that there is one primary instance in each bag that is responsible for the output of the bag, which is set to the one with the highest output value. By introducing slack variables δ_i, ξ_i, ξ_i^* the following optimization problem is presented:

$$\begin{aligned} \min_{w,b,\xi,\xi^*,\varepsilon,\delta} \quad & \tfrac{1}{2}\|w\|^2 + C\nu\varepsilon + \tfrac{C}{m}\textstyle\sum_i \xi_i + \tfrac{C}{m}\sum_i \xi_i^* + \tfrac{C\lambda}{m}\sum_i \delta_i \\ s.t. \quad & f(X_i) - y_i \le \varepsilon + \xi_i, \\ & y_i - f(X_i) \le \varepsilon + \xi_i^*, \\ & -\delta_i \le f(X_i) - \max_j f\left(x_{ij}\right) \le \delta_i \\ & \xi_i, \xi_i^*, \delta_i, \varepsilon \ge 0. \end{aligned} \tag{6.6}$$

Different loss functions for MI regression can be used. Both bags and instances directly participate in the optimization process.

6.3.1.3 Probabilistic Prime-MIR Algorithm

The above prime-instance methods make strong assumptions on the prime instances. The probabilistic method of Wang et al. [19] assumes that each instance has a certain probability to be the prime instance of the bag. Under this assumption, the bag label is treated as a random variable described by the mixture model

$$p(y_i|X_i) = \sum_{j=1}^{n_i} \pi_{ij} p\left(y_i|x_{ij}\right), \tag{6.7}$$

where π_{ij} is the prior probability that the jth instance is the prime instance of the ith bag. The value $p\left(y_i|x_{ij}\right)$ is the label probability in case the jth instance is the prime instance. In the mixture model (6.7), the contribution of each instance to the bag label is proportional to its probability of being the prime instance. The label of the ith bag can be predicted as the expected value of the mixture model, which is the weighted sum of the label probabilities for individual instances

$$\hat{y}(X_i) = \sum_{j=1}^{n_i} \pi_{ij}\left(\boldsymbol{\theta}_g\right) f\left(x_{ij}, \mathbf{w}\right),$$

where the prior probability π_{ij} is a function of the model parameters $\boldsymbol{\theta}_g$ and $f\left(x_{ij}, \mathbf{w}\right)$ is a regression function with parameters $\mathbf{w}$. The parametrized probabilities π_{ij} and $p\left(y_i|x_{ij}\right)$ are learned from the data using the expectation maximization algorithm. Model parameters are randomly initialized at first. Later, expectation and maximization steps are alternated until convergence. In the expectation step, the algorithm evaluates the expected value of the log-likelihood of the training data with respect to the current estimate of the model parameters. In the maximization step, the algorithm updates these parameters to maximize the expectation. The framework allows for the prime instance probability to be modeled as appropriate to the application at hand. The method was successfully applied to two remote-sensing applications [19].

6.3.1.4 Prime Instance-Based Applications

The prime instance assumption has been used in several real-life applications. In [15], it was applied to the prediction of protein-ligand binding affinities to guide optimization in structure-based drug discovery. A bag corresponds to a protein-ligand pair and its instances are binding poses. The binding affinity of the ligand to a particular protein is to be predicted. The gradient boosting approach [6] is used, where an additive model $f(x) = \sum f_k(x)$ is constructed to minimize a squared loss function. In each boosting iteration, the instance with maximal output x_i^{max} is selected from the ith bag. It is based on the domain-specific assumption that the most plausible binding pose is that with the maximum predicted binding score. A new training set $R = \left\{\left(x_i^{max}, r_i\right), i = 1, \ldots N\right\}$ is arranged from the N training bags, where $r_i = y_i - f\left(x_i^{max}\right)$ is the pseudo-residual from the bag label and the predicted output. This step effectively converts the initial MIR problem into a single-instance regression problem. A single-instance regression model $f_k(x)$ is trained on R and added to the boosting function $f(x)$ to decrease the value of the loss function.

Another application of the prime instance assumption has been in the prediction of polyp size in computed tomography images (CT) [8]. Polyps are precursors of cancer tumors and their dimension indicates cancer staging. A polyp is represented by a bag of polyp-like candidates extracted from a 3D CT scan. The polyp size y needs to be predicted. It is approximated by a hyperplane $w^T x$ described by the weight vector w. To make the size prediction, it is desirable to use the candidate whose segmentation is the closest to the actual layout of the polyp. Therefore, the authors assume that the primary instance is that whose estimated output $f(x) = w^T x$ differs the least from y, i.e., $\min |y - f(x)|$. Under this assumption, a ridge regression model is defined as

$$\min_w \sum_j \min_{i \in I_j} \left(y_j - w^T x_{ji}\right)^2 + \lambda \|w\|^2,$$

where I_j is the index set of instances belonging to the jth bag. Like the prime-instance algorithm described above, an EM approach is used to solve the optimization problem.

6.3.2 *Collective Assumption*

In Sect. 3.4.3, we described the collective MIL assumption for classification, which states that all instances in the bag contribute equally to the bag label. An extension of the collective assumption allows that each instance contributes independently, but not necessarily equally, to the class label of the bag. In this section, we consider a similar idea applied to the regression setting. The notion of a primary instance inside each bag is abandoned. Instead, all instances contribute equally to the bag label. We present an implementation of this assumption in Sect. 6.3.2.1. In the extended

collective MIR assumption, not all instances have the same contribution to the bag label. Some instances are noisy and should be discarded, which is the idea behind the method described in Sect. 6.3.2.2. Alternatively, we can try to determine the weight that each instance has in the bag label formation. The algorithm described in Sect. 6.3.2.3 exploits that information to make its bag label predictions.

6.3.2.1 Instance-MIR Algorithm

The Instance-MIR algorithm is the regression counterpart of the wrapper classifier discussed in Sect. 4.2. Each instance x_{ij} from each bag X_i receives the label y_i of its bag. All instances are joined into a single-instance dataset $D = \left\{\left(x_{ij}, y_i\right), i = 1 \ldots m, j = 1 \ldots n_i\right\}$. To ensure that all bags are represented with the same importance in D, independently of their size, each bag is sampled with replacement and added to D the same number of times. An ordinary regression model f is trained on D. As in the collective assumption for classification (Sect. 3.4.3), the bag label probability is the expected outcome value of the instance population estimated by the sample mean (6.3). Alternatively, the median of the instance predictions (6.4) can be used in some applications to prevent outliers. The Instance-MIR algorithm has been used as benchmark in several studies [9, 11, 19, 20], showing competitive results on many datasets despite its simplicity.

6.3.2.2 Pruning-MIR Algorithm

The Instance-MIR algorithm, described in the previous section, uses all available training instances to construct the regression model. When bags contain many noisy instances, their inclusion can have a detrimental effect. On the other extreme, the Prime-MIR algorithm selects a single instance from each bag, which makes it highly probable that informative instances are discarded. The Pruning-MIR algorithm proposed in [20] is a compromise solution between these two extremes. The assumption is that each bag is generated by some random noise around a prime point in instance space. Bag labels are assumed to be generated by some function of the prime instances with added noise.

The algorithm aims to keep relevant instances from each bag, while removing those that seem noisy. It starts from the Instance-MIR solution. In each iteration, it discards a small fraction of the noisiest instances in each bag and trains a new predictor (using Instance-MIR) on the remaining instances. The noisiest instances in a bag are defined as those whose predictions are the farthest away from the median prediction over the non-pruned instances. In this way, noise is gradually removed and the quality of the training data is improved. The algorithm runs for as long as there is an improvement in prediction accuracy.

6.3.2.3 Weighted-MIR Algorithm

The weighted collective assumption states that each instance has a particular relevance or weight in the bag label generation. Under this assumption, the label $\hat{y}$ of a bag X_i can be calculated as the weighted aggregation of the prediction of bag instances, namely

$$\hat{y}\left(X_i\right) = \frac{\sum_j f\left(x_{ij}\right) w_{ij}}{\sum_j w_{ij}}, \tag{6.8}$$

where the denominator is present for normalization. However, expression (6.8) gives rise to problems. We need to determine both the instance weights w_{ij} and the regression model f.

Regression Based on Instance Weights

Wagstaff and Lane [16] develop a method to estimate instance weights in MIR under the assumption that each instance contributes independently to the bag label. Given a set of m bags $\{X_1, \ldots, X_m\}$ and their respective labels $Y = \{y_1, \ldots, y_m\}$, it is assumed that an exemplar p_i exists inside each bag X_i that can accurately predict the bag's true label, that is, $y_i = f\left(p_i\right)$. The exemplar can be described as a convex combination of instances, namely $p_i = \sum_j^{n_i} \psi_{ij} x_{ij}$, where $\psi_{ij} \geq 0$ and $\sum_j \psi_{ij} = 1$. Note that these two restrictions enforce p_i to fall within the convex hull of the points in X_i. The authors assume a linear regression $\hat{y}\left(p_i\right) = \Phi^T p_i$, where Φ is the vector of regression coefficients and p_i is a column vector. An optimization problem is defined according to the least squares objective dependent on $P = \{p_1, \ldots, p_m\}$, Y, Φ and the set of weight vectors $\Psi = \{\psi_1, \ldots, \psi_m\}$, $\psi_i = [\psi_{i1}, \ldots, \psi_{im}]^T$. The L_2 loss is used with regularization terms ε_1 and ε_2 for each ψ_i and Φ, respectively, yielding

$$\underset{\psi_i, \ldots, \psi_m, \Phi}{\arg\min} \sum_{i=1}^{m} \left[\left(y_i - \Phi^T X_i \psi_i \right)^2 + \varepsilon_1 \left\| \psi_i \right\|^2 \right] + \left\| \Phi \right\|^2$$

$$s.t.\ (\forall i, j)(\psi_{ij} \geq 0);\ (\forall i)(\sum_{j=1}^{n_i} \psi_{ij} = 1),$$

where the factor $X_i \psi_i$ represents the aggregation of the instances in the ith bag to one exemplar. This is a non-convex and difficult to optimize objective, because the minimization is with respect to both Φ and $\{\psi_i\}$ simultaneously. Wagstaff and Lane [16] propose an alternating projections solver, that alternates between two projection steps. First, the Φ values are fixed and each ψ_i is solved, which can be seen as a projection of Φ on the ψ_i space. Next, the ψ_i vectors are fixed and projected back onto the Φ space. The two steps are alternated until convergence.

Predicting Instance Weights

With the Wagstaff and Lane [16] optimization method we can obtain the coefficients Φ of the regression hyperplane as well as the instance weight vectors ψ_i for each

training bag. Using the regression model Φ, we could compute the output of a new bag X_z as

$$\hat{y}(X_z) = \Phi^T X_z \psi_z,$$

where ψ_z is the weight vector representing the contribution of each instance in X_z to the label of X_z. Unfortunately, the method of Wagstaff and Lane [16] is unable to produce predictions. In their model, ψ_z is unknown, because the method can only find ψ_i for training bags. Pappas and Popescu-Belis [9] present a simple solution to this problem. They formulate another regression problem to predict instance weights of unlabeled bags as $\hat{\psi}_z = \Omega^T X_z$ where Ω are the coefficients of a linear regression model optimized over the weight vectors ψ_i of training bags. Assuming an ℓ_2-norm for the regularization with an ε_3 term, the optimization objective is

$$\arg\min_{\Omega} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\psi_{ij} - \Omega^T x_{ij}\right)^2 + \varepsilon_3 \|\Omega\|^2 .$$

Well-known least squares solving techniques can be used to solve this minimization task. The method allows to determine instance weights of an unlabeled bag and to predict its label. It was successfully applied to a sentiment analysis application [9]. In this case, a text is a bag of sentences and each sentence is modeled as a word vector. The desired prediction is a real-valued rating of the overall sentiment of the text with respect to a specific aspect.

6.4 Bag-Based Regression Methods

In bag-based regression methods the main learning process occurs at bag level. Like bag-based classifiers, regression methods that fall in this category can be further divided into two groups:

- **Bag-based regression methods that work in the original bag space**: these methods rely on a metric function defined over bags, which is used in a distance-based regression algorithm, e.g., a nearest neighbor algorithm. We refer to these methods as *original bag space regression methods* (original-BS methods, for short) and discuss them in more depth in Sect. 6.4.1.
- **Bag-based regression methods that work in a mapped space**: these methods transform the multi-instance data into a single-instance representation and train a single-instance regression algorithm in this transformed space. The same transformation is applied to an unseen bag and its outcome is predicted by the single-instance regression model learned in the mapped space. We refer to these methods as *mapped bag space regression methods* (mapped-BS methods, for short). They are discussed in Sect. 6.4.2.

6.4.1 Original Bag Space Methods

Any single-instance distance-based regression method can be upgraded to a MIR method by using an appropriate MIL metric. By virtue of this measure, the MIR model can be learned and used to make predictions in the original bag space.

The most popular algorithm in this category is RCitationKNN [1]. Its name makes reference to CitationKNN [18] for regression. The minimal Hausdorff distance (Sect. 3.5) is plugged into the CitationKNN algorithm. To obtain the prediction of a new bag, the closest neighbors (citers and references) are consulted and their outcomes averaged. The traditional KNN is also upgraded to MIL in [1].

As an alternative, a support vector regression model [14] can be transferred to the MIL setting by replacing its instance-level kernel by a bag-level kernel, for example the MI kernel [7].

6.4.2 Mapped Bag Space Methods

Mapping methods (Chap. 5) allow the transformation of bags into single-instance vectors. Once the MIL data has been mapped to a single-instance representation, any traditional classifier can be learned on the data. The same mapping methods can be used for regression and traditional regression models can be learned over the single-instance representations afterward. For example, in [21], the BARTMIP mapping is applied to MIR benchmark problems. Another illustrative example is the bioinformatic application described in [5] where the MILES mapping [3] is first applied and a support vector regression (SVR) model [14] is subsequently trained on the mapped data. We refer the interested reader to Sect. 5.3. Below, we discuss two mapping methods that are of particular interest, because they have been used as benchmarks in several studies on MIR [9, 19, 20].

6.4.2.1 Aggregate-MIR

The mapping step of the Aggregate-MIR algorithm [20] is similar to the average mapping described in Sect. 5.3.1. Each bag X_i is mapped to a single instance $(\bar{x}_i, y_i)$ where $\bar{x}_i$ is obtained by averaging all its instances, namely

$$\bar{x}_i = mean\left(\left\{x_{ij}, j = 1 \ldots n_i\right\}\right).$$

A single-instance set D is obtained after mapping all training bags. A traditional regression model is trained on it. To predict the label of a new bag X_i, we apply the mapping method to X_i and obtain the corresponding $\bar{x}_i$. The bag label is predicted as

$$\hat{y}\left(X_i\right) = f\left(\bar{x}_i\right).$$

The method can be more appropriate when the dataset has a low noise level and large bags. This simple method may be useful to initialize more advanced MIR methods.

6.4.2.2 Cluster-MIR

The Cluster-MIR algorithm [17] was set up for MIR problems with a structured instance space. The assumption is that instances in each bag are drawn from different underlying data distributions and that only one distribution is responsible for the bag label. The method can be considered as a generalization of Aggregate-MIR and is related to the stratified bag statistic mapping described in Sect. 5.3.1. The first step is to determine the space structure by the use of unsupervised learning. A soft clustering algorithm is applied over all instances of all training bags to identify k clusters. Training bags are mapped with respect to each cluster. A bag X is mapped with respect to a cluster θ as $\mathscr{M}(X, \theta) \mapsto \langle a_1, \ldots, a_d \rangle$, where $a_j = \sum_{i=1}^{m} r_{\theta i} x_{ij}$, x_{ij} is the value of the jth attribute in the ith instance of the bag X and $r_{\theta i}$ is the relevance of x_i with respect to the jth group. A total number k of single-instance datasets are obtained from the mapping of training bags with respect to each cluster. A regression model is constructed in each mapped dataset. The best regression model is selected based of the training data. At prediction time, a new bag is mapped with respect to the cluster corresponding to the selected regression model. Note that for $k = 1$, Cluster-MIR reduces to Aggregate-MIR.

6.5 Summarizing Comments

MIR is an important task within the MIL paradigm. Although the number of studies on MIR methods is small compared to the abundant literature on multi-instance classifiers, an increased interest on MIR is apparent in recent years. Current MIR methods can be categorized into two groups, instance-based methods and bag-based methods, much like the categories of classification methods. In a few cases, a traditional regression method is modified to the MIL setting. Mostly, the data is manipulated (e.g., an instance is selected, many instances are aggregated to a single instance per bag or bags are mapped to single vectors), such that traditional regression methods can be applied without modifications.

References

1. Amar, R., Dooly, D., Goldman, S., Zhang, Q.: Multiple-instance learning of real-valued data. In: Brodley, C., Danyluk, A. (eds.) Proceedings of the 18th International Confernce on Machine Learning (ICML 2001), pp. 3–10. Morgan Kaufmann Publishers, San Francisco (2001)
2. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Becker, S., Thrun, S., Obermayer, K. (eds.) Advances in Neural Information Processing Systems, pp. 561–568. MIT press, Cambridge (2002)

3. Chen, Y., Bi, J., Wang, J.: MILES: Multiple-instance learning via embedded instance selection. IEEE Trans. Pattern Anal. Mach. Intell. **28**, 1931–1947 (2006)
4. Cheung, P., Kwok, J.: A regularization framework for multiple-instance learning. In: Cohen, W., Moore, A. (eds.) Proceedings of the 23rd International Conference on Machine learning (ICML 2006), pp. 193–200. ACM, New York (2006)
5. EL-Manzalawy, Y., Dobbs, D., Honavar, V.: Predicting MHC-II Binding Affinity Using Multiple Instance Regression. IEEE/ACM Trans. Comput. Biol. Bioinf. **8**(4), 1067–1079 (2011)
6. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001)
7. Gärtner, T., Flach, P.A., Kowalczyk, A., Smola, A.: Multi-instance kernels. In: Sammut, C., Hoffmann, A. (eds.) Proceedings of the 19th International Conference on Machine Learning (ICML 2002), pp. 179–186. Morgan Kaufmann Publishers, San Francisco (2002)
8. Lu, L., Bi, J., Wolf, M., Salganicoff, M.: Effective 3D object detection and regression using probabilistic segmentation features in CT images. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011), pp. 1049–1056. IEEE, Los Alamitos (2011)
9. Pappas, N., Popescu-Belis, A.: Explaining the stars: weighted multiple-instance learning for aspect-based sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 455–466. The Associations for Computational Linguistics, Stroudsburg (2014)
10. Ray, S.: Learning from data with complex interactions and ambiguous labels. PhD Thesis, University of Wisconsin at Madison, United States of America (2005)
11. Ray, S., Craven, M.: Supervised versus multiple instance learning: an empirical comparison. In: De Raedt, L., Wrobel, S. (eds.) Proceedings of the 22nd International Conference on Machine Learning (ICML 2005), pp. 697–704. ACM, New York (2005)
12. Ray, S., Page, D.: Multiple instance regression. In: Brodley, C., Danyluk, A. (eds.) Proceedings of the 18th International Confernce on Machine Learning (ICML 2001), pp. 425–432. Morgan Kaufmann Publishers, San Francisco (2001)
13. Schölkopf, B., Smola, A.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT press, Cambridge (2002)
14. Shevade, S., Keerthi, S., Bhattacharyya, C., Murthy, K.: Improvements to the SMO algorithm for SVM regression. IEEE Trans. Neural Netw. **11**(5), 1188–1193 (2000)
15. Teramoto, R., Kashima, H.: Prediction of protein-ligand binding affinities using multiple instance learning. J Mol. Gr. Model. **29**(3), 492–497 (2010)
16. Wagstaff, K., Lane, T.: Salience assignment for multiple-instance regression. In: Proceedings of the ICML 2007 Workshop on Constrained Optimization and Structured Output Spaces, Citeseer (2007)
17. Wagstaff, K.L., Lane, T., Roper, A.: Multiple-instance regression with structured data. In: Bonchi, F., Berendt, B., Giannotti, F., Gunopulos, D., Turini, F., Zaniolo, C., Ramakrishnan, N., Wu, X. (eds.) Proceedings of the 2008 IEEE International Conference on Data Mining Workshops (ICDMW 08), pp. 291–300. IEEE, Los Alamitos (2008)
18. Wang, J., Zucker, J.: Solving the Multiple-Instance Problem: a Lazy Learning Approach. In: Langley, P. (ed.) Proceedings of the 17th International Conference on Machine Learning (ICML 2000), pp. 1119–1126. Morgan Kaufmann Publishers, San Francisco (2000)
19. Wang, Z., Lan, L., Vucetic, S.: Mixture model for multiple instance regression and applications in remote sensing. IEEE Trans. Geosci. Remote Sens. **50**(6), 2226–2237 (2012)
20. Wang, Z., Radosavljevic, V., Han, B., Obradovic, Z., Vucetic, S.: Aerosol optical depth prediction from satellite observations by multiple instance regression. In: Apte, C., Park, H., Wang, K., Zaki, M. (eds.) Proceedings of the 2008 SIAM International Conference on Data Mining, pp. 165–176. SIAM, Philadelphia (2008)
21. Zhang, M., Zhou, Z.: Multi-instance clustering with applications to multi-instance prediction. Appl. Intell. **31**(1), 47–68 (2009)