Multiple Instance Ranking

Charles Bergeron CHBERGERON@GMAIL.COM
Mathematical Sciences Department, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180 USA

Jed Zaretzki ZARETJ@RPI.EDU
Curt Breneman BRENEC@RPI.EDU
Chemistry Department, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180 USA

Kristin P. Bennett BENNEK@RPI.EDU
Mathematical Sciences Department, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180 USA

Abstract

This paper introduces a novel machine learning model called multiple instance ranking (MIRank) that enables ranking to be performed in a multiple instance learning setting. The motivation for MIRank stems from the hydrogen abstraction problem in computational chemistry, that of predicting the group of hydrogen atoms from which a hydrogen is abstracted (removed) during metabolism. The model predicts the preferred hydrogen group within a molecule by ranking the groups, with the ambiguity of not knowing which hydrogen atom within the preferred group is actually abstracted. This paper formulates MIRank in its general context and proposes an algorithm for solving MIRank problems using successive linear programming. The method outperforms multiple instance classification models on several real and synthetic datasets.

1. Introduction

This paper introduces a new machine learning paradigm called multiple instance ranking (MIRank), bringing the concept of ranking to the framework of multiple instance learning. Some problems that MIRank could potentially solve based on prior data are:

1. For a given country, predict the city that contains the most profitable store.
2. For a given state, predict the congressional district that contains the politician that delivers the most subsidies.
3. For a given document, predict the paragraph/passage that contains the most pertinent sentence/phrase/word.
4. For a given molecular class, predict the molecule with the conformation having the highest human immunodeficiency virus (HIV) inhibition efficacy.
5. For a given state, predict the division that contains the town with the highest median housing unit price.
6. For a given molecule, predict the site of metabolism from which a hydrogen atom is abstracted (removed).

It is this last application, that of hydrogen abstraction from the field of computational chemistry, that motivated this work. The fifth application, which involves making predictions from the census, is also explored here. Later in this paper, a general formulation for multiple instance ranking is provided, an algorithm for MIRank is proposed, and this algorithm is tested on datasets that stem from both applications as well as synthetic data.

As introduced by Dietterich et al. (1997), the setup for multiple instance learning differs somewhat from the standard learning framework. In standard classification, the task is to predict the class of each item. Each item has a corresponding binary classification label, and features defined for each item are used to build the model. In multiple instance classification (MIC), each item belongs to a bag. The task is to predict the class of each bag of items. Features are defined for

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

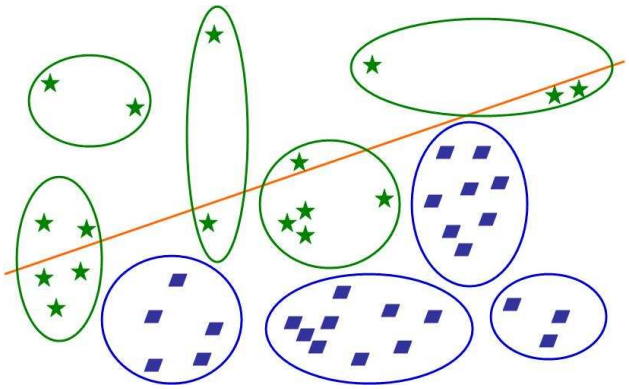

Figure 1. Schematic of multiple instance classification. Bags are ellipses, active bags contain stars and inactive bags contain parallelograms.

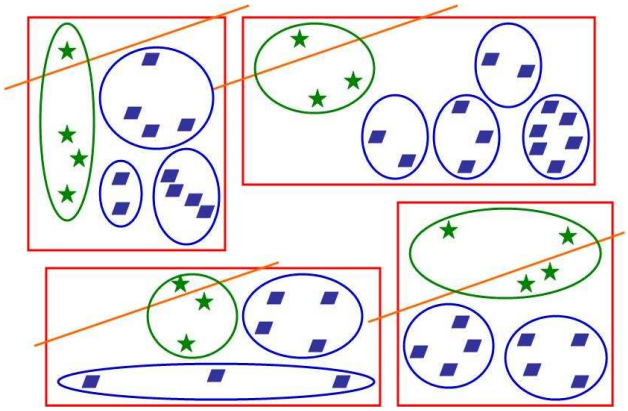

Figure 2. Schematic of multiple instance ranking. Boxes are rectangles, bags are ellipses, preferred bags contain stars, and other bags contain parallelograms.

each item, but the class label is assigned to each bag. For simplicity of presentation, assume there are two classes: active and inactive. By definition, an active bag must contain at least one active item, while an inactive bag contains exclusively inactive items. It is not known which item is active.

Figure 1 illustrates MIC, in which bags are ellipses, items in active bags are represented as stars, and items in inactive bags are marked as parallelograms. The straight line is the separating line representing the classification function. Notice that at least one item from each active bag is found above the line, while all items in inactive bags are located below the line.

The difficulty is that there exists an ambiguity as to which items in an active bag are actually active. For example, consider the drug discovery application (Dietterich et al., 1997), with molecules as bags and conformations (three-dimensional molecular shapes that differ from each other by the rotation of atom groups about one or more bonds) as items. If a molecule possesses one—or possibly several—conformations that are active, then it is known that the molecule is active. However, it is not known which conformation is active. On the other hand, if none of a molecule's conformations are active, then the molecule is deemed inactive, and in this case, it is inferred that all of that molecule's conformations are inactive.

Other applications of MIC include automatic image annotation (Andrews et al., 2003), context-based image indexing (Maron & Ratan, 1998), text categorization (Andrews et al., 2003) and hard-drive failure prediction (Murray et al., 2005). Algorithms for MIC stem from diverse density (Maron & Ratan, 1998; Zhang & Goldman, 2001), neural networks (Ramon & Raedt, 2000), and generalisations of support vector machines (Andrews et al., 2003; Mangasarian & Wild, 2008). The drug discovery application later inspired Ray & Davis (2001) to formulate multiple instance regression, where this time the response assigned to each bag is a real number quantifying the activity of the molecules.

Multiple instance ranking differs in that a classification label is not known for each bag. Rather, some preference information is available for pairs of bags. For example, it may be known that bag A ranks higher than both bags B and C, while the relative ranking of bags B and C may not be known. In many applications, even more structure exists. In these cases, it is convenient to think of every bag as belonging to a box. Within each box, exactly one bag ranks higher than the other ones in the box, and this bag is designated the preferred bag. It is not known how the other bags in the box rank with respect to each other. Further, it is not known how bags rank with respect to each other across boxes. Additionally, there remains the ambiguity of which items in the preferred bags are preferred and which ones are not preferred. Figure 2 illustrates the situation. Large rectangles represent boxes. As was the case in Figure 1, bags are ellipses, items in preferred bags are represented as stars and items in the other bags are marked as parallelograms. Instead of being fixed, the separating line (representing the ranking function) slides from one box to the next. For each box, the ranking function separates at least one item of the preferred bag from the remaining items of the box.

The hydrogen abstraction application fits perfectly into this framework. For each molecule (box), the task

is to find the group (bag) from which a hydrogen atom (item) is abstracted. It is not known which hydrogen atom is abstracted, only to which group it belongs.

The organization of this paper is as follows. Section 2 defines some mathematical notation. Section 3 motivates multiple instance ranking through the computational chemistry problem of hydrogen abstraction. Multiple instance ranking is formulated, and an algorithm for MIRank is proposed, in Section 4. The model and algorithm are generalized to nonlinear MIRank problems in Section 5. Section 6 describes the datasets used in this paper, and Section 7 specifies the modeling results. Finally, Sections 8 and 9 constitute a discussion and outlook, respectively.

2. Notation

Let $\boldsymbol{x}$ denote a vector in R^n and let $\boldsymbol{x}^T$ mark the transpose of $\boldsymbol{x}$. Let $\mathbf{0}$ denote the vector of all zeros and $\boldsymbol{e}$ denote the vector of all ones. Let $|\boldsymbol{x}|$ denote the cardinality of $\boldsymbol{x}$, that is, the number of entries in the vector. Let $\|\boldsymbol{x}\|_1$ denote the 1-norm of $\boldsymbol{x}$, equal to the sum of the absolute values of the entries of the vector. If $\boldsymbol{x}$ has nonnegative entries, then this equals $\boldsymbol{e}^T\boldsymbol{x}$. Let $X \in R^{k\times n}$ and $H \in R^{m\times n}$ denote matrices. I and J indicate index sets. The cardinality of the set I is indicated by $|I|$. The matrix X_I indicates the matrix in $R^{|I|\times n}$ with rows restricted to the index set I. A kernel matrix $K(X, H')$ maps $R^{k\times n}$ and $R^{n\times m}$ into $R^{k\times m}$. Each entry of the mapping results from a function (such as the radial basis function) applied to one row of X applied to one row of H.

3. Motivating application

Bioavailability of a drug, or its ability to be administered orally, is a major concern to the pharmaceutical industry. This characteristic depends on a drug's capability to withstand degradation by intestinal and hepatic enzymes during first-pass metabolism in order to cross the intestinal lining and make it into the bloodstream so that its medicinal effect may be felt (Thummel et al., 1997). Hence, this process of drug metabolism needs to be better understood. More specifically, it is important to discover the attributes of molecules that identify sites which are vulnerable to enzymatic degradation.

Cytochrome CYP3A4 is but one of many metabolising enzymes found in the human liver and small intestine, yet this enzyme metabolises nearly 50% of marketed drugs (Guengerich, 1999; Rendic, 1997). For CYP3A4 substrates, approximately half of the known metabolism reactions occur via hydroxylation, the rate limiting step of which is hydrogen atom abstraction (Sheridan et al., 2007). Knowing where a molecule is preferentially oxidized by this cytochrome would aid the modification of compounds to improve their kinetic or pharmacological profiles (Afzelius et al., 2007).

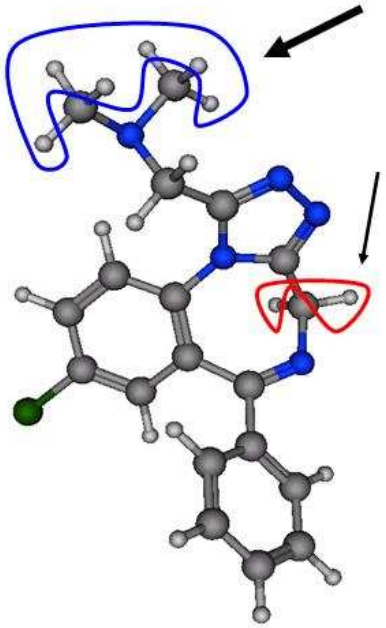

Figure 3. Stick model of an Adinazolam molecule. Large spheres represent nonhydrogen atoms while small spheres represent hydrogen atoms. Two groups of hydrogens are evidenced. The top group, indicated by a thick arrow, has a hydrogen abstracted during metabolism. The lower group, indicated by a thin arrow, does not.

Normally, experimental techniques are used to identify the molecular sites susceptible to metabolism. This is a time- and labor-intensive process. While *in vitro* studies are increasingly high throughput, the *in silico* identification of metabolic liability early on in the drug discovery process will help prevent taking forward poor drug candidates. In addition, the constraints of the biological problem fit perfectly into the framework of a MIRank application, leading to a potential *in silico* solution.

The goal is to build a model that predicts, for each molecule, the site of abstraction of a hydrogen atom during metabolism. In order to accomplish this, individual hydrogen atoms are first grouped together according to molecular equivalence: hydrogens are placed within the same group if and only if the abstraction of any hydrogen from within the group would result in the same metabolised molecule. In this way, groups are equivalent representations of potential sites

of metabolism. Note that experimental data do not show which individual hydrogen is abstracted during metabolism, but rather to which group this hydrogen atom belongs. This setup perfectly fits that of multiple instance ranking. Molecules can be viewed as boxes, groups as bags, and individual hydrogens as items. Figure 3 illustrates these using a stick representation of a molecule.

Two prior modeling attempts are described. Firstly, Singh et al. (2003) chose the hydrogen atom that has the minimum estimated abstraction energy, with a sufficiently large surface area (of 8 squared Angstroms), as the abstracted hydrogen. Allowing 1 guess per molecule, their rule-based model performed correctly in 44% of molecules. Sheridan et al. (2007) later reported that this model has a prediction rate of 51%, allowing for 2 guesses per molecule. Secondly, Sheridan et al. (2007) assembled a database of 316 molecules (including the 50 molecules used by Singh et al. (2003)). They used a random forest applied to molecular descriptors, and found a model that correctly predicted the site of abstraction for 77% of molecules, allowing for 2 guesses per molecule.

4. Formulation

Let (I, J) denote an ordered pair of bags where I and J are lists of indices referring to their items. Let $\boldsymbol{x}_i$ denote a vector of n features for an item i, and let matrix X_I's rows contain the features for each index in I. Further let f denote the ranking function. Then the statement *bag I is preferred over bag J* is expressed mathematically as

$$\max_{i \in I} f(\boldsymbol{x}_i) > \max_{j \in J} f(\boldsymbol{x}_j).$$

The maximum operator on the right hand side can be replaced with all of the items it operates over, hence the inequality is rewritten as

$$\max_{i \in I} f(\boldsymbol{x}_i) > f(\boldsymbol{x}_j) \ \ \forall \ \ j \in J.$$

The maximum operator on the left hand side is also replaced. A convex combination of the items in bag I is taken, following the lead of Mangasarian & Wild (2008) in their formulation of MIC. This convex combination is achieved through vector $\boldsymbol{v_{I,J}}$ whose cardinality is that of I. In a slight abuse of notation, $\boldsymbol{v_{I,J}}$ means the vector corresponding to the pair of bags (I, J). This vector is nonnegative $\boldsymbol{v_{I,J}} \geq 0$, and its entries sum to one: $\boldsymbol{e}^T \boldsymbol{v_{I,J}} = 1$. This vector multiplies matrix X_I:

$$f(X_I^T \boldsymbol{v_{I,J}}) > f(\boldsymbol{x}_j) \ \ \forall \ \ j \in J.$$

Let the model be linear defined by vector $\boldsymbol{w}$, i.e.

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{w}. \tag{1}$$

In this case, we have

$$\boldsymbol{v_{I,J}}^T X_I \boldsymbol{w} > \boldsymbol{x}_j^T \boldsymbol{w}.$$

This paper focuses on linear models, because chemists are interested model interpretation. However, this formulation is readily kernelized, as discussed in Section 5.

Now introduce an empirical risk scalar $\xi_{I,j}$ based on the hinge-loss, allowing for errors in training the model:

$$\boldsymbol{v_{I,J}}^T X_I \boldsymbol{w} - \boldsymbol{x}_j^T \boldsymbol{w} \geq 1 - \xi_{I,j}.$$

This inequality resembles the main constraint in Joachims' ranking support vector machine (2002). It is also the key constraint in an optimization problem whose objective function is to minimize

$$\nu \mathcal{L}_{emp}(\boldsymbol{\xi}) + \mathcal{L}_{reg}(\boldsymbol{w})$$

where $\nu > 0$ is the tradeoff parameter and $\mathcal{L}_{emp}$ and $\mathcal{L}_{reg}$ are arbitrary loss functions.

Choosing the 1-norm for both loss functions makes the objective linear in the variables, a choice that was also made by Mangasarian & Wild (2008). Furthermore, using the 1-norm on $\boldsymbol{w}$ makes for sparse models, facilitating the interpretability of linear models. Thereupon, the 1-norm MIRank optimization problem is

$$\min_{\boldsymbol{\xi}, \boldsymbol{w}, \boldsymbol{v_{I,J}}} \quad \nu \boldsymbol{e}^T \boldsymbol{\xi} + \|\boldsymbol{w}\|_1 \tag{2}$$

subject to

$$\boldsymbol{v_{I,J}}^T X_I \boldsymbol{w} - \boldsymbol{x}_j^T \boldsymbol{w} \geq 1 - \xi_{I,j} \ \ \forall \ \ (I, J, j) \tag{3}$$

$$\boldsymbol{e}^T \boldsymbol{v_{I,J}} = 1 \ \ \forall \ \ (I, J) \tag{4}$$

$$\boldsymbol{v_{I,J}} \geq \boldsymbol{0} \ \ \forall \ \ (I, J) \tag{5}$$

$$\boldsymbol{\xi} \geq \boldsymbol{0}. \tag{6}$$

The entries of empirical risk vector $\boldsymbol{\xi}$ are $\xi_{I,j}$ as they appear in the first constraint. This notation signifies that, for each pair (I, J), there is an empirical risk contribution from each item $j \in J$. These are nonnegative quantities, as per 6. Note that there are as many vectors $\boldsymbol{v_{I,J}}$ as there are pairs (I, J). These vectors are forced to be nonnegative and to sum to one in constraints 4 and 5.

Since the first constraint is linear and the remaining terms are linear, this is a bilinear optimization problem. We use the successive linear programming algorithm given in Algorithm 1 to find a locally optimal

Algorithm 1 Multiple instance ranking algorithm

Select tolerance τ and tradeoff parameter ν.
Initialise $\boldsymbol{u_{I,J}} = \frac{\boldsymbol{e}}{|I|} \ \forall \ (I, J)$.
repeat
  Set $\boldsymbol{v_{I,J}} = \boldsymbol{u_{I,J}} \ \forall \ (I, J)$.
  Fix the $\boldsymbol{v_{I,J}}$'s and solve the linear program 2-6 for $\boldsymbol{\xi}$ and $\boldsymbol{w}$.
  Fix $\boldsymbol{w}$ and solve the linear program 2-6 for $\boldsymbol{\xi}$ and the $\boldsymbol{u_{I,J}}$'s.
until $\|\boldsymbol{v_I} - \boldsymbol{u_I}\|_1 \leq \tau \ \forall \ (I, J)$

solution of the bilinear problem. This proposed MI-Rank algorithm belongs to a family of algorithms that has proven to find good local solutions on a variety of bilinear machine learning problems. The subproblem solutions are not necessarily unique, but this has no impact on algorithm convergence.

The convergence proof for the MIC algorithm in Mangasarian & Wild (2008) can be readily adapted to Algorithm 1. Specifically, the algorithm converges because the sequence of objective function values

$$\{\nu \boldsymbol{e}^T \boldsymbol{\xi} + \|\boldsymbol{w}\|_1\}$$

at each iteration is nonincreasing and bounded below by zero, and every accumulation point satisfies a local minima property. The formal proof is omitted for brevity; see Mangasarian & Wild (2008).

Algorithm 1, as well the Mangasarian & Wild (2008) algorithm for MIC, were implemented in Matlab using the linear programming solver MOSEK (`www.mosek.com`).

5. Nonlinear Formulation

A nonlinear MIRank function can be generated by kernel transformations (Shawe-Taylor & Cristianini, 2004). We adopt the notation and direct kernel approach used for MIC in Mangasarian & Wild (2008). The linear ranking function 1 is replaced by the nonlinear function:

$$f(\boldsymbol{x}) = K(\boldsymbol{x}^T, H^T)\boldsymbol{\alpha} \tag{7}$$

where $\boldsymbol{x} \in R^n$ is an item, $\boldsymbol{\alpha} \in R^m$ are the dual variables and the matrix $H \in R^{n \times m}$ has as its rows all of the m items found collectively in all of the bags and boxes, and $K(\boldsymbol{x}^T, H^T)$ is an arbitrary kernel map. The bilinear program generating the nonlinear MI-Rank function becomes:

$$\min_{\boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{v_{I,J}}} \quad \nu \boldsymbol{e}^T \boldsymbol{\xi} + \|\boldsymbol{w}\|_1 \tag{8}$$

subject to

$$\boldsymbol{v_{I,J}}^T K(X_I, H^T)\boldsymbol{w} - K(\boldsymbol{x}_j^T, H^T)\boldsymbol{\alpha} \geq 1 - \xi_{I,j} \ \forall \ (I, J, j) \tag{9}$$

$$\boldsymbol{e}^T \boldsymbol{v_{I,J}} = 1 \ \forall \ (I, J) \tag{10}$$

$$\boldsymbol{v_{I,J}} \geq \boldsymbol{0} \ \forall \ (I, J) \tag{11}$$

$$\boldsymbol{\xi} \geq \boldsymbol{0}. \tag{12}$$

The kernel formulation remains a bilinear program and thus can be solved using Algorithm 1 by substituting $\boldsymbol{\alpha}$ for $\boldsymbol{w}$ and bilinear program 8-12 for bilinear program 2-6.

6. Datasets

In addition to the hydrogen abstraction dataset, several additional datasets are used in modeling experiments. All three are described here.

6.1. CYP3A4 substrate dataset

The CYP3A4 substrate dataset is made up of 227 small drug-like compounds. A series of 36 descriptors for each hydrogen atom for all molecules are calculated:

- the charge of the hydrogen;
- the surface area of the hydrogen;
- the non hydrogen surface area of the base atom the hydrogen is attached to;
- the hydrophobic moment: the hydrogen's location with regards to the hydrophobic or hydrophilic end of the molecule;
- the span: a measure of whether the candidate hydrogen is located at the end or within the middle of the molecule.
- the topological neighborhood: the distributions of atom types within a various topological distances from the hydrogen.

Recall that, for each molecule, the goal is to predict from which group a hydrogen atom is abstracted, and it is not known which hydrogen from the abstracted site is removed.

These 227 molecules form are a subset of the 305 non-proprietary molecules used by Sheridan et al. (2007), and represent all those for which descriptor generation could be completed.

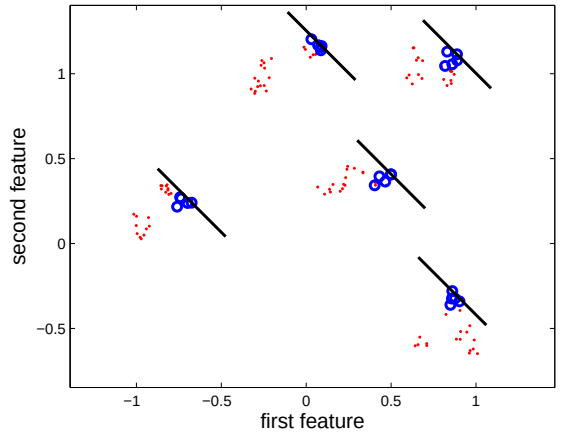

Figure 4. Synthetic dataset visualisation. Preferred bags contain circles and other bags contain dots. Sliding line represents the ranking function found by MIRank that separates at least one circle from remaining items in each box.

6.2. Synthetic datasets

This dataset consists of 227 boxes, five bags per box and five items per bag. There are two features. Each feature is calculated as follows:

$$\mu_i^{box} + \mu_j^{bag} + \mu_k^{item}$$

with μ_i^{box} drawn from the uniform distribution $\mathcal{U}(-1, 1)$, μ_j^{bag} drawn from the distribution $\mathcal{U}(-A, A)$ and μ_k^{item} drawn from the distribution $\mathcal{U}(-B, B)$. Put in words, the center of each box is chosen from a uniform distribution, and the center of each bag with respect to its box is chosen from a different uniform distribution, and each item with respect to its bag is chosen from yet another uniform distribution. Parameters A and B characterize these synthetic datasets. For each item, the response is the sum of the features. The goal is, for each box, to find the bag containing the item of greatest response. Five boxes of this dataset are portrayed as Figure 4. It illustrates the difficulty in constructing a linear function separating at least one circle from each box from the remaining circles and dots, as MIC attempts to do. On the other hand, it it possible to find a ranking function (the sliding line) that does this for each box, as MIRank does.

Different values for dataset parameters A and B were attempted:

- Synthetic-1 set $A = B = 0.01$.
- Synthetic-2 set $A = 0.1$ and $B = 0.01$.
- Synthetic-3 set $A = 0.01$ and $B = 0.1$.
- Synthetic-4 set $A = B = 0.1$.

Table 1. Prediction accuracies

Dataset	MIC	MIRank
CYP3A4 substrate	$67.1\% \pm 7.1$	$70.9\% \pm 6.9$
Synthetic-1	$90.8\% \pm 8.6$	$99.8\% \pm 0.53$
Synthetic-2	$96.8\% \pm 4.6$	$99.1\% \pm 1.8$
Synthetic-3	$95.5\% \pm 8.3$	$99.9\% \pm 0.38$
Synthetic-4	$95.7\% \pm 5.2$	$99.7\% \pm 0.91$
Census-16h	$52.8\% \pm 17.4$	$60.3\% \pm 15.1$
Census-16l	$46.2\% \pm 17.7$	$57.5\% \pm 16.0$

6.3. Census datasets

The two census datasets (census-16h and census-16l) belong to the *Data for Evaluating Learning in Valid Experiments* (DELVE, `http://www.cs.toronto.edu/~delve/`) repository. It consists of 22784 towns spread amongst the 50 states of the United States of America. This study only considered the 3054 towns of more than 10000 inhabitants. Each town is assigned a 5-digit Federal Information Processing Standard (FIPS) place code (that is not a zip code). Typically, this dataset is used in a regression setting to model the response—which is the town's median housing unit price. The census-16h and census-16l datasets differ in their features: each consists of 16 features drawn from the 1990 census.

These datasets are fitted into the multiple instance ranking framework as follows. States are boxes, divisions of towns are bags and towns are items. For each state, towns whose place code begin with the same number are assigned to the same division. As no place code commences with the number 9, there are up to 9 divisions per state. The task is to predict, for each state, the division that contains the town with the highest median housing unit price.

7. Results

For each dataset, results were obtained using both the MIC and MIRank algorithms. For MIC, preferred bags were treated as active bags and other bags were treated as inactive bags. All results reported are for linear functions.

The experimental design is as follows. Each dataset was randomly split into training, validation and testing subsets consisting of 60%, 20% and 20% of the boxes, respectively. The training subset was used to

Table 2. Hypothesis testing

Dataset	P-value
CYP3A4 substrate	$5.59 \cdot 10^{-3}$
Synthetic-1	$1.62 \cdot 10^{-6}$
Synthetic-2	$1.31 \cdot 10^{-2}$
Synthetic-3	$5.84 \cdot 10^{-3}$
Synthetic-4	$1.46 \cdot 10^{-4}$
Census-16h	$4.51 \cdot 10^{-2}$
Census-16l	$3.92 \cdot 10^{-4}$

train both MIC and MIRank models for 19 values of tradeoff parameter ν spread logarithmically over the range $[10^{-3}, 10^{6}]$. The model corresponding to the value of ν that resulted in the best prediction accuracy over the validation set was retained, and a prediction using this model was obtained for the testing subset. This process was repeated 32 times, and the average performance across these 32 testing subsets is reported in Table 1, along with the standard deviation as a measure of spread.

All results in Table 1 are presented as a percentage of boxes for which the preferred bag was accurately predicted, allowing for 2 guesses per box, which is the metric employed by Sheridan et al. (2007). The algorithm tolerance τ defined in Algorithm 1 was set to 10^{-3}.

For all datasets, the hypothesis that MIC and MIRank results are statistically equal is dismissed using paired t-testing at a 5% significance level. The p-values are reported in Table 2.

8. Discussion

The results of Section 7 make a strong case supporting the hypothesis that these problems, when framed in a multiple instance ranking paradigm, are better solved by an algorithm that is designed to solve problems of that paradigm over one that is not. Forcing MIRank problems into a MIC paradigm was not as successful. In other words, the improvement is due to choosing a model that better fits the problem.

The MIRank result for the CYP3A4 substrate dataset reported in this paper compare favourably with existing approaches to hydrogen abstraction. It clearly outperforms the results of Singh et al. (2003). Their results are reproducible and their reported error holds on new molecules. Comparison with Sheridan et al. (2007) is more difficult. Reproduction of their results is challenging since since their descriptors are not public and the details of the learning and model selection methods they used are not entirely clear. Our descriptors attempt to reproduce those of Sheridan et al. (2007), but could not be generated for all molecules. Hence, we regard their results as optimistic.

A future controlled experiment is needed to fully compare the approaches of Sheridan et al. (2007) and those of this paper. This experiment would validate which descriptor set and modeling paradigm is most well suited for this chemistry application. To facilitate future investigations into MIRank and hydrogen abstraction, the datasets and Matlab source codes used in this paper are available from `http://www.rpi.edu/~bennek/MIRank/`.

9. Conclusion

This paper introduced a framework that tackles a novel machine learning question arising from an important chemistry problem. A first working algorithm produces excellent results on it and other problems. We believe that this first paper for MIRank will generate future research into new algorithms and applications. This section explores several possible extensions.

In the chemistry domain, we often restrict ourselves to sparse and linear models because model interpretability is a desired property in the particular application of drug discovery. However, this interpretability analysis is a paper of its own, and does not appear here.

Hydrogen abstraction is an important application of MIRank modeling of great practical value for drug discovery. We are working to expand the efficacy and applicability of the MIRank hydrogen abstraction models in several ways. First, we are increasing the number of molecules in the database of CYP3A4 substrates that can be used to develop and test new MIRank models. Second, we hope to build databases and models for new substrates, such as CYP2D6 and CYP2C9. Third, we are developing novel descriptors that are believed to be indicative of hydrogen abstraction.

We are working to improve the MIRank modeling paradigm and investigating other potential multiple instance ranking problems. Reports here are limited to the linear MIRank models, but as discussed the approach can be readily applied with nonlinear models using kernel functions. Research is needed to investigate how modeling results are affected by changing the loss functions in the empirical risk and/or regularization terms of the optimization problem.

Finally, further improvements to the MIRank algo-

rithm are possible. More scalable and efficient algorithms for finding locally optimal solutions could be developed by exploiting recent developments in large scale support vector machine algorithms. In addition, integer programming or cutting plane algorithms could be used to find global minima of the optimization problem, but at much greater computational cost.

Acknowledgments

Charles Bergeron is under fellowship from the *Fonds québécois de la recherche sur la nature et les technologies*. This work was done under, and all authors belong to, the Rensselaer Exploratory Center for Cheminformatics Research (RECCR, `reccr.chem.rpi.edu`).

References

Afzelius, L., Arnby, C. H., Broo, A., Carlsson, L., Isaksson, C., Jurva, U., Kjellander, B., Kolmodin, K., Nilsson, K., Raubacher, F., & Weidolf, L. (2007). State-of-the-art tools for computational site of metabolism predictions: Comparative analysis, mechanistical insights, and future applications. *Drug Metabolism Reviews*, *39*, 61–86.

Andrews, S., Tsochantaridis, I., & Hofmann, T. (2003). Support vector machines for multiple-instance learning. *Advances in Neural Information Processing Systems 15*.

Dietterich, T. G., Lathrop, R. H., & Lozano-Perez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, *89*, 31–71.

Guengerich, F. P. (1999). Cytochrome p-450 3A4: Regulation and role in drug metabolism. *Annual Review of Pharmacology and Toxicology*, *39*, 1–7.

Joachims, T. (2002). Optimizing search engines using clickthrough data. *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 133–142).

Mangasarian, O. L., & Wild, E. W. (2008). Multiple instance classification via successive linear programming. *Journal of Optimization Theory and Applications*, Accepted.

Maron, O., & Ratan, A. L. (1998). Multiple-instance learning for natural scene classification. *Proceedings of the 15th International Machine Learning Conference*.

Murray, J. F., Hughes, G. F., & Kreutz-Delgado, K. (2005). Machine learning methods for predicting failures in hard drives: A multiple-instance application. *Journal of Machine Learning*, *6*, 783–816.

Ramon, J., & Raedt, L. D. (2000). Multi instance neural networks. *Proceedings of the 17th International Machine Learning Conference*.

Ray, S., & Page, D. (2001). Multiple instance regression. *Proceedings of the 18th International Machine Learning Conference*.

Rendic, S. (1997). Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Metabolism Reviews*, *34*, 83–448.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press.

Sheridan, R. P., Korzekwa, K. R., Torres, R. A., & Walker, M. J. (2007). Empirical regioselectivity models for human cytochromes P450 3A4, 2D6, and 2C9. *Journal of Medicinal Chemistry*, *50*, 3173–3184.

Singh, S. B., Shen, L. Q., Walker, M. J., & Sheridan, R. P. (2003). A model for predicting likely sites of CYP3A4-mediated metabolism on drug-like molecules. *Journal of Medicinal Chemistry*, *46*, 1330–1336.

Thummel, K. E., Kunzea, K. L., & Shen, D. D. (1997). Enzyme-catalyzed processes of first-pass hepatic and intestinal drug extraction. *Advanced Drug Delivery Reviews*, *27*, 99–127.

Zhang, Q., & Goldman, S. A. (2001). EM-DD: An improved multiple-instance learning technique. *Advances in Neural Information Processing Systems 14*.