Algorithms **2010**, *3*, 92-99; doi:10.3390/algor3010092

OPEN ACCESS

algorithms

ISSN 1999-4893

www.mdpi.com/journal/algorithms

Article

Base Oils Biodegradability Prediction with Data Mining Techniques

Sihem Ben Abdelmelek [1], Saloua Saidane [2,*] and Malika Trabelsi [1]

[1] University of Bizerte, Bizerte, Zarzouna 7021, Tunisia; E-Mails: sihembenabdelmelek@yahoo.fr (S.B.A.); malikatrabelsi_ayadi@yahoo.fr (M.T.)

[2] San Diego Mesa College, 7250 Mesa College Drive, Room K202, San Diego, CA 92111, USA

* Author to whom correspondence should be addressed: E-Mail: ssaidane@sdccd.edu; Tel.: +1-619-388-5821

Received: 30 December 2009; in revised form: 24 January 2010 / Accepted: 28 January 2010 / Published: 23 February 2010

Abstract: In this paper, we apply various data mining techniques including continuous numeric and discrete classification prediction models of base oils biodegradability, with emphasis on improving prediction accuracy. The results show that highly biodegradable oils can be better predicted through numeric models. In contrast, classification models did not uncover a similar dichotomy. With the exception of Memory Based Reasoning and Decision Trees, tested classification techniques achieved high classification prediction. However, the technique of Decision Trees helped uncover the most significant predictors. A simple classification rule derived based on this predictor resulted in good classification accuracy. The application of this rule enables efficient classification of base oils into either low or high biodegradability classes with high accuracy. For the latter, a higher precision biodegradability prediction can be obtained using continuous modeling techniques.

Keywords: base oils; biodegradability; classification models; data mining; multiple linear regression; machine learning models; predictive models

---

1. Introduction

The interest in the prediction of biodegradability of base oils (e.g., motor oils and lubricants) using their chemical structure and/or chemical and physical characteristics stems from a threefold motivation. First, the scientific interest makes modeling of the biodegradability process as an effect

caused by their chemical and physical attributes worth pursuing for theoretical purposes. Second, the economic incentives to save time and material costs when biodegradability can be directly predicted are highly valuable. Third, environmental imperatives for designing, producing, and using environmentally friendly base oils are becoming a higher priority. Predicting base oil biodegradability before they are produced, tested and used will make these imperatives easier to meet, and the development of environmentally friendlier oils all the more feasible.

However, this problem has so far eluded the search for an adequately accurate solution. The objective of this paper is to uncover new modeling techniques that would improve base oils biodegradability through testing of a large variety of data mining techniques using the chemical and physical characteristics of 63 base oils' data analyzed by Haus *et al.* [1]. The state of the art of biodegradability modeling is first reviewed in the following section.

2. Literature Review

Data mining techniques are being aggressively adopted by the hard sciences mostly in biology and increasingly in chemistry research [2,3]. Early research focused on quantitative structure-activity relationships (QSARs) models that correlate molecular structure to compound activity [4]. One of the earlier efforts used a survey of expert knowledge to build a "screening-level" model for aerobic biodegradability [5]. Original studies that applied Artificial Intelligence for such expert judgment in the modeling of biodegradability used the Inductive Logic Programming to predict the half-life for aerobic aqueous biodegradation [6,7]. Evaluation and application of models for the prediction of ready biodegradability are reported in the MITI-I test where two methods in which the probability of rapid biodegradation is estimated by Multivariate Linear Regression (MLR) models of biodegradability and counts of fragments of structure and molecular weight. [6]. Another Artificial Intelligence type of modeling biodegradation is an example-based learning system, instead of expert systems [7]. A third effort [8] also applied other Artificial Intelligence models for biodegradability prediction to generate predictive rules using the inductive machine learning approach with structural features as variables, and discretized biodegradability comprising two classes (slow and fast biodegradation). The IUPAC study presented a review and a critical analysis of modeling and estimating the degradability of chemicals in the environment [9]. More recently, various QSARs-based classification techniques to classify different types of base oils were tested [2]. However, available QSAR models have so far proved to be of limited effectiveness since their achieved predictive accuracy varied widely, from 40% to 90% [8].

An alternative to QSAR modeling used chemical and physical characteristics with analytical modeling of biodegradability as a process assumed to be driven by such attributes. Two studies of this type are of particular interest to our research. The first used Artificial Neural Network models to predict biodegradability of base oils using some of their chemical composition and their viscosity [10]. The second used different characteristics of a variety of oils to investigate the impact of chemical and physical characteristics on the biodegradability of 63 base oils using MLR models [1]. The focus of the former research was to uncover the most significant factors having the highest impact on biodegradability through a global predictive (MLR) model. Their research showed that the main characteristics that significantly affect biodegradability are Paraffinic Carbon (PC) content, Kinematic

Viscosity (as Log), and Noak Volatility (as Log). The resulting MLR models showed that the average error is 30%, and can be as high as 88%.

Thus, based on the results of modeling efforts to-date, it can be concluded that the biodegradability problem has not been satisfactorily solved. Due to the highly adverse effects of base oils on the environment and human well-being, higher prediction accuracy is of critical significance. The present research investigates the performance of various data mining techniques in predicting and possibly improving the prediction of base oils biodegradability, using the data and the relevant variables reported [1]. The focus here is not on finding general theoretical explanatory models of biodegradability, but it is rather directed toward the identification of modeling techniques that would improve the accuracy of biodegradability prediction.

3. Modeling Methodology

The data mining techniques used in this research are comprised of methods with numeric continuous biodegradability, and methods with discrete biodegradability. The former type includes Multiple Regression Models using a number K of most similar base oils referred, to here as K-Nearest Neighbors Multiple Linear Regression (KNN-MLR) models, Artificial Neural Networks (ANN), and a continuous type of Decision Trees known as Classification and Regression Trees (CART). The discrete approaches include the unsupervised classification techniques of K-means and Two-step clustering methods, and the supervised classification techniques of Logistic Regression, and Decision Trees.

In order to use the largest data set available for the estimation/training phase, the "Leave-One-Out-Cross-Validation (LOOCV)" approach is used. Each of the oils considered is in turn left out of the training /model building phase, and then its biodegradability is predicted using the resulting model. Except for trained (traditional) MLR, and clustering, LOOCV with a number K of most similar oils is used in model building and prediction. The problem of the optimal number N of nearest neighbors is addressed by estimating biodegradability models for a number M of different values of K and taking the median of the K predictions as the final prediction. The median serves as a shield against "extreme" predictions. With the ANN technique, the M predictions are derived using M different two-hidden-layer networks, where the starting weights are selected randomly in each of the M runs.

The performance and testing of each technique in terms of prediction accuracy is analyzed using the predicted median values for all the oils considered following the methodology above. All the tested techniques are evaluated for their performance at three different partitions of the biodegradability spectrum in an attempt to uncover any significant differences in their predictive accuracy between these biodegradability partitions. In addition to the whole data set, two biodegradability classes are considered -low biodegradable (less than 50%), and high biodegradable (over 50%) for model evaluation. This two-class partitioning is supported by the Two-step automatic cluster detection technique which uncovered only two clusters in the available data. This classification was also adopted by some of the previous research practice under the umbrellas of readily or highly biodegradable class and not readily or slowly biodegradable class [11,4].

Model accuracy is measured by the Mean Absolute Percent Error (MAPE) for the continuous models, and by the percent of misclassified cases for the discrete models. As mentioned earlier, we use the base oils data reported by Haus *et al.* [1] and the properties they found to have the most significant impact on biodegradability, namely PC, Log(KV), and Log(NV). The available data is deemed to be representative of the base oil biodegradability process since it includes a wide range of biodegradability as well as a wide range of KV and NV values. The results and findings of the various modeling techniques using this data set are discussed in the following section.

4. Results and Comparative Analysis

Modeling and testing were carried out using SPSS Clementine commercial software package. Although the MLR model of Haus *et al.* [1] showed a high adjusted R^2 of 95.6%, the training and prediction errors in MAPE terms were not as good. The training (using 38 out of 63 oils) and prediction (using the remaining 25 oils) had average MAPEs of 21% and 30%, and maxima of 42% and 89% respectively. The results of our testing are discussed in the following sections.

Table 1. MAPE for continuous models.

Method	Average			Max		
	Low-B	**High-B**	**All**	**Low-B**	**High-B**	**All**
Trained MLR	30.9	27.6	29.9	88.6	52.7	88.6
LOOCV MLR	21.5	6.1	13.8	70.1	21.2	70.1
LOOCV K-NN MLR	22.1	6.6	14.2	79.2	25.7	79.2
LOOCV ANN	22.9	7.1	16.1	87.2	27.0	87.2
LOOCV Continuous MBR	10.6	37.5	23.8	47.6	146.5	146.5
LOOCV CART	24.8	11.1	17.8	66.6	30.8	66.6
Basu-ANN	26.2	9.2	17.4	133.64	24.7	133.5

4.1.Continuous biodegradability models

The prediction accuracy results obtained by the continuous modeling techniques tested are reported in Table 1, which shows the MAPE obtained with continuous biodegradability models in terms of average and maximum values for the low and high biodegradability sets, and for the whole data set.

For (Trained) MLR with the training set used by Haus *et al.* [9], the average MAPEs for the low, high and whole set are 30.88%, 27.59% and 29.96%, respectively. The maxima are 88.63%, 52.73% and 88.63% respectively. This shows that the Trained MLR model has similar predictive ability for the three sets, with a higher maximum error for the low set (88% *vs.* 52%). LOOCV MLR fared better than trained MLR, with average errors of 21.51%, 6.10% and 13.80%, and maxima of 70.12%, 21.26% and 70.12%, for the three sets respectively. Thus LOOCV MLR has a much higher predictive ability for the high end than trained MLR (6% *vs.* 27%). LOOCV K-Nearest Neighbors MLR performed similarly to LOOCV MLR with averages of 21.51%, 6.60% and 13.8%, and maxima of 79.2%, 25.73% and 79.2% respectively. LOOCV ANN had a little higher accuracy than Trained MLR for the high end, but not better than the previous methods. The averages were 22.98% for the low end, 7.11% high end,

and 16.05% overall, while the maxima were 87.27%, 27.04%, and 87.27%. Continuous LOOCV MBR fared better than all of the previous methods on the low end with an average of 10.63% and a maximum of 47.62%. For the high and overall sets, MBR's average errors were 37.57% and 23.89%, and maxima of 146.57% and 66.67%, respectively. Although Basu *et al.* [2] used ANN with a different set of oils and different attributes, we were curious to compare their results to ours. The prediction accuracy obtained was inferior to all of our tested methods. Their methods had averages of 26.29%, 9, 2%, and 17%, and maxima of 133.64%, 24.76% and 133.64%, respectively. Compared to trained MLR [1], the errors are higher, especially in terms of maxima. Finally, CART had a little easier time predicting the high end, but was inferior to the other methods that favored the high end. CART averages were 24.82%, 11.15%, and 17.88%, with maxima at 66.67%, 30.8% and 66.67% for the low, high, and overall data sets respectively.

4.2. Discretized biodegradability results derived from continuous models

The results from the continuous models reported in Table 2 were converted into binary equivalents using the partitioning of biodegradability values and associated predictions into Low end (Biodegradability < 50%), and High end. The expectation was that better class prediction would be achieved, although at the cost of loss of precision.

Table 2 shows that the first four methods (Trained MLR, MLR, K-NN MLR, and ANN) predicted better the Low end set, while the last three favored the High end. For the low end, K-NN MLR was the best with perfect prediction (0%). ANN came in second place with 3.23% error. The third and fourth were MLR and Trained MLR with errors of 3.32% and 5.88%, respectively. For the whole, K-NN MLR, and ANN are best with 4.76% error for both. Trained MLR and CART were the least accurate overall with errors of 8% and 9.52% respectively. The high and overall averages ranged from 4.76% to 12.25%. MBR did well on the high end, with 3.13% error. MBR's average low and high have maximum errors of 9.68% and 6.35% respectively. Basu *et al.*'s converted results had much inferior predictive accuracies with errors of 26.67%, 11.76% and 18.75% for the Low, High and overall respectively.

Table 2. Percent Misclassification derived from continuous results.

Technique	Low-B	High-B	All
Trained MLR	5.8	12.2	8.0
LOOCV MLR	3.3	9.3	6.3
LOOCV NN-MLR	0.0	9.3	4.7
LOOCV ANN	3.2	6.2	4.7
LOOCV Continuous MBR	9.6	3.1	6.3
LOOCV CART	9.6	9.3	9.5
Basu-ANN	26.6	11.7	18.7

4.3. Discrete biodegradability models

The results from the tested discrete models are presented in Table 3. The methods of Logistics Regression, Logistic-KV (using Log (KV) only), MBR and K-Means had the best accuracy for the low end, with 3.23% error. For the high end, Logistic Regression, Two-Step, and K-means-KV achieved the best result, with a 3.13% error. For the whole set, Logistic Regression, Logistics with KV, and K-Means-KV had the best prediction with a 4.76% error. Overall, mismatch errors ranged from 4.76% to 12.70%. However, it is worth noting that the powerful method of DT did not achieve the highest performance for all the sets with 9.68%, 6.25% and 7.94% errors for the low, high, and overall sets, respectively.

Haus *et al.*'s findings that Log (KV) is the most significant predictor suggested the use of Log (KV) as the sole predictor. With this single variable, DT uncovered an important result given by the classification Rules below:

1. If Log (KV) ≥ 1.9 (KV≥ 96), then Biodegradability is Low (<50%),
2. If (Log (KV) < 1.9, (KV < 96) then Biodegradability is High (≥50%).

This simple Rule derived from C5.0 DT provides an efficient classification approach that achieves the same best performance level obtained by more powerful methods for the high-end and overall data sets with a 3.13% and 4.76% error respectively, while providing the second lowest error of 6.25% for the low-end data set.

Table 3. Percent Misclassification from Binary Prediction.

Technique	Low-B	High-B	All
LOOCV DT C5.0	9.6	6.2	7.9
LOOCV Logistic Reg.	3.2	3.1	4.7
LOOCV MBR	3.2	21.8	12.7
K-Means	3.2	6.2	4.7
Two-Step	16.1	3.1	9.5
Logistics-KV	3.2	6.2	4.7
Rule 1	6.4	3.1	4.7

Furthermore, most modeling techniques with this single variable approach generated prediction classification accuracies similar to models using the three original predictors. K-Means and Logistics Regression provided the best performing classification results using this variable only. The results are reported in Table 3 under Logistics-KV and K-Means-KV, which were the best performers with 3.23%, 3.13%, and 4.76% errors for the low, high and the overall sets respectively.

4.4. Comparative analysis and summary of results

The results of this research show a prevailing difference between continuous and discrete prediction techniques. Most continuous methods tend to be more accurate with the high end of the biodegradability spectrum for which LOOCV MLR is the best performer (6.10% error), closely followed by LOOCV K-NN MLR (6.60% error). The best performer for the low end is MBR

with 10.63% error, while for overall, it is either trained MLR (12.99% error) or LOOCV MLR (13.8%). The latter is to be preferred since it enables the use of the largest data set available.

All these continuous methods resulted in comparable maximum errors around 70%. This is due to the nonlinearity of the biodegradability process that is hard to predict using the linear methods applied. A more plausible alternative for improvement is provided by the K-NN MLR, which can capture a "local" linear model around the case to be predicted, thus capturing most of the biodegradability behavior locally. This state of highly nonlinear biodegradability process is a further rationale for using the binary dependent variable classification techniques, which were expected to be more accurate, although less precise. The best classification results derived from continuous models were achieved by K-NN MLR and ANN for the overall set with a 4.76% error. Both favored the low end with 0% and 3.23% error, respectively. The results from the direct binary methods show that the best method is the three-variable Logistic Regression, closely followed by Logistics-KV which performed better for the low end (3.23%), and K-Means-KV which fared better for the high end (3.13%). The simple classification Rule derived from this single-variable DT model achieved comparable results. All of these one-variable methods achieved the same accuracy for the overall set with 4.76% error. The application of this simple rule enables an effective classification of base oils into the slowly/highly biodegradability classes. This result is significantly superior to the one achieved by QSARs methods that achieved misclassification errors in the 40% to 90% [8].

5. Conclusions

In this research, several data mining techniques including continuous and binary classification techniques were applied to the prediction of base oil biodegradability using three of their most significant predictive characteristics, namely PC, KV and NV. Following suggestions from research practice and clustering results, all the tested techniques were evaluated for their performance at three different partitions of the biodegradability spectrum in an attempt to uncover any significant differences in their predictive accuracy over these biodegradability partitions. Thus, in addition to the whole data set, two biodegradability classes are considered: low biodegradable (less than 50%), and high biodegradable (50% or higher) for model evaluation. Most continuous methods tended to be more accurate with the high end for which MLR is the best performer (6.10% error), closely followed by NN-MLR (6.60% error). Classification techniques had a mixed of performance. The best classification techniques resulted in misclassification errors in the 3.2% to 6.5% range. These include Logistic Regression, the simple K-Means clustering, and binary methods derived from continuous models. The best derived binary results were achieved by NN-MLR and ANN for the overall set with a 4.76% error. Both favored the low end with 0% and 3.23% error respectively. The best derived binary results were achieved by NN-MLR and ANN for the overall set with a 4.76% error. Both favored the low end with 0% and 3.23% error respectively.

Although the powerful technique of Decision Trees did not perform as well, it provided a simple but valuable classification rule using only the KV variable. Based on the available data, the quick test defined by the derived using a threshold KV value (KV = 96), can provide an efficient prediction accuracy, which is second only to Logistics Regression with three variables. The discrete methods resulted in significantly better accuracy than QSARs methods with misclassification errors in the 3.2%

to 6.25% range compared to QSAR's 40% to 90% range. The application of Rule 1 enables effective classification of base oils into the slowly/highly biodegradability classes. A higher accuracy can be pursued for oils predicted to have a high biodegradability (KV < 96), using the continuous predictive methods of MLR or NN-MLR.

References

1. Haus, F.; Boissel, O.; Junter, J.A. Multivariate regression modeling of mineral base oil biodegradability based on their physical properties and overall chemical composition. *Chemosphere* **2003**, *50*, 993-948.
2. Kapur, G.S.; Sastry, M.I.S.; Jaiswal A.K.; Sarpal, A.S. Establishing structure-property correlations and classification of base oils using statistical techniques and artificial neural networks. *Anal. Chim. Acta* **2004**, *506*, 57-69.
3. Zeroski, S.; Blockeel, H.; Kompare, P.; Pfahringer, B.; Laer, W.V. Experiments in predicting biodegradability. In *Proceedings of the 9th International Workshop on Inductive Logic Programming*, Bled, Slovenia, June 1999; pp. 80-91.
4. Cambon, B.; Devillers, J. New trends in structure-biodegradability relationships. *Quant. Struct. Act. Relat.* **1993**, *12*, 49-56.
5. Boethling, R.S.; Sabljic, A. Screening-level model for aerobic biodegradability based on a survey of expert knowledge. *Environ. Sci. Technol.* **1989**, *23*, 672-679.
6. Amberger, D.; Sekusak, S.; Sabljic, A. Modelling biodegradation by an example based learning system. *Informatica* **1993**, *17*, 157-166.
7. Amberger, D.; Horvatic, D.; Sekusak, S.; Sabljic, A. Application of expert judgment to derive structure- biodegradation relationships, Environ. *Sci. Poll. Res.* **1996**, *3*, 224-228.
8. Baker, R.; Gamberger, D.; Mihelsic, J.R.; Sabljic, A. Evaluation of artificial intelligence based models for chemical biodegradability prediction. *Molecules* **2004**, *2*, 989-1004.
9. Sabljic, A.; Peijnenburg, W. Recommendations on modeling lifetime and degradability of organic compounds in air, soil and water systems. *Pure Appl. Chem.* **2001**, *73*, 1331-1348.
10. Basu, B.; Singh, M.P.; Kapur, G.S.; Ali, N.; Sastry, M.I.S.; Jain, S.K.; Srivastava, S.P.; Bhatnagar, A.K. Prediction of biodegradability of mineral base oils from chemical composition using artificial neural networks. *Tribol. Int.* **1998**, *31*, 159-168.
11. Rorije, E.; Loonen, H.; Müller, M.; Klopman, G.; Peijnenburg, W.J.G. Evaluation and application of models for the prediction of ready biodegradability in the MITI-I test. *Chemosphere* **1999**, *38*, 1409-1417.