Cent. Eur. J. Chem. • 10(4) • 2012 • 1149-1156
DOI: 10.2478/s11532-012-0038-x

VERSITA

Central European Journal of **Chemistry**

Instance-based regression with missing data applied to a photocatalytic oxidation process

Research Article

Florin Leon[1], Ciprian George Piuleac[2], Silvia Curteanu[2*], Ioannis Poulios[3]

[1]*"Gheorghe Asachi" Technical University Iasi, Department of Computer Science and Engineering,700050 Iasi, Romania*

[2]*"Gheorghe Asachi" Technical University Iasi, Department of Chemical Engineering, 700050 Iasi, Romania*

[3]*Aristotle University of Thessaloniki, Department of Chemistry, Laboratory of Physical Chemistry, 54124 Thessaloniki, Greece*

Received 28 October 2011; Accepted 25 January 2012

Abstract: In this paper, a modified nearest-neighbor regression method (kNN) is proposed to model a process with incomplete information of the measurements. This technique is based on the variation of the coefficients used to weight the distances of the instances. The case study selected for testing this algorithm was the photocatalytic degradation of Reactive Red 184 (RR184), a dye belonging to the group of azo compounds, which is widely used in manufacturing paint paper, leather and fabrics. The process is conducted with TiO_2 as catalyst (an inexpensive semiconductor material, completely inert chemically and biologically), in the presence of H_2O_2 (with the role of increasing the rate of photo-oxidation), at different pH values. The final concentration of RR184 is predicted accurately with the modified kNN regression method developed in this article. A comparison with other machine learning methods (sequential minimal optimization regression, decision table, reduced error pruning tree, M5 pruned model tree) proves the superiority and efficiency of the proposed algorithm, not only for its results, but for its simplicity and flexibility in manipulating incomplete experimental data.

Keywords: *Modified nearest-neighbor regression method • Prediction • Photocatalytic degradation • Reactive Red 184*

1. Introduction

The current requirements provided from an extensively social-economic development to a clean environment demand new approaches, focus on principles based on total elimination of toxic elements. A major source of pollution of ground and underground waters is represented by the textile and industries dyes [1].

The classical physico-chemical methods of treating water and wastewater (membrane filtration, precipitation, adsorption on activated carbon, incineration and others) are no longer sufficient because of incomplete destruction of the pollutants at its transfer from one phase to another. Advanced oxidant pollution control methods (POMA) have been widely studied and applied in the last years. Several examples are: photolysis (UV-C, UV-B), ozonolysis O_3/UV-B, the method based on Photo-Fenton Fe^{+3}/H_2O_2/UV/Vis reagent or heterogeneous photocatalysis TiO_2/UV-A [2]. The idea behind this concept is that the exposure of a strong oxidizing agent to UV light generates highly reactive radical species (especially the hydroxyl radical HO·), which can react with a wide spectrum range of compounds.

Heterogeneous photocatalytic oxidation, the method approached in this paper, has a variety of advantages as enumerated in [3-6]. The illumination of the particles with light energy, higher than the bandgap energy of the semiconductor (hu > Eg = 3.2 eV), produces excited high energy states of electron and hole pairs (e-/h+). These pairs can migrate to the surface of the particle and initiate a wide range of chemical redox reactions, which can lead to complete mineralization of the organic pollutants [7]. Among various semiconducting materials (oxides, sulfides *etc.*) most attention has been

* *E-mail: silvia_curteanu@yahoo.com*

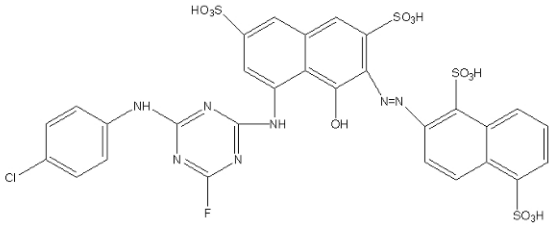

Figure 1. Reactive Red 184 (Cibacronrot F-B).

given to TiO_2 (anatase) or ZnO because of their highly photocatalytic activity, resistance to photocorrosion, biological immunity and low cost.

The photocatalytic degradation of commercial dye Reactive Red 184 (RR184) has been investigated under various experimental conditions, using a regression-based simulation method. This procedure is recommended when insufficient data is available and the domain of interest is not uniformly covered by the experiment.

Beside analytical curve-fitting models, machine learning techniques are valuable tools for regression. In this paper, we propose the use of a modified nearest-neighbor regression method, which proves to be superior to other models, especially in the presence of incomplete information of the measurements.

2. Experimental procedure

RR184 (Fig. 1) belongs to the group of azo compounds, which is widely used in manufacturing paint, paper, leather and fabrics.

In a previous study [8], the photocatalytic oxidation of RR184, in aqueous solution and in the presence of semiconductor powder of TiO_2-P25 was studied. The experimental investigations were conducted under various conditions represented by different values of initial concentrations of the dye, TiO_2, H_2O_2, pH and time. After 60 minutes illumination, complete decolorization of the solution occurs, due to the disruption of the color former group in the best accepted conditions. The photodegradation kinetics followed pseudo-first order being in agreement with the model of Langmuir-Hinselwood.

The combination of photocatalytic oxidation with TiO_2, which is an inexpensive semiconductor material, completely inert chemically and biologically, offers a "clean" technology oxidative destruction of various organic compounds found in water and wastewater. The addition of hydrogen peroxide (H_2O_2) leads to an increase in the rate of photo-oxidation, having a dual role in the process of photocatalytic degradation: it accepts a photogenerated electron from the conduction band and thus promotes the charge separation and forms ·OH radicals via superoxide, while a possible reaction of H_2O_2 with the photogenerated intermediates cannot be excluded. An excess amount of H_2O_2 could act as a hole or ·OH scavenger or react with TiO_2 to form peroxo compounds, which is detrimental to the photocatalytic action.

A number of 240 experimental data describes the heterogeneous photocatalytic oxidation of RR184 over semiconducting powder (TiO_2) under various operating conditions: initial concentrations of TiO_2 and H_2O_2, pH and time of the process. But within the domain of these data there are some places with missing data; our regression procedure is especially designed for this kind of situations, fruent in chemical experimental practice.

The experimental data are visible on the diagrams that show the influence of different reaction conditions on the final concentration of RR184, in the section Results and Discussion.

3. Regression analysis

Regression analysis includes any techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables, by estimating the conditional expectation of the dependent variable given the independent variables. It is widely used for prediction and forecasting, and it has substantial overlap with the field of machine learning [9].

The general regression model is [10]:

$$y_i = g(\mathbf{x}_i) + e_i \tag{1}$$

where g is the regression model (the approximating function), y_i is the output (the desired output value that corresponds to the $\mathbf{x}_i$ input of the training set), and e_i is a residual whose expected error given the sample point $\mathbf{x}_i$ is:

$$E(e_i \mid \mathbf{x}_i) = 0 \tag{2}$$

3.1. Regression based on k-nearest neighbor method

k-Nearest Neighbor (kNN) is a simple, efficient way to estimate the value of the unknown function in a given point using its values in other points. Let $S = \{\mathbf{x}_1,...,\mathbf{x}_m\}$ be the set of training points. The kNN estimator is defined as the mean function value of the nearest neighbors [11]:

$$\tilde{g}(\mathbf{x}) = \frac{1}{k}\sum_{\mathbf{x}'\in N(\mathbf{x})} g(\mathbf{x}') \tag{3}$$

where $N(\mathbf{x}) \subset S$ is the set of *k* nearest points to **x** in *S*.

Another version of the method considers a weighted average, where the weight of each neighbor depends on its proximity to the reference point:

$$\tilde{g}(\mathbf{x}) = \frac{1}{z}\sum_{\mathbf{x}'\in N(\mathbf{x})} g(\mathbf{x}')K\big(d(\mathbf{x},\mathbf{x}')\big) \tag{4}$$

where:

$$d(\mathbf{x},\mathbf{x}') = \left(\sum_{i=1}^{n}\left|x_i - x'_i\right|^p\right)^{1/p} \tag{5}$$

is the *Minkowski distance* between the two points **x** and **x**', *K* (the *kernel*) is a monotonically decreasing function of the distance and *z* is a normalization factor:

$$z = \sum_{\mathbf{x}'\in N(\mathbf{x})} f\big(d(\mathbf{x},\mathbf{x}')\big) \tag{6}$$

The *Euclidian distance* (*p* = 2 in Eq. 5) is commonly used as a metric.

A classic, often used kernel, is the *inverse power distance* function, which raises the distance to a negative power [12,13], such that the magnitude of the power determines how local the regression is:

$$K(d) = \frac{1}{d^q} \tag{7}$$

This type of weighting function goes to infinity as the query point approaches a training point and forces the locally weighted regression to exactly match the training point. If the training data contain noise, an exact interpolation may give suboptimal results and thus a limited magnitude is more appropriate [14]:

$$K(d) = \frac{1}{1+d^q} \tag{8}$$

Another weighting function with an infinite extent is the *Gaussian* kernel [15,16]:

$$K(d) = \mathrm{e}^{-d^2} \tag{9}$$

In a cognitive psychology study, Aha and Goldstone [17] discovered that a related distance-weighting formula can better model the human classification behavior. This is the *exponential* kernel:

$$K(d) = \mathrm{e}^{-d} \tag{10}$$

A comprehensive review of locally weighted regression and classification methods is given by Atkeson, Moore and Schaal [18].

3.2. Selected kernels

As stated in the previous theoretical section, the information about the value of the function in a certain test (or query) point is given by the values of the available points in the training set. Depending on the problem, it may be beneficial to take into account only the closest training point (simple nearest neighbor), several *k* nearest points (k-nearest neighbor) or all the points in the training set (when *k* equals the number of training instances). It is logical that the closer a training point is to the query point, the stronger the influence it will have on the result. Therefore, the weight of a training point is high when the distance between the query point and the training point is small, and conversely, the weight is low when the distance is large, and it must converge to 0 when the distance tends to infinity. This is the role of the kernel: to specify the relation between the distance and the weight of a training point, while complying with the requirements mentioned above.

We considered all the training instances for our model, and therefore the computed value for a query point **x** is:

$$\tilde{g}(\mathbf{x}) = \frac{\sum_{i=1}^{n} g(\mathbf{t}_i)\cdot K\big(d(\mathbf{x},\mathbf{t}_i)\big)}{\sum_{i=1}^{n} K\big(d(\mathbf{x},\mathbf{t}_i)\big)}, \tag{11}$$

where *n* is the number of training instances, *K*(*d*) is the kernel, and *d* is the Euclidian distance between the query point **x** and the training point $\mathbf{t}_i$.

Based on the functions presented above, we used the following two kernels for our experiments:

$$K_1(d) = \frac{1}{d^{\mathrm{a}}} \tag{12}$$

and

$$K_2(d) = \mathrm{e}^{-d/\mathrm{b}} \tag{13}$$

where *α* and *β* are real numbers, parameters which control the balance between a strictly local behavior and a more global one. By varying their values, the

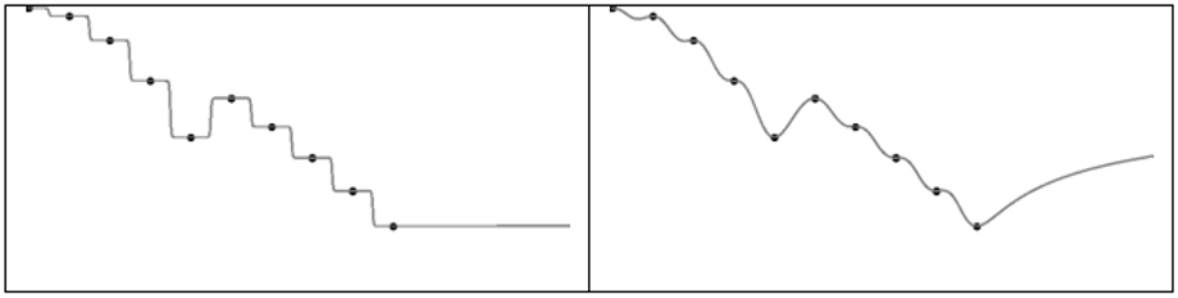

Figure 2. Regression with kernel K_1 when: a) $\alpha = 20$; b) $\alpha = 2$.

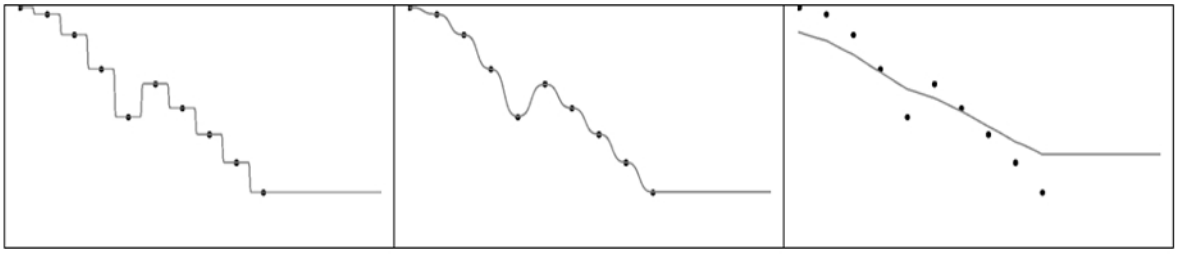

Figure 3. Regression with kernel K_2 when: a) $\beta = 1$; b) $\beta = 10$; c) $\beta = 100$.

regression model can adapt to the nature of the problem under study. All the training instances take part in the regression process, and their influence is weighted by the kernel value.

As an example, Figs. 2 and 3 show the regression results for a set of arbitrary points which define a function $y = g(x)$, with $x, y \in \Re$. Fig. 2 displays two situations for kernel K_1 when $\alpha = 20$ and $\alpha = 2$, respectively. In the first case, the regression is purely local and it basically reduces to simple nearest neighbor (NN) behavior (when $k = 1$). In the second case, the regression shows better generalization capability, and interpolates the training points.

Fig. 3 displays three situations for kernel K_2 when $\beta = 1$, $\beta = 10$ and $\beta = 100$, respectively. It can be noticed that the influence of β is contrary to that of α. When β is small, the regression is local, and when it increases, the regression becomes global. One can find a proper value of the parameter that provides optimal regression results. In this situation, the second case appears to be closer to the desired behavior.

3.3. Advantages and disadvantages of the kNN method

Beside its simplicity, one of the advantages of the *kNN method* is the capability to learn complex functions, because no information is lost in the training process. "Eager learners" such as algorithms for building decision trees or neural networks create global approximations of the model, while "lazy learners" such as kNN create many local approximations. As the number of training points approaches infinity and k gets large, the method is guaranteed to have an error rate no worse than twice the Bayes error rate, which is the minimum achievable error rate given the distribution of the data [19].

The main reason for choosing kNN for our problem over other methods is the manner in which missing values can be handled. Many high-quality algorithms rely on the completeness of the dataset. The missing values are handled as a preprocessing step, *e.g.* they are assigned average values of the corresponding attribute. However, this can introduce erroneous information into the computation. On the other hand, the designed kNN algorithm simply ignores the missing values, and computes the distance between instances only on the defined attributes. In this way, it relies only on the actual information provided by the experiments.

Among the disadvantages of kNN we can mention the high computation time needed to achieve a query, especially when the number of training points is large, and its sensitivity to irrelevant attributes, because all the attributes contribute to the calculation of the distance between two points. Several methods have been proposed to handle these deficiencies, *e.g.* the use of k-dimensional trees [20], which can speed up the finding of the neighbors, for the former, and various feature selection or attribute weighting schemes for the latter.

In our case, these problems are alleviated by the fact that the number of instances is only 240 and thus the algorithm responds very quickly, and the dataset contains no irrelevant attributes.

4. Alternative methods

In our paper, the regression procedure designed for experiments with missing data was compared with other classification algorithms, enumerated below.

Presently, Support Vector Machines (SVM) [21] are one of the best classification methods available, which ensure a good generalization capability of the learned model and is based on solid mathematical foundations. This technique uses an optimization algorithm to find a "large margin" of separation between classes, and the Sequential Minimal Optimization (SMO) algorithm [22] is usually employed in this respect. This technique has also been adapted for regression [23], and the SMO algorithm has been improved to deal with the continuous values of the output [24].

The Decision Table Method [25] is a simple method of data representation, which consists of a schema (a set of features) and a body (a multi-set of labeled instances, such that each instance consists of a value for each of the features in the schema and a value for the label). It also has a default mapping when there are no instances to support a decision. The IDTM (Inducer of DTMs) algorithm has been proposed to find the optimal feature subset needed for a decision table, and incremental cross-validation is used for finding (near-) optimal solutions.

The Reduced Error Pruning Tree, REPTree [26] is a fast algorithm that builds a decision tree using information gain or variance reduction and prunes it using reduced-error pruning with back-fitting. It sorts the values for numeric attributes only once. Missing values are dealt with by splitting the corresponding instances into pieces.

A further development is the M5 pruned model tree [27]. While regression trees have values as their leaves, M5 builds trees with multivariate linear models and these model trees are analogous to piece-wise linear functions. M5 trees are generally much smaller than regression trees, are applicable to larger problems involving more than 20 attributes, and can also provide increased accuracy.

Artificial neural networks can be considered as efficient alternatives for process modeling, with the main feature (advantage) deriving from their empirical character. They work only with input-output data and don't need significant knowledge about the phenomenology of the process. Our group previously applied neural network modeling for different photocatalytic processes such as: the removing of the azo dye Reactive Black 5 from wastewater [28], the photocatalytic oxidation of Triclopyr [29], photodegradation of the cationic dye Alcian Blue 8 GX [30] or the photocatalytic degradation of sulfamethazine in aqueous heterogeneous solutions containing oxide semiconductors [7]. On the other hand, different machine learning methods such as neural networks and categorization algorithms have been applied to various processes and classes of compounds such as copolyethers with mesogene groups in the main chain [31], ferrocene derivatives [32] or azo aromatic compounds [33].

5. Results and discussion

The experiments evaluated the influence of the main parameters on the photodegradation process: initial concentration, pH, catalyst concentration and H_2O_2 on the final concentration of the dye. The variation in time of these parameters was represented in Figs. 4-7.

The total dye removal at different initial concentrations of RR184 (in a range of 5-60 ppm), using 1 g L^{-1} TiO_2 P-25 is represented as a function of time in Fig. 4. It is evident that low initial amount of dye determines a rapid removal.

The quantity of catalyst plays a major role in the photodegradation rate, reducing the time significantly (Fig. 5). A concentration of 6 g L^{-1}, where removal of colorant takes 20 minutes is not acceptable from an economic point of view. For this reason, the experiments conducted to study the influence of other parameters could be considered at 1 g L^{-1} TiO_2 P-25.

pH plays an important role if we take into account the initial colorant concentration. At low concentrations, using the same concentration of catalyst of 1 g L^{-1}, dye photodegradation occurs at pH values 3-4. For initial concentration values above 20 mg L^{-1}, total removal takes place in a pH range of 9-12 (Fig. 6).

Fig. 7 shows the influence of H_2O_2 concentration, considering the same catalyst concentration of 1 g L^{-1} TiO_2 P-25. Using more than 100 ppm H_2O_2, the efficiency of the process is highly improved.

For both kernels presented in Section 3.2, the corresponding parameter was varied in order to maximize the correlation coefficient between the actual output and the output provided by our algorithm. We considered two situations. First, the algorithm was applied to the whole dataset, in order to analyze its approximation behavior. Second, the dataset was split into 2/3 of the instances used for training and the accuracy of the algorithm was measured for the remaining 1/3 test instances, in order to investigate its generalization capabilities. The training and validation phases were designed in this way.

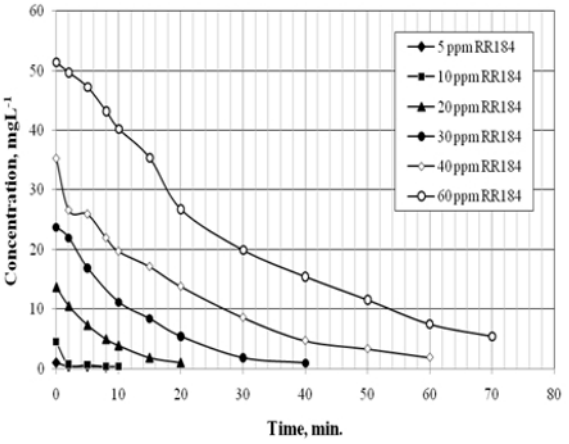

Figure 4. The influence of initial concentration of RR184 on the photodegradation process.

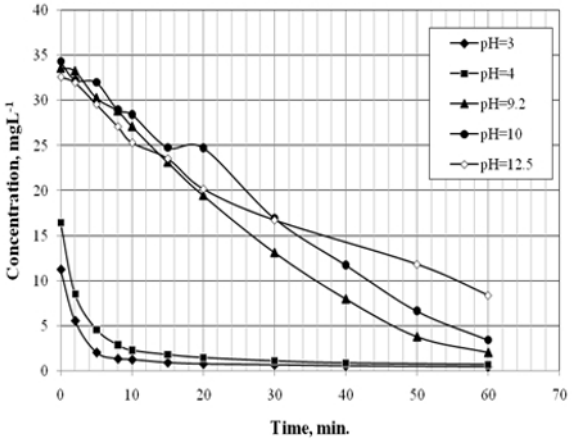

Figure 6. The influence of pH on the photodegradation process.

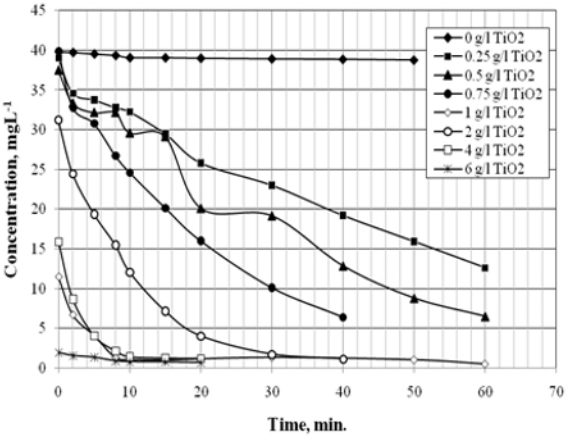

Figure 5. The influence of TiO_2 P-25 concentration on the photodegradation process

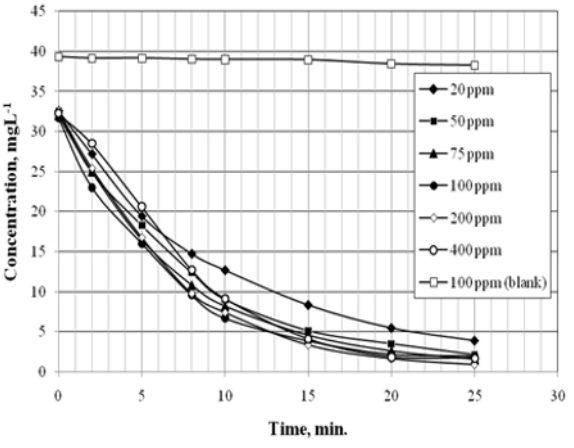

Figure 7. The influence of H_2O_2 concentration on the photodegradation process

When using kernel K_1, the generalized inverse power distance for the whole database, all the values of parameter α for $\alpha \in [1, 50]$ provide the same value for the correlation coefficient: $r = 0.9599$.

For the train-test split case, a maximum $r = 0.8633$ was obtained for $\alpha = 17.1$. The evolution of the correlation coefficient for the test set as a function of α is displayed in Fig. 8.

For the second kernel K_2, the generalized exponential kernel, the optimal value of the parameter was found to be $\beta = 10^{-6}$ for the whole database case, yielding a correlation coefficient $r = 0.9599$.

The value for the train-test split case was $\beta = 2.4 \times 10^{-3}$, which provided a correlation coefficient $r = 0.8626$. The evolution of the correlation coefficient for the test set as a function of β is displayed in Fig. 9.

Table 1 presents the results obtained by our algorithm with the best parameters compared to the results obtained by several other regression methods. For this purpose, we used the versions implemented in the Weka software [34], a free collection of machine learning algorithms. Out of the many regression algorithms available, we only included those which provided the best correlation coefficients for our data.

As the dataset is composed of data provided by different experiments, and each experiment has different attributes with missing values, the splits into training and testing sets were made in such a way that the proportion (*e.g.* 2/3 *vs.* 1/3) was constant for each data subset.

The best values for each case are highlighted. It can be seen that our algorithm clearly outperforms the others. In both situations, the values of the correlation coefficients are greater than 0.95 on the training set and greater than 0.86 on the testing set.

Regarding the proposed regression method, its main advantage lies in its simplicity and flexibility, because it involves only one parameter to be determined by the user. This parameter depends on the problem at hand. In our case, it seems that the preferred values are the ones approaching a pure NN behavior, when the value

Table 1. Correlation coefficients obtained by different regression algorithms.

Test Case	SMO Regression	Decision Table	REPTree	M5 pruned model tree	kNN Regression IPD (K_1)	kNN Regression Exp (K_2)
All data	0.8109	0.9388	0.9073	0.8137	**0.9599**	**0.9599**
2/3 – 1/3 split	0.7552	0.8132	0.6203	0.7947	**0.8633**	0.8626
Crossvalidation	0.8001	0.8576	0.8714	0.8478	**0.9114**	0.9061

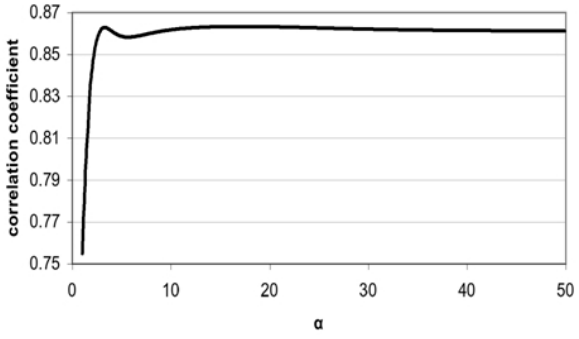

Figure 8. The correlation coefficient for the test set as a function of α.

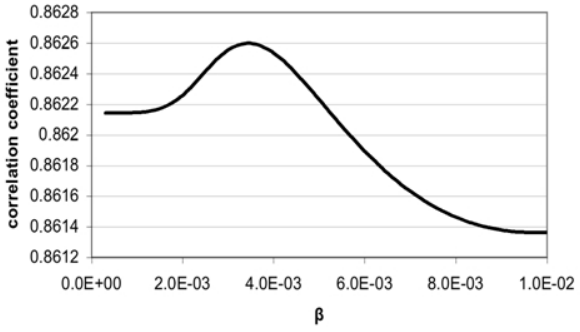

Figure 9. The correlation coefficient for the test set as a function of β.

of a query point is computed by taking into account only the value of the closest point in the training set.

The method presented in this paper can also be considered as a regression methodology for any other function. It is a general method which can be applied to different case studies selected from the real world problems. Its main advantage comes from the accurate results provided even when working with incomplete series of data.

6. Conclusions

The main goal of this paper was to develop and test a regression algorithm to model experimental data which do not uniformly cover the domain investigated. When a function is too complex or there are missing data in the training set, a parametric curve fitting becomes very difficult or even impossible. In these cases, non-parametric methods such as those provided by machine learning prove to be very useful. Since an instance-based approach does not explicitly build a model (the training data itself is the model), it does not have any constraints regarding the difficulty of the learned function. The way in which missing values are treated makes it applicable to situations with partial information, where our algorithm uses only the available data, without "filling in" additional computed values (*e.g.* mean values), or ignoring the incomplete instances.

The photocatalytic degradation of commercial dye Reactive Red 184 was chosen as case study to prove the efficiency of kNN algorithm because incomplete information is available concerning the experiments. The final concentration of the dye was modeled as a function of reaction conditions (time, initial concentrations of RR184, TiO_2 catalyst, H_2O_2, pH).

The generalization capability of the kNN method was tested by splitting data into training and validation sets. Accurate results were obtained in the validation phase with correlation coefficient being values greater than 0.9. Other similar algorithms - sequential minimal optimization regression, decision table, reduced error pruning tree, M5 pruned model tree - had more precise results when simulating the same approached photocatalytic process.

The method presented in this paper, which extends the nearest neighbor paradigm for regression, demonstrates superior results compared to other algorithms and it is also much more flexible, with fewer restrictions than other multidimensional curve fitting techniques. It only involves one parameter to be determined and it can provide satisfactory results when the experimental dataset is incomplete. Therefore, it can also be considered as a regression methodology for any sampled function.

Acknowledgement

This work was supported by CNCSIS-UEFISCSU, project number PNII-IDEI 316/2008, *Behavioural Patterns Library for Intelligent Agents Used in Engineering and Management*.

References

[1] P. Anjali, S. Poonam, I. Leela, Int. Biodeter. Biodegr. 59, 73 (2007)
[2] K.H. Gregor, In: W. Wesley Eckenfelder and A. Bowers (Eds.), Chemical Oxidation (J. Roth, Lancaster, Pensilvania, USA, 1994) Vol. 1-6
[3] D. Bahnemann, J. Cunningham, M.A. Fox, E. Pelizzetti, P. Pichat, N. Serpone, In: G. Helz, R. Zepp, D . Crosby (Eds.), Photocatalytic treatment of waters, in Aquatic and Surface Photochemistry (Lewis Publs., Boca Raton, FL, 1994) 261
[4] M.R. Hoffman, S. Martin, W. Choi, D.W. Bahnemann, Chem. Rev. 95, 69 (1995)
[5] D.Y. Goswami, In: K.W. Boer (Ed.), Engineering of the Solar Photocatalytic Detoxification and Disinfection Processes, in Advances in Solar Energy (American Solar Energy Society Inc., Boulder, Colorado, 1995) Vol. 10, 165
[6] S. Malato, J. Blanco, C. Richter, M. Maldonado, Appl. Catal. B: Environ. 37, 1 (2002)
[7] C.G. Piuleac, I. Poulios, S. Curteanu, Env. Eng. Manag. J. 8, 439 (2009)
[8] I. Poulios, A. Papathanasiou, E. Ntarakas, H. Xatziefangelou, E. Papachristou, In: Seventh National Conference on Renewable Energy Sources, 6-8 November 2002, Patras, Greece, (Patras 2002)
[9] R.D. Cook, S. Weisberg, Sociol. Methodol. 13, 313 (1982)
[10] Q. Li, J. S. Racine, Nonparametric Econometrics: Theory and Practice (Princeton University Press, Princeton, New Jersey, USA, 2006)
[11] A. Navot, L. Shpigelman, N.Tishby, E. Vaadia, Advances in Neural Information Processing Systems18, 995 (2006)
[12] D. Shepard, In: 23rd ACM National Conference, 27-29 Aug. 1968, New York, USA (Brandon Systems Press, Princeton, New Jersey, USA, 1968) 524
[13] D. Ruprecht, H. Müller, In: 5th Eurographics Workshop on Visualization in Scientific Computing, 30 May – 1 Jun. 1994, Rostock, Germany (Eurographics Association and Blackwell Publishers, Oxford, UK 1994) 517
[14] G. Wolberg, Digital Image Warping (IEEE Computer Society Press, Los Alamitos, Californica, USA, 1990)
[15] P. Deheuvels, RSA 25, 5 (1977)
[16] M.P. Wand, W.R. Schucany, Can. J. Stat. 18, 197 (1990)
[17] D.W. Aha, R.L. Goldstone, In: Proceedings of the 14th Annual Conference of the Cognitive Science Society, 29 July – 1 August (Indiana University, Bloomington, USA, 1992) 534
[18] G.C. Atkeson, A.W. Moore, S. Schaal, J. Artif. Intell. Rev. 11, (1997)
[19] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory 13(1), 21 (1967)
[20] J.L. Bentley, Comm. ACM 18, 509 (1975)
[21] V.N. Vapnik, The Nature of Statistical Learning Theory (Springer-Verlag, Berlin, Germany 1995)
[22] J.C. Platt, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Advances in Kernel Methods - Support Vector Learning, Technical Report MSR-TR-98-14, Microsoft Research (Microsoft Press, Redmond, Washington, USA, 1998)
[23] B. Schölkopf, A.J. Smola, Learning with Kernels (MIT Press, 2002)
[24] S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy, IEEE Transactions on Neural Networks 11, 1188 (2000)
[25] R. Kohavi, In: 8th European Conference on Machine Learning, 25-27 Apr. 1995, Heraclion, Crete, Greece (Springer, Berlin-Heidelberg-New York 1995) 174
[26] R.J. Quinlan, Mach. Learn. 1, 81 (1986)
[27] R.J. Quinlan, In: 5th Australian Joint Conference on Artificial Intelligence, 16-18 Nov. 1992, Hobart, Tasmania, Australia (World Scientific, Singapore 1992) 343
[28] G.D. Suditu, M. Secula, C.G. Piuleac, S. Curteanu, I. Poulios, Rev. Chim.-Bucharest 59, 816 (2008)
[29] C.G. Piuleac, I. Poulios, F. Leon, S. Curteanu, A. Kouras, Sep. Sci. Technol. 45, 1644 (2010)
[30] F.A. Caliman, S. Curteanu, C. Betianu, M. Gavrilescu, I. Poulios, J. Adv. Oxid. Technol. 11, 316 (2008)
[31] F. Leon, S. Curteanu, C. Lisa, N. Hurduc, Mol. Cryst. Liq. Cryst. 469, 1 (2007)
[32] C. Lisa, S. Curteanu, V. Bulacovschi, D. Apreutesei, Rev. Roum. Chim. 53(4), 283 (2008)
[33] C. Lisa, S. Curteanu, Comp. Aided Chem. Eng. 24, 39 (2007)
[34] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, SIGKDD Explorations 11 (2009)