SPC

Advanced Journal of Chemistry-Section A

Journal homepage: www.ajchem-a.com

Original Research Article

Kernelled Naïve Bayes Using a Balanced Dataset for Accurate Classification of the Material Toxicity

Ali Ekramipooya[1,2], Davood Rashtchian[1,2], Mehrdad Boroushaki[3]*

[1] *Department of Chemical and Petroleum Engineering, Sharif University of Technology, Tehran, Iran*
[2] *Center for Process Design, Safety and Loss Prevention (CPSL), Sharif University of Technology, Tehran, Iran*
[3] *Department of Energy Engineering, Sharif University of Technology, Tehran, Iran. P.O.Box 14565-114*

ARTICLE INFO

Article history

Submitted: 22 February 2021
Revised: 24 March 2021
Accepted: 03 April 2021
Available online: 04 April 2021
Manuscript ID: AJCA-2102-1244

DOI: 10.22034/AJCA.2021.274570.1244

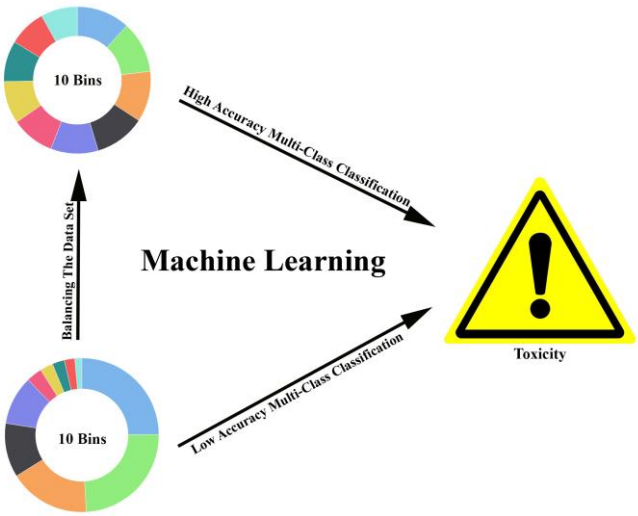
KEYWORDS

QSAR
Machine Learning
Accurate Toxicity Prediction
Imbalanced Dataset
Multi-Class Classification

ABSTRACT

In this work, a new multi-class classification approach was employed in the QSAR model to assess chemical toxicity prediction through handling the imbalanced dataset as the critical preprocessing step in the training dataset. Various classifiers of the decision tree, K-NN, naïve Bayes, kernelled naïve Bayes, and SVM and two distinct acute aquatic toxicity datasets towards Daphnia Magna and Fathead Minnow Fish were used to evaluate the generality of the approach. The quantitative response (LC_{50}) was discretized into ten bins. Imbalanced dataset classification leads to a high level of errors since the classifier tends to learn from the majority class more than the minority class. Each training dataset was specified by different weights related to the class population. These datasets were then bootstrapped based on their weights to convert the imbalanced dataset into a balanced one. This approach enhanced the accuracy of classification of material toxicity dramatically (up to 99%). Balanced dataset classification had high overall accuracy when correlated attributes were removed. Therefore, fewer attributes are sufficient to predict material toxicity. The overall accuracy improvement of the decision tree, K-NN, naïve Bayes, kernelled naïve Bayes, and SVM for the Daphnia Magna dataset after balancing the data set are 58.03%, 55.08%, 9.09%, 72.48%, and 53.05%, respectively.

GRAPHICAL ABSTRACT



* Corresponding author: Boroushaki, Mehrdad
E-mail: boroushaki@sharif.edu
Tel number: +982166166136

Introduction

Risk assessment consisting of risk analysis and evaluation is an essential prerequisite for risk management. The analysis step identifies the potential event's negative impact on the individuals, assets, and the environment. In the evaluation step, judgments are carried out based on the risk-based risk analysis's tolerability while considering practical factors [1, 2]. The environmental risk assessment focuses on the nature and likelihood of hazardous effects in organisms such as humans, animals, and plants due to their exposure to hazards. Different chemicals based on their diverse structure can hurt the organisms. The upsurge of chemical production requires a risk assessment method, which is fast, cost-effective, and accurate[3]. The European Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) regulation suggested using computer-aided methods for hazard identification and risk assessment of the chemicals instead of the experimental approaches. Moreover, computer-aided methods comply with no animal testing and sustainability definition [4-7]. Organization for Economic Co-operation and Development (OECD) set five principles for considering a Quantitative structure-activity relationship (QSAR) model for regulatory application: 1-a defined endpoint. 2- an unambiguous algorithm. 3-a defined domain of applicability. 4-appropriate measures of goodness-of-fit, robustness, and predictability. 5-a mechanistic interpretation, if possible [8].

Machine learning, an *in silico* method, gives computers the capability to learn from the data and make fast and accurate predictions[9, 10]. Various research groups have widely used machine learning methods for toxicity prediction [11-16]. QSAR is an approach based on the idea that chemical activity is related to the structure of a molecule [17, 18]. Chemical descriptors are specific numbers attributed to chemical structures so that computers can use them; they are the link between the activity and the structure of molecules. Different machine learning methods and descriptors have various efficiencies for predicting toxicity[19]. In machine learning, supervised learning algorithms use a mathematical model to link inputs to the desired outputs. The training dataset consists of examples in which each training example has one input and the desired output. Supervised machine learning algorithms include classification and regression. Classification and regression algorithms are used when outputs are restricted to a limited set of values and numeric, respectively[20-22]. Computational methods based on QSAR predictions of the toxicity can be generally divided into two classes: quantitative regression and qualitative classification. After determining the chemical descriptors, the vital step in QSAR is using machine learning methods to predict the toxicity based on chemical descriptor inputs. Most studies used regression methods to predict toxicity in the literature because the attributes (chemical descriptors) and labels (toxicity) are numeric. However, some researchers preferred to use qualitative classification. In the qualitative classification, the label is binary (toxicant and non-toxicant) or ordinary (low, medium, and high toxicity)[23-36].

In this research study, the quantitative response (LC_{50}) in the range of 0 to 10 is divided into 10 bins. Since these datasets are rigorously imbalanced, the previous studies in this field are limited to a low number of bins (≤ 5) with an overall accuracy of about 80% [29, 30]. In this study, the imbalanced datasets were handled by a balancing technique. Each training data was specified by different weights considering their class population and then were bootstrapped based on their weights to convert the imbalanced dataset into a balanced one. Various classification algorithms were used in this study, including decision tree, K-nearest neighbors (K-NN), naïve

Bayes, Kernelled naïve Bayes, and support vector machine (SVM).

Preliminaries

Classification

Classifiers use machine learning algorithms to predict categorical (discrete) class labels. Classification is a two-step process consisting of a learning step in which a classifier is built up and a classification step in which the classifier is used to predict class labels. Classifiers are divided into eager and lazy learners. Eager learners construct the classifier when given a set of training datasets; on the contrary, the lazy learners store the training dataset until it is given test data and then classify it based on its similarity to the stored training dataset. Among classifiers used in this study, just K-NN is a lazy learner [37, 38].

Decision tree

As shown in **Figure 1**, A decision tree has a flowchart structure, consisting of an internal node, branch, and leaf node where each internal node stands for a test on an attribute. Each branch denotes a result of the test, and each leaf node represents a class label placed at the terminal of the tree. In the decision tree, the topmost node is called the root node. The decision tree adopts a top-down approach that uses criteria such as information gain (IG) to select attributes to decrease the entropy (E) of the tree. The expected information required to classify an example is equal to entropy. Entropy is the measure of the amount of uncertainty in data. Information gain is also called Kullback-Leibler divergence, which is the effective change in entropy after deciding on a particular attribute. Information gain calculates the relative change in entropy. The first step is to select the attribute that leads to the highest possible information gain, which will be used as the root node[37, 38].

$$IG(S,A) = E(S) - E(S|A) \qquad (1)$$

where, $E(S)$ is the entropy for the dataset before deciding on a particular attribute and $E(S|A)$ is the conditional entropy for the dataset considering a particular attribute.

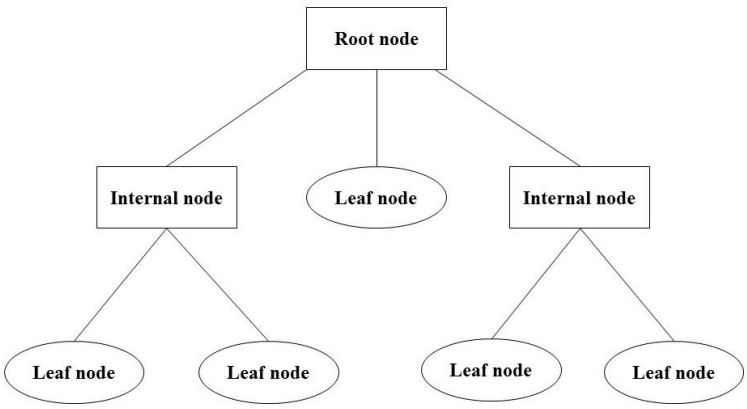


Figure 1. The simple structure of the decision tree

K-NN

K-NN classifier is based on learning by comparing the similarity between a given test data and the training dataset. Similarity measurement is defined based on the closeness of the data. The Euclidean distance was used for

estimating the similarity. The Euclidean distance between two points is calculated by:

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \quad (2)$$

where, $X_1 = (x_{11}, x_{12}, \dots, x_{1n})$ and $X_2 = (x_{21}, x_{22}, \dots, x_{2n})$. For K-NN classification, the anonymous example is designated as the most common class among its k nearest neighbors. When k = 1, the anonymous example is named the class of the training example closest to it[37, 38].

Naïve bayes and kernelled naïve bayes

Naïve Bayes classifier predicts the probability that an example belongs to the class. It is named naïve because the attributes are considered conditionally independent. However, this assumption is not correct in the real world, but the naïve Bayes classifier has good accuracy in predicting the class. Bayes classifier uses Bayes' theorem for classification:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (3)$$

where, $P(C_i|X)$ is the posterior probability indicating example X belongs to the class C. $P(X)$ and $P(X|C_i)$ are the prior probability of X and posterior probability of X conditioned on H, respectively. Based on the Bayesian classifier method, when there are multiple classes, the naive Bayes classifier assigns example $\mathbf{X}$ to the class C_i if:

$$P(C_i|X) > P(C_j|X) \text{ for } 1 \le j \le m,\ j \ne i \quad (4)$$

The attributes of the dataset are assumed to be independent, thus:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad (5)$$

In this study, for estimation of the $P(x_k|C_i)$ in the numeric attributes, the Kernel density estimation was used with the n number of Gaussian functions.

$$P(x_k|C_i) = \frac{1}{n}\sum_{i=1}^{n} g(x_k, \mu_{C_i}, \sigma_{C_i}) \quad (6)$$

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (7)$$

where μ and σ are the mean and standard deviation of the dataset[37, 39, 40].

SVM

The SVM classifier usually is used for the classification of binary classes with the separating hyperplane:

$$W.X + b = 0 \quad (8)$$

here $W = \{w_1, w_2, \dots, w_n\}$, n stands for the number of attributes, and b represents the bias. The best hyperplane is when the distance from it to the nearest data point on each side is maximized.

In this study, there is a multi-class classification, so the method of one-versus-all was used. This method is implemented in two steps. In the first step, *m* binary classifiers are trained for the given *m* classes; then, a voting process is applied to the ensemble SVMs. The new example is applied to an ensemble structure to classify an unseen example, then each classifier votes, while the majority of the votes is selected as the example class [37, 38, 41-43]

Balancing dataset

The main goal of many classifiers is to maximize their overall accuracy, but in the situation in

which the data samples are distributed unequally among the different classes, the error originates from learning from the minority class samples because the classifier tends to favor the majority class. There are two standard methods for handling the imbalanced data, including oversampling and the minority and majority class's under-sampling. It has been revealed that oversampling is a lot more effective than under-sampling in maximizing overall accuracy, even for complex data[44, 45]. Random oversampling of the minority class is the random sampling of the minority class with replacement, indicating that the samples are placed back in the minority sample set after each sampling and can be selected again until the minority class size is the same as the majority class. Random under-sampling of the majority class can be defined as removing samples until the number of samples of the minority class equals the majority class samples. However, it may increase the classifier's variance and remove useful or necessary samples[46, 47].

The mentioned oversampling and under-sampling methods are usually used for binary class classification; however, an innovative method is required for the multi-class classification. This paper's balancing method is a cost-sensitive algorithm that weights the examples to rebalance the importance of the examples for the different classes on an imbalanced data set. This approach attempts to equilibrate the classes such that the effect of different classes on the learning step is proportional to their costs. A typical process is to assign different weights to training examples of different classes in proportion to their misclassification costs; then, the weighted examples are given to a cost-blind learning algorithm[48, 49].

Balancing the dataset was carried out with two steps, including weight generation for different examples and then bootstrapping (random sampling with replacement) considering the weights, leading to oversampling and under-sampling the minority class and majority class by default, respectively. At first, weights were specified for each example, such that weights sum up equally per range. The total weight parameter is set to 1 such that the sum of all weights is equal to 1. the weighted sum of examples with different ranges has been the same. For the n number of the bins, the weights are calculated with the 9 equation.

$$\sum_{i}^{i=n} w_i m_i = 1 \qquad (9)$$

where, w_i is the weight allocated to each example and m_i is the number of examples in each bin. For illustrating the generating weights well, an artificial example is solved with two bins. In bin 1 there are two examples, and in bin 2 there is just one example. The results of equation 9 are reported in **Table 1**.

Table 1. weighting the dataset for the artificial example

Bin	Weight	Count	Weight*Count
1	0.25	2	0.5
2	0.5	1	0.5
			Sum = 1

The bootstrapped sampling is based on the replacement. Thus, all examples have an equal chance of being chosen at every step. Example weights are considered during the bootstrapping, so the balanced dataset is obtained [37]. The steps of balancing the dataset are illustrated in **Figure 2**.

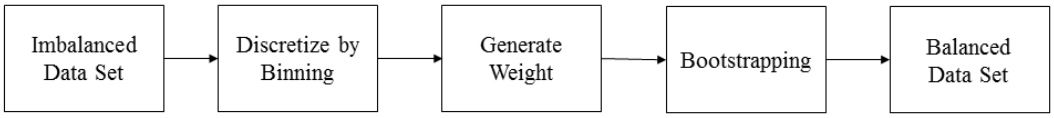


Figure 2. Block diagram of balancing the dataset.

Validation

One of the most critical steps of the classification study is validation to estimate the classifier's reliability and accuracy for the present data and the future application. The n-fold cross-validation is the well-accepted approach in the QSAR study. In this method, the given dataset is divided into n subsets, and one subset is used for the testing, and the others are used for training the classifier. This process is repeated n times until all the datasets are covered and used for testing and training. The n-folded cross-validation is fit for the QSAR study because no data is wasted. A classifier's accuracy on a given dataset is the percentage of correctly predicted examples by the classifier. In this study, 10-fold cross-validation was used. 10 accuracy values were generated, and the mean with the standard deviation was reported[37, 50].

Attribute reduction

Pearson correlation coefficient (r) can be used for estimating the linear correlation between attributes (chemical descriptors)[37]:

$$r = \frac{cov(A_i, A_j)}{\sigma_{A_i}\sigma_{A_j}} \qquad (10)$$

where, cov and σ are the covariance and standard deviation, respectively. The selection of chemical descriptors is an essential part of the QSAR study. The correlation coefficient can be a proper measurement for removing redundant chemical descriptors with a strong correlation. Attribute reduction is valuable, particularly in large datasets, to save computational time and attribute selection techniques such as genetic algorithms[51-53].

Simulation

The QSAR Daphnia Magna Toxicity Dataset, including numerical values for the eight chemical descriptors of the 546 molecules with the quantitative response (LC_{50}), was used to predict the acute aquatic range toxicity towards Daphnia Magna. In toxicology, LC_{50} (lethal concentration, 50%) is the concentration of a chemical, given over a period of time, that causes the death of an individual organism[54]. For further evaluation of the approach used in this study, another dataset named QSAR Fathead Minnow Fish Toxicity Dataset was used. This dataset has six chemical descriptors of the 908 molecules with the quantitative response (LC_{50})[55]. Both of the datasets were downloaded from the UC Irvine Machine Learning Repository [56]. In **Figures 3** and **4,** the quantitative response (LC_{50}) is illustrated for two datasets by histograms, a type of bar plot for numeric data that group the data into bins. 10 bins were used. RapidMiner Studio version 9.7 with the educational license was used throughout this study.

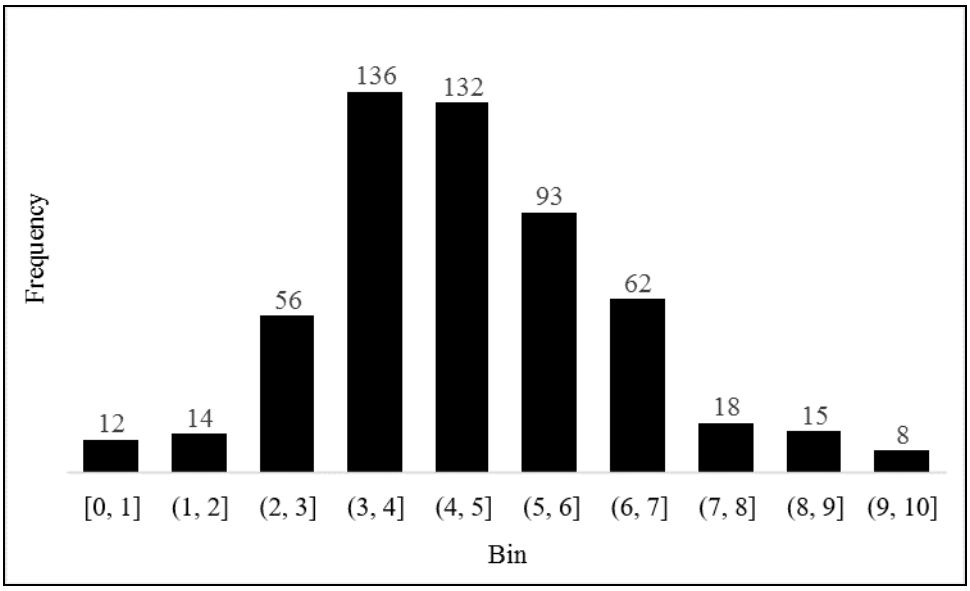


Figure 3. Histograms of the quantitative response (LC_{50}) of the QSAR Daphnia Magna Toxicity Dataset for 10 bins

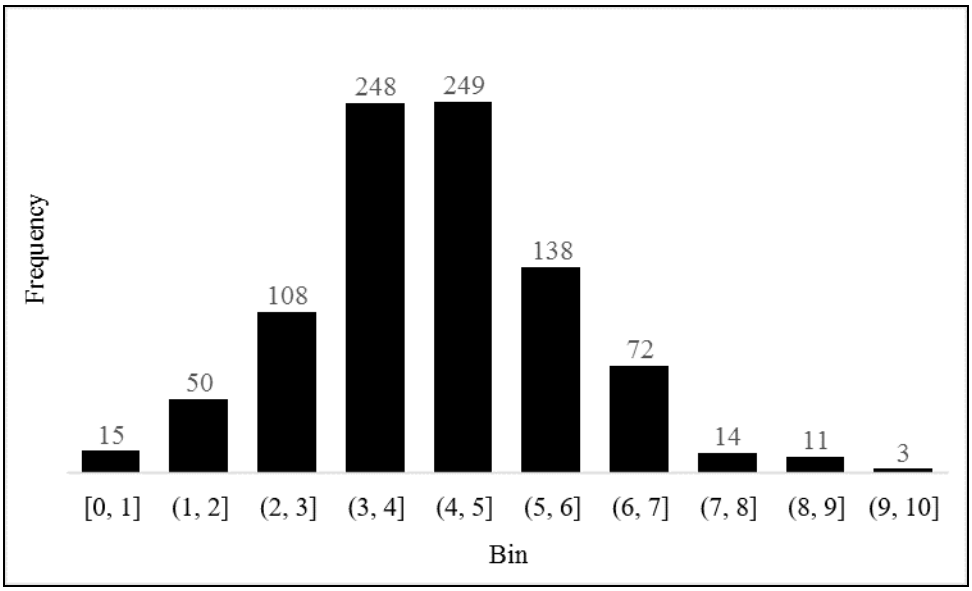


Figure 4. Histograms of the quantitative response (LC_{50}) of the QSAR Fathead Minnow Fish Toxicity Dataset for 10 bins

Results and Discussion

Balancing dataset

Data distribution before and after balancing is shown in **Figure 5**. The balancing algorithm used in this study resulted in oversampling the minority class and under-sampling the majority class. The details of the calculation for balancing the two datasets are reported in **Tables 2** and **3**.

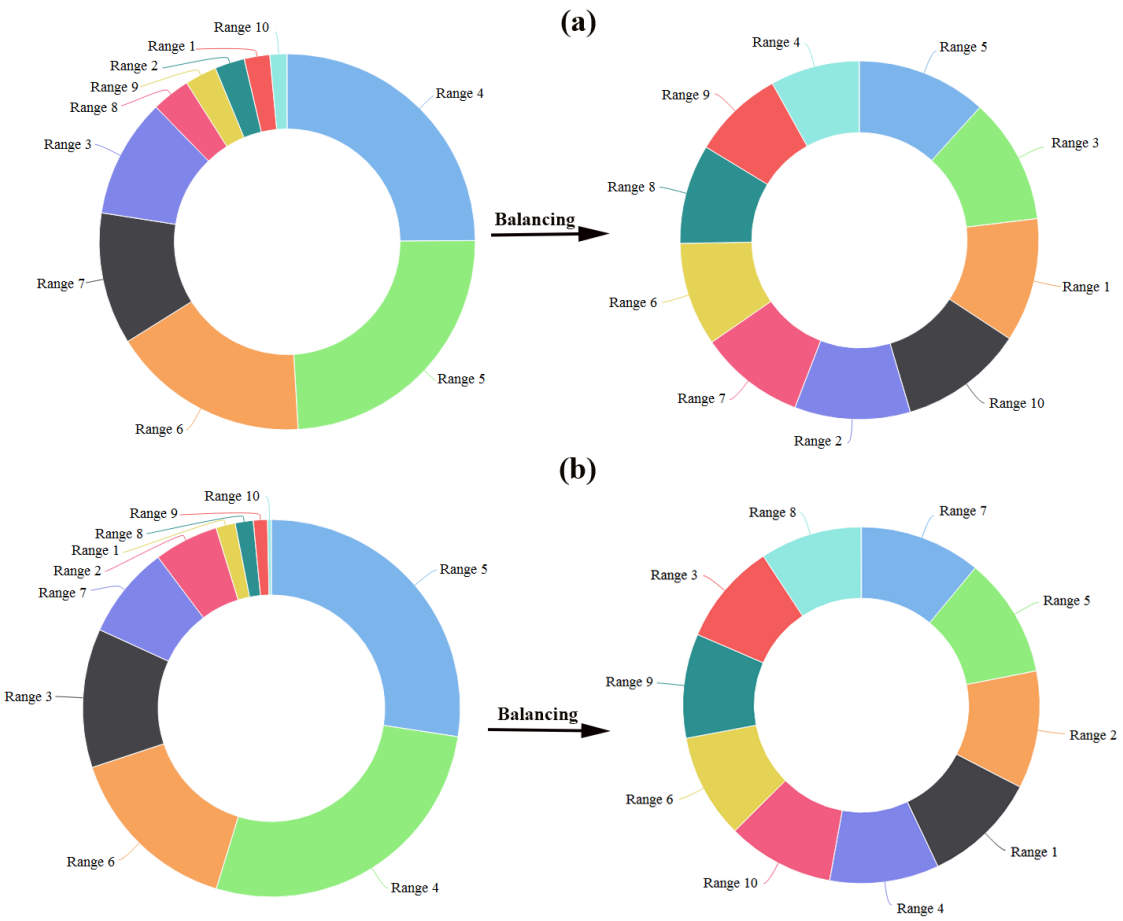


Figure 5. Transformation of the imbalanced dataset to a balanced one for the Daphnia Magna (a) and Fathead Minnow Fish (b) Toxicity Dataset

Table 2. balancing the dataset for the Daphnia Magna dataset

Range	bin	weight	Count (Before balancing)	Weight* Count	Count (After balancing)
1	[0-1]	0.008333	12	0.1	61
2	(1-2]	0.007142	14	0.1	57
3	(2-3]	0.001785	56	0.1	62
4	(3-4]	0.000735	136	0.1	44
5	(4-5]	0.000757	132	0.1	64
6	(5-6]	0.001075	93	0.1	51
7	(6-7]	0.001612	62	0.1	52
8	(7-8]	0.005555	18	0.1	49
9	(8-9]	0.006666	15	0.1	45
10	(9-10]	0.0125	8	0.1	61
			Sum = 546	Sum = 1	Sum = 546

Table 3. balancing the dataset for the Fathead Minnow Fish Dataset

Range	bin	weight	Count (Before balancing)	Weight* Count	Count (After balancing)
1	[0-1]	0.006666	15	0.1	94
2	(1-2]	0.002	50	0.1	97
3	(2-3]	9.2592592	108	0.1	84
4	(3-4]	4.0322580	248	0.1	90
5	(4-5]	4.0160642	249	0.1	99
6	(5-6]	7.2463768	138	0.1	86
7	(6-7]	0.0013888	72	0.1	100
8	(7-8]	0.0071428	14	0.1	84
9	(8-9]	0.0090909	11	0.1	86
10	(9-10]	0.0333333	3	0.1	88
			Sum = 908	Sum = 1	Sum = 908

Classification

As reported in **Table 4**, the classification is carried out for the different classifiers for the Daphnia Magna dataset before and after balancing. According to the results, balancing the datasets significantly enhanced the overall accuracy and decreased the standard deviation. Another dataset named QSAR Fathead Minnow Fish Toxicity Dataset was used to evaluate the generality of the approach used in this study. The results are reported in **Table 5**, indicating that balancing the dataset improved the classification's overall accuracy.

As reported in **Tables 4** and **5**, before balancing the dataset, Naïve Bayes had weaker performance in multi-class classification than other classifiers used in this study; however, after balancing the dataset, the overall accuracy was improved. Additionally, Kernel functions improved the overall accuracy of the naïve Bayes. As illustrated in **Figure 6,** the naïve Bayes classifier's accuracy depends on the number of bins. An increase in the number of bins decreases the accuracy. Generally, using a higher number of bins in the multi-class classification takes advantage of toxicity range prediction in smaller scopes. Kernelled naïve Bayes had an excellent performance in the 10 bins classification, but the naïve Bayes classifier's accuracy decreased with increasing the number of bins.

Table 4. The overall accuracy of the different classifiers for the Daphnia Magna dataset

Classifier	Balanced dataset?	Accuracy	Improvement
Decision Tree	No	34.27±4.72%	58.03%
Decision Tree	Yes	92.3±3.26%	
K-NN	No	37.74±4.49%	55.08%
K-NN	Yes	92.82±2.81%	
NB	No	22.55±4.83%	9.09%
NB	Yes	31.64±8.31%	
Kernelled NB	No	27.23±4.63%	72.48%
Kernelled NB	Yes	99.71±0.53%	
SVM	No	39.22±5.24%	53.05%
SVM	Yes	92.27±3.82%	

Table 5. The overall accuracy of the different classifiers for the Fathead Minnow Fish dataset

Classifier	Balanced dataset?	Accuracy	improvement
Decision Tree	No	43.18±5.83%	54.23%
Decision Tree	Yes	97.41±0.51%	
K-NN	No	42.62±5.46%	55.17%
K-NN	Yes	97.79±0.59%	
NB	No	30.94±3.32%	22.46%
NB	Yes	53.4±1.37%	
Kernelled NB	No	39.48±4.55%	59.56%
Kernelled NB	Yes	99.04±0.74%	
SVM	No	44.06±3.4%	53.52%
SVM	Yes	97.58±0.59%	

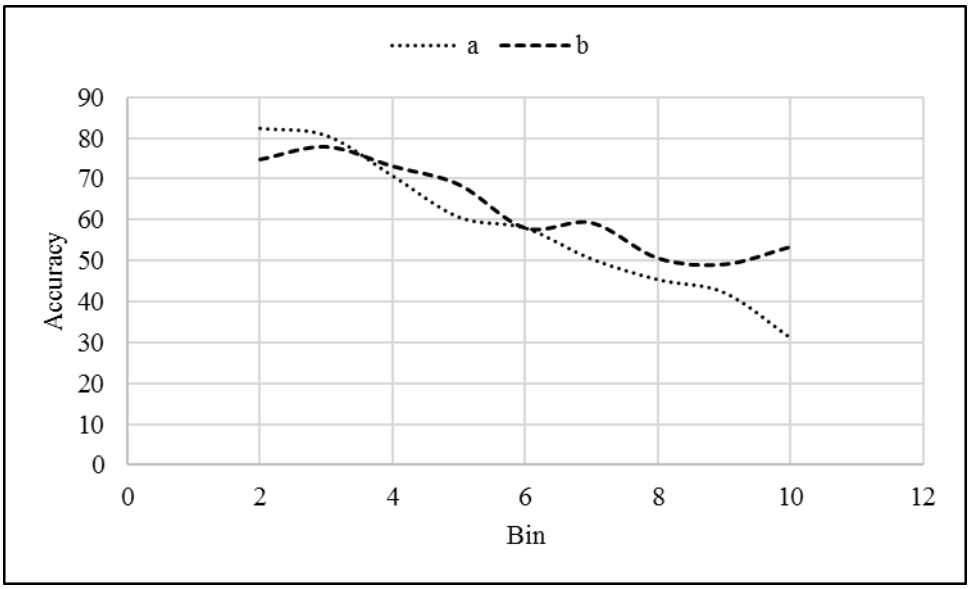


Figure 6. Effect of the number of bins on the naïve Bayes classifier's accuracy before balancing the Fathead Minnow Fish dataset (a) and after balancing the Fathead Minnow Fish dataset (b)

Attribute reduction

As can be seen in **Table 6**, based on the attribute reduction method illustrated in section **2.4**, the overall accuracy did not change with the five attributes (removing three redundant attributes with correlation coefficients greater than 0.5 for the Daphnia Magna dataset). In the literature, attributes with a correlation greater than 0.8 were usually excluded[52]. Removing attributes with the correlation coefficients less than 0.8 usually weakened the strength of the classifier's performance. However, in this study, the balancing the dataset technique results in excellent accuracy even with the attribute reduction indicating the robustness of the approach. Attribute reduction was also tested on the QSAR Fathead Minnow Fish Toxicity Dataset. Three redundant attributes with a correlation coefficient greater than 0.2 were removed. Overall accuracy still revealed robustness. Therefore, balancing the dataset proved its advantage in improving the strengths of the multi-class prediction generally. Among the Daphnia Magna dataset, the MLOGP and RDCHI molecular descriptors encode the information about narcosis. SAacc, TPSA, H-050, and nN

accounts for hydrogen bonding. C-040 encode information about the electrophilic features of the molecules. GATS1p encodes information on molecular polarizability. Thus, there are four groups of descriptors. Removing four redundant attributes with correlation coefficients greater than 0.5 in the Daphnia Magna dataset leads to selecting one attribute from each molecular descriptor group. As a result, MLOGP, TPSA, C-040, and GATS1p molecular descriptors are sufficient to predict molecules' toxicity with an overall accuracy of 99.04%. In the Fathead Minnow Fish dataset, there are six molecular descriptors, MLOGP and GATS1i encode the information about narcosis, CIC0, and SM1_Dz(Z) encode information regarding heteroatoms, and NdssC and NdsCH account for a variety of functional groups with double bonds. Thus, there are three groups of descriptors. Removing three redundant attributes with correlation coefficients greater than 0.2 in the Fathead Minnow Fish dataset leads to selecting one descriptor from each group. In this regard, GATS1i, C0, and NdsCH molecular descriptors can predict molecules' toxicity with an overall accuracy of 92.9%.

Table 6. The Kernelled NB classifier's overall accuracy for two datasets after removing correlated attributes before and after balancing the datasets

Dataset	Before balancing	After balancing
Daphnia Magna	27.71%±4.66%	99.04%±1.2%
Fathead Minnow Fish	37.83%±4.62%	92.9%±1.38%

Comparison of QSAR models

The overall efficiency of the QSAR models depends on the molecular descriptors and classifiers. The ultimate performance of the classification methods is reported in **Table 7**. As listed in **Table 7**, Kernelled naïve Bayes using a balanced dataset increased the overall accuracy, indicating balancing the dataset as a preprocessing step in machine learning can enhance the accuracy of the toxicity prediction. Although the datasets are different in these studies, comparing the toxicity prediction results before and after balancing the dataset (as presented in **Tables 4** and **5**) indicates that balancing the dataset can dramatically improve the accuracy.

Table 7. Classification accuracies of toxicity from various studies reported in the literature.

Method	Accuracy	Year	References
MultiCASE, multiple computer automated structure evaluation	81.6%	2008	[57]
linear discriminant analysis	80%	2011	[58]
Naïve Bayes	95%	2017	[28]
Artificial Neural Network based on MACCS fingerprints	83.9%	2018	[27]
support vector machine based on MACCS fingerprints	84.9%	2019	[59]
Bayesian Network	80%	2020	[30]
Naïve Bayes	91.8%	2020	[29]
Kernelled naïve Bayes using balanced dataset	99%		This study

Conclusion

Specifying different weights to training datasets and then bootstrapping them resulted in the balanced dataset. Two datasets and five classifiers were used to demonstrate the importance of balancing the dataset in enhancing the accuracy of the toxicity material classification. This approach increased the overall accuracy up to 99%. In previous studies, increasing the maximum number of bins for the

material toxicity classification decreased the overall accuracy. However, this approach increased the overall accuracy for 10 bins. Kernelled naïve Bayes had better performance than naïve Bayes in the classification. Removing four attributes from Daphnia Magna and three attributes from the Fathead Minnow Fish dataset did not change the overall accuracy because of balancing the dataset, indicating this approach's robustness. Therefore, this study revealed the importance of balancing the dataset on the accuracy of the toxicity prediction in the multi-class classification, suggesting that the preprocessing step always plays an essential role in machine learning.

References

[1] M. Rausand, *Risk assessment: theory, methods, and applications,* John Wiley & Sons, **2013**.

[2] G. Popov, B.K. Lyon, B. Hollcroft, *John Wiley & Sons*, **2016**.

[3] Development (OECD) Staff, Development. *Working Party on Environmental Performance, & United Nations. Economic Commission for Europe. Committee on Environmental Policy.* OECD environmental performance reviews: Canada, **2004**.

[4] CEC, Regulation (EC) No. 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), EU CEC Brussels, **2006**.

[5] M.T. Cronin, J.D. Walker, J.S. Jaworska, M.H. Comber, C.D. Watts, A.P. Worth, *Environ. Health Perspect.,* **2003**, *111*, 1376–1390.

[6] R. Combes, C. Grindon, M.T. Cronin, D.W. Roberts, J.F. Garrod, *Alt. Lab. Anim.,* **2008**, *36*, 45–63.

[7] M. Cronin, *Chemical Toxicity Prediction: Category Formation and Read-Across*, Royal Society of Chemistry, **2013**, pp 155–167.

[8] H. Liu, E. Papa, P. Gramatica, *Chem. Res. Toxicol.*, **2006**, *19*, 1540–1548.

[9] T.M. Mitchell, *Machine Learning*, McGraw-Hill, **1997**.

[10] D.E. Jones, H. Ghandehari, J.C. Facelli, *Comput. Methods Programs Biomed.*, **2016**, *132*, 93–103.

[11] H. Yang, L. Sun, W. Li, G. Liu, Y. Tang, *Front. Chem.*, **2018**, *6*, 30.

[12] A.A. Toropov, A.P, Toropova, I. Raska Jr, D. Leszczynska, J. Leszczynski, *Comput. Biol. Med.,* **2014**, *45*, 20–25.

[13] S. Schmidt, M. Schindler, D. Faber, J. Hager, *SAR QSAR Environ. Res.*, **2021**, *32*, 151–174.

[14] O. Tinkov, V.Y. Grigorev, A.N. Razdolsky, L.D. Grigoryeva, J.C. Dearden, *SAR QSAR Environ. Res.*, **2020**, *31*, 615–641.

[15] F. Lunghini, G. Marcou, P. Azam, M.H. Enrici, E. Van Miert, A. Varnek, *SAR QSAR Environ. Res.*, **2020**, *31*, 655–675.

[16] M. Marzo, G.J. Lavado, F. Como, A.P. Toropova, A.A. Toropov, D. Baderna, C. Cappelli, E. Benfenati, *SAR QSAR Environ. Res.*, **2020**, *31*, 227–243.

[17] J. Polanski, *Chemoinformatics: From Chemical Art to Chemistry in Silico*, Elsevier, **2019**, pp 601–618.

[18] S.R. Kazmi, R. Jun, M.S. Yu, C. Jung, D. Na, *Comput. Biol. Med.*, **2019**, *106*, 54–64.

[19] Y. Wu, G. Wang, *Int. J. Mol. Sci.*, **2018**, *19*, 2358.

[20] S.J. Russell, P. Norvig, *Artificial Intelligence-A Modern Approach*, 3rd Ed., Pearson Education: London, **2010**.

[21] E. Alpaydin, *Introduction to machine learning*, MIT press, **2020**.

[22] M. Mohri, A. Rostamizadeh, A. Talwalkar, *Foundations of Machine Learning*, MIT Press, **2018**.

[23] S. Cassani, S. Kovarich, E. Papa, P.P. Roy, L. van der Wal, P. Gramatica, *J. Hazard. Mater.*, **2013**, *258*, 50–60.

[24] F. Abbasitabar, V. Zare-Shahabadi, *Chemosphere*, **2017**, *172*, 249–259.

[25] B. Giner, C. Lafuente, D. Lapeña, D. Errazquin, L. Lomba, *Ecotoxicol. Environ. Saf.*, **2020**, *191*, 110004.

[26] R. Aalizadeh, C. Peter, N.S. Thomaidis, *Environ. Sci. Process Impacts*, **2017**, *19*, 438–448.

[27] T. Fan, G. Sun, L. Zhao, X. Cui, R. Zhong, *Int. J. Mol. Sci.*, **2018**, *19*, 3015.

[28] H. Zhang, P. Yu, J.X. Ren, X.B. Li, H.L., Wang, L. Ding, W.B. Kong, *Food Chem. Toxicol.*, **2017**, *110*, 122–129.

[29] H. Zhang, C. Shen, R.Z. Liu, J. Mao, C.T. Liu, B. Mu, *J. Appl. Toxicol.*, **2020**, *40*, 1198–1209.

[30] A. Lillicrap, S.J. Moe, R. Wolf, K.A. Connors, J.M. Rawlings, W.G. Landis, S.E. Belanger, *Integr. Environ. Assess. Manag.*, **2020**, *16*, 452–460.

[31] J. Roy, P.K. Ojha, E. Carnesecchi, A. Lombardo, K. Roy, E. Benfenati, *J. Hazard. Mater.*, **2020**, *386*, 121660.

[32] N. Abramenko, L. Kustov, L. Metelytsia, V. Kovalishyn, I. Tetko, W. Peijnenburg, *J. Hazard. Mater.,* **2020**, *384*, 121429.

[33] L. Yang, Y. Wang, J. Chang, Y. Pan, R. Wei, J. Li, H. Wang, *Chemosphere*, **2020**, *258*, 127217.

[34] K. Khan, P.M. Khan, G. Lavado, C. Valsecchi, J. Pasqualini, D. Baderna, E. Benfenati, *Chemosphere*, **2019**, *229*, 8–17.

[35] S.K. Pandey, P.K. Ojha, K. Roy, *Chemosphere*, **2020**, *252*, 126508.

[36] X. Jin, M. Jin, L. Sheng, *Comput. Biol. Med.*, **2014**, *51*, 205–213.

[37] J. Han, J. Pei, M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier Science, ProQuest Ebook Central, **2011**.

[38] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*, 2nd Ed., IEEE: Wiley, **2019**.

[39] A. Pérez, P. Larrañaga, I. Inza. *Int. J. Approx. Reason.*, **2009**, *50*, 341–362.

[40] G.H. John, P. Langley, *Estimating continuous distributions in Bayesian classifiers.* Morgan Kaufmann: Montr´eal, **1995**, pp 338–345.

[41] K.P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press: Cambridge, **2012**.

[42] L. Michielan, L. Pireddu, M. Floris, S. Moro, *Mol. Inform.*, **2010**, *29*, 51–64.

[43] V. Kotu, B. Deshpande, *Predictive analytics and data mining: concepts and practice with rapidminer*, Elsevier Science, Morgan Kaufmann: USA, **2014**.

[44] N. Japkowicz, S. Stephen, *Intell. Data Anal.*, **2002**, *6*, 429–449.

[45] S. Barua, M.M. Islam, X. Yao, K. Murase, *IEEE Trans. Knowl. Data Eng.*, **2012**, *26*, 405–425.

[46] C.X. Ling, C. Li. *Data mining for direct marketing: Problems and solutions*, AAAI Press**:** New York, **1998**, pp 73–79.

[47] A. Fernández, S. Garcia, F. Herrera, N.V. Chawla, *J. Artif. Intell. Res*, **2018**, *61*, 863-905.

[48] Z.H. Zhou, X.Y. Liu, *Comput. Intell.*, **2010**, *26*, 232–257.

[49] A. Fernández, V. López, M. Galar, M.J. Del Jesus, F. Herrera, *Knowl.-Based Syst.*, **2013**, *42*, 97–110.

[50] D.M. Hawkins, S.C. Basak, D. Mills, *J. Chem. Inform. Comput. Sci.*, **2003**, *43*, 579–586.

[51] S.K. Jha, T.H. Yoon, Z. Pan, *Comput. Biol. Med.***,** **2018**, *99*, 161–172.

[52] A. Rácz, D. Bajusz, K. Héberger, *Mol. Inform.*, **2019**, *38*, 1800154.

[53] R. Todeschini, V. Consonni, *Handbook of molecular descriptors,* Wiley-VCH: Weinheim, **2008**.

[54] M. Cassotti, D. Ballabio, V. Consonni, A. Mauri, I.V. Tetko, R. Todeschini, *Altern. Lab. Anim.*, **2014**, *42*, 31–41.

[55] M. Cassotti, D. Ballabio, R. Todeschini, V. Consonni, *SAR QSAR Environ. Res.*, **2015**, *26*, 217–243.

[56] D. Dua, C. Graff, *UCI Machine Learning Repository*. School of Information and Computer Science, University of California, Irvine, CA, USA. **2019**.

[57] G.E. Jensen, J.R. Niemelä, E.B. Wedebye, N.G. Nikolov, *SAR QSAR Environ. Res.*, **2008**, *19*, 631–641.

[58] M.T. Martin, T.B. Knudsen, D.M. Reif, K.A. Houck, R.S. Judson, R.J. Kavlock, D.J. Dix, *Biol. Reprod.* **2011**, *85*, 327–339.

[59] C. Jiang, H. Yang, P. Di, W. Li, Y. Tang, G. Liu, *J. Appl. Toxicol.*, **2019**, *39*, 844–854.

HOW TO CITE THIS ARTICLE

Ali Ekramipooya, Davood Rashtchian, Mehrdad Boroushaki*. Kernelled Naïve Bayes Using a Balanced Dataset for Accurate Classification of the Material Toxicity. Adv. J. Chem. A., 2021, 4(2), 138-151.

DOI: 10.22034/AJCA.2021.274570.1244
URL: http://www.ajchem-a.com/article_128746.html