International Journal of Combinatorial Optimization Problems and Informatics, 13(2), May-Aug 2022, 58–64. ISSN: 2007-1558

www.editada.org

A Design and analysis of classification models for the gelification of alkoxybenzoates using the *k*NN algorithm

Virginia Loredo-Pong, María Lucila Morales-Rodríguez, Nancy Patricia Díaz-Zavala, Nelson Rangel-Valdez, Jaime E. Sosa-Sevilla*

Centro de Investigación en Petroquímica, División de Estudios de Posgrado e Investigación. Tecnológico Nacional de México-Instituto Tecnológico de Ciudad Madero. Prolongación Bahía de Aldair, Ave. De las Bahías, Parque de la Pequeña y Mediana Industria, 89600, Altamira, Tamaulipas, México.

*Correspondence: vlpong83@gmail.com

Abstract. Classification models of the states produced in the gelation tests of organic molecules require designing several corpora of data based on their characteristics. This work studies 15 solvents characterized by Hansen Solubility Parameters and a series of alkoxybenzoates. Characterization of the alkoxybenzoates has as distinctive feature the number of carbons on its alkyl tails. Solvent and molecule properties were evaluated as attributes on corpus through the kNN algorithm. Three corpora were tested in different algorithm configurations, varying each corpus content according to the solvents and molecules attributes. Relevance of some attributes over others on the performance prediction of the products class can be appreciated. Significant instances were correctly classified on corpora when the HSP and the alkoxybenzoate alkyl ether tail length were considered, thus, stating the influence of these properties on classification accuracy. The most suitable configurations in kNN as metric, k value, and attribute weight were determined according to each corpus.

Keywords: machine learning, predictive models, alkoxybenzoates, HSP.

Article Info

Received: August 31, 2021
Accepted: October 23, 2021

1. Introduction

Organogels are thermally reversible viscoelastic solid-like materials composed of an organic liquid and small quantities of low molecular mass molecules. These materials have gained much interest and are currently highly investigated because of the wide variety of applications that organogels have shown.

Organogel molecules can entrap organic liquids through self-assembly in a solid three-dimensional network and return to the solution state upon heating. The diverse applicability of these materials includes electronics, food, drugs, plastics, contaminant removal agents like fuels [1],[2],[3].

Many molecules with different chemical structures capable of congealing organic solvents and acting as organogelators have been reported both in the industry and research fields; however, the methods to know a priori the suitable design of gelators and the solvents to match and produce the gel are shallow. Designing a molecule is still a restricted and tedious limited-step process to achieve the gel state. Generally, they are discovered experimentally modulating slight changes in the chemical structure, testing several solvents, and changing the concentration ratio.

Based on the previously mentioned weakness, many analog molecules within a chemical structure design fail to congeal. In order to understand the lack of gel formation, several studies report the relevance of the non-covalent forces balance leading to self-assembly. It has been reported that slight modifications on the number of carbons in alkyl tails can interrupt gel formation due to the fragile balance of these forces [4], [5], [6].

Several approaches have been used to elucidate the design requirements of an efficient organogelator. Most attempts use artificial intelligence algorithms to process databases with suitable molecular properties to represent aggregation phenomena. Some of these computational tools relate specific, measurable properties to a given physical product state obtained from a chemical experiment, based on the principle that these products depend on the components' molecular properties [7].

Commonly, the properties used to describe a physicochemical phenomenon like gelification have different dimensions; zero-dimensional (like molecular weight), uni-dimensional (sum of specific molecular fragments), and bi-dimensional (number and types of atoms describing molecular constitution). Machine learning methods can process the databases designed with these properties to classify the aggregation state produced with a given molecule [8].

Hansen Solubility Parameters (HSP) are molecular properties currently used as attributes in gelification prediction models [9], [10], [11]. The HSP are matter cohesion energy parameters; they can represent the physical interactions that maintain gelator molecules self-assembled in a gel. Diehn [12] recently reported the products formed with 1,3: 2,4-dibenzylidene sorbitol (DBS) in a set of solvents, linking the solvents HSP with particular moieties of the gelator. The energy cohesion is decomposed according to three contributions: dispersive interactions, polar interactions, and H-bonds [7]. Every interaction contribution of solvent and gelator HSP values can be related to molecular behavior and products obtained in the gelification test.

2. Design of a predictive model

In computational intelligence, classification is an area that searches patterns in data corpora [13]. One of the challenges of creating classification or predictive models is defining the appropriate tool and its configuration to produce a precise assignation of a given target. Predictive analytics take databases through data mining and Machine Learning to search and modulate the specific characteristics needed in an optimal model [14]. Machine Learning includes artificial intelligence algorithms with the necessary skills to identify qualitative and quantitative data patterns and model the representative phenomena, in this case, a chemical nature phenomenon the gelification.

The supervised algorithms in machine learning consist in making predictions based on stored data in training sets. These algorithms can make predictions based on labeled data characterized by a series of specific attributes. A dataset trains the algorithm with input information that teaches it to search for patterns in the values given and correlate them to a specific label [15][16].

The chosen algorithm for this study case was *k*NN. This nonparametric classification algorithm estimates an element X to belong to a class C from the prototype set. Training instances build vectors in a multidimensional space; p attributes describe every instance considering q classes. The attribute values of the ith example (where $1 \leq i \leq n$) are represented by the p-dimensional vector $X_i = (X_{1i}, X_{2i}, ..., X_{pi})$. The multidimensional space is partitioned into regions by locations and labels of the training instances; if a point in the space belongs to the closest most frequent class among the k neighbors in the training instances, it is assigned with the class C. *k*NN assumes that the closest neighbors select the best class for every instance using all the attributes. However, this assumption does not take into account the possibility of having irrelevant attributes dominating the ranking, causing a misclassification because the relevant attributes could lose weight among a set of irrelevant attributes.

To correct a possible slant, the attributes are assigned by weight in two different types: by the distances of each attribute, preselecting attributes, or by uniform weight, where all attributes have equal relevance [17].

An optimal k selection depends on the type and quantity of instances. Large k values could produce noise in classification but are beneficial in establishing limits between similar classes. Commonly, the adequate k values are selected by optimization in the model.

Evaluation methods are used to corroborate that the algorithm is working correctly concerning the data sets designed and applied in *k*NN. The evaluation method has to demonstrate adequate prediction capability; to accomplish this, learning algorithms are evaluated and compared by cross-validation dividing data into segments named folds, one used to train and the others to test. The training and test sets cross over in successive rounds, so each data point has a chance of being validated [14]. Cross-validation gives representative results about the models' ability and the suitable configurations to classify new instances unknown to the algorithm.

The present work evaluates a series of alkoxybenzoate derivates and their aggregation behavior. Uni-dimensional (moieties of the alkoxybenzoates structure) and bi-dimensional (Hansen solubility parameters of solvents and alkoxybenzoates) properties to characterize gelification components were defined for the study. The contribution of the properties is evaluated by designing a series of corpora and validating them on the *k*NN algorithm, trying different configurations to fit the optimum model combination.

3. Experimental procedures

3.1 Corpora design

The studied alkoxybenzoates were previously synthesized and submitted to gelification tests to define the aggregation state produced with every tested solvent. The studied moieties of the Alkoxybenzoates derivatives consist of a 12 carbon ether alkyl tail and a variable ester chain of one (1-12), three (3-12), and four (4-12) carbons attached to a p-substituted aromatic ring. The produced aggregation states were gel (G), solution (S), and precipitate (P); these were the defined target classes for every instance.

The attributes used to characterize the data sets were alkoxybenzoates moieties, ether carbon number (12), and ester carbon number (1, 3, 4), the HSP of all species of both solvents and derivatives (H-bond interactions, polar interactions, and dispersive interactions). Three different corpora structures were designed, changing the attributes in each one. Table 1 shows the corpora created and their attribute content.

Table 1. Corpora designed and their attributes

Corpus	Attributes	
	Alkoxybenzoates	**Solvents**
A	Ether carbon number Ester carbons number	Dispersive I. H-bond Polar I.
B	Dispersive I. H-bond Polar I.	
C	Dispersive I. H-bond Polar I. Ether carbons Ester carbons	

Every corpus is composed of 45 instances. The number of instances distributed according to their class is gel (G)-16, solution (S)-27, and precipitate (P)-2.

3.1. Configuration hyperparameter values

The corpora assessment was cross-validation with three stratified folds. The test sets consist of 15 and 30 instances in the training sets.

Table 2. Configuration hyperparameters values tested in the kNN algorithm by cross-validation

Corpus	k	Metric	Attribute weight
A B C	3 5	Euclidean Chebyshev	Uniform By distance

The chosen metrics for the neighbor distance estimation were Euclidean and Chebyshev. To evaluate the suitable assignation of attributes weight, two types were estimated: uniform and by distance. The values of k neighbors appointed were 3 and 5. Every variable in this configuration was tested for each corpus, see Table 2.

4. **Results**

4.1 Configuration evaluation

The performance results of configurations in kNN are presented in %CA (classification accuracy percent). This value represents the ratio of predicted and actual classes. The corpora were evaluated in *k*NN according to the proposed configurations changing *k* value, attribute weight, and metric. The results are shown in Table
Table *3*.

Table 3. % Classification Accuracy achieved in kNN with the proposed configurations

Metric	**Attribute weight**	***k***	**%CA**		
			A	**B**	**C**
Euclidean	Distance	3	80	80	80
		5	82	80	85
	Uniform	3	82	85	82
		5	87	87	87
Chebyshev	Distance	3	80	76	80
		5	87	76	87
	Uniform	3	82	80	82
		5	85	82	85

The highest values of %CA were obtained when applying the Euclidean metric with a uniform weight assigned for the attributes, a *k* = 5 value, and no influence of the type of corpus; this can be observed in Table 3.

According to the weight attribute, establishing a uniform relevance for all attributes offers a higher %CA in all cases rather than applying the distance attribute to preselect features. A comparison of the *k* values tested and their performance on the classification accuracy is shown in Figure 1.

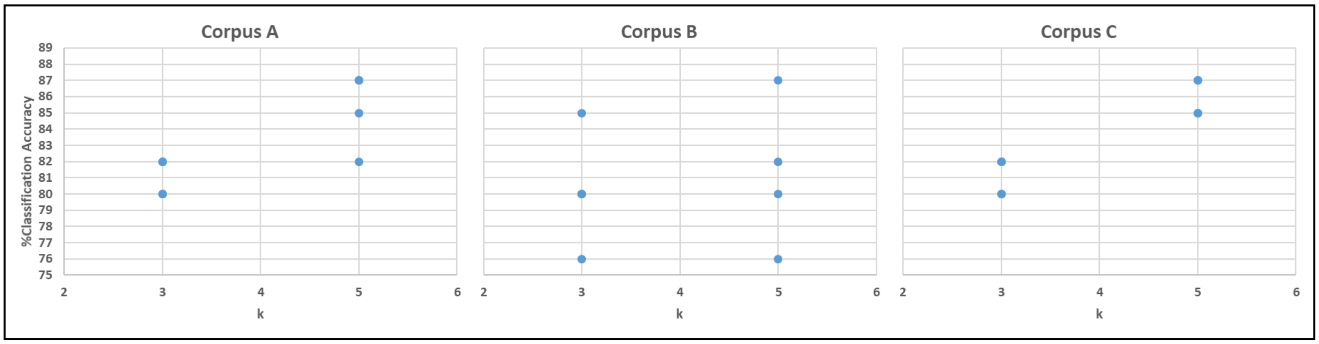


Fig. 1. Comparison of the performance of *k* neighbors in the precision of the classification for each corpus

The performance in Figure 1 indicates that a value of *k*=5 neighbors is required to obtain the higher classification values with every corpus, and this applies to all configurations. This effect was observed when no attributes were discriminated against, and more neighbors were selected. The algorithm was not affected by wrong value noise, which causes misclassification of the instances; it was observed that the algorithm could choose the correct class in a range of neighbors broader than 3, which is more notable in corpus A and C results.

Statistically, the characterization of corpus C produces a higher classification compared to the other two corpora. C is the corpus that possesses the total attributes tested, suggesting that the composition of C is the adequate studied corpora for classification purposes.

Concerning the metric, the Chebyshev equation application shows the same %CA values in all cases for configurations using A and C corpora. The lower performance of *k*NN is due to corpus B. This corpus lacks ether and ester attributes in its

composition; therefore, its low performance can be associated with this absence, suggesting that these attributes raise classification accuracy.
It was observed that classification performance according to the weight for attributes is influenced by the value of *k* and the corpus structure. This effect was detected by comparing configurations with the maximum CA values of each metric, observing a fluctuation according to A, B, and C corpora, and k values.

Also, the Euclidean metric produces higher classification values assigning uniform weight according to the metrics development in conjunction with attributes weight. Lower values in the corpus B results were obtained for the Chebyshev metric, contrasting with those obtained by the Euclidean metric. Although the classification with corpora A and C produces the same results, except for a value increase when *k*=5. To assess which attribute has a significant contribution effect on the classification, an ANOVA was performed for every corpus.

4.2 One-way variance analysis (ANOVA)

Table 4 shows the difference between average values of the attributes in different classes, represented by the ANOVA. The results provided by the ANOVA are F-value; these were used to rank the other attributes. The evaluation makes it possible to estimate whether any of the characteristics affect the accuracy of the classification.

Table 4. One-way ANOVA for corpora A, B and C according to the significance of the attributes over the classification

Corpus A		Corpus B		Corpus C	
Attribute	**ANOVA**	**Attribute**	**ANOVA**	**Attribute**	**ANOVA**
Solvent H-bond	5.798	Solvent H-bond	5.798	Solvent H-bond	5.798
Solvent Polar I.	5.325	Solvent Polar I.	5.325	Solvent Polar I.	5.325
Ester	4.128	Alcoxy H-bond	4.103	Ester	4.128
Solvent Dispersive I.	0.625	Alcoxy dispersive I.	4.094	Alcoxy H-bond	4.103
Ether	NA	Alcoxy polar I.	4.091	Alcoxy dispersive I.	4.094
		Solvent Dispersive I.	0.625	Alcoxy polar I.	4.091
				Solvent Dispersive I.	0.625
				Ether	NA

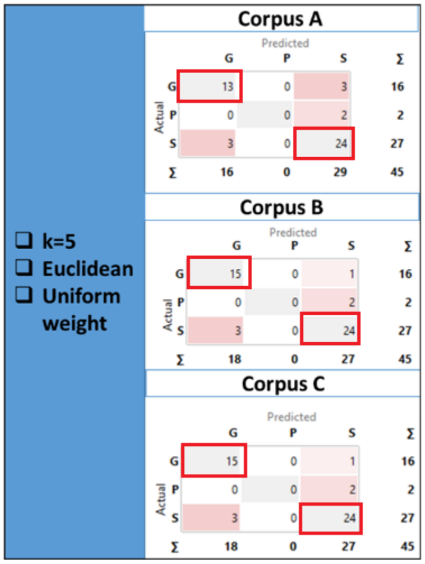
According to observations in Table 4, a significant difference in the average values of all attributes exists, except the Ether one. Two facts can be concluded: in the corpus with the studied structure; the Ether feature is not relevant for classification due to its constant value of 12; furthermore, the other attributes present evidence of affecting differentiating class type. The attributes that contribute significantly to the classification, coinciding for the three corpora, are the H-bond and polar interactions, both from the solvent characterization.

The corresponding feature to the carbons in the alkoxybenzoates ester group, with values of 1, 3, and 4, occupies third place in the corpora containing it. This attribute distinguishes each of the three studied molecules 1-12, 3-12, and 4-12, indicating that features with significant differences in numeric values promote the correct classes assignation. The attributes characterizing alkoxybenzoates: dispersive, H-bond, and polar interactions, are located in the same position inside the corpus. While solvent interactions are found in the three corpora, dispersive interactions present the lowest contribution due to its ANOVA F value, putting it in the last place.

As noted above, two of the solvents' HSP properties and one of the alkoxybenzoates, the number of carbons in the ester group, present the higher significance as attributes for classification. Confusion Matrixes were developed to identify which instances were correctly classified by each class.

4.3 Confusion Matrixes

Figure 2 shows the confusion matrixes of the highest classification values (87%). First of all, a coincidence in the configuration of the three maximum CA can be observed. This configuration applies the Euclidean metric, five neighbors, and uniform weight to all attributes. Corpora B and C produce more actual instances of class G than corpus A. These results show an influence of the alkoxybenzoates HSP as attributes over the classification accuracy because corpus A lacks the presence of these attributes and produces more classification fails of class G.



- ❑ k=5
- ❑ Euclidean
- ❑ Uniform weight

Corpus A

		Predicted			
		G	P	S	Σ
Actual	G	13	0	3	16
	P	0	0	2	2
	S	3	0	24	27
	Σ	16	0	29	45

Corpus B

		Predicted			
		G	P	S	Σ
Actual	G	15	0	1	16
	P	0	0	2	2
	S	3	0	24	27
	Σ	18	0	27	45

Corpus C

		Predicted			
		G	P	S	Σ
Actual	G	15	0	1	16
	P	0	0	2	2
	S	3	0	24	27
	Σ	18	0	27	45

Fig. 2. Confusion matrixes for experiments with the highest classification accuracy (87%) and their configuration

Meanwhile, the number of instances correctly classified as S is the same for the three corpora. However, instances with class P could not be classified accurately; this can be attributed to a minority number of cases within this class, the same ones that the model misclassified as S, as the predominant class in the training set.

5. Conclusions

Data corpora based on structural properties and HSP of gelators and solvents demonstrate a classification capability in gelification prediction models. From the tested configurations, the impact of the different variable values was identified: *k* neighbors, the metric, the corpus structures, and the attributes.

The most advisable configuration identified for its classification results is composed as follows. The Euclidean metric presented superior performance than the Chebyshev metric by producing higher values of %CA (Table 1). The assignation of 5 neighbors and uniform weight for the attributes applied to this equation showed the best performance. For the corpora composition, alkoxybenzoates HSP and ester feature presented an influence raising instance classification with the Gel target. For the case of solvents characterization, H-bond and polar interactions showed higher contributions over classification accuracy. The less significant attributes for classification were: alkoxybenzoates polar interactions and solvents dispersive interactions. The attribute ether with a constant value does not show relevance for classifying the instances. Based on the previous observation, it is demonstrated that using the same value to characterize an instance is not advisable for a classification corpus.

However, less relevant attributes in these corpora are not dismissed for future studies with different alkoxybenzoate families because they were not analyzed as a corpus applied variable of distinct composition. Similarly, the number of instances according to their class in the data sets will be analyzed as a possible factor affecting the classificatory skills of the corpus.

The cross-validation results obtained in this study show advisable configurations of data corpora and algorithm parameters to be applied in a test with data sets composed of new alkoxybenzoate molecules or solvents and achieve high accuracy predictions.

6. Acknowledgements

The authors thank CONACyT for the support granted through the Scholarship for Postgraduate Studies with CVU 660121, and the project Cathedra with number 3058. They also thank the support from Tecnológico Nacional de México-Instituto Tecnológico de Ciudad Madero (TecNM- ITCM) and the project of TecNM number 10772.21-P.

References

1. Weiss, R. and Terech, P.: Molecular gels: Materials with self-assembled fibrillar networks. Springer (2006).
2. Trong, W., Lewis, L., Thordarson, P.: Functional molecular gels. 1st edn. The Royal Society of Chemistry, Cambridge (2014).
3. Weiss, R. The Past, Present, and Future of Molecular Gels. What is the status of the field, and where is it going. *J. Am. Chem. Soc.* 136, 7519–7530 (2014). doi.org/10.1021/ja503363v
4. Iqbal, S., Miravet, J., Escuder, B.: Biomimetic Self-assembly of Tetrapeptides into Fibrillar Networks and Organogels. *Eur. J. Org.* 27, 4580–4590 (2008). doi.org/10.1002/ejoc.200800547
5. Sato, H., Nogami, E., Yajima, T., Yamagishi, A.: Terminal effects on gelation by low molecular weight chiral gelators. *RSC Adv.* 4, 1659–1665 (2014). doi.org/10.1039/C3RA44070B
6. Suzuki, M., Yumoto, M., Shirai, H., Hanabusa, K.: A family of low-molecular-weight organogelators based on Na,N3-diacyl-L-lysine: effect of alkyl chains on their organogelation behavior. *Tetrahedron* 64, 10395–10400 (2008). doi.org/10.1016/j.tet.2008.08.061
7. Raynall, M., Bouteiller, L.: Organogel formation rationalized by Hansen solubility parameters. *Chem. Comm.* 47, 8271–8273 (2011). doi.org/10.1039/C1CC13244J
8. Bonnet, J., Suissa, G., Raynall, M., Bouteiller, L.: Organogel formation rationalized by Hansen solubility parameters: dos and don’ts. *Soft Matter.* 10, 3154–3160 (2014). doi.org/10.1039/C4SM00244J
9. Bonnet, J., Suissa, G., Raynall, M. Bouteiller, L.: Organogel formation rationalized by Hansen solubility parameters: influence of gelator structure. *Soft Matter.* 11, 2308–2314 (2015). doi.org/10.1039/C5SM00017C
10. Diehn, K., Oh, H., Hashemipour, R., Weiss, R., Raghavan, S.: Insights into organogelation and its kinetics from Hansen solubility parameters. Toward a priori predictions of molecular gelation. *Soft Matter.* 10, 2632–2640 (2014). doi.org/10.1039/C3SM52297K
11. Katritzky, A.R., Kuanar, M., Slavov, S., Hall, C.D., Karelson, M., Kahn, I. Dobchev, D.A.: Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chem. Rev.* 10, 5714–5789 (2010). doi.org/10.1021/cr900238d
12. Gupta, J.K., Adams, D.J., Berry, N. G.: Will it gel? Successful computational prediction of peptide gelators using physicochemical properties and molecular fingerprints. *Chem. Sci.* 7, 4713–4719 (2016). doi.org/10.1039/C6SC00722H
13. Pedrycz, W., Martínez, L., Espin-Andrade, R. A., Rivera, G., Gómez, J. M (Eds.): *Computational Intelligence for Business Analytics*. Springer, Cham, 2021. doi.org/10.1007/978-3-030-73819-8
14. Rivera, G., Florencia, R., García, V., Ruiz, A., Sánchez-Solís, J. P: News classification for identifying traffic incident points in a Spanish-speaking country: A real-world case study of class imbalance learning. *Applied Sciences* 10(18), 6253 (2020). doi.org/10.3390/app10186253
15. García, V., Sánchez, J. S., Marqués, A. I., Florencia, R., & Rivera, G.: Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data. *Expert Systems with Applications* 158 (2020): 113026. doi.org/10.1016/j.eswa.2019.113026
16. Cambronero, C. G., Moreno, I. G.: Algoritmos de aprendizaje: KNN & KMEANS. *Inteligencia en Redes de Telecomunicación*. 1, 1–8 (2016).
17. Hodges, J.L.: An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges. *Int. Stat. Rev. Revue Internationale de Statistique* (1951).