UNIVERSITY OF CALGARY | *Program for Undergraduate Research Experience*

FINAL REPORT AND REFLECTION

"*Prediction of Emulsion Phase Equilibria with Machine Learning Classification Models"*

Rylan Marianchuk

Project Duration: May 1, 2019 - Aug 28, 2019.

Supervisor Name: Dr. Sonny Chan

INTRODUCTION

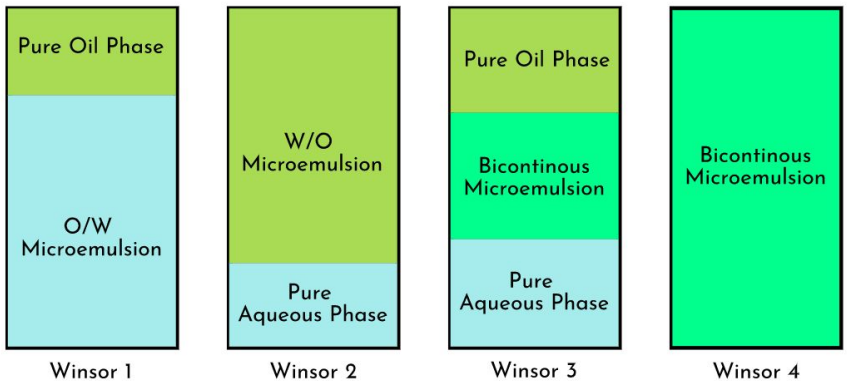
Annually, the international Genetically Engineered Machine (iGEM) competition brings teams across the world to present projects that apply the principles of genetic engineering and synthetic biology to real-world challenges. The themes of the competition are open ended, allowing teams to tackle any problem facing a chosen industry. This year, the iGEM Calgary 2019 team aims to tackle the issue of green seed within the canola agricultural industry.

During the natural maturation cycle of canola seed, chlorophyll is produced and broken down as it acts in the energy harvesting mechanisms of the plant (Luciński & Jackowski, 2006). However when unfavorable weather conditions occur (specifically frost), the catabolic breakdown of chlorophyll is interrupted (Diosady, 2005). This directly leads to the accumulation of chlorophyll, and other intermediates within the seeds; so when the seeds are crushed for purposes of oil extraction, the pigments are retained within the canola oil (Diosady, 2005). The chemical structure of chlorophyll pigments causes the oil to oxidize rapidly, reducing the shelf-life of the final product (Wrolstad, 2004). The major canola oil producing provinces of Alberta and Saskatchewan are notorious for their weather instability which exacerbates this issue to a larger degree.

Existing methods of chlorophyll removal (often referred to as ‘Bleaching’) are inefficient, as they can cause up to a 50% loss of oil produced (Diosady, 2005; Ramamurthi & Low, 1995; Srinivasan, 2011). Our team has proposed an alternative: the production of chlorophyll-binding proteins, suspended in water-in-oil (W/O) microemulsions, as a way to capture and remove the chlorophyll pigments from the oil. The work presented here is a component of this overall project, focusing specifically on modelling and validating the variables of the emulsion formulation.

Since the goal of the iGEM Calgary project is to propose an alternative step in the processing of canola oil, the overall system should be able to filter out chlorophyll without interrupting existing downstream-processing operations. This can be achieved by leveraging the equilibrium behavior of our emulsion solution (Abdulkarim et al., 2011). Generally, emulsions are mixtures of two naturally immiscible liquids and a surfactant, which is partially miscible in both fluids. The surfactant then reduces the surface tension of immiscible fluids, allowing them to be homogeneously mixed. The type of equilibrium that an emulsion solution reaches is regularly characterized under the Winsor classifications (Figure 1) (A Winsor & Hahn, 1932).

The aim of this work is to develop supervised machine-learning based models, to predict the formulation conditions at which the used emulsion system will produce purified oil, as shown in the Winsor type 1 and 3 classifications. Machine learning methods, which generally aim to *approximate functions* between classes of data, may prove to be a powerful and efficient tool in the design of emulsion systems. Our proposed method can look to describe the phase class boundaries with only relatively few samples of data gathered in vitro, which can reveal greater practicality when compared to other phase prediction models.



Winsor Phase Classification of Emulsion Equilibria

Figure 1: A visual of the various Winsor types which classify phase equilibria that can form after emulsifying oil, water, and surfactant. The type of equilibrium depends on the ratios of the chemical species and the temperature of the solution.

METHODS

Computational Approach:

Samples of data were gathered *in vitro* and prediction of the decision boundaries between phases was completed with Support Vector Classification (SVC) and K-Nearest Neighbours (KNN) – both machine learning methods. The two approaches were compared and evaluated on their ability to find the decision boundaries at five different temperatures. The parameters of both methods are validated using cross-validation, which attempts to eliminate bias by resampling the data to train and test the model.

Problem Outline — Description of Data

The data given is four dimensional, containing three compositions of oil, water, and surfactant, and its equilibrium phase (classification label). Our model is looking to find a function $F : \vec{v} \longrightarrow y$ mapping a given vector to a winsor phase class such that,

$$\vec{v} = \begin{pmatrix} r_{oil} \\ r_{water} \\ r_{surfct} \end{pmatrix},$$

$$r_{oil} + r_{water} + r_{surfct} = 1 \ ,$$

$$y \in \{\text{Winsor 1, Winsor 2, Winsor 3, Winsor 4}\}$$

The training data set was given at each temperature, where its phase y is the classification label:

$$\{(\vec{v}_i, y_i)\} \quad i = 0, 1, 2...60$$

where *n* was 60.

K-Nearest Neighbours Clustering

The aim of a general classification model is to provide the likelihood a new unlabelled vector lies within a given predefined class. The $\mathcal{K}$-Nearest Neighbours method is a non-parametric approach which looks at the $\mathcal{K}$ nearest vectors (in terms of distance) within the space and assigns a label based on those closest neighbours. The probability given a vector $\vec{v}$ (from described above) will be labeled with phase y can be calculated with KNN by:

$$Pr(\ Y = y \mid X = \vec{v}\ ) = \frac{1}{\mathcal{K}} \sum_{i \in \mathcal{N}} I(\ y_i = y\ )$$

where *i* indexes through the $\mathcal{K}$ nearest vectors in $\mathcal{N}$ and *I* is the identity function which outputs a 1 if the label y_i of the neighbour is equal to y and 0 otherwise (James et al. 2017).

Support Vector Classification

Support Vector Classification (SVC) provides a classification approach which finds a hyperplane that divides two classes of vectors within a space. The goal is to find the maximum margin between the labelled data and generate parameters for a hyperplane that would divide this margin. The optimization problem of generating a separating hyperplane between two classes holding n data points can be summarized:

$$\max_{\beta_0, \beta_1, \beta_2, \beta_3, \epsilon_i \ldots \epsilon_n} \mathcal{M} \quad \text{subject to}$$

$$\beta_0^2 + \beta_1^2 + \beta_2^2 + \beta_3^2 = 1$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) \geq \mathcal{M}(1 - \epsilon_i)$$

$$\sum_{i=0}^{n} \epsilon_i \leq \mathcal{C} \ , \quad \epsilon_i \geq 0 \quad \text{and} \quad y_i \in \{1, -1\}$$

Where $\mathcal{M}$ is the size of the margin, β_i are the parameters defining the hyperplane, y_i is the label of each vector which can only be 1 or -1. ϵ_i is the error for each vector which is constrained by $\mathcal{C}$, the cost parameter (James et al. 2017).

Since we have four classes to be separated, we applied the one-versus-one approach, where divisions were constructed for each pair of classes, meaning this optimization was solved 6 times. It was obvious that the data is not linearly separable, so a non-linear radial basis function (RBF) was used as a kernel. A kernel is generally a function used to quantify the similarity between vectors . The RBF kernel is defined as:

$$K(\vec{v_0}, \vec{v_i}) = e^{-\gamma \vec{v_0} \cdot \vec{v_i}}$$

where $\vec{v_0}$ is the vector to be labelled, and the kernel is applied on each training vector $\vec{v_i}$ for this test observation. γ is a parameter subject to choice (James et al. 2017). This is a more useful approach because it considers the training data that are more local.

Cross-Validation:

K-fold cross validation provides some reasoning for the parameter choices within the model. To attempt to validate the model's classifications, the data is split into two partitions: train and test. K-fold cross validation outlines the way to choose such partitions. The model is trained and fed the test observations without a label. The model predicts each unlabelled test observation and is compared to its actual label (class), giving a quantification on the models accuracy. For K partitions of the data, the mean error rate of a model with given parameters is calculated as:

$$\text{Mean Error Rate} = \frac{1}{N} \sum_{i=1}^{K} \sum_{y_p \in Y_i} I(\, y_p \neq y_a \,)$$

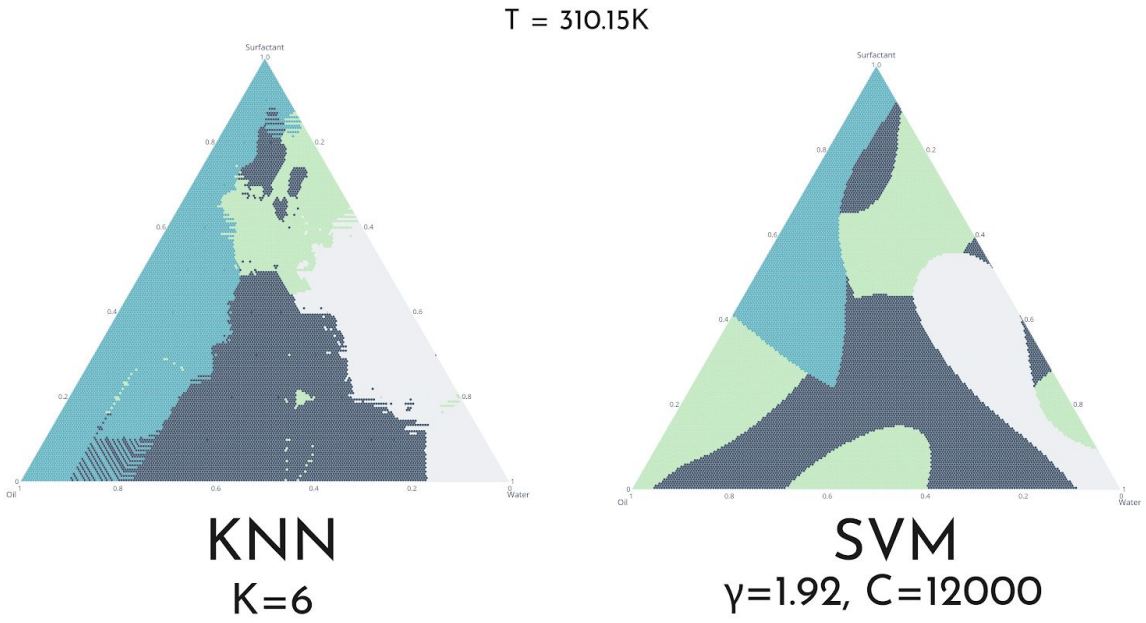
Where N is the number of data points (including train and test), y_p is the predicted label calculated by the model for all test vectors in partition Y_i, and y_a is the actual label of the vector (James et al. 2017).

It is not sufficient to merely choose a single subset for testing because it can be an inaccurate representation of which points the model can predict well. K-fold Cross Validation attempts to address this by using K partitions of the data to resample. The error rate is then determined by averaging each of the partition error, providing a more representative reading of the model's true error rate.

RESULTS

Comparison of K-Nearest Neighbours and Support Vector Classification:

At each temperature the optimal parameters for both KNN and SVC were found and the average K = 10 fold cross validation error was calculated for each method. SVC had an average error of 0.17 over all temperatures, meaning there would be a 17% chance of a new test observation being classified incorrectly. KNN had an average error of 0.33. Thus SVC proves to provide better phase classification on the data obtained (Figure 2). The difference in performance may be due to the tuning of the radial basis kernel parameters in the SVC, as it provides an advantage over KNN which does not consider distance calculations on the nearest vectors.

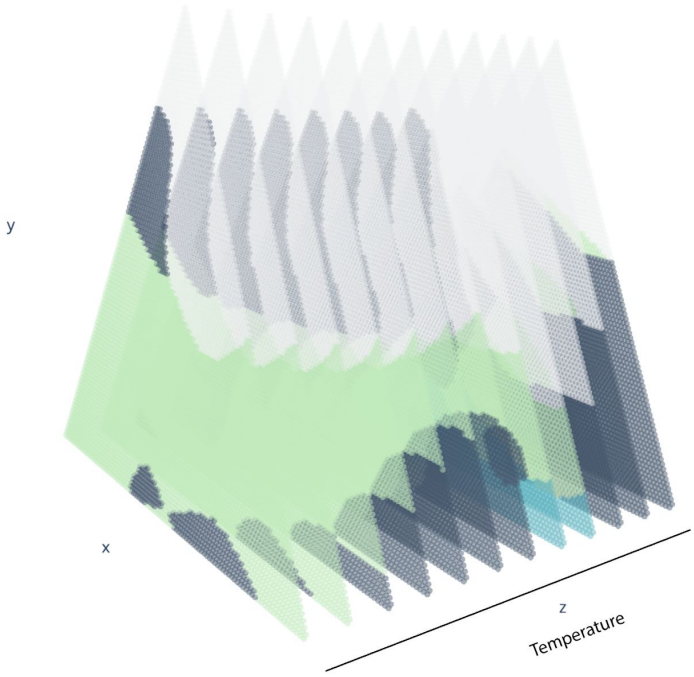


Visual Comparison of two Classification Approaches

Figure 2: Comparison of the phase boundaries at a single temperature of the K-Nearest Neighbours and Support Vector Classification method. Parameters were chosen based on the lowest test error rate found with Cross-Validation. (Navy = Winsor 1, Grey/White = Winsor 2, Green = Winsor 3, Light Blue = Winsor 4)

Visualization in 3D

Multiple ternary plots representing phase changes were generated over a range of temperatures using SVC with independently optimized parameters for each temperature value. The data within the ternary plots were projected into two dimensional cartesian coordinates (see methods). Temperature was added as a third dimension for each point within the triangular slice, allowing phase change to be viewed in three dimensional space as a function of temperature (Figure 3).



3D Plot of Phase Change as a Function of Temperature

Figure 3: Display of phase equilibria changing as a function of temperature in three dimensional cartesian space.

CONCLUSION

This model provides knowledge on emulsion formulation, which is an indispensable part of the UCalgary 2019 iGEM project, allowing for viable refinement of canola oil by dispensing chlorophyll that would be detrimental to its production. It is desirable to find the correct ratios of the emulsion constituents that lead to a Winsor 1 type equilibrium because the pure oil phase is left without a microemulsion. Determining these phase equilibria boundaries is therefore an imminent task which we provided a solution with the application of K-Nearest Neighbours and Support Vector Classification, both machine learning classification models. With merely sparse data samples gathered *in vitro*, Support Vector Classification predicted phase boundaries at its best with an error rate of 17% which could be improved with more samples taken in regions of low classification confidence. Hence, this method provides a practical alternative to other deterministic methods which require rigorous effort and time.

REFERENCES

A Winsor, B. P., & Hahn, von. (1932). *HYDROTROPY, SOLUBILISATION AND RELATED EMULSIFICATION PROCESSES. PART I. Aqueous Solutions of Parajin Chain Salts* (Vol. 62). Retrieved from https://pubs-rsc-org.ezproxy.lib.ucalgary.ca/en/content/articlepdf/1948/tf/tf9484400376

Abdulkarim, M. F., Abdullah, G. Z., Sakeena, M. H. F., Chitneni, M., Yam, M. F., Mahdi, E. S., … Noor, A. M. (2011). Study of Pseudoternary Phase Diagram Behaviour and the Effect of Several Tweens and Spans on Palm Oil Esters Characteristics. *International Journal of Drug Delivery*, *3*, 95–100. https://doi.org/10.5138/ijdd.2010.0975.0215.03058

Diosady, L. L. (2005). Chlorophyll Removal From Edible Oils. *International Journal of Applied Science and Engineering*, *3*(2), 81–88.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: With applications in R*. Springer.

Luciński, R., & Jackowski, G. (2006). *The structure, functions and degradation of pigment-binding proteins of photosystem II*. Retrieved from www.actabp.pl

Ramamurthi, S., & Low, N. H. (1995). *Effect of Possible Chlorophyll Breakdown Products on Canola Oil Stability*. *J. Agrie. FoodChem* (Vol. 43). Retrieved from https://pubs.acs.org/sharingguidelines

Srinivasan, R. (2011). Advances in application of natural clay and its composites in removal of biological, organic, and inorganic contaminants from drinking water. *Advances in Materials Science and Engineering*, *2011*. https://doi.org/10.1155/2011/872531

Wrolstad, R. E. (2004). Symposium 12 : Interaction of Natural Colors with Anthocyanin Pigments — Bioactivity and Coloring Properties. *Journal of Food Science*, *69*(5), 419–421.

LEARNING SKILLS AND DEVELOPMENT

Assessing Solutions of Issues

Our team spent a month during the early part of the year discussing issues within the community and economy which could be addressed and mediated by applications of synthetic biology. The time allowed us to think of solutions to meaningful problems, only to come to the conclusion that many people have already done work or attempted the same solution. I realized that if one wants to solve some of the most vexing concerns within society, years of work dedicated to thinking, collaborating, and developing solutions is the most productive approach to propose a truly novel idea.

Attempting to think of novel scientific applications is also beneficial in that one starts to find which fields of academics are meaningful. For example, to research in physics would most likely reveal one is concerned with understanding the complexity of the inanimate (although not always the case), whereas the interest in life sciences should reveal one to find health and identity of human life (and other life) of greater importance. The field of study one pursues tends to reveal the values and objectives they seek in their work. The exposure to more specific work done in other fields lead me to reassess my interests to see what I hope to study going forward. This realization reaffirmed why studying computer science and math is important. The technical skills needed for such breakthroughs in science disciplines from physics to medicine will be rigorously practised in the study of both math and computer science, presenting the opportunity for oneself to develop a breadth of understanding when applied to these specific sciences.

Integrating within a Team

Being chosen to work as a member of the iGEM Calgary team has exposed me to an interdisciplinary work environment. Our team is split into a “wetlab” composed of mainly biology majors and a “drylab” where mainly engineering, computer science and math students work on different models for the project. The challenge of explaining our work to other team members who are not as specialized in our field is important if we hope to have a successful integration of our work. The drylab had initially thought of a clever model in the sense that it would be intriguing math and engineering work, however after discussing with wet lab members, the model seemed to provide no beneficial knowledge. Identifying and communicating aspects of the project that can be informed with “drylab” work is critical for overall project success. .