Enzyme Promiscuity Model Formulation

# Problem Statement

Given an enzyme-substrate pair and information about the primary sequence of the enzyme and chemical formula of the substrate, the model aims to predict the propensity of the given enzyme acting upon the substrate as a probability measure.

# Primary Challenge

Unavailability of instances that represent lack of activity between enzyme and substrates restrict us from creating a training set with both positive and negative enzyme-substrate interaction instances, thereby preventing the use of traditional Machine Learning based discriminative classifiers.

# Modeling paradigms capable of solving the task

Although discriminative Machine Learning models in the supervised learning paradigm require both positive and negative training instances, there have been efforts to train a classifier using only positively labeled and unlabeled training examples [1]–[3], called PU learning. PU learning assumes that the unlabeled dataset contains both positive and negative training instances and tries to identify a set of reliable negative examples from the unlabeled set either automatically or through manual interference. For example, to ensure that the model can correctly discriminate between the two classes, the authors in [1] introduced a large number of irrelevant training examples which can be considered negative without any restrictions since they are not related to the problem they tried to solve. In our case however, such irrelevant examples may be hard to conjure.

A different approach to work with unlabeled or mislabeled training examples was proposed by Scholkopf et. al. [4], [5] where they adopted the discriminative classification algorithm, Support Vector Machines to the unsupervised learning domain and used it to detect outliers in their data. Although the algorithm does not require any labeled examples, the training set should contain both positive and negative data. Thus, along with the inclusion of active enzyme-substrate pair, we can also include random enzyme-substrate pair in our training set among which we hope that there are enough negative instances that would allow the model to detect an outlier enzyme-substrate pair.

While discriminative models such as SVMs try to learn the decision boundary between classes, generative models assume that the data comes from a probability distribution and tries to estimate that distribution [6]. For our purposes, generative models can be used to learn the probability distribution of substrates that lead to a certain enzyme. We could train a Generative Adversarial Network to predict encoded enzyme representations from numerical vector representation of substrates. Thereafter, we can use the encoded enzyme representation retrieved for a novel substrate to create a ranked list of enzymes by selecting those enzymes whose actual encoded representations are the most similar to the predicted encoded representation based on a distance metric such as Euclidean Distance.

Finally, among the unsupervised learning techniques, probability density estimation techniques such as the Kernel Density Estimation [7] can be used to learn the probability density function of our enzyme-substrate pair samples and given an unknown sample, we can use the estimated density function to predict how likely it is to observe that sample.

We believe the kernel density estimation might be the easiest to implement followed by outlier detection using unsupervised SVM algorithm. Please find below a brief sketch of the two methods.

# Kernel Density Estimation

Kernel Density Estimation (KDE) is a non-parametric method which can be used to estimate the probability density function of a multivariate random variable. Given a sample of independent and identically distributed observations, , of a random variable, the estimated density function at a point is:

Here, is the kernel function, is the bandwidth, a smoothing parameter, and is the number of observations. We could estimate the best kernel function and the ideal bandwidth required for our task using cross-validation. Compared to a Tanimoto index based search of relevant enzymes for a given substrate, KDE also not only incorporates substrate encodings but also includes enzyme features and is capable of ranking enzyme-substrate pairs based on the density function estimate. Unlike some of the other methods discussed in the above section that are capable of solving the task, KDE can work with only positive instances of enzyme-substrate interaction.

# One-Class SVM for outlier detection

The novelty detection using SVM algorithm, termed One-Class SVM, aims to estimate a function that is positive on a small subset of the dataset where most of the training data points belong and negative elsewhere. It uses a kernel function to map data into a feature space where it creates a separating hyperplane of maximal margin between the positive and negative regions. For a new instance, , the value of the function is determined based on which side of the hyperplane it belongs in the feature space. For our purposes of predicting propensity of enzyme-substrate activity, we could use the positive enzyme-substrate interaction instances along with manually created random enzyme-substrate instances and feed it to the One-Class SVM algorithm. Ideally, the algorithm will learn the positive instances as the primary subset where most of the data points would lie and define most of the random instances as outliers. Given a new enzyme-substrate instance from the test set, the algorithm will thus be able to predict whether it is most likely to belong to the positive instance subset or the random subset. One-Class SVMs are known to be highly sensitive to outliers and there are other outlier detection algorithms such as Isolation Forests, based on Random Forests, which have been shown to yield better performances and can be used to solve our task.

Diagram

Description automatically generated

Figure : The workflows of both Kernel Density Estimation (above) and Novelty Detection (below) using One-Class Support Vector Machines are illustrated here. Given information about interacting Enzyme-Substrate pairs, we can either estimate the density function of the concatenated feature space using Kernel Density Estimation or estimate the region where most of the datapoints lie using One Class SVMs. Henceforth, using the trained algorithms, we can score novel enzyme-substrate pairs according to their propensity of interaction.

# References

[1] X. L. Li and B. Liu, “Learning from positive and unlabeled examples with different data distributions,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2005, vol. 3720 LNAI, doi: 10.1007/11564096\_24.

[2] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, “Building text classifiers using positive and unlabeled examples,” 2003, doi: 10.1109/icdm.2003.1250918.

[3] J. Bekker and J. Davis, “Learning from positive and unlabeled data: a survey,” *Mach. Learn.*, vol. 109, no. 4, 2020, doi: 10.1007/s10994-020-05877-5.

[4] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Piatt, “Support vector method for novelty detection,” 2000.

[5] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, “Estimating the support of a high-dimensional distribution,” *Neural Comput.*, vol. 13, no. 7, 2001, doi: 10.1162/089976601750264965.

[6] G. Harshvardhan, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, “A comprehensive survey and analysis of generative models in machine learning,” *Comput. Sci. Rev.*, vol. 38, 2020, doi: 10.1016/j.cosrev.2020.100285.

[7] G. R. Terrell and D. W. Scott, “Variable Kernel Density Estimation,” *Ann. Stat.*, vol. 20, no. 3, 2007, doi: 10.1214/aos/1176348768.