Postulating Support Vector Regression Model to Measure the Effect of Protein, Carbohydrate and Fats on the Weight of Carb Fish

Rawa saman maaroof,
Department of Tourism, College of Commerce, University of Sulaimani, rawa.maaroof@univsul.edu.iq

Sham azad Rahim
Department of Finance and Banking College of Commerce, University of Sulaimani, sham.rahim@univsul.edu.iq

Shahla Othman salih
Department of Statistics and Informatics, College of Administration Economic University of Sulaimani, shahla.salih@univsul.edu.iq

Hindreen Abdullah Taher
Department of Information Technology, college of Commerce, University of Sulaimani, Hindreen.taher@univsul.edu.iq

Abstract

In this paper we studied the combination among the levels of carbohydrate (20%, 30%, 40% and 50%), protein (8%, 12%, 16% and 20%) and fats (5%, 10%, 15% and 20%), where all possible combinations are 64. We gave each combination of the aforementioned elements to an aquarium fish with volume of 1.92 m3, each aquarium contained 5 fishes, the aim of our study is to detect which combination the three elements are record a high weight of the fishes, here we depend on the average of the fishes weight and the results clarified that the combination (15%, 12% and 10%) of 1kg for carbohydrate, protein and fats respectively are gave average weight of 2.312 kg, for this purpose SVR has been used. According to the results radial kernel function gave highest performance compared to the other kernel functions, the R2 = 89% this implies the factors capable of explaining 92% of fishes weight with MSE and RMSE of (0.000506 and 0.02249) respectively. And p-values of the three aforementioned variables are less than the significant level of 0.01, implying that the three factors have a statistically significant impact on the fishes weight. Where carbohydrate has an impact of 0.0021 on the fishes weight, in another word if carbohydrate increase by 1% unite, then the weight of fishes increase by 0.002 grams, also both protein and fats have a significant positive effect on the response variable, and the amount of impacts are (0.0136 and 0.0014) respectively.

Keywords: *Regression Model, Support vector regression, kernel functions.*

INTRODUCTION

Carp are several species of oily freshwater fish in the Cyprinidae family, a vast group of fish endemic to Europe and Asia. While carp is eaten in many parts of the world, it is considered an invasive species in sections of Africa, Australia, and the majority of the United States. Because these groups share

similar characteristics, the cypriniformes (family Cyprinidae) are conventionally grouped with the Characiformes, Siluriformes, and Gymnotiformes to form the superorder Ostariophysi. These characteristics include being found primarily in fresh water and having Weberian ossicles, which are an anatomical structure derived from the first five anterior-most vertebrae. Ribs and neural crests correlate. Humans have historically relied on carp as a source of protein. Numerous species have been popular decorative fishes, including various goldfish breeds and the domesticated common carp variant known as koi. As a result, carp have been brought to a variety of areas, with varying degrees of success. Many carp species are considered invasive in the United States[2], and considerable quantities of money are spent on carp control globally.

Literature Review

To classify thermography images into normal or abnormal categories for the detection of canine bone cancer disease, canine anterior cruciate ligament rupture, and feline hyperthyroid disease, Lama (2017) employed SVM models as binary classifiers using gray level co-occurrence matrix texture features extracted from the thermographs[1]. Also based on parallel factor analysis coupled with support vector regression (SVR), Gu and Sun (2019) designed a probe-based fluorescence spectroscopy for the quick detection of lysed and oxidized chemicals (i.e., acids, aldehydes, alcohols, ketones, hydrocarbons, etc.) in frying palm oil. Characteristic fluorescence peaks were identified using loading scores at relevant components with the help of the parallel factor analysis technique. Then, a variety of preprocessing algorithms were combined with the SVR algorithm. Grid search performed better than the other three methods in a regression test using four distinct SVM models. The final SVR models' performance was evaluated using the following metrics: R2 = 0.9753, P = 0.9724, MSE = 0.0089, and P = 0.0088 for the calibration and prediction sets, respectively[2]. And Gu et al. (2020) invented probe-based three-dimensional fluorescence spectroscopy using parallel factor analysis and support vector regression (SVR) to identify, discriminate, and quantify dissolved organic materials in frying oil. Compared to time-consuming and expensive chemical procedures, the proposed methodology improved the rapid assessment of frying oil quality and other high-oil food and beverages. Considering time and model robustness, parallel factor analysis combined with analysis of characteristic peaks data may be better for model creation[3].

Methodology

Support Vector Regression (SVR)

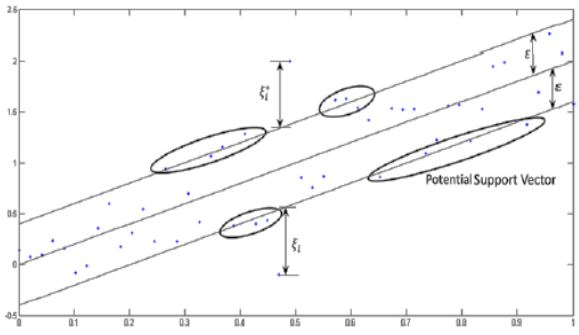
Linear regression is the most statistical model used in practical applications because these types of models are linearly dependent on their unknown parameters. This can be fitted much more easily than the other models which response have a non-linear relationship with their unknown parameters and because the properties of statistical estimators are easier to explain. But the assumption of OLS method cannot be achieved easily[4,5]. Support vector machines (SVMs) are well-suited to generalizing on unseen data due to their statistical learning or Vapnik-Chervonenkis (VC) foundations[6]. Kernels, sparse solutions, VC margin, and SVR control are similar to categorization. SVR estimates real-valued functions better than SVM, despite its lesser fame. SVR's loss function punishes over and under-estimates equally during training. SVR is supervised learning. In his e-insensitive technique, Vapnik builds a flexible tube with a minimal radius symmetrically around the estimated function to reject absolute values of errors below a predefined threshold e in both the upper and lower regions of the estimate. This approach affects the region above and below the function but

not the tube[6,9,11]. SVR's computational complexity is independent of input space size, which is a significant benefit. It can predict and generalize well. Thus, this chapter will cover SVR and Bayesian regression. An adjusted SVR can be used to avoid underestimating a function. Figure(1) depicts a one-dimensional problem that can be viewed geometrically to help establish the best formulation for an SVR problem. Equation 4-1 provides a convenient form for approximating continuous-valued functions. Simplifying the mathematical terminology, we may construct the multivariate regression from Equation 4-2 by increasing x by one and adding b to the w vector.

$$y = f(x) = <w, x> + b = \Sigma_{j-1}^{m} w_j \; x_j + \text{b} \text{ , y ,b } \in \text{R ,x , w } \in R^m \quad (4\text{ -1})$$

$$f(x) = \begin{bmatrix} w \\ b \end{bmatrix}^r \begin{bmatrix} x \\ 1 \end{bmatrix} = w^T \; x + b \; x, w \; \in R^{m+1} \quad (4\text{-2})$$

Figure (1) shows one dimension SVR



By framing the work as an optimization issue, support vector regression (SVR) seeks to find an approximation for a function that minimizes the prediction error, or the difference between the expected and the intended outputs, where ||w|| is the magnitude of the normal vector to the surface being approximated, the objective function is given by Equation 4-3:

$$min_w \; \frac{1}{2} \; |w|^2$$

Here's an illustration of how the sum of the weights might be used as a proxy for levelness:

$$f(x\,, w) = \sum_{t-1}^{M} w_t \; x^t \;\;, x \in R, w \in R^M \quad (4\text{-}3)$$

There is a clear indication of the approximate polynomial's order, M. As the size of the vector w grows. The horizontal line stands for a significantly off-ideal 0th order polynomial solution. While the 1st-order polynomial linear function better approximates portions of the data, it still doesn't produce a satisfactory match to the training data as a whole. The 6th-order solution provides an acceptable compromise between function flatness and prediction error. The immense complexity of the highest-order solution means it will likely overfit the answer on unseen data even though it has zero error. The size of the regularizing term w determines the extent to which the flatness of the solution can be manipulated in an optimization problem. The constraint is to restrict the value of the function to be as close as possible to the expected value for a particular input[10,12]. The SVR algorithm penalizes predictions that are more than e distant from the target value by using a loss function that does not consider e. In addition to affecting the number of support vectors and, by extension, the sparsity of the solution, the value of e determines the width of the tube. A smaller value suggests a lower tolerance for error. Figure 4-1 provides a visual representation of this latter concept[7]. If e is lowered, the tube's boundary creeps inward. Since there are more data points near the boundary, more support vectors exist. Increased e also reduces the number of locations near to international borders. The model's resilience is enhanced by the e-insensitive zone, which makes it less vulnerable to perturbations in the data. Equations 4-4, 4-5, and 4-6 illustrate the linear, quadratic, and Huber e loss functions, respectively, and can be applied. The Huber loss function, as seen in Figure 4-3, is more lenient on minor deviations from the desired

output than linear and quadratic loss functions, but it still penalizes any and all outliers. Which loss function to employ depends on the available computational resources for training, the desired degree of model sparsity, and a priori knowledge of the noise distribution affecting the data samples.

$$L_\delta(y, f(x,w)) = \begin{cases} 0 & |y - f(x,w)| \le \delta; \\ |y - f(x,w)| - \delta & otherwise, \end{cases} \quad (4-4)$$

$$L_\delta(y, f(x,w)) = \begin{cases} 0 & |y - f(x,w)| \le \delta; \\ (|y - f(x,w)| - \delta)^2 & otherwise, \end{cases} \quad (4-5)$$

$$L(y, f(x,w)) = \begin{cases} c|y - f(x,w)| - \frac{c^2}{2} & |y - f(x,w)| > c \\ \frac{1}{2}|y - f(x,w)|^2 & |y - f(x,w)| \le c \end{cases} \quad (4-6)$$

Kernel SVR and Different Loss Functions

Before, we assumed that f(x) was linear and focused on data in the feature space. When dealing with nonlinear functions, it is possible to improve classification accuracy by mapping the data into a higher-dimensional space (called kernel space) using kernels that satisfy Mercer's condition[8,10]. Substituting k(xi, xj) for x in Equations 4-1-4-18 results in the fundamental formulation illustrated in Equation 4-19, where Φ(.) denotes the transformation from feature to kernel space. The reformulated weight vector, in terms of the original input, is defined by Equation 4-20.

Symmetric and convex loss functions are provided. To ensure that the optimization problem has a unique solution that can be found in a finite number of iterations, the loss function used to correct for under- or over-estimation must be convex. To begin deriving the topics of this chapter, we will use Equation 4-4's linear loss function.

Equation 4-21 represents the dual problem, while Equation 4-22 represents the function approximation f (x), where k(.,.) is the kernel, as shown in Equation 4-23.

$$min \frac{1}{2}\|w\|^2 + C\sum_{t-1}^{N} \xi_i + \xi_i^n \quad (4-7)$$

Subject to

$$y_t - w^T\Phi(x_i) \le \xi_i + \xi_i^n \quad i = 1, \dots\dots, N$$
$$w^T\Phi(x_{i)} - y_i \le \xi + \xi_i \quad i = 1, \dots\dots, N$$
$$\xi_i, \xi_i^n \ge 0 \quad i = 1 \dots . . N$$

$$w = \sum_{i-1}^{Nsv} (\alpha_i^n - \alpha_i)\phi(x_i\ ) \quad (4-8)$$

$$max_{n,m} - \varepsilon\sum_{i=1}^{Nsv}(\alpha_i + \alpha_i^n) + \sum_{i=1}^{Nsv}(\alpha_i^n - \alpha_i)y_i - \frac{1}{2}\sum_{j-1}^{Nsv}\sum_{i-1}^{Nsv}(\alpha_i^n - \alpha_i)\left(\alpha_j^n - \alpha_j\right)k(x_i, x_{j)} \quad (4-21)$$
$$\alpha_i, \alpha_i^n \in [0, C], i = 1, \dots\dots . N_{sv}, \sum_{i-1}^{Nsv}(\alpha_i^n - \alpha_i) = 0$$

$$f(x) = \sum_{i=1}^{Nsv}(\alpha_i^n - \alpha_i)k(x_i, x) \quad (4-9)$$

$$k(x_i, x) = \emptyset(x_i).\emptyset(x) \quad (4-10)$$

Applications

Data Description

The data of our study is an agricultural experiment, three factors have been used to measure the weight of fishes as response variable and each factors has four percentage levels in each kilograms of fishes food, which are carbohydrate (20%, 30%, 40% and 50%), protein (8%, 12%, 16% and 20%) and fats (5%, 10%, 15% and 20%), where all possible combinations are 64. We gave each combination of the aforementioned elements to an aquarium fish with volume of 1.92 m3, each aquarium contained 5 fishes

Results of SVR

For performing the SVR model we used all data as training data set which are 64 observations, because our data set is rather small.

Table-1 Shows the performance of SVR for each kernel functions.

Kernel	number of SVR	R^2	MSE	RMSE
Linear	40	70%	0.002344	0.04841
Polynomial	32	62%	0.004118	0.06417
Radial	51	89%	0.000506	0.02249
Sigmoid	35	37%	1.378911	1.17426

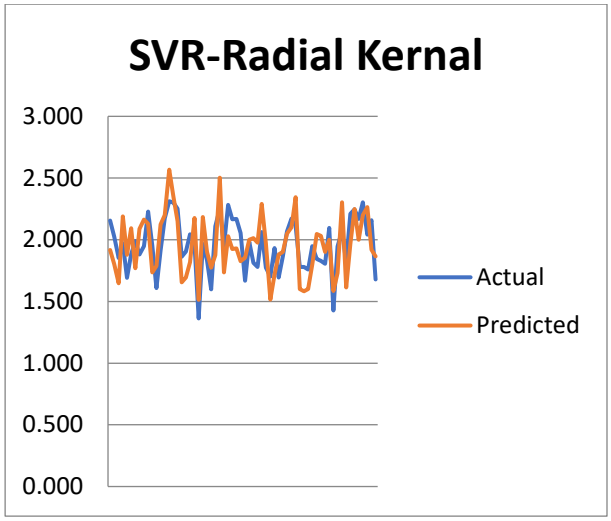
Summing up to the table-1, which represents the application of the epsilon support vector regression model with selecting the best kernel function, it is clear that the radial kernel function has the highest performance among the other kernel functions. Furthermore, the R2 of the best kernel is equal to 89% with minimum MSE and RMSE (0.000506 and 0.02249), respectively.

Table-2 Displays the test of the estimators and their impacts.

Explanatory variables	Estimated	S.E	P -Value
Carbohydrate	0.0021	0.00037	0.000
Protein	0.0136	0.0041	0.000
Fats	0.0014	0.00053	0.000

The table-3 clarifies the estimated values of the parameters for the features and their test to check whether they have an impact on fishes weight or not. The three values of the p-value column are less than the significant level of 0.01, implying that the three factors have a statistically significant impact on the response variable.

Figure (2)



Conclusions

This paper estimated the impact of carbohydrate, protein and fats of 64 datasets, using radial kernel functions. Results showed that the three factors had a statistically significant impact on fishes weight. And protein has highest impact.

Reference

Lama, N. (2017). Optimized Veterinary Thermographic Image Classification using Support Vector Machines and Noise Mitigation (Doctoral dissertation, Southern Illinois University at Edwardsville).

Gu, H., & Sun, Y. (2019). Enhancing the fluorescence spectrum of frying oil using a nanoscale probe. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 218, 27-32.

Gu, H., Huang, X., Chen, Q., & Sun, Y. (2020). Rapid Assessment of Total Polar Material in Used Frying Oils Using Manganese Tetraphenylporphyrin Fluorescent Sensor with Enhanced Sensitivity. Food Analytical Methods, 13(11), 2080-2086.

Ahmed, N. M., & Taher, H. A. (2018). Multi-response Regression Modeling for an Agricultural Experiment. Journal of University of Human Development, 4(2), 46-52.

Taher, H. A., & Ahmed, N. M. (2023). Using Bayesian Regression Neural Networks Model to Predict Thrombosis for Covid-19 Patients. resmilitaris, 13(1), 2077-2087.

Vapnik, V. (2000). The nature of statistical learning theory. Springer, 314. doi: https://doi.org/10.1007/978-1-4757-3264-1

Mechelli, A., Vieira, S. (Eds.) (2019). Machine learning: methods and applications to brain disorders. Academic Press. doi: https://doi.org/10.1016/C2017-0-03724-2

Blanco, V., Puerto, J., Rodriguez-Chia, A. M. (2020). On lp-Support Vector Machines and Multidimensional Kernels. Journal of Machine Learning Research, 21 (14). Available at: https://jmlr.org/papers/volume21/18-601/18-601.pdf

Astuti, W., Adiwijaya (2018). Support vector machine and principal component analysis for microarray data classification. Journal of Physics: Conference Series, 971, 012003. doi: https://doi.org/10.1088/1742-6596/971/1/012003

Chowdhury, U. N., Rayhan, M. A., Chakravarty, S. K., Hossain, M. T. (2017). Integration of principal component analysis and support vector regression for financial time series forecasting. International Journal of Computer Science and Information Security (IJCSIS), 15 (8), 28–32.

Naik, G. R. (Ed.) (2018). Advances in Principal Component Analysis. Springer. doi: https://doi.org/10.1007/978-981-10-6704-4.

Arık OA (2020) Comparisons of metaheuristic algorithms for unrelated parallel machine weighted earliness/tardiness scheduling problems. Evol Intel 13:415–425.