Using Optimizing Parameters Support Vector Regression Model to Predict Potassium Ratio in Carb Fish

Azhy Akram Aziz
Asst Lecturer, College of Business and Administration, Erbil polytechnic university, azhy.aziz@epu.edu.iq

Heshu Othman F. Mahmood
Asst Lecturer, Department of Statistics and Informatics, College of Administration & Economics, University of Sulaimani, heshu.faqe@univsul.edu.iq

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

Rawa Saman Maaroof
Asst Lecturer, Department of Tourism, College of Commerce, University of Sulaimani, rawa.maaroof@univsul.edu.iq

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

Abstract

In this paper, we studied the combination of the levels of carbohydrates (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 a volume of 1.92 m3, each aquarium contained 5 fishes, the aim of our study is to detect which combination of the three elements recorded a high potassium ratio of the fishes, here we depend on the average of the fishes weight and the results clarified that the combination (30%, 8% and 15%) of 1kg for carbohydrate, protein, and fats respectively are given average potassium ratio of 503mg/kg, for this purpose optimized parameters SVR has been used. According to the results radial kernel function with optimized parameter (gamma = 1.341 and cost = 0.844) gave the highest performance compared to the other kernel functions, the R2 = 91% this implies the factors capable of explaining 91% of fishes weight with MSE and RMSE of (0.000438 and 0.02092) 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 fish's weight. Where carbohydrate has an impact of 0.12 on the fish's potassium ratio, in another word if carbohydrate increase by one unit, then the fish's potassium ratio increases by 0. 12 mg, also both protein and fats have a significant positive effect on the response variable, and the amount of impacts are (1.015 and 0.117) respectively [13].

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

INTRODUCTION

The carp fish lives in fresh waters such as riverbeds and water reservoirs, submerged areas, and shallow waters, where it settles at the bottom but moves to the middle and upper regions for food, and although this fish has a flavor that everyone likes, it is very much What is caught and raised for enjoyment is considered a large fish, and its breeding depends on several factors, including the environment and its life cycle. The most important foods are carnivores and consumers of animal food, such as aquatic insects, larvae, worms, and plankton. They also feed on small fish, eggs of other fish, and plant seeds. Balanced feeds can be added, but they must be small in size, with the possibility of increasing the size of the feed depending on the increase in the size of the mouth, and it is preferable that these feeds contain a high percentage of protein, as protein constitutes approximately 40% of its diet, meaning that it needs 12 grams of protein per kilogram of weight. They also eat corn and bread, cheese and berries, cat food, potatoes, beans, oats, flaxseed, and canola, all of which can be used to make healthy and nutritious meals for fish. These fish need fats by 5-15% of the diet, and carbohydrates by 30-40%, Because carp fish do not have a stomach to store food, it is preferable to always provide them with food in small quantities. During the fall season, carp fish must be provided with food rich in fats to be stored in the muscles and used in the winter. It is not necessary to provide carp fish with sources of vitamin C, as they manufacture it inside their intestines. Protein should be reduced in the diet in the winter season and replaced with carbohydrates, as the decrease in its metabolism in this season due to low temperatures makes it difficult to digest protein. Staying away from the immune-enhancing ingredients in the feed just because the winter season ends when the water temperature starts to rise, because of this effect on the depletion of the immune response.

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], also Rawa S. Maaroof and et al (2023) studied the combination of carbohydrates, protein, and fats in an aquarium fish with a volume of 1.92 m3. The results showed that the combination (30%, 8%, and 15%) had an average potassium ratio of 503mg/kg. Radial kernel function with optimized parameters (gamma = 1.341 and cost = 0.844) gave the highest performance compared to other kernel functions. Carbohydrate had an impact of 0.12 on the fish's potassium ratio, while protein and fats had a significant positive effect on the response variable [13].

Methodology

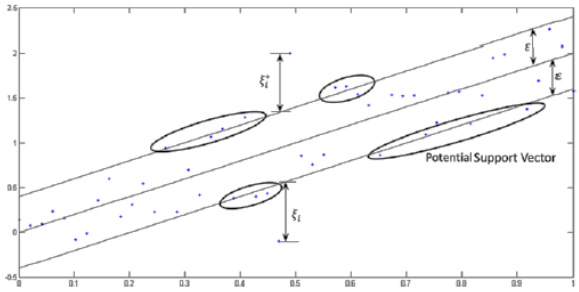
Support Vector Regression (SVR)

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 whose responses have a non-linear relationship with their unknown parameters because the properties of statistical estimators are easier to explain. But the assumption of the OLS method cannot be achieved easily [4,5,13]. 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 2-1 provides a convenient form for approximating continuous-valued functions. Simplifying the mathematical terminology, we may construct the multivariate regression from Equation 2-2 by increasing x by one and adding b to the w vector.

$$y = f(x) = <w, x> + b = \Sigma_{j-1}^{m} w_j \; x_j + \text{b, y, b } \in \text{R, x, w } \in R^m \qquad (2\text{ -1})$$

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

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 2-3:

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

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

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

(2-3)

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 over fit 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 [10,12,13]. 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 2-1 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 2-4, 2-5, and 2-6 illustrate the linear, quadratic, and Huber e loss functions, respectively, and can be applied. The Huber loss function, as seen in Figure 2-3, 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.

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 2-4's linear loss function.

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

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

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

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 [8,10,13]. Substituting k(xi, xj) for x in Equations 2-1-2-2 results in the fundamental formulation illustrated in Equation 2-9, where Φ(.) denotes the transformation from feature to kernel space. The reformulated weight vector, in terms of the original input, is defined by Equation 2-4. Equation 2-5 represents the dual problem, while Equation 2-6 represents the function approximation f (x), where k(.,.) is the kernel, as shown in Equation 2-7.

$$min\,\tfrac{1}{2}\|w\|^2 + C\textstyle\sum_{t-1}^{N}\ \xi_i + \xi_i^n \qquad (2-7)$$

Subject to

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

$$w = \textstyle\sum_{i-1}^{Nsv}\ (\alpha_i^n - \alpha_i)\phi(x_i\ ) \qquad (2-8)$$

$$max_{n,m} - \varepsilon\textstyle\sum_{i=1}^{Nsv}(\alpha_i + \alpha_i^n) + \sum_{i=1}^{Nsv}(\alpha_i^n - \alpha_i)y_i - \frac{1}{2}\sum_{j-1}^{Nsv}\sum_{i-1}^{Nsv}(\alpha_i^n - \alpha_i)\big(\alpha_j^n - \alpha_j\big)k(x_i, x_{j)} \qquad (2\text{-}9)$$
$$\alpha_i, \alpha_i^n \in [0, C], i = 1, \ldots\ldots N_{sv}, \textstyle\sum_{i-1}^{Nsv}(\alpha_i^n - \alpha_i) = 0$$

$$f(x) = \textstyle\sum_{i=1}^{Nsv}(\alpha_i^n - \alpha_i)k(x_i, x) \qquad (2-10)$$

$$k(x_i, x) = \emptyset(x_i).\emptyset(x) \qquad (2-11)$$

Optimize Parameters of SVR

Support vector regression is here. SVR allows us to choose a tolerance for model error and then finds a line (or hyperplane in higher dimensions) that best fits the data. SVR differs from OLS in that rather than minimizing the squared error, its goal is to reduce the coefficients themselves, or more precisely the l2-norm of the coefficient vector. The error term is instead dealt with in the constraints, where we require that the absolute error be smaller than or equal to a maximum error (epsilon). Depending on how precise we need our model to be, we can adjust epsilon. Here are our revised objectives and limitations:

Objective function:

MIN ½ $||w||^2$ (2 – 12)

Constraints:

$|y_i - w_i x_i| \le \varepsilon$ (2 – 13)

$y_i, x_i \ge 0$

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 kilogram 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	42	73%	0.001904	0.04363
Polynomial	28	57%	0.005207	0.07216
Radial	53	91%	0.000438	0.02092
Sigmoid	22	42%	1.553711	1.24648

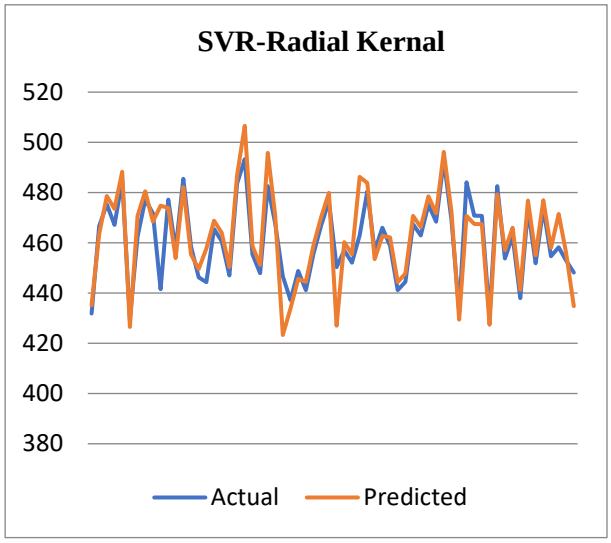
Summing up to the table-1, which represents the application of the epsilon support vector regression model with selecting the best kernel function, it is clear that the radial kernel function has the highest performance among the other kernel functions. Furthermore, the R2 of the best kernel is equal to 91% with minimum MSE and RMSE (0.000438 and 0.02092), respectively.

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

Explanatory variables	Estimated	S. E	P -Value
Carbohydrate	0.120	0.00011	0.000
Protein	1.015	0.00280	0.000
Fats	0.117	0.00039	0.000

The table-2 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 potassium ratio.

Figure (2)



The above figure shows the line graph between the actual and predicted values of the best kernel function.

Conclusions

This paper we estimated the effect of carbohydrate, protein and fats of 64 observations, by using radial kernel functions with optimized SVR parameters. The results showed that the three explanatory variables

had a statistically significant effect on the potassium ratio. 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.

azad Rahim, S., & Taher, H. A. (2023). Postulating Support Vector Regression Model to Measure the Effect of Protein, Carbohydrate and Fats on the Weight of Carb Fish. Journal of Survey in Fisheries Sciences, 10(3S), 4099-4104.