CHAPTER

7

Support vector regression

Fan Zhang, Lauren J. O'Donnell

Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States

7.1 Introduction

Support vector regression (SVR) is a supervised machine learning technique to handle regression problems (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997; Vapnik, 1998). Regression analysis is useful to analyze the relationship between a dependent variable and one or more predictor variables. SVR formulates an optimization problem to learn a regression function that maps from input predictor variables to output observed response values. SVR is useful because it balances model complexity and prediction error, and it has good performance for handling high-dimensional data. SVR is an extension to the Support Vector Machine (SVM) classification algorithm (Boser, Guyon, & Vapnik, 1992) (see Chapter 6). However, unlike SVM classification that produces binary output (i.e., a class label), SVR handles a regression problem that allows for a real-valued function estimation (e.g., continuous score in a clinical scale). SVR applies the basic idea of SVM, i.e., a sparse kernel machine that performs classification using a hyperplane defined by a few support vectors. As a result, the optimization in SVR is represented in terms of support vectors (a small set of training data samples), where the optimization solution does not depend on the dimension of the input data but only depends on the number of support vectors.

SVR has additional advantages when compared to other regression methods. With the use of a kernel, SVR can provide an efficient way to handle a nonlinear regression problem by projecting the original feature into a kernel space where data can be linearly discriminated (Ben-Hur, Ong, Sonnenburg, Schölkopf, & Rätsch, 2008; Muller, Mika, Ratsch, Tsuda, & Scholkopf, 2001; Orrù, Pettersson-Yeo, Marquand, Sartori, & Mechelli, 2012; Schölkopf, Smola, & Bach, 2002). Another benefit of SVR is that it learns a model to describe a variable's importance in characterizing

Machine Learning
https://doi.org/10.1016/B978-0-12-815739-8.00007-9

the relationship between input and output, whereas in a traditional data regression method, one needs to assume a model that might not be accurate. For example, a linear regression (e.g., least squares regression) makes assumptions regarding a linear distribution of input data, without providing meaningful coefficients or confidence intervals when the underlying relationship between inputs and outputs is actually nonlinear. SVR, on the other hand, is a machine learning technique, which seeks to maximize predictive accuracy from computation of a confidence interval for a variable's importance to describe the relationship between inputs and outputs (Glaser, Benjamin, Farhoodi, & Kording, 2019).

SVR has been shown to be an effective tool in many applications to study brain disorders. In brain disorder studies, output response measures can be clinical and/or behavioral outcomes that are continuous real-valued numbers, such as disease symptom severity scores (Moradi, Khundrakpam, Lewis, Evans, & Tohka, 2017), chronological ages (Koutsouleris et al., 2014), and mathematical ability measures (Iuculano et al., 2014). Because of its good performance on high-dimensional data, SVR has been widely applied as a tool for multivariate pattern analysis to understand complex disorder effects across the brain (Linn, Gaonkar, Doshi, Davatzikos, & Shinohara, 2016). SVR has been most often applied in studies using magnetic resonance imaging (MRI). MRI is widely used to investigate changes in the brain structure and function in neurological and psychiatric disorders, as described in multiple reviews (Li, Karnath, & Xu, 2017; Lorenzetti, Allen, Fornito, & Yücel, 2009; Shenton, Dickey, Frumin, & McCarley, 2001; Yu-Feng et al., 2007). Disorder effects may be manifested as spatially distributed patterns across multiple brain regions on MRI (Craddock, Holtzheimer, Hu, & Mayberg, 2009; Cuingnet et al., 2011; Fan, Shen, Gur, Gur, & Davatzikos, 2007; Zhang et al., 2018). SVR is one widely used technique to characterize the relationship between such complex patterns and the underlying disorder. SVR is applied to build a predictive model from imaging features to clinical and/or behavioral outcomes. For example, SVR has been used to study neurodevelopment in schizophrenia by predicting chronological ages from gray matter density and volume features computed from MRI (Koutsouleris et al., 2014). As a second example, SVR has been used to study attention-deficit/hyperactivity disorder (ADHD) by predicting childhood aggression using brain white matter fiber tract features extracted from diffusion MRI (Cha et al., 2015).

This chapter is designed to provide an overview of the SVR algorithm, followed by a description of SVR applications to study brain disorders. In the rest of the chapter, we first describe the basics of SVR, including underlying concepts, extended models, and model validation. Then, we describe how SVR can be used in studies of brain disorders, focusing on the SVR applications that use MRI data. Finally, conclusions are provided for a summary of the chapter, followed by several key points of SVR.

7.2 Method description

7.2.1 Overview

SVR was developed by Vapnik and co-workers (Drucker et al., 1997; Vapnik, 1998) by extending their SVM algorithm for classification (Boser et al., 1992). In machine learning, SVM is well known for its good performance to handle high-dimensional data. SVM is grounded in the framework of statistical learning theory (or Vapnik-Chervonenkis [VC theory]) and it offers a principled approach to machine learning problems due to this mathematical foundation. The basic idea of SVM was originally proposed in the 60s by Vapnik et al. (1963, 1964), following which the algorithm was largely developed in the next decades. The entire system of SVM was considered to be officially published in 1992 for classification (Boser et al., 1992) and then for regression (known as the ε-SVR model) (Drucker et al., 1997; Vapnik, 1998). Because SVR extends the SVM classification algorithm, we give a brief description for an SVM classification task (see Chapter 6 for details). In SVM classification, given a training dataset, each labeled sample is treated as a data point in a multidimensional feature space, and a hyperplane in this feature space is computed to correctly classify as many training samples as possible. New samples are then classified based on which side of the hyperplane they fall in the multidimensional feature space. To find a good hyperplane, optimization is performed by maximizing the margin between the support vectors (i.e., the data points nearest to the hyperplane).

For data regression, instead of finding a hyperplane that can largely separate the training samples, SVR introduces an ε-insensitive loss function to compute a hyperplane such that the predicted response values of the training samples have at most an ε deviation from their observed (actual) response values (Fig. 7.1). The hyperplane plus ε define an

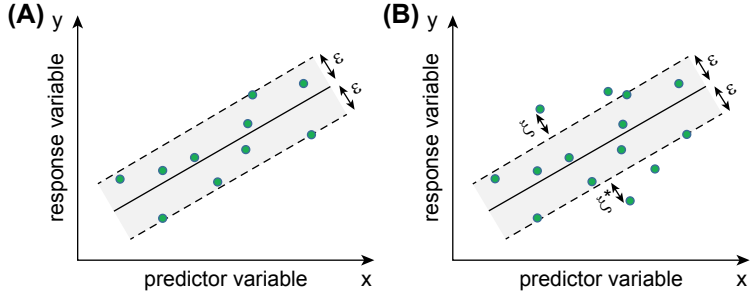

FIGURE 7.1 Graphic representation of linear ε-SVR models. Subfigure (A) shows a case of the simplest SVR model where the regression function can approximate all input data points. Subfigure (B) shows a model with slack variables ξ and ξ^*. These variables can account for noisy data at the hyperplane boundaries. *SVR*, support vector regression.

ε-insensitive tube (or band) for computing generalization bounds for regression. Optimization is performed by minimizing the ε-insensitive tube to be as flat (narrow) as possible while containing most of the training samples. In this case, the hyperplane is represented in terms of a few support vectors, i.e., training samples that lie outside the boundary of the ε-insensitive tube. As a result of the SVR training, a regression model is learned for prediction of a response output for a new sample.

7.2.2 Linear ε-SVR model

The goal of ε-SVR is to estimate a function with a constraint that the estimation of each input data point has at most ε deviation from its actual response value, by forming an ε-insensitive tube symmetrically abound the estimated function (Fig. 7.1A). In this section, we start with a simple case of a linear ε-SVR model. The mathematical formulation of a linear ε-SVR can be expressed as follows. Suppose we have a set of training data $\{(x_1, y_1), \ldots, (x_l, y_l)\}$, where x_i is the input data and y_i is the target output. The case of a linear function f takes the form:

$$y = f(x) = w, x + b = w^T x + b \tag{7.1}$$

where $w,\ x$ denotes the dot product of input data x and the weight vector w. In ε-SVR, approximation of function f is performed by finding an ε-insensitive tube as flat as possible, which is formally referred to as *flatness*, i.e., seeking a small w. This can be done by minimizing the norm of w. Therefore, we can write the approximation of f as follows:

$$\min_w \frac{1}{2}||w||^2,$$
$$\text{subject to } \begin{cases} y_i - w^T x_i - b \leq \varepsilon \\ w^T x_i + b - y_i \leq \varepsilon \end{cases} \tag{7.2}$$

From (7.2), we can see that the essence of ε-SVR is to perform a linear regression with an ε-insensitive loss function, penalizing predictions that are farther than ε from the desired output. The value of ε is a key factor affecting the flatness of the tube, where a small value leads to a narrow tube, thus a low tolerance for prediction errors, and a large value leads to a broad tube, thus a high error tolerance. There are several popular ε-insensitive loss functions, e.g., the linear and quadratic functions shown, respectively, in (7.3) and (7.4) below.

$$L(y, f(x)) = \begin{cases} 0 & \text{if } |y - f(x)| \leq \varepsilon \\ |y - f(x)| - \varepsilon & \text{otherwise} \end{cases} \tag{7.3}$$

$$L(y,f(x)) = \begin{cases} 0 & \text{if } |y - f(x)| \le \varepsilon \\ (|y - f(x)| - \varepsilon)^2 & \text{otherwise} \end{cases} \tag{7.4}$$

In the rest of this chapter, we will keep using this linear loss function for illustration.

The above optimization (7.2) is feasible in the case that a function f actually exists that approximates all input data points (x_i, y_i) with ε precision. However, in practical applications, there are usually outlier points that deviate from the majority of the data, as illustrated in Fig. 7.1B. Therefore, there is a need for a model that allows for prediction errors. Similar to the idea from a soft-margin SVM classification approach that introduces slack variables to account for noisy data at the hyperplane boundaries (Cortes & Vapnik, 1995), slack variables ξ and ξ^* can be added to (7.2) to guard against the outliers, as shown in Fig. 7.1B. These two slack variables determine how many data points can be tolerated outside the ε-insensitive tube. The original optimization problem in (7.2) is now written as a multiobjective optimization problem with additional parameters ξ and ξ^*

$$\min_w \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*),$$

$$\text{subject to } \begin{cases} y_i - w^T x_i - b & \le & \varepsilon + \xi_i \\ w^T x_i + b - y_i & \le & \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \ge & 0 \end{cases} \tag{7.5}$$

where $C > 0$ is a regularization parameter that determines the trade-off between the flatness of function f and the prediction errors. A large C value gives more weight to minimizing the prediction errors, while a small C value gives more weight to minimizing the flatness. In this case, the ε-insensitive loss function

$$L(y, f(x)) = L(\xi) = \begin{cases} 0 & \text{if } |\xi| \le \varepsilon \\ |\xi| - \varepsilon & \text{otherwise} \end{cases} \tag{7.6}$$

7.2.3 Kernel SVR

The above section describes a linear ε-SVR model, dealing with input data in their feature space and assuming function $f(x)$ is a linear function. To allow ε-SVR to handle nonlinear data, we can introduce a kernel function that transforms the original input data to a higher-dimensional space, referred to as a *kernel space*. In machine learning and SVM, the "kernel trick" is well known because it can be used to learn nonlinear

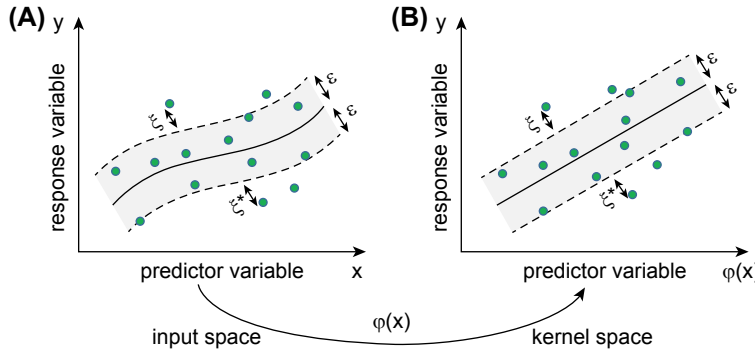

FIGURE 7.2 Graphic representation of nonlinear ε-SVR. A mapping function φ is used to transform the data from the input space (A), where no linear separation of the data is possible, to a higher-dimensional kernel space (B), where the data can be separated by a linear hyperplane. *SVR*, support vector regression.

decision boundaries (Ben-Hur et al., 2008; Muller et al., 2001; Orrù et al., 2012; Schölkopf et al., 2002). In this case, a nonlinear kernel is used to implicitly map the data from the input space (where no linear separation of the data is possible) to a higher-dimensional kernel space (where the data can be separated by a linear hyperplane).

In SVR, data can be discriminated using a linear function in the kernel space and the optimization can be solved following the same computation for the above nonlinear model. Fig. 7.2 illustrates a graphic representation of nonlinear ε-SVR. Here, we have a mapping function to transform the input feature R^d (Fig. 7.2A) into a kernel space F (Fig. 7.2B). Using kernels is one of the most common approaches in SVM (for regression and classification) because there is no need of solving a high-order separating hypersurface in the input space, which is highly complicated compared to solving a linear optimization in the kernel space.

$$\varphi(\bullet): R^d \rightarrow F \tag{7.7}$$

Given (7.7), we can write a linear function $f(x)$ in terms of $\varphi(x)$ as follows:

$$y = f(x) = \langle w, \varphi(x) \rangle + b = w^T \varphi(x) + b \tag{7.8}$$

Then, the optimization problem of $f(x)$ can be written, corresponding to (7.3),

$$\min_w \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*),$$

$$\text{subject to } \begin{cases} y_i - w^T\varphi(x_i) - b & \leq & \varepsilon + \xi_i \\ w^T\varphi(x_i) + b - y_i & \leq & \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq & 0 \end{cases} \tag{7.9}$$

Note that in the nonlinear setting, this optimization problem is to solve the flatness in the kernel space (Fig. 7.2B), not in the input data space. Following the Karush-Kuhn-Tucker (KKT) conditions (referred to (Karush, 1939; Kuhn & Tucker, 1951) for details), we can have (7.9) represented, in terms of support vectors, as the following dual optimization problem

$$\max_{\alpha,a^*} -\varepsilon\sum_{l}^{l_{sv}}(a_i+\alpha_i^*)+\sum_{l}^{l_{sv}}y_i(a_i-\alpha_i^*)-\frac{1}{2}\sum_{i}^{l_{sv}}\sum_{j}^{l_{sv}}(\alpha_i-\alpha_i^*)$$
$$\times\left(\alpha_j-\alpha_j^*\right)k(x_i,x_j)$$

where $k(x_i,x_j)=\varphi(x_i)\varphi(x_j)$, subject to

$$\sum_{i}^{l_{sv}}(a_i-\alpha_i^*)=0, a_i,\alpha_i^*\in[0,C] \tag{7.10}$$

Here, $k(.)$ is the kernel function. Then, the expansion of w and f can be written as

$$w=\sum_{i}^{l_{sv}}(a_i-\alpha_i^*)\varphi(x_i) \tag{7.11}$$

and

$$f(x)=\sum_{i}^{l_{sv}}(a_i-\alpha_i^*)k(x_i,x)+b \tag{7.12}$$

As for the kernel function $k(.)$, there are several popular functions, such as linear kernels (corresponding to the linear ε-SVR model), polynomial kernels, radial basis function (RBF) kernels, and ANOVA RB kernels. Selection of a kernel function depends on distribution of the input data. For example, the linear kernel, which is the simplest of all, is useful when the input is large sparse data vectors. The polynomial kernel is widely used in image processing. The RBF kernel is a general-purpose kernel that is mostly applied in the absence of prior knowledge. The ANOVA RB kernel is usually reserved for regression tasks (Awad & Khanna, 2015).

7.2.4 V-SVR model

Many studies have been done to extend the ε-SVR model for algorithm improvements. In this section, we focus on the v-SVR model, which is one of the most popular modifications proposed by Schölkopf, Bartlett, Smola, and Williamson (1999). The benefit of v-SVR is that it provides a way to automatically minimize ε. In ε-SVR, selection of a proper ε value is

essential for an accurate regression approximation. However, it is difficult to specify ε beforehand, other than an empirical choice. In v-SVR, a new parameter of a prior $v \in (0, 1)$ is introduced to automatically adjust a flexible tube by controlling the number of support vector and tolerated training errors. Then, the parameter ε becomes a variable in the optimization process and is controlled by the new parameter v.

In v-SVR, the optimization problem can be written, given a function $\varphi(x)$ to the kernel space for a nonlinear case, as follows

$$\min_w \frac{1}{2}||w||^2 + C\left(v\varepsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi + \xi^*)\right),$$

$$\text{subject to } \begin{cases} y_i - w^T\varphi(x_i) - b & \leq & \varepsilon + \xi \\ w^T\varphi(x_i) + b - y_i & \leq & \varepsilon + \xi^* \\ \xi_i, \xi_i^*, \varepsilon & \geq & 0 \end{cases} \tag{7.13}$$

Here, the newly introduced constant variable $v \in (0, 1)$ is used as a trade-off against model complexity and slack variables. Forming a Lagrangian formulation from (7.7) by introducing positive multipliers α, α^*, η, η^*, and β gives

$$\begin{aligned} L(w, \xi, \xi^*, \alpha, a^*, \eta, \eta^*, \beta) = {} & \frac{1}{2}||w||^2 + Cv\varepsilon + \frac{C}{l}\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right) \\ & + \sum_{i=1}^{l} a_i^*\left(y_i - w^T x_i - b - \varepsilon - \xi_i\right) + \sum_{i=1}^{l}\alpha_i\left(w^T x_i + b - y_i - \varepsilon - \xi_i^*\right) \\ & - \sum_{i=l}^{l}\left(\eta_i\xi_i + \eta_i^*\xi_i^*\right) - \beta\varepsilon \end{aligned} \tag{7.14}$$

Following the KKT conditions that partial derivatives with respect to the variables w, b, ξ, ξ^*, and ε are equal to be zero and the products of the Lagrange multipliers and the constraint are equal to zero, we have the following dual optimization problem of v-SVR

$$\max_{\alpha, a^*} \sum_{l}^{l_{sv}} y_i\left(a_i - \alpha_i^*\right) - \frac{1}{2}\sum_{i}^{l_{sv}}\sum_{j}^{l_{sv}}\left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right)k(x_i, x_j)$$

where $k(x_i, x_j) = \varphi(x_i)\varphi(x_j)$ subject to

$$\sum_{i}^{l_{sv}}\left(a_i - \alpha_i^*\right) = 0, a_i, \alpha_i^* \in \left[0, \frac{C}{l}\right], \sum_{l}^{l_{sv}}\left(a_i + \alpha_i^*\right) \leq Cv \tag{7.15}$$

Then, the regression estimate takes the form

$$f(x) = \sum_{i}^{l_{sv}} (a_i - \alpha_i^*) k(x_i, x) + b \tag{7.16}$$

Compared to the optimization problem in ε-SVR (7.14), we can see that the parameter ε vanishes but instead there is the new parameter v in v-SVR (7.15). Schölkopf et al. (1999) had proved that $v \in (0, 1)$ is an upper bound on the fraction of errors (i.e., data points outside of the tube divided by the total number of data points l) and a lower bound on the fraction of support vectors (i.e., the numbers of support vectors divided by the total number of data points l). Therefore, in practical applications, a user only needs to specify a fraction of data points that is allowed to become errors using v-SVR, other than to give a specific level of accuracy as a priori.

7.2.5 Parameter selection using cross-validation

Multiple parameters are involved in SVR models. For example, in ε-SVR, there are the regularization parameters C and ε, and in v-SVR, there are the regularization parameters C and v. For nonlinear models, there may be additional kernel function parameters, e.g., σ in an RBF kernel. In practice, the most widely used approach to determine these parameters is cross-validation, which assesses how the results of a statistical analysis can generalize to an independent dataset (Devyver & Kittler, 1982; Kohavi, 1995) (see Chapter 2). For example, in a linear ε-SVR, given certain values of parameters C and ε, an SVR model is trained on a subset of the dataset (training dataset). Then, the remaining data (testing set) are used to test the trained model by comparing the predicted output y_i' and the known target output y_i. The differences between y_i' and y_i provide a quantitative assessment of the prediction accuracy. There are many measurements to quantify the difference between the actual and predicted outputs, including mean squared error, mean absolute error, R^2 (explained variation/total variation), and so on (see Chapter 2).

7.3 Applications to brain disorders

In this section, we describe how SVR can be used for study of brain disorders. We focus on the SVR applications that use neuroimaging information, in particular MRI data. MRI provides important information about brain anatomy and has been widely used in neuroscience and the study of brain disorders (Li et al., 2017; Lorenzetti et al., 2009; Shenton et al., 2001; Yu-Feng et al., 2007). There are many MRI acquisition

techniques, where each technique provides specific anatomical information about the brain. For example, T1-weighted or T2-weighted structural MRI can be used for brain gray matter segmentation. This allows measurement of cortical thickness, which has been studied during brain maturation using SVR, e.g., in schizophrenia patients (Koutsouleris et al., 2014; Schnack et al., 2016). In addition to T1- and T2-weighted images, additional acquisitions such as diffusion MRI and functional MRI (fMRI) allow investigation of brain white matter structural connectivity and brain functional connectivity, respectively. Many studies have applied SVR with diffusion or fMRI to investigate different brain disorders (Cha et al., 2015; Iuculano et al., 2014). Below, we describe studies that use SVR with three different MRI acquisition techniques (T1-weighted MRI, diffusion MRI, and fMRI) to investigate brain disorders including schizophrenia, autism spectrum disorder, and ADHD. For each study, we give a brief outline of the aim, the methodology, and the results, while the readers are referred to the published papers for details.

7.3.1 Predicting deviations from "brain age" in schizophrenia

Koutsouleris et al. (2014) applied SVR to study whether patients with schizophrenia deviated from the trajectory of normal brain maturation. The authors measured this deviation for a subject by computing the difference between the subject's chronological age and the subject's neuroanatomical age estimated from structural MRI images. To do this, an SVR model of neuroanatomical brain maturation was first trained using MRI data from 800 healthy controls and then applied to a new subject (e.g., a schizophrenia patient) for estimation of a neuroanatomical age. In more details, the authors segmented gray matter voxels from the T1-weighted MR image and computed a gray matter density map and a gray matter volume map of each healthy control. A dimensionality reduction of each of the gray matter maps was performed using principal component analysis (PCA), resulting in a relatively low-dimensional feature per map ($D = 400$). The two low-dimensional features were concatenated together as one feature ($D = 800$), which was then used as the input feature. The target output was the chronological age. A linear ν-SVR algorithm was used to learn the model from the input feature to the respective target output. The authors used a nested cross-validation (Filzmoser, Liebmann, & Varmuza, 2009) to select optimal parameters C and ν in ν-SVR, where the out-of-training prediction performance was quantified using the mean absolute error and explained variance between the predicted age and the subject's actual chronological age. Then, given a new subject (e.g., a schizophrenia patient), the same feature extraction step was performed, including gray matter voxel segmentation, gray matter density and volume map computation, and PCA feature reduction. The obtained

feature was fitted into the leaned SVR model for age prediction. The difference between the predicted age and the actual chronological age was used for an estimation of deviation from the trajectory of normal brain maturation of this subject. Using this SVR-based approach, the authors compared the computed brain maturation deviation in three different populations including patients with schizophrenia ($n = 141$), major depression ($n = 104$), and borderline personality disorder ($n = 57$). They found that the schizophrenia patients had significantly higher brain maturation deviations than the other compared diseases.

In another investigation of schizophrenia, Schnack et al. (2016) applied SVR to study whether the progressive brain loss in schizophrenia patients reflect accelerated aging of the brain, using longitudinal (baseline and follow-up) neuroimaging data. The authors applied a similar idea to that in the above study to predict a neuroanatomical age[1] using a T1-weighted MR image. Specifically, a neuroanatomical age prediction model was trained using input high-dimensional gray matter density maps to predict chronological age. The model was trained using baseline T1-weighted images of healthy subjects ($n = 386$). The difference between the predicted age and the chronological age was recorded as the "brain age gap" at baseline for each healthy subject. Then, this trained model was subsequently applied to all follow-up images of the healthy subjects for a computation of the brain age gap at follow-up. In a similar way, the authors applied the trained model to compute brain age gaps from the baseline and follow-up scans of the schizophrenia patients ($n = 341$). The authors found that the brain age gaps at baseline and follow-up in the healthy group were -0.0017 and -0.045 years (not significantly different from 0). However, in the schizophrenia group, the brain age gaps at baseline and follow-up were 3.36 and 4.72 years (both significantly different from 0). These results suggested that progressive changes in the gray matter morphology of schizophrenia patients resemble, and possibly reflect, an accelerated aging process.

7.3.2 Predicting symptoms severity and mathematical abilities in autism spectrum disorder

Moradi et al. (2017) applied SVR to predict symptom severity of individuals with autism spectrum disorder based on cortical thickness measurements computed from MRI data. The authors were motivated by evidence that cortical thickness measurements provide an index of the maturation of cortex and cortico-cortical connectivity (Raznahan et al.,

[1] Here, "neuroanatomical age" is used to be consistent with the term used in the previous study from Koutsouleris et al. (2014). In the original paper from Schnack et al. (2016), they refer to this as "brain age."

2011; Shaw et al., 2008), and autism spectrum disorder may be characterized by delayed maturation (Johnson, Gliga, Jones, & Charman, 2015; Webb et al., 2011) In this study, a dataset of 156 subjects from the Autism Brain Imaging Data Exchange (ABIDE) project (Di Martino et al., 2014) was analyzed, in which each subject was associated with an autism severity score based on behavioral evaluations of social interaction and communication. The goal of this study was to build a predictive model of the severity score using cortical thickness measurements. Specifically, T1-weighted MR images were used to segment 78 cortical regions for each subject, where a cortical thickness measurement was computed for each region. During the learning stage, an ε-SVR model was trained for each of these regions separately to predict a severity score from the computed cortical thickness measurements. This resulted in a total of 78 trained models, where each allowed for a region-specific severity score prediction. These 78 predictions were concatenated into a feature vector ($D = 78$), which provided an overall severity description across all cortical regions for one subject. This feature vector, along with the severity score, was used to build a least squares linear regression with elastic net penalty (Zou & Hastie, 2005) for prediction of a final estimated severity score.[2] Then, for severity score prediction of a new subject, region-specific predictions were performed using the 78 trained SVR models based on the cortical thickness measurements computed from the subject's T1-weighted image, and the results were concatenated to obtain an overall severity feature vector ($D = 78$). A final severity prediction score was computed by feeding the feature vector into the learned elastic net model. For experimental evaluation, the authors applied two nested cross-validation loops, where the outer loop was used to train and test the elastic net model and the inner loop used the training data of the outer loop to train and test the SVR models. The performance was evaluated based on the Pearson correlation coefficient, mean absolute error, and the coefficient of determination between estimated and actual severity score. One of the benefits of this study in using SVR was to provide region-specific models for regional predictions. In the results, the authors showed increased prediction performance when including the regional predictions compared to only using cortical thickness information from the whole brain.

In a second study of autism, Iuculano et al. applied SVR to investigate whether functional brain activations could predict mathematical abilities of children with autism using functional MRI (Iuculano et al., 2014). A population of 18 children with autism and 18 typically developing children

[2] Notice that in this study the authors had two regression steps. SVR was used for a region-specific severity score prediction, and least squares linear regression was used for a final severity score prediction based on the predicted region-specific severity scores.

was studied. Each subject had a *Numerical Operations* score that provided a standardized measure of their math abilities. Task fMRI data were acquired from an experiment consisting of two arithmetic conditions and two nonarithmetic conditions. A multivariate pattern analysis (Cho, Ryali, Geary, & Menon, 2011) was used to identify brain regions that had spatial functional activation patterns that could discriminate between the children with autism and the typically developing children. Then, for each group, SVR was used to train a model to predict *Numerical Operations* scores from voxel values of each identified brain region. A leave-one-subject-out cross-validation was used, where, in each group, each subject was designated as the test data in turns while the remaining subjects were used to train the SVR model. R^2 was computed based on the actual scores and predicted scores. To test whether there was a statistical significance of the SVR prediction result, a nonparametric analysis was used. A null distribution of R^2 was generated for each discriminative brain region for a certain group by permuting the *Numerical Operations* scores (10,000 times) across all subjects in this group. The actual R^2 was compared to this null distribution to calculate a P-value, i.e., the number of permutations that had R^2 greater than the R^2 value divided by the total number of permutations ($N = 10{,}000$). In the results, the authors found that numerical abilities in the autism group were predicted by the pattern of neural activity in an area of the left ventral temporal-occipital cortex encompassing the left fusiform gyrus and lateral occipital cortex, with a significance of $P = .04$ and $R^2 = 0.69$.

7.3.3 Predicting aggression in attention-deficit/hyperactivity disorder

Cha et al. (2015) applied SVR to investigate how abnormalities of white matter structural connectivity within the fronto-accumbal circuitry relate to aggression in children with ADHD. A population of 30 children with ADHD was studied, where each subject was associated with an aggression score. The authors used diffusion MRI that enables noninvasive mapping of the brain's white matter connections via a computational process called tractography (Basser, Pajevic, Pierpaoli, Duda, & Aldroubi, 2000). They performed probabilistic tractography of each subject (Behrens et al., 2003), and they extracted the white matter fibers connecting to the nucleus accumbens and the ventral prefrontal cortex.[3] The number of fibers was computed to measure white matter connectivity of the fronto-accumbal circuit. Then, a linear SVR model was built to predict aggression

[3] Six cortical regions were included the medial orbitofrontal (mOFC), lateral orbitofrontal (lOFC), pars orbitalis (Pars Orb), rostral anterior cingulate (rACC) cortex, frontal pole (FP), and medial prefrontal cortex (mPFC).

scores from the number of fibers. A two-layer nested leave-one-subject-out cross-validation was used, where within each run a predicted aggression score was computed for the left-out subject. Across all runs, a vector of predicted scores of all subjects was obtained and compared with the actual aggression scores using a correlation analysis for a correlation coefficient. The significance of this correlation was assessed through bootstrapping where we permuted the aggression scores randomly 1000 times and then applied the same CV procedure to each of the resulting datasets to yield an empirical estimate of the significance of the predication (this is similar to the significance computation of R^2 in Iuculano et al. (2014)). The authors found that the left fronto-accumbal tract measures had significant correlation with aggression with a P-value $= .04$. This suggested that the patterns of fronto-accumbal white matter connectivity in children with ADHD were predictive of the level of aggression in these children.

7.4 Conclusion

In this chapter, we have described the underlying principles of SVR. SVR is a regression tool to analyze the relationship between a continuous dependent variable and one or more predictor variables. The optimization in SVR is represented in terms of support vectors, where the optimization solution does not depend on the dimension of the input data but only depends on the number of support vectors. As a result, SVR provides an effective tool to handle high-dimensional data. In addition, SVR is a machine learning method that learns a model to describe a variable's importance in characterizing the relationship between input and output, unlike a traditional regression method that depends on the assumption of a model (e.g., linear data distribution) that might not be accurate. Owing to these properties, SVR has been adopted in many studies of brain disorders. In particular, SVR has been most often applied in studies using MRI for a multivariate pattern regression analysis. It enables the analysis of spatially distributed patterns across multiple brain regions that may be related to certain brain disorder effects. In this chapter, we have discussed multiple applications of SVR to study different kinds of brain disorders using MRI.

We suggest that the following points should further be considered beyond the content introduced in this chapter. First, while we have described the underlying principles of SVR, more information about SVR and SVM can be found in Smola and Schölkopf (2004) and Vapnik (1998). These can help a reader to have a better understanding of theoretical and mathematical details. Second, in addition to the classical ε-SVR model, we have discussed an improved model, v-SVR. While these two are the most

widely used models in current brain disorder applications, there are many other advanced models that can potentially be more useful and accurate in certain types of applications. For example, advanced SVR models have been proposed and demonstrated to have a good performance in studying Alzheimer's disease (Wang, Fan, Bhatt, & Davatzikos, 2010; Zhang, Shen & Alzheimer's Disease Neuroimaging Initiative, 2012). Third, we have discussed the applications of SVR in studying brain disorders using MRI data. Readers desiring to apply SVR in their research may also be interested in its applications to other types of biomedical data, including genetic (Chen et al., 2010) and histological (Du & Dua, 2011) data. Overall, SVR is a useful tool for regression analysis in the study of brain disorders.

7.5 Key points

- SVR is a machine learning regression method which allows for prediction of continuous real-valued variables.
- SVR provides a sparse solution for optimizing of regression loss function using a subset of input data, i.e., support vectors.
- SVR provides good performance to handle high-dimensional data.
- SVR allows multivariate pattern regression analysis of different regions across the brain for neuroimage-based studies of brain disorders.
- SVR performance depends on the choice of kernels when handling nonlinear data.
- SVR has been applied to study many brain disorders, including schizophrenia, autism, and ADHD.

References

Awad, M., & Khanna, R. (2015). *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*. Apress.

Basser, P. J., Pajevic, S., Pierpaoli, C., Duda, J., & Aldroubi, A. (2000). In vivo fiber tractography using DT-MRI data. *Magnetic resonance in medicine: Official Journal of the Society of magnetic resonance in medicine/Society of Magnetic Resonance in Medicine, 44*(4), 625–632.

Behrens, T. E. J., Woolrich, M. W., Jenkinson, M., Johansen-Berg, H., Nunes, R. G., Clare, S., et al. (2003). Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magnetic Resonance in Medicine: Official Journal of the Society of Magnetic Resonance in Medicine/Society of Magnetic Resonance in Medicine, 50*(5), 1077–1088.

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., & Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS Computational Biology, 4*(10), e1000173.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 144–152). New York, NY, USA: ACM.

Cha, J., Fekete, T., Siciliano, F., Biezonski, D., Greenhill, L., Pliszka, S. R., et al. (2015). Neural correlates of aggression in medication-naive children with ADHD: Multivariate analysis of morphometry and tractography. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology, 40*(7), 1717–1725.

Chen, L., Xuan, J., Riggins, R. B., Wang, Y., Hoffman, E. P., & Clarke, R. (2010). Multilevel support vector regression analysis to identify condition-specific regulatory networks. *Bioinformatics, 26*(11), 1416–1422.

Cho, S., Ryali, S., Geary, D. C., & Menon, V. (2011). How does a child solve 7 + 8? Decoding brain activity patterns associated with counting and retrieval strategies. *Developmental Science, 14*(5), 989–1001.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Craddock, R. C., Holtzheimer, P. E., III, Hu, X. P., & Mayberg, H. S. (2009). Disease state prediction from resting state functional connectivity. *Magnetic Resonance in Medicine: Official Journal of the Society of Magnetic Resonance in Medicine/Society of Magnetic Resonance in Medicine, 62*(6), 1619–1628.

Cuingnet, R., Rosso, C., Chupin, M., Lehéricy, S., Dormont, D., Benali, H., et al. (2011). Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome. *Medical Image Analysis, 15*(5), 729–737.

Devyver, P. A., & Kittler, J. (1982). *Pattern recognition: A statistical approach*. Prentice-Hall.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., et al. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry, 19*(6), 659–667.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). Support vector regression machines. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.), *Advances in neural information processing systems* (pp. 155–161). MIT Press.

Du, X., & Dua, S. (2011). Cancer prognosis using support vector regression in imaging modality. *World Journal of Clinical Oncology, 2*(1), 44–49.

Fan, Y., Shen, D., Gur, R. C., Gur, R. E., & Davatzikos, C. (2007). Compare: Classification of morphological patterns using adaptive regional elements. *IEEE Transactions on Medical Imaging, 26*(1), 93–105.

Filzmoser, P., Liebmann, B., & Varmuza, K. (2009). Repeated double cross validation. *Journal of Chemometrics, 23*(4), 160–171.

Glaser, J. I., Benjamin, A. S., Farhoodi, R., & Kording, K. P. (2019). The roles of supervised machine learning in systems neuroscience. *Progress in Neurobiology, 175*, 126–137.

Iuculano, T., Rosenberg-Lee, M., Supekar, K., Lynch, C. J., Khouzam, A., Phillips, J., et al. (2014). Brain organization underlying superior mathematical abilities in children with autism. *Biological Psychiatry, 75*(3), 223–230.

Johnson, M. H., Gliga, T., Jones, E., & Charman, T. (2015). Annual research review: Infant development, autism, and ADHD–early pathways to emerging disorders. *The Journal of Child Psychology and Psychiatry and Allied Disciplines, 56*(3), 228–247.

Karush, W. (1939). *Minima of functions of several variables with inequalities as side constraints, M. Sc. Dissertation*. Univ. of Chicago: Dept. of Mathematics. Retrieved from https://ci.nii.ac.jp/naid/10027639655/.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *, Vol. 14. Ijcai* (pp. 1137–1145). Montreal: Canada.

Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., et al. (2014). Accelerated brain aging in schizophrenia and beyond: A neuroanatomical marker of psychiatric disorders. *Schizophrenia Bulletin, 40*(5), 1140–1153.

Kuhn, H. W., & Tucker, A. W. (1951). Nonlinear programming. In *Proceedings of the second berkeley symposium on mathematical statistics and probability*. The Regents of the University of California . Retrieved from https://projecteuclid.org/euclid.bsmsp/1200500249.

Li, D., Karnath, H.-O., & Xu, X. (2017). Candidate biomarkers in children with autism spectrum disorder: A review of MRI studies. *Neuroscience Bulletin, 33*(2), 219–237.

Linn, K. A., Gaonkar, B., Doshi, J., Davatzikos, C., & Shinohara, R. T. (2016). Addressing confounding in predictive models with an application to neuroimaging. *International Journal of Biostatistics, 12*(1), 31–44.

Lorenzetti, V., Allen, N. B., Fornito, A., & Yücel, M. (2009). Structural brain abnormalities in major depressive disorder: A selective review of recent MRI studies. *Journal of Affective Disorders, 117*(1–2), 1–17.

Moradi, E., Khundrakpam, B., Lewis, J. D., Evans, A. C., & Tohka, J. (2017). Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data. *NeuroImage, 144*(Pt A), 128–141.

Muller, K., Mika, S., Ratsch, G., Tsuda, K., & Scholkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks, 12*(2), 181–201.

Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neuroscience and Biobehavioral Reviews, 36*(4), 1140–1152.

Raznahan, A., Lerch, J. P., Lee, N., Greenstein, D., Wallace, G. L., Stockman, M., et al. (2011). Patterns of coordinated anatomical change in human cortical development: A longitudinal neuroimaging study of maturational coupling. *Neuron, 72*(5), 873–884.

Schnack, H. G., van Haren, N. E. M., Nieuwenhuis, M., Hulshoff Pol, H. E., Cahn, W., & Kahn, R. S. (2016). Accelerated brain aging in schizophrenia: A longitudinal pattern recognition study. *American Journal of Psychiatry, 173*(6), 607–616.

Schölkopf, B., Bartlett, P. L., Smola, A. J., & Williamson, R. C. (1999). Shrinking the tube: A new support vector regression algorithm. In M. J. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems* (pp. 330–336). MIT Press.

Schölkopf, B., Smola, A. J., & Bach, F. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press.

Shaw, P., Kabani, N. J., Lerch, J. P., Eckstrand, K., Lenroot, R., Gogtay, N., et al. (2008). Neurodevelopmental trajectories of the human cerebral cortex. *Journal of Neuroscience, 28*(14), 3586–3594.

Shenton, M. E., Dickey, C. C., Frumin, M., & McCarley, R. W. (2001). A review of MRI findings in schizophrenia. *Schizophrenia Research, 49*(1), 1–52.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199–222.

Vapnik, V. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control, 24*, 774–780.

Vapnik, V. (1964). *A note one class of perceptrons. Automation and Remote Control*. Retrieved from https://ci.nii.ac.jp/naid/10021840590/.

Vapnik, V. N. (1998). *Statistical learning theory*. Wiley.

Wang, Y., Fan, Y., Bhatt, P., & Davatzikos, C. (2010). High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables. *NeuroImage, 50*(4), 1519–1535.

Webb, S. J., Jones, E. J. H., Merkle, K., Venema, K., Greenson, J., Murias, M., et al. (2011). Developmental change in the ERP responses to familiar faces in toddlers with autism spectrum disorders versus typical development. *Child Development, 82*(6), 1868–1886.

Yu-Feng, Z., Yong, H., Chao-Zhe, Z., Qing-Jiu, C., Man-Qiu, S., Meng, L., et al. (2007). Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI. *Brain and Development, 29*(2), 83–91.

Zhang, F., Savadjiev, P., Cai, W., Song, Y., Rathi, Y., Tunç, B., et al. (2018). Whole brain white matter connectivity analysis using machine learning: An application to autism. *NeuroImage, 172*, 826–837.

Zhang, D., Shen, D., & Alzheimer's Disease Neuroimaging Initiative. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage, 59*(2), 895–907.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320.