Feature Selection Using Gram-Schmidt Orthogonalization For Support Vector Regression – A Case Study Of Mortality Rate Prediction Caused By Pneumonia

Yuni R Dewi¹, Hendri Murfi²

Department of Mathematics Universitas Indonesia, Depok 16424, Indonesia

E-mail: ¹yuni.rosita@sci.ui.ac.id, ²hendri@ui.ac.id

Abstract. Feature selection is a technique for finding optimal features among original features by eliminating irrelevant features. Besides improving the learning accuracy and facilitate a better understanding of the model, feature selection may reduce the cost of building, storing and processing models. Recently, a Gram-Schmidt Orthogonalization-based feature selection is proposed for unstructured data. In this paper, we extend this Gram-Schmidt Orthogonalization-based feature selection for structured data. Our simulation shows that this Gram-Schmidt Orthogonalization-based feature selection improves the accuracy of Support Vector Regression in the average of 1.384925% for the case study of the prediction of mortality rates caused by pneumonia.

1. Introduction

The current developing science and technology are used to help human work in many disciplines, i.e., data science. Data science is a new field of science that grows very fast and has many fields of applications. In data science, the existing data is used to show information or mine knowledge from the data. The core model to mine knowledge of a collection of data is machine learning, i.e., support vector machine (SVM) [1]. SVM was originally used for classification problems and then was developed to solve regression problems. The model is known as support vector regression (SVR). SVR solves the regression problem by solving quadratic programming. The advantages of this model are its goal to minimize the generalization error, and the quadratic programming is convex. Several studies also show that SVR can provide higher accuracy than other machine learning models for several real-world problems [2, 3].

The data are usually characterized by some features. However, some features of the existing data are often not suitable for specific knowledge. Therefore, we need a feature selection to overcome the problem. Feature selection is a technique to find an optimal set of features among the set of original features by removing irrelevant and redundant data. Therefore, feature selection can improve the learning accuracy and facilitate a better understanding of the model. Moreover, by reducing the dimensions, feature selection may give other benefits including the cost of building, storing and processing models [4, 5, 6].

There are some feature selection methods for SVR, i.e., univariate selection [7], recursive feature elimination (RFE) [7,8], and naïve bayes [9]. We consider another new approach for feature selection called the Gram-Schmidt method. In algebra, the Gram-Schmidt method is a method to orthogonalize a set of vectors [10]. Arora et al. used the orthogonalization method to find anchor words for separable
nonnegative matrix factorization [11]. Next, Wang et al. continued the work by using the Gram-Schmidt method for feature selection from a word co-occurrence of textual data [12].

In this paper, we extend the Gram-Schmidt Orthogonalization based feature selection method for structured data. We use the problem of mortality rate prediction as our case study. Mortality rate defined as a measure of the death of the population in a particular place, time, and condition [13]. The mortality rate can be influenced by various factors, i.e., age, gender, environmental conditions, economic conditions, diseases. In this case study, we consider the prediction of mortality rates based on pneumonia. Our simulation shows that the feature selection using the Gram-Schmidt orthogonalization increases the accuracy of SVR for mortality rate prediction caused by pneumonia in the average of 1.384925%.

This paper is organized as follows: Section 2 describes the related research and the differences in this study from previous research. Section 3 explains the method used. Section 4 shows the data used and the steps taken during the study. Sections 5 shows the result and discussion about the results. Finally, Section 6 presents the conclusion.

2. Related Work
In recent years, the feature selection method was studied widely. There are many feature selections method that combines with Support Vector Machine (SVM) [14, 15, 16] such as Recursive Feature Elimination (RFE), Generalized Benders Decomposition, and proposed a novel strategy based on the evaluation of local kernel gradients in the vicinity of decision hyperplane.

Cinelli et al. [9] used feature selection methods to improve the accuracy of SVM to classify the CDR3 repertoires. As a result, SVM can achieve a high level of classification by testing the validity of more than 90% of these problems. In other words, the accuracy of SVM in classifying CDR3 increases if the feature selection process uses one dimension classification naïve Bayes.

Other studies from Goli et al. [7] have researched the survival prediction and feature selection in the patient with breast cancer using Support Vector Regression. In their research, they used univariate feature selection, RFE and another method (a combination of SVR and statistical tests) method for feature selection. Based on the three feature selection methods, metastasis status, progesterone receptor status, and human epidermal growth factor receptor two status were the best features associated with survival.

Meanwhile, Arora et al. used the orthogonalization method to find anchor words for separable nonnegative matrix factorization [11]. Next, Wang et al. extended the orthogonalization method from Arora to be Gram Schmidt Orthogonalization for feature selection in the case of the perspective of the appearing words that inspired from the anchor word [12]. It means in the previous study Gram-Schmidt Orthogonalization is used to unstructured data, specifically to classifying text and grouping test. In this paper, we examine the Gram-Schmidt Orthogonalization as a new feature selection method for structured or tabular data.

3. Method
This section explains the method related to the prediction of mortality rates using Support Vector Regression with the Gram-Schmidt Orthogonalization-based feature selection.

3.1. Support Vector Regression (SVR)
Support Vector Machine (SVM) which was first presented by Vapnik. There are two main categories of Support Vector Machines: Support Vector Classification (SVC) and Support Vector Regression (SVR). SVM is a learning system using a high dimensional feature space. It yields prediction functions that expanded on a subset of support vectors. SVM can generalize complicated gray level structures with only a very few support vectors and thus provides a new mechanism for image compression. A version of SVM for regression has been proposed in 1997 by Vapnik, Steven Golowich, and Alex Smola.

Suppose a training data \( \{x_n, t_n\}, n = 1 \text{ to } N \), with \( x_n \) is \( n \)-th training data and \( t_n \in \mathbb{R} \). The problems are how to get model that represented training data and how to find model \( y(\alpha) \) that give a little error for
all training data $x_n$ [1]. The general form of the linear model that will be used as a regression function defined by:

$$y(x) = w^T \phi(x) + b$$

(1)

where $x$ is an input vector, $w$ is a weight parameter, $\phi(x)$ is a feature transformation function, and $b$ is a bias. In this case, we will minimize the regularized error function:

$$E(w, b) = \frac{1}{2} \sum_{n=1}^{N} (y(x_n) - t_n)^2 + \frac{\lambda}{2} ||w||^2$$

(2)

In SVR model, we will minimize the regularized error function:

$$E(w, b) = C \sum_{n=1}^{N} E_{\varepsilon}(y(x_n) - t_n)^2 + \frac{1}{2} ||w||^2$$

(3)

where $E_{\varepsilon}(\cdot)$ is a $\varepsilon$- sensitive error function that defined below:

$$E_{\varepsilon}(y(x_n) - t_n) = \begin{cases} 0, & \text{if } |y(x_n) - t_n| < \varepsilon \\ |y(x_n) - t_n| - \varepsilon, & \text{others} \end{cases}$$

(4)

After that, determination of SVR model parameters can be formulated as an optimization problem below:

$$\arg \min_{w,b} C \sum_{n=1}^{N} (\xi_n + \tilde{\xi}_n) + \frac{1}{2} ||w||^2$$

s. t. \begin{align} \xi_n &\geq 0, \tilde{\xi}_n \geq 0 \\ \tilde{\xi}_n &\leq y(x_n) + \varepsilon + \xi_n \\ \xi_n &\leq y(x_n) - \varepsilon - \tilde{\xi}_n \end{align}$$

(5) (6) (7) (8)

From the data points and the value $\varepsilon$, it has been received for the loss function, one constructs two new datasets, one by shifting the response up by $\varepsilon$, and the other by shifting it down by $\varepsilon$. For a large enough $\varepsilon$ for which hard tubes exist, these shifts amount to move the first dataset to stand fully above the regression hyperplane, and the second dataset to stand fully below it. Thus, the regression hyperplane has been effectively transformed into a classification hyperplane. Algorithms for training support vector machines in classification problems have been extended into algorithms for support vector regression problems under $\varepsilon$-insensitive loss [17].

3.2. Gram Schmidt Orthogonalization

Like many feature selection methods, Gram-Schmidt Orthogonalization has procedure and algorithm to get optimal feature and reduce irrelevant or annoying feature. Therefore, each step or the procedure of the Gram-Schmidt Orthogonalization method is the following [18]:

- After orthogonalization (by classical of modified Gram-Schmidt), pick the projected candidate feature (not selected at previous steps) that has the smallest angle with the projected output,
- Compute the value of the cumulative distribution function as described in the next section,
- If that value is smaller than the risk, keep the feature and perform the next step of Gram-Schmidt orthogonalization
- If that value is larger than the risk, discard the feature under consideration and terminate the procedure

Gram-Schmidt Orthogonalization that used in Feature Selection showed in Algorithm 1 by Arora et al. [11]. In this study, the algorithm used for structured data. The input contains M features and W country.

| Algorithm 1. General algorithm for Feature Selection |
| --- |
| **INPUT** : |
| $D$: The structured data, containing M feature and W country. |
| $K$: The number of basis vectors (basis features). |
| $T$: The column dimension of random matrix $R$, and $T \ll W$ |
| **OUTPUT** : |
| $S$: The $K$ feature indices (indices of basic features) |
4. Simulation

We examined Gram-Schmidt Orthogonalization as feature selection for SVR in a case study of mortality rate prediction caused by pneumonia in this section.

4.1. Data and programming

We use publicly available datasets from Pneumonia through Kaggle to build and evaluate the feature selection and prediction. The kind of data is .csv in Microsoft Excel, and all of the data was processed in Spyder (Python 3.6). By obtaining data, the data is divided into training data and testing data. The training data is used to build a model as a predictor. Using the testing data, we estimate the accuracy of the model. Some elementary data information from these datasets are:

- There are 140 countries of observations in the training set.
- There are 24 features that affect the mortality rate
- The target is a regression value that represented the mortality rate
- Feature titles are country, year, deaths.per.1k, code, total.usd.2013, total.usd.2012, total.usd.2000, total.ppp.2013, total.ppp.2012, total.ppp.2000, govt.usd.2013, govt.usd.2012, govt.usd.2000, govt.ppp.2013, govt.ppp.2012, govt.ppp.2000, pct.deaths.5y, pct.pop.vaccinated, grouping, in.epi, PCV10.Price, PCV13.Price, gni.2012, gni.2013, sanitation.2012, sanitation.2013, birth.rate.2012, birth.rate.2013. Each feature titles means:

| Table 1. Feature description on Kaggle [1]. |
|-------------------------------------------|
| **Feature**                               | **Description**                      |
| total.usd.2013                            | Per capita total health expenditure on average USD for 2013 |
| total.usd.2012                            | Per capita total health expenditure on average USD for 2012 |
| total.usd.2000                            | Per capita total health expenditure on average USD for 2000 |
| total.ppp.2013                            | Per capita total health expenditure in USD scaled for purchasing power parity for 2013 |
| total.ppp.2012                            | Per capita total health expenditure in USD scaled for purchasing power parity for 2012 |
| total.ppp.2000                            | Per capita total health expenditure in USD scaled for purchasing power parity for 2000 |
| govt.usd.2013                             | Per capita total government health expenditure on average USD for 2013 |
| govt.usd.2012                             | Per capita total government health expenditure on average USD for 2012 |
| govt.usd.2000                             | Per capita total government health expenditure on average USD for 2000 |
| govt.ppp.2013                             | Per capita total government health expenditure in USD scaled for purchasing power parity for 2013 |
| govt.ppp.2012                             | Per capita total government health expenditure in USD scaled for purchasing power parity for 2012 |
| govt.ppp.2000                             | Per capita total government health expenditure in USD scaled for purchasing power parity for 2000 |
4.2. Preprocessing data
The purpose of this stage is to make the data complete and ready for processing because not all the data was complete or ready to process. The provided data contains missing data, so that the preprocessing data stage is needed, such as overcoming missing data using fillNaN, which is to fill in the empty data with a value of 0 (zero), or use the average value (mean), the middle value (median) and the values that often appear or most frequent (modus).

The other step in preprocessing data is changing the categorical data "Yes" and "No" to numerical data. The categorical data are replaced by "1" and "0". Besides that, in this step, the feature that contained character type will be deleted. In this case study, the deleted feature is “grouping”.

4.3. Processing data
Data processing uses the Support Vector Machine (SVM) method. SVM was developed by Boser, Guyon & Vapnik which was first presented in 1992 at the Annual Workshop on Computational Learning Theory. Because the data used a regression target, we use Support Vector Regression (SVR). The basic idea of SVR is to find the \( f(x) \) model function to represent the relationship between features and targets \(^1\).

4.4. Learning process
The learning process aims to determine the parameters of the method in the training data provided. In SVM it takes two parameters, namely the parameters \( C \) and \( \gamma \) (gamma). Parameter \( C \) is a parameter that is used to measure the trade-off from estimating the target training data, while the parameter \( \gamma \) if the value is getting bigger, it indicates that some training data is increasingly affecting other training data. The approach used to obtain the optimal combination of \( C \) and \( \gamma \) is the Grid Search approach. In practice, optimal parameters \( C \) and \( \gamma \) with intervals of \( 10^{-3} \) to \( 10^{3} \) is sufficient \(^1,19\).

4.5. Model evaluation
The model evaluation process aims to determine the accuracy of the regressor function model in predicting new data that does not include training data. Cross-Validation which is part of Grid Search is used to share data into training data and testing data, and then calculate the accuracy of the regressor function model for new data. In this study, we use cross-validation (cv) = 5.

The accuracy unit used is explained variance, with calculations:

\[
\text{explained}_\text{variance}(y, \hat{y}) = 1 - \frac{\text{var}[y-\hat{y}]}{\text{var}[y]} \tag{9}
\]

where \( \hat{y} \) is the estimated target output, \( y \) the corresponding (correct) target output, and \( \text{Var} \) is variance, the square of the standard deviation. The best possible score of \( \text{explained}_\text{variance}(y, \hat{y}) \) is 1.0, and the smaller score means weak accuracy.

---

\(^1\)https://www.kaggle.com/c/pneumonia-child-mortality-data624-16a/data
4.6. Feature selection
Feature selection is the feature determination process which is most suitable for representing data. The features or characteristics of the data will be the difference between data with each other. So, if the data is represented by accordingly features, the accuracy will get better.

In this case, the feature selection method is Gram-Schmidt Orthogonalization. By using algorithm 3 in the Method section, the result that obtained is in the form feature indices list and accuracy score to mortality rate prediction.

5. Result and discussion
We compare the accuracy of SVR with and without the Gram-Schmidt Orthogonalization based feature selection. The accuracy result of mortality rate prediction using SVR without feature selection is shown in table 2. The best accuracy of SVR without the Gram-Schmidt Orthogonalization based feature selection happens when it used mean as handle missing data. It can be seen in table 2 below.

Table 2. The accuracy of SVR without feature selection.

| Handle missing data | Accuracy  |
|---------------------|-----------|
| FillNaN             | 0.69957   |
| Mean                | 0.75217   |
| Median              | 0.74233   |
| Most Frequent (Modus)| 0.72645   |

Table 2 shows the accuracy of mortality rates prediction using SVR without feature selection. The best accuracy is when using the mean as handle missing data method. The second best accuracy is median, the third is most frequent (modus) and the worst is fillNaN as handle missing data method.

We also find the accuracy of SVR after Gram-Schmidt Orthogonalization process. Table 3 below gives the selected features in form of index lists by the Gram-Schmidt Orthogonalization-based feature selection for each number of features.

Table 3. The selected features in form of index lists by the Gram-Schmidt Orthogonalization-based feature selection. “…” means the indices are like indices of the previous numbers of features.

| Number of features | FillNaN | Mean | Median | Modus |
|--------------------|---------|------|--------|-------|
| 2                  | 0, 12   | 3, 12| 0, 12  | 0, 12 |
| 3                  | 0, 12, 15| 3, 12, 13| 0, 12, 13| 0, 12, 13 |
| 4                  | 0, 12, 15, 21| 3, 12, 13, 15| 0, 12, 13, 15| 0, 12, 13, 15 |
| 5                  | 0, 12, ..., 21, 20| 3, 12, ..., 15, 14| 0, 12, ..., 15, 14| 0, 12, ..., 15, 14 |
| 6                  | 0, 12, ..., 20, 13| 3, 12, ..., 14, 21| 0, 12, ..., 14, 20| 0, 12, ..., 14, 20 |
| 7                  | 0, 12, ..., 13, 16| 3, 12, ..., 21, 20| 0, 12, ..., 20, 16| 0, 12, ..., 20, 16 |
| 8                  | 0, 12, ..., 16, 14| 3, 12, ..., 20, 16| 0, 12, ..., 16, 22| 0, 12, ..., 16, 22 |
| 9                  | 0, 12, ..., 14, 11| 3, 12, ..., 16, 11| 0, 12, ..., 22, 11| 0, 12, ..., 22, 11 |
| 10                 | 0, 12, ..., 11, 17| 3, 12, ..., 11, 18| 0, 12, ..., 11, 18| 0, 12, ..., 11, 18 |
| 11                 | 0, 12, ..., 17, 18| 3, 12, ..., 18, 2| 0, 12, ..., 18, 8| 0, 12, ..., 18, 17 |
| 12                 | 0, 12, ..., 18, 2| 3, 12, ..., 2, 7| 0, 12, ..., 8, 10| 0, 12, ..., 17, 2 |
| 13                 | 0, 12, ..., 2, 4| 3, 12, ..., 7, 17| 0, 12, ..., 10, 17| 0, 12, ..., 2, 4 |
| 14                 | 0, 12, ..., 4, 19| 3, 12, ..., 17, 10| 0, 12, ..., 17, 5| 0, 12, ..., 4, 19 |
| 15                 | 0, 12, ..., 19, 9| 3, 12, ..., 10, 5| 0, 12, ..., 5, 3| 0, 12, ..., 19, 9 |
| 16                 | 0, 12, ..., 9, 5| 3, 12, ..., 5, 4| 0, 12, ..., 3, 19| 0, 12, ..., 9, 5 |
| 17                 | 0, 12, ..., 5, 7| 3, 12, ..., 4, 19| 0, 12, ..., 19, 9| 0, 12, ..., 5, 7 |
| 18                 | 0, 12, ..., 7, 8| 3, 12, ..., 19, 8| 0, 12, ..., 9, 7| 0, 12, ..., 7, 8 |
| 19                 | 0, 12, ..., 8, 3| 3, 12, ..., 8, 9| 0, 12, ..., 7, 2| 0, 12, ..., 8, 3 |
| 20                 | 0, 12, ..., 3, 22| 3, 12, ..., 9, 0| 0, 12, ..., 2, 21| 0, 12, ..., 3, 21 |
| 21                 | 0, 12, ..., 22, 6| 3, 12, ..., 0, 22| 0, 12, ..., 21, 4| 0, 12, ..., 21, 6 |
Table 3 explains the number of selected features starts from two features until 23 features. It can be seen how the Gram-Schmidt Orthogonalization get the number of features that we want will be influenced by the previous features. As an example to get three selected features, it will be the same as the two previous features that have been obtained and then added by one new feature. This table 4 will represent each feature indices list:

**Table 4. Feature indices list from mortality rate prediction.**

| Feature          | Feature indices list | Feature          | Feature indices list |
|------------------|----------------------|------------------|----------------------|
| total.usd.2013   | 0                    | pct.deaths.5y    | 12                   |
| total.usd.2012   | 1                    | pct.pop.vaccinated| 13                   |
| total.usd.2000   | 2                    | grouping         | -                    |
| total.ppp.2013   | 3                    | in.epi           | 14                   |
| total.ppp.2012   | 4                    | PCV10.price      | 15                   |
| total.ppp.2000   | 5                    | PCV15.price      | 16                   |
| govt.usd.2013    | 6                    | gni.2012         | 17                   |
| govt.usd.2012    | 7                    | gni.2013         | 18                   |
| govt.usd.2000    | 8                    | sanitation.2012  | 19                   |
| govt.ppp.2013    | 9                    | sanitation.2013  | 20                   |
| govt.ppp.2012    | 10                   | birth.rate.2012  | 21                   |
| govt.ppp.2000    | 11                   | birth.rate.2013  | 22                   |

Feature ‘grouping’ is omitted because it contains characters and cannot be processed in feature selection. That is why only 23 features that use to feature selection using Gram-Schmidt Orthogonalization. After we choose the number of feature, we search the accuracy using SVR with each handle missing data as shown in table 5.

**Table 5. The accuracy of SVR with feature selection.**

| Number of features | FillNaN | Mean | Median | Modus   |
|--------------------|---------|------|--------|---------|
| 2                  | 0.52724 | 0.503| 0.52724| 0.52724 |
| 3                  | 0.46269 | 0.44089| 0.46508| 0.43423 |
| 4                  | 0.66791 | 0.37425| 0.38749| 0.4296  |
| 5                  | 0.72164 | 0.37248| 0.35642| 0.40737 |
| 6                  | **0.73258** | 0.72264| 0.63181| 0.63974 |
| 7                  | 0.72839 | 0.74501| 0.63558| 0.6408  |
| 8                  | 0.71391 | 0.74349| 0.74426| 0.71898 |
| 9                  | 0.7143  | 0.74161| 0.74358| 0.71766 |
| 10                 | 0.71206 | 0.74152| 0.74582| 0.71717 |
| 11                 | 0.71119 | 0.73907| 0.74539| 0.71754 |
| 12                 | 0.71118 | 0.74278| 0.74508| 0.71653 |
| 13                 | 0.70882 | 0.74209| 0.74588| 0.71636 |
| 14                 | 0.70096 | 0.74168| 0.74577| 0.7283  |
| 15                 | 0.69729 | 0.74033| 0.74222| **0.72837** |
| 16                 | 0.69789 | 0.73984| 0.74242| 0.7283  |
| 17                 | 0.69687 | 0.74262| 0.74283| 0.72686 |
| 18                 | 0.69635 | 0.74171| 0.74262| 0.72642 |
| 19                 | 0.69431 | 0.74218| 0.74139| 0.72462 |
| 20                 | 0.7006  | 0.74288| **0.74606** | 0.72744 |
| 21                 | 0.69998 | 0.75193| 0.74378| 0.72714 |
| 22                 | 0.69961 | **0.75258** | 0.74165| 0.72642 |
According to table 5, the best accuracy of SVR with the Gram-Schmidt Orthogonalization-based feature selection happens when it used fillNaN as handle missing data. Figure 1 gives this best accuracy in graphical form.

Table 5 represents the accuracy of SVR with the Gram-Schmidt Orthogonalization-based feature selection for mortality rate prediction with each handle missing data method. This indicates the number of selected features affects the accuracy of the method. Figure 1 also shows the accuracy of fillNaN as handle missing data with each feature that used. Based on figure 1, the accuracy decreases from two features until three features. After that, the accuracy increases from four features until six features and monotonically decreasing to fifteen features. Then, the accuracy increases in sixteen features and monotonically decreasing to nineteen features. In twenty features, the accuracy increases and monotonically decreasing to twenty-three features.

Based on previous paragraphs, it can be concluded that the best feature selection using Gram Schmidt combine SVR is when used fillNaN as the handle missing data. From table 5, the maximum accuracy is gained with six features, and the score is 0.73258. This means that the accuracy is increased 4.7186% before the feature selection. While using mean as the handle missing data, the maximum is when the feature is 22 with an accuracy level 0.75258 which means that the accuracy is increased 0.0545% from the original. If we used median as the handle missing data method, it will maximum when the feature is 20 with the accuracy level 0.74606. This shows that the accuracy is the increase of 0.5024%. And finally using the handle missing data most frequent or modus is maximum in feature 15, with the accuracy level 0.72837 which means an accuracy increase of 0.2642%.

Based on each accuracy, feature selection using the Gram-Schmidt orthogonalization increases the accuracy of SVR of mortality rate prediction caused by pneumonia in average 1.384925%. The accuracy is small because the size of data is small. This means the 23 features that are used to predict the mortality rate are indeed important features. If the data used is large data with many features, feature selection using the Gram Schmidt Orthogonalization will be more visible to increase accuracy, because the possibility of some features is not needed to predict mortality rates.

6. Conclusion
Feature selection using Gram-Schmidt Orthogonalization can be used in structured data. It affects to increase the accuracy on SVR. In this case, the accuracy increases in average 1.384925% to predict mortality rates caused by pneumonia.

Acknowledgments
This paper was supported by Universitas Indonesia under PITTA 2018 grant. Any opinions, findings, conclusion, and recommendations are the authors’ and do not necessarily reflect those of the sponsor.
References
[1] Bishop C M 2006 *Pattern Recognition and Machine Learning* (Berkeley: Springer)
[2] Li G-Z, Meng H-H, Yang M Q and Yang J Y 2009 Combining support vector regression with feature selection for multivariate calibration *Neural Computing and Applications* 18 pp 813–820
[3] Khan F M and Zubek V B 2008 Support Vector Regression for censored data (SVRc): a novel tool for survival analysis *Eighth IEEE International Conference on Data Mining* pp 863–868
[4] Guyon I and Elisseeff A 2003 An introduction to variable and feature selection *Journal of Machine Learning Research* 3 pp 1157–82
[5] Cai J, Luo J, Wang S and Yang S 2018 Feature selection in machine learning: a new perspective *Neurocomputing* 300 pp 70–79
[6] Yang J-B and Ong C-J 2010 feature selection for support vector regression using probabilistic prediction *IEEE Transactions on Neural Networks* 22 pp 954–962
[7] Goli S, Manjub H, Faradmal J, Mashayekhi H and Soltanian A-R 2016 Survival prediction and feature selection in patients with breast cancer using Support Vector Regression *Computational and Mathematical Methods in Medicine* 2016 pp 1–12
[8] Huang M-L, Hung Y-H, Lee W M, Li R K and Jiang B-R 2014 SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier *The Scientific World Journal* 2014 pp 1–10
[9] Cinelli M, Sun Y, Best K, Heather J M, Reich-Zeliger S, Shifrut E, Friedman N, Shawe-Taylor J and Chain B 2017 Feature selection using a one dimensional Naive Bayes’ classifier increases the accuracy of Support Vector Machine Classification of CDR3 repertories. *Oxford University Press* 33 pp 951–955
[10] Golub G H and Loan C F V 2013 *Matrix Computation 4th Edition* (USA: Johns Hopkins University Press)
[11] Arora S, Ge R, Halpern Y, Mimno D, Moitra A, Sontag D, Wu Y and Zhu M 2013 A practical algorithm for topic modeling with provable guarantees *Proceedings of International Conference on Machine Learning* 28 pp 280–288
[12] Wang D, Zhang H, Liu R, Liu X and Wang J 2015 Unsupervised feature selection through Gram Schmidt Orthogonalization – a word co-occurrence perspective *Neurocomputing* 143 pp 845–854
[13] Porta M 2001 *A Dictionary of Epidemiology* (Oxford: Oxford University Press)
[14] Liu Q, Chen C, Zhang Y and Hu Z 2011 Feature selection for Support Vector Machine with RBF kernel *Publish Online in Springer Science and Business* 36 pp 99–115
[15] Aytug H 2015 Feature selection for Support Vector Machine using generalized benders decomposition *European Journal of Operational Research* 244 pp 210–218
[16] Hermes L and Buhmann J M 2000 Feature selection for Support Vector Machines. *IEEE Proceedings 15th International Conference on Pattern Recognition* pp 712–715
[17] Basak D, Pal S and Patranabis D C 2007 Support Vector Regression *Neural Information Processing* 11 pp 203–224
[18] Stoppiglia H, Dreyfus G, Dubois R and Oussar Y 2003 Ranking a random feature for variable and feature selection *Journal of Machine Learning Research* 3 pp 1399–1414
[19] Amami R, Ayed D B and Ellouze N 2013 Practical Selection of SVM Supervised Parameters with Different Feature Representations for Vowel Recognition *International Journal of Digital Content Technology and its Applications (JDCTA)* 7 pp 418–424