# PRODUCT QUALITY ANALYSIS USING SUPPORT VECTOR MACHINES

**A. Nachev, B. Stoyanov**

***Abstract**: This paper presents an exploratory study of the effectiveness of support vector machines in the prediction of a product quality based on its characteristics. The study answers the following three questions: how does the choice of kernel and model parameters affect the predictive abilities of support vector machines; can an alternative subset of variables be unearthed that can be used in order to increase the predictive abilities of the data mining model; how will the removal of potential outliers affect the predictive abilities of the data mining model. We used a dataset of red and white wine samples presented by their physiochemical characteristics. Findings show that a correct selection of kernel and appropriate variable selection technique may have a significant impact on the prediction ability of the data mining model. Certain model settings can even make it to outperform the best technique reported thus far in the application area.*

***Keywords**: data mining, support vector machines, sensory preferences, variable selection, wine classification.*

***ACM Classification Keywords**: I.5.2- Computing Methodologies - Pattern Recognition – Design Methodology - Classifier design and evaluation.*

## Introduction

Many studies have used data mining methods to predict quality outcomes of a range of products, based on data available for these products. A variety of domains and applications range from wine quality prediction [Cortez et al., 2009, Beltran et al., 2008, Bapna and Gangopadhyay, 2006, Fei et al., 2008, Li et al., 2010], manufacturing [Xiaoh, 2009, Deh, 2008], water quality prediction [Wang et al., 2010], textiles quality prediction [Selvanayaki et al., 2010], to image quality prediction and image steganography [Hsien-Chu et al., 2008, Narwaria and Lin, 2010]. Many of these studies have used support vector machines (SVM) for data analysis. Researchers report that SVM show superior predictive power to other data mining methods and techniques used in the domains.

This paper presents an exploratory study of the effectiveness of SVM in the prediction of wine quality, based on the physiochemical components thereof. Within the creation and marketing of wine, certification and quality assessment is of great importance, for both health considerations and quality assurance. Quality assessment, in effect is a contributing factor used in determining the price of wine. According to a study conducted by “Wine Business Monthly”, the salaries of wine tasters at vineyards accounts for a quantifiable proportion of expenditure [Tinney, 2006]. Yet human error can be a diminishing factor in the accuracy of this assessment. This opens up avenues for data mining as a good quality control process in the assessment of wine [Cortez et al., 2009]. It is based on these observations that the inherent business value of data mining physiochemical characteristics for predicting product quality becomes evident. Using data mining in the field would allow wine producers to migrate this expensive job function over to a technological platform.

The use of support vector machines in the prediction of wine quality is still in an early stage, yet initial studies within this domain have yielded promising results. Bapna and Gangopadhyay [2006] displayed that SVM exceed both Naive Bayes and Adaptive Bayes in the classification of wine with results based on performance estimation by the classification accuracy metric alone.

Beltran et al. [2008] utilise SVM in addition to, and in comparison with, radial basis function neural networks (RBFNN) and linear discriminant analysis (LDA), in the classification of Chilean wine. The analyses are carried out on data derived from wine aroma chromatograms of three different Chilean wine varieties. Two dimensionality reduction techniques were incorporated, namely principal component analysis (PCA), and wavelet transformation (WT). This work can also be extedned towards using various preformance metrics and different kernel types. Li et al. [2010] proposed use of star-graphs to study behaviour of variables in wine classification. These graphs provided a means of visualization of an instance, taking into account all variables simultaneously. Fei et al., [2008] utilized least squares support vector machines (LS-SVM) on physiochemical data of red wine samples obtained obtained through the use of visible and near infrared (Vis/NIR) transmittance spectroscopy. Cortez et al., [2009] discussed data mining techniques to be used in the prediction of wine taste preferences also. Utilizing a large dataset of Portuguese “vinho verde” samples, three regression techniques were used, namely, SVM, multiple regression (MR) and backpropagation neural networks (BPNN). In utilising the SVM technique, the authors adopted the Gaussian kernel, yet there is little descripton on how kernel type, as a hyperparameter, influences the model preformance. This work can also be extended towards study of different techniques for reduction of dimensionality and selection of optimal subset of variables.

Finally, whilst using the SVM model, many authors state that three issues play significant role in the model preformance: attaining the optimal input subset, correct kernel function, and the optimal parameters of the selected kernel [Fei et al., 2008]. This provides implications to future work, which is addressed in this study.

The structure of this paper is as follows: Section 2 describes SVM as data mining tools; Section 3 describes the dataset used in the study and outlines the variable selection techniques as part of the data pre-processing; Section 4 briefly outlines the role of outliers in data mining; Section 5 describes the experimental results and discusses their meaning.

## Support Vector Machines

Support vector machines have grown in status over the past decade due to the satisfactory results returned over a diverse range of fields. SVM are data analysis techniques categorised within the domain of supervised machine learning [Dash and Singhania, 2009, Salfner et al., 2010], whereby the learning process results in a function being contingent on the supervised training data. Through this supervised machine learning process, the algorithm returns either a classification function, or a regression function. A support vector regression procedure suggests an optimal trade off between complexity and learning ability in order to achieve a strong generalization of accuracy [Xiaoh, 2009].

For a two-class, separable training data set, such as the one in Figure 1, there are lots of possible linear separators. Intuitively, a decision boundary drawn in the middle of the void between data items of the two classes seems better than one which approaches very close to examples of one or both classes. While some learning methods such as the perceptron algorithm find just any linear separator, others, like Naive Bayes, search for the best linear separator according to some criterion. The SVM in particular defines the criterion to be looking for a decision surface that is maximally far away from any data point. This distance from the decision surface to the closest data point determines the margin of the classifier. This method of construction necessarily means that the decision function for an SVM is fully specified by a subset of the data points, which defines the position of the separator. These points are referred to as the support vectors. Figure 2 shows the margin and support vectors for a sample problem. Other data points play no part in determining the decision surface that is chosen.

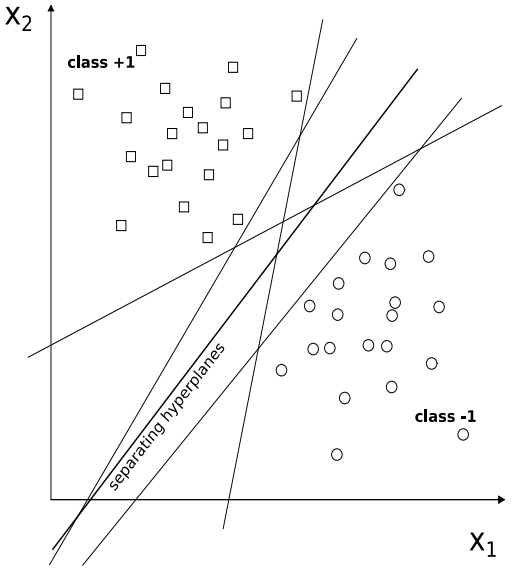


*Figure 1. Separating lines for a two-class separable dataset*

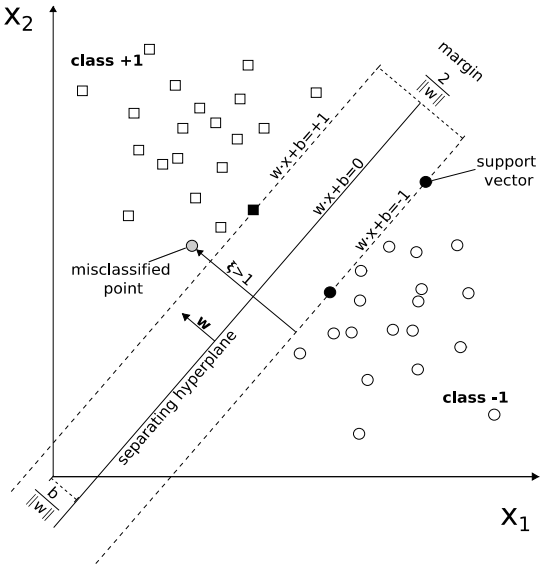


*Figure 2. Geometry of support vector machines.*

SVM can be formalized as follows. Training data of n samples is a set of pairs of data points $\vec{x}_i$ (p-dimensional vectors) and class labels $y_i$ where -1 indicates one class; +1 the other class.

$$D = \{(\vec{x}_i, y_i) \mid \vec{x}_i \in \Re^p, y_i \in \{-1,+1\}\}_{i=1}^n \quad (1)$$

During training a SVM builds a decision boundary that separates the classes. The decision boundary is a p-1 – dimensional hyperplane (a line in the 2D case, a plane in the 3D case, etc.). A decision hyperplane can be defined by a normal vector $\vec{w}$ perpendicular to the hyperplane and a term $b$. The vector $\vec{w}$ is often called weight vector. The term $b$ specifies the choice of hyperplane among all perpendicular to the normal vector. Because the hyperplane is perpendicular to the normal vector, all points x on the hyperplane satisfy

$$\vec{w}^T\vec{x} + b = 0. \quad (2)$$

Data points would fall into one or another side of the decision hyperplane turning the above equality into inequality, therefore the decision function of a linear SVM classifier can be defined as

$$f(\vec{x}) = sign(\vec{w}^T\vec{x} + b). \quad (3)$$

Class labels are +1, -1. The points closest to the separating hyperplane are called support vectors. The margin of a classifier is the maximum width of the band that can be drawn separating the support vectors of the two classes. It can be shown that maximizing the margin is the following minimization problem: find $\vec{w}$ and $b$ such that

$$\frac{1}{2}\vec{w}^T\vec{w} \text{ is minimized and for all } \{(\vec{x}_i, y_i)\} \quad y_i(\vec{w}^T\vec{x}_i + b) \geq 1 \quad (4)$$

This task is optimization of a quadratic function subject to linear constraints. The solution of that problem involves constructing a dual form of the optimization problem where a Lagrange multiplier $\alpha_i$ is associated with each constraint $y_i(\vec{w}^T\vec{x} + b) \geq 1$ in the primal problem. The dual problem is: find $\alpha_i, \ldots, \alpha_N$ such that

$\sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \vec{x}_i^T \vec{x}_j$ is maximized and $\sum_i \alpha_i y_i = 0$; $\alpha_i \geq 0$ for all $1 \leq i \leq N$ (5)

A solution of that problem allows building the decision hyperplane:

$$\vec{w} = \sum_i \alpha_i y_i \vec{x}_i \; ; \; b = y_k - \vec{w}^T \vec{x}_k \text{ for any } \vec{x}_k \text{ such that } \alpha_k \neq 0 \quad (6)$$

Most Lagrange multipliers found by the optimization problem are zero. Each non-zero indicates that it corresponds to a support vector. The classification function (2) can be presented in the form

$$f(\vec{x}) = sign(\sum_i \alpha_i y_i \vec{x}_i^T \vec{x} + b) \quad (7)$$

The above formulas that contain vectors also use dot product operation between them.

The simplest way to divide two classes is with a straight line in 2D, flat plane in 3D or an (N-1)–dimensional hyperplane in an N-dimensional attribute space. Sometimes, however, such a separation is impossible (as shown in Figure 3). Instead of fitting nonlinear curves (hyper-surfaces) to the data, an SVM can handle this using a kernel function that maps the data to a different higher dimensional space where a hyperplane can be used to do the separation. Indeed, if there are two data attributes (2D data points) and data set is not linearly separable by a line, the kernel function can add a third attribute in order to map the points into 3D, so that the data set could be linearly separable by a flat plane in 3D. It can be generalised that the kernel function transforms the data into a higher dimensional space to make separation by hyperplanes possible.

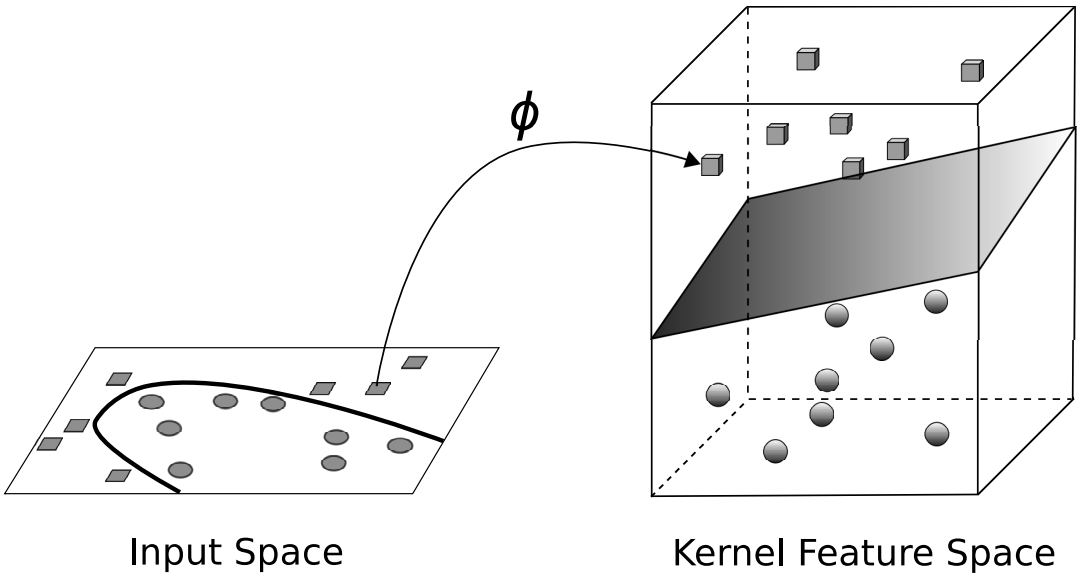


*Figure 3. The kernel trick: a linearly inseparable input space can be mapped to a higher dimensional space, which is linearly separable.*

The kernel function can be defined as

$$K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i)\Phi(\vec{x}_j), \quad (8)$$

where $\Phi(\vec{x})$ maps the vector $\vec{x}$ to some other Euclidean space. The dot product $\vec{x}_i \cdot \vec{x}_j$ in the formulas above is replaced by $K(\vec{x}_i, \vec{x}_j)$ so that the SVM optimization problem in its dual form can be redefined as: maximize (in $\alpha_i$)

$$\widetilde{L}(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j K(\vec{x}_i, \vec{x}_j), \text{ subject to } \sum_i \alpha_i y_i = 0; \quad \alpha_i \geq 0 \text{ for all } 1 \leq i \leq N \quad (9)$$

Various kernel functions can be used with SVM and perhaps their number is infinite. But a few of them have been found to work well for a wide variety of applications. These are:

*Linear:* $K(\vec{x}_i, \vec{x}_j) = \vec{x}_i^T \vec{x}_j$ *(10)*

Polynomial: $K(\vec{x}_i, \vec{x}_j) = (\gamma \vec{x}_i^T \vec{x}_j + r)^d, \gamma > 0$ (11)

Radial Basis Function (RBF) a.k.a. Gaussian kernel: $K(\vec{x}_i, \vec{x}_j) = \exp(\gamma \left\| \vec{x}_i - \vec{x}_j \right\|^2), \gamma > 0$ (12)

Sigmoid: $K(\vec{x}_i, \vec{x}_j) = \tanh(\gamma \vec{x}_i^T \vec{x}_j + r), \gamma > 0, r < 0$ (13)

Ideally, an SVM analysis should produce a hyperplane that completely separates the feature vectors into two non-overlapping groups. However, perfect separation may not be possible, or it may result in a model in so high dimensional space that the model does not generalize well. To allow some flexibility in separating the classes, the soft-margin SVM proposed by Cortes and Vapnik [1995] permit some misclassifications. The method chooses a hyperplane that splits data points as clean as possible while still maximizing the distance to the nearest cleanly split points. The method introduces slack variables $\xi_i$ in $y_i(\vec{w}^T \vec{x}_i + b) \geq 1 - \xi_i$, $1 \leq i \leq n$, which measure the degree of misclassification of the points $\vec{x}_i$. If a training example lies on the 'wrong' side of the hyperplane, the corresponding $\xi_i$ is greater than 1. Therefore, the primal form of the optimization problem is

$$\min_{w,\xi,b} \left\{ \frac{1}{2}\left\| w \right\|^2 + C\sum_{i=1}^{n} \xi_i \right\}, \text{ subject to } \forall_{i=1}^{n}: y_i(\vec{w}^T \vec{x}_i + b) \geq 1 - \xi_i; \forall_{i=1}^{n} \ \xi_i > 0 \quad (14)$$

The factor C in the formula is a parameter that represents the cost of misclassification. A small value of C will increase the number of training errors, while a large C will lead to a behavior similar to that of a hard-margin SVM. In that sense the cost parameter C that controls the trade-off between allowing training errors and forcing rigid margins.

The soft-margin optimization problem along with the constraint can be solved using Lagrange multipliers (as before) so that in a dual form it can be formulated as follows: minimize

$$\widetilde{L}(\alpha) = -\sum_{i=1}^{n} \alpha_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j), \text{ subject to: } \forall_{i=1}^{n}: \sum_{i=1}^{n} y_i \alpha_i = 0; \quad \forall_{i=1}^{n}: 0 \leq \alpha_i \leq C \quad (15)$$

The advantage of the dual form is that the slack variables vanish, with the parameter C appearing only as an additional constraint on the Lagrange multipliers.

The SVM method can also be applied to the case of regression. A version of SVM for regression, called support vector regression (SVR), was proposed by Drucker et al. [1997]. The basic idea of SVR is that a non-linear

function learns by a linear learning method in a kernel-induced higher dimensional space. Similarly to how SVM classification ignores data points that are not support vectors, the SVR depend on a small subset of training data points.

The SVM's major advantage lies with their ability to map variables onto an extremely high feature space. This, in essence facilitates a means for the exploration of nonlinear kernel-based classifiers [Oladunni and Singhal, 2009, Burges, 1998], however, they have been discovered to not favour large datasets, due to the demands it imposes on virtual memory, and the training complexity resultant from the use of such a scaled collection of data [Cortez et al., 2009, Horng et al., 2010].

Work from Fei et al. [2008] highlighted three "crucial problems" in the use of support vector machines. These are attaining the optimal input subset, correct kernel function, and the optimal parameters of the selected kernel, all of which are prime considerations within this study. Multiple authors also echoed sentiments of kernel selection problems [Wang et al., 2010, Selvanayaki et al., 2010, Petrujkic et al., 2008], which further indicated the importance of this factor for this research.

## Dataset and Variable Selection

The data used in this study consists of two distinct sets, which represent the two most common variants of Vinho Verde wines, white and red. With regard to the red sample collection, data instances numbered 1599, while white instances totaled 4898. These instances held 12 variables respectively, relating to the physiochemical breakdown, namely: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and a quality rating. This quality rating was based on a sensory taste test carried out by at least three wine experts grading the wine quality on a scale between "0 (very bad) and 10 (very excellent).

The selection of a subset of variables, that return the best prediction rates in the SVM task, is another important consideration within an exercise in support vector machine prediction. Its importance lies in the need to develop predictive models from a reduced, minimal number of input variables which best summarize the overall input data resulting in maximal predictive power. Dimensionality reduction is the process undertaken in order to reduce the number of independent variables utilised within a data mining exercise. The variables within the dataset are examined to see if and how they relate to, and influence other variables. Fayyad et al., [1996] describe data reduction as the most tedious stage of data analysis. It is at this stage that the cataloguing, classification, segmentation and partitioning of data occur. However, in the aim of effectively discovering knowledge within datasets, reducing the dimensionality of the data is a required step and indeed, a fundamental tool for many data mining tasks. A synergetic benefit of dimensionality reduction is its tendency to reduce overfitting. Datasets possessing a high degree of dimensionality, i.e. a large quantity of variables, will be hindered by the choice of data mining method available to them. Effectively, the method used to reduce the dimensionality of a collection of data will influence the overall accuracy and effectiveness of the data mining exercise. Dimensionality reduction is considered application specific problem, which is not backed by a universal theory. It is as a major challenge in the data mining process as the 'best' variables in one data subset may not necessarily be the best in another; these best variables are, for the most part dependent on the model under employ.

As an indication of its importance within the realm of data analysis, there are many dimensionality reduction techniques, which have been proposed by researchers. These include, but are not limited to, Discrete Fourier Transformation (DFT), Singular Value Decomposition (SVD), Discrete Wavelet Transformation (DWT), Piecewise Aggregate Approximation (PAA), and Adaptive Piecewise Constant Approximation (APCA), etc. LDA and Principal Component Analysis (PCA) are widely popular methods of dimensionality reduction due to its simplicity

and effectiveness in comparison to others. Petrujkic et al., [2008] employ a particle swarm optimization (PSO) based cross-validation method for reducing dimensionality.

There are two distinct groupings of variable selection algorithms, specifically wrapper methods and filter methods. The wrapper methods employ the feature subset selection algorithm in unison with an induction algorithm. The selection algorithm proceeds to unearth a favorable subset of data whilst using this induction algorithm to evaluate proposed subsets. The filter methods use a preprocessing step and autonomously select variables independent of the induction algorithm. There are a number of algorithms that fall under the umbrella of the filter approach, such as the FOCUS algorithm, which inspects all subsets of features in a brute-force fashion in order to unearth a minimal subset of variables that adequately represent the whole; the relief algorithm, which assigns a weighting of relevance to each feature, that is, the relevance of the selected variable to the target output; and the decision tree algorithm, which is used to select feature subsets for the nearest neighbor algorithm [Kohavi and John, 1997].

Rueda et al., [2004] highlight a particular strength possessed by wrapper algorithms. The authors state that if variables are highly correlated with the response, the filter algorithm would typically include them, even if they diminished the overall algorithm performance. While in the wrapper approach, the induction algorithm may discover these diminishing effects, and exclude them.

## Outliers

As previously mentioned, the effect of outliers (a.k.a. noisy data) can have diminishing effects on the accuracy of a data mining and analysis exercise. Many factors serve as the causes of these anomalies including human error/maliciousness, system faults, erroneous measurements or innate deviations [Hodge and Austin, 2004]. The exceptional behaviors of these datapoints go a long way in damaging the accuracy of a given experimentation if overlooked and included incorrectly in the mix. They contribute little or no relevant information to the overall model, and indeed, can be detrimental to the data mining process [Tang et al., 2007]. Detection and removal of detrimental outliers is a key component of this process. The method we use is commonly referred to as the Quartile or Fourth-Spread method [Devore, 2000]. Essentially, we identified the boundaries of each of the quartiles in your data set, measure the fourth-spread (fs), which is the distance between the lower and upper quartiles, and set the upper and lower outlier boundaries as a function of fs. A quartile is any of the three values that divide an ordered data set into four approximately equal parts. Quartiles are a particular type of quantiles, which divide the data into some given number of equal parts.

## Experiments and Discussion

This study requires multiclass classification, as the training set consists of data points belonging to 10 different quality classes. From another hand, the SVM are inherently binary classifiers, which means that their usage is to discriminate between two classes. There are different strategies to make multiclass classification via SVM, such as one-vs.-all (OVA), one-vs.-one (OVO), or using SVM regression with error-tolerance mapping. The OVA technique presumes that binary classifiers are built for each class so that each of them distinguishes between one of the labels and the rest of labels, that is one-versus-all. The OVO technique presumes that a separate classifier is built for each pair of classes. By using a voting technique, the class with most votes is the winner. Due to the complexity of those techniques with regard to the nature of the task solved, the strategy applied in this study was using SVM for regression, which outputs wine quality as a real value. Values were then mapped to integer class labels by the error tolerance technique. The error tolerance $\tau$, a positive real number, defines the interval $[X-\tau, X+\tau]$. A regression output is hit for a class X, if the value belongs to the interval, or miss

otherwise. This approach also preserves the order of preferences. For example, if the true quality class is 5, a model prediction 6 is better than prediction of class 8.

In order to build the SVM model, the dataset was divided into three separate subsets, namely training (50%), validation (25%) and testing (25%).

A number of metrics were used to estimate model performance. These include:

Prediction accuracy at certain error tolerance values were calculated. For the sake of consistency with previous studies [Cortez et al., 2009], the error tolerance thresholds used for experiments were 0.25, 0.5, 1, and 2.

- Mean Absolute Deviation (MAD) represented by (16) is a robust performance measure of the model variability

$$MAD = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \quad (16)$$

where $\hat{y}_i$ is the predicted value.

- Area over the regression error characteristic curve. The regression error characteristic (REC) curve plots the error tolerances along the horizontal axis versus the prediction accuracy on the vertical. The area over the REC curve (AOC) is a scalar value that estimates the overall model performance regardless of the error tolerance values applied to each model instance. The 'ideal' classifier is represented by the most north-west point indicating 100% accuracy and zero error tolerance. Therefore, the model with the least AOC is best performing. REC and AOC applied to SVM regression are analogous to the receiver operating characteristics (ROC) curve and the area above the ROC curve (AOC) metrics used to estimate binary classifiers.

The effectiveness and accuracy of an SVM model is largely dependent on the selection of kernel, the kernel's parameters and value of the cost parameter C. This is an empirical task as there is no theory that can suggest optimal parameter values and also those values strongly depend on features of the training data set and the nature of the task solved. There are a number of parameters that can control the learning and performance of SVM/SVR. The two most relevant are the insensitivity zone $\varepsilon$ and the penalty parameter C, both selected by the user. The former parameter is a positive real number that controls the width of the insensitive zone used to fit the training set, controlling in that way the number of support vectors and model complexity. It also determines the level of accuracy of the approximated classification/regression function and finally the generalization capabilities of the model. An increase in $\varepsilon$ means a reduction in requirements for the accuracy of approximation, but at the same time decreases the number of support vectors, which reduces the model complexity. If $\varepsilon$ is larger than the range of the target values, results are poor. On the other extreme, if $\varepsilon$ is zero, we can expect overfitting. Some studies report that a good empirical rule is that value for $\varepsilon$ is one that leads to percentage of support vectors to be about 50% of the number of dataset samples. The latter parameter C is a penalty factor that can control the tradeoff between the training error and model complexity, which is the number of support vectors. If C is too large, we have high penalty for non-separable data points and many support vectors, which in fact which turns a soft-margin SVM into hard-margin SVM. This leads to overfitting. On the other extreme when C is zero, we have no penalty for misclassifications, few support vectors, and model underfitting. A reasonable proposal for value of C is to be close to the upper bound of the output values, i.e. if the model outputs in [0,B], a value close to B would be a robust choice.

In order to cast the broadest possible catchment area in search of the best performing SVM, four kernels were tested: linear; polynomial with parameter ranges d=[0,5], gamma=[0,5], and r=[0,5]; RBF with gamma=[-5,5]; and

sigmoid with parameters gamma=[-5,5], and r=[-5,5]. Also, the parameter C from (14) was explored in [0,10], and insensitivity zone epsilon=[0,1].

Popular techniques for finding optimal parameter values are grid search and a pattern search. A grid search tries values of each parameter across the specified search range using geometric progression, e.g. $C \in \{2^{-4}, 2^{-2}, 1, 2^{2}, \ldots, 2^{8}, 2^{10}\}$. Similarly, $\gamma$ can take a range of values. The grid search tests the model with each pair of values. Obviously, the method can be computationally expensive in some cases, as it must be evaluated with many parameter values in the grid. The things can even get worse if cross-validation (CV) for each trail is applied. Another search technique called pattern search (also known as a compass search or a line search) can be applied. It starts at the center of the search range and makes trial steps in each direction for each parameter. If the model accuracy improves, the search center moves to the new point and the process is repeated. If no improvement is found, the step size is reduced and the search is tried again. If no step improves the model, the step size is reduced and the process is repeated. The search stops when the search step size is reduced to a specified tolerance. This method requires fewer evaluations but a weakness is that it may find a local rather than global optimal point for the parameters (local minimum problem). A combination of the two techniques is possible, e.g. grid search optimum is further refined by pattern search. Neither technique, however guaranties that the search will end up with a global optimum instead of a local one. For the purposes of this study we applied the pattern search technique. The results of the experiments displayed the polynomial kernel as performing best.

After SVM regression was carried out, k-fold cross-validation (CV) was applied to ensure the integrity of the experiments. This study uses k=5 as many authors, including Cortez et al. [2009], recommend five-fold CV as more robust than other validation techniques. The dataset was split into five subsets, each holding 20% of the instances. Each chunk was used for testing, while the 80% chunk was used in training. This process was then iterated 5 times, with each of the K subsamples used once as the testing data.

Two different attribute evaluation techniques were used, evaluation on either a subset or individual basis. Attribute subset evaluation techniques were classifier subset evaluation; consistency subset evaluation; and wrapper subset evaluation. The single attribute selection techniques used were: chi-squared evaluation, which is based on the chi-squared statistics; gain ratio attribute evaluation; information gain attribute evaluation; principal component analysis evaluation; relief attribute evaluation; and symmetric uncertainty attribute evaluation. A brief description of those techniques can be found in [Hall et al., 2009] and [Witten and Frank, 2005]. After a multitude of attribute evaluation runs and counting the AOC, it was found that the best performing attribute selection technique for red wine is chi-squared evaluation. Table 1 shows the pre-cross-validation top performers.

*Table 1. Pre-CV red wine attribute selection techniques.*

| Attribute Selection | AOC |
|---|---|
| ChiSq+3+4+11+12 | 50.55938 |
| ChiSq+2+3+4+6+8+9+11+12 | 50.73438 |
| ClassifierSubsetEvaluatorRandomSearch+2+3+4+6+11+12 | 50.975 |
| CfsSubsetEvalRandomSearch+3+4+8+9+11+12 | 51.0125 |
| ChiSq+3+4+8+9+11+12 | 51.0125 |
| OriginalPolynomial | 52.375 |

The best post-cross-validation model, however, is the second in Table 1. It suggests using 8 attributes, namely: alcohol, volatile acidity, sulphates, citric acid, total sulfur dioxide, density, chlorides, and fixed acidity. Table 2 and Figure 4 show the worth value, i.e. the percentage of importance of each attribute as proposed by the chi-squared attribute evaluation.

It was also explored how removal of outliers affects the predictive capabilities of the model. For each individual attribute, boundaries were quantitatively set which excluded outliers that resided outside the assigned boundary. This boundary was set using the fourth-spread method. This method entailed identifying the boundaries of the quartiles of each attribute within the wine quality dataset, identifying the range between the upper and lower quartiles. Upper and lower outlier boundaries were set as a function of this fourth-spread. After outliers had been removed it was found that the model improved slightly its performance upon its predecessor, by 0.39%, which is insignificant. This shows empirically that the SVM technique is robust in the studied application area and works well with noisy data.

*Table 2. Chi-square attribute importance, red wine.*

| Attribute | Chi-Squared Worth | Importance % |
|---|---|---|
| alcohol | 497.7464 | 29.61 |
| volatile acidity | 354.4793 | 21.09 |
| sulphates | 252.0535 | 15.00 |
| citric acid | 169.8607 | 10.11 |
| total sulphur dioxide | 145.3958 | 8.65 |
| density | 130.73 | 7.78 |
| chlorides | 82.6207 | 4.92 |
| Fixed acidity | 48.0288 | 2.86 |
| ph | 0 | 0.00 |
| residual sugar | 0 | 0.00 |
| free sulphur dioxide | 0 | 0.00 |

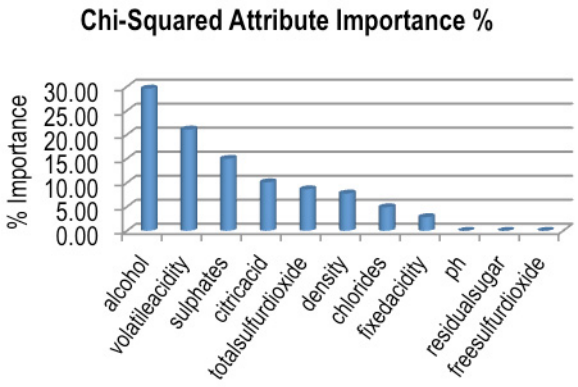


*Figure 4. Chi-square attribute importance, red wine.*

It was found that by using polynomial kernel, the model could be optimized so that in certain conditions it can outperform previously reported models. A combination of the attribute selection described above with SVM parameters C = 1.397998, epsilon = 0.745744, d = 1, gamma = 0.571688, and r = 0.529951 leads to mean squared error (MSE) reduced to 0.4369, which results in a confusion matrix presented by Table 3.

*Table 3. Confusion matrix for SVM red wine prediction model. Bold writing denotes accurate predictions.*

| Actual Class | Red Wine Predictions | | | | |
|---|---|---|---|---|---|
| | **4** | **5** | **6** | **7** | **8** |
| **3** | 0 | 9 | 1 | 0 | 0 |
| **4** | 1 | 35 | 17 | 0 | 0 |
| **5** | 2 | 466 | 210 | 3 | 0 |
| **6** | 0 | 193 | 413 | 32 | 0 |
| **7** | 0 | 11 | 130 | 58 | 0 |
| **8** | 0 | 0 | 13 | 5 | 0 |

Results register that under those conditions and error tolerance $\tau$=1, the model reaches prediction accuracy of 89.5%, outperforming the best model reported by Cortez et al. (2009).

Similar considerations were made regarding the white wine quality prediction task. In summary, the best attribute selection technique found was symmetrical uncertainty ranking [Hall et al., 2009]. It registered the lowest AOC upon cross-validation and suggests 7 attributes, presented and plotted in Table 4 and Figure 5. These are alcohol, density, chlorides, total sulfur dioxide, citric acid, free sulfur dioxide, and volatile acidity.

*Table 4. Symmetrical uncertainty ranking of attributes, white wine.*

| Attribute | Symmetrical Uncertainty Ranking | % Importance |
|---|---|---|
| alcohol | 0.08998 | 26.46626272 |
| density | 0.06524 | 19.18936408 |
| chlorides | 0.04878 | 14.34790282 |
| total sulphur dioxide | 0.03513 | 10.33296076 |
| citric acid | 0.03468 | 10.20060004 |
| free sulphur dioxide | 0.03376 | 9.929995882 |
| volatile acidity | 0.03241 | 9.532913701 |

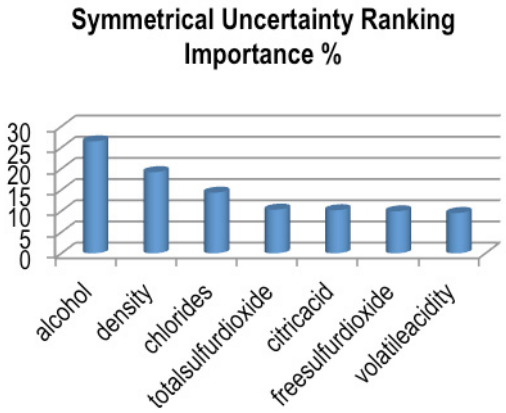


*Figure 5. Symmetrical uncertainty importance of attributes, white wine*

Similarly to the red wine task, part of the research was to combat the detrimental effects of outliers, which resided within the white wine dataset. This process, similar to the method used with the red wine dataset, was conducted through the use of the fourth-spread method. Once these outliers had been eradicated, a re-optimization procedure was conducted in order to attempt to find a better combination of parameters, which would improve the

overall performance of the SVM. This re-optimization failed to improve upon the previously attained MSE of 0.555 unearthed in the initial optimization runs.

It was found that by using polynomial kernel, the white wine model could be optimized so that a combination of the abovementioned attribute selection with certain SVM parameters leads to a confusion matrix presented by Table 5. The experiments clearly showed that the performance of the model built with a reduced attribute selection significantly outperforms its pre-reduced counterpart. Across all error tolerance levels there is a substantial prediction accuracy improvement held by the reduced model. With regards to the overall performance of these models, best depicted by the AOC metric, the reduced model holds a strong 9.61% improvement over its original complete state.

*Table 5. Confusion matrix for SVM white wine prediction model. Bold writing denotes accurate predictions.*

| Actual Class | White Wine Predictions | | | | |
|---|---|---|---|---|---|
| | 4 | 5 | 6 | 7 | 8 |
| 3 | 0 | 2 | 17 | 0 | 0 |
| 4 | 19 | 55 | 88 | 1 | 0 |
| 5 | 7 | 833 | 598 | 19 | 0 |
| 6 | 0 | 235 | 1812 | 144 | 3 |
| 7 | 0 | 18 | 414 | 441 | 7 |
| 8 | 0 | 3 | 71 | 43 | 59 |
| 9 | 0 | 1 | 3 | 2 | 0 |

## Conclusion

The goal of this study was to explore what factors affect the quality of the SVM model in the prediction of wine quality. I was found that the choice of kernel function greatly affects the model predictive abilities. The kernels explored were linear, radial basis function, polynomial, and sigmoid. It was only the polynomial kernel that returned workable results due to its abilities to transform the input space into a much higher dimensional one, thus improving the discriminatory power of the model. It was also found that an appropriate reduction of variables and finding an optimal subset greatly improves the predictive power of the model. An improvement of 9.61 % was found when comparing the pre-reduced model with the post-reduced model in the case of the white wine dataset. One of the primary contributions of this study is improvement of the model performance with regard to error tolerance 1 in the case of red wine dataset. By using variable selection technique based on the chi-squared attribute evaluation, the model outperforms that of Cortez et al. [2009]. Variables eliminated during the model constructions were residual sugar, free sulfur dioxide, and pH. It was also found that removal of outliers, which are anomalous in nature, can improve the overall performance, but marginally, which draws to the conclusion that SVM is a robust data mining technique in this application area and that eliminating outliers influences little the predictive abilities of the model.

## Bibliography

[Agrawal et al., 1993] Agrawal, R., Faloutsos, C. & Swami, A. Efficient similarity search in sequence databases. Foundations of Data Organization and Algorithms, 69-84, 1993.

[Bapna and Gangopadhyay, 2006] Bapna, S. and Gangopadhyay, A. A Wavelet-Based Approach to Preserve Privacy for Classification Mining. Decision Sciences, 37, 623-642, 2006.

[Beltran et al., 2008] Beltran, N. H., Duarte-Mermoud, M. A., Soto Vicencio, V. A., Salah, S. A. & Bustos, M. A. Chilean Wine Classification Using Volatile Organic Compounds Data Obtained With a Fast GC Analyzer. IEEE Transactions on Instrumentation and Measurement, 57, 2421-2436, 2008.

[Burges, 1998] Burges, C. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2, 121-167, 1998.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. Support-vector networks. Machine Learning, 20(3): 273-297, 1995.

[Cortez et al., 2009] Cortez, P., Cerdeira, A., Almeida, F., Matos, T. & Reis, J. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47, 547-553, 2009.

[Dash and Singhania, 2009] Dash, M. & Singhania, A. Mining in Large Noisy Domains. J. Data and Information Quality, 1, 1-30, 2009.

[Deh, 2008] Deh, W. 2008. Surface Hardness Intelligent Prediction in Milling Using Support Vector Regression. Fourth International Conference on Natural Computation, ICNC '08, 2008

[Devore , 2000] Devore, J.L. Probability and Statistics for Engineering and the Sciences. Pacific Grove, CA, 2000.

[Drucker et al., 1997] Drucker, H. Burges, C., Kaufman, L., Smola, A., and Vapnik, V., Support vector regression machines, Advances in Neural Information Processing Systems 9, pages 155-161, Cambridge, MA, MIT Press, 1997.

[Fayyad et al., 1996], Fayyad, U., Haussler, D. & Stolorz, P. Mining scientific data. Commun. ACM, 39, 51-57, 1996.

[Fei et al., 2008] Fei, L., Li, W. & Yong, H. Application of least squares support vector machines for discrimination of red wine using visible and near infrared spectroscopy. Intelligent System and Knowledge Engineering, ISKE' 08, 2008.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten. I., 2009. The WEKA data mining software: an update. SIGKDD Explor. Newsl. 11, 1, 10-18, 2009.

[Hodge and Austin, 2004] Hodge, V. & Austin, J. A survey of outlier detection methodologies. Artificial Intelligence Review, 22, 85-126, 2004.

[Horng et al., 2010] Horng, S., Su, M., Chen, Y., Kao, T., Chen, R., Lai, J. and Perkasa, C. A novel intrusion detection system based on hierarchical clustering and support vector machines. Expert Systems with Applications, 38, 306-313, 2010.

[Hsien-Chu et al., 2008] Hsien-Chu, W., Kuo-Ching, L., Jun-Dong, C. & Ching-Hui, H. An Image Steganographic Scheme Based on Support Vector Regression. Eighth International Conference on Intelligent Systems Design and Applications, 2008. ISDA '08, 2008.

[Kohavi and John, 1997] Kohavi, R. & John, G. H. 1997. Wrappers for feature subset selection. Artificial Intelligence, 97, 273-324, 1997.

[Li et al., 2010] Li, J., Wang, J.-J., Zhang, T., Ma, C.-X. & Hong, W.-X. The Graphical Feature Extraction of Star Plot for Wine Quality Classification. First International Conference on Pervasive Computing Signal Processing and Apps., 2010.

[Narwaria and Lin, 2010] Narwaria, M. & Lin, W. Objective image quality assessment based on support vector regression. IEEE Transactions on Neural Networks, 21, 515-519, 2010.

[Oladunni and Singhal, 2009] Oladunni, O. O. & Singhal, G. 2009. Piecewise multi-classification support vector machines. International Joint Conference on Neural Networks, IJCNN'09, 2009.

[Petrujkic et al., 2008] Petrujkic, M., Rapaic, M. R., Jakovljevic, B. & Dapic, V. Electric energy forecasting in crude oil processing using Support Vector Machines and Particle Swarm Optimization. 9th Symposium on Neural Network Applications in Electrical Engineering, NEUREL, 2008.

[Rueda et al., 2004] Rueda, I. E. A., Arciniegas, F. A. & Embrechts, M. J. SVM sensitivity analysis: an application to currency crises aftermaths. IEEE Transactions on Systems, Man and Cybernetics, 34, 387-398, 2004.

[Salfner et al., 2010] Salfner, F., Lenk, M. & Malek, M. A survey of online failure prediction methods. ACM Comput. Surv., 42, 1-42, 2010.

[Selvanayaki et al., 2010] Selvanayaki, M., Vijaya, M. S., Jamuna, K. S. & Karpagavalli, S. An Interactive Tool for Yarn Strength Prediction Using Support Vector Regression. Conference on Machine Learning and Computing (ICMLC), 2010.

[Tang et al., 2007] Tang, J., Chen, Z., Fu, A. & Cheung, D. Capabilities of outlier detection schemes in large datasets, framework and methodologies. Knowledge and Information Systems, 11, 45-84, 2007.

[Tinney, 2006] Tinney, M.-C. 2006. Wine Business Monthly Salary Survey Report [Online]. Wine Business Monthly. Available: http://www.winebusiness.com/wbm/?go=getArticle&dataId=45483.

[Wang et al., 2010] Wang, X., Lv, J. & Xie, D. A hybrid approach of support vector machine with particle swarm optimization for water quality prediction. 5th International Conference on Computer Science and Education (ICCSE), 2010.

[Witten and Frank, 2005] Witten, I. H. & Frank, E. Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann Pub, 2005.

[Xiaoh, 2009] Xiaoh, W. Intelligent Modeling and Predicting Surface Roughness in End Milling. Fifth International Conference on Natural Computation, ICNC '09, 2009.

## Authors' Information

***Anatoli Nachev*** *– Business Information Systems, Cairnes Business School, National University of Ireland, Galway, Ireland; e-mail: anatoli.nachev@nuigalway.ie*

*Major Fields of Scientific Research: data mining, neural networks, support vector machines, adaptive resonance theory.*

**Borislav Stoyanov** – Department of Computer Science, Shumen University, Shumen, Bulgaria; e-mail: *borislav.stoyanov@shu-bg.net*

*Major Fields of Scientific Research: artificial intelligence, cryptography, data mining.*