University of Alberta

Processability Analysis using Principal Component Analysis and Support Vector Machine

by

Yixin Zhang

A thesis submitted to the Faculty of Graduate Studies and Research

in partial fulfillment of the requirements for the degree of

Master of Science

In

Control Systems

Department of Electrical and Computer Engineering

© Yixin Zhang

Spring 2014

Edmonton, Alberta

Abstract

The method of Support Vectors Machine has become a well-established tool in machine learning. In practice, it has found a wide range of applications from handwritten digits recognition, to face identification, text categorization, bioinformatics and database marketing. Recent research has been conducted on finding key parameters in the oil sands processing that affect the oil sands recovery rate. The application of online visualization and analysis of bitumen exaction allows study of bitumen-air attachment under different operations. As a consequence, we have great confidence that the oil sands processing can be made more effective, efficient, and profitable. In this study, support vector machine models are developed to understand the processability by several case studies.

Due to limited access to bitumen recovery data, for the validation purpose, the methodology is applied to a wine selection process using datasets constructed for classification of wine quality. The obtained model developed outperforms the existing linear and logistic prediction methods in terms of content prediction error. As the proof of concept, the methodology is applied to an oil sands processing dataset created using an artificial model with such variables as bitumen content and fines content of ores, along with the processing variables such as pH and

temperature. The model established using the PCA and SVM methods matches well with the original model used to generate the artificial datasets.

Acknowledgements

I would like to thank my supervisors: Dr. Qing Zhao and Dr. Zhenghe Xu, for their support, guidance and invaluable suggestions. I am grateful for the financial support of National Sciences and Engineering Research Council of Canada (NSERC).

Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

Symbol and Abbreviations	Definition
ANN	Artificial Neural Network
SVM	Support Vector Machine
SVR	Support Vector Regression
RMS	Root Mean Square
PCA	Principal Component Analysis
ICA	Independent Component Analysis
KKT	Karush Kuhn Tucker
RBF	Radial Basis Function
GA	Genetic Algorithm
KPCA	Kernel Principal Component Analysis
PLS	Partial Least Square
MLP	Multilayer Perceptron

Chapter 1

Introduction

1.1 Motivation

Canada is one of the largest sources of oil sands, having extremely tremendous resources compared to other deposit places of the world. The recoverable resource of oil sands in Canada, especially the northeast part of Alberta, can produce millions of barrels of oil per day [1].

Although the recovery rate of oil sands processes is relatively high, the processability of oil sands extraction remains as a challenging research. Laboratory experiments, which are excellent substitutions for traditional large-scale experiments, have conducted, leading to new analytical techniques for oil sands.

Under the circumstances of bench scale techniques, industrial procedures can be closely imitated, with well controlled environment to simulate and record commercial processes, and even recently hydro-transport processes can be mimicked in the laboratory.

However, even with lab-scale units on well-designed experiments, in the complex interaction of oil sands processability including oil sands, chemical additions, operating temperatures, markers that determine the chemical mechanisms of extraction processes are not fully understood. For this purpose, advanced pattern recognition methods, such as support vector machine approach, are investigated to examine oil sands extraction processes. Although SVM has been used in many applications in recent years, its applications in the oil sands industry are relatively new, and the documented results are scarce.

1.2 Research Objectives

Our research is one of the first attempts to explore and investigate the application of machine learning methods, e.g. SVM modeling to the processability analysis of the oil sands extraction process. A case study is presented in this work to identify the key parameters affecting the oil sand recovery rate, and the prediction of the recovery rate is performed based on the established SVM. Due to the limited access to bitumen recovery data, we have used both experiment data and artificial data obtained from empirical models to demonstrate a procedure to build SVM models. In addition, another case study about the wine process analysis and modelling is presented. PCA based classification of wine of different quality is performed based on a sample dataset (real wine data). The objective of our research is to exploit the applications of SVM and PCA to processability analysis of complex engineering processes. Based on the case studies presented in our research, the potential and the effectiveness of these methods are demonstrated.

1.3 Organization of the Thesis

The thesis is organized as follows. Chapter 1 provides a basic introduction and states the objectives. In chapter 2, a review on oil sands industry and statistical learning theory is provided, which introduces background knowledge for the subsequent chapters.

Chapter 3 presents the basic concepts for principal component analysis and support vector machine, and their advantages for data analysis. PCA algorithm is used to classify/identify wine location for wine dataset. Based on the artificial data, a SVM model is built to identify important markers of oil sands. The results are compared with experimental data points for oil sands processability. With respect to wine quality prediction, three multivariate analysis methods are compared for validation purposes.

Chapter 4 concludes the current work and proposes future work, followed by references.

Chapter 2

Background of Oil Sands Industry and Statistical Learning Theory

2.1 Oil Sands Industry Overview

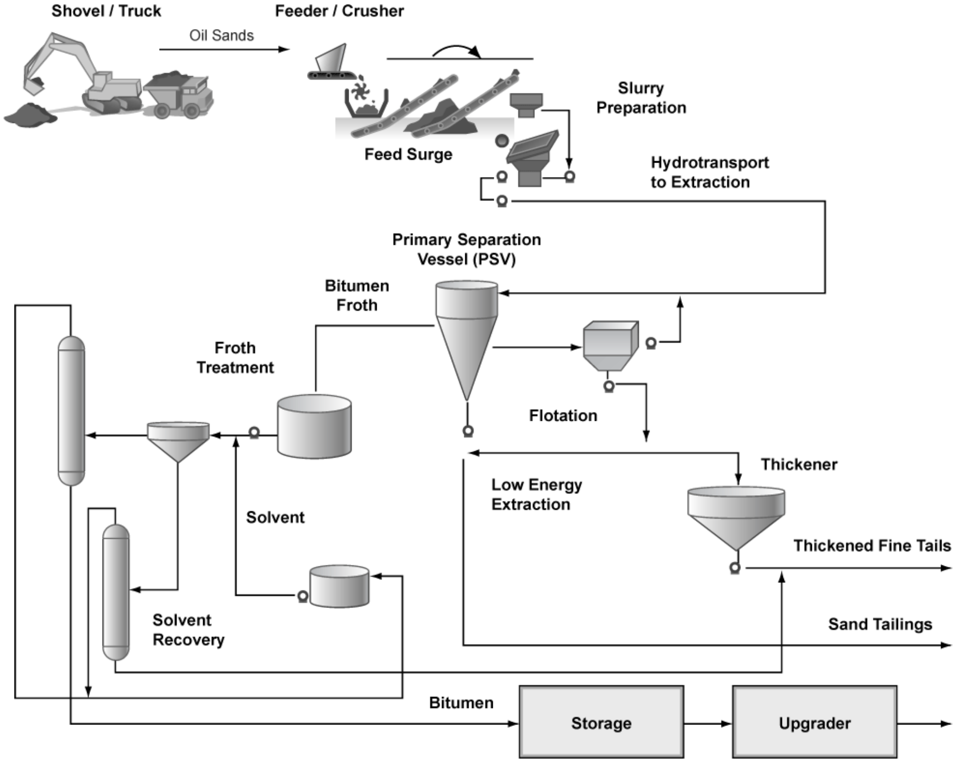


Figure 1 Oil sand processibility overview [1]

Bitumen is a kind of petroleum with high molar mass and in oil sands, and it accounts for 8-15 wt% of mass. At a first glance, the bitumen is very similar to cold molasses, except it is darker. In a way, high viscosity is one of the most critical characteristics of bitumen, which makes it hard to extract and difficult to move. Bitumen can be upgraded to produce various sorts of fuel, such as gasoline, heating oil, etc. Canada has one of the largest sources of bitumen as compared to other deposits around the world. The recoverable resource of Canada, especially the northeast part of Alberta, can serve Canadians and the world for centuries at the present.

2.2 Fundamentals of Oil Sands Processability

The most important reason for using the water-based method in the extraction of bitumen from oil sands is that the sand grains in the oil sands of Alberta are hydrophilic. As a result, a water film surrounding sand grain facilitates them to separate from bitumen. The current commercial hot water extraction process relies on oil sand grains. This water layer is predicted to be about 10nm, but without direct experimental proof.

The bitumen recovery from oil sands requires several steps:

1. Lump size reduction;
2. Bitumen liberation with the help of a combination of several factors;
3. Area of liberated bitumen; and
4. Flotation of aerated bitumen.

Recovery of the bitumen from the oil sands has been improved in efficiency and become more economical. Understanding how to improve the existing process is a challenging research field. Pioneers in the 1920’s laid a foundation that led to our current oil sands technology. Hot water with caustic addition was traced back to their studies in the bitumen liberation and flotation processes. Furthermore, the first successful application based on their research was established in 1967 and it was estimated that oil sands would account for half the amount of oil production with the investment of more than 50 billion by 2010[2]. On the other hand, nowadays, new technologies have been developed in order to contend against the high operating temperature and consumption of water, taking up mostly operating costs, as well as carbon dioxide emission. Several remarkable developments have come into existence to reduce operating costs and intensity of the extraction process. In addition, laboratory scale experiments have led to developments in new analytical techniques for oil sands that are truly great and acceptable substitutions for traditional large-scale experiments. Under the condition of bench scale tests, new techniques allow the industrial procedures to be closely imitated under well controlled environment. Additionally, surfaces and colloid techniques have had tangible impact on improving our understanding of oil sands extraction procedures at the molecular level. Interfacial properties of bitumen can be experimentally observed and accurately measured, and hetero-coagulation phenomena of fine solids with bitumen are detected with the advanced applications of zeta potential distributions measurement on well-designed experimental systems. More importantly, new equipment, such as atomic force microscopy (AFM), allows us to practically measure the cohesion and adhesion forces among bitumen, silica and fines. Consequently, our understanding for oil sands processing has been substantially improved and augmented, helping industries to reap more long-term economic benefits.

The typical unit operations of bitumen production are shown: 1. Mining; 2. Utilities; 3. Extraction; 4. Froth treatment; 5. Water management; and 6.

Upgrading. Each and every unit is related to and affects each other. Maximum recoverable product is the primary objective of bitumen production. So how to control these unit operations in order to operate efficiently and reduce the environmental impact remains research in progress.

Take ore preparation as an example. As shown in Figure 1, firstly trucks mine oil sands from open-pit mine, and then oil sand lumps are crushed and mixed with process water in mixing boxes. In the next stage, the oil sands slurry is transported by hydro transport pipelines to reduce lump size, while bitumen is released in this procedure from sand grains. Chemical additives are often added in this preparation process. Air is introduced into the pipelines to make air bubbles that attach to bitumen for flotation. Then, the aerated bitumen floats from the oil sands slurry in large gravity separation vessels such as PSV or PSC. At last, the bitumen droplets that fail to attach to air bubbles within the slurry are further recovered using flotation cells and tails oil recovery vessels. The temperature of the process is between 40-55 ℃ and bitumen recovery in practice is 88-95%, in the form of froth that contains 60% bitumen. Meanwhile, solvents are added into the bitumen froth to increase the difference in density between bitumen and water and to reduce bitumen viscosity. A separation of organic phase from aqueous is accomplished through inclined plate settlers, cyclones and/or centrifuges. As for the choice, naphtha is being chosen by both Syncrude and Suncor. However, paraffinic solvent is the first option for Albian’s froth treatment plant. No matter what the solvent they choose, both of them can speed up the aggregation of emulsified water and solids. Gravity separation is often preferred over cyclones or centrifuges.

Solid-liquid separation takes place in tailing ponds. The warm water can be recycled from the pond for later use. Suncor and Syncrude use consolidated (composite) tailings (CT) process to accelerate solid-liquid separation. The purpose of CT procedure is to enhance further de-watering and ultimate recovery. Albian uses cyclones to deal with tailings, separating fine tailings as overflow to thickeners and coarse tailings as underflow to tailings pond. Flocculants addition, CT, and the combination of the two are considered by all four oil sand operators of bitumen recovery from oil sands. For oil sands processability, some experiences coming from large-scale commercial operations are summarized and proven to be invaluable. However, deeper understanding for every step in bitumen extraction is still needed.

Recent studies are focused on revealing the key parameters of oil sands components and the interfacial reactions between them. Meanwhile, the observation level develops to the molecular level. The application of online visualization of bitumen exaction allows study of bitumen-air attachment under different operations. As a consequence, we have great confidence that the oil sands processing can be made more effective, efficient, and profitable.

2.3 Literature Review for Statistical Learning Theory

The statistical learning theory provides a framework for studying the problem of gaining knowledge, making predictions, and making decisions from a set of data[3][4].

In statistical learning theory supervised learning is formulated as follows: a set of training data $\{(\mathbf{x}_1,y_1)... (\mathbf{x}_i,y_i)\}$ is given according to unknown probability distribution $P(\mathbf{x},y)$, and $V(y,f(\mathbf{x}))$ which measures the error. The issue is to get a model that minimizes the error expectation:

$$\int V(\mathrm{y},\mathrm{f}(\mathrm{x}))\,\mathrm{P}(\mathrm{x},\mathrm{y})\,\mathrm{dxdy} \qquad (1)$$

In statistical modeling a model from the hypothesis space could be chosen, which is closest to the underlying function in the target space.

Early machine learning algorithms aim to learn representations of simple functions. Hence, the goal of learning is to output a hypothesis that performs the correct classification of the training data. The early learning algorithms are designed to find such an accurate fit to the data. Support vector machine performs better in terms of not over generalization while the neural networks could end up over generalizing easily. Another important thing to consider is to find where the best trade-off in trading complexity with the number of epochs can be made.

Multilayer perceptron (MLP) uses feed forward and recurrent networks[5]. The properties of MLP include universal approximation of continuous nonlinear functions and learning with input-output patterns. MLP also involves advanced network architectures of multiple inputs and outputs.

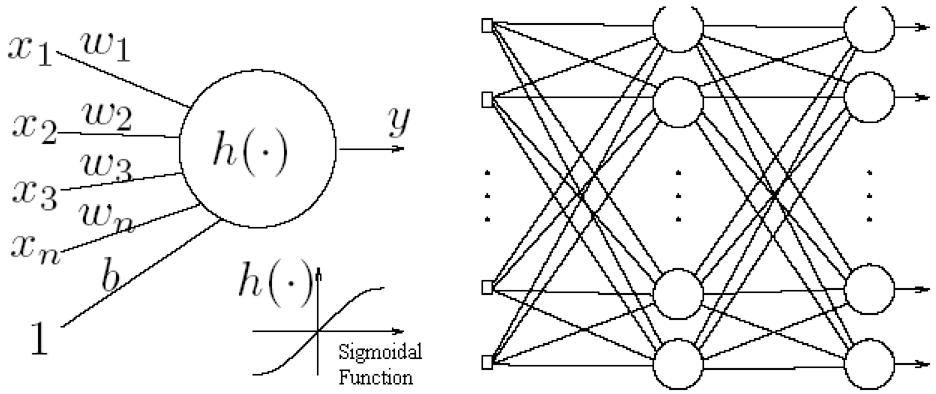


Figure 2 a) Simple neural network b) Multilayer perceptron [5]

There can be some issues in MLP. Finding how many neurons that might be needed for a task is an issue which determines whether the optimality of the neural network is reached. In some cases even if the neural network solutions tend to converge, the results cannot be in a unique solution [6].

2.4 Background for Support Vector Machine

Support vector machine (SVM), based on the structural risk minimization (SRM) principle [7], is a promising method for data mining and knowledge discovery. It stems from the frame work of statistical learning theory of Vapnik-Chervonenkis (VC) theory [8] The SVM is originally developed for pattern recognition. VC theory is the most successful tool for accurately describing the capacity of a learned model by controlling the capacity of a model and ensuring the generalization performance for future samples [9]. VC theory is mainly based on the consistency of a learning process, the rate of convergence of a learning process, the control of the generalization performance of a learning process, and the construction of learning algorithms.

Compared with the traditional learning methods, SVM is firmly rooted in VC theory and hence has its superiority. By minimizing the structural risk, SVM works well not only in classification but also in regression. It has been introduced into many other research fields, e.g. image analysis [10], drug design [11], time series analysis [12-13], quality control of food [14,15], protein structure and function prediction [16–19], genomics [20], and it usually outperforms the traditional statistical learning methods [21,22]. Thus, SVM has been received increasing attention and quickly becomes quite an active research field.

The applications of SVM cover a wide range of chemical problems, e.g. food quality control, chemical reaction monitoring, chemical time series analysis [23], metabolite analysis [24], tea classification [25], QSAR/QSPR studies of retention time [26], toxicity [27], surface tension [28] etc. In general, the SVM models in the above applications, either for classification or for regression, are demonstrated to be superior to other machine learning methods, such as multivariate linear regression (MLR), partial least squares (PLS) [29], artificial neural networks (ANN) [30-31], projection pursuit (PP) methods [32]. However, the parameters of SVM, which have great influence on the predicting ability of SVM, are usually not well refined. For this reason it is expected that better models could be constructed under the condition that a deeper understanding of the theory, performing the parametric tuning in a more effective and global manner.

Here are the literature reviews related to SVM:

A) Machine Reliability Forecasting Applications

The prediction of machine reliability is typically nonlinear. Several traditional and ANN approaches have been studied regarding this type of applications. However, the use of SVM for this particular application has not been widely studied.

Yang and Zhang [33] compared the use of an SVR and LS-SVM with a back propagation neural network (BPNN), an RBF network, and a GRNN for predicting vibration time series signals related to the mechanical condition of machinery. For short term prediction (one step ahead prediction), the SVR using a Gaussian kernel outperformed all of the other methods including the LS-SVM. Compared with the two SVM methods, for long term prediction (24 samples), the RBF network performed better with respect to the NMSE.

Hong et al. [34] discussed the use of SVMG and RSVMG [75] models to predict the "period reliability ratio" for the automotive industry based on time series data containing vehicle damage incidents and the number of damages repaired. For one-step ahead forecasting, the RSVMG model outperformed ARIMA, BPNN, ICBPNN and SVMG (no feedback) methods. The key to this approach was the use of both a genetic algorithm and the use of feedback (recurrent network architecture) to aid in the selection of the free parameters of the SVR.

Hong and Pai [35-36] compared the SVR to three other models (Duane, ARIMA, and GRNN) for the prediction of engine failures. The authors noted that the prediction of engine failure is critical in both the repair and design process of mechanical engines. The dataset used as input was the engine age at the time of unscheduled maintenance actions, and the outputs of the different models were the predicted engine age of the next unscheduled maintenance action per maintenance period. The authors noted that with respect to the NRMSE, the use of SVR exceeds the performance of other models.

B) Control System and Signal Processing Applications

There are several research papers using SVR for time series prediction in the fields of control systems and signal processing. These applications include: mobile position tracking, Internet flow control, adaptive inverse disturbance cancelling, narrow band interference suppression, antenna beam forming, elevator traffic flow prediction, and dynamically tuned gyroscope drift modeling. These diverse applications face the same nonlinear prediction challenges. In addition, some of these applications face an additional challenge of being highly sensitive to computation timing, as expected in real time signal processing applications.

Suykens et al. [37] provided a detailed summary with real world (simplified) examples of nonlinear control system theory using Least Squares Support Vector Machines (LS-SVM). Important discussion topics related to closed loop control theory such as local stability analysis were included. Several real world examples were given, including state space estimation for nonlinear system, inverted pendulum problem, and a ball and beam example.

Gezici et al. [38] proposed the use of SVR to improve the position estimation of users of wireless communications devices. Multi-path, non-line-of-sight propagation, and multiple access interference are the main sources of geo-location error. They proposed the use of a two steps process to estimate the position of the mobile user. First, an SVR (e-insensitive loss function and Gaussian kernel function) is used to predict an initial location. This process is followed by a Kalman-Bucy (K-B) filter to refine the geo-location.

Liu et al. [39] discussed methods to control plant responses and plant disturbances, which were treated as separate process using LS-SVM. The goal was to combine the plant output, which includes the plant disturbance, with the output of the LS-SVM (plant model approximation) to produce an estimate of the disturbance, and feedback of this estimate through an “inverse” LS-SVM to negate the disturbance via the input of the actual plant. For a nonlinear modeled plant and a one-step-ahead prediction horizon, the authors successfully demonstrated the use of both SVR and an adaptive method for determining the free parameters of the SVM. The key aspect of this approach was the use of a Bayesian Evidence Framework for the adaptive selection of LS-SVM free parameters.

Yang and Xie [40] proposed the use of SVR to reduce the effects of high-power narrowband interference (NBI) in spread spectrum systems. Adaptive filters used to solve this problem were time-domain nonlinear LMS adaptive filters (TDAF) and frequency-domain nonlinear LMS adaptive filters (FDAF), both exhibiting sensitivity to noise in estimating NBI. For this specific application, cross-validation methods were too time costly to train the SVR and to determine the SVR free parameters. Using a Gaussian kernel function, the authors noted that using SVR was a viable approach for NBI suppression where computational time was a more critical aspect of this application.

Ramon et al. [41] used an SVR approach to adaptively change antenna beam patterns (beam-forming) in the presence of interfering signals arriving at an arbitrary angle of arrival. This particular application requires the use of complex variables (real and imaginary components of the objective function associated with the signal weighting for the individual antenna elements) for the solution which required separate Lagrange multipliers for the real and imaginary components of the solution. Because this was an adaptive beam forming problem, there was also a computational time constraint. The authors used an alternative optimization method, known as the iterative re-weighted least squares (IWRLS). Using a modified cost function (quadratic for data and linear for "outliers"), the authors demonstrated a significant decrease in bit error rate (BER) as compared to a minimum mean square error-based algorithm.

Luo et al. [42] proposed the use of an LS-SVM for the prediction of elevator traffic flow. ANNs had been used to study this problem and the LS-SVM was used here to improve the ability of control system to predict traffic flow in order to improve elevator service quality. Using three different groups of elevator traffic data, the authors demonstrated the feasibility of the LS-SVM for predicting traffic flow. There was a significant computational trade-off between the computational complexity associated with the training of the LS-SVM and the sparseness of the LS-SVM solution compared to a standard SVR using other non-quadratic loss functions.

Xu et al. [43] compared with SVR using accumulated generated operation (AGO) based on grey theory to an RBF neural network, a grey model, and a standard SVR to predict the drift of a dynamically tuned gyroscope. The AGO algorithm was used to pre-process the drift data in order to reduce noise and complexity of the original dataset. Then, the SVM was trained and an inverse AGO algorithm (IAGO) was applied after the SVM training. A B-spline kernel function was used for this application. As compared to the RBF network, the AGOSVM approach showed superior performance in both the MAE and NMSE by almost an order of magnitude.

C) Atmospheric Distillation

Yafen et al. [44] used least square support vector machine (LS-SVM), RBF neural network and square SVM regression methods to control quality of dry point of aviation kerosene in the atmospheric distillation column. The authors adopted a method based on LS-SVM regression to implement online estimation of aviation kerosene dry point, and compared this method with RBF neural network and SVM regression. They claimed that better abilities of model generalization and real time character by using LS-SVM regression-based soft sensing.

Yan et al. [45] introduced SVM into soft sensor modeling and proposed a SVM-based soft sensing modeling method. A model selection method within the Bayesian evidence framework was proposed to select an optimal model for a SVM-based soft sensor. In their case study, they applied the SVM-based soft sensors estimating the freezing point of light diesel oil in distillation column. They showed that the estimated outputs of SVM soft sensors with the optimal model matched the real values of the freezing point of light diesel oil, which followed the varying trend of the freezing point of light diesel oil very well. Experimental results demonstrated that SVM provided a new and effective method for soft sensing modeling and had promising applications to industrial processes.

Petkovi et al. [46] used support vector machines for predicting electrical energy consumption in the atmospheric distillation of oil refining at a particular oil refinery. During cross-validation process of the SVM training, Authors used particle swarm optimization (PSO) algorithm for selecting free SVM kernel parameters and showed that incorporation of PSO into SVM training process greatly enhanced the quality of prediction.

D) Enterprise Performance and Prediction

Xie et al. [47] proposed a method for predicting crude oil price. The authors used SVM and compared its performance with auto regressive integrated moving average (ARIMA) and back-propagation neural network (BPNN). Authors concluded that SVM outperformed the other two methods.

In order to evaluate and predict the performance of an enterprise such as the oil refinery, Jiekun and Zaixu [48] proposed a model which used data envelopment analysis (DEA) and SVM. The authors used DEA method to first evaluate DEA efficiency of all the oil refining enterprises performance. Then, the input/output data and results of the decision making units (DMUs) were used as the learning examples to train the SVM network and test the network. The results of performance exhibited high accuracy.

E) Flare Analysis using Support Vector Machine

Venkoparao et al.[49] proposed an algorithm using SVM to analyze videos captured from flares. Refineries flared up the exhaust gases prior to releasing them in to the atmosphere. The model proposed intended to reduce environment pollution. Authors argued that the area or volume of the flare and its color could be interpreted as the quantity of released gas during refining process. These parameters of flare were found to indirectly indicate the performance of refining process.

F) Classification of Gasoline

Balabin et al.[50] conducted near infrared (NIR) spectroscopy for gasoline classification and compared the abilities of nine different multivariate classification methods: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), regularized discriminant analysis (RDA), soft independent modeling of class analogy (SIMCA), partial least squares (PLS) classification, K-nearest neighbor (KNN), support vector machines (SVM), probabilistic neural network (PNN) and multilayer perceptron (ANN-MLP). They found that KNN, SVM and PNN techniques for classification were the most effective ones.

Chapter 3

Application of PCA and SVM to Processability Analysis

3.1 The Advantages of PCA and SVM

PCA is purely a descriptive technique. In itself PCA makes no prediction about what future data will look like. That is why we need to use SVM. On the other hand, all classification methods have advantages and disadvantages, which could be relative important depending on the data which are being analyzed. SVM can be a useful tool for data analysis, in particular for the case of non-regularity in the data, for example when the data are not regularly distributed or have an unknown distribution.

The advantages of the SVM can be summarised as follows:

1. By introducing the kernel, SVM gains flexibility in the choice of the form of the threshold separating original data. In fact, it does not need to be linear and even needs not have the same functional form for all the data, since its function is non-parametric and operates locally.

2. Since the kernel implicitly contains a nonlinear transformation, no assumptions about the functional form of the transformation are necessary. The transformation occurs implicitly on a robust theoretical basis and human expertise judgement is not needed.

3. SVM provides a good out-of-sample generalization, if the parameters are appropriately chosen. This means that, by choosing an appropriate generalization grade, SVM can be robust, even when the training samples have some bias.

4. SVM delivers a unique solution, since the optimality problem is convex. This is an advantage compared to neural networks, which have multiple solutions associated with local minima and for this reason may not be robust over different samples.

3.2 Principal Component Analysis Theory

Introduced by Pearson (1901) and Hotelling (1933), Principal Component Analysis has become a popular data-processing and dimension-reduction technique, with numerous applications in engineering, biology, economy and social science.

Consider a dataset consisting of p variables observed on n subjects. Variables are denoted by $(\mathrm{x}_1, \mathrm{x}_2, \ldots \mathrm{x}_p)$. In general, data is in a table with the rows representing the subjects (individuals) and the variables. The dataset can also be viewed as a $n \times p$ rectangular matrix X. The variables are also normalized.

We can represent these data in two graphs: one, in a subject graph where we try to find similarities or differences between subjects, and other in a variable graph where we try to find correlations between variables. Subjects graph belongs to a p-dimensional space, while variables graph belongs to an n-dimensional space. We have two clouds of points in high-dimensional space. The PCA will give us a subspace of reasonable dimension so that the projection onto this subspace retains more information present in the dataset. In other words, the goal of PCA is to compute another dimension that best represents the dataset. The purpose is that this new dimension will filter out the noise and reveal hidden structure.

$$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ . \\ . \\ . \\ x_{pi} \end{pmatrix} \rightarrow \text{reduce} \quad \text{dimensionality} \rightarrow z_i = \begin{pmatrix} z_{1i} \\ z_{2i} \\ . \\ . \\ . \\ z_{pi} \end{pmatrix} with(q < p) \qquad (2)$$

We assume that the data points are projected on a 1-dimensional space. The principal component corresponding to this axis is a linear combination of the original variables and can be expressed as follows:

$$z_1 = \alpha_{11}\,\mathrm{x}_1 + \alpha_{12}\,\mathrm{x}_2 + ... + \alpha_{1p}\,\mathrm{x}_p = Xu_1 \qquad (3)$$

where $u_1 = \left(\alpha_{11}, \alpha_{12}, ...\alpha_{1p}\right)'$ is a column vector of weights. The principal component z_1 is determined by the overall variance of the resulting points. Of course, the variance of z_1 could be made as large as possible by choosing large values for the weights $\alpha_{11}, \alpha_{12}, ...\alpha_{1p}$. To prevent this, weights are calculated with the constraint that their sum of squares is 1, so that u_1 which is a unit vector subjects to the constraint:

$$\alpha_{11}^2 + \alpha_{12}^2 + ... + \alpha_{1p}^2 = \left\|u_1\right\|^2 = 1 \qquad (4)$$

Eq.(4) is also the projection of the n subjects on the first component. PCA finds u_1 to maximize $Var(\mathrm{z}_1)$ which is given by

$$Var(\mathrm{z}_1) = \frac{1}{n}\sum_{i=1}^{n} z_{1i}^2 = \frac{1}{n}\|\mathrm{z}_1\|^2 = \frac{1}{n}u_1^{'}X^{'}Xu_1 \tag{5}$$

The matrix $C = \frac{1}{n}X^{'}X$ is the correlation matrix of the variables. The optimization problem is to maximize $u_1^{'}Cu_1$ with the constrain of $\|u_1\|^2 = 1$, i.e.:

$$\underset{\|u_1\|^2=1}{Max}\ u_1^{'}Cu_1 \tag{6}$$

This condition means that there is a unit vector u_1 so as to maximize the variance of the projection on the first component. The technique for solving such optimization problems (linearly constrained) involves a construction of a Lagrangian function:

$$\Im_1 = u_1^{'}Cu_1 - \lambda_1(u_1^{'}u_1 - 1) \tag{7}$$

Taking the partial derivative $\partial\Im_1 / \partial u_1 = Cu_1 - \lambda_1 u_1$ and solving the equation $\partial\Im_1 / \partial u_1 = 0$ yields:

$$Cu_1 = \lambda_1 u_1 \tag{8}$$

By pre-multiplying each side of this condition by $u_1^{'}$ and using the condition $u_1^{'}u_1 = 1$

$$u_1^{'}Cu_1 = \lambda_1 u_1^{'}u_1 = \lambda_1 \tag{9}$$

It is known that the parameters u_1 and λ_1 that satisfy conditions (8) and (9) are the maximum eigenvalue and the corresponding eigenvector of the correlation matrix C. Thus the optimum coefficients of the original variables generating the first principal component z_1 are the elements of the eigenvector corresponding to the largest eigenvalue of the correlation matrix. These elements are also known as loadings.

The second principal component is calculated in the same way, with the condition that it is uncorrelated (orthogonal) with the first principal component and that it accounts for the largest part of the remaining variance.

$$Z_2 = \alpha_{21}x_1 + \alpha_{22}x_2 + ... + \alpha_{2p}x_p = Xu_2 \qquad (10)$$

where $u_2 = \left(\alpha_{21}, \alpha_{22}, ...\alpha_{2p}\right)'$ is the direction of the component. This axis is constrained to be orthogonal to the first one. Thus, the second component is subject to the constraints:

$$\alpha_{11}^2 + \alpha_{12}^2 + ... + \alpha_{1p}^2 = \|u_1\|^2 = 1, \qquad u_1'u_2 = 0 \qquad (11)$$

The optimization problem is therefore:

$$\underset{\|u_2\|^2=1, u_1'u_2=0}{Max} \quad u_2'Cu_2 \qquad (12)$$

Using the Lagrangian function, the following conditions:

$$Cu_2 = \lambda_2 u_2 \qquad (13)$$

$$u_2'Cu_2 = \lambda_2 \qquad (14)$$

are obtained again. The second vector comes to be the eigenvector corresponding to the second highest eigenvalue of the correlation matrix.

Using induction, it can be proven that PCA is a procedure of eigenvalue decomposition of the correlation matrix. The coefficients generating the linear combinations that transform the original variables into uncorrelated variables are the eigenvectors of the correlation matrix.

The key property of the principal components is that they are all uncorrelated (orthogonal) to one another. Because C is a covariance matrix, it is a positive matrix in the sense that $u^{'}Cu \geq 0$ for any vector u.

This condition implies that the eigenvalues of C are all non-negative.

$$\mathrm{var(z)} = \begin{bmatrix} \lambda_1 & 0 & . & 0 \\ 0 & \lambda_2 & & \\ . & & . & \\ 0 & & & \lambda_p \end{bmatrix} \qquad (15)$$

The first principal component is the direction along which the data have the most variance. The second principal component is the direction orthogonal to the first component with the most variance. It is clear that all components explain together 100% of the variability in the data. Analyzing the original data in the canonical space yields the same results than examining it in the components space. However, PCA allows us to obtain a linear projection of our data, originally in R^p, onto R^q, where q < p. The variance of the projections on to the first q principal components is the sum of the eigenvalues corresponding to these components.

Summarizing the computational steps of PCA:

Step1. Compute mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Step2. Normalize the data: $\Phi_i = \frac{x_i - \overline{x}}{\sigma_x}$

Step3. Form the matrix A= $\left[\Phi_1, \Phi_2, ..., \Phi_p\right]$ ($p \times n$ matrix), then compute: $C = \frac{1}{n}\sum_{i=1}^{n} \Phi_i^{'}\Phi_i$

Step4. Compute the eigenvalues of C: $\lambda_1 > \lambda_2 > ... > \lambda_p$

Step5. Compute the eigenvectors of C: $u_1, u_2, ..., u_p$

Step6. Proceed to the linear transformation $R^p \to R^q$ that performs the dimensionality reduction.

In principal component analysis the number of components extracted is equal to the number of variables being analyzed (under the general condition $n > p$). However, since PCA aims at reducing dimensionality, only the first few components will be important enough to be retained for interpretation and used to present the data. It is therefore reasonable to decide how many principal components are necessary to best describe the data.

Eigenvalues are quantitative assessments of how much a component represents the data. The higher the eigenvalues of a component, the more representative of the data. Eigenvalues are therefore used to determine the meaningfulness of components. The Kaiser method can be used to retain meaningful components. This rule suggests keeping only components with eigenvalues greater than 1. This method is also known as the eigenvalue-one criterion. When a covariance matrix is used, this criterion retains components whose eigenvalue is greater than the average variance of the data (Kaiser-Guttman criterion).

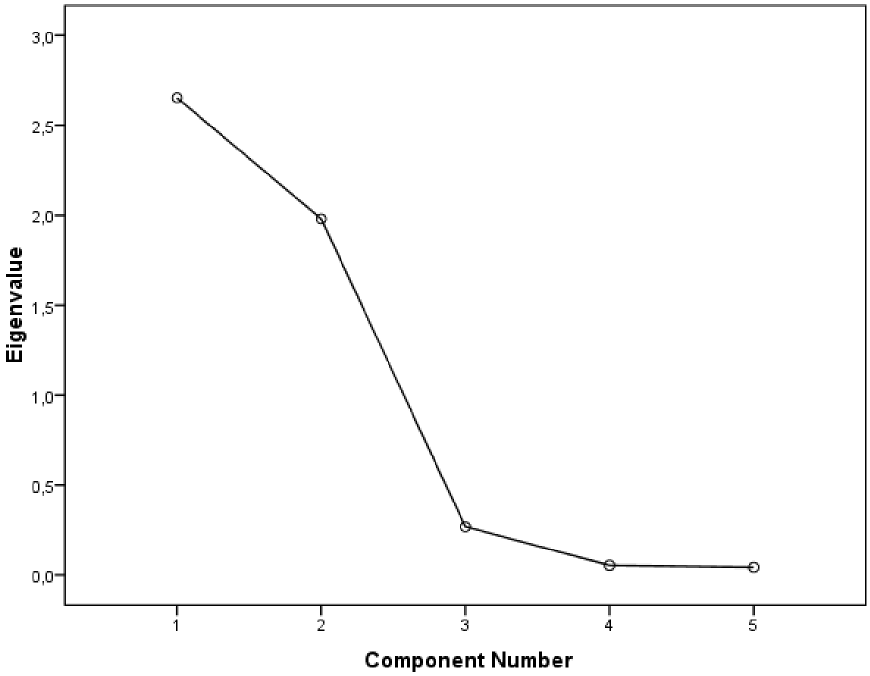


Figure 3 Plot for eigenvalues

How many principal components we should use depends on how big R_q we need. This criterion involves retaining all components up to a total percent variance. It is recommended that the components retained account for at least 60% of the variance. The principal components that offer little increase in the total variance are ignored and those components are considered to be the noise.

3.3 Kernel Principal Component Analysis

The kernel method for PCA provides a bridge between PCA and SVM. Kernel Principal Component Analysis (KPCA) can be represented as an introduction for kernel method which is used in SVM.

Our aim is to find meaningful projections of the data. However, we are facing an unsupervised problem where we do not have access to any labels. Due to this lack of labels, our aim will be to find the subspace of largest variance, where we choose the number of retained dimensions. This is clearly a strong assumption, because there could exist outlier data points in the directions of small variance. In this case PCA is not a suitable technique. However, usually it is true that the directions of smallest variance represent noise.

The sample-covariance matrix C is given by:

$$C = \frac{1}{n}\sum X^{'}X \tag{16}$$

The eigenvalues of this matrix represent the variance in the eigen directions of data space. The eigenvector corresponding to the largest eigenvalue is the direction in which the data is most stretched out. The second direction is orthogonal to it and picks the direction of the largest variance in that orthogonal subspace. Thus, to reduce the dimensionality of the data, the data onto the retained eigen-directions of largest variances are projected:

$$U\Lambda U^{'} = C \Rightarrow C = \sum_{\alpha} \lambda_{\alpha} U_{\alpha} U_{\alpha}^{'} \tag{17}$$

The projection is given by

$y_i = U^{'}_{\kappa} x_i$ for all i (18)

where $U^{'}_{\kappa}$ means the $d \times k$ sub-matrix containing the first k eigenvectors as columns. It shows that the projected data are de-correlated in the new dimension:

$$\frac{1}{N}\sum_i y_i y_i^{'} = \frac{1}{N}\sum_i U^{'}_k x_i x_i^{'} U_k = U^{'}_k C U_k = U^{'}_k U \Lambda U^{'} U^{'}_k = \Lambda_k \qquad (19)$$

where Λ_k is the diagonal $k \times k$ sub-matrix corresponding to the largest eigenvalues. Another convenient property of this procedure is that the reconstruction error in L_2 norm is minimal, i.e.

$$\sum_i \| x_i - P_k x_i \|^2 \qquad (20)$$

where $P_k = U_k U^{'}_k$ is the projection onto the subspace spanned by the columns of minimal U_k.

It shows that the eigenvectors that span the projection space must lie in the subspace spanned by the data-cases. This can be seen as follows:

$$\lambda_a U_a = C u_a = \frac{1}{N}\sum_i x_i x_i^{'} u_a = \frac{1}{N}\sum_i (\mathrm{x}_i^{'}\mathrm{u}_a) x_i \Rightarrow u_a = \sum_i \frac{(\mathrm{x}_i^{'}\mathrm{u}_a)}{N\lambda_a} x_i = \sum_i \alpha_i^a x_i \qquad (21)$$

where u_a is some arbitrary eigenvector of C. From this equation the coefficients

α_i^a can be computed efficiently over a space of dimension N as follows:

$$x_i^{'} C u_a = \lambda_a x_i^{'} u_a \Rightarrow$$

$$x_i^{'} \frac{1}{N} \sum_k x_k x_k^{'} \sum_j \alpha_j^a x_j = \lambda_a x_i^{'} \sum_j \alpha_j^a x_j \Rightarrow$$

$$\frac{1}{N} \sum_{j,k} \alpha_j^a [\mathrm{x}_i^{'} \mathrm{x}_k][\mathrm{x}_k^{'} \mathrm{x}_j] = \lambda_a \sum_j \alpha_j^a [\mathrm{x}_i^{'} \mathrm{x}_j] \qquad (22)$$

Rename the matrix $[\mathrm{x}_i^{'} \mathrm{x}_j] = K_{i,j}$ to arrive at,

$$K^2 \alpha^a = N \lambda_a K \alpha^a \Rightarrow K \alpha^a = (\tilde{\lambda}_a) \alpha^a \quad \text{with} \quad \tilde{\lambda}_a = N \lambda_a \qquad (23)$$

By requiring u being normalized; we obtain

$$u_a^{'} u_a = 1 \Rightarrow \sum_{i,j} \alpha_i^a \alpha_j^a [\mathrm{x}_i^{'} \mathrm{x}_j] = \alpha_a^{'} K \alpha_a = N \lambda_a \alpha_a^{'} \alpha_a = 1 \Rightarrow$$

$$\|\alpha_a\| = \frac{1}{\sqrt{N \lambda_a}} \qquad (24)$$

When a new data case t is received and computed its projections onto the new reduce space, $u_a^{'} t$ becomes:

$$u_a^{'} t = \sum_i \alpha_i^a x_i^{'} t = \sum_i \alpha_i^a K(\mathrm{x}_i, \mathrm{t}) \qquad (25)$$

It is central to most kernel methods.

3.4 Support Vector Machine Theory

Now SVM can be represented through kernel methods which have been introduced above:

Expression for maximum margin is given as:

$$m\arg in \equiv \arg\min_{x\in D} d(\mathrm{x}) = \arg\min_{x\in D} \frac{|\,\mathrm{x}\cdot w + b\,|}{\sqrt{\sum_{i=1}^{d} w_i^2}} \qquad (26)$$

The goals of SVM are to separate the data and to extend this to nonlinear boundaries using kernel methods. For calculations we have:

$$(\mathrm{wx}_i + \mathrm{b}) \geq 1 \qquad y_i = +1$$

$$(\mathrm{wx}_i + \mathrm{b}) \leq 1 \qquad y_i = -1$$

$$y_i(\,\mathrm{w}\,\mathrm{x}_i + \;\mathrm{b} \geq) \qquad \text{for all i} \qquad (27)$$

In the above equation x is a vector point, and w is constant weight and b is also a constant. If the training dataset of SVM is properly chosen, the testing dataset will be located in an acceptable distance from training vector. The selected hyperplane to maximize the margin is located among the closest points of original dataset, which is canonically represented as $wx_i + b = 1$ and $wx_i + b = -1$. The maximum margin is

maximum margin = M = 2 / ||w|| (28)

This is a quadratic optimization problem. The solution involves constructing a dual problem where a Langlier’s multiplier α_i is associated. We need to find w and b in such a way that Φ (w) =½ |w’||w| is minimized for all $\{(\mathrm{x}_i, \mathrm{y}_i)\} : \mathrm{y}_i(\mathrm{w} \cdot \mathrm{x}_i + b) \geq 1$.

We get that $w = \sum a_i \cdot x_i$; $b = y_k - w \cdot x_k$ for any x_k

The classifying function will have the following form:

$$f(\mathrm{x}) = \sum a_i y_i x_i \cdot x + b \tag{29}$$

the QP formulation for SVM classification is presented:

$$\min_{f, \xi_i} \| f \|_K^2 + C \sum_{i=1}^{l} \xi_i \;\; y_i f(\mathrm{x}_i) \geq 1 - \xi_i \;\; \text{for all} \;\; i \;\; \xi_i \geq 0 \tag{30}$$

Dual formulation for SVM classification:

$$\min_{\alpha_i} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathrm{x}_i, \mathrm{x}_j) \qquad 0 \leq \alpha_i \leq C$$

$$\text{for all} \;\; i \;\; \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{31}$$

Variables ξ_i is called slack variable. It measures the error made at point $(\mathrm{x}_i, \mathrm{y}_i)$.

In real world problem, a curved decision boundary is required to separate the data. Moreover, it is better for the smooth boundary under the conditions that original data has noise or outliers. For this situation, the term slack variable S_k is introduced: $\mathrm{y}_i(\mathrm{w} \cdot \mathrm{x} + \mathrm{b}) \geq 1 - \mathrm{S}_k$. This allows a small distance S_k on the wrong side of the hyperplane without violating the constraint.

$$\min \mathrm{L} = \frac{1}{2} w \cdot w - \sum \lambda_k (\mathrm{y}_k (w \cdot x_k + \mathrm{b}) + \mathrm{S}_k - 1) + \alpha \sum \mathrm{S}_k \tag{32}$$

Reducing α allows more data to lie on the wrong side of hyperplane, which would be treated as outliers and give smoother decision boundary.

If the data is linear, a separating hyperplane may be used to divide the data. However, it is often the case that the data is far from linear and the datasets are inseparable (Figure 4). Kernel methods are used to non-linearly map the input data to a high dimensional space. The new mapping is then linearly separable.

This mapping is defined by the kernel

$$K(\mathrm{x}, \mathrm{y}) = \Phi(\mathrm{x}) \cdot \Phi(\mathrm{y}) \tag{33}$$

Transforming the data into feature space makes it possible to define a similarity measure on the basis of the dot product.

$$\langle \mathrm{x}_1 \cdot \mathrm{x}_2 \rangle \leftarrow K(\mathrm{x}_1, \mathrm{x}_2) = \langle \Phi(\mathrm{x}_1) \cdot \Phi(\mathrm{x}_2) \rangle \tag{34}$$

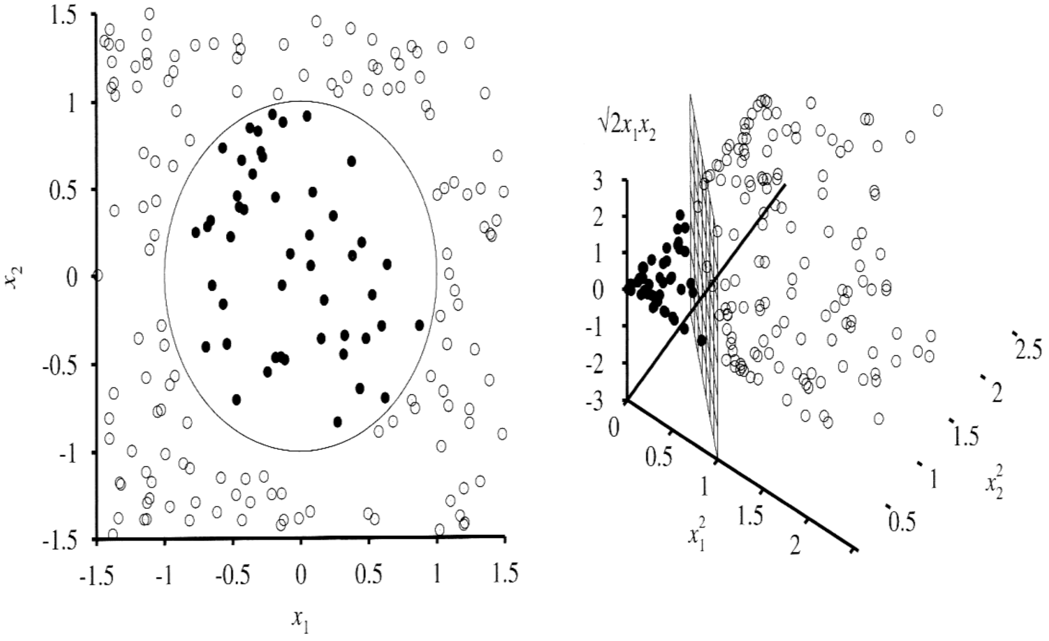


Figure 4: Feature space representation [51]

The kernel function is very critical to SVM. The purpose of kernel function is to perform operations in the input space. Thus the inner product can be operated in the feature space rather than high dimensional space.

$$K(\mathrm{x},\mathrm{x}') = \left\langle \phi(\mathrm{x}), \phi(\mathrm{x}') \right\rangle \tag{35}$$

K is a symmetric positive definite function.

$$\iint K(\mathrm{x},\mathrm{x}')\,\mathrm{g}(\mathrm{x})\,\mathrm{g}(\mathrm{x}')\,\mathrm{dxdx}' > 0 \tag{36}$$

$$K(\mathrm{x},\mathrm{x}') = \sum_{m}^{\infty} \alpha_m \phi_m(\mathrm{x}) \phi_m(\mathrm{x}') \qquad \alpha_m \geq 0 \tag{37}$$

then the kernel represents a legitimate inner product in feature space. In this case, the training set is not linearly separable in an input space but in the feature space.

The different kernel functions are listed below:

Polynomial: a polynomial mapping is a method for nonlinear modeling.

$$K(\mathrm{x},\mathrm{x}^{'})=\left\langle x,x^{'}\right\rangle^{d} \qquad (38)$$

Gaussian Radial Basis Function: Radial basis function is expressed most commonly with a Gaussian form

$$K(\mathrm{x},\mathrm{x}^{'})=\exp(-\frac{\left\|x-x^{'}\right\|^{2}}{2\sigma^{2}}) \qquad (39)$$

Exponential Radial Basis Function: A radial basis function produces a piecewise linear solution which can be effective when discontinuities are acceptable.

$$K(\mathrm{x},\mathrm{x}^{'})=\exp(-\frac{\left\|x-x^{'}\right\|}{2\sigma^{2}}) \qquad (40)$$

Multi-Layer Perceptron: The long established MLP with a single hidden layer also has a valid kernel representation.

$$K(\mathrm{x},\mathrm{x}^{'})=\tanh(\rho\left\langle x,x^{'}\right\rangle+\ell) \qquad (41)$$

Summarizing the computational steps of SVM:

Step1. Define the training, validation and test sets in a clear way.

Step2. Perform multiple random data splits into training, validation and test sets and repeat the analysis for each partition.

Step3. Consider data standardization (i.e. mean subtraction and variance scaling) or scale the features within the same range.

Step4. Create binary classifiers for each pair of classes for exploratory purposes and multiple combination schemes for multi-class classification.

Step5. Include a simple and filter method as a baseline for comparison with more sophisticated feature selection procedures.

Step6. Realize that wrapper and embedded methods select features that are relative to the model and for predictive purposes.

Step7. Use a global optimization method or define a coarse grid of values for the parameters on a logarithmic scale and evaluate their performance in a cross-validation setting.

The advantages of PCA and SVM can be represented on case studies given below.

3.5 PCA on Wine Dataset

Wine is a beverage that is complementary to food consumption. Overtime, wine made by a certain winery or from a certain region can develop a greater reputation, and this wine can be more desired and thus more costly than those from other sources.

Quality as a concept within the discipline of marketing is a vast subject, characterized by its complexity. A more relevant aspect of quality is how quality is evaluated and how it is conceptualized. Quality of wine is a characteristic involving the combination of different components of the wine to give a sensory experience. All of these components have a strong influence on the quality and character of wine, and are critical for the characterization and differentiation of wines.

The wine input attributes of measured data are listed in Table 3.1.

Attributes	Red Wines			White Wines		
	Min	Max	Average	Min	Max	Average
Fixed acidity	4.6	15.9	8.3	3.8	14.2	6.9
Volatile acidity	0.1	1.6	0.5	0.1	1.1	0.3
Citric acid	0.0	1.0	0.3	0.0	1.7	0.3
Residual sugar	0.9	15.5	2.5	0.6	65.8	6.4
Chlorides	0.01	0.61	0.08	0.01	0.35	0.05
Free sulfur oxide	1	72	14	2	289	35
Total sulfur oxide	6	289	46	9	440	138
Density	0.99	1.004	0.996	0.987	1.039	0.994
pH	2.7	4.0	3.3	2.7	3.9	3.1
Sulphates	0.3	2.0	0.7	0.2	1.1	0.5
Alcohol	8.4	14.9	10.4	8.0	14.2	10.4

Table 3.1 Input attributes on wine data

The first data subset describes the red wine with 1599 samples. The second data subset describes the white wines with 4898 samples. Each sample is measured by eleven physic and chemical parameters (inputs of model) and recorded by its quality

However it assumes that the different outcomes are classified nominally and they are mutually exclusive. It is possible to model multinomial data in an ordinal way, but if the wrong method is used this may introduce bias or loss of efficiency and information.

Residual tests are carried out to determine whether the normality, constant variance and independence assumptions are satisfied. Scatterplots of the original datasets are drawn (Figure 5) to determine if there exists a relationship between the various variables involved.

For the identification of dependencies among the attributes of processed data, correlation matrices are created for wine data that describe the interdependence of individual attributes. Mutual correlation dependencies are shown in Table 3.2, which shows that the correlation of certain attributes reaches values outside the interval (-0.5, 0.5), indicating considerable interdependence.

Attributes	Fixed acidity	Volatile acidity	Citric acid	Residual sugar
Fixed acidity	1	-0.023	0.289	0.089
Volatile acidity	-0.023	1	-0.149	0.064
Citric acid	0.289	-0.146	1	0.094
Residual sugar	0.089	0.064	0.094	1

Table 3.2 Correlation dependence on the first four wine attributes

The original dataset is not separable (Figure 5). The relationship among each input variables is not clear. On the other hand, we can easily see the relationship among those input variables in Figure 6. The loadings of the input variables show the original axes and the new (rotated) axes derived from covariance.

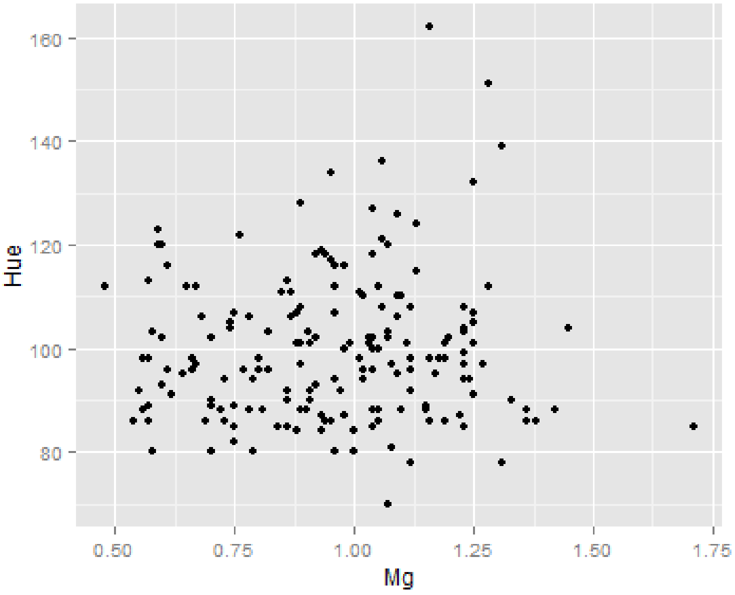


Figure 5 Original wine dataset

The directions of all input variables that use to separate data are shown in Figure 6. The loadings of the input variables show the original axes and the new (rotated) axes derived from covariance. The first component separates wine from Barolo and wine from Barbera, while the second component separates wine from Grignolino. The examination of the values of the contributions complements and refines this interpretation because the contributions suggest that the first component highly contributes to Barbera wine, while the second component contributes most to Barolo wine.

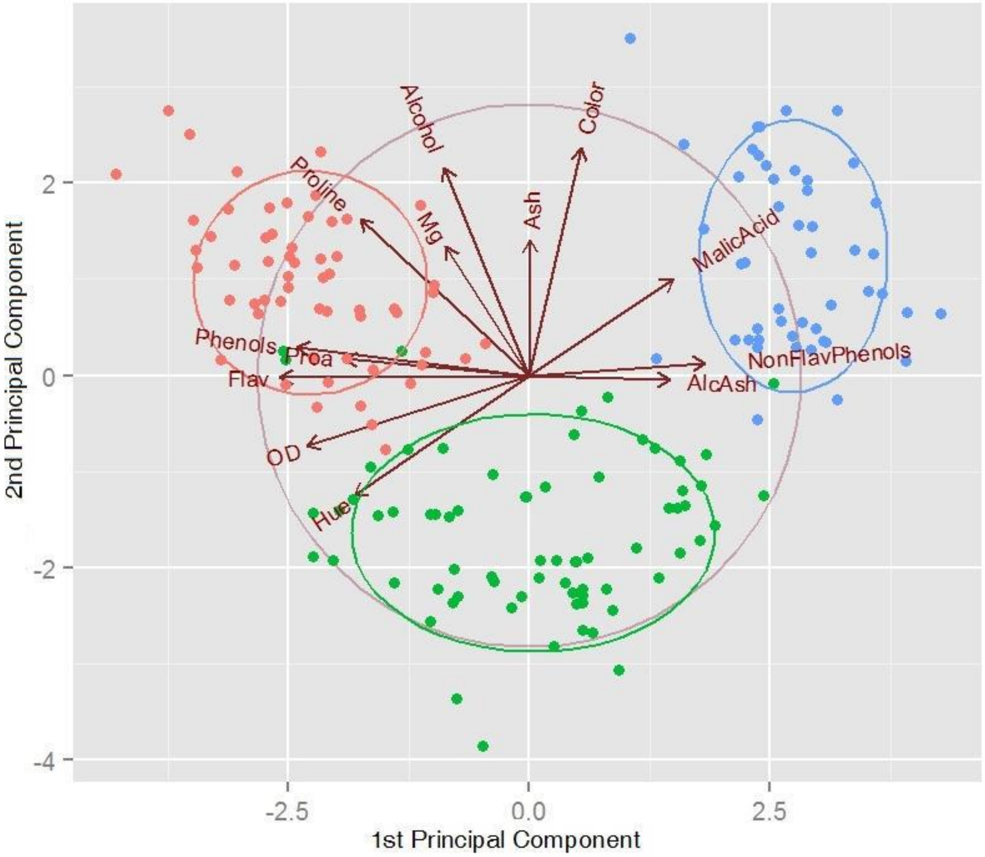


Figure 6 PCA for wine dataset (circles represent maximum possibility)

For comparison, we use KPCA method to apply to wine dataset in order to make a comparison to evaluate superiority of kernel method (Figure 7). In Figure 7, the three input variables are fully separable although there have some overlaps between the second and third original input data.

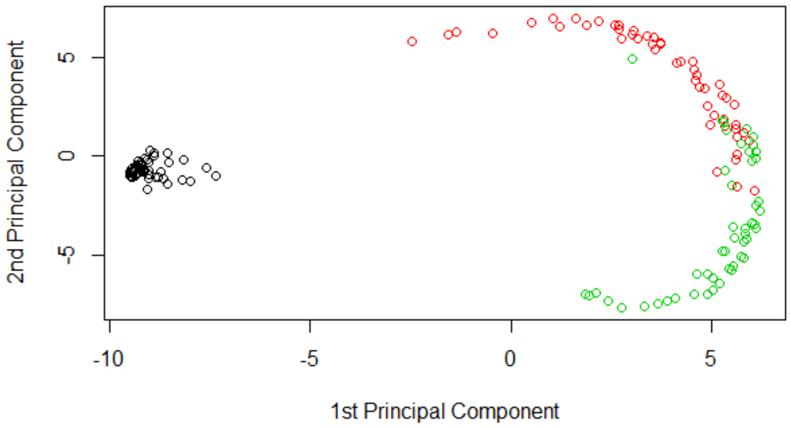


Figure 7 KPCA for wine dataset

3.6 SVM on Oil Sands Dataset

Hot water based bitumen production process from mineable oil sands is extremely complex in nature and highly sensitive to variability of oil sands ores. Understanding ore processability and developing a sensible marker for ore processability have been proven to be a very challenging task. In addition to processing variables such as temperature, hydrodynamics, process water chemistry and chemical additives, ore characteristics, such as bitumen content, connate water content and chemistry, fines content and more importantly types of fines play a decisive role in determining the processability of oil sands ores. Due to limited access to bitumen recovery data, SVM is applied to an oil sands processing dataset created using an artificial model with such variables as bitumen content and fines content of ores, along with the processing variables such as pH and temperature.

The function is generated as follows:

$$R = \frac{B^{\frac{1}{2}}}{F^{\frac{1}{3}}} \times W \times pH^{3} \times [Mg^{2+} + Ca^{2+}]^{-\frac{1}{2}} \tag{42}$$

R: bitumen recovery rate (%)

B: bitumen content (%)

F: fines content (%)

W: water content (%)

Mg^{2+}: Magnesium (ppm)

Ca^{2+}: Calcium (ppm)

The steps to use SVM on oil sands dataset are as follows:

Step1. Generate data from above function.

Step2. Divide data into 80% training dataset and 20% testing dataset.

Step3. Process to make parse parameters classification.

Step4. Choose a selected kernel and load the kernel model.

Step5. Calculate the score of each iteration using the average accuracy of prediction by SVM using 10 folds and repeat 50 times.

Step6. Return the predicted label and remove the 10% lower scores.

Step7. Evaluate classification accuracy on test dataset.

Step8. Accumulate statistics about a single training curve.

Three kernel methods are used to build SVM models, and the comparisons are shown below.

3.6.1 Linear Kernel Results

A linear kernel is tested for cost ranging from 0.0001 to 2. As is shown below, this model produces reasonable accuracy results. Only 20 support vectors are required to describe this data with an accuracy of 1.0 while a cost value of 1.0 is used. Cost values as small as 0.01 also result an accuracy of 1.0:

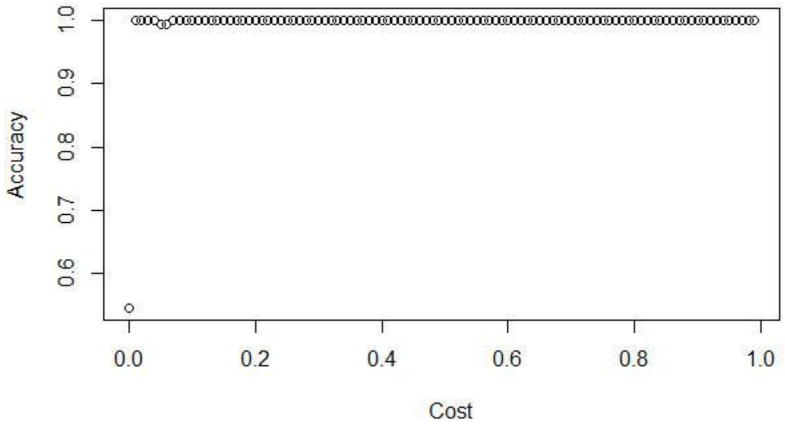


Figure 8 Linear Kernel cost and accuracy

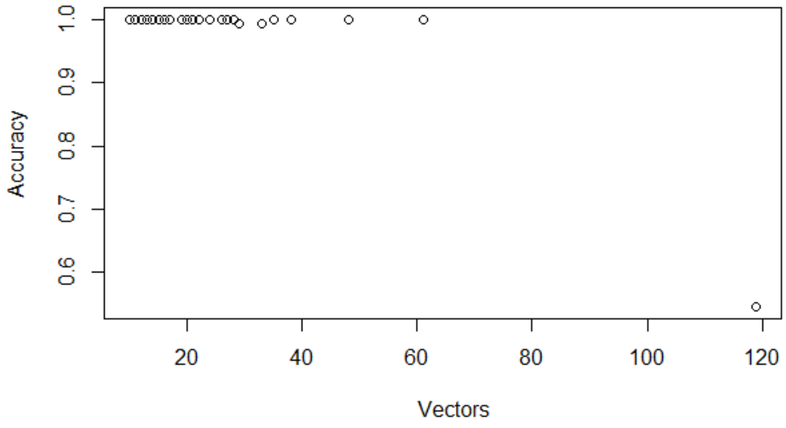


Figure 9 Linear Kernel vectors generation and accuracy

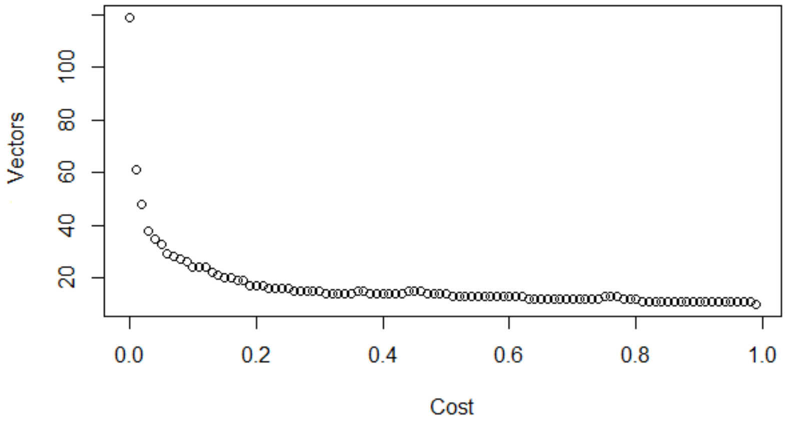


Figure 10 Linear Kernel cost and vectors

3.6.2 Polynomial Kernel Results

A polynomial kernel is tested for degrees ranging from 1 to 5 and for cost ranging from 0.0001 to 2. As expected, Figure 11 shows that the results of polynomial model (degree=1, gamma=1) outperform the linear result in Figure 8.

The SVM model allows for the selection of a gamma parameter. The results for degree of 11 and gamma from 0.01 to 1.0 are shown in Figure 11 through Figure 13. As gamma decreases, the accuracy and cost-curve show degradation at low cost and the number of vectors and cost curve show an increasing trend which implies that the number of vectors needed increases with decreasing gamma. Figure 13 presents the data for gamma of 0.01, which shows a continuation of this trend.

While the accuracy is 1 for a cost of 1, the number of support vectors increases by a factor of approximately 6 (10 to 60). The polynomial model produces excellent accuracy results. With a cost of 1 and gamma of 1, only 10 support vectors are required to describe this data with an accuracy of 1.0. However, as the degree of the polynomial increases, the number of support vectors also increases. For even degrees, the number of support vectors needed is larger than for odd degrees. This suggests an anti-symmetric behavior that odd degree polynomials can work with a linear fit.

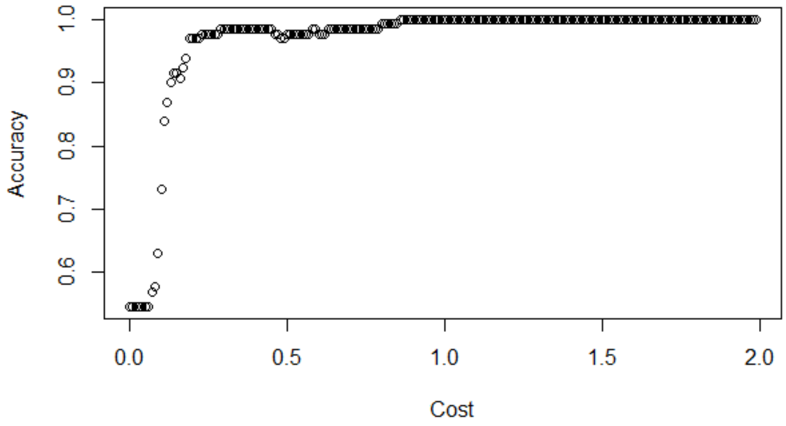


Figure 11 Polynomial Kernel cost and accuracy

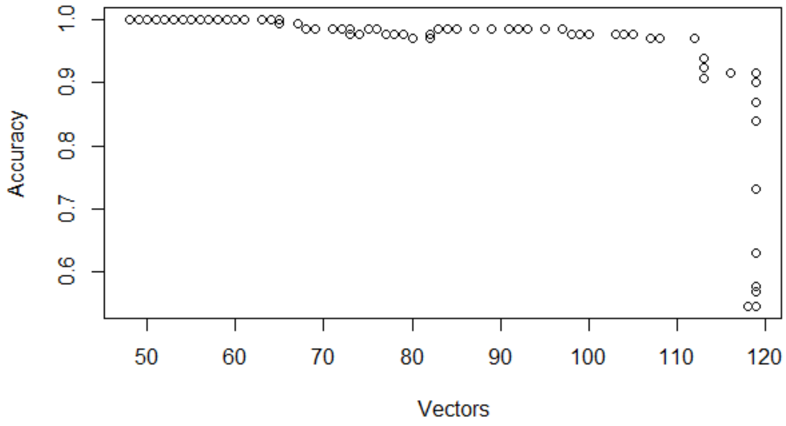


Figure 12 Polynomial Kernel vectors generation and accuracy

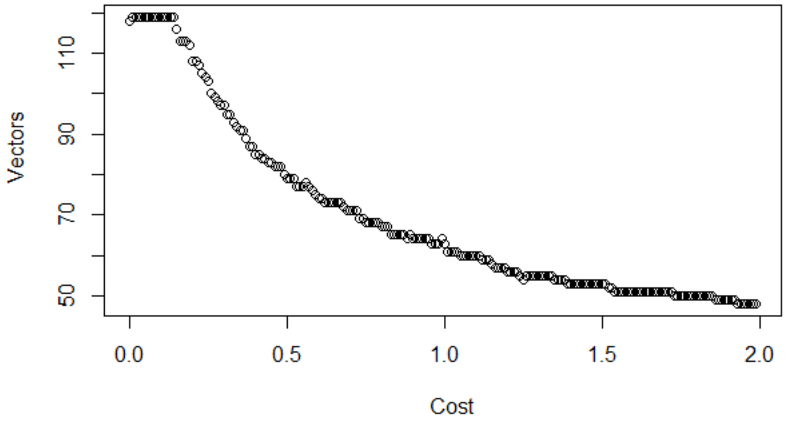


Figure 13 Polynomial Kernel cost and vectors

3.6.3 Gaussian Kernel Results

A Gaussian kernel is tested for cost ranging from 0.0001 to 2 and for gamma from 0.001 to 1.0. As is shown below, this model produces excellent accuracy results but the number of support vectors required is larger. For a cost of 1, the number of support vectors that are required to describe this data ranges from 50 to 130 for gamma ranging from 0.01 to 1.

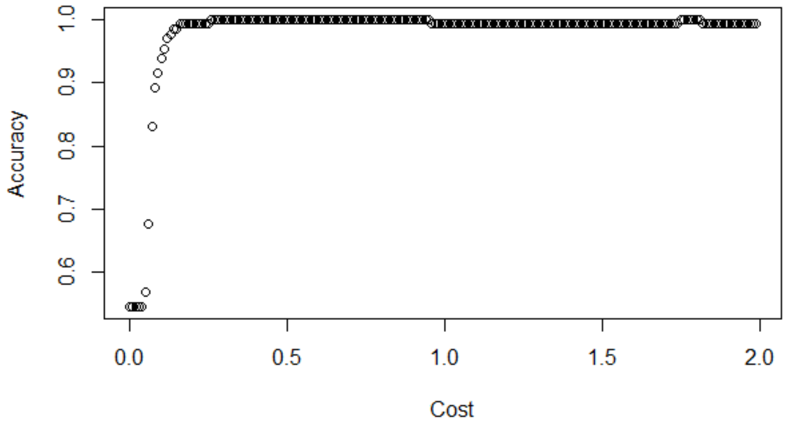


Figure 14 Gaussian Kernel cost and accuracy

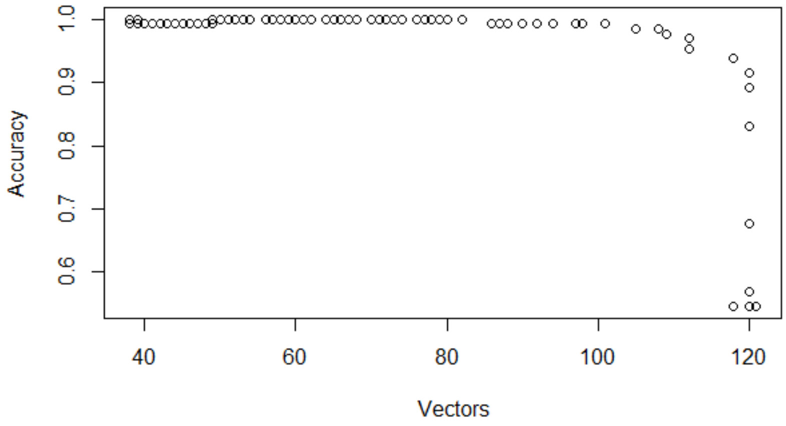


Figure 15 Gaussian Kernel vectors generation and accuracy

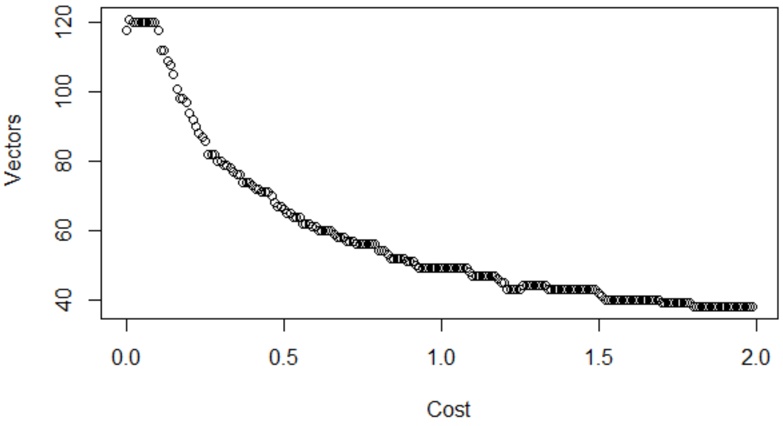


Figure 16 Gaussian Kernel cost and vectors

3.6.4 Discussion of Full Training Set Results

As a result, the best kernel for the oil sands dataset is the polynomial kernel. This kernel provides accuracy of 1.0 with only 29 support vectors. Only 10 vectors are required if the cost is significantly increasing. The success of the polynomial kernel, however, is not obviously apparent from the raw data since all of the independent variables have significant overlap. The SVM model (median $\gamma = 2-3$, C= 3, total execution time 148s) obtains the best predictive results. Figure 17 shows the error rate for the oil sands artificial dataset. SVM provides a valid out-of-sample generalization under the conditions that the parameters are appropriately chosen.

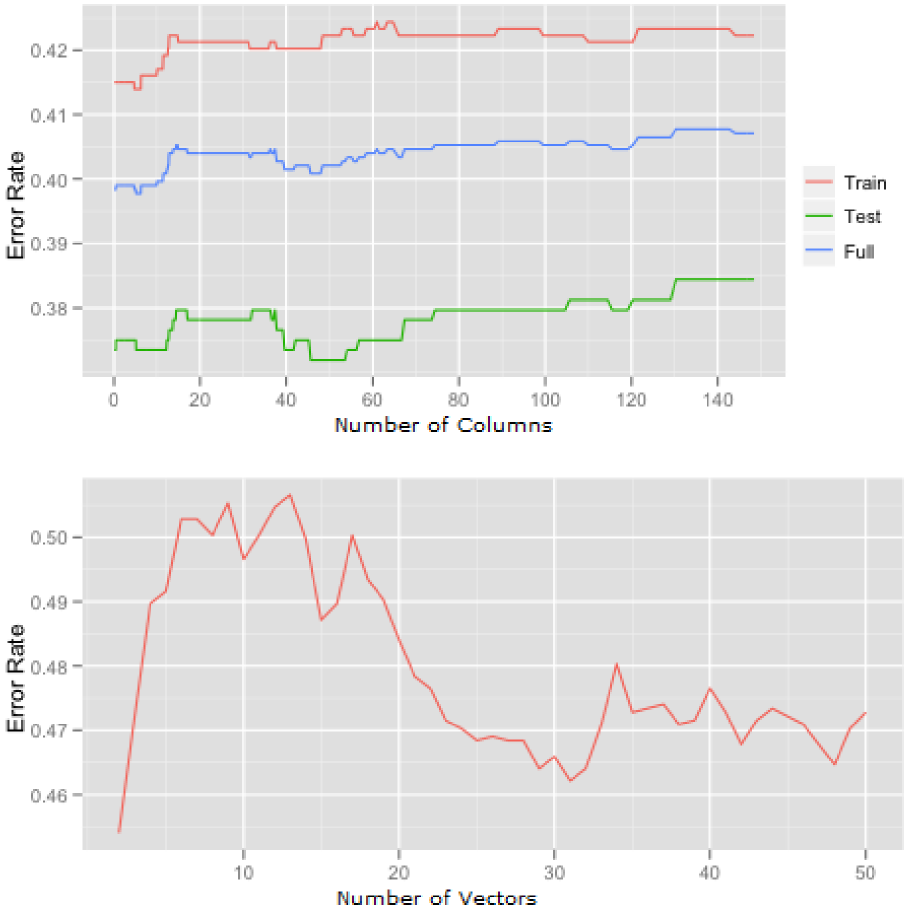


Figure 17 Error rate for oil sands dataset

Figure 18 plots and confirms the SVM performance superiority. In this case, the relative input importance of the SVM model (ordered by importance) shows that the fines content weights as the most relevant input variable and the importance of the rest input variables is shown in decreasing order.

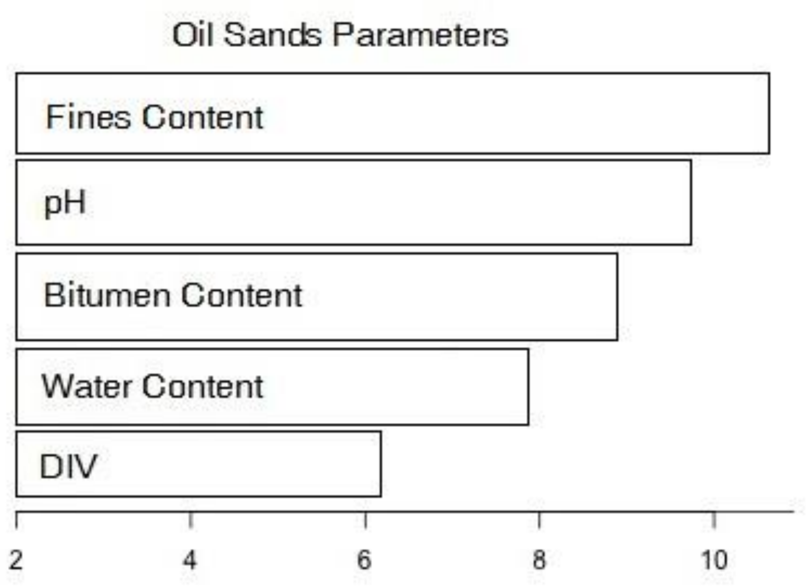


Figure 18 Oil sands variables priority for processibility

The relationship between pH and recovery rate is shown in Figure 19. The model generated data curves show a positive increasing rate, which matches laboratory data. The average DIV with recovery curve is plotted in Figure 20, showing a positive effect followed by a negative effect, where an increasing trending of the weight leads to a higher recovery to a threshold. The same training result to make a comparison is Figure 21, which shows the relationship between fines content with output recovery rate.

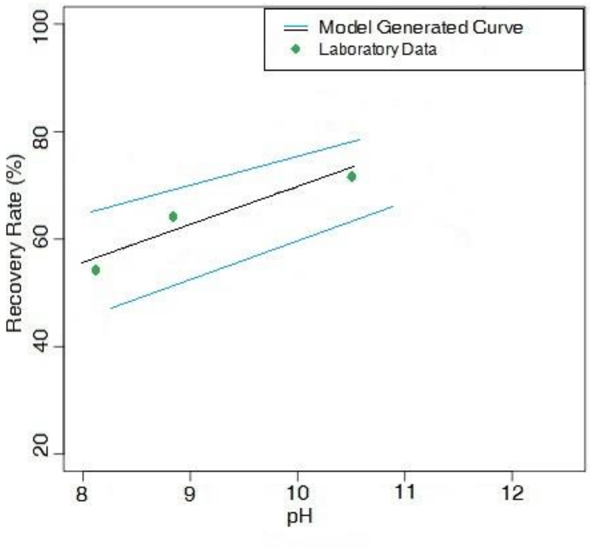


Figure 19 SVM model curve and simulated data curve comparison

(The optimized condition: Mg^{2+}:6.7; Ca^{2+}:11.2; Fines content:21%; Bitumen content: 15%; water content: 4.7%)

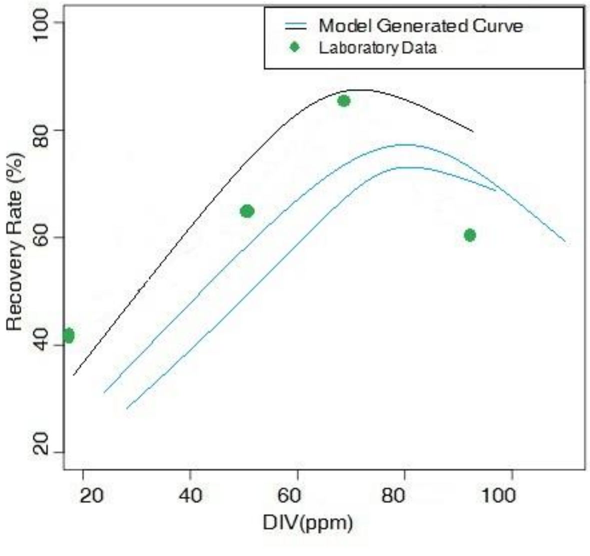


Figure 20 SVM model curve and simulated data curve comparison

(The optimized condition: pH: 8.3; Fines content: 38.7%; Bitumen content: 11.8%; water content: 4.2%)

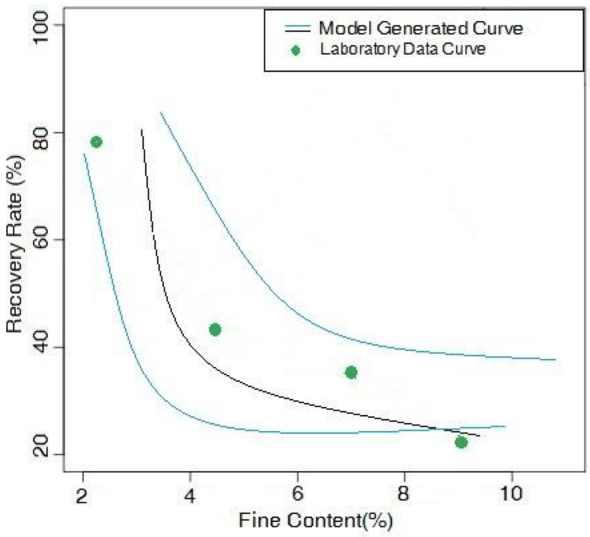


Figure 21 SVM model curve and simulated data curve comparison

(The optimized condition: Mg^{2+}:3.4; Ca^{2+}:7.0; Bitumen content: 9.3%; water content: 8.4%; pH: 8.6)

We compared three different kernels on oil sands dataset. The results of SVM prediction on oil sands dataset are indicated that SVM prediction has great impact on processability analysis. We continue to use prediction models on wine dataset to prove the superiority of SVM.

3.7 SVM on Wine Dataset

In order to compare the performance of proposed models, comparison of models are made using the mean absolute deviation (MAD), which is obtained from the difference between actual and predicted values of quality wines.

The mean absolute deviation is given by:

$$MAD = \sum_{i}^{N} \left| y_i - \hat{y}_i \right| / N \tag{43}$$

$\hat{y}_i$ the estimation of wine quality

y_i real quality for the sample of wine

N number of submitted sample of wine

The models are compared using the accuracy of classification in each quality class of wine samples. Classes are chosen with a tolerance of T=0.25, 0.5, 1.

SVM model is compared with two other models (Linear Regression (LR) and Neural Network (NN)). Before entering into the model, backward selection is performed to normalization. Train and test datasets are the entire suite of samples and only may vary by the deleted attributes.

In Table 3.3, the best accuracy of prediction (90.12 ± 0.44%) for wine quality is from the model SVM-1, which uses ten input attributes. The second best accuracy of prediction (90.51 ± 0.46%) is the model NN-1, which uses all eleven input attributes.

Attributes	LR-1	NN-1	NN-2	SVM-1	SVM-2
Fixed acidity	-	1	1	1	1
Volatile acidity	1	1	1	1	1
Citric acid	-	1	-	1	-
Residual sugar	-	1	1	-	-
Chlorides	-	1	-	1	-
Free sulfur oxide	-	1	1	1	1
Total sulfur oxide	-	1	1	1	1
Density	-	1	-	1	-
pH	1	1	1	1	-
Sulphates	1	1	1	1	1
Alcohol	-	1	1	1	1

Table 3.3 Attributes used to create models (1-attributes used)

In order to understand the processability, models for predicting wine quality are used by linear regression, neural network and support vector machine.

In Tables 3.4, the errors, the accuracy of models and selected attributes are shown in design of models. SVM models can reach accuracy 90.12 ±0.44% of prediction for T=1. This model has a lower cost for measuring physic and chemical parameters of each sample since it needs the small number of input attributes and is thus very suitable for determining the quality of wine.

Model	LR	NN-1	NN-2	SVM-1	SVM-2
MAD	0.59	0.48	0.47	0.59	0.5
T=0.25(%)	31.1±0.7	38.9±0.09	38.02±0.08	35.4±0.08	34.46±0.08

T=0.5(%)	59.1±0.3	60.85±0.2	61.23±0.2	58.97±0.2	58.22±0.2
T=1(%)	78.8±0.2	88.56±0.44	88.12±0.44	90.12±0.44	88.18±0.45

Table 3.4 Comparison of models

The relative importance of the SVM input variables is shown in Figure 22. A more detailed analysis will be given to sixth most relevant analytical tests (Figure 23). For a given input, each plot shows that analytical test values (x-axis) are changed. For a given test, we built a curve with L=6 points (the sensitivity levels).

The obtained results confirm the empirical outcomes from wine industry. For instance, an increase in the alcohol (the most relevant factor) tends to result in a higher quality wine. Figure 23 shows that this is true among the range from 9 to 13 % (which is related to most samples). In addition, the volatile acidity has a negative impact within the range that corresponds to the majority of the examples. This outcome is expected, since acetic acid is the key ingredient in vinegar. Moreover, residual sugar levels are important. The most intriguing result is the high importance of sulphates, ranked second. An increase in sulphates might be related to the fermenting nutrition, which is very important to improve the wine quality, in an effect that occurs within the range 0.4 to 0.7.

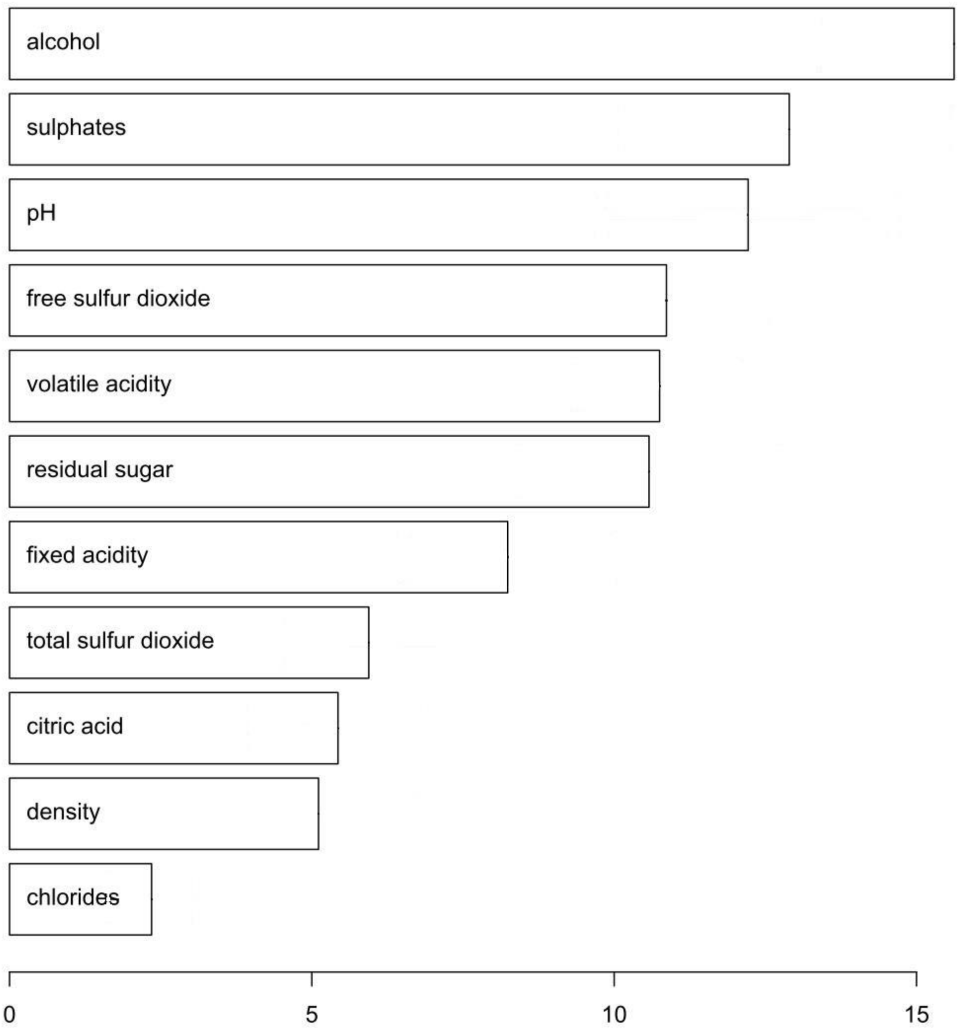
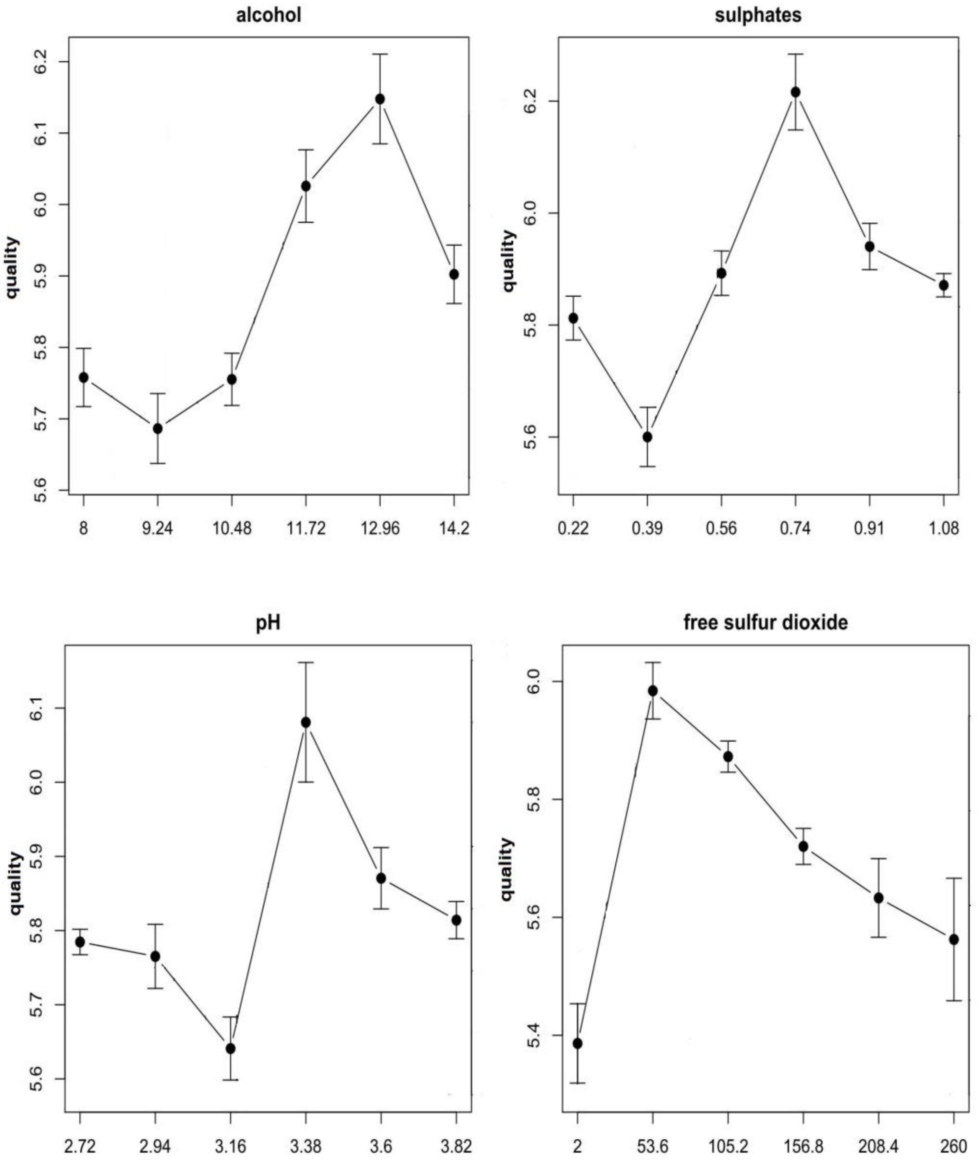


Figure 22 Wine variables priority for quality

This study aims at the prediction of wine preferences from objective analytical tests. This case study is addressed by a regression task, where wine preference is modeled in a continuous scale, from 0 (very bad) to 10 (excellent). The approach preserves the order of the classes, allowing the evaluation of distinct accuracies, according to the degree of error tolerance T.


alcohol
quality
5.6
5.7
5.8
5.9
6.0
6.1
6.2
8
9.24
10.48
11.72
12.96
14.2
sulphates
quality
5.6
5.8
6.0
6.2
0.22
0.39
0.56
0.74
0.91
1.08
pH
quality
5.6
5.7
5.8
5.9
6.0
6.1
2.72
2.94
3.16
3.38
3.6
3.82
free sulfur dioxide
quality
5.4
5.6
5.8
6.0
2
53.6
105.2
156.8
208.4
260

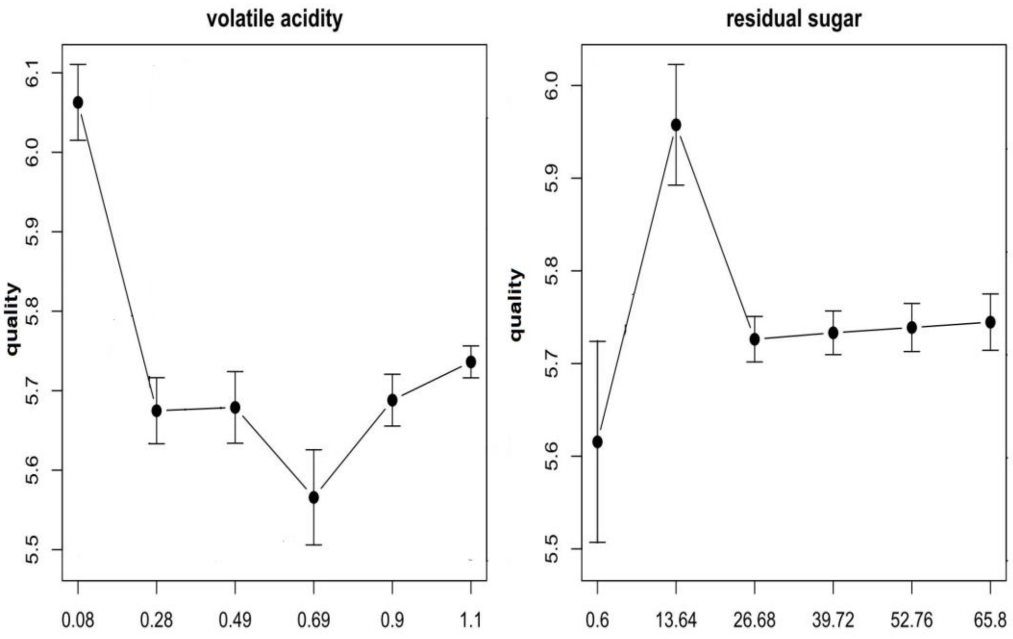


Figure 23 Wine variable effect curves for the SVM model

The performance of SVM model depends on a correct setting of parameters (e.g. SVM kernel parameter) and the input variables used by the model. In this study, we present an integrated and computationally efficient approach that simultaneously addresses both issues. Sensitivity analysis is used to extract knowledge from the SVM models, given in terms of the effect on the responses when one input is varied, leading to the proposed variable priority curves, and relative importance of the inputs (measured by the variance of the response changes). The variable selection is guided by sensitivity analysis and the model selection is based on trial and error search that starts from a reasonable value and is stopped when the generalization estimate decreases.

Encouraging results are achieved, with the SVM model providing the best performances and outperforming the NN and LR techniques. The overall accuracies are 58.97% (T= 0.5) and 90.12% (T= 1.0). It should be noted that the datasets contain six/seven classes (from 3 to 8/9) and these accuracies are much better than the ones conducted by a random classifier. While requiring more computation, the SVM model can still reach a high accuracy within a reasonable time range. Furthermore, the relative importance of the inputs brings interesting insights regarding the impact of the analytical tests. Since some variables can be controlled in the production process, the results of SVM model can be used to improve the wine quality.

Chapter 4

Conclusion and Future Work

4.1 Conclusion

In this thesis, statistical learning techniques are applied to the field of wine and oil sands datasets. The thesis consists of three main parts. In the first part of the thesis, after the introduction of oil sands industry and statistical learning theory, PCA and SVM are detailed explained in order to build classification and regression models for further prediction on experimental and artificial datasets.

In the second part, PCA methods are used to real world wine dataset to identify three different wine locations. Only with the first two principal components, the results of PCA show clearly identification of wine locations. Compared with the results of Kernel Principal Component Analysis, PCA proves itself as an efficient dimensional reduction technique, especially working on datasets with low covariance inputs.

In the last part of the thesis, SVM model is built for artificial oil sands dataset. Three different kernel methods, linear, polynomial and Gaussian kernel, are utilized to get the best accuracy results. The comparisons show that the polynomial kernel outperforms the other two kernel methods with less support vectors and lower cost. SVM model is also used to predict the important markers for oil sands dataset and to make comparison between laboratory data and model generated data, which shows SVM model successfully matches the original data curve. With respect to the prediction of wine quality, three multivariate analysis methods are compared. In this comparison, SVM model proves itself as an efficient method to find the key attributes and predict wine quality with high accuracy.

4.2 Future work

PCA is particularly useful if a limited amount of data is available. Once the data has been transformed to a low dimensional space, data is easier to handle with different methods, such as classification and regression.

The result of applying SVM method to the artificial dataset is encouraging. The results of this research can have a positive impact on the oil sands industry. The proposed data-driven approach is based on objective tests and thus it can be integrated into a decision support system, aiding the speed and high bitumen recovery rate and improving procedure performance. Furthermore, the relative importance of the inputs brings interesting insights regarding to the impact of the analytical tests. Since some variables can be controlled in the production process and more data can be added into oil sands datasets, the current results can be highly improved to enhance the oil sands recovery rate. If oil sands dataset from real world is applied to SVM model, the results will be very encouraging for oil sands industry.

The steps for future work:

1. Obtain a learning dataset containing ore analysis and processing conditions with corresponding recovery and froth quality.

2. Categorize ores in terms of bitumen content, water content, solids content and fines content (in three different classes: -44 um, -5 um and -2 um), type of clays (methylene blue index) and wettability, if all possible.
3. Categorize operating conditions in terms of temperature, pH, divalent cation concentration, energy input, and chemical additives.
4. Build customized SVM model to get realistic results of oil sands processability.

Bibliography

[1] J.Masliyah, Z.Xu, and J. Czarnecki, Handbook on Theory and Practice of Bitumen Recovery from Athabasca Oil Sand, *Kingsley Publishing Services*, 2011.

[2] J. Masliyah, Z.Zhou, Z.Xu, J.Czarnecki, and H.Hamza, "Understanding Water-Based Bitumen Extraction from Athabasca Oil Sands", *The Canadian Journal of Chemical Engineering*, Volume 82, August 2004.

[3] E.Theodoros and P.Massimilliano, Statistical Learning Theory: a Primer, *Springer*,1998.

[4] B.Olivier, B.Stephane, and L.Gabor, “Introduction to Statistical Learning Theory”, *Advanced Lectures on Machine Learning Lecture Notes in Computer Science,* Volume 3176, 2004.

[5] D. Skapura, Building Neural Networks, *ACM press*, 1996.

[6] T.Mitchell, Machine Learning, *McGraw-Hill Computer science series*, 1997.

[7] N.Cristianini and J.Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, *Cambridge University Press*, 2000

[8] V. Vapnik, Statistical Learning Theory, *Wiley*, New York, 1998.

[9] C. Cortes, “Prediction of generalization ability in learning machines”, PhD thesis, department of computer science, university of Rochester, USA, 1995.

[10] J.Liu, M.Bharati1, K.Dunn, and J.MacGregor, Chemometrics Intelligence Laboratory System, 2005.

[11] R. Burbidge, M. Trotter, B. Buxton, and S. Holden, Computer Chemometrics,2001.

[12] J.Rodr ǵuez, C.Alonso, and J.Maestro, Knowledge-based System, 2005.

[13] U.Thissen, R.Brakel, A.Weijer, W.Melssen, and L.Buydens, Chemometrics Intelligence Laboratory System, 2003.

[14] A.Borin, M.Ferrao, C.Mello, D.Maretto, and R.Poppi, Analytica Chimica Acta,2006.

[15] D.Wu, Y.He, S.Feng, and D.Sun, Journal of Food Engineering, 2008.

[16] S.Hua, and Z.Sun,Journal of Molecular Biology, 2001.

[17] Y,Cai, X.Liu, X.Xu, and K.Chou, Computer and Chemistry, 2003.

[18] A.Lorena, and A.Carvalho, Computers in Biology and Medicine,2007.

[19] C.Cai, W.Wang, L.Sun, and Y.Chen, Mathematical Biosciences, 2003.

[20] R.Aoki, T.Kimura, M.Kanaoka, FEBS Letter, 2005.

[21] S.Amendolia, G.Cossu, M.Ganadu, B.Golosio, G.Masala, and G.Mura, Chemometrics Intelligence Laboratory System, 2009.

[22] H.Liu, X.Yao, R.Zhang, M.Liu, Z.Hu, and B.Fan, Chemosphere, 2006.

[23] U.Thissen, M.Pepers, B.Ustun, W.Melssen, and L.Buydens, Chemometrics Intelligence Laboratory System, 2004.

[24] B.Ustun, W.Melssen, and L.Buydens, Analytica Chimica Acta,2007.

[25] J.Zhao, Q.S.Chen, X.Y.Huang, and C.H.Fang, Journal of Pharmaceutical and Biomedical Analysis,41 (2006)1198–1204.

[26] B.Kermani, I. Kozlov, P.Melnyk, C.Zhao, J. Hachmann, D. Barker, and M. Lebl, Sensors and Actuators B, Chemical. 125 (2007) 149–157.

[27] N.Ali, J.Saeed, and N.Davood, Journal of Hazardous Materials. 151 (2008) 603–609.

[28] J.Wang, H.Du, H.Liu, X.Yao, Z.Hu, and B.Fan, Talanta, 73 (2007) 147–156.

[29] S.Wold, A.Ruhe, H.Wold, and W. Dunn III, Journal on Scientific and Statistical Computing. 5 (1984) 735–743.

[30] F.Schwenker, H.Kestler, and G.Palm, Neural Networks, 14 (2001) 439–458.

[31] D.Rummelhart, and J.McClelland, Parallel Distributed Processing, *M.I.T Press, Cambridge*, 1986.

[32] J.Friedman, and W.Stuetzle, Journals of American Statistical Association. 76 (1981) 817–823.

[33] J. Yang and Y. Zhang, "Application research of support vector machines in condition trend prediction of mechanical equipment", *Lecture Notes in Computer Science*, May 30–June 1, 2005, vol. 3498, pp. 857–864.

[34] W.Hong, P.Pai, C.Chen, and P.Chang, "Recurrent support vector ma- chines in reliability prediction", *Lecture Notes in Computer Science*, vol. 3610, pp. 619–629, 2005.

[35] P.Pai and W.Hong, "Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms", *Electric Power System Research*., vol. 74, no. 3, pp. 417–425, 2005.

[36] W.Hong and P.Pai, "Predicting engine reliability using support vector machines", The International Journal of Advanced Manufacturing Technology, vol. 28, no. 1/2, pp. 154–161, Feb. 2006.

[37] J.Suykens, J.Vandewalle, and B.Moor, "Optimal control by least squares support vector machines", *Neural Network*., vol. 14, no. 1, pp. 23–35, 2001.

[38] S.Gezici, H.Kobayashi, and H.Poor, "A new approach to mobile position tracking", *Advances in Wired and Wireless Communications*, March 2003, pp. 204–207.

[39] X. Liu, J. Yi, and D. Zhao, "Adaptive inverse disturbance cancelling control system based on least square support vector machines", American Control Conference, June 2005, pp. 2625–2629.

[40] Q. Yang and S. Xie, "An application of support vector regression on narrow-band interference suppression in spread spectrum systems", *Lecture Notes Computer Science*, vol. 3611, pp. 442–450, August 2005.

[41] M.Ramon, N. Xu, and C.Christodoulou, “Beamforming using support vector machines”, *Antennas and Wireless Propagation Letters*, vol. 4, pp. 439–442, 2005.

[42] F.Luo, Y.Xu, and J.Cao, “Elevator traffic flow prediction with least squares support vector machines”, *Machine Learning and Cybernetics*, August 2005, pp. 4266–4270.

[43] G.Xu, W.Tian, and Z.Jin, “An AGO-SVM drift modelling method for a dynamically tuned gyroscope”, *Measurement of Science and Technology*, vol. 17, no. 1, pp. 161–167, Jan. 2006.

[44] L.Ya, L.Qi, and N.Meng, “Soft Sensing Based on LS-SVM and Its Application to a Distillation Column”, *Intelligent Systems Design and Applications*, 2006

[45] W.Yan and H.Wang, “Soft sensing modeling based on support vector machine and Bayesian model selection”, *Computer&Chemical Engineering*, 2004.

[46] M.PetkoviC, M.RapaiÇ, and B. JakovljeviÇ ,“Electrical Energy Consumption Forecasting in Oil Refining Industry Using Support Vector Machines and Particle Swarm Optimization”, *WSEAS Transactions on Information Science and Applications*, 2009.

[47] W.Xie, L.Yu, S.Xu, and S.Wang, “A New Method for Crude Oil Price Forecasting Based on Support Vector Machines”, *International Conference on Computational Science*, 2006.

[48] Jiekun S, Zaixu Z, “Oil Refining Enterprise Performance Evaluation Based on DEA and SVM”, *Knowledge Discovery and Data Mining*, 2009.

[49] V.Venkoparao, R.Hota, V.Rao, and M.Gellaboina, “Flare monitoring for petroleum refineries”, *Industrial Electronics and Applications*, 2009.

[50] R.Balabin, R.Safieva, and E.Lomakina, “Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques”, *Analytica Chimica Acta*, 2010.

[51] T.Verplancke, S.Van Looy, and D.Benoit, “Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies”, *BMC Medical Informatics and Decision Making*, 2008