GAZİ UNIVERSITY

Gazi

1926

A NEW OUTLIER DETECTION METHOD BASED ON PROBABILISTIC OUTPUTS OF SUPPORT VECTOR MACHINES IN BINARY CLASSIFICATION

A THESIS SUBMITTED TO THE GRADUATE SCHOOL OF
NATURAL AND APPLIED SCIENCES
OF
GAZİ UNIVERSITY

BY
Habib CEESAY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
DEPARTMENT OF STATISTICS

JULY 2019

The thesis study titled " A NEW OUTLIER DETECTION METHOD BASED ON PROBABILISTIC OUTPUTS OF SUPPORT VECTOR MACHINES IN BINARY CLASSIFICATION" is submitted by Habib CEESAY in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistics, Gazi University by the following committee.

Supervisor: Doç. Dr. Filiz KARDİYEN

Statistics, Gazi University

I certify that this thesis is a Master of Science thesis in terms of quality and content ………………………….

Chairman: Prof. Dr. M. Akif BAKIR

Statistics, Gazi University

I certify that this thesis is a Master of Science thesis in terms of quality and content ………………………….

Member: Doç. Dr. Rukiye DAĞALP

Statistics, Ankara University

I certify that this thesis is a Master of Science thesis in terms of quality and content ………………………….

Date: 12/06/2019

I certify that this thesis, accepted by the committee, meets the requirements for being a Master of Science Thesis.

…………………….………

Prof. Dr. Sena YAŞYERLİ

Dean of Graduate School of Natural and Applied Sciences

ETHICAL STATEMENT

I hereby declare that in this thesis study I prepared in accordance with thesis writing rules of Gazi University Graduate School of Natural and Applied Sciences;

- All data, information and documents presented in this thesis have been obtained within the scope of academic rules and ethical conduct,
- All information, documents, assessments and results have been presented in accordance with scientific ethical conduct and moral rules,
- All material used in this thesis that are not original to this work have been fully cited and referenced,
- No change has been made in the data used,
- The work presented in this thesis is original,

or else, I admit all loss of rights to be incurred against me.

Habib CEESAY
12/06/2019

A NEW OUTLIER DETECTION METHOD BASED ON PROBABILISTIC OUTPUTS OF SUPPORT VECTOR MACHINES IN BINARY CLASSIFICATION

(M. Sc. Thesis)

Habib CEESAY

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE

July 2019

ABSTRACT

With data growing so rapidly, classification has become one of the most important and effective tools in Machine Learning and Applied Statistics, in which a given observation can be predicted in the right class given some features. Classification is used in most sectors such as; Biomedical Studies, Genetics, Social Science, Marketing, etc. Data are said to be binary when each observation falls into one of two categories, such as: alive or dead, positive or negative, etc. Support Vector Machines are a class of statistical models first developed in the mid-1960s by Vladimir Vapnik and they are very flexible due to the incorporation of Kernel Functions which can help separate and classify data that are not linearly separable. However, Support Vector Machines can suffer a lot from unclean data containing, for example, outliers or mislabeled observations. The goal of this thesis is to compare the classification accuracy of the SVM on both clean and contaminated data and also a new method based on the probabilistic outputs of SVM (PoC) is proposed. The outlier detection rate for this new method and the Robust Mahalanobis distance (MCD) are compared. The results show that PoC performs better than MCD at detecting outliers.

Science Code	:	20513
Key Words	:	Support Vector Machine, Logistic Regression, Classification, Kernel Function, Outliers
Page Number	:	73
Supervisor	:	Assoc. Dr. Filiz KARDİYEN

İKİLİ SINIFLAMA PROBLEMİNDE AYKIRI GÖZLEM TESPİTİ İÇİN DESTEK VEKTÖR MAKİNELERİ OLASILIKSAL ÇIKTILARINA DAYALI YENİ BİR YÖNTEM

(Yüksek Lisans Tezi)

Habib CEESAY

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Temmuz 2019

ÖZET

Hızla büyüyen veri teknolojisi ile, belirli özelliklere sahip bir gözlemin doğru sınıfa atanması bağlamında sınıflandırma Makine Öğrenmesi ve Uygulamalı istatistik alanlarında en önemli ve etkin araçlardan biri haline gelmiştir. Sınıflandırma, biyomedikal çalışmalar, genetik, sosyal bilimler, pazarlama gibi pek çok alanda kullanılmaktadır. Her bir gözlemin sağ-ölü, pozitif-negatif gibi iki kategoriden birine ait olduğu veriye ikili very denir. Destek Vektör Makineleri ilk olarak 1960'ların ortasında Vladimir Vapnik tarafından geliştirilen doğrusal olarak ayrılamayan veriyi sınıflandırmaya yardımcı Kernel fonksiyonlarının da kullanımı ile oldukça esnek bir istatistiksel modeller sınıfıdır. Ancak SVM verinin aykırı gözlem veya yanlış veri gibi kirlenmiş gözlem içermesinden olumsuz yönde etkilenebilir. Bu tez çalışmasında amaç, SVM'nin temiz ve kirli veri için sınıflandırma kesinliğini karşılaştırmak olup, çalışmada Destek Vektör Makinelerinin olasılıksal çıktılarına dayanan (PoC) yeni bir aykırı değer tespit yöntemi önerilmiştir. Önerilen yöntem ile Sağlam Mahalanobis uzaklığı (MCD) yönteminin aykırı gözlem tespit oranları karşılaştırılmıştır. Sonuçlar, önerilen yöntemin daha iyi performans gösterdiğini göstermiştir.

Bilim Kodu	:	20513
Anahtar Kelimeler	:	Destek vektör makineleri, lojistik regresyon, sınıflandırma, Kernel fonksiyonları, aykırı gözlemler
Sayfa Adedi	:	73
Danışman	:	Doç. Dr. Filiz KARDİYEN

ACKNOWLEDGEMENT

I will first of all thank God for giving me the life, health and strength to complete this thesis. It would not have been possible without the prayers, encouragement, sacrifice and support of my family from the first day especially my wonderful, amazing and caring mother Mrs. Mariama Manneh. I could not have done it alone successfully without the support and never-ending understanding of my lovely, supportive and caring wife, Mrs. Mariama Njie Ceesay.

I wish to send a special thanks to my able supervisor Doç. Dr. Filiz KARDİYEN for her endless support, advice and guidance during this period, I appreciate all the humble academic discussions we had, and I will forever appreciate the positive impacts they have on me, not only on my career but my life in general. I cannot forget the immense support, discussion and advice from Prof. Dr. M. Akif BAKIR. Despite his busy schedule, he is always available whenever needed. Thank you prof.

My heartful gratitude and respect to Mr. Serign Modou Njie (Deputy Ambassador in Turkey during this period) for his endless support in all forms with his entire family. Also, Mr. Musa Jobarteh (Financial attaché at the Embassy during this period) and his family for their support during this period. Special thanks to Dr. Bumi Camara and his humble wife Mrs. Fatou Susso Camara for their support and advice. Thank you Mr. Kutubo Gitteh and Mr. Abdou Darboe for all your support. I will take this opportunity to thank every person I have interacted with at Gazi university who has helped me directly or indirectly; students, Lecturers and admin staffs. It is all appreciated.

Finally, I will like to thank all my friends for their support and prayers.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

SYMBOLS AND ABBREVIATIONS

The symbols and abbreviations used in this thesis are given below:

Symbols	**Explanation**
ρ	Rho
$\boldsymbol{\alpha}$	Alpha
$\boldsymbol{\beta}$	Beta
$\boldsymbol{\gamma}$	Gamma
$\boldsymbol{\varepsilon}$	Epsilon
$\boldsymbol{\lambda}$	Lambda
$\boldsymbol{\chi}$	Chi

Abbreviations	**Explanation**
D	Deviance
DK	Dragon Kings
DNA	Deoxyribonucleic acid
GRPF	Gaussian Radial Basis Polynomials Function
LR	Logistic Regression
MCR	Misclassification Rate
MDA	Multiple Discriminant Analysis
MRS	Max-Robust-Sum
MS	Max-Sum
PSSP	Protein Secondary Structure Production
RBF	Radial Basis Function
RDM	Robust Mahalanobis Distance
RFE	Recursive Feature Elimination
RKHS	Reproducing Kernel Hilbert space
SRS	Sum-Robust-Sum
SS	Sum-Sum Test Statistics

Abbreviations	**Explanation**
SVM	Support Vector Machine
TWSVM	Twin Support Vector Machine

1. INTRODUCTION

With the rapid growth of data, classifying a given data in its right class known as classification has become one of the most important statistical tools that benefits the society. Every sector in the society requires a correct classification of a given observation when their attributes are given. One of the most important field in which classification has a huge significance in our lives is in the medical field. For instance, predicting the right disease given all the symptoms affecting the patient plays a vital role in the process of treating the patient, especially when they suffer from different illnesses. In such a case, a doctor will be interested to correctly classify a patient in a right class after the diagnosis. In a bank, suppose they are giving some loans to some of their customers, the manager will be interested in deciding and classifying a customer into the right class depending on whether the customer will repay the loan or not given some attributes of the customer (which can be the previous financial dealings of the customer or the customer's current financial situation, etc.,). In general, it is necessary for almost all sectors in our societies to have an accurate and reliable classification tool which can be used in order to make the correct decision by classifying every observation into their right class given some attributes of the observations. Wrong classification of observations can be disastrous; as in the example discussed above, if the doctor classified the patient in the wrong class, the first error will be that the prescribed medicines will be the wrong medicines for the patient's current sickness and it can even give the patient some more serious complications than the patient's current sickness; if the bank manager did not classify the customers correctly, he/she will either end up giving the loan to the wrong customers who will not repay the loan by the deadline or deny the right customers who are qualified to have the loan.

In this thesis, we are dealing with a binary classification of observations with the presence of outliers and the methods we are going to use to determine the effectiveness of the classification are the Logistic Regression and Support Vector Machine. Logistic Regression has been widely used by statisticians for a long time now before the invention of the SVM for binary classification. Its simplicity and accuracy made it a very popular technique of binary classification. The log of the odds of an event will give us the logistic regression of the model. The odd of an event is the ratio of the probability of an event happening to the probability of the event not happening.

The Support Vector Machine has been developed in the 1960s by Vladimir Vapnik and it has become a very effective and reliable classification tool for binary data since its invention. It has the maximal margin hyperplane which perfectly linearly separates the data into two separate classes and the separating line between the two classes is called the hyperplane. The closest observations from either of the classes to the hyperplane are called the support vectors while the interval between the support vectors to the hyperplane is what is known as the margin. Unfortunately, not all data in the world are perfectly linearly separable. Nonetheless, some can be linearly separable but there will be a misclassification of some other observations, and this is due to the fact that the observation that is not correctly classified will either be on the wrong side of the margin or even on the wrong side of the hyperplane. Some data cannot be separated by the liner hyperplane at all, in those cases the Kernel functions will help to give a smooth non-linear boundary for the separation of the non-separable cases by expanding the feature space of the variables.

An observation is said to be an outlier if it is completely different from all the remaining observations in the data and they are usually formed by a completely different function or distribution and are mostly found at the extreme ends of the distributions. For any given data, mining the availability of an outlier is a very important primary step to take since the presence of an outlier can give a poor estimate of parameters which are used to describe the general behavior of the data. In short, the outliers will attract or skewed the data towards themselves which will lead to a poor estimate of the parameters and a wrong conclusion about the data. Outliers can occur in a data either due to the bad recording of the data (error in coding) or to an experimental error. In cases where it is certain that a mistake has occurred, the mistake can either be corrected in the coding or experiment or be completely deleted from the data since it is an outlier. But there are some cases where deleting the outlying observations from the data is not permissible since they may carry some useful information about the data, in such cases, the data is further examined. In this thesis, the focus is on classifying a binary data in the presence of outliers. Sometimes some outliers can mask other outliers and it will not be able to detect them unless the first outlier is deleted. This is what is known as the masking effect. On the other hand, a good observation can be mistakenly taken as an outlier and this is known as the swamping effect. Later in the discussion, the two binary classification methods: SVM and logistic Regression are compared, by using some of the outlier detection methods such as the Robust Mahalanobis distance, multivariate outlier detection, masking and swamping effect, etc. to see their classification accuracy and their resistance to the presence of outliers in a data by using different statistical distributions.

2. LITERATURE REVIEW

Guyon, Weston, Barnhill and Vapnik (2002) discussed the problem of selecting small subset of genes from a larger pattern of gene expression data which were recorded on DNA micro-arrays. With the available training examples from patients who are having cancer and normal patients, they built a classifier suitable for genetics diagnosis and as well as drug discovery. With the use of Support Vector Machine based on Recursive Feature Elimination (RFE), they have shown that it yields a better classification performance and are biologically related to cancer as compared to preceding efforts to resolve the same problem select gene with correlation techniques.

Sun, Lim and Ng. (2002) proposed using Support Vector Machine (SVM) classifiers to categorize web pages using both their text and content feature sets. This is because web pages not only plain text documents, and that web classification methods should involve the use of supplementary context features of web pages such as hyperlinks and HTML tags. When they compared the result with other methods on the same data as FOIL-PILES method, the results indicated that the SVM method yield a better result. Their research also indicated that the use of other context features like hyperlink can meaningfully expand the classification performance.

Min J.H. and Lee Y.C. (2005) applied Support Vector Machine (SVMs) to bankruptcy prediction problem in order to propose a new model that has a better explanatory power and stability since it has drawn a lot of research interest in past literatures. Researches conducted recently have indicated that machine learning techniques have accomplished better performance than traditional statistical techniques. They used a grid search technique using 5-fold cross validation to find out the optimal parameter values of Kernel Functions of SVM. To evaluate the prediction accuracy with other classification methods like Multiple Discriminant Analysis (MDA) and Logistic Regression analysis. Their experimental result showed that SVM outperformed the other methods such as. Multiple Discriminant Analysis and Logistic Regression.

Wang, L. ed., (2005) used Support Vector Machines (SVM) to compare and contrast two bioinformatic problems, i.e., the cancer diagnosis based on gene expression data and protein secondary structure prediction (PSSP). The conclusion of their research indicates that SVM performs well in both bioinformatic problems. In the case of the cancer diagnosis based on micro-array data, the SVM they used performed better with a more accurate result in relations to the number of genes required as compared to alternate methods that were proposed in the past. For that being the reason, they concluded that the SVM makes highly reliable prediction and that it also reduces unnecessary genes. For the PSSP problem, the SVM also obtained results that could be compared to results obtained by other methods.

Joachims. T. (1998) explored the use of Support Vector Machines (SVMs) for learning text classifiers. For example, they analyze the properties of learning with text data and identifies why SVMs are appropriate for this task. Their research finding showed that SVMs constantly achieved good performance on text categorization tasks, and they substantially outperform existing methods in a significant way. SVM makes the application of text categorization remarkably easier and as such it eradicates the need for feature selection. Also, their ability to automatically find parameter settings eliminates the need for parameter tuning.

Tsai. C. F. (2005) presented a two-level stacked generalization scheme composed of three generalizers of Support Vector Machines (SVMs) for image classification - the color, texture, and high-level concept SVMs. Their focus is to examine two training approaches based on two-fold cross-validation and non-cross validation for the proposed classification scheme by gauging their classification performances, margin of the hyperplane, and numbers of support vectors of SVMs. And their result showed that the non-cross validation training methods performed better, having higher correct classification rate, larger margin of the hyperplane, and smaller numbers of support vectors.

In another study, Zanaty, E.A., (2012) evaluated and compared Support Vector Machines with dissimilar Kernel Functions and multi-layer neural networks to a different type of non-separable dataset with numerous attributes. Zanaty introduced a new kernel function that can improve the correctness of the support vector machines (SVMs) and the proposed kernel function is called Gaussian Radial Basis Polynomials Function (GRPF) which combines both Gaussian Radial Basis Function (RBF) and Polynomial kernels. In the end, in almost all the data sets, it turns out that the proposed kernel function gives a better accuracy especially of those in high dimensions. It also gave a better performance than those with existing kernels.

Este, Gringoli, and Salgarelli (2009) described a method to traffic classification based on Support Vector Machine (SVM). Since it is designed for binary classification, their generalization to multi-class problems is still under investigation. And their performance is highly vulnerable to the correct optimization of their working parameters. The accuracy of the proposed classifier is then evaluated and measured over three sets of traffic traces, coming from different topological points in the internet. Their task was to solve a traffic classification problem by applying one of the methods to solving multi-class problems with the use of SVM and describe a simple optimization algorithm that allows the classifier to perform correctly with just a few hundred samples. In their results, they have concluded that even with reduced training set sizes, SVM-based classifiers can be very effective at discriminating traffic generated by different applications.

Qi, Tian and Shi, (2013) designed a new structural Twin Support Vector Machine (TWSVM) since the structural information of data may contain a useful prior domain knowledge for training a classifier. The common method of factoring all structural information within classes into one model is the all existing structural large margin methods, but this method failed to balance infra-class and inter-class and as a result prior information is not efficiently exploited. TWSVM uses two hyperplanes to decide the class of New data, of which each model only considers one class structural information and closer to the class at the same time far away from the other class. This makes it fully exploit this prior knowledge to directly improve the algorithm’s capacity of generalization. They concluded that the proposed method is rigidly superior based on structural information of data in both computation time and classification history.

Salazar, D.A., Vélez, J.I. and Salazar, J.C (2012) compared the effective performance of classification of Support Vector Machine to logistic Regression in a statistical simulation study by using different statistical distributions. They have concluded that to envisage the class of an observation on the basis of a single variable, the SVM model is better as compared to LR by using the Poisson, Exponential and normal distribution but the polynomial SVM model is not recommended since its misclassification rate is higher. SVM has a better performance to LR when there is a high correlation in the data set in the case of a multivariate and mixture of distribution.

Wheatley, and Sornette, (2015) conducted a simulation study to examine the extent to which different statistics models are affected by the masking and swamping effect in which the tests are done on a synthetic data for a range of block sizes. The tests are carried out in three cases:

(i) swamping due to a single outlier,
(ii) swamping without masking due to dispreads outliers, and
(iii) swamping with masking due to clustered outliers.

From the conclusion based on their simulation study, in cases where large observations are closely clustered, masking effect is more problematic, i.e. it suffers masking effect more for example as in case III. Fortunately, the test statistics which are based on sums overcome masking effect and they recover from swamping faster than those based on spacing and maxima; the robust test statistics are less expose or less affected by the masking effect as intended. As the size of the block becomes bigger than the actual size of the block, the rate of the rejection decays slowly and this indicate that the smallest p-value in the sequence of estimates will not give the true block size. This leads them to compare the inward and outward sequential procedures in four scenarios:

(i) outward test with Max-Sum (MS), Max-Robust-Sum (MRS), Sum-Sum (SS) and Sum-Robust-Sum (SRS) test statistics,
(ii) the inward procedure with only the MRS test statistics that is essential to evade masking and swamping,
(iii) the mixture model, and
(iv) the SRS block test, given the correct number of outliers.

The fourth case was the best performing block test and it gives them the benchmark. i.e. they compare the performance of the other tests to the fourth case. The inward and mixture models gave a false positive number of small outliers when there were no outliers while the outward procedure gives false large number of outliers when there was no outlier. In the case where they identified a single outlier, the inward test is most powerful, it is even close to the power of the block test and it provides a higher estimation of outliers where the other tests tend to overrate. The outward test and mixture model perform best in cases with cluster outliers, even better than the block test. All the inward and outward are almost the same effective in cases of a multiple dispersed outliers and they are slightly more powerful than the block test.

They have concluded that: in the case of identifying a single and multiple dispersed outlier, the inward procedure with the MRS test statistics is more powerful than the outward procedure. The mixture approach can be more powerful and effective in identifying a dense cluster of outliers than the outward approach. The MS statistics was superior and more powerful compared to the all the outward approach, and the robust modification performs similarly.

Musa, A.B. (2013) conducted a comparison study between Support Vector Machine and Logistic Regression in a different manner that is different from most of the machine learning comparisons by using bagging and ensemble on different sizes of balanced and unbalanced data set. SVM is comparatively a new machine learning algorithm that is used for classification while on the other hand, LR is an old standard statistical classification method. Different statistical analysis was used on several algorithm performance measure which enabled them to come up with a strong conclusion. The study includes several measures to assess the classification performance of the learning methods: accuracy, sensitivity, precision and specificity. The study indicated that both SVM and LR on the different performance measures have the same performance for both balanced and unbalanced date sets but SVM is a better choice to work with when there is a high unbalanced data. The interpretability is higher in LR while the SVM is a black box predictor i.e. it neither makes its prediction understood nor gives the cause in the rules governing its prediction. Either of the method can be used when the primary interest is just for classification but LR should be preferred above SVM when explanation is essential such as in the medical field.

Khanna, Sahu, Baths and Deshpande (2015) also embarked on a comparison study on the commonly used machine learning algorithm of classification (SVM, LR and Neural Networks) in the field of medicine by predicting the existence of a heart disease in a patient, primarily focusing on the data from the University of California heart disease dataset. They indicated patients with heart disease to be 1 and the healthy individuals 0. The data is split into two halves, the first half was used to train the model and the parameters are selected through cross-validation and the accuracy of the model was tested using the remaining half of the data, i.e. the testing data. This was done to avoid the model from biasing and to give a fresh perspective for testing the model. Their results have shown that SVM turns out to be a very good method and approach for precise predictions of heart disease especially seeing classification accuracy as a performance measure. The neural network also gives good results as compared to the classical methods. Generally, for this heart disease dataset, simpler methods like Logistic Regression and SVM with linear kernel turns out to be more impressive. The results of their study have also been used in making technologies for precise prediction of heart disease in hospitals thereby contributing to the science of medical diagnosis and analysis.

Salazar, D.A., Vélez, J.I. and Salazar, J.C (2012) engaged in a statistical simulation study between SVM and LR to determine which one is better to discriminate in the classification of a new observation into either of the two groups. LR have been used in most of the classification problems in different fields. With the invention of the machine learning approaches such as SVM and Neural Networks, in these methods, a data is given to the machine and the results are obtained at the end without knowing exactly what happened in between, and therefore they are considered as “black box”. In their simulation study, they determined the Mis-Classification Rates (MCR) of SVM and LR of observations that come from a population to be determined into which of the two groups it belongs to by using different probability distributions such as: Poisson, Exponential, Normal Cauchy-Normal, Normal-Poisson, Bivariate-Normal in which the training data set and other functional parameters are controlled. From their results and conclusion, the MCR for the polynomial SVM is higher in the Normal distribution while the performance of LR, linear and radial SVM are almost all the same. When the two groups have the same number of observations and they both come from a Poisson distribution, other methods perform better than the polynomial SVM kernel. However, in the case where the sample sizes are not equal in the two groups, LR is preferable to SVM methods. When the number of observations is equal in both groups in the Exponential case, SVM models perform equally well than LR except for the polynomial kernel. The polynomial SVM is not recommended in both the Normal and Poisson distribution.

Kannan, K.S. and Manoj, K. (2015) used several distance measure techniques such as: Mahalanobis Distance, Cooks Distance, Leverage Point and DFFITS in their research to detect outliers in a multivariate data but unfortunately most of those measures depends on the sample mean and covariance matrix which leads to a poor result due to the fact that those measurements themselves are affected by outliers and they tend to pull the results of the regression closer to themselves. In some cases, one outlier tends to hide (mask) another outlier, and the hidden outliers can only be detected after deleting the other outlier. An appropriate method is implemented to identify the unmasking outliers and to compare the various distance measures. From their results, they concluded that the outlier detection level for both Mahalanobis Distance and Leverage Point are almost the same, 9, but the outlier detection sensitivity of the DFFITS are very low, 5, while it is very high using the Cooks Distance since it identified the maximum number of outliers in the data set, 11.

3. CLASSIFICATION

In most of the statistical applications, the observations that are to be examined take one of the two possible outcomes such as, a success or failure, when they are exposed to a certain experimental condition. For example, a seed planted by a farmer will either germinate or fail to germinate; a device produced by an Engineer may be defective or non-defective; an insect that is exposed to an insecticidal trial may either survive it or die from it; a student may be accepted or not accepted into a specific program under certain conditions. These kinds of data are known as binary data. The Linear Regression Models work well with cases where the response variable Y is quantitative (discrete or continuous). For example, to find the cost of producing an equipment given the prices of the raw materials involved in the production, but what happens when the response variable Y is qualitative? in this case, the Linear Regression Models will not be able to give a proper solution to the question. Qualitative data are said to be classified into classes or categories. For instance, an emergency room in a hospital, a patient arrives with a set of medical symptoms that could possibly be attributed to one of several medical conditions. Determining which of the several medical conditions the patient has requires a proper classification in the right class *(James, Witten, Hastie, and Tibshirani, 2013).* The prediction of such cases is what is known as classification. In this case, the patient must be diagnosed and classified correctly so that the right and proper medications can be given to the patient. One of the most widely used and studied method for classification in machine learning is the supervised learning in which a learning algorithm uses labeled instances to formulate a predictive model. There are many ways for classification of data, but the commonly used methods are: The Logistic Regression, Support Vector Machine, Decision tree, Linear Discriminant Analysis and K-nearest neighbors.

The classification setting is similar to a regression setting in a sense that it also has a set of training observations (x_1, y_1) ... $(x_n\ , y_n)$ used to build a classifier. The classifier is not only built to perform on the training data, but also to test observations that were not used to train the classifier.

3.1. Decision Tree

A decision tree is just like the beginning of a road that leads to different destinations, and a choice is to be made on which road to be taken because they all have their different rewards at the end of the road. It is used to elucidate and establish answers to a sophisticated problem. As progress is made from the main decision to the conclusion, there exist branches that classify your data into their respective classes. The structure allows a problem with numerous possible solutions to be displayed in a simple readable format. Each branch of the decision tree represents a possible decision, occurrence or reaction and at the same time it classifies the data with the same characteristics on the same branch. The tree is designed such that it depicts by what means and why a choice may lead to the other. The furthest branches of the tree represent the result (https://www.investopedia.com/terms/d/decision-tree.asp) [38].

3.2. The Support Vector Machine

Vladimir Vapnik was the first to have developed a class of statistical models referred to as Support vector machines in mid-1960s. The model has progressed substantially into one of the most flexible and effective machine learning tools widely used in classification and recognition task in supervised learning. Their strong theoretical background makes them very important and necessary in the field of classification. SVM is a classification and regression technique that has both a good computational algorithm and a strong theoretical result that gives it a good reputation and increases its use in different fields. (Salazar, D.A., Vélez, J.I. and Salazar, J.C., 2012.). Support Vector Machines are proposed for binary classification settings in which there are two classes. Maximum margin classifier is a classifier that is applied on data that can be linearly separable. However, this limits its usage into most of the data set. Fortunately, there exists the support vector classifier which is an extension of the maximum margin classifier. This extension allows the misclassification of the data during classification. Support vector machine is an extension of the support vector classifier and it can separate data that are not linearly separable. The idea of a support vector machine is to find a line called the hyperplane which will divide or separate the data in the best possible correct ways, so that different classes of the data will be far apart as much as possible.

3.2.1. The hyperplane

The line separating the classes of the data is called a hyperplane since it can work with any number of dimensions. Classifying the data with one dimension, then the hyperplane is a point. If it is a data with two dimensions, then the hyperplane is a line, and data with three dimensions has a hyperplane called plane. But if a data has four or more dimensions then the line that separates the different classes of data might be hard to visualize, so it is just called a hyperplane. In general, if the data is having a p dimension, then the hyperplane will have p-1 dimension.

Since the hyperplane for a data with two dimensions is a line, then the equation of the hyperplane is represented as below:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0 \tag{3.1}$$

From the equation above, when the solutions of the values of the parameters β_0, β_1 and β_2 are obtained, then for any $X = (x_1, x_2)^T$that satisfies the equation above is a point on the hyperplane. But the equation can also be extended if it is a data with p dimensions:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \cdots + \beta_p x_p = 0 \tag{3.2}$$

If a point $X = (x_1, x_2, x_3, \ldots x_p)^T$ satisfies the equation for the p dimensional data, then the point X lies on the hyperplane.

Unfortunately, this is not always the case when we substitute a point X, there are cases when a point does not satisfy the hyperplane equation. In such cases, there exist two cases:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \cdots + \beta_p x_p > 0 \tag{3.3}$$

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \cdots + \beta_p x_p < 0 \tag{3.4}$$

From the two inequalities above, all the points that are greater than zero will be on one side of the hyperplane while all the points that are less than zero will be on the other side of the hyperplane provided that the hyperplane linearly divides the data into two classes. Suppose the data has n training observations in p-dimensional space whose outcomes fall into two classes as {-1,1} where -1 represent one class and 1 represent the other class. Based on these training data, a classifier can be developed that will correctly classify the test observations in their right class. The inequalities below demonstrate the decision rule for this case:

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \cdots + \beta_p x_{ip} > 0 \; if \; Y_i = 1 \tag{3.5}$$

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \cdots + \beta_p x_{ip} < 0 \; if \; Y_i = -1 \tag{3.6}$$

A separating hyperplane has the property below:

$$Y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \cdots + \beta_p x_{ip}) > 0. \tag{3.7}$$

After having a separating hyperplane, this can now help in the construction of a classifier. A test observation will now be used to classify it on the right side of the hyperplane. Suppose the test observation is labelled as x^*, it can be classified on the correct side of the hyperplane based on the sign of the equation below:

$$f(x^*) = \beta_0 + \beta_1 {x_1}^* + \beta_2 {x_2}^* + \cdots + \beta_p {x_p}^* \tag{3.8}$$

If $f(x^*)$ is greater than zero, then the test observation is assigned to Class 1, if $f(x^*)$ is less than zero, then it is assigned to Class -1.

One important point that can be concluded from the value of $f(x^*)$ is that it gives the researcher more confidence about the classification accuracy of the test observation if its value is not close to zero. The suggestion is that, the test observation is not close to the hyperplane and is correctly classified. On the other hand, it gives less confident about the classification of the test observation if it is closer to zero, i.e. it is close to the hyperplane.

3.2.2. Optimal separating hyperplane

Since there exist infinitely many hyperplanes that can separate a linearly separable data into two classes, the hyperplane that best separate the data must be chosen. The optimal separating hyperplane separates the data into two classes and maximizes the distance to the closest point from either class (Vapnik, 1996). It does not only give a unique solution, but it also maximizes the distance between the two classes on the training data, which gives a better classification performance on test data. The distance between the observation points and the hyperplane are measured and the smallest distance is known as the MARGIN. The goal is to find the farthest minimum distance that will give the classifier the maximum margin classifier. The classifier will likely classify the test data correctly into their right classes if it gives a large margin on the training data because that will also provide a large margin on the test data. The movement of just any data observation will not affect the position of the hyperplane unless and until if it crosses the boundary set by the margin or enter the wrong side of the hyperplane or if that observation is a support vector. SUPPORT VECTORS are observations that gives the width of the margin since they are the closest observation of a class to the hyperplane, any movement or shift of those points will affect the hyperplane since they help to determine the width of the margin and the hyperplane directly depends on them.

In the other hand, the solution to the optimization problem below is the maximal margin classifier:

$$maximize\ M\ \beta_0, \beta_1, \beta_2, \dots, \beta_p$$

$$subject\ to\ \textstyle\sum_{j=1}^{p} {\beta_j}^2 = 1$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M\ \ \forall\ \ i = 1, \dots, n. \tag{3.9}$$

Given that *M* is a positive value, this will give the researcher the assurance that every observation will be in the right class of the hyperplane. “M represent the margin of the hyperplane, and the optimization problem chooses $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ to maximize *M*. This is exactly the definition of the maximal margin hyperplane.” (James *et al*, 2013)

The equation can also be transformed so that the maximum margin classifier can be written in terms of each data point in the data in the sample as:

$$D(x_i) = \beta_0 + \textstyle\sum_{i=1}^{n} y_i \alpha_i x_i {x_i}' \tag{3.10}$$

With $\alpha_i \geq 0$ and it is exactly equal to zero for all the samples that are not on the margin. For this reason, the prediction equation is the function that contain only the observations that are on the margin and they are called the support vectors (James *et al*, 2013).

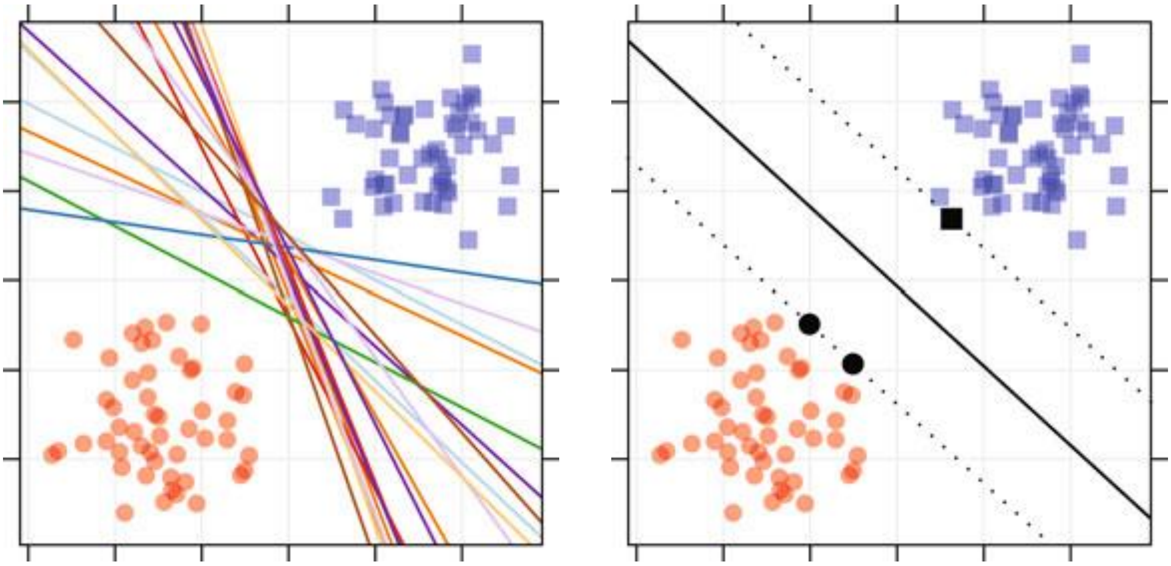


Figure 3.1. A perfectly linearly separable case with the maximal margin classifier

3.2.3. The non-separable case

Sometimes the available data cannot be perfectly separated into the different classes by the maximal margin classifier since there will exist some misclassifications in the data. In this case, a different separating hyperplane should be used that will be able to separate almost all data sets correctly, one of the main methods is the soft margin classifier or support vector classifier. In this case, some of the training data are allowed to be misclassified to be either on the wrong side of the margin or sometimes even on the wrong side of the hyperplane to avoid overfitting and this will help to have a better classification on the remaining observations. It is therefore the solution of the optimization problem below:

$$maximize\ M\ \beta_0, \beta_1, \beta_2, \dots, \beta_p, \epsilon_1, \dots, \epsilon_n$$

$$subject\ to\ \sum_{j=1}^{p} {\beta_j}^2 = 1$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \varepsilon_i) \tag{3.11}$$

$$\varepsilon_i \geq 0, \qquad \sum_{i=1}^{n} \varepsilon_i \ \leq C$$

Where C represents a non-negative turning parameter. $\epsilon_1, \dots, \epsilon_n$ are slack variables that permit a given observation to be classified wrongly, either on the wrong side of the margin or the hyperplane. And it also tells us the position of the given observation. $\varepsilon_i = 0$, this means that the i^{th} observation is classified correctly, and it is on the correct side of the margin. For the case where $\varepsilon_i > 0$, this means that the i^{th} observation is wrongly classified, and, in this case, it is classified on the wrong side of the margin. If $\varepsilon_i > 1$, then that observation is not only wrongly classified but it is also on the wrong side of the hyperplane. The value of *C* helps to determine the number of observations that can be compromised to be wrongly classified, either on the wrong side of the margin or the hyperplane. If *C* = 0, this is also equivalent to the case where $\epsilon_1 = \cdots = \epsilon_n = 0$, in this case, it can therefore be concluded that all the observations are correctly classified into their right classes. If *C* > 0, in this case it will only allow the maximum of *C* observations to be on the wrong side of the hyperplane. There exist a tradeoff in the bias-variance relation for any chosen value of *C*. if the value of *C* is getting bigger, the classification boundary will move to control itself so that many of the training data points will be correctly classify as much as possible, this means the margin will be bigger and many observations will fall in the region as a support vector since they will have an impact on the position of the hyperplane. In short, there will be a high bias classifier with a low variance. If *C* is small, this means that the margin will be smaller and only few observations will serve as a support vector in determining the position of a hyperplane. In short, there will be a low bias classifier with a high variance (James *et al*, 2013).

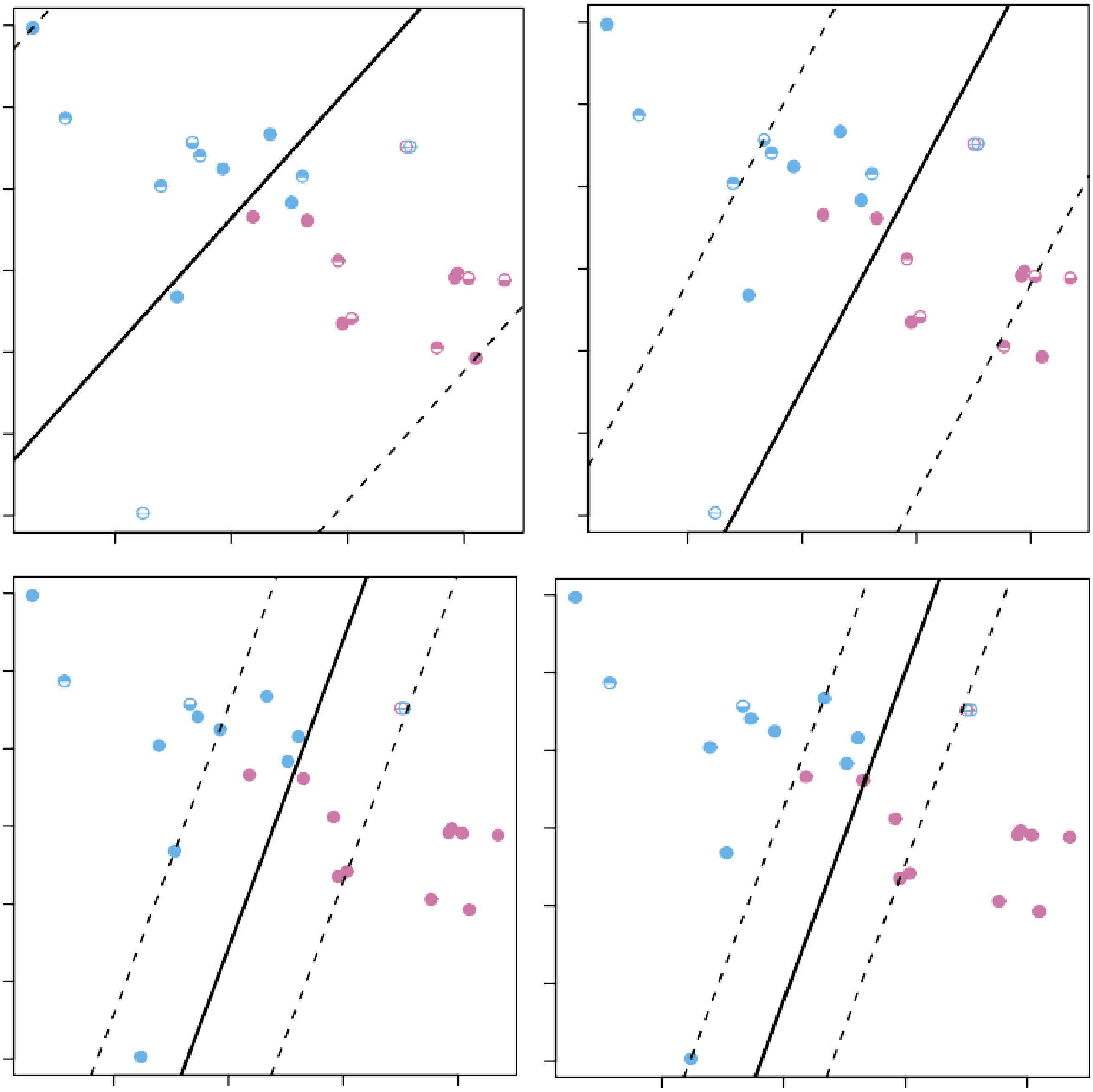


Figure 3.2. Tradeoff between the tolerance for observation to be on the wrong side of the margin C and the size of the margin

In fact, misclassification of some observation in the data becomes unavoidable if the data cannot be separated perfectly by the maximal margin classifier. In such cases, the problem of non-linear separation of the data into their respective classes can be solved by expanding the feature space using quadratic, cubic, and even higher order polynomial functions of the predictors or sometimes the interaction of terms in the function rather than creating a support vector classifier using only p features as in:

$$x_1, x_2, x_3, x_4, x_5, \dots, x_p \tag{3.12}$$

There are numerous methods to expand the feature space, one of the approaches is to enlarge it to $2p$ features as below:

$$x_1, {x_1}^2, x_2, {x_2}^2, \dots, x_p, {x_p}^2 \tag{3.13}$$

In this case, the optimization problem will be:

$$maximize\ M\ \beta_0, \beta_{11}, \beta_{12}, \dots, \beta_{p1}, \beta_{p2}, \epsilon_1, \dots, \epsilon_n$$

$$subject\ to \qquad y_i\left(\beta_0 + \sum_{j=1}^{p} \beta_{j1} x_{ij} + \sum_{j=1}^{p} \beta_{j2} {x_{ij}}^2\right) \geq M(1 - \varepsilon_i) \tag{3.14}$$

$$\sum_{j=1}^{p} \sum_{k=1}^{2} {\beta_{jk}}^2 = 1$$

$$\varepsilon_i \geq 0, \qquad \sum_{i=1}^{n} \varepsilon_i \leq C$$

Sometimes the feature space can be enlarged by including the interaction of terms of the form $x_i x_j$, for $i \neq j$ but a proper selection of the feature space must be made to avoid having a complex computation which can become unmanageable.

3.2.4. The kernel functions

After enlarging the feature space of the non-linear separable case by using specific kernels, this will result in the use of the support vector machine. The main objective is to classify classes which are not linearly separable, and only the inner products of the observations are important in this process and not just the observations themselves alone. “A kernel is a function that quantifies the similarity of two observations”. According to (Max Khun and Kjell Johnson,2013), the kernel trick helps the support vector machine to come up with a smooth flexible decision boundary for the non-linear separable classes but the choice of the parameters of the kernel function and the cost values have a high effect on the solution and a proper selection should be considered to avoid overfitting the training data (James *et al*, 2013).

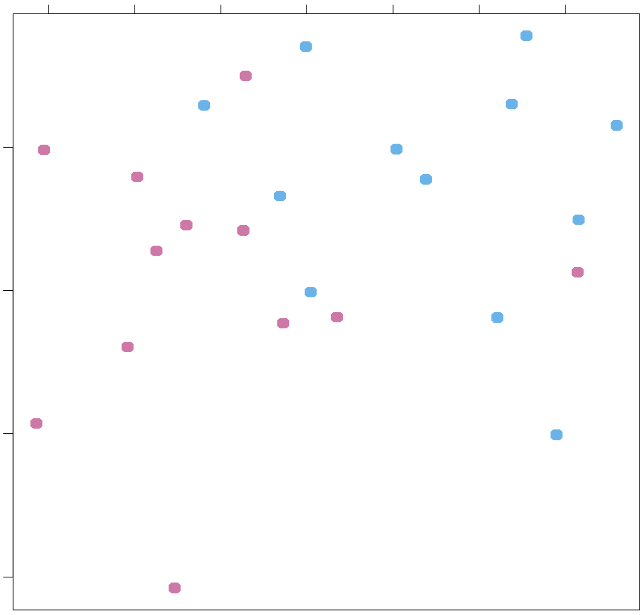


Figure 3.3. Two non-linearly separable classes.

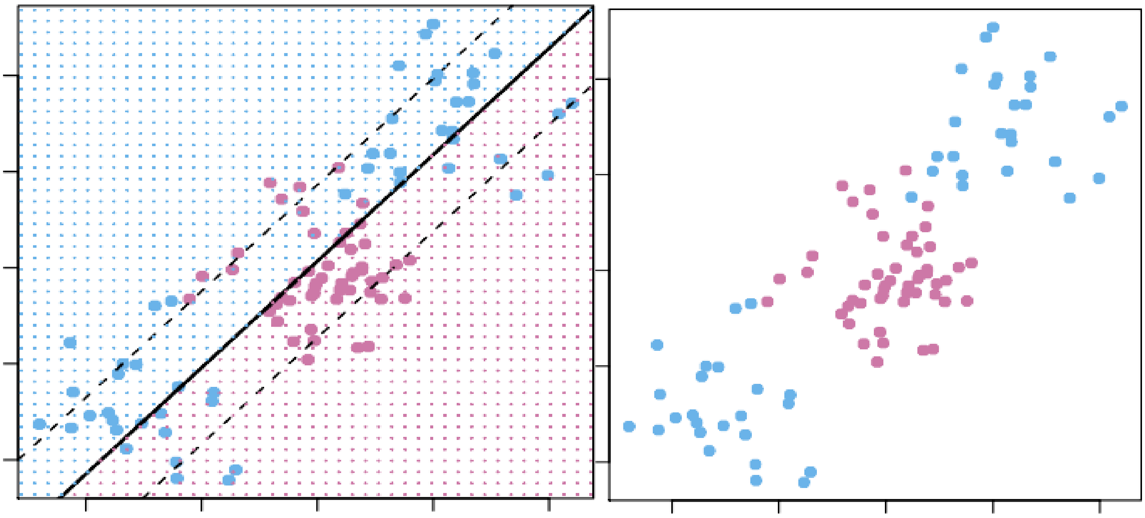


Figure 3.4. Application of the maximal margin classifier on a non-separable case.

Having a large value for the kernel parameters, the model will under fit the data if the cost value is low, and it will over fit the data if the cost value is high. So, there should be a fine tune in between the values of the cost from underfitting and overfitting the data so that it will give a model that will be balance and fit the data properly. There is a relationship between the error of classification and the cost parameter. Hence, the main tool in adjusting and simplifying the complexity of the model is the cost parameter, it provides more flexibility for tuning the model. For that being the reason, it is suggested to fix the error value and tune over the other kernel parameters. It is recommended scaling the predictors before building the SVM model since differences in the predictors scale can affect the model.

The inner product of two observations say $x\ and\ x'$ is given by the following formula:

$$[x_i, x_i'] = \sum_{i=0}^{p} x_i x_i' \tag{3.15}$$

and it can be more generalized as a linear support vector as below:

$$f(x) = \beta_0 + \sum_{i=0}^{p} \alpha_i \, (x_i \, x_i') \tag{3.16}$$

and each α_i corresponds to one training data. At first sight it looks a bit scary and complex but the good thing about the equation is that the value of α_i is zero for all the non-support vectors, so all that should be done is to find the summation of the inner product of the new observation and the support vectors. The generalization of the inner product can be written in the form of a Kernel as:

$$K(x_i, x_i') = \sum_{i=0}^{p} x_i \, x_i' \tag{3.17}$$

and this is called the linear kernel since the features of the support vector classifier is linear.

The linear kernel function can be extended into much more flexible decision boundary like the polynomial and radial kernels. The polynomial kernel has the form as:

$$K(x_i, x_i') = (1 + \sum_{i=0}^{p} x_i \, x_i')^d \tag{3.18}$$

where *d* represent the degree of the polynomial kernel and it must be a number greater than zero. The support vector machine is equivalent to the support vector classifier when d=1. The non-linear functions have the form:

$$f(x) = \beta_0 + \sum_{i=0}^{p} \alpha_i K(x_i \, x_i') \tag{3.19}$$

α_i is non-zero only for the support vectors.

Another method will be the radial kernel and it has the form:

$$K(x_i\, x_{i'}) = \exp(-\gamma \textstyle\sum_{j=0}^{p} (x_{ij} - x_{i'j})^2) \qquad (3.20)$$

where γ is a positive constant. This is a very good and effective method in separating two classes that are not linearly separable. Given a test observation $x^* = (x_1{}^* \ldots x_p{}^*)$, the training data that are very far away from the test observation will have little or no help in classifying the test observation. This is because $\sum_{j=1}^{p} (x_j{}^* - x_{ij})^2$ will be very large and the radial kernel will be very small. Therefore, the closer training values to the test observation will help us classify a test observation (James et al, 2013)

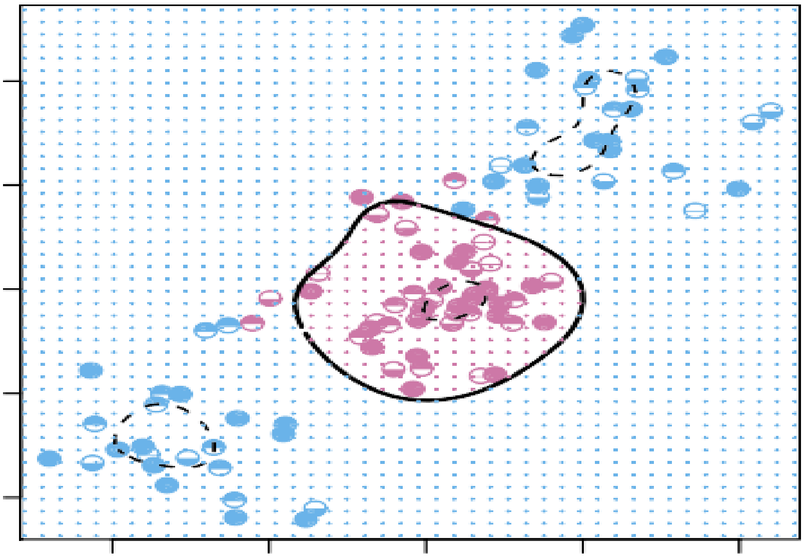


Figure 3.5. Using kernel functions to separate non-linear separable classes.

Given these several kernel functions that can be applied to a support vector machine, the question that need to be answered is: which kernel should be used in the classification process? This actually rest on the problem at hand. The radial basis function has been very effective and accurate in classification especially if the classes cannot be linearly separated, but the linear kernel function will be the better function to use when we are sure that the data can be linearly separated.

3.3. Logistic Regression

3.3.1. Bernoulli and binomial distribution

The binomial distribution is the most widely probability distribution that is used when there are two classes. It has a single parameter, p, which determines the probability of an event or class for a particular binary response that is either success or failure. The binomial distribution is built from the Bernoulli distribution. Suppose that this response concerns whether a student will be admitted into a specific program under certain conditions. There will be varieties of factors that will determine the success of the student. Suppose that there is a probability, *p*, that the student will be admitted. This probability is called the success probability or response probability. If its small, the student may not likely be admitted, while if *p* is close to one, the student is likely to be admitted. Whether the student is or not admitted can be denoted as R and the two possible values of R are 0 and 1. R = 1 corresponds successful admittance and *R* = 0 is a failure to be admitted.

The probability that *R* = 1 is the success probability, *p*, and so $P(R = 1) = p$. The probability of an event not happening (failure) is

$P(R = 0) = 1 - p$
$P(R = r) = \ p^r(1 - p)^{1-r}\ , r = 0{,}1$

This expression defines how the probability of the two events *R* = 0 and *R* = 1 are distributed, and so it expresses the probability distribution of R. this particular distribution is known as the Bernoulli distribution.

The Mean: E (*R*) = *p*
The variance: V (*R*) = *p* (*1*-*p*)

It can be seen that both the mean and variance of the Bernoulli distribution are affected by the success probability *p*, so any factor that changes the success probability will automatically change the mean and variance.

Suppose that a given sequence of *n* binary observation contains y successes and *n*-*y* failures. Assuming that each of the binary response is independent of the others and they have a common probability p of having the attribute of interest. The probability of y successes in n observations is therefore given by:

$$P(Y = y) = nC_y p^y (1 - p)^{n-y} \tag{3.21}$$

For y = 0, 1, …, n. The Radom variable Y is said to have a binomial distribution. It depends on two parameters, the total number of observation n and the success probability p. The p and 1-p reflect the frequencies of the classes in the observed data. With the help of the Maximum Likelihood estimator, it will be able to find the value of p that will produce the largest value for f (p). $Y \sim B(n, p)$.

The Mean: $E(Y) = np$
The variance: $V(Y) = np(1 - p)$

3.3.2. The linear logistic regression model

Regression has become one of the effective methods in statistics and data analysis where the relationship between a dependent and a set of independent (one or more explanatory variables) are established. One of the most important models for the data whose response is categorical, in this case a binary response is Logistic regression. (Agresti, 2013). When a linear logistic model is built to establish the relationship between a binary response variable (0,1) and one or more explanatory variables which are sometimes called predictors, the model is also referred to as a logistic regression. The standard classification method which is based on the probabilistic statistics of the data which is used to predict a binary response from a binary predictor is known as the logistic regression. (Khanna *et al*, 2015.).

One of the main differences between a linear regression and a logistic regression model is that the dependent variable in the later is binary. When a binary response data is represented on a scatter plot, all the points will fall on either of the parallel line, i.e. one of the two possible outcomes, but it does not provide a clear relationship between the response variable and the independent variables. In the linear regression model, the error term which is the deviation of an observation from its conditional mean follows a normal distribution with the mean value zero and it a constant variance across the level of independent variable; while it is a different case for the binary outcome variable, here the expression of the outcome variable is written below:

$$y = p_i + \varepsilon \tag{3.22}$$

If y=1, the error term takes the value $\varepsilon = 1 - p_i$. If y = 0, then the error term takes the value $\varepsilon = -p_i$. From here we can conclude that the error term for a binary outcome variable follows a binomial distribution with mean zero and variance $p_i(1 - p_i)$ given the conditional mean p_i.

It is used in most of the applications in life such as biomedical studies, Genetics, social science, marketing, etc. (Agresti, 2013). Suppose that we have *n* binomial observation of the form $\frac{y_i}{n_i}$ $for\ i = 1,2, \dots, n$ where $E(y_i) = n_i p_i\ and\ p_i$ is the probability**.**

If the probabilities p_i depends on a vector of observed covariates X_i. The linear probability model can be written as follows:

$$p_i = x_i{}'\beta \tag{3.23}$$

Where β is a vector of regression coefficient. The issue in this case is that the left-hand side of the equation is only limited to certain range of values i.e. from zero to one while the other side of the equation can take every real value. The first approach into solving this issue is to transform the probability into odds instead of p_i. The odd of an event is the ratio of the success probability of the event to failure. It can take any positive value depending on the value of the success probability, this makes it have no ceiling restrictions and one end is solved. In such cases, how is the floor restriction removed? Taking the log of the odds, which is also called the logit and it can give any negative value and the formula is written below:

$$Logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) \tag{3.24}$$

The logit transforms probabilities that lies in the range of zero to one to have any real number on the number line. If the probability is half (0.5), the odds are neutral, and the logit value is zero. Probabilities less than half will give negative logit values and probabilities above half will give positive logit values.

The linear logistic model of p_i on the values of the k explanatory variables $X_{1i}, X_{2i}, \dots, X_{ki}$ related with the observation is:

$$Logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \tag{3.25}$$

The log odd of the model can take values from $-\infty$ to ∞.

for a model to fit the data well. These three assumptions below should be met:

1. There should be no correlation between the predictors.
2. The predictors should have a significance on the response binary data.
3. The observations are also uncorrelated.

With some mathematical application and rearrangements on the above formula, we can solve for p_i as below:

$$p_i = \frac{\exp(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots+\beta_k X_{ki})}{1+\exp(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots+\beta_k X_{ki})} \tag{3.26}$$

Due to its simplicity and ability of the Logistic regression to make inferential statements about the model terms, this makes it a very popular model. It is very effective when the target is just for prediction if you have provided quality set of good predictors that will yield a better performance. It predicts the probability that the response value has a value 1 given a specific set of predictor values. That is:

$$P(Y = 1 \,|X = x) = 1 - P(Y = 0 \,|X = x)$$

Suppose that the binomial data are from y_i successes out of n_i trials, i = 1,2, . . ., n are available. Given set of data, in order to fit a linear logistic model, the k + 1 unknown parameters $\beta_0, \beta_1, \dots, \beta_k$ have first to be estimated. In a simple linear logistic model:

$$Logit(p_i) = \alpha + \beta x \tag{3.27}$$

Like in the case of a linear regression setting, the logistic regression has an intercept and a slope parameter. The sign of β determines if p_i is either increasing or decreasing as x

increase. Exponentiating both sides shows that the odds are an exponential function of *x*. LR has the power to handle a non-linear relationship between the response variable and independent variables since it applies a non-linear log transformation of the linear regression.

3.3.3. Odds and odd ratio

The Odd of a success of an event is defined to be the ratio of the success probability to the probability of a failure. Thus, if p is the true success probability of an event, the odd of a success is $\frac{p}{1-p}$. If the observed binary data consist of y successes in n observation, the odd of a success can be estimated by $\frac{y}{n-y}$.

When two sets of binary data are to be compared, a relative measure of the odds of a success in the first set relative to that in the second set is the Odds Ratio.

Suppose that p_1 and p_2 are the success probabilities in the first and second set respectively, so that the odds of success in the ith set is $\frac{p_i}{1-p_i}$ for $i = 1,2$. The odd ratio is:

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

When the odds of success in each of the two binary data sets are identical, then the odd ratio in this case is equal to one. When the values of odd ratio are less than one, this suggest that the odd of success are less in the first set of data than in the second set of data, while an odd ratio greater than one indicates that the odd of success are greater in the first set of data.

“The odd ratio can also be used to determine whether a particular exposure is a risk factor for a particular outcome and to compare the magnitude of various risk factors for that outcome. Odd ratio equal to 1 indicates exposure does not affect odd of outcome. Odd ratio greater than 1 indicates exposure associated with higher odds of outcome. Odd ratio less than 1 indicates exposure associated with lower odds of outcome.” (Park, H., 2013). To understand the concept better, the example below will elaborate more. For example, “if the variable smoking is coded as (0=no smoking) and (1=smoking) and the odd ratio for this variable is 3.2. then, the odds for a positive outcome in smoking cases are 3.2 times higher than in non-smoking” (Park, H., 2013).

3.3.4. The confusion matrix

This is one of the techniques that is used to measure the accuracy and the performance of the algorithm which is based on the accuracy, recall (sensitivity) and precision. In the binary classification case, the confusion matrix is illustrated in Table 3.1:

Table 3.1. The confusion Matrix

CORRECT CLASSIFICATION	PREDICTED CLASSIFICATION	
	Positive (+)	Negative (-)
Positive (+)	True Positive TP (+, +)	False Negative FN (+, -)
Negative (-)	False Positive FP (-, +)	True Negative TN (-, -)

From the table above, the two correct classification decisions are TP and TN. An observation is classified as a TP if it is a positive value and classified as positive. An observation is labelled a TN if it is negative and is classified as negative. FP and FN are both occasions where an observation is misclassified i.e. when a positive observation is classified as negative it is called a FN, when a negative observation is classified as positive it is called a FP. Most of the data should fall in the region of TP and TN if the logistic regression model has a good fit. We can also know that the model has a good fit when the sensitivity and specificity values are high.

Different measures of performance such as accuracy, recall and precision can be calculated from the confusion matrix table as below:

Accuracy: The accuracy gives the performance measure of instances in which the observations are correctly classified. It tells us how accurate or close is the predicted value to the actual value. i.e. it gives the general effectiveness of the classifier in classifying instances correctly. The formula is as below:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \qquad (3.28)$$

The misclassification rate of the model which is sometimes called the error rate can be written as "$1 - Accuracy$".

Specificity: Given an observa.tion that is truly negative, the conditional probability that it is predicted as negative is known as specificity. i.e. it measures how effectively the negative instances are classified correctly. Its formula is written below:

$$Specificity = \frac{TN}{FP+TN} \tag{3.29}$$

Presicion: This gives the performance of instances in which the true positive observations are classified positive. i.e. the true positive observations predicted by the classifier divided by the number of observations predicted positive by the classifier. The formula is as below:

$$Precision = \frac{TP}{TP+FP} \tag{3.30}$$

Recall (Sensitivity): This gives the performance of the instances of the positive observations that are classified correctly. In other words, given an observation that is truly positive, the conditional probability that it is predicted as positive is known as sensitivity. i.e. the true positive observations predicted by the classifier divided by the amount of positive observations in the data. The formula is given below:

$$Recall\ (Sensitivity) = \frac{TP}{TP+FN} \tag{3.31}$$

This help us measure how many true samples are predicted from all the samples.

F1 Score: This is a performance measure that combines both the precision and recall (sensitivity) and tries to balance between them. The formula is given below:

$$F1\ Score = \frac{2*Precision*Recall}{Precision+Recall} \tag{3.32}$$

It tells how good a classifier is considering both precision and recall.

Nearly almost every field in life such as: medical diagnosis, financial forecasting, credit card fraud detection, etc. uses classification approach as a data mining technique to correctly classify the data used in the respective fields according to the features of the items with respect to the predefined set of classes. Generally, classification approaches use the objects that are already associated with known class labels and they are called the training set, the classification algorithm learn on the training data set to build a model which is later used to classify new objects into their right classes.

3.3.5. Goodness of fit of linear logistic model

Given a sample of *n* independent pair of observations (x_i, y_i) where y_i is the binary response variable and x_i are the independent variables. To fit a logistic regression model, the first approach is to find the estimates of the unknown parameters; for the purpose of simplicity, say β_0 and β_1. Least square is the most common method used in the linear regression for estimating the unknown parameters in the model, the best estimates that dives the smallest value of the sum of squared deviation of the observed values *y* from the predicted values are chosen. Unfortunately, the estimators do not have the same properties when the method is applied to a model with a binary outcome variable. For that been the case, the maximum likelihood estimate method will be applied in logistic regression to give the estimate of the unknown parameters.

This method gives estimate values for the unknown parameters that will maximize the probability of obtaining the observed set of data. The first step in constructing this method is to create the likelihood function which is just stated below:

$$l(\beta) = \prod_{i=1}^{n} p_i{}^{y_i}(1 - p_i)^{1-y_i} \tag{3.33}$$

Due to complexity of the equation above, it will be easier to work with its log function and that equation is written below:

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n}\{y_i \ln(p_i) + (1 - y_i)\ln(1 - p_i)\} \tag{3.34}$$

Now differentiate $L(\beta)$ with respect to $\beta_0\ and\ \beta_1$ and set them to zero to give the estimates of the unknown parameters that maximizes $L(\beta)$.

The significance of the estimated parameters is now tested based on the log likelihood function above by comparing the observed values of the response variable to the predicted values from the model with or without the variable in the model. This can be done using the statistics below:

$$G = -2\left[\frac{likelihood\ without\ the\ variable}{likelihood\ with\ the\ variable}\right] \tag{3.35}$$

To test the hypothesis on whether the parameter is equal to zero or not, compare the *G* statistic above with $\chi^2(\alpha, 1)$. If $G > \chi^2(\alpha, 1)$, then the parameter is said to be insignificant and the model is better off when it is removed. Among the several approaches that could be used to build a best fitting model are the forward and backward variable selection. In the forward selection approach, this method looks at each explanatory variable separately and selects the first single explanatory variable that has the highest significant to the model on its own. This process is repeated on the remaining explanatory variables and the best explanatory variable among the rest is selected and added to the model. This continues until the none of the remaining explanatory variables can add a significant to the model. For the backward selection, it includes all the explanatory variables into the model, and choose to delete the variable from the model that when removed will cause the least change in the overall fit of the model. The process is repeated to remove the next insignificant variable in the model. The process continues until all the remaining explanatory variables in the model are significant.

After fitting a model to a set of data. If the agreement between the observations and the corresponding fitted value is good, the model may be acceptable at this stage. If the value is bad, the current form of the model will certainly not be acceptable as the correct model, and the model will need to be revised again to give it a better fitted value. This is referred to as goodness of fit.

When the unknown parameters are set to equal to their maximum likelihood estimates, this is maximized likelihood under the current model. If the fitted values coincide with the actual observations, that is, a model that fits the data perfectly. Such a model will have the same number of unknown parameters as there are observations. This model is termed the full or saturated model.

To compare the current and full model, the statistics D therefore measures the extent to which the current model deviates from the full model and is termed the deviance. It has the same importance as the residual sum of squares has in linear regression. Large values of D indicate that the current model is a poor one. Small values of D indicate that the current model is a good one. The equation for the deviance is as below:

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{p_i}{y_i}\right) + (1 - y_i)\ln\left(\frac{1-p_i}{1-y_i}\right)\right] \tag{3.36}$$

3.3.6. Comparing linear logistic model

When a new model comprises of terms that are additional to already existing term in the old other, the two models are said to be nested. The difference in the deviance of two nested models measure the amount to which this additional term in the new model improves the fit of the old model to the observed response variable.

For example :

$Model\ (1): logit(p) = \beta_0$
$Model\ (2): logit(p) = \beta_0 + \ \beta_1 X_1$
$Model\ (3): logit(p) = \beta_0 + \ \beta_2 X_2$
$Model\ (4): logit(p) = \beta_0 + \ \beta_1 X_1 + \beta_2 X_2$

The difference in deviance between Model (4) and Model (2) and between Model (3) and Model (1) both measures the effect of including X_2 in the model. Between Model (4) and Model (2), the effect of X_2 is adjusted for X_1 since X_2 is being added to a model that already include X_1. Between Model (3) and Model (1), the effect of X_2 is unadjusted for X_1. The extent to which the deviances, on including X_2 in the model with or without X_1, depends on the extent to which X_1 and X_2 are associated. If X_1 and X_2 are acting independently, the difference in deviance will be quite similar. Whereas if X_1 and X_2 are highly correlated, they will be very different.

In linear modelling, when the effects of two factors can be estimated independently, they are said to be orthogonal. Similarly, if two explanatory variates have zero correlation, they can be termed orthogonal.

4. OUTLIER DETECTION METHOD

4.1. Definition of Outliers

When a data is available, the primary step into mining the data to come up with some useful information or knowledge about the data is to check the availability of an outliers and to solve them since their presence in a data set can usually give a false information about the general information about the data which results to the misspecification of the data, giving biased parameter estimation and incorrect results. “An outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism.” (Hawkins, 1980). Generally, an observation is said to be an outlier when it is totally different from all the remaining observations in the data set (the average data set in general) and they are usually at the extremes of the data. Generally, the statistical problem to the researchers is to identify the outlying observations in the data so that they can carry out their statistical analysis on the remaining set of the stable sample data. Even though an outlier can be an error in the data, some of the mechanisms that generated the outliers and the outliers themselves gives them a particular interest to have a detail analysis to them as they sometimes carry some useful information about the data at large. There are two basic methods which can cause an outlier(s) in a data set:

It can occur due to bad data entry, that is if some of the data values are wrongly entered (coded) and sometimes due to the fact that the experiment was wrongly conducted. In either case, since it is certain that the outlier was a mistake in the data then it is better to delete the outliers or rectify the mistakes in the experiment so that the data will only contain correct observations. However, sometimes it cannot be determined if an outlier is a coding error or due to wrong experiment. In this case it is not advisable to just delete the outlier(s) because they might be showing some interesting pattern or event in the data or it might be a random variation in the data, and it can be caused by heavy tail distribution.

In most of the data mining process, outlier detections are usually determined by distance measures, clustering and partial methods. Given the target distribution of the data F which follows the assumption of normality, one of the methods of identifying the outliers in a data is to assume that the observations that are considered to be outliers to have a different distribution compared to the target distribution and they are found in a supposed outlier region. For any confidence coefficient α, 0 < α < 1, the α-outlier region of the $N(\mu, \sigma^2)$ is defined by:

$$out(\alpha, \mu, \sigma^2) = \{x : |x - \mu| > \ z_{1-\frac{\alpha}{2}} \sigma\} \tag{4.1}$$

Where $z_q$is the q quantile of the $N(0,1)$. A number x is an α-outlier with respect to F if $x \ \in out(\alpha, \mu, \sigma^2)$ (Davies and Gather, 1993).

In the process of identifying all the outliers in our data, there are two approaches to handle this. First, it to identify all the outliers in the data at one step and it is called a *single step procedure* while there can also be repeated processes to identify outliers where one observation is tested and identified as an outlier or not per step and this process is called the *sequential procedure.* Generally, the mean of the sample and sample standard deviation gives the estimates of the population mean and standard deviation respectively and they also give an estimate on hoe the shape of the data looks and also its location, but these values are very sensitive to the presence of an outlier which can totally give a false estimate for the population. Therefore, the multiple-comparison correction is used in the process where a multiple statistical test is being performed simultaneously. The α-value gives the outlier region for a single observation in a single comparison test. Unfortunately, the α-value will be inappropriate for the entire set of *n* multiple comparison test, so it therefore must be adjusted, in this case to be reduced to give a new value for the outlier region for each comparison to be α/*n,* and this approach is called the Bonferroni's correction. In a sequential procedure, an inward and an outward method is applied. In an inward method or a forward selection method help us identify an outlier. In every step on this process, the observation which is the most extreme value i.e. the one which is most different from the other remaining observations in the data is tested for being an outlier, that observation is deleted if it is an outlier and the same procedure is repeated until at a point where the observation tested turns out to be a non-outlying observation. In an outward procedure, "The sample of observations is first reduced to smaller sample while the removed sample are kept in a reservoir. The statistics ae calculated based on the reduced sample and then the removed observation in the reservoir are tested in reverse order to indicate whether they are outliers" (Ben-Gal, 2005). An observation is deleted from the reservoir if it is confirmed it is an outlier. An observation is transferred from the reservoir to the reduced sample if it is determined to be a non-outlying observation and the statistics are recalculated and the procedure continues with a new observation until there is no more observation left in the reservoir.

4.2. Masking and Swamping Effects

In cases where the exact number of outliers is not known, then there exist two possible scenarios, that is either the masking effect or swapping effect has occurred.

The masking effect occurred in cases where an outlier is detected due to the presence of another close outlying observations. Also, the adjacent observation turns to be an outlier after the deletion of the first outlier which means it can only be considered as on outlier by itself. “masking occurs when a cluster of outlying observations skews the mean and the covariance estimates towards it, and the resulting distance of the outlying point from the mean is small” (Ben-Gal, 2005).

The swamping effect occurred in cases where an observation that is considered clean to be incorrectly identified as an outlier because of the presence of another clean subset. Here, the adjacent observation turns to be a non-outlying observation after the deletion of the first observation considered to be an outlier which means the adjacent observation can only be considered an outlier in the presence of the first outlier. “Swamping occurs when a group of outlying instances skews the mean and the covariance estimate toward it and away from other non-outlying instances, and the resulting distance from these instances to the mean is large, making them look like outliers” (Ben-Gal, 2005).

4.3. Multivariate Outlier Detection

When a data consists sets of measurements on a number of individuals such as the height, weight drawn from a population is said to be concerned with a multivariate statistical analysis and the measurements made on a single individual can be put together into a column vector. The means and variances of separate measurements have a corresponding relevance and the dependence between two variables may involve covariance between them, i.e. the average products of their deviations from their respective means. (Anderson T.W, 1958) Observations that diverge from the usual assumption or from the pattern suggested by the majority of the data and are not consistent to the general correlational trend of the data set are said to be a multivariate outlier. There are various methods in identifying outliers in a multivariate data which mostly rely on the mean and covariance matrix and correlation but the problem we face in this case is that those measurements are affected by outliers. Outlier detection is usually the initial process of cleaning the dataset. Multivariate outlier detection is a very important method to help detect the presence of outliers in the data since most of the data contain outliers and their presence can give a false and misleading results which leads to a wrong conclusion. Because the results will lean more to the outlier(s) which does not necessarily gives the best description of the general data.

Suppose that X = [2, 3, 5, 9, 7, 8, 25, 96] is a dataset. From this data set it is seen that the only authentic outlier present is the value (96). But what happened when the first outlier is deleted from the data and examine it again? Now it can be seen that the value (25) is the only authentic outlier in the dataset. Why was this not shown in the first case? This was because it was masked by the presence on the value (96) and it can only be identified as an outlier after the deletion of the value (96). We say the value (25) is masked by the value (96) and is referred to as the masking effect. That is in general, one outlier can be too extreme that it will mask the adjacent outlier. “Masking occurs when a cluster of outlying observations skews the mean and the covariance estimates towards it, and the resulting distance of the outlying point from the mean is small” (Ben-Gal, 2005).

Most of the outlier detection methods uses the distance measure which rely on the sample mean and variance, but those values are heavily affected by the presence of an outlier in the data set. The multivariate outlier detection methods that the paper focus on are: The Mahalanobis Distance and the Robust Mahalanobis Distance. Since the focus of the paper is on a binary data, we want to correctly classify a given data into the right class. The closer the given observation is to the mean of a class, the higher possibility of it belonging to that class, and the further away an observation is from the mean of a class, the more likely that the given observation can be an outlier and should not be classified as belonging to that set. This method can simply be defined as the distance between two objects and is known as *The Eucledian Distance.* But the problem in this case is that the unit of the variables can affect the distance of measurement and if the variables are correlated among themselves also will mean that more calculation than necessary have been done since those observations are in the same direction and have similar impact.

4.4. The Mahalanobis Distance and The Robust Mahalanobis Distance

The Mahalanobis distance is a classical way of identifying Multivariate outliers and it serves as the modified Eucledian distance since it is unitless and it takes the covariance matrix into account. Based on a chi square distribution and by considering the shape of the cloud data, the distance from the center of the cloud to the point helps to determine if an observation is an outlier. However, the spread of the data i.e. the variance (standard deviation) from the mean is also a very important feature to decide if an observation is an outlier. If the given observation falls within a region of one standard deviation from the mean, this can give us more confidence that it belongs to that particular class but if it falls beyond the region of three standard deviation from the mean then we can be somehow suspicious that it might be an outlier. "Mahalanobis distance identifies observations that lie far away from the center of the data cloud, giving less weight to the variables with large variances or to groups of highly correlated variables" (Joliffe). Their disadvantage is they do suffer from the masking effect and they do not resist well to outliers. The modified classical method is known as the robust Mahalanobis distance, "Rousseeuw and Leroy (1987) recommend using distance based on robust estimators of multivariate location and scatter (μ_R, V_R) to avoid masking effect" and they do resist to outliers better than the Mahalanobis distance in a multivariate data set.

The Mahalanobis distance for each multivariate data point $i, i = 1, \dots, n$ is denoted by M_i and is given as below:

$$M_i = \left(\sum_{i=1}^{n}(x_i - m)^T V_n{}^{-1}(x_i - m)\right)^{1/2}$$
$$V_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)(x_i - m)^T \tag{4.2}$$

$x_i =$ Vector of data of independent variables
$m =$ vector of mean values of independent variables
$T =$ it indicates that the vector should be transposed
$V_n =$ Sample covariance matrix

The Robust Mahalanobis Distance can be written as below:

$$RDM(x_i) = \sqrt{(x_i - \mu_R){V_R}^{-1}(x_i - \mu_R)^t} \tag{4.3}$$

μ_R, is our first moment vector and V_R, is the robust covariance matrix. Accordingly, those observations with a large Mahalanobis distance are indicated as outliers. All the points that fall beyond the point $\sqrt{{\chi^2}_{p,0.975}}$ are determined to be outliers. “Robust estimation finds a ‘robust fit’ which is similar to the fit we would have found without the outliers” (Hubert, Rousseeuw and Van Aelst, 2005).

4.5. Inward and Outward Outlier Detection Methods

A block test is a simple approach method in detecting outliers in a data where the number of outliers is specified before applying the method to detect the available outliers. In this case, at the first step, either all the outliers are identified, or none is identified. It can also suffer a masking effect when too many points are included in the block but can also suffers a swamping effect when just few of the points are included in the block. It is a good method when the number of outliers is well specified. But when a proper specification on the number of outliers is not possible, then either the inward or outward test of detecting outliers can be applied.

In the inward test, the most outlying observation is tested to be an outlier or not, if it is rejected (that is it is an outlier), it is removed from the data and test for the next most outlying observation in the remaining data. The process is repeated on until where the first non-outlying observation is detected (that is we fail to reject) and that is where the process is stopped. The numbers of the estimated outliers are equal to the number of rejections. It’s *type 1* error is equal to the first rejection of the outlying observation since that is the only time the null hypothesis can be rejected. But there is likelihood that this test can suffer both masking and swamping effect.

In the outward test, first specify the maximum number of outliers, *r*, to be tested. the smallest among the outliers (the rth observation) is tested to be an outlier, if it is rejected then the next smallest observation in the *r*-1 observation is tested. This process is repeated until at the point where an observation if fail to be rejected and all the remaining observations specified to be outliers are now outliers. In the case where all the observations specified to be outliers are fail to be rejected, then we can conclude that the data is free from outliers. This test provides a less chance of having both the masking and swamping effect, thus that makes it preferable and more efficient than the inward test. However, it is difficult to control the *type* 1 error in this method.

Generally, the test statistics provides a comparison between the outlyingness (the extremeness of the observation compared to the remaining observations in the data) of the suspected outliers in the data to the measure of the spread within the remaining set of the data. Spacings is one of the measures used in some test statistics while others are based on the sums of observation sizes. Some of the methods for outlier detection are discussed below:

The sum-sum (SS) test statistics is a popular method used to detect the *r* upper outliers in the dataset. The test suffers from swamping effect as the number of *r* increases. The mathematical notation is as below:

$$T_r{}^{SS} = \frac{\sum_{i=1}^{r} x_{(i)}}{\sum_{i=1}^{n} x_{(i)}} \tag{4.4}$$

The numerator of the equation measures the magnitude of the *r* upper outliers rather than the differences. In short, it does not compare the distance between two observations. It is very powerful and effective in cases where the outliers are clustered. If all the outliers are not identified in the numerator, the remaining number of outliers included in the denominator might cause a masking effect in the denominator.

To give robustness in the denominator, the sum-robust-sum (SRS) can be applied and the notation is below:

$$T_r{}^{SS} = \frac{\sum_{i=1}^{r} x_{(i)}}{\sum_{i=m+1}^{n} x_{(i)}}\ ,\ \ m \geq 1 \tag{4.5}$$

Where *m* is the number of identified outliers in the dataset before applying the test. “Here, the choice of *m* is a tradeoff between sample size (power) and sample purity (masking avoidance)”.

The max-sum (MS) is for the jth rank of outliers identified. The index *j* makes it possible to use the outward procedure. The notation is written below:

$$T_j{}^{MS} = \frac{x_{(j)}}{\sum_{i=j}^{n} x_{(i)}} \tag{4.6}$$

In this method the numerator is the maximum of outliers instead of the sum in the SS and this will help us avoid the swamping effect but in cases where the outliers are clustered, the SS/SRS is more powerful. To avoid masking effect in the denominator, the max-robust-sum (MRS) is applied and is below:

$$T_{jm}{}^{MRS} = \frac{x_{(j)}}{\sum_{i=m+1}^{n} x_{(i)}},\ m \geq 1 \tag{4.7}$$

The *Dixon* statistics is also used for detecting the *r* upper outliers in a dataset and it is often used as a substitute to the SS, its notation is written below:

$$T_r{}^{D} = \frac{x_1}{x_{(r+1)}} \tag{4.8}$$

The advantage it has over the other methods is that it does not suffer much swamping or masking effect.

Another method for detecting the *r* upper outliers is the Dragon King (DK) statistics, it does not treat the extremeness of each point equally since it sums the weighted spacings instead of the absolutes. This makes it very vulnerable to both swamping and masking effect and is less powerful to case where the outliers are clustered. The formula is as below:

$$T_r{}^{DK} = \frac{\sum_{i=1}^{r} z_i}{\sum_{i=r+1}^{n} z_i} \sim F_{2r,2(n-r)} \tag{4.9}$$

Where $z_i = i\left(x_{(i)} - x_{(i+1)}\right), i = 1, \ldots, n-1, z_n = nx_{(n)}$ and it has an F distribution.

4.6. Pareto and Dragon King

Extreme values of earth such as landslides, earthquakes, volcanic eruptions, biological, financial crises and such similar extreme events often forms outliers to heavy tails of empirical frequency distributions which are mostly approximated by the stretched exponential, log-normal or power functions. Most of the engineering applications rely to an Exponential df, $E\, iid\sim Exp(\alpha)$ for outlier detection. When the concept changed from reliability to risk, it is transformed with the variable X such that $X = \mu \exp\{E\}\ iid \sim Pareto\ (\alpha, \mu)$ has the heavy tail pareto df:

$$F(x) = 1 - \left(\frac{x}{\mu}\right)^{-\alpha},\ x \geq \mu, \alpha > 0 \tag{4.10}$$

Which is generally used for modelling extremes in both natural and social sciences. The logarithm of the Pareto tail is the Exponential tail.

The scale invariant of the pareto df makes it unique such that it suggests events of all sizes including the extremely large ones are all generated by a single mechanism operating at different scales, and this gives it the power to represent a different range of event sizes. Due to its scale invariant, this makes it hard to predict the extreme events since they are only different from all the other small events based on only their resultant size. “according to the approximate scale invariance of the Gutenberg-Richter law, large earthquakes are just earthquakes that started small… and did not stop growing.”

One of the most noticeable properties of the natural and social sciences is the occurrence of rare large abnormal events which often dominate their organization and leads to a very huge loss or damage and this is usually quantified by the heavy-tailed distribution of the event sizes. With a series of studies having enough strong evidence that the extreme events (outliers) beyond the Pareto sample and this gives the introduction of the Dragon King (DK). “the concept of the Dragon king is to refer to the existence of transient organization into extreme events that are statistically and mechanistically different from the rest of their smaller siblings. This realization opens the way for a systematic theory of predictability of catastrophes” (Sornette, 2009). The name DK gives its own definition of the extreme event that the event is both extremely large (a king) and born of unique origins (dragon) compared to its peers. For a basic demonstration, the term “king” is like the king of a country whose wealth is way more than the remaining inhabitants in the country. The distribution of the wealth of the population excluding the king follows the well-known power law Pareto distribution while the wealth of the king is an “outlier” compared to the remaining population wealth. The term “dragon” is describing an animal but with a higher supernatural power compared to the rest of the other animals. Because of some strong external forces or extreme parameter excursions of the governing mechanism, this can cause the unusual extreme events which can be named Dragon King, (Sornette, 2009).

Due to their special status and significance, DK can be of a very special interest and they may be subjected to deeper investigation of their origin, understanding their generating mechanism, and developing forecasting methods and control. For this being the reason, there should be a proper selection of the DK by a special statistical technique from the sample for further study since it is always not easy and simple to collect a good number of DK for reliable statistical analysis.

The frequent occurrence of financial crashes in the financial market does not only have its negative impact on the market participants alone but the whole economy at large. It will be a big gain for the private companies if the central bank can come up with better and efficient financial policies regarding these risks so that they can be predicted so and give warning to the population about the future economic crises, as this will help them to prepare well for the crises ahead. In the financial market, “a drawdown is the total cumulative return of a negative run in price over time with some specified tolerance for small positive changes along the way. A draw up is its positive counterpart.” (Spencer and Sornette, 2015). This is an important measure of the risk in the financial market as it captures the short time dependence of price change to time. DK comes into effect when the market dynamics is becoming more complicated than ever before such that future trades are not triggered by news but only by the previous trades which makes the financial market essentially self-referential in these periods. From these observations, some of the outliers diagnosed can be classified as DK drawdowns.

The rank-ordering plots is known to be one of the most commonly used method in identifying DKs in a sample. In this method, events are arranged in a decreasing magnitude (from highest to lowest) and plotted versus rank-order on semi-log or log-log scale. The events with the highest rank and that deviates clearly well away from the model fitted to most of the extreme values, usually a power model, are identified as DKs. (Riva, Neuman and Guadagnini, 2013).

4.7. Detection of Outliers Using Clustering

In some instances, there is not just a single observation that is completely different from all other observation in the data, i.e. an outlier, but instead there are several observations which are regarded as outliers in a dataset and are usually scattered outside the normal observations in the data and sometimes these outliers form a cluster. In such cases, the technique that can be used to detect such outliers is called the clustering technique. There are several clustering techniques that are used but the focus of this paper will be on the Partitioning Around Medoid (PAM) method. This method usually works well and efficient in cases where the sample size is small, but it is very costly in cases when we have a large dataset. For this being the reason, another technique called Clustering Large Applications (CLARA) was developed (Kaufman and Rousseuw,1990) where several samples of dataset are generated and them PAM is applied to the samples. Its robustness in the presence of an outlier makes it a reliable method and it does not depend on the order in which instances are examined.

To begin with, to construct k partitions, the PAM method creates a first partition by forming groups from where to start with. In this case, each of the clusters will have a representative observation which is called the medoid, and it should generally be the most centrally located observation within the cluster and on average its dissimilarity to all the remaining observations within the cluster to be minimal. After this process, all the remaining observations in the dataset are then assigned to the nearest medoid. It then uses iteration technique that will help to adjust to give a proper partitioning by moving observations from one group to another and this will provide a better partition. The general rule for a good partitioning at the end is to have observations that are close to one another or share a similar criterion in the same cluster, on the other hand, observations that are dissimilar or do not share similar criteria to be far apart as much as possible, i.e. to be in different clusters.

After establishing the first initial k clusters, the separation of the cluster C is defined as the smallest dissimilarity between two objects, in which one of the objects belong to the cluster C and the other does not. All instances that do belong to a specific cluster are considered to be outliers if the separation of an outlier is huge enough. In order to detect the clustered outliers, one must vary the number k of clusters until clusters of small size are obtained that have a large separation from other clusters.

4.8. Probabilistic outputs for support vector machines

Generally, support vector machines do not evaluate the probabilities of classifying an observation in the right class but instead they provide a classifier for us that will help us classify an observation in the right class. So, what do we do if we want to find the probability that an observation is correctly classified given the attributes (Inputs)? Here, Bayes theorem is to be used to calculate P (Class | Inputs). And this called the Posterior probability. “Posterior probabilities are important when a classifier is making a small part of an overall decision, and the classification outputs must be combined for the overall decision”. (John C Platt, 1999). Extracting probabilities from SVM outputs is important for classification post processing. The SVM is trained to minimize the error function which will also minimize the rate of the misclassification error. Minimizing the error function will also provide a sparse machine where only a subset of kernels is used in the final machine.

$$C \sum_i (1 - y_i f_i) + \frac{1}{2} ||h|| F \tag{4.11}$$

According to Plat (1999), Wahba used the logistic link function to provide the probabilistic output from a kernel machine. The logistic function is as below:

$$\mathrm{P\ (Class\ |\ Input)} = \mathrm{P\ (y = 1\ |\ X)} = \mathrm{P(X)} = \frac{1}{1+\exp(-f(X))} \tag{4.12}$$

And then he proposed minimizing a negative log multinomial likelihood plus a term that penalizes the norm in a Reproducing Kernel Hilbert Space (RKHS).

$$-\frac{1}{m} \sum_i \left(\frac{y_i+1}{2} \log(p_i) + \frac{1-y_i}{2} \log(1-p_i) \right) + \lambda ||h||^2 F \tag{4.13}$$

Where $p_i = p(x_i)$. The result $p(x)$ of such machine will produce the posterior probability but unfortunately it will not maintain its sparseness unless and until it is modified.

A sigmoid function is a mathematical function having the characteristics “S” shaped curve and is sometimes referred to as the special case of the logistic regression. It is a bounded differentiable real function that is defined for all real input values and has a positive derivative at each point with a return value monotonically increasing most often from 0 to 1.

The modification of the SVM (SVM plus Sigmoid) provides the posterior probability and at the same time it also maintains its sparseness. Vapnik proposed a method of producing the probability of SVM by dividing the feature space into two in which one-dimension falls into a direction orthogonal to the separating hyperplane parameterized by t and the remaining dimensions are parameterized by the vector u. His function is as below:

$$P(y = 1 \,|t, u) = a_0(u) + \sum_{n=1}^{N} a_n(u)\cos(nt) \tag{4.14}$$

In this method the initial results were good, but its limitations are that the sum of the cosine function does not fall between 0 and 1, and it is not forced to be monotonic in *f*. Also, for every new evaluation of the system is requires a new linear system. Another method was proposed by Gaussian in which a single tie variance was estimated for both class conditional densities $P(f \,|y = 1)$ and $P(f \,|y = -1)$ which also determines the slope of the posterior probability $P(y = 1 \,|f)$ which is thus a sigmoid. This method was first proposed by Hastie and Tibshirani who have adjusted the bias of the sigmoid so that $P(y = 1 \,|f) = 0.5$ occurs at 0. The Bayes rule used is as below:

$$P(y = 1 \,|f) = \frac{P(f|y=1)\,P(y=1)}{\sum_{i=-1,1} P(f|y=i)\,P(y=i)} \tag{4.15}$$

where $P(y = i)$ are the prior probabilities that can be calculated from the training data. The model was simplified as below:

$$P(y = 1 \,|f) = \frac{1}{1+\exp(af^2+bf+c)} \tag{4.16}$$

Unfortunately, the function is non-monotonic, and it does not follow the strong monotonic prior and sometimes the assumption of Gaussian class-conditional densities is violated.

In fitting the sigmoid after the SVM, they used a parametric model to fit the posterior probability $P(y = 1 \,|f)$ directly instead of estimating the class conditional densities $P(f \,|y)$. The parametric form of the sigmoid using the Bayes rule on two exponentials suggests:

$$P(y = 1 \,|f) = \frac{1}{1+\exp(Af+B)} \tag{4.17}$$

This sigmoid fit well and is similar to saying that the output of the SVM is relative to the log odds of a positive example and it has two parameters trained discriminatively instead of one parameter trained from a tied variance. The monotonicity is always assured as long as A < 0 and the model is still appropriate even if the class conditional densities are close to Gaussian.

The error rate of classification was compared between the raw SVM, an SVM plus Sigmoid, and a regularized likelihood kernel method. It was shown that the addition of a sigmoid sometimes improves the error rate of the raw SVM, and Bayes optimal is not essentially a zero threshold. Also, the SVM plus sigmoid maintains its sparseness and the posterior probability is almost as good as the regularized likelihood kernel method.

The Probability of Class PoC is the proposed method for this thesis, this method has an advantage over other classification methods due to its ability to give us the probability of classification of an observation that we already knew belongs to a particular class. When the probability of class for the clean data are calculated, conducting bootstrap resampling on the clean data and finding the average mean of all the means will give a better threshold for determining if an observation is an outlier or not. The probability of class of observation which are more than the threshold probability will give us more confident that they belong to that class, and if they are less than the threshold probability then they can be determined as outliers. Methods such as the SVM using the Mahalanobis distance to detect the outliers will fall short to the PoC they will predict the belonging of an observation to a particular class while the PoC will give you the probability of an observation after knowing that it belongs to that class.

5. SIMULATION STUDY AND DISCUSSION

In this chapter, we conducted a simulation study to compare the accuracy of the support vector machine with and without outliers in the data and also the performance on effectiveness on detecting outliers in a data set by using the two outlier detection methods for this thesis; Support Vector Machine with the Mahalanobis distance and the Support Vector Machine with the Probability of Class (PoC) in which the posterior probabilities of observations are calculated, i.e. The probability of an observation belonging to a particular class after knowing that it is from that class. We compare the performance of the two methods of classification with the presence of outliers within the data.

Two different classes of data set are generated under different mechanism. The sample size of the data is the total number of of observations from the two classes. The sample sizes of 100, 200 and 300 were used in the simulation study. For each sample size, the number of observations in the two classes are equal. The two classes of data are generated from a multivariate normal distribution; the first class of distribution has a mean vector of $[0 \quad 0]$ while the second class has a mean vector of $[-1 \quad -1]$ but both of them have the same variance covariance matrix $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, just that we can vary the value of ρ (which determines the correlation between variables) so that to have different variance covariance matrix for different values of ρ. For every value of ρ used, we evaluate the classification accuracy of the classifier by applying SVM classification on the clean data (the data without outliers) and we define this as the SVM accuracy of the clean data.

At this point, outliers were generated which are to be added to the first class and it has a mean vector which is different from the mean vector of the two classes. These generated outliers form a percentage of the first class and these percentages are what we defined as the contamination rate. We generated two scenarios:

Scenario 1: the mean vector of the distribution used to generate the outliers is further away from the mean vector of the group it should be added to i.e. the first class

Scenario 2: the mean vector of the distribution used to generate the outliers is closer to the mean vector of the group it should be added to i.e. the first class

After adding the outliers to the first class, the two classes now have equal number of observations in their classes.

We now apply the SVM classification process on the contaminated data (the data that contains outliers) and the classification accuracy of the classifier is evaluated which we called the SVM accuracy of the contaminated data. The same time, the two different outlier detection methods that are used to determine the rate at which they can detect the outliers present in the data. By using the Robust Mahalanobis distance, we first of all determine the mean of the first class by only using the clean observations since the mean is very sensitive to the presence of outliers and that is why it has a robust mean. The distance between the outliers and the mean of the class determines whether the observation is an outlier. The threshold (cut-off mark) that is used in this method is the Chebyshev inequality method, and the range that defines the lower and upper cut-off mark is the distance three standard deviation from both side of the mean. If any observation has a distance from the mean that is beyond this range, it is said to be identified as an outlier. The rate at which this method detects the outliers in the data is what we called the MCD average outlier detection rate.

The posterior probabilities of classification are calculated for each observation in the first class of the data, the value of the posterior probabilities is evaluated, and we applied the Chebyshev inequality on the data. The range of probabilities that is considered a clean observation are the observations which falls in the range of three standard deviation from both side of the mean. Observations with distances that is beyond this limit are considered to be an outlier(s). The rate which this method detects the outliers is evaluated and we define this method as the PoC average outlier detection rate.

This process is repeated for 1000 times for different sample sizes with different number of outliers. The correlation between the variables also has three levels: no correlation ($\rho = 0$), average correlation ($\rho = 0.5$) and high correlation ($\rho = 0.9$). The simulation process runs for each of the levels combined with different sample sizes and different contamination rate included.

5.1. First Scenario

The two classes of the data are generated from a multivariate normal distribution where the mean vector of the first class is [0 0] and the second class has a mean of [-1 -1], they both have the same variance-covariance matrix $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, this is the clean part of the data but the number of observations in the second class is more than those in the first class. This clean data is what is used as our train data to perform the classification of the observation using the SVM. Next we generate the outliers by using the same distribution but with a different mean vector of [-2 -2] and its variance covariance matrix is a two by two identity matrix $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, these outliers are attached to the observations in the first class.

The value of ρ in the variance covariance matrix can be varied on its three levels for this simulation: The first level is when there is no correlation between the two variables ($\rho = 0$), the second level is when the correlation between the two variables is weak ($\rho = 0.5$) while the third level is when the correlation between the two variables is high ($\rho = 0.9$). Varying both the levels of correlation and the number of outliers present in the data set (the first class), different sample sizes are used to evaluate the classification accuracy of the trained SVM. The classification performance of the SVM using the PoC and the SVM using the Mahalanobis distance are compared and their ability to detect the outliers present in the data are also evaluated.

The visual presentation of the impact of a high correlation between observations is demonstrated in Figure 5.1. When the correlation between the two variables is high and the data contains small number of outliers, the observations are stretched outward positively, good part of the observations of both classes are clearly apart and the outliers do not only have high outlying values, but they also seem misclassified. Hence the detection rate for outliers is higher using the PoC compared to the Mahalaobis distance and the SVM performs better on the clean data compared to the contaminated data.

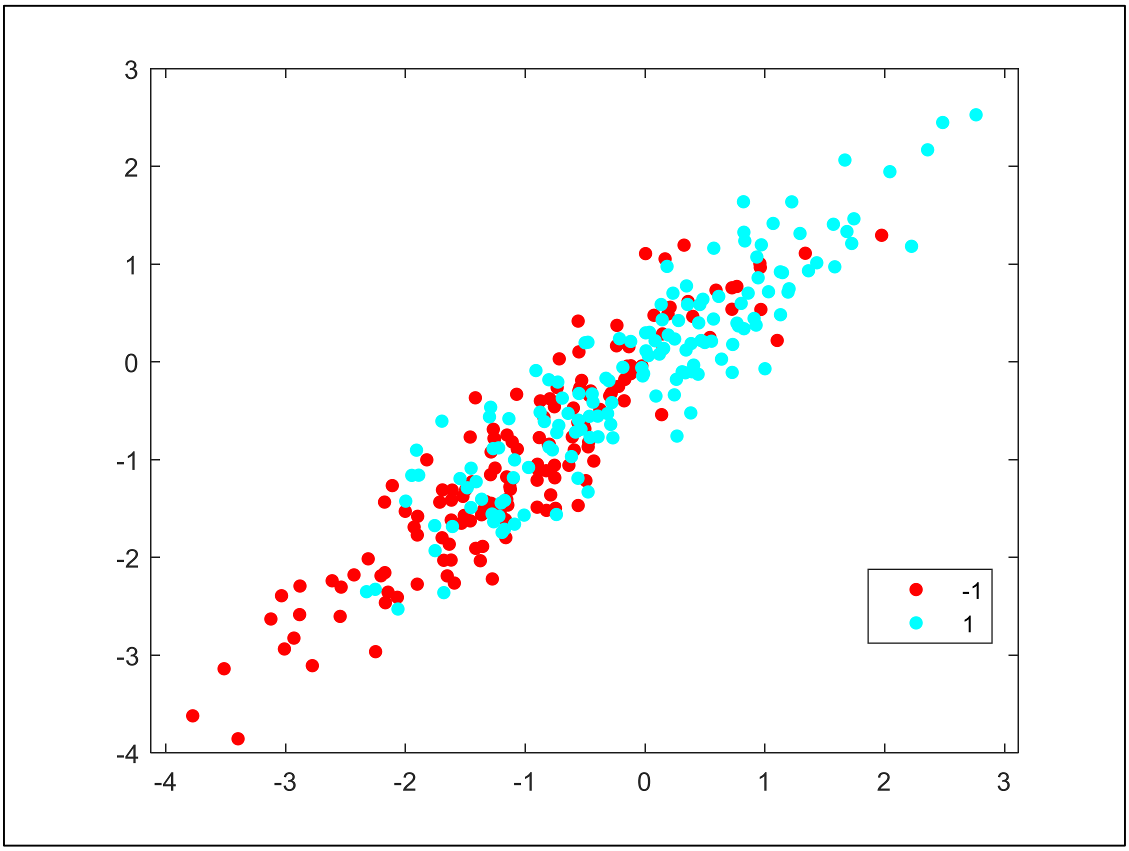


Figure 5.1. Scenario 1: n1=150 n2=150, contamination rate=0.10 and ρ=0.90

Suppose we decrease the correlation between the variables from high to weak, major part of the data lies at the center of the graph. The detection of Outliers is still higher in the PoC method compared to the Mahalanobis distance method, but both are lower than the case when the correlation is High. Also, the SVM classification accuracy is higher when the data is clean. Figure 5.2 illustrates the situation clearly.

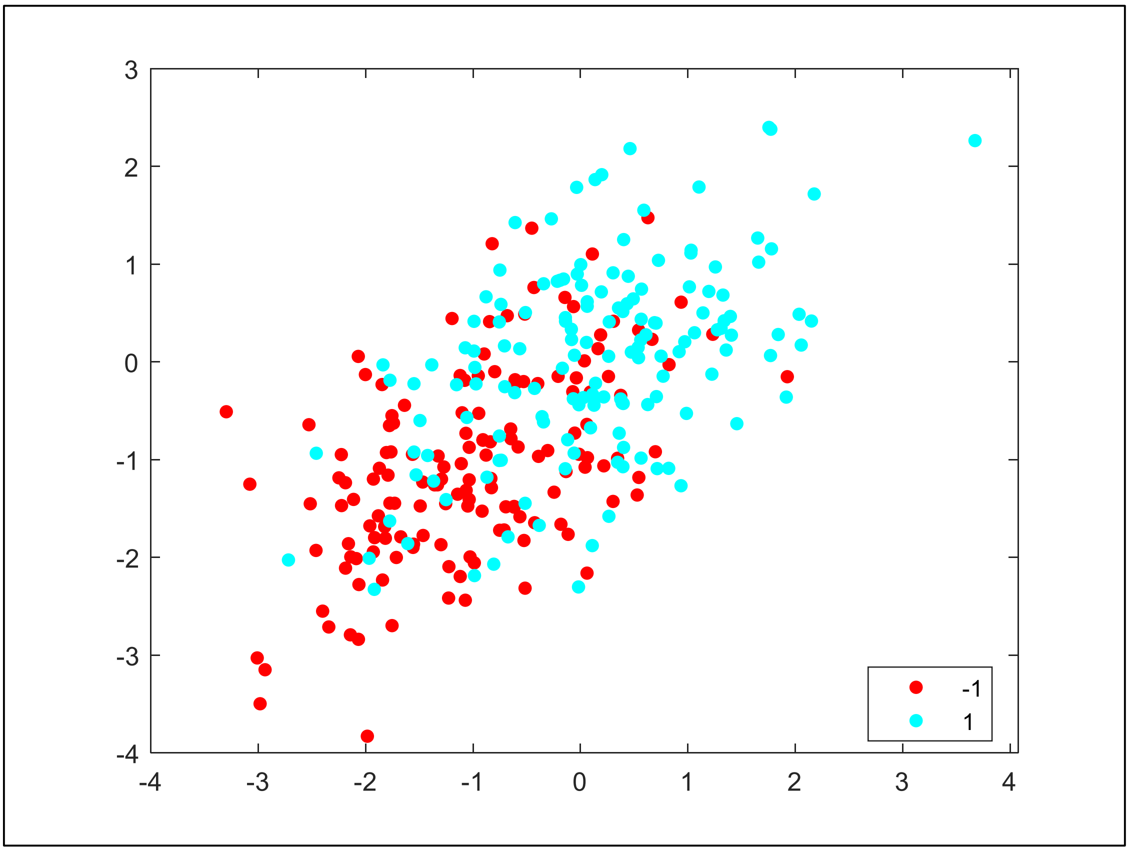


Figure 5.2. Scenario 1: n1=150 n2=150, contamination rate= 0.20 and =0.5

The final case demonstrated what happens when there is no correlation between the variables and the number of outliers in the data increased. Here most of the data set lies at the center of the graph since there is no linear correlation between the classes. The PoC detects outliers better than the Mahalanobis distance but the detection rate is poor compared to the two previous cases. It is also noted that the accuracy of the SVM is higher on the clean data. Figure 5.3 illustrates the idea.

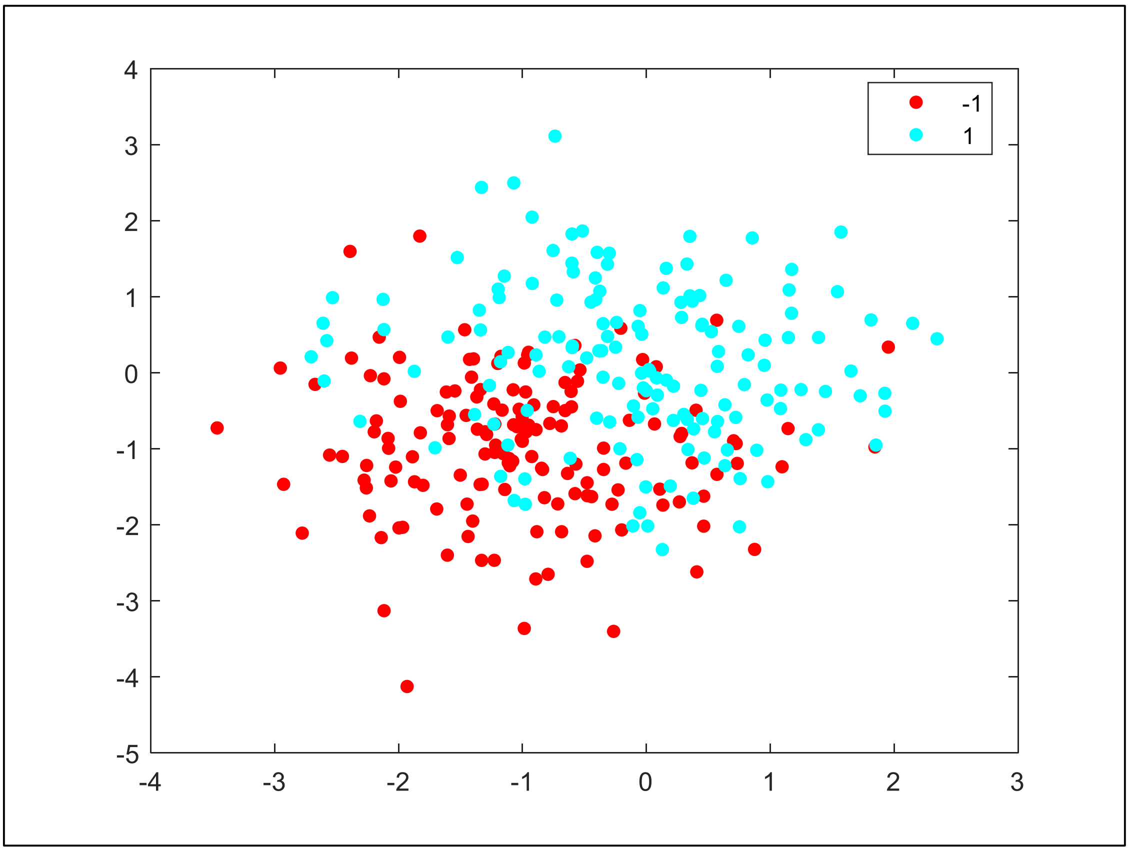


Figure 5.3. Scenario 1: n1=150 n2=150, contamination rate=0.30 and ρ=0

We now present the summary of all the sample sizes generated from the first scenario with different number of outliers and also different factor of the correlation between the two classes, together with the accuracy of the SVM when the data is clean(without outliers) or contaminated (with outliers) and the performances of both PoC and the Mahalanobis distance at detecting the outliers in the data in Table 5.1.

Table 5.1. The summary of classification accuracy of the SVM with clean and contaminated data and also the PoC and Mahalanobis distance in detecting outliers for scenario 1.

Total sample size N	Contamination rate	ρ	SVM accuracy of the clean data	SVM accuracy of the contaminated data	MCD average outlier detection rate	PoC average outlier detection rate
		0	0.7980	0.7788	0.9608	0.9908
100	%10	0.5	0.7565	0.7403	0.9630	0.9960
		0.9	0.7217	0.7138	0.8904	0.9986
		0	0.7977	0.7662	0.9352	0.9887
100	%20	0.5	0.7592	0.7322	0.9308	0.9977
		0.9	0.7237	0.7122	0.7305	0.9991
		0	0.7973	0.7562	0.9187	0.9907
100	%30	0.5	0.7594	0.7235	0.8927	0.9980
		0.9	0.7225	0.7153	0.6387	0.9985
		0	0.7846	0.7614	0.9689	0.9995
200	%10	0.5	0.7420	0.7193	0.9717	0.9999
		0.9	0.7130	0.7018	0.8980	1.0000
		0	0.7817	0.7412	0.9452	0.9994
200	%20	0.5	0.7413	0.7047	0.9406	1.0000
		0.9	0.7112	0.6969	0.7400	0.9998
		0	0.7842	0.7328	0.9325	0.9991
200	%30	0.5	0.7412	0.6949	0.9126	0.9999
		0.9	0.7119	0.6986	0.6428	1.0000
		0	0.7767	0.7505	0.9710	0.9998
300	%10	0.5	0.7357	0.7110	0.9715	1.0000
		0.9	0.7073	0.6945	0.9045	1.0000
		0	0.7762	0.7320	0.9485	0.9999
300	%20	0.5	0.7358	0.6935	0.9455	1.0000
		0.9	0.7076	0.6914	0.7449	1.0000
		0	0.7767	0.7205	0.9400	1.0000
300	%30	0.5	0.7350	0.6827	0.9228	1.0000
		0.9	0.7065	0.6906	0.6483	1.0000

The graphical representation of table 5.1 is shown in Figure 5.4. It compares the SVM accuracy on clean vs contaminated data for different sample sizes with different ρ values, shown as bars. In the same charts, the line graphs compare the outlier detection rate of MCD vs PoC for different sample sizes with different ρ values.

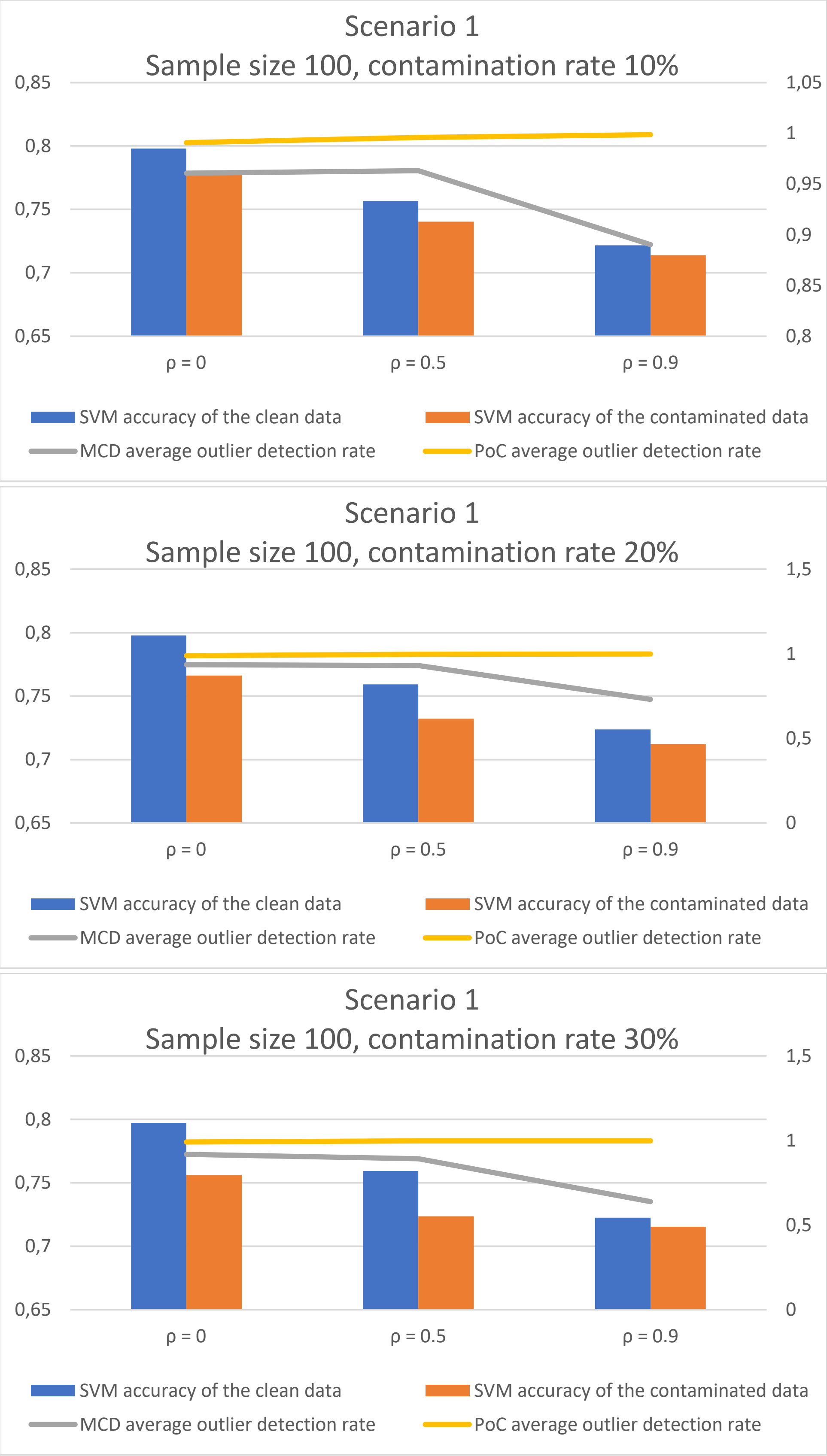
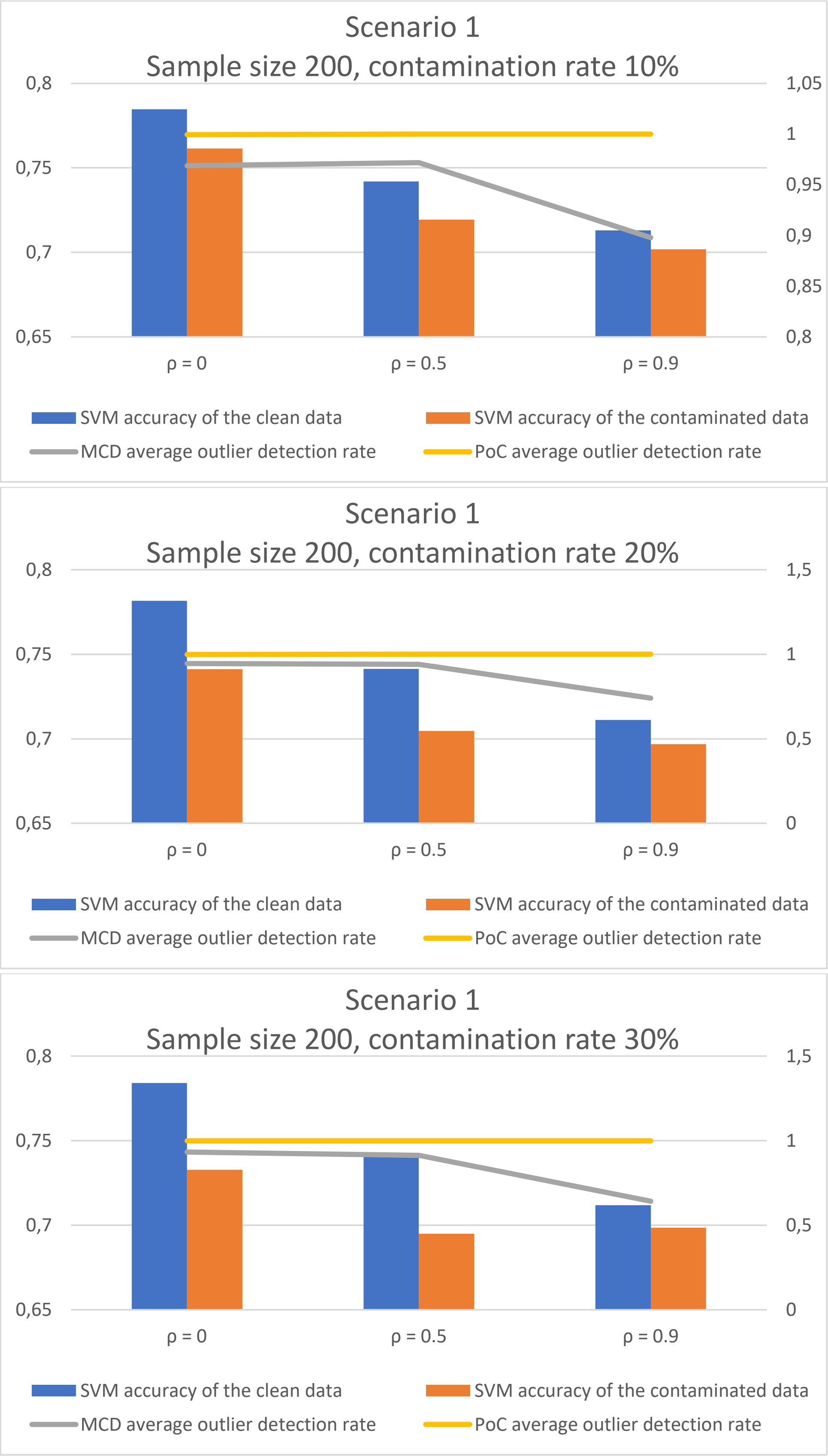


Figure 5.4. Scenario 1; Comparison of SVM accuracy on clean vs contaminated data and Outlier detection rate of MCD vs PoC for different sample sizes with different levels of correlation



Scenario 1
Sample size 200, contamination rate 10%

Scenario 1
Sample size 200, contamination rate 20%

Scenario 1
Sample size 200, contamination rate 30%

Figure 5.4. (continuation) Scenario 1; Comparison of SVM accuracy on clean vs contaminated data and Outlier detection rate of MCD vs PoC for different sample sizes with different levels of correlation

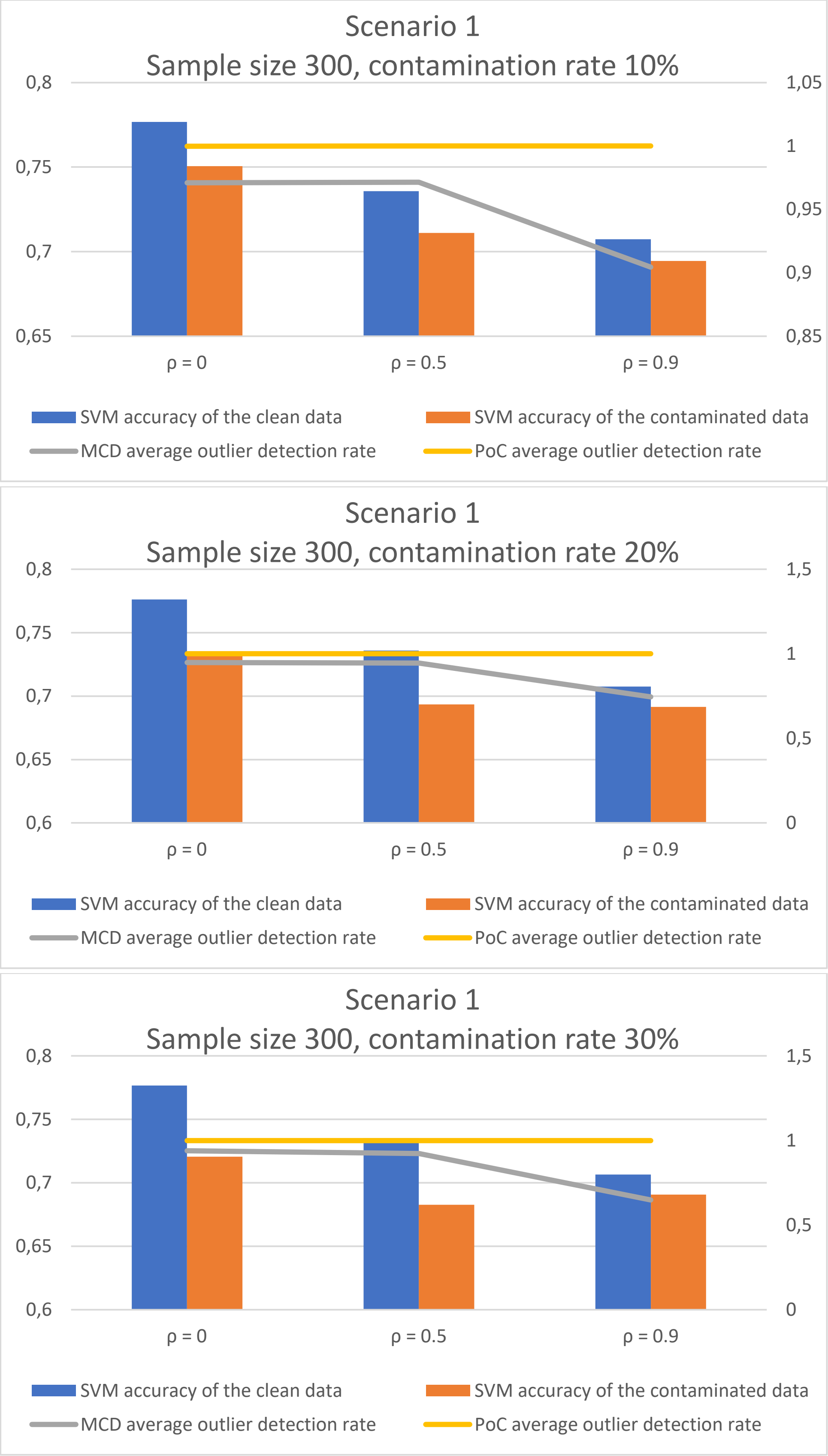


Figure 5.4. (continuation) Scenario 1; Comparison of SVM accuracy on clean vs contaminated data and Outlier detection rate of MCD vs PoC for different sample sizes with different levels of correlation

5.2. Second Scenario

In the second scenario, it is also similar to the first scenario just that it has a different mean vector. The first class is generated from a multivariate normal distribution with mean vector [0 0] while the second class has a mean vector of [1 1], but they both have the same variance covariance matrix $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. The outliers are generated by using the same distribution but with a mean vector [-1 -1] and an identity variance covariance matrix $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ which is added to the first class. The three levels of ρ are all examined with different sample sizes and different number of outliers.

When there is no correlation between the variables and the number of outliers in the data is small. Interestingly, the classification accuracy is higher when the contaminated data (the data with outliers) are used for classification. This can be due to the fact that the outliers are closer to the class they are included compared to the other class, this will make the classifier to shift its hyperplane closer to the first class and hence classifying most of them correctly. Unfortunately, the detection rate of outliers is reduced yet still the PoC outperforms the Robust Mahalanobis distance. Figure 5.4 demonstrates the scatter plot.

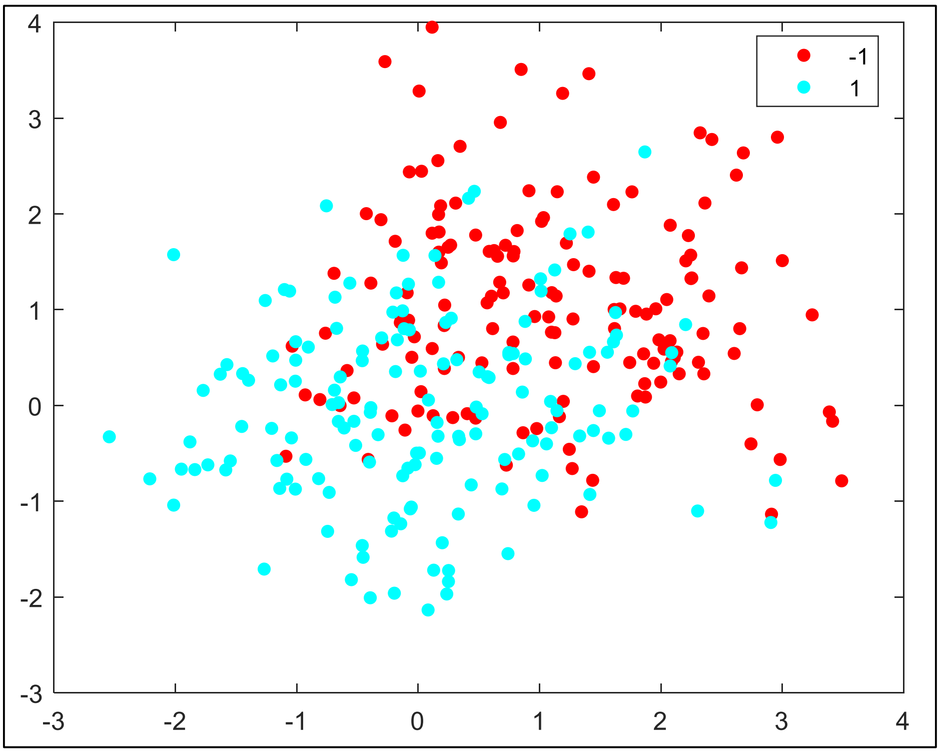


Figure 5.5. Scenario 2: n1=150 n2=150 contamination rate=%10, ρ=0

When there is a weak correlation between the variables, the data somehow stretched and the separation of the two classes can be visible. The PoC again outperforms the Mahalanobis in detecting outliers but not very high even though it is better than the previous case when there is no correlation between the variables. We indicate the scatter plot in Figure 5.5

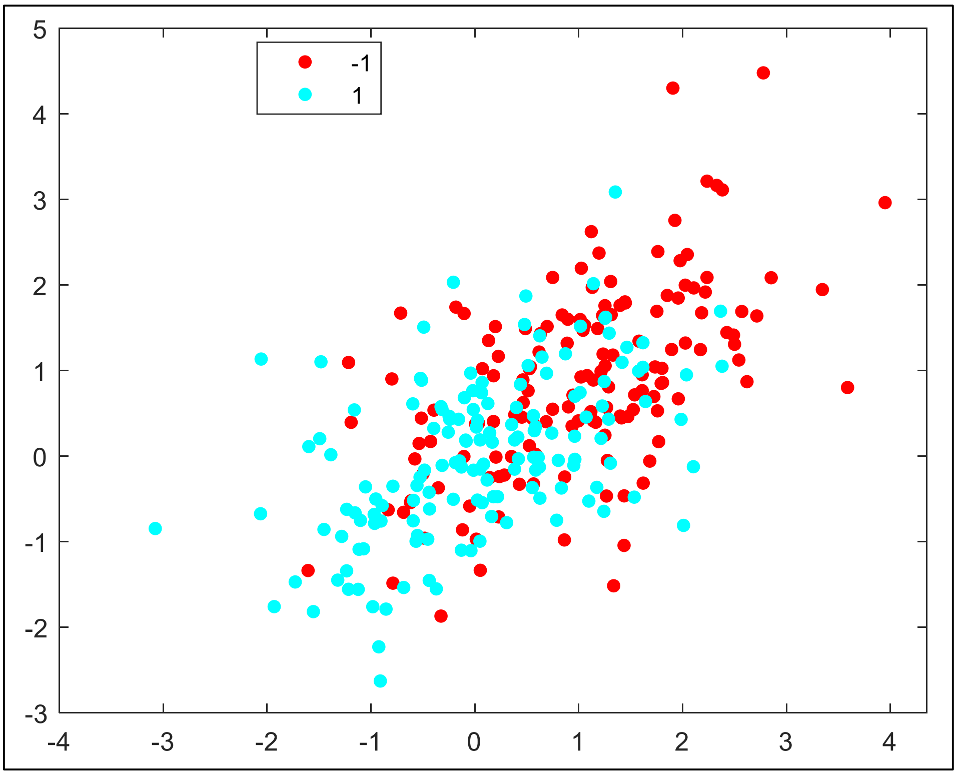


Figure 5.6. Scenario 2: n1=150 n2=150, contamination rate=%10, ρ=0.5

The final case for the second scenario is when the correlation between the variables is high, the separation of the two classes is clearly more visible compared to the previous two cases in scenario 2. This case gives a better detection rate of outliers and once again the PoC has a higher detection rate but still the classification accuracy of the SVM is better with the contaminated data. Figure 5.6 shows the scatter plot.

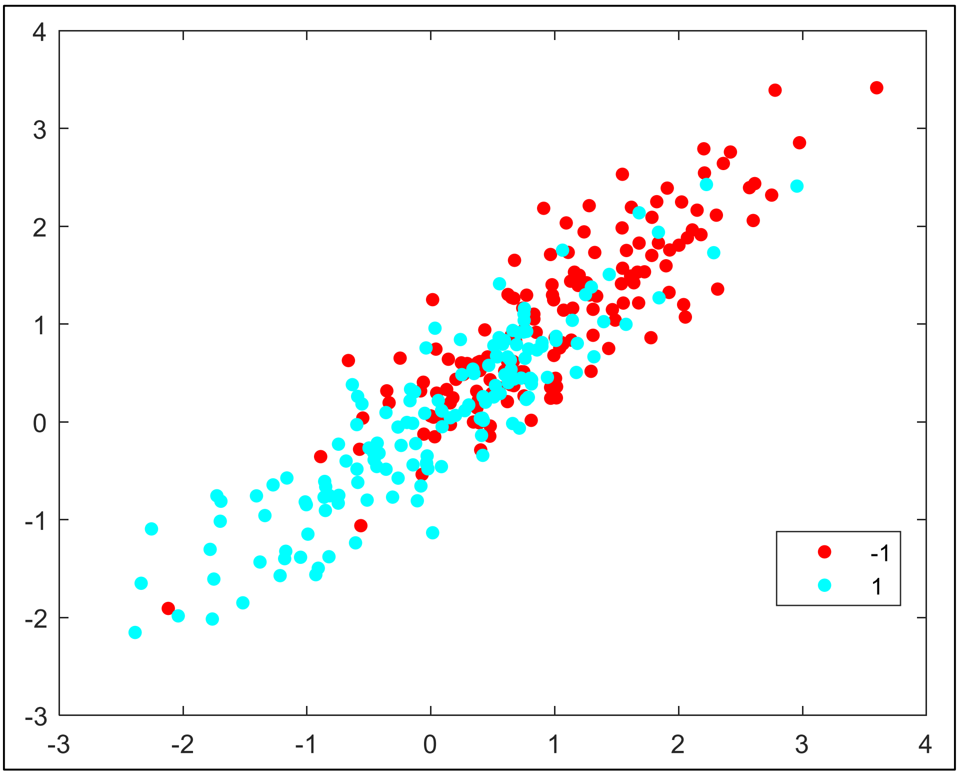


Figure 5.7. Scenario 2: n1=150 n2=150 contamination rate=%10, ρ=0.9

All the sample sizes generated from the second scenario are shown in Table 5.2 with different number of outliers and also different factor of the correlation between the two classes, together with the accuracy of the SVM when the data is with or without an outlier(s) and the performances of both PoC and the Mahalanobis distance at detecting the outliers in the data.

Table 5.2. The summary of classification accuracy of the SVM with clean and contaminated data and also the PoC and Mahalanobis distance in detecting outliers for scenario 1.

Total sample size N	Contamination rate	ρ	SVM accuracy of the clean data	SVM accuracy of the contaminated data	MCD average outlier detection rate	PoC average outlier detection rate
		0	0.7992	0.8087	0.8566	0.9998
100	%10	0.5	0.7581	0.7699	0.8530	1.0000
		0.9	0.7249	0.7378	0.7904	1.0000
		0	0.8015	0.8234	0.8272	0.9995
100	%20	0.5	0.7575	0.7815	0.8301	0.9999
		0.9	0.7233	0.7513	0.6771	1.0000
		0	0.7968	0.8303	0.8062	0.9993
100	%30	0.5	0.7581	0.7952	0.7995	0.9995
		0.9	0.7241	0.7700	0.6148	0.9998
		0	0.7820	0.7930	0.8667	1.0000
200	%10	0.5	0.7416	0.7544	0.8772	1.0000
		0.9	0.7108	0.7248	0.7932	1.0000
		0	0.7843	0.8069	0.8441	0.9998
200	%20	0.5	0.7433	0.7692	0.8453	1.0000
		0.9	0.7115	0.7407	0.6721	1.0000
		0	0.7834	0.8196	0.8268	0.9997
200	%30	0.5	0.7429	0.7837	0.8160	0.9999
		0.9	0.7103	0.7582	0.6118	1.0000
		0	0.7774	0.7890	0.8765	1.0000
300	%10	0.5	0.7361	0.7490	0.8783	1.0000
		0.9	0.7073	0.7225	0.7918	1.0000
		0	0.7774	0.8001	0.8526	1.0000
300	%20	0.5	0.7353	0.7618	0.8537	1.0000
		0.9	0.7075	0.7379	0.6750	1.0000
		0	0.7785	0.8136	0.8241	1.0000
300	%30	0.5	0.7349	0.7773	0.8213	1.0000
		0.9	0.7079	0.7563	0.6129	1.0000

The graphical representation of table 5.2 is shown in Table 5.8. It compares the SVM accuracy on clean vs contaminated data for different sample sizes with different ρ values, shown as bars. In the same charts, the line graphs compare the outlier detection rate of MCD vs PoC for different sample sizes with different ρ values.

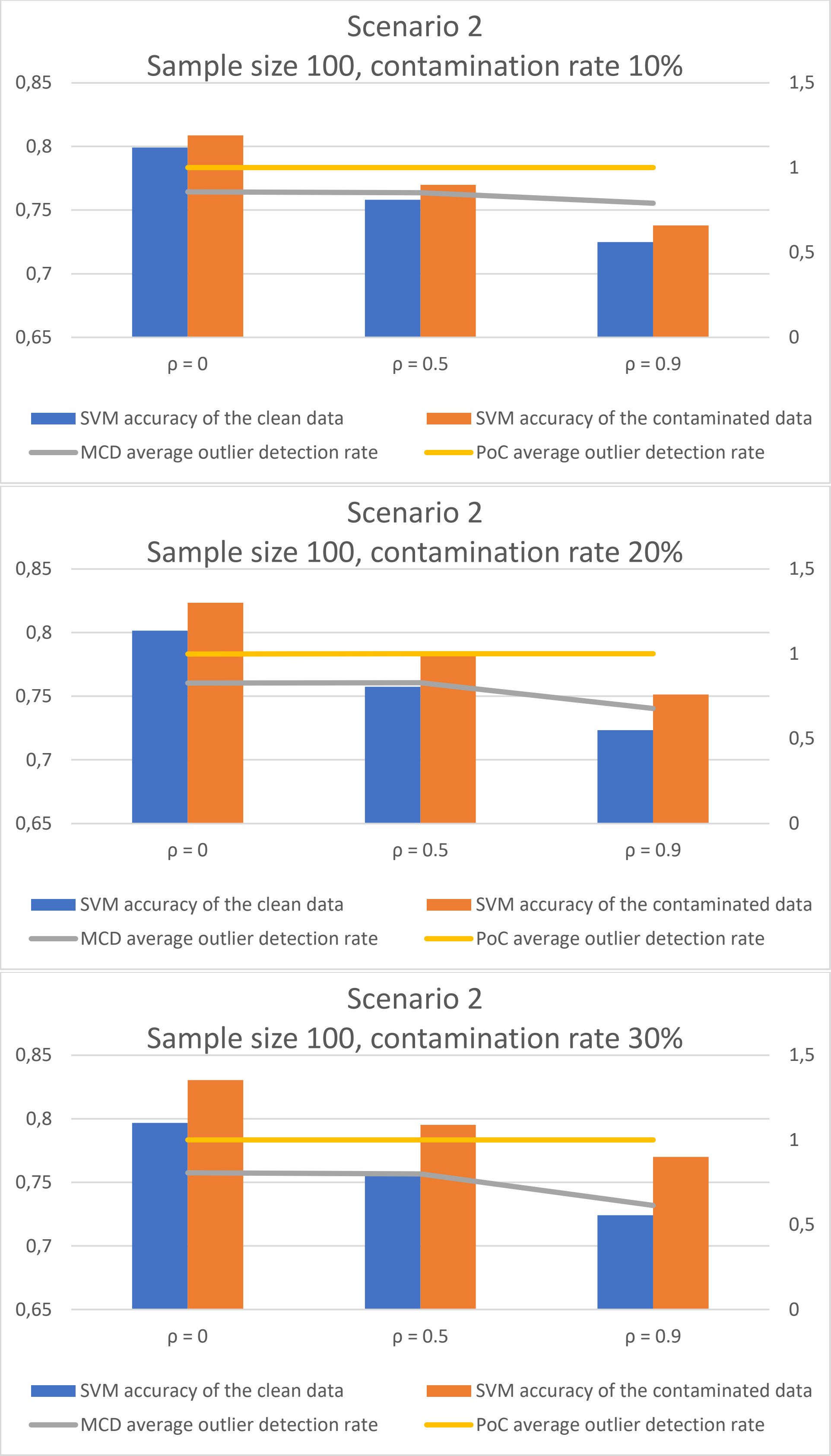


Figure 5.8. Scenario 2; Comparison of SVM accuracy on clean vs contaminated data and Outlier detection rate of MCD vs PoC for different sample sizes with different levels of correlation

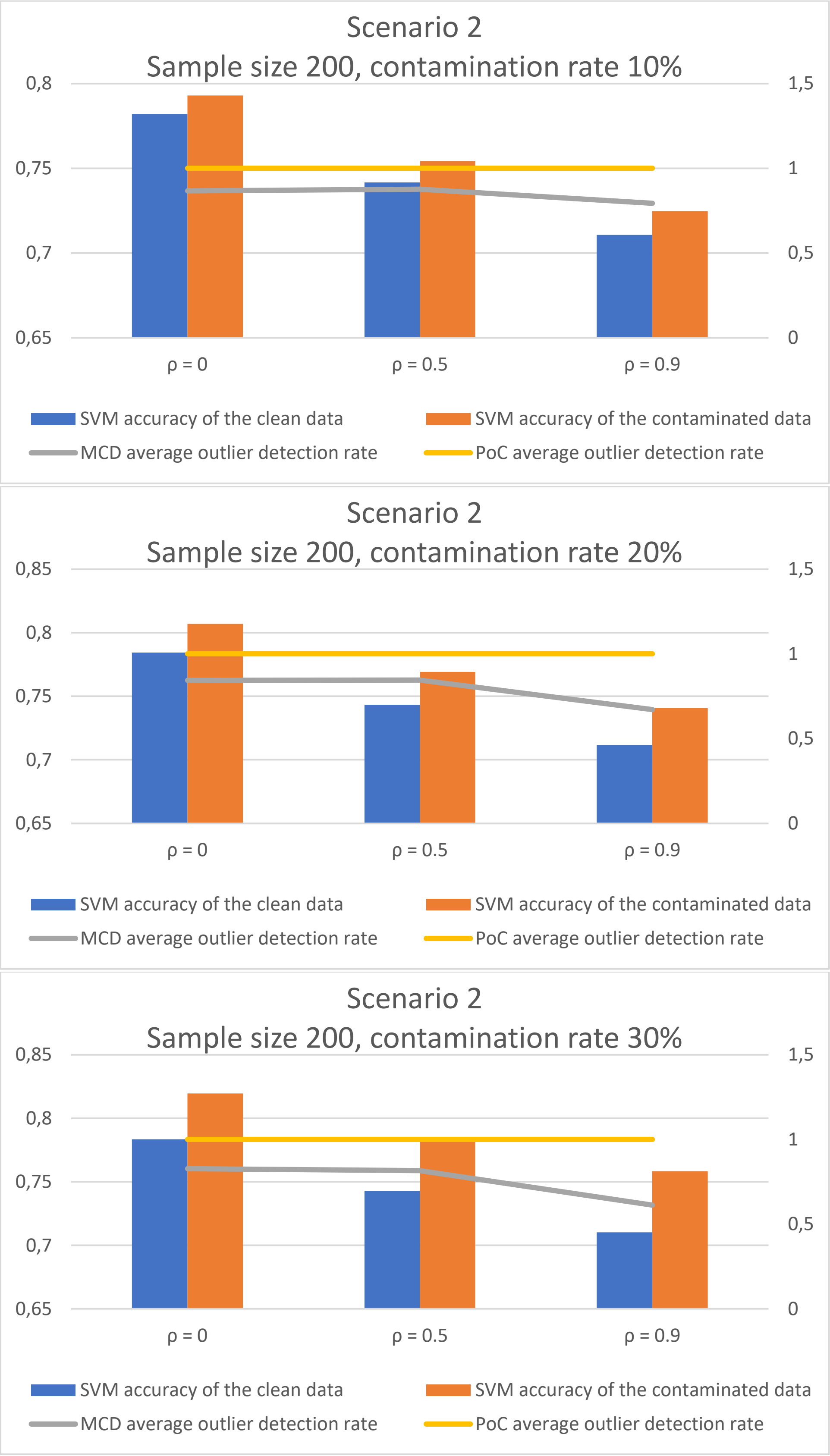


Figure 5.8. (continuation) Scenario 2; Comparison of SVM accuracy on clean vs contaminated data and Outlier detection rate of MCD vs PoC for different sample sizes with different levels of correlation

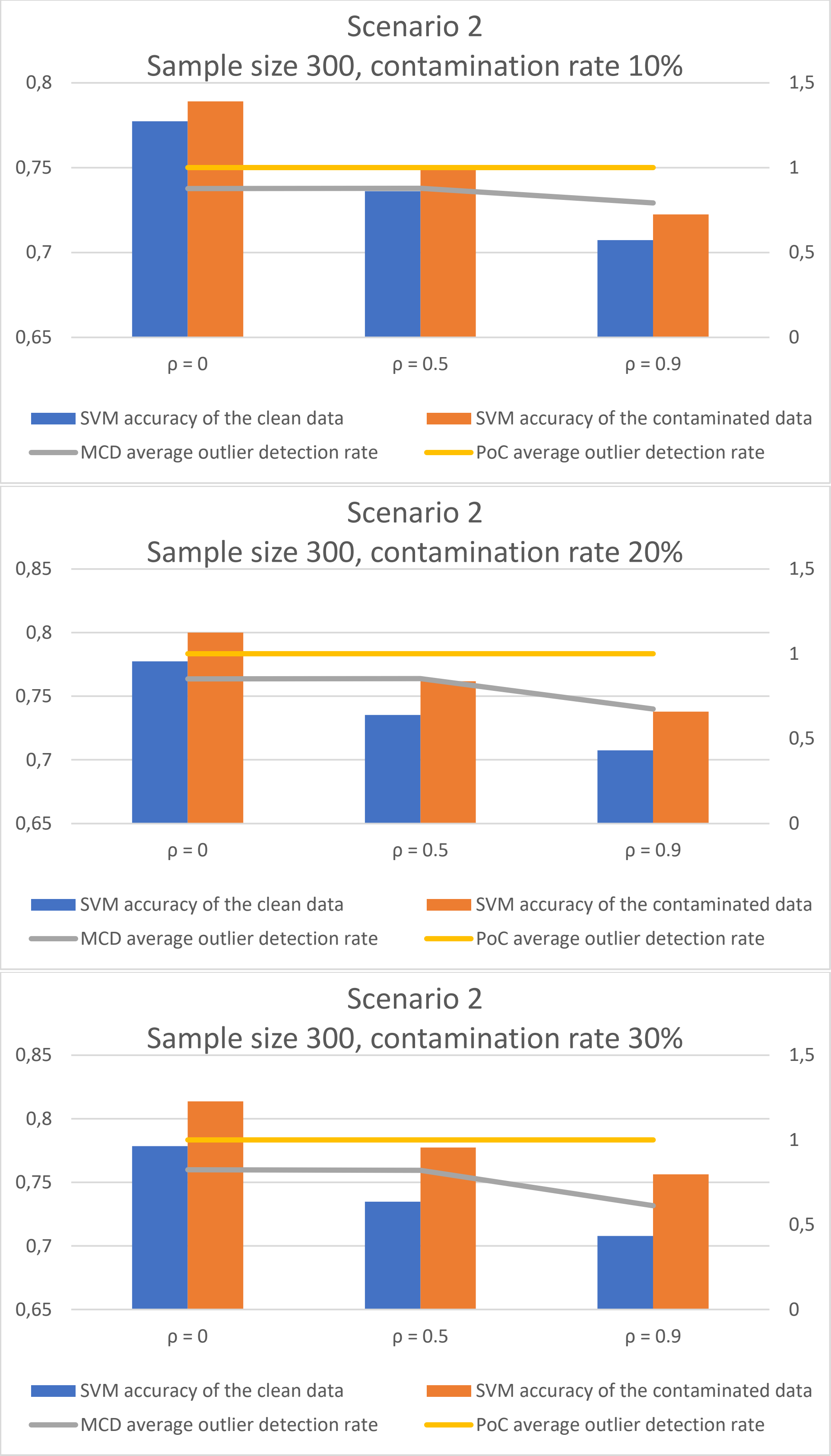


Figure 5.8. (continuation) Scenario 2; Comparison of SVM accuracy on clean vs contaminated data and Outlier detection rate of MCD vs PoC for different sample sizes with different levels of correlation

6. CONCLUSION

Scenario 1

For a given sample size and contamination rate, the classification accuracy for both the clean and contaminated data reduces as the correlation between the observations increases but the classification accuracy with the clean data is higher at all levels of correlation. Also, the outlier detection rate for PoC increases as the value of ρ increases.

For a given sample size and a given correlation level between observations, the classification accuracy for both the clean and contaminated data decreases as the contamination rate increases. Also, the outlier detection rate for PoC increases as the contamination rate increases.

As the sample size increases for a given contamination rate, the outlier detection rate for the PoC increases for the corresponding values of ρ. Thus, the classification accuracy for both the clean and contaminated data decreases for the corresponding ρ values.

For a given sample size, the outlier detection rate for the MCD decreases as the correlation between observations increases.

The MCD outlier detection rate decreases as for a given sample size as the contamination rate increases.

As the sample size increases, The PoC outlier detection method detects all the outliers present in the data.

For all cases, the outlier detection rate for the PoC performs better than the MCD

Scenario 2

For a given sample size and contamination rate, the classification accuracy for both the clean and contaminated data decreases as the correlation between the observations increases but the SVM performs better with the contaminated data at all correlation levels.

For a given level of correlation between variables and a given contamination rate, the classification accuracy for both the clean and contaminated data decreases as the sample size increases.

When the correlation between variables increase for a given sample size, the outlier detection rate for MCD decreases while that of the PoC increases.

As we increase the contamination rate for a given sample size, the outlier detection rate for MCD decreases while the outlier detection rate for PoC increases on all levels of correlations between the variables

For a given correlation level and a given contamination rate, both the outlier detection rate for both MCD and PoC increases as the sample size increases.

The PoC was able to detect almost all the outliers present in the data with all the correlation levels and the contamination rates.

For all cases, the PoC performs better than the MCD in detecting the outliers present in the data.

REFERENCES

1. Acuña, E. and Rodriguez, C. (2004). On detection of outliers and their effect in supervised classification. *University of Puerto Rico at Mayaguez*, *15*.

2. Agresti, A. and Kateri, M. (2013). *Categorical data analysis*. Springer Berlin Heidelberg, 206-208.

3. Anderson, T.W. and Mathématicien, E.U. (1958). *An introduction to multivariate statistical analysis.* New York: Wiley, (2), 3-5.

4. Ben-Gal, I. (2005). *Outlier detection.* In Data mining and knowledge discovery handbook. Springer, Boston, MA, 131-146.

5. Chiang, J.T. (2007). The masking and swamping effects using the planted mean-shift outliers models. *International Journal Contemporary Mathematics Sciences, 2*(7), 297-307.

6. Collett, D. (2002). *Modelling binary data*. Chapman and Hall/CRC.

7. Davies, L. and Gather, U. (1993). The identification of multiple outliers. *Journal of The American Statistical Association, 88*(423), 782-792.

8. Este, A., Gringoli, F. and Salgarelli, L. (2009). Support vector machines for TCP traffic classification. *Computer Networks*, *53*(14), 2476-2490.

9. Franklin, S., Thomas, S. and Brodeur, M. (2000, June). Robust multivariate outlier detection using Mahalanobis' distance and modified Stahel-Donoho estimators. In *Proceedings of the second international conference on establishment surveys*. American Statistical Association Buffalo, NY, 697-706.

10. Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1-3), 389-422.

11. Hawkins, D.M. (1980). *Identification of outliers.* London: Chapman and Hall, 11.

12. Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*(2), 85-126.

13. Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X. (2013). *Applied logistic regression*. John Wiley & Sons, 389.

14. Hubert, M., Rousseeuw, P.J. and Van Aelst, S. (2005). Multivariate outlier detection and robustness. *Handbook of Statistics*, *24*, 263-302.

15. Hussain, S., Mohamed, M.A., Holder, R., Almasri, A. and Shukur, G. (2008). Performance evaluation based on the robust Mahalanobis distance and multilevel modeling using two new strategies. *Communications in Statistics-Simulation and Computation®, 37*(10), 1966-1980.

16. İnternet: Logit Models for Binary Data. URL: http://data.princeton.edu/ wws509/notes/ c3.pdf, Son Erişim Tarihi: 14.07.2019.

17. İnternet: Decision Tree, URL: https://www.investopedia.com/terms/d/decision-tree.asp, Son Erişim Tarihi: 14.07.2019.

18. James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *Statistical learning.* In An Introduction to Statistical Learning. Springer, New York, NY, 15-57.

19. Joachims, T. (1998, April). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*. Springer, Berlin, Heidelberg, 137-142.

20. Jolliffe, F.R. (1986). *Survey design and analysis*. Ellis Horwood.

21. Kannan, K.S. and Manoj, K. (2015). Outlier detection in multivariate data. *Journal of Applied Mathematical Sciences, 9,* 2317-2324.

22. Kaufman, L. and Rousseeuw, P.J. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, *344*, 68-125.

23. Khanna, D., Sahu, R., Baths, V. and Deshpande, B. (2015). Comparative study of classification techniques (SVM, logistic regression and neural networks) to predict the prevalence of heart disease. *International Journal of Machine Learning and Computing*, *5*(5), 414.

24. Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling.* New York: Springer, 26.

25. Leroy, A.M. and Rousseeuw, P.J. (1987). Robust regression and outlier detection. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley*.

26. Min, J.H. and Lee, Y.C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems With Applications*, *28*(4), 603-614.

27. Musa, A.B. (2013). Comparative study on classification performance between support vector machine and logistic regression. *International Journal of Machine Learning and Cybernetics*, *4*(1), 13-24.

28. Park, H. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing*, *43*(2), 154-164.

29. Patil, T. R. and Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications*, *6*(2), 256-261.

30. Pisarenko, V.F. and Sornette, D. (2012). Robust statistical tests of Dragon-Kings beyond power law distributions. *The European Physical Journal Special Topics*, *205*(1), 95-115.

31. Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers, 10*(3), 61-74.

32. Qi, Z., Tian, Y. and Shi, Y. (2013). Structural twin support vector machine for classification. *Knowledge-Based Systems, 43,* 74-81.

33. Riva, M., Neuman, S.P. and Guadagnini, A. (2013). On the identification of Dragon Kings among extreme-valued outliers. *Nonlinear Processes in Geophysics*, *20*(4), 549-561.

34. Salazar, D. A., Vélez, J. I. And Salazar, J. C. (2012). Comparison between SVM and logistic regression: Which one is better to discriminate?. *Revista Colombiana de Estadística*, *35*(2), 223-237.

35. Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427-437.

36. Sornette, D. (2009). Dragon-kings, black swans and the prediction of crises. *arXiv preprint arXiv, 0907.4290*.

37. Sornette, D. and Ouillon, G. (2012). *Dragon-kings: mechanisms, statistical methods and empirical evidence.* The European Physical Journal Special Topics, *205*(1), 1-26.

38. Sun, A., Lim, E.P. and Ng, W.K. (2002). Web classification using support vector machine. In *Proceedings of the 4th international workshop on Web information and data management*, 96-99. ACM.

39. Tsai, C.F. (2005). Training support vector machines based on stacked generalization for image classification. *Neurocomputing*, *64*, 497-503.

40. Wang, L., Zhang, Z. and Design, C. X. R. C. (2005). Theory and applications. *Support Vector Machines,* Springer-Verlag, Berlin Heidelberg, 177.

41. Wheatley, S. and Sornette, D. (2015). Multiple outlier detection in samples with exponential & pareto tails: Redeeming the inward approach & detecting dragon kings. *Swiss Finance Institute Research Paper,* 15-28.

42. Zanaty, E.A. (2012). Support vector machines (SVMs) versus multilayer perception (MLP) in data classification. *Egyptian Informatics Journal*, *13*(3), 177-183.

CURRICULUM VITAE

Personal Information

Surname, Name	: CEESAY, Habib
Nationality	: Gambian
Date and Place of Birth	: 28.12.1988, New Jeshwang
Marital status	: Married
Phone number	: +(220) 7449533
E-Mail	: noblehabz@gmail.com

Education

Degree	School/ Program	Graduation Date
Master's Degree	Gazi University/Statistics	Ongoing
Undergraduate	University of The Gambia/Mathematics	2012

Professional Experience

Year	Place of Work	Position
2014 – Ongoing	University of The Gambia	Graduate Assistant
2012 – 2013	Standard Chartered Bank (Gambia)	Cashier
2011 – 2012	Nusrat Senior Secondary School	Mathematics Teacher

Foreign Language

English, Turkish

Hobbies

Reading, Soccer, Swimming

GAZİ GELECEKTİR…