UNIVERSITÉ DE MONTRÉAL

A NOVELTY DETECTION TOOL BASED ON PARALLEL COORDINATES PLOT

SHEIDA SHAMS SHIRAZI
DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE INDUSTRIEL)
DÉCEMBRE 2017

© Sheida Shams Shirazi, 2017.

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

A NOVELTY DETECTION TOOL BASED ON PARALLEL COORDINATES PLOT

présenté par : SHAMS SHIRAZI Sheida
en vue de l'obtention du diplôme de: Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de :

M. ROBERT Jean-Marc, Doctorat, président
M. BASSETTO Samuel Jean, Doctorat, membre et directeur de recherche
M. FRAYRET Jean-Marc, Ph. D., membre

DEDICATION

to my husband, Vahid;
I am truly thankful for having you in my life....

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Prof. Samuel Bassetto for his continuous support and commitment which had a great impact on my work and for all the wonderful lessons that he has taught me during my master program.

My sincere thanks also goes to my thesis committee: professor Jean-Marc Frayret and professor Jean-Marc Robert for kindly accepting to be my juries.

Additionally, I would like to thank all of my friends and colleagues for their help throughout this journey.

Finally, but most importantly, I wish to thank my parents and my husband for their endless supports.

RÉSUMÉ

La détection de nouveauté est le problème de trouver des événements inconnus ou des échantillons nommés nouveautés quand il y a peu d'informations disponibles ou même aucune information disponible les concernant. En fait, un classificateur de détection de nouveauté est entraîné par les données historiques. L'ensemble de données historiques contiennent les données normales attendues. Ensuite, les nouveautés sont détectées alors qu'elles sont inconnues du classificateur. Il y a beaucoup de cas dans différentes industries où la collecte de données anormales devient un problème paralysant. Par exemple, parfois dans l'étude des soins de santé, ce n'est pas faisable de collecter les échantillons anormaux, parce que fournir les conditions de survenue d'un nouvel échantillon peut nuire aux individus et à l'environnement. La visualisation des données peut avoir un impact efficace sur la détection de ces nouveaux comportements et les analyser et aider à améliorer le processus de prise de décision. Mais aucun des détecteurs de nouveauté n'a pas été établi sur la base d'un graphique de visualisation de données multivariées. Dans cette étude, nous visons à développer un classificateur visuel pour le problème de détection de nouveauté. Ceci est réalisé en développant un outil de détection de nouveauté basé sur le puissant potentiel géométrique de coordonnées parallèles combinées avec le clustering K-medoids. Cet outil, appelé NDTool, pourrait être facilement utilisé dans diverses industries, y compris les soins de santé. Les résultats sur les jeux de données réelles montrent que NDTool a un rendement efficace pour résoudre les problèmes de détection de nouveauté et produit des résultats compétitifs par rapport aux autres algorithmes étudiés dans le travail actuel. Puis NDTool est utilisé pour certaines études sur le cancer du sein comme un moyen à bas coût de suivre la masse suspecte dans le sein. Il aide à détecter précocément de la masse cancéreuse qui est le facteur le plus important du taux de survie des patients.

ABSTRACT

Novelty detection is the problem of finding unknown events or samples named novelties when there is a limited information or even no information available about them. In fact, a novelty detection classifier is trained by the historical data. The historical dataset contains the expected "normal" data. The novelties are detected while they are unknown to the classifier. There are a lot of cases in different industries where collecting abnormal data becomes a crippling problem. For example, sometimes in the healthcare studies, it is not feasible to collect the abnormal samples, because providing the conditions of occurring of a novel sample may harm the individuals and environment. Data visualization can have an effective impact on detecting such novel behaviors and analyze them and help to improve the process of decision-making. But none of the novelty detectors has been established based on a multivariate data visualization. This study aims to develop an innovative classifier for novelty detection problem. This is achieved by developing a novelty detection tool based on the powerful geometric potential of parallel coordinates plot combined with k-medoids clustering. This tool, named NDTool, could be easily used in various industries. Computational results on real-life datasets show that NDTool is efficient for solving novelty detection problem and produce competitive results compared to the other investigated algorithms in this work. Then NDTool is employed for breast cancer studies as a low-cost way of tracking suspicious masses in the breast. It helps to an early detection of the cancerous mass which is one of the most important factors of the survival rate of patients.

TABLE OF CONTENTS

DEDICATION iii

ACKNOWLEDGEMENTS iv

RÉSUMÉ v

ABSTRACT vi

TABLE OF CONTENTS vii

LIST OF TABLES ix

LIST OF FIGURES x

LIST OF SYMBOLS AND ABBREVIATIONS xi

CHAPTER 1 INTRODUCTION 1
- 1.1 Problem Definition 1
- 1.2 A brief overview of the proposed methodology 3

CHAPTER 2 LITERATURE REVIEW 4

CHAPTER 3 MATERIALS AND METHOD 8
- 3.1 Parallel coordinates 8
- 3.2 Clustering 10
  - 3.2.1 Average silhouette method 11
- 3.3 NDTool 13

CHAPTER 4 NDTOOL PERFORMANCE 20
- 4.1 Performance Measures 20
- 4.2 Data description 23
- 4.3 Comparative study with state-of-art algorithms 24
- 4.4 Breast Cancer Experiments 27

CHAPTER 5 DISCUSSION 32

CHAPTER 6 CONCLUSION 34

6.1 Synthesis of the work 34
6.2 Limitations of the proposed solution 34
6.3 Future Work 35

REFERENCES 36

LIST OF TABLES

Table 4.1 The benchmark datasets 23
Table 4.2 Comparisons of AUC performance of different methods for benchmark datasets 26
Table 4.3 Performance of NDTool 26
Table 4.4 Characteristics of breast cancer datasets 28
Table 4.5 Definition of attributes in breast cancer datasets 28
Table 4.6 Performance of NDTool for Breast Cancer Studies 29
Table 4.7 Participation percentage of each attribute in novel behaviour for Diagnostic Breast Cancer dataset 31

LIST OF FIGURES

Figure 2.1 Support Vector Machines 6
Figure 2.2 k-NN 6
Figure 2.3 Gaussian mixture models 6
Figure 2.4 K-means 6
Figure 3.1 Example of parallel coordinates plot for Ionosphere dataset. 8
Figure 3.2 Illustrating Cartesian coordinates and their parallel coordinates plots 9
Figure 3.3 Difference between K-means and K-mediods for assigning a representative for a cluster which contains outlier 10
Figure 3.4 Cohesion versus Separation in clustering 12
Figure 3.5 silhouette factor plot of "normal" observations of Ionosphere dataset, when the number of clusters is three.. 13
Figure 3.6 Example of finding upper bound and lower bound in each dimension for each cluster. 14
Figure 3.7 Example of finding tangent of the angles in each dimension. 15
Figure 3.8 Numerical example of calculating tangents of the angles in a dimension. 15
Figure 3.9 A new observation and clustered data by K-medoid 16
Figure 3.10 Visualizing novel behaviours in parallel coordinates plot 17
Figure 3.11 Parallel coordinates of a newcomer data versus the standardized historical "normal" data for Diagnostic Breast Cancer dataset 18
Figure 3.12 Flowchart of NDTool 19
Figure 4.1 Confusion Matrix 20
Figure 4.2 Receiver Operating Characteristics curve 22
Figure 4.3 Strategy of dividing the dataset into train and test 24
Figure 4.4 ROC plots of benchmark datasets 25
Figure 4.5 ROC plots of Breast Cancer datasets 29
Figure 4.6 Parallel coordinates of Diagnostic Breast Cancer dataset with the cluster of novel behaviours which is in dark grey 30
Figure 4.7 Pie Chart for Diagnostic Breast Cancer dataset 30

LIST OF SYMBOLS AND ABBREVIATIONS

NDTool	Novelty Detection Tool
SVDD	Support vector data description
K-NN	K-nearest neighbours
GMM	Gaussian mixture models
AUC	Area Under Curve
ROC	Receiver Operating Characteristics
UCI	University of California at Irvine
WBCD	Wisconsin Breast Cancer Database

CHAPTER 1 INTRODUCTION

"Numbers have an important story to tell. They rely on you to give them a voice." - Stephen Few[1]

Today, one of the greatest challenges faced by industries and businesses is data explosion (BI-Survey, 2017). Data analysis tools help extracting more precious information from data where achieving a fast and efficient decision-making is a constant need. There is thus a growing demand for developing new algorithms and tools to perform effective analysis and improve the decision-making process. Translating obtained information into visuals can help the decision-making process, since the human brain is able to process images much faster than text (Pant, 2015). This talent help us to get more from data visualization and digest more information in a glance in the era of the data analysis.

1.1 Problem Definition

In many real-world cases, some kind of samples rarely occur. These samples are difficult or even impossible to be obtained while the information about "normal"[2] behaviour are accessible. For example in aircraft engine condition monitoring (Hayton et al., 2007), the data of abnormality in engine condition is under-represented and it is not possible to destroy jet engines just to see how they fail. In domain of IT system security, the abnormal observations are from system crash. Then, while the number of abnormal sample (crashes) are low, it is not possible to crash down the system in order to collect abnormal samples. Another example is medical diagnosis, where relatively few observations of a special disease are available in medical tests (Marsland, 2003). Lacking enough sample of a possible category (class), diminish the performance of an algorithm whose function is distinguishing the category of each new sample.

So, what are the solutions of aforementioned condition and how can these solutions be used to aid non expert people in different industries?

The above cases are in the context of *novelty detection* problem. Novelty detection is the task of identifying new and unknown data that differ from the "normal" behavior of historical data. Novelty detection has attracted considerable attention as shown by the increasing number of publications (Pimentel et al., 2014).

[1] Innovator, consultant, and educator in the fields of business intelligence

[2] For clarity, the term "normal" indicates the expected regular samples. To create a distinction between normal distribution and normal samples we used the quotation marks.

There are a lot of real life application of novelty detection in different domains, such as industrial health monitoring and fault detection (Hu et al., 2012; Hayton et al., 2007), IT security systems (Helali, 2010; Jyothsna et al., 2011), healthcare and medical diagnostics (Kuijf et al., 2016), environmental monitoring (Worden et al., 2002) and video surveillance (Diehl and Hampshire, 2002; Owens et al., 2002).

For example in the healthcare domain, a novelty detection problem could be defined as a medical diagnosis when a patient has irregular test records. In cancer studies, Tarassenko et al. (1995) used novelty detection for identification of masses in mammograms. In IT security domain, novelty detection is applied for finding the network intrusion, detecting unauthorized access to a system or a computer network. Finding such novel behaviour can help intrusion prevention. In fraud detection domain, a novel behaviour could interpret as a stolen credit cards, a bogus insurance claim or a financial transactions. Detecting the above case and analyzing the information they contain can help preventing their frequent occurrence.

The above mentioned applications usually contains large volumes of registered information about patient, customers, systems, etc. in multi-dimentional datasets. Understanding complex high-dimensional datasets is an important yet challenging problem. Multivariate visualization such as parallel coordinates (Inselberg, 1985) enables us to visualize multidimensional data into the 2D-space (Tilouche et al., 2017). While parallel coordinates plot facilitates multivariate data exploration, only a few of classification studies are built on the basis of the potential power of this multivariate visualization technique (Xu et al., 2007). There are studies that contain a parallel coordinates illustration of the result, but they need a human administrative to constantly monitor the system and detect the drifts and abnormalities (Azhar and Rissanen, 2011; Choi et al., 2009).

To the best of our knowledge, none of the novelty detection classifiers utilize the capability of a multivariate visualization technique. This study considers the problem of novelty detection, presenting an algorithm based on the parallel coordinates plot combined with K-medoid clustering algorithm, named *NDTool*. There is no need to monitor NDTool for detecting the novel behaviour as it provides the report of detected novel observations. This essential characteristic prevents the human error in the process of detecting novel behaviours. In order to let human sight infer to the detected cases intuitively, visual supports are also provided via the algorithm. This is a help for the human brain visual analysis power to gain a deeper insight into the novel behaviors. The ultimate goal of this study is creating an innovative novelty detector tool that does not require specific machine learning knowledge to use it besides having an eye on industrial use.

1.2 A brief overview of the proposed methodology

To meet the stated goal, we present a novelty detection tool (NDTool) based on parallel coordinates combined with a clustering algorithm. NDTool gets the historical "normal" data without a need to have abnormal samples. The new coming observations will be checked and illustrated via multidimensional data visualization plot. In the case of detecting a different behavior compared to the most similar cluster in historical data, it will alert along with the exact address of detected novel behaviors (row and column).

The performance of NDTool is evaluated by comparing the obtained result with other novelty detection algorithms. We report the results of experiments conducted on three well-known datasets, derived from the University of California – Irvine Machine Learning Repository (UCI); "Ionosphere", "Spambase" and "Breast Cancer". These results allow us to compare the quality of solutions in term of AUC (area under ROC curve) produced by NDTool and with state-of-art algorithms presented in the literature, including SVDD, K-NN, GM and K-means based algorithm.

Finally, NDTool is used in breast cancer studies. It diagnoses whether a sample breast tumor is benign or malignant. The motivation of choosing breast cancer data is the fact that it is the most common cancer in women worldwide. According to the reports of Canadian Cancer Society on Cancer Statistics, it is estimated that $26,300$ women in Canada will be diagnosed with breast cancer in 2017 (Canadian Cancer Society, 2017). The early diagnosis of cancer increases the chances of a complete treatment. So any study that accelerates the process of cancer diagnosis can save patients' lives. Two different breast cancer datasets (WBCD) are chosen from the University of California at Irvine (UCI) machine learning repository in order to show the abilities of NDTool in the era of the cancer diagnosis.

The reminder of this study is structured as follows: Chapter 2 gives the review of related literature. Chapter 3 describes the developed algorithm. Chapter 4 presents the performance of the proposed algorithm on different benchmark datasets compared to different methods. Section 5 provides a discussion and Chapter 6 contains the conclusions and limitations of this study and also proposes the future work.

CHAPTER 2 LITERATURE REVIEW

Novelty detection could be considered as *one-class classification* problem because all possibilities are compared with one class (Moya et al., 1993). The literature states one-class classification, *anomaly detection* and *outlier detection* are the terms used interchangeably with novelty detection. It is due to the fact that their solutions and methods are often the same (Ding et al., 2014). But to clarify their subtle differences, some definitions referred to the literature are provided as follows:

According to Hawkins (Hawkins, 1980) "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism". Anomalies are the observations that are not compliant with expected behaviour (Chandola et al., 2009). Novelty detection is also concerned with identifying the observations that are not consistent with the expected behaviours. Detecting the novelties is based on "normal" data and the algorithm looks for the unobserved data points that are not available during training. Generally, the novelty detection approaches are applied when the number of abnormal data is not sufficient for learning their behaviour (Pimentel et al., 2014).

In summary, we can conclude that an exact definition of the problem depends on the assumptions regarding the data structure and the goal of detecting them. For example, the outlier are not of interest and may consider as some noises in data with destructive effect on the results. These outliers may be removed in the data preprocessing. While, anomalies are of great interest and the goal is to detect and analyze them. Novelties are also the points that have great significance.They are of great interest due to their valuable information, like anomalies. But in a novelty detection scenario, we need to detect the unknown cases just based on "normal" observations.

During recent decades various methods have been proposed for novelty detection problem. We refer the readers to the surveys provided by Ding et al. (2014), Pimentel et al. (2014) and Markou and Singh (2003). Ding et al. (2014) categorizes the novelty detection methods into two main approaches: semi-supervised and unsupervised training.

Semi-supervised training approach employs artificially generated abnormal data as well as "normal" data. Blanchard et al. (2010) developed a statistical-based semi-supervised approach which assigns an upper threshold on the false alarms. Their method assumes that the novelty samples are available during the training phase. Surace and Worden (2010) applied the negative selection algorithm for novelty detection problem which is inspired by the function of immune system in human bodies.

One of the most common techniques for novelty detection is *support vector machines* (SVMs) (Schölkopf et al., 2000) which is applied for both unsupervised and semi-supervised approaches. Support Vector Machines separate a dataset into multiple classes via decision boundaries. *Support vectors* are the data points in training data that are close to the boundary and define the separating margin; See Figure 2.1. SVMs look for a decision boundary which separate the majority of "normal" points, with a maximum margin between "normal" and abnormal observation in a multidimensional space. Any data point placed out of this boundary is detected as abnormality. Then, class membership of a new observation is determined according to its location versus the defined boundary. If it is placed outside the defined boundary, it will be labeled as abnormal.

De Morsier et al. (2013) presented two methods for semi-supervised novelty detection problem, based on a SVM methodology named "cost sensitive Support Vector Machine" which assigns different error costs for both classes in order to simplify the process of parameter selection.

The main weak point of semi-supervised novelty detection approach is the fact that their performance depends upon the quality of the collected abnormal data; despite the fact that obtaining abnormal data is a big challenge in the real-life applications of novelty detection. For example, for monitoring the health of aircraft, which abnormal data means a crash, it is not possible to destroy them for the sake of collecting abnormal data (Hayton et al., 2007).

Unsupervised training methods only employ unlabeled "normal" data without using the abnormal samples. The most wildly-used methods in unsupervised category are Support Vector Machines, K-nearest neighbor, probability density based methods and clustering based method.

Support vector data description (SVDD), proposed by Tax and Duin (2004), aims to define a hyperplane boundary with minimum volume that surround most of the "normal" data (or even all of them) in training phase. This method has a more flexible spheres compared to original SVM. A new observation is labeled abnormal, if it lies within the mentioned hypersphere.

K-nearest neighbours (K-NN) approach is another well-known unsupervised solution for novelty detection (Pimentel et al., 2014). In K-NN, the majority votes of neighbors determine if the new data is a member of them or not. This technique assumes that normal data lie near their neighbours, while abnormal observations are located far from their neighbours. In fact, the distance between a new observation is compared with its k nearest neighbours in training dataset. For example if $K = 1$, then we have a 1-NN problem. See Figure 2.2 for illustrated example of 1-NN. Let d_1 the distance between a new observation y and its nearest neighbour

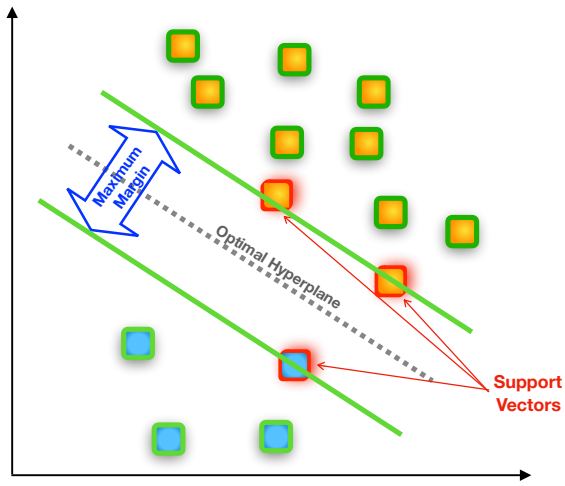


Figure 2.1 Support Vector Machines

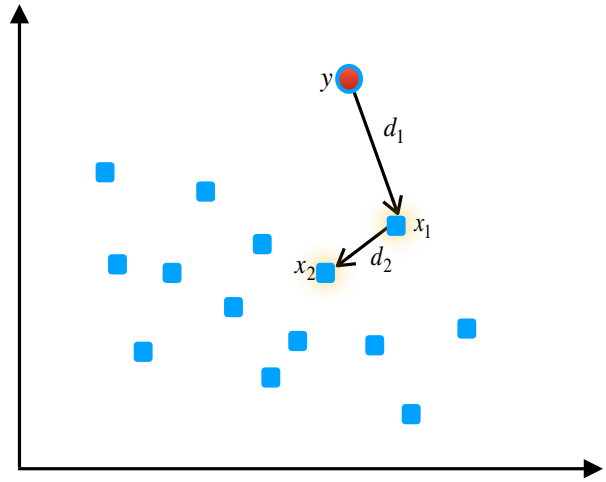


Figure 2.2 k-NN

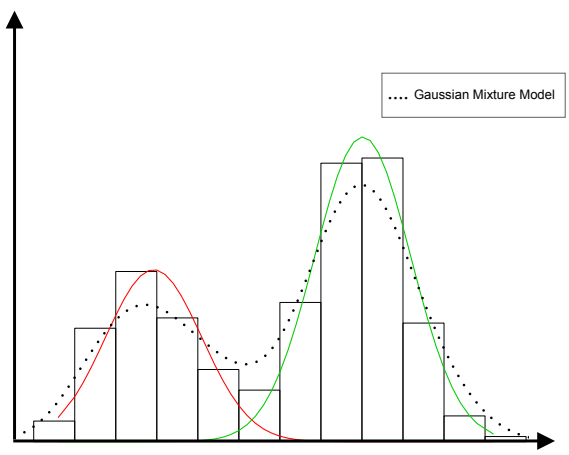


Figure 2.3 Gaussian mixture models

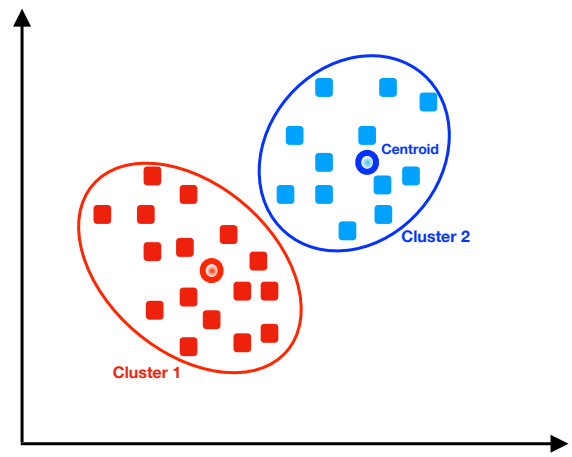


Figure 2.4 K-means

x_1; and d_2 the distance between the point that is assigned as the nearest neighbour of new data (x_1) and its nearest neighbour x_2. The new observation (y) is diagnosed as "normal" if d_1 is less than or equal to d_2. Otherwise, y is labeled as a novelty. For more information one can be referred to Duda et al. (1973) and Cunningham and Delany (2007). However, the performance of K-NN based methods decrease for high-dimensional datasets (Pimentel et al., 2014).

Gaussian mixture models (GMM) is another popular approach in studies for novelty detection (Pimentel et al., 2014). GMM is a parametric probability density function made up of the combination of Gaussian component densities (Figure 2.3). GMM takes less number of kernel

than existing pattern in the training phase. GMM in novelty detection problem estimate the probability density of the "normal" class. A major disadvantage of GMM for novelty detection problem is the high dependence of the algorithm to the sufficient training samples. The other disadvantage is reducing its performance by increasing the number of dimension in datasets.

In clustering-based method, *K-means* is one of the most popular as it is easy to understand and easy to implement. For solving a novelty detection problem by K-means, the training data need to be divided to k clusters – See Figure 2.4. Each cluster is defined by a prototype named *centroid* in k-means algorithm. Then if the test data fall into any of these clusters, it will be considered as a "normal" observation. A major disadvantage of K-means is its sensitivity to the noise or outliers.

The solution for novelty detection problem may seem difficult to understand in real world situation. For example in healthcare era, the physician can take advantage of a novelty detection algorithm, but lacking the knowledge of machine learning could be restrictive. Another similar condition may happen in industrial monitoring which is an application of novelty detection. It seems that in the literature, there is a lack of connecting theoretical approach of novelty detection to an easy to understand way of data analysis techniques. What's more, to the best of our knowledge, none of the above-mentioned categories contains the visualization-based approaches for novelty detection. This study presents a novelty detection tool that follows the unsupervised approach; as it doesn't need any abnormal samples during its training phase. The proposed tool is based on combination of the potential geometric power of parallel coordinates plot and K-medoids clustering method.

CHAPTER 3 MATERIALS AND METHOD

In this chapter, we describe proposed novelty detection tool (NDTool) which consists of two major components: parallel coordinates plot and K-medoids clustering algorithm. In Section 3.1 and 3.2 we describe the parallel coordinates plot and K-medoids clustering algorithm, respectively. Finally, Section 3.3 explains the proposed algorithm.

3.1 Parallel coordinates

Visualizing multidimensional data has huge effects on data analysis and decision making. *Parallel coordinates* (Inselberg, 1985), is a powerful plot that enables us to visualize multi-dimensional data into a two-dimensional space. This plot has been the focus of attention of many researchers. We refer the reader to (Johansson and Forsell, 2016) and (Heinrich and Weiskopf, 2013). These plots are not complex while there is no theoretical limit to the number of dimensions that can be visualized. So it is possible to add up new dimensions to the plot as far as necessary. In fact, the number of variables that can be displayed is only limited by the horizontal resolution of the display device (Choi et al., 2009).

Parallel coordinates structure consists of two major components: parallel vertical axes representing the features within the data and polylines which represent data rows.

Figure 3.1 illustrates the parallel coordinates plot of Ionosphere dataset (Lichman, 2013). This dataset, which is investigated in our experiments in Chapter 4, is about the radar returns from the ionosphere.

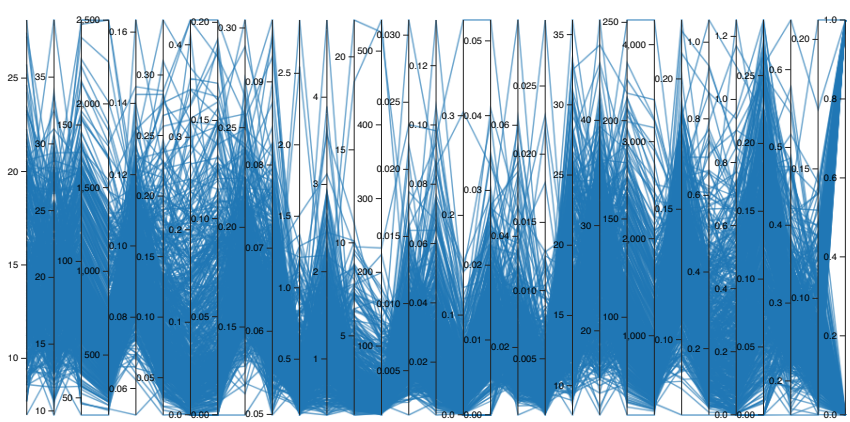
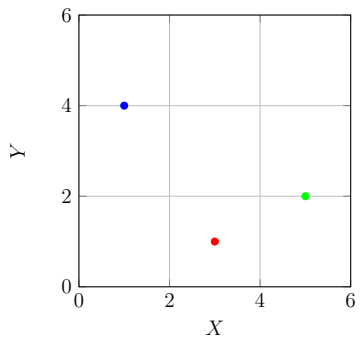


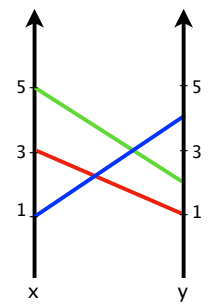
Figure 3.1 Example of parallel coordinates plot for Ionosphere dataset.

For each data point, the position of polyline on the i-th axis corresponds to the scalar value

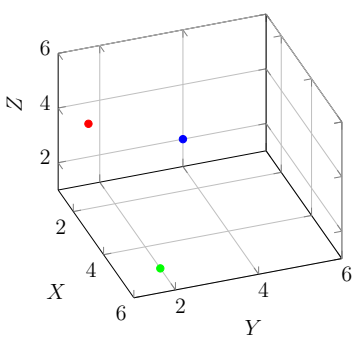
of the i-th dimension. Figure 3.2 (a) to (d) illustrate the equivalent display of the two-dimensional and three-dimensional space into the parallel coordinates charts. Figure 3.2 (e) displays the lack of a four-dimensional illustration while it is presented by a parallel coordinates plot, simply and efficiently conveying the concept; see Figure 3.2 (f).



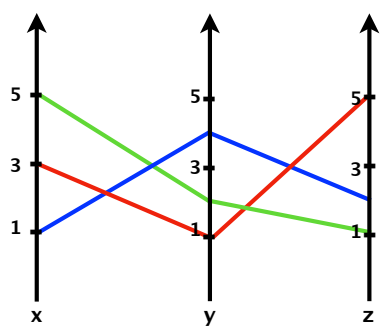
(a) 2D space



(b) Parallel coordinates plot of 2D space



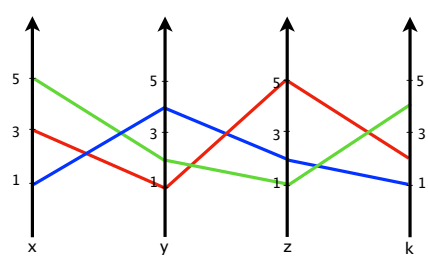
(c) 3D space



(d) Parallel coordinates plot of 3D space

$$\begin{bmatrix}5\\2\\1\\4\end{bmatrix} \quad \begin{bmatrix}3\\1\\5\\2\end{bmatrix} \quad \begin{bmatrix}1\\4\\2\\1\end{bmatrix}$$

(e) 4D points



(f) Parallel coordinates plot of 4D space

Figure 3.2 Illustrating Cartesian coordinates and their parallel coordinates plots

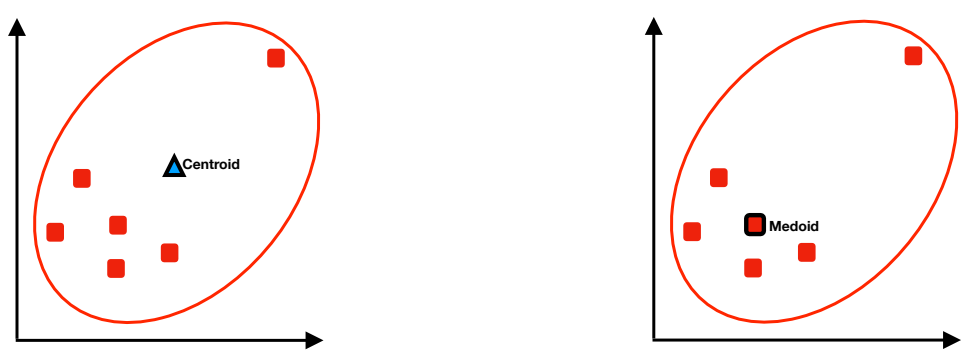
Our search through the relevant literature yielded there are few studies that are mainly based on the potential of parallel coordinates for classification. For example, Xu et al. (2007) proposed a visual classifier using parallel coordinates combined with decision trees and linear discriminant analysis. Andrienko and Andrienko (2001) used parallel coordinates just as a

visualization support for similarity-based classification. Edsall (2003) linked parallel coordinates plot to the map for geographic data exploration. Azhar and Rissanen (2011) proposed using parallel coordinates plot for analyzing the alarms based on human visual perception with the purpose of detecting false alarms versus true alarms. They set up an experiment involving 12 participants with an engineering background to use a parallel coordinate plot of alarms and declare their diagnosis. Undoubtedly these methods that are dependent on human are not safe from the human error. In this study, we investigate on parallel coordinates plot as an important aspect of NDTool for detecting the novel behavior and also visualizing them without a need for human intervention.

3.2 Clustering

NDTool benefits from *K-medoids* which is a partitioning-based clustering algorithm. For a review of different clustering algorithms, one can be referred to Berkhin (2006).

K-medoids (Kaufman and Rousseeuw, 1987) is a version of *K-means* clustering algorithm. K-means and K-medoids are usually the initial choices of clustering because they are easy to implement and easy to understand. Like the other partitioning algorithm, K-medoids constructs K partitions and each partition has a representative. It uses the actual data point, called *medoid*, as the representative of the cluster rather than using the conventional mean or centroid for each cluster. A medoid is the most centrally located point in its respective cluster. Unlike other partitioning algorithms, K-medoids is robust toward the existence of outliers and noise compared to its similar algorithm, K-means (Jin and Han, 2011). For more clarity, a visual example is provided for K-means and K-medoid in Figure 3.3(a) and 3.3(b), respectively. If the data contains outlier, it highly affects the place of the centroid in



(a) in K-means (b) Medoid in K-medoid

Figure 3.3 Difference between K-means and K-mediods for assigning a representative for a cluster which contains outlier

K-means. With the same condition, the place of medoid in K-medoid is more representative of the cluster members.

K-medoids starts with K random medoids and assigns the data points to their closest cluster medoid. In each iteration, a medoid (m) and a random data point (x) associated to m are swapped, if the objective function 3.1 can be reduced.

$$S = \sum_{i=1}^{N} min(d(x_i, m_1), ..., d(x_i, m_k)) \tag{3.1}$$

Where:
x_i: data point i
m_i: medoid of cluster i
N: Number of data points in the dataset

The algorithm repeats the aforementioned steps alternatively until the minimum amount of S can no longer be decreased.

K-medoid clustering technique is used in NDTool in order to divide the training data into multiple groups (clusters) with the most similar data points. The strategy toward a new observation is to find the best possible cluster for that.

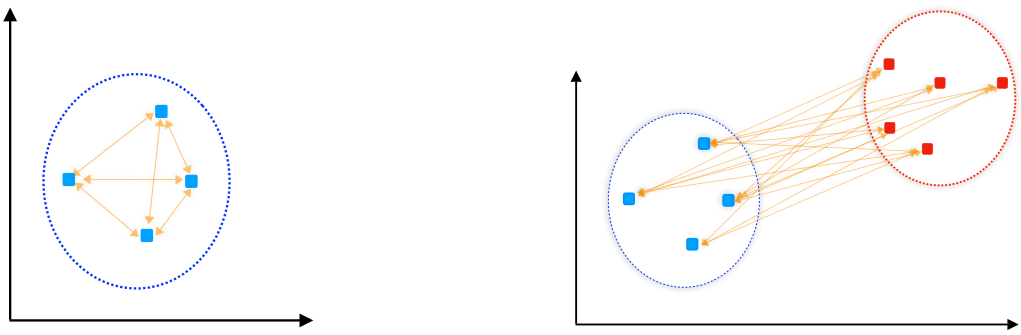
3.2.1 Average silhouette method

Clustering algorithms like k-means and k-medoid are unable to determine the number of clusters (K). There are some methods to determine the number of clusters. Silhouette method (Rousseeuw, 1987), is a popular method for determining the optimal number of clusters. This method is based on the concepts of cohesion and separation. Cohesion is the compactness within each cluster and separation indicates the external separation of the clusters. Figure 3.4 illustrates the concepts cohesion in a cluster and separation between two clusters.

For each data point i, the silhouette value is calculated via Formula 3.2. The average silhouette value for the whole dataset is calculated by Formula 3.3.

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{3.2}$$

$$\bar{s} = \frac{\sum_{i=1}^{N} s(i)}{N} \tag{3.3}$$



(a) Cohesion between cluster members (b) Separation between different clusters

Figure 3.4 Cohesion versus Separation in clustering

Where:

$a(i)$: Average distance of point i from all other objects in its correspondent cluster

$b(i)$: Minimum average distance of point i from the objects of other clusters

N: Number of all data points in the dataset

The amount of $s(i)$ for each point i is in range of $[-1, 1]$ and shows how well it fits into the correspondent cluster. The values near to one indicate a high homogeneity and the negative values indicate that the data points are not fit to the assigned cluster. So a negative $s(i)$ indicates a low quality clustering.

The average silhouette ($\bar{s}$) for different values of K are calculated. Since $\bar{s}$ indicates the quality of clustering, it can be used for the selection of the optimal number of cluster. In this study, K varies from 2 to 10 clusters that were chosen based on the results of preliminary tests. The maximum amount of average silhouette specifies the appropriate number of clusters for each given dataset.

Figure 3.5 [1] illustrates the cluster silhouette plot for the all "normal" observations of Ionosphere dataset. Each vertical pixel indicates a $s(i)$ for each data point. The red dotted line is the average silhouette value for the all the points which is 0.46 for $K = 3$.

NDTool contains K-medoids algorithm, which needs the number of clusters as an input. So by employing Silhouette factor algorithm, the process of finding an optimal number of cluster is automated.

[1] For better viewing the figures, please see the electronic copy of the thesis.

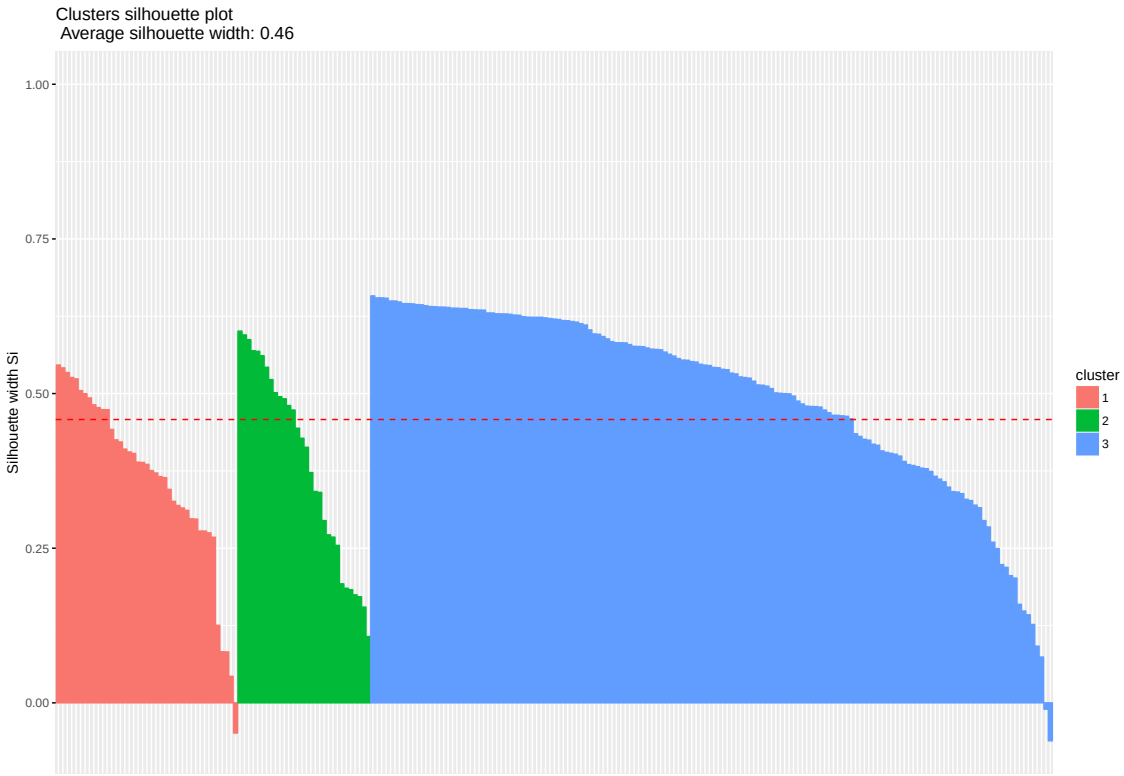


Figure 3.5 silhouette factor plot of "normal" observations of Ionosphere dataset, when the number of clusters is three..

3.3 NDTool

In previous sections, we introduced the components that are applied in NDTool. In this section, we explain the algorithm of NDTool which consists of four main phases. Figure 3.12 shows the flowchart of NDTool. Here, we describe each phase step by step as follow:

Phase I: NDTool needs a historical dataset (M_{mn}) from "normal" expected events. Let the data matrix is given by:

$$M = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix} \tag{3.4}$$

where m is the number of observation and n is the number of dimension. In the preprocessing step, the rows contain missing values are removed. Since the range of values of data may vary widely, it can affect the performance of NDTool. Hence, the given dataset is standardized using the mean and standard deviation of each dimension.

For given matrix M, the standardization formula is defined as:

$$Z(x_{ij}) = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{3.5}$$

$$\mu_j = \frac{\sum_{i=1}^{m} x_{ij}}{N}, \sigma_j = \sqrt{\frac{\sum_{i=1}^{m}(x_{ij} - \mu_j)^2}{N}} \tag{3.6}$$

where μ_j and σ_j are the mean and standard deviation of the *j-th* attribute, and N is the number of observations.

The next step is clustering the standardized historical dataset by K-medoid algorithm as described in Section 3.2. This algorithm minimizes the sum of dissimilarities between data points in the cluster and the representative point of that cluster which is called *medoid*. Like other partitional clustering methods, K-medoids needs the initial amount of K. NDTool is able to define the best number of cluster for the given dataset via Average silhouette method as described in Subsection 3.2.1. In this way, the operator doesn't need to choose a random number of cluster.

Phase II: After identifying the clusters, NDTool calculates the lower bound and upper bound of each dimension for each cluster. Let U_j^c and L_j^c indicate the lower bound and upper bound of dimension j, respectively, in cluster c. Figure 3.6 shows a parallel coordinate plot with three dimensions and two clusters. For example in this Figure, vectors $U = [U_i^1, U_j^1, U_k^1]$ and $L = [L_i^1, L_j^1, L_k^1]$ indicate the upper bound and lower bound for each dimension in the cluster 1 (red cluster), respectively.

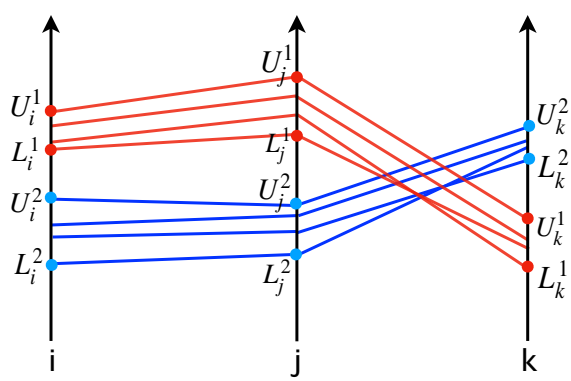


Figure 3.6 Example of finding upper bound and lower bound in each dimension for each cluster.

Then for each cluster, all existing angles of each dimension are found via calculating tangent.

The tangent of an angle at each dimension is calculated by:

$$tan\theta = \frac{opposite}{adjacent} \tag{3.7}$$

Figure 3.7 shows how we use the concept of tangent in NDTool. The adjacent side is equal for all axis, so we consider it as 1. Accordingly, the angles are represented by their opposite lines. The longer opposite side implies the wider angle. The range of acceptable tangent for each cluster is belonged to the maximum and minimum amount of tangent.

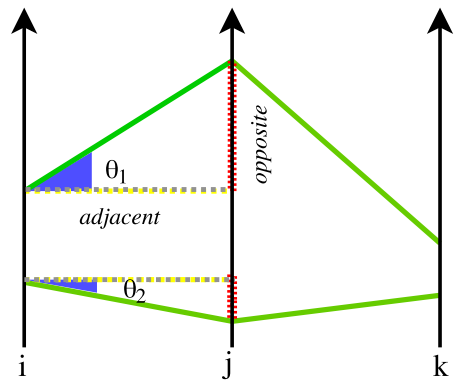
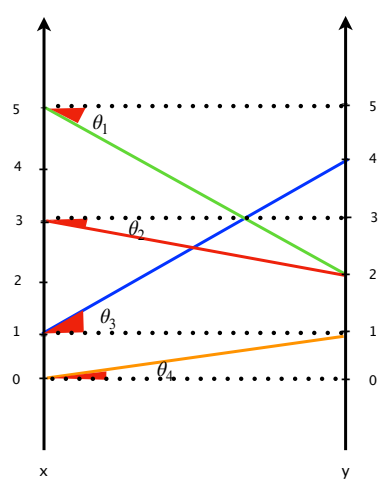


Figure 3.7 Example of finding tangent of the angles in each dimension.

Figure 3.8 is a numerical example of calculating the angle rage in a cluster, for a considered dimension. In this example, the angle range is $[-3, 3]$. Any newcomer data row, with an angle outside this range, is considered as a non-conformity.



$$tan\theta_1 = \frac{2-5}{1} = -3 \quad \textbf{MIN}$$

$$tan\theta_2 = \frac{2-3}{1} = -1$$

$$tan\theta_3 = \frac{4-1}{1} = 3 \quad \textbf{MAX}$$

$$tan\theta_4 = \frac{1-0}{1} = 1$$

Figure 3.8 Numerical example of calculating tangents of the angles in a dimension.

Phase III: At this phase, NDTool calls a new row of data. This newcomer data should be standardized using the mean (μ_i) and standard deviation (σ_i) of historical data . After that, the euclidean distance of the newcomer data and each medoid are calculated. Then the newcomer data candidates for becoming a member of the cluster with the most nearest medoid. For more clarity, following example is given:

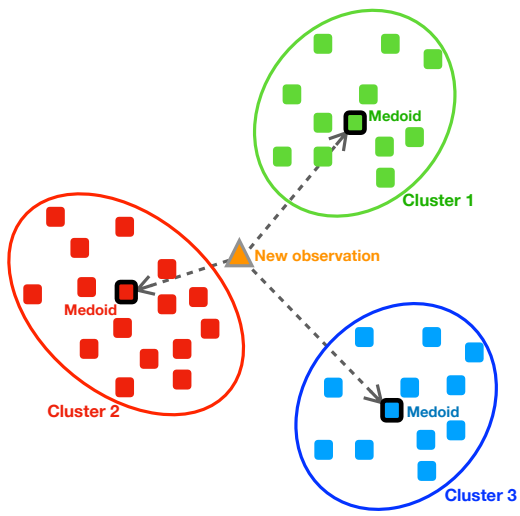


Figure 3.9 A new observation and clustered data by K-medoid

Figure 3.9 shows three cluster with their medoids and a newcomer data. Let M^i indicate medoid of cluster i. Now, we calculate the Euclidean distance of the new observation (Y) from each medoid using Equation 3.10.

$$Y = (y_1, y_2, ..., y_n) \tag{3.8}$$

$$\begin{aligned} M^1 &= \left(m_1^1, m_2^1, ..., m_n^1\right) \\ M^2 &= \left(m_1^2, m_2^2, ..., m_n^2\right) \\ M^3 &= \left(m_1^3, m_2^3, ..., m_n^3\right) \end{aligned} \tag{3.9}$$

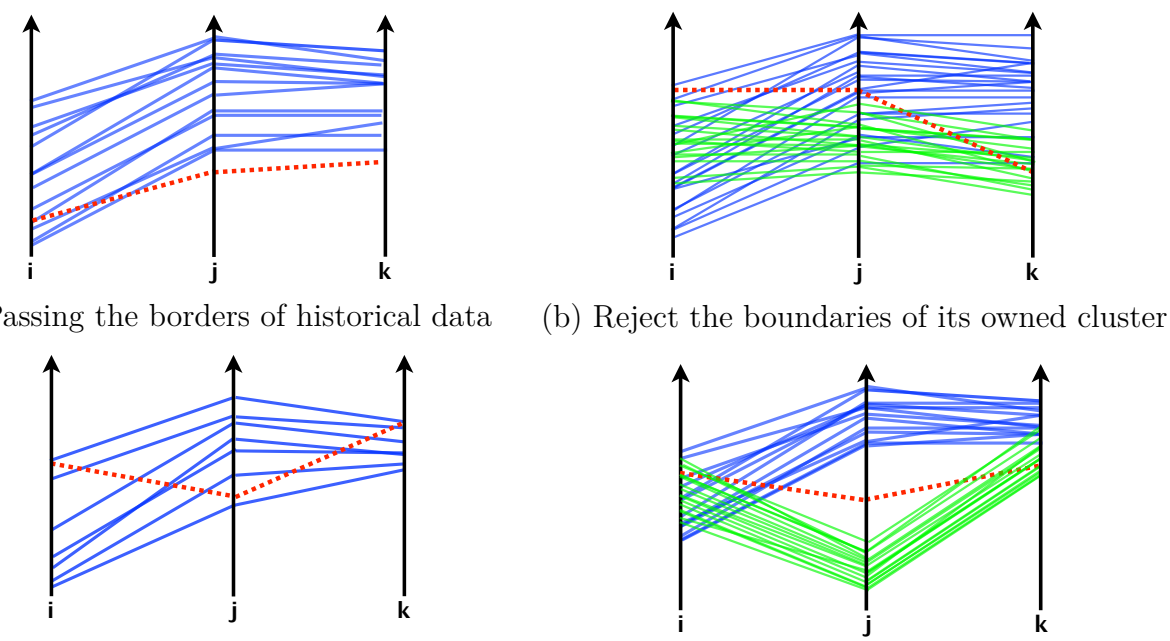
$$t^k = \sqrt{\left(y_1 - m_1^k\right)^2 + \left(y_2 - m_2^k\right)^2 + ... + \left(y_n - m_n^k\right)^2} \quad \forall k \in \{1, 2, 3\} \tag{3.10}$$

where k is the index of each cluster. Without loss generality, we assume that the Euclidean distance of the new data (Y) from the medoid M^2 is less than the other medoids, i.e., $t^2 = min\{t^1, t^2, t^3\}$. Hence, the new data is candidate for being a member of the cluster 2.

Phase IV: After going through the above steps, NDTool starts the phase of evaluating newcomer rows. It checks the new data, row by row, as long as they exist.

In this phase, each dimension of the newcomer is compared with the relevant cluster in the correspondent dimension. Let c indicate the relevant cluster of newcomer data and $[L_j^c, U_j^c]$ be lower bound and upper bound of cluster c in dimension j. If jth dimension of newcomer data doesn't belong to $[L_j^c, U_j^c]$, it is considered as a candidate of novel behaviour.

For an explicit explanation, we provide visual examples. The illustrated cases in Figure 3.10 are the events that may cause a new observation to be considered as a novel behavior. Figure 3.10 (a) shows a newcomer data which pass the lower bound of the standardized historical data. This event is detected by comparing the lower bound and upper bound values versus the newcomer's values over each dimension. Figure 3.10 (b) shows an observation which is between the lower bound and upper bound of the whole standardized historical data. But it passed the upper bound of its correspondent cluster which is in green.



(a) Passing the borders of historical data

(b) Reject the boundaries of its owned cluster

(c) Noncompliance with the trend of historical data

(d) Crossing the enclosed cavity

Figure 3.10 Visualizing novel behaviours in parallel coordinates plot

In another case illustrated in Figure 3.10 (c), the newcomer data may lie in the frame of its own cluster without passing any lower bound and upper bound, nevertheless it behaves differently versus the historical dataset. For detecting such kind of novel behaviour, NDTool uses the angles of data in each dimension. In each dimension, the angle of newcomer data is compared with the acceptable angle range (see phase II) in the corresponding cluster.

In Figure 3.10 (d), the new data passed the enclosed cavities inside the range of historical dataset. Passing through such spaces considers as an novel behaviour versus to the historical data. This newcomer data doesn't respect the upper bound or lower bound of its candidate cluster.

If one of the above-mentioned conditions are not satisfied, the same process from the beginning of phase IV will be done for the next nearest clusters to the newcomer row. If the newcomer was not compatible to any clusters, it is labeled as a novel behaviour. NDTool announces the existence of a novelty by making an alarm while a parallel coordinates of the novel observation is stored in the system. According to the information that NDTool provided about the exact place of novel behaviour (relevant columns and the type of behaviour), it is much easier for the operators to have a fast and detail diagnosis in a glance.

For example, Figure 3.11 provides an example of a parallel coordinates plot of a novel behaviour and the "normal" historical data for Diagnostic Breast Cancer dataset [2]. This dataset is investigated in our experiments in Chapter 4. The novel behaviour is in violet and the clusters of historical data are in three different colors (red, green, blue).

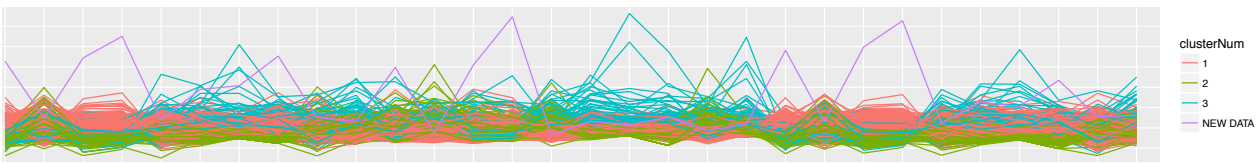


Figure 3.11 Parallel coordinates of a newcomer data versus the standardized historical "normal" data for Diagnostic Breast Cancer dataset

After each working period (a day, a week or even a group of patients), NDTool provides a *pie chart* which shows the portion of each attribute in making novelties. This pie chart helps to analyze the obtained results via defining the role of each attribute in the creation of novel behaviours. More detailed information with an example of pie chart related to the Diagnostic Breast Cancer dataset is provided in Section 4.4.

The summary of all phases is given in the flowchart of NDTool; illustrated in Figure 3.12. In this flowchart, set C indicate set of clusters of the historical data.

[2]For better viewing the figures, please see the electronic copy of the thesis.

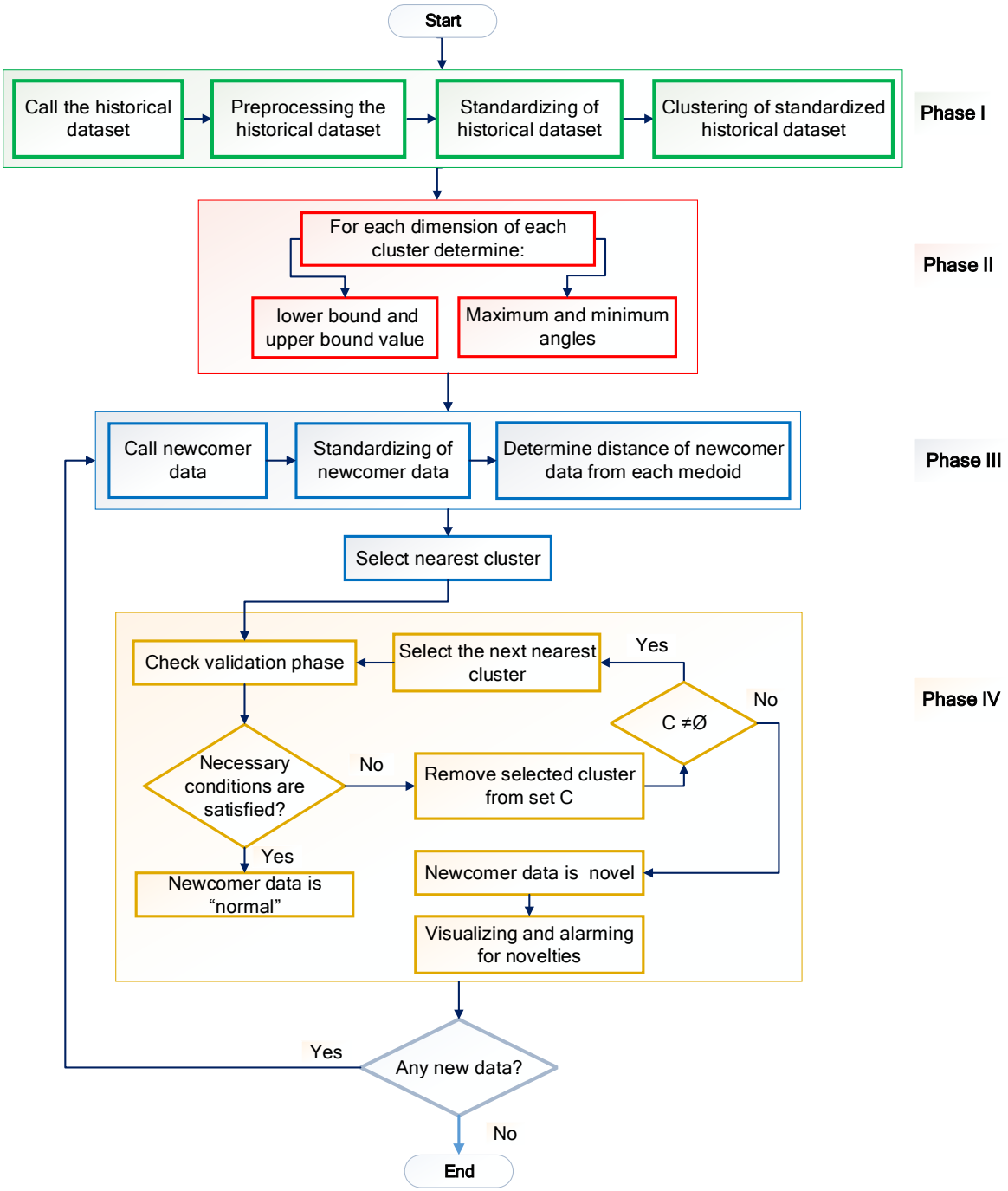


Figure 3.12 Flowchart of NDTool

CHAPTER 4 NDTOOL PERFORMANCE

This chapter presents the experiments carried out on the proposed novelty detector: NDTool. Section 4.1 introduces the well-known performance measures for novelty detection problem. The data description is given in Section 4.2. Then in Section 4.3, the performance of NDTool is evaluated via represented measures on the benchmark datasets. The results are compared with the most popular methods presented in the literature (Ding et al., 2014). At the end in Section 4.4, we applied NDTool for detecting cancerous mass in two datasets in breast cancer studies. Our implementation is coded in RStudio, version 1.0.136.

4.1 Performance Measures

As mentioned before, novelty detection is a binary classification problem (also known as one-class classification) with the same evaluating measures as classification system (Khan and Madden, 2009). Here, we introduce the popular evaluation criteria for one-class classifiers which are as follows: Accuracy, Precision, Sensitivity, AUC (Area Under ROC Curve). All of these criteria are founded on Confusion Matrix (Fawcett, 2006); see Figure 4.1. In the following, we describe them:

Confusion Matrix is the foundation of above-mentioned measures. In fact, the obtained information from Confusion Matrix are applied for making metrics to assess the performance of a classifier. In general, a Confusion Matrix contains four scenarios which are derived from comparing the classification results and the reality of the observation.

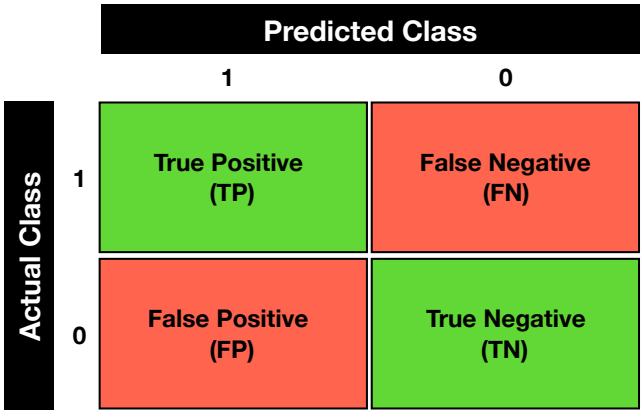


Figure 4.1 Confusion Matrix

Two of these scenarios which are in the first row of Confusion Matrix happen when an observation is from class 0 [1]. If it is labeled 0, it is called *true negative* (TN) and if it is labeled as 1, it is called *False Positive* (FP). Two other scenarios are in the second row of Confusion Matrix which happen when an observation is from class 1 [2]. If the classifier labels it as 0, it is called *False Negative* (FN) and if it is labeled as 1, it is called *True Positive* (TP).

As mentioned above, there are measures driven from Confusion Matrix which are defined as follows:

Accuracy (Acc) indicates the closeness of prediction to the reality by calculating the ratio of correct prediction to the total number of conducted prediction. It is calculated by Formula 4.1.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Negative} + \text{True Positive} + \text{False Positive} + \text{False Negative}} \tag{4.1}$$

Precision, calculated by Formula 4.2. To make it more clear, Precision is the ratio of the number of rainy days that are correctly predicted to the number all days that we make a rain prediction. The more False Positive causes the lower Precision rate.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{4.2}$$

True Positive rate (sensitivity) and False Positive rate are calculated by Formula 4.3 and 4.4, respectively. True Positive rate shows the ability of a classifier in detecting the abnormal samples. For example, in weather condition studies, Sensitivity shows the ratio of rainy days which are correctly predicted as rainy to the number of all rainy days.

$$\text{True Positive rate} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{4.3}$$

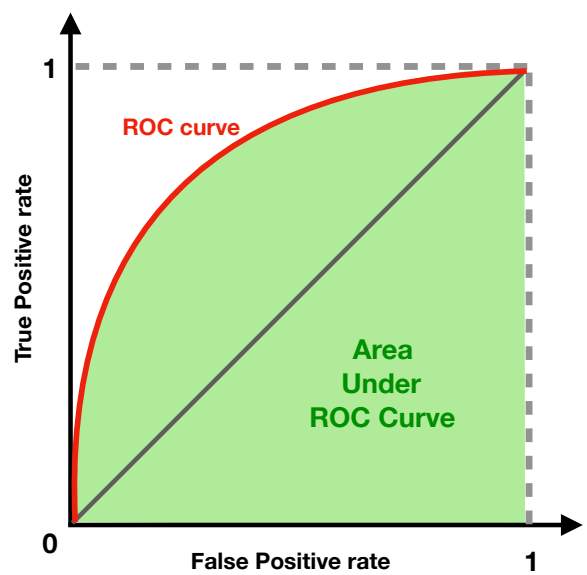
$$\text{False Positive rate} = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}} \tag{4.4}$$

Accuracy, Precision, Sensitivity and False Positive rate take a value between 0 and 1 and they can also be presented as a percentage.

[1] Class 0 indicates the "normal" data.
[2] Class 1 indicates abnormal data.

Receiver operating characteristics (*ROC plot*) is a well-known performance metric for classification. The name of "Receiver Operating Characteristic" is taken from Signal Detection Theory during second world war. At that time, radar waves were used to detect enemies. The radar operator had to distinguish whether the spots on the radar screen are enemies, friend, or just a noise (like birds). So, signal detection theory was used as a measure for evaluating the ability of receiver operators of radars in such sensitive distinction. Their detection ability was called the Receiver Operating Characteristics. Afterward, in the 1970's, signal detection theory was applied for medical test and diagnosis (Sonntag, 2010).

The Area Under the ROC Curve (*AUC*) is the most commonly used measure of evaluating the one-class classifiers; see Figure 4.2. ROC curve is a graphical display of the trade-off between True Positive rate (Sensitivity) on the Y-axis and False Positive rate on the X-axis for every cut-off of a test; or better said the trade-off between benefit and cost. It takes value between 0 and 1. The bigger values indicate better performance. In other words, a ROC plot near to the upper left indicates a better performance of the classifier. One can be advised to refer to detailed literature on ROC analysis by Fawcett (2006).



.7

Figure 4.2 Receiver Operating Characteristics curve

4.2 Data description

The experiments are conducted on three well-known datasets derived from the University of California – Irvine Machine Learning Repository (UCI). The datasets are selected in different size of dimension and observation. The characteristics of each dataset are given in Table 4.1. The last column of this table indicates if there exists missing value in observations or not. NDTool contains a preprocessing phase which ignores the observations with missing value.

Table 4.1 The benchmark datasets

	Number of		
Datasets	Dimensions	Observations	Missing Values
Ionosphere	34	351	No
Spambase	57	4601	Yes
Breast Cancer	9	286	Yes

The definition of each dataset are described as follow:

Johns Hopkins University Ionosphere is a dataset collected from the radar system that is located in Goose Bay. This data is prepared by the Applied Physics Laboratory of Space Physics Group at Johns Hopkins University return the information about the condition of ionosphere. The ionosphere is the highest layer of the earth's atmosphere. It consist of electrons and electrically charged atoms and molecules. This layer absorbs the dangerous rays of the sun and prevents them from entering the earth like a ceiling to allow life on Earth. Instances in this database are belong to two classes: "Good" with 225 and "Bad" with 126 observations. Good radar returns indicate the existence of expected structure in the ionosphere, while Bad returns are those observation that their signals pass through the ionosphere.

Spambase E-mail is a dataset containing 4601 observations with attributes that indicate if the e-mails contain some special words or not. Spambase dataset is composed of two classes: Spam with 1813 observations and not Spam with 2788 observations.

The other benchmark is a *breast cancer* dataset which is obtained from the University Medical Centre at institute of oncology of Yugoslavia. It is consist of 9 attributes, 286 observation and two classes. One class of no-recurrence events with 201 instances and the other class contains recurrence events with 85 instances.

4.3 Comparative study with state-of-art algorithms

The comparative study investigates performance of the NDTool against other state-of-art algorithms on a set of well known benchmarks. NDTool is compared with SVDD, K-NN, GMM and K-means based, in terms of the AUC, presented in Section 4.1, which is the most common measure for evaluating the performance of classifiers. The experimental result for SVDD, K-NN, GMM and K-means based algorithms are provided by Ding et al. (2014). We also provide the experimental result of the other measures: Accuracy, Precision and Sensitivity.

The result of these experiments are reported in Table 4.2. To make a fair comparison between proposed algorithm and studied algorithms, for all datasets, we separate the class 0 ("normal") and class 1 (abnormal) samples while the training set is consist of 50% of class 0 samples which is chosen randomly and called as historical dataset. The test data set consist of the rest of class 1 samples (50%) in addition to all class 1 samples; see Figure 4.3.

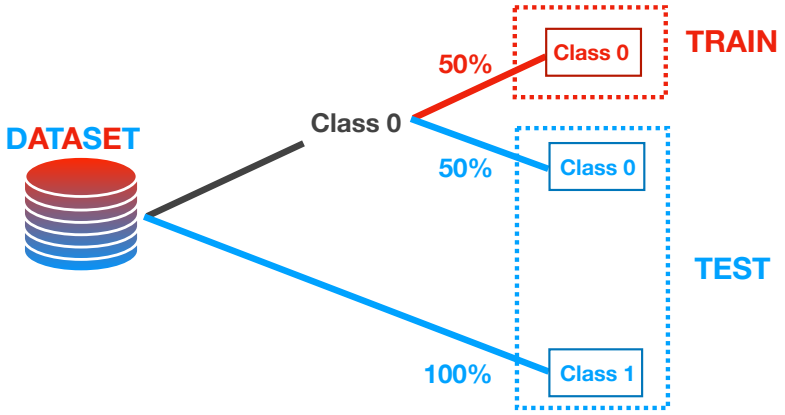
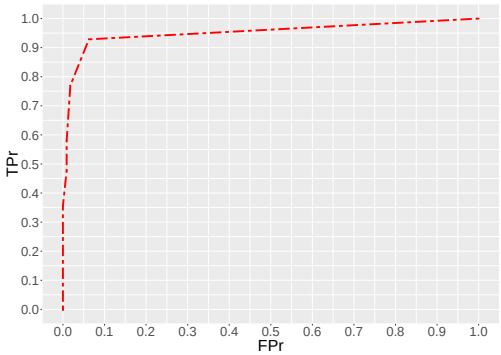


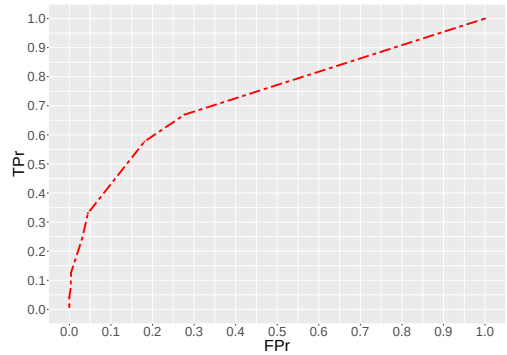
Figure 4.3 Strategy of dividing the dataset into train and test

We applied an innovative method to obtain the trade-off between True Positive rate and False Positive rate in ROC plot. We put a threshold for the range of $[0, n]$, where n is the number of dimension of a historical data. A threshold equals to 0 indicates that NDTool should alarm even it finds no error in the newcomer data. So the ROC curve meets the point $(1, 1)$. Next, for a threshold equals to 1, NDTool alarms as soon as it finds an error in newcomer data, and so on. In the other hand, a threshold equals to n indicates a condition where NDTool alarms when there are errors in all dimensions of a newcomer data.

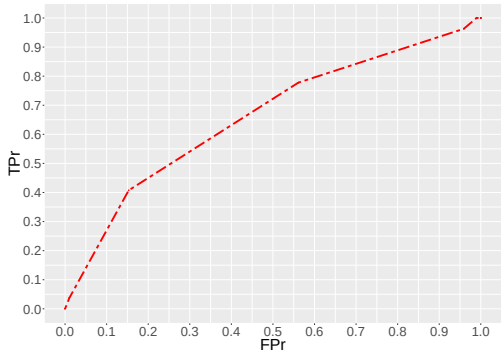
The ROC curves are plotted in Figure 4.4 for the three benchmark datasets. These ROC plots show how well NDTool can diagnose "normal" and abnormal observations. As explained earlier in Section 4.1, the area under a ROC curve is a popular measure for evaluating the diagnosis ability of a classifier (see Figure 4.2).



(a) ROC plot for Ionosphere with 34 dimensions



(b) ROC plot for Spambase with 57 dimensions



(c) ROC plot for Brest Cancer dataset with 9 dimensions

Figure 4.4 ROC plots of benchmark datasets

It is apparent from the results that NDTool performs well in terms of AUC measure compared to the other algorithms. For Ionosphere dataset, NDTool surpasses all rival algorithms with a value of 0.9502. For Spambase dataset, NDTool has a competitive performance as K-NN has the best performance with a value of 0.7766 whereas NDTool has a AUC value of 0.7345. For Breast Cancer dataset, NDTool performs almost the same as the best algorithm; NDTool with the AUC of 0.6605 and K-NN with an AUC of 0.6755. The last row of Table 4.2 indicate the average performance of each algorithm in terms of AUC. The results show that, on average, NDTool obtained the second rank between 5 investigated methods: NDTool with an average of 0.7817 has a very close performance to the first rank algorithm K-NN with an average of 0.7939.

Table 4.2 Comparisons of AUC performance of different methods for benchmark datasets

Datasets	NDTool	SVM	K-NN	GMM	K-means
Ionosphere	**0.9502**	0.8342	0.9296	0.8059	0.9191
Spambase	0.7345	0.6870	**0.7766**	0.7546	0.7289
Breast Cancer	0.6605	0.6678	**0.6755**	0.6662	0.6547
Average	0.7817	0.7296	0.7939	0.7422	0.7675

The performances of NDTool in terms of accuracy, sensitivity, precision, training time and detection time are given in Table 4.3. As we didn't have access to the results of these measures and CPU time (training time and detection time) for the other algorithms, we couldn't make a comparison between the obtained result and the investigated algorithms. However, we provide the whole experimental results. As an example, the results show that NDTool has a classification Accuracy of 0.9327, Sensitivity of 0.9285, and a precision value of 0.9435 for Ionosphere dataset. The last two columns indicate training time and detection time. The training time is the time that NDTool needs to perform Phase I and II. The detection time is the average time for detection of each newcomer data in Phase III and IV. The result shows that NDTool is not a time consuming algorithm and it performs in an acceptable time duration.

Table 4.3 Performance of NDTool

Datasets	ACC	Sensitivity	Precision	Training time (s)	Detecting time (s)
Ionosphere	0.9327	0.9285	0.9435	0.14	0.01
Spambase	0.6866	0.7137	0.7269	7.16	0.04
Breast Cancer	0.6480	0.4074	0.6875	0.11	0.01

4.4 Breast Cancer Experiments

This section provides a brief summary of breast cancer statistics and the urgent need for automated haelthcare systems. Then we use NDTool as a medical diagnostic support system. Two real world datasets about barest cancer are being used to show the accuracy of NDTool in detection of malignant masses.

In the past, the health and disease conditions were constantly monitored by physicians and technicians. Over time by increasing the number of patients, the necessity of developing automated diagnostic systems was aroused.

Unfortunately, Cancer is the main cause of death in Canada. According to annual Canadian Cancer publication in the year 2017, almost 206200 Canadians will be diagnosed with cancer while an approximate of $80,800$ of Canadian will lose their life as a result of a malignant cancer (Canadian Cancer Statistics, 2017).

The official reports state that breast cancer is the second most common cancer and is the most common cancer in women worldwide (World Cancer Research Fund International, 2012). As reported by Canadian Cancer Society, in the year 2017, an approximate of $26,300$ women and 230 men will be diagnosed with breast cancer. Among them 5000 women and 43 men will lose their life. This is despite the fact that global statistics are even worse. For example, the rate of breast cancer reaches 89.7 per $100,000$ women in Western Europe. An early detection of barest cancer can reduce the number of victims of this disease. In the developing countries, the number of individuals who suffer from breast cancer is increasing while the majority of cases are diagnosed in late stages(World Health Organization, 2008).

Mammography screening for checking up the breast condition is very costly. Low-cost screening approaches, such as clinical breast examination, could be implemented instead of the expensive imaging tests. So in countries with low-income can not afford an screening program for their population. Apart from the cost problem, human error can have a negative affect on the accuracy of detection. All the aforementioned facts necessitate the need for automated tool that can help low-cost and early diagnosis.

We applied NDTool on two real-world datasets about breast cancer, derived from the University of California – Irvine Machine Learning Repository (UCI) (Lichman, 2013). The datasets are from Wisconsin Breast Cancer Databases obtained from the University of Wisconsin Hospitals. All features are computed from a digitized image of biopsy of a breast mass. These features describe elements of the cell nuclei exist in the the digitized image. The characteristics of each instance are given in Table 4.4.

The first dataset named Original Breast Cancer is consist of 10 dimensions and 699 obser-

Table 4.4 Characteristics of breast cancer datasets

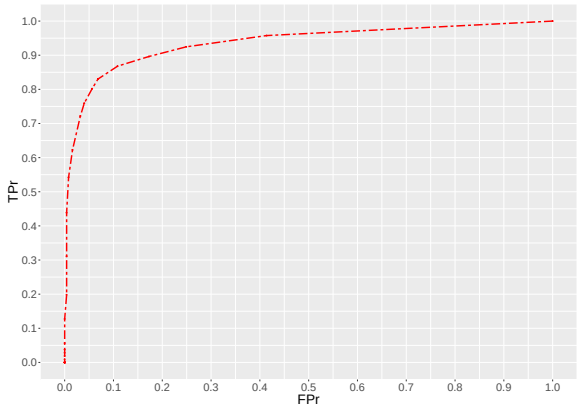
Datasets	No. of Dimensions	No. of Instances	Missing Values
Original Breast Cancer	10	699	Yes
Diagnostic Breast Cancer	32	569	No

vation. The second dataset is Diagnostic Wisconsin Breast Cancer with 569 rows and 32 dimensions. Both datasets include a dimension of ID number of the patients and a diagnosis decision dimension. The reminder of dimensions are related to the features of breast cancer experiments. In both datasets, the ID column which shows the patient admission number and the rows with missing data are ignored. Table 4.5 provides a summary of feature description for both breast cancer datasets.

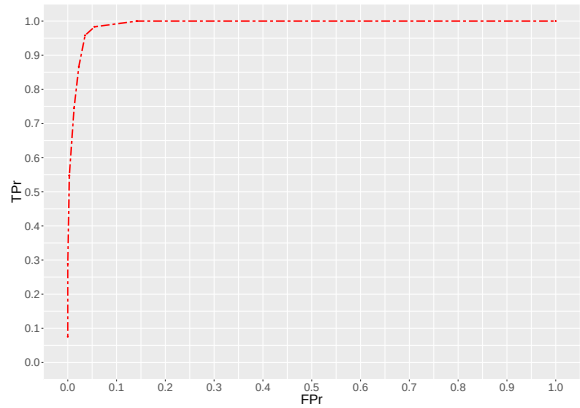
Table 4.5 Definition of attributes in breast cancer datasets

Attributes	**Definition**
radius	mean of distances from center to included points in the perimeter
texture	standard deviation of gray-scale values
perimeter	boundary size of affected area
area	surface size of affected area
smoothness	local variation in radius lengths
compactness	$(perimeter^2)/(area - 1.0)$
concavity	severity of concave portions of the contour
concave points	number of concave portions of the contour
symmetry	equality of breast breasts size
fractal dimension	coastline approximation-1

We pursue the same strategy as in Section 4.3 to separate a dataset into train and test subsets. The performance of NDTool is shown via the well-known measures mentioned earlier in Section 4.1. Figure 4.5 shows the ROC plots for both datasets. The area under ROC plots (AUC) indicate the good performance of NDTool.



(a) Roc plot for Diagnostic Wisconsin Breast Cancer with 32 dimensions



(b) Roc plot for Wisconsin Breast Cancer with 10 dimensions

Figure 4.5 ROC plots of Breast Cancer datasets

Table 4.6 shows a summary of NDTool performance on these two breast cancer datasets [3]. The obtained result shows an Accuracy of 0.9392 for Original Breast Cancer dataset and 0.8743 for Diagnostic Breast Cancer dataset. The training time and detection time are also provided in the last two columns.

Table 4.6 Performance of NDTool for Breast Cancer Studies

Datasets	ACC	Sensitivity	Precision	AUC	Training time (s)	Detection time (s)
Original Breast Cancer	0.9392	0.9581	0.9271	0.9605	0.09	0.009
Diagnostic Breast Cancer	0.8743	0.8254	0.9358	0.8993	0.12	0.012

After illustrating each novel behaviour and alarming during the process time, two visual report are provided by NDTool. A parallel coordinates plot and a pie chart.

The parallel coordinates plot contains the "normal" historical data and the cluster of novel behaviour that are detected during the working period. Figure 4.6 is an example of the parallel coordinates for Diagnostic Breast Cancer dataset with the historical data and the novelties which are in gray [4].

[3] We couldn't find any experiment in the literature, with the same scenario for these two dataset. That is why we couldn't make a comparison between the obtained result and other algorithms.

[4] For better viewing the figures, please see the electronic copy of the thesis.

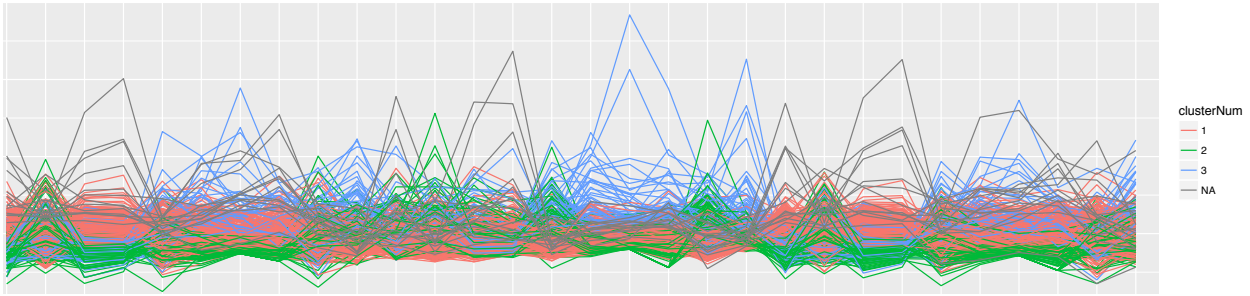


Figure 4.6 Parallel coordinates of Diagnostic Breast Cancer dataset with the cluster of novel behaviours which is in dark grey

As explained in Section 3.3, NDTool provides a pie chart after each working period. The pie chart shows the participation of each data feature in the total novelty observations that occur during a work period. For example, Figure 4.7 is the obtained pie chart for Diagnostic Breast Cancer dataset.

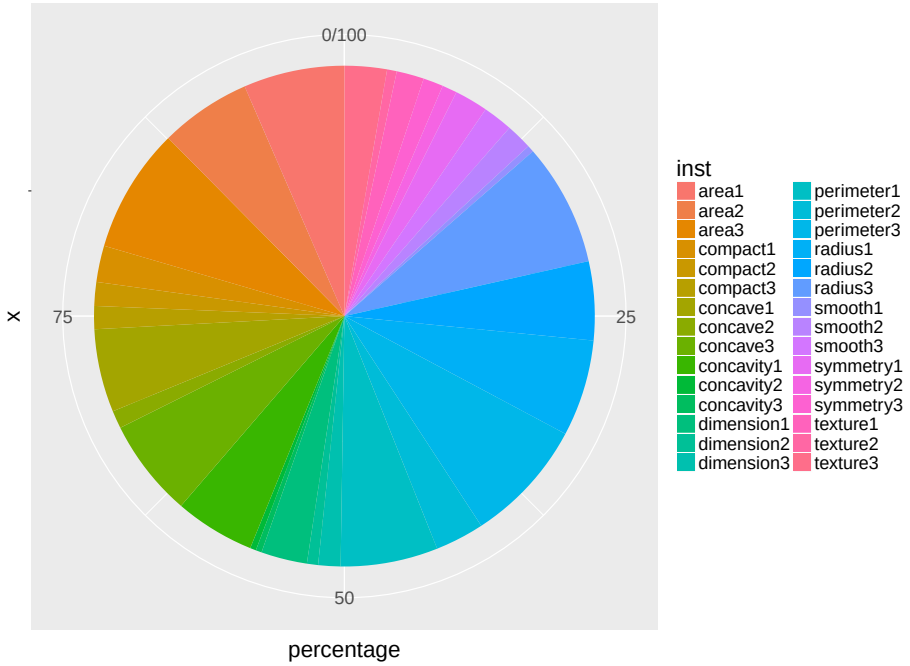


Figure 4.7 Pie Chart for Diagnostic Breast Cancer dataset

For more convenience in reading the result, Table 4.7 is provided. In this table the first and third columns contain the names of attributes. The "Percentage" columns indicates the participation rates of each attribute in the total novelties cases. The results show that in total, the "*area*" attributes (including area1, area2 and area3) have the most participation

in novel behaviours. In contrast, "*smooth*" attributes (including smooth1, smooth2 and smooth3) have the minimum portion of participation novelties.

Table 4.7 Participation percentage of each attribute in novel behaviour for Diagnostic Breast Cancer dataset

Attribute	**Percentage**	**Attribute**	**Percentage**
area1	6.51	perimeter1	6.27
area2	5.89	perimeter2	3.16
area3	8.05	perimeter3	8.00
compact1	2.34	radius1	6.23
compact2	1.53	radius2	5.08
compact3	1.48	radius3	7.90
concave1	5.36	smooth1	0.43
concave2	1.15	smooth2	1.67
concave3	6.32	smooth3	1.91
concavity1	5.17	symmetry1	2.15
concavity2	0.38	symmetry2	0.95
concavity3	0.38	symmetry3	1.29
dimension1	2.97	texture1	1.77
dimension2	0.71	texture2	0.62
dimension3	1.43	texture3	2.73

CHAPTER 5 DISCUSSION

The initial motivation of working on this study is coming from a real experience in an IT company. They were interested to study about system intrusion. But they had difficulties of preparing a decent dataset containing enough sample of the system crash. Obviously, they could not attack their systems in order to register the information about system failure. This real experience demonstrates the need for investigating on such scenarios that may arise in different industries. The way to deal with these kinds of scenarios can be framed in the context of novelty detection problem, where preparing the data from a specific class is kind of impossible.

Novelty detection is the problem of finding unknown samples where limited information or even no information is available about them. In this research, we focus on developing a novelty detection tool to help people in different industries who have no information in the domain of data mining and machine learning. This tool, named NDTool, is based on parallel coordinates plot and K-medoid clustering.

This tool contains four phases that are explained in Section 3.3. To sum up, NDTool takes a dataset which only consists of "normal" samples. Based on these available "normal" data, NDTool try to diagnose if a newcomer data is a novel or "normal" behavior. We evaluate the performance of NDTool via famous metrics on different datasets with different structures and dimensions from different industries. Then the results are compared with four most widely used methods, including support vector machine, K-nearest neighbor, Gaussian mixture and a K-means based method; see Table 4.2. NDTool has a slight difference (0.012), in terms of AUC, to the best performance algorithm, K-Nearest Neighbor.

NDTool is applied for the breast cancer studies through real-life datasets. The accuracy of detecting a malignant mass in Original Breast Cancer dataset and Diagnostic Breast Cancer dataset are 0.9392 and 0.8743, respectively. For all datasets, other metrics including Accuracy, Sensitivity, and Precision are measured; see Tables 4.3 and 4.6. The conducted experiments show that NDTool has a good performance compared to competitors.

In this study, NDTool has the most conservative approach toward non-conformities in each newcomer data point. It means that NDTool alarms if a data point contains just one error of any kind. We deliberately took this approach because of the application which is related to human life (breast cancer). But depends on the application, NDTool could be changed to a less conservative mode toward the non-conformities. For example, it could alarm after finding two or three non-conformities in a newcomer data point. In fact, the desirable levels

of False Positive alarm and False Negative errors are different for industries. For example in healthcare studies related to patients'lives, False Positive alarms are tolerated, while a False Negative error may result in the death of a patient. So a conservative and rigorous approach is safer for these kinds of applications. On the other hand, in an example of manufacturing production line, a false alarm may be equivalent to putting a unit of product out of a production line as waste or even a pause in a production line.

The experiments show the trade-off between False Positive alarms and False Negative errors via ROC plots. But as previously explained, in real-world scenarios of novelty detection, the abnormal samples are not at hand. So in real cases, it is not possible to conduct an experiment like ROC plot curve to test and decide about the threshold of alarming. We suggest beginning with the most conservative approach and then if it was necessary the number of errors for an alarm could be increased.

The visual reports at the end of each work period help to draw attention to the attributes that have a significant portion to make the novel behaviors. Prevention of non-conformities in these attributes can help decrease the number of novel behaviours. Depends on the industry and the case at hand, different strategies could be taken, from solving the problem to accepting the current novelties as the "normal" behavior.

In a parallel coordinates plot, the order of coordinates affects the data structure (Tilouche et al., 2017). This structure can effect on defining the angle ranges in phase II of NDTool. NDTool and the project of Tilouche et al. (2017) have been performed in the same laboratory, CIMAR-LAB and they can be merged in order to improve the novelty detection process.

With all this said, the proposed tool is still an experimental prototype and it is needed to examine the usability of this tool via applying that in real scenarios in different industries.

CHAPTER 6 CONCLUSION

6.1 Synthesis of the work

We have proposed a novelty detector tool based on combined parallel coordinates and K-medoids clustering algorithm for novelty detection problem. The proposed algorithm has four main phases. Phase I contains some preprocessing tasks. Afterward, Phase II gets information about the structure of the normal historical dataset. Then in phase III, a new observation come into the process and the same preprocessing tasks are done for this data. Phase IV is where the newcomer observation get evaluated versus the historical data. It alarms in case of finding a novel behavior in new observation while illustrating the historical data and newcomers observation on the parallel coordinate plots. At the end, a pie chart and a final parallel coordinates are illustrated and stored in the system. The pie chart shows the portion of each feature in novel behaviour and the final parallel coordinates contains the historical data plus all of the new observations including novel behaviors.

The performances of the NDTool have been investigated on a set of well-known benchmarks and compared with state-of-art algorithms, including K-NN, SVDD, GMM, and K-means based algorithms. The results show that NDTool has a very close performance to the best algorithm, K-NN, and surpass SVDD, GMM, and K-means based algorithm. We also used NDTool for detection of cancer in two real breast cancer dataset. The overall results show that NDTool provides efficient performance in solving novelty detection problem.

6.2 Limitations of the proposed solution

A limitation is that NDTool algorithm is just compatible with numerical values. In the case of data with categorical values, a data preprocessing for converting the text to numerical and a code modification are necessary.

We also didn't have access to the result of the test and train times of any experiments with the scenario of novelty detection for the applied data sets. Besides that, we didn't have access to the result of conducted experiments for the two breast cancer datasets (4.4). So we couldn't make a comparison in the mentioned experiments.

Another limitation is the existence of noise and outlier in the historical data. It may reduce the accuracy of the detection process.

Another limitation is the slowness of R in visualization. The visualization for each data point

and alarming take a relatively long time.

6.3 Future Work

This work can be further expanded to apply reordering and dimension reduction techniques as a preprocessing phase. It helps NDTool to handle big datasets, where the number of data can be reduced. This part of work has been done by another project (Tilouche et al., 2017) in our laboratory, CIMAR-LAB. These two projects could be merged to build a more efficient tool.

Other clustering methods could be tested for different datasets as a preprocessing phase. It may improve the result depending on the available data structure. Also, hybridization of NDTool with the other machine learning algorithm, such as K-NN, may provide a way to improve the efficiency of the algorithm.

Furthermore, designing an ergonomic interface containing the controlling features is a necessary need for evolution NDTool. It is very advantageous to design an HTML-based interface for NDTool to make it more convenience and ease of use for different industries.

Also, it is necessary to conduct more experiments in real scenarios in order to justify the usefulness of NDTool.

REFERENCES

G. Andrienko et N. Andrienko, “Constructing parallel coordinates plot for problem solving”, *1st International Symposium on Smart Graphics*, pp. 9–14, 2001.

S. B. Azhar et M. J. Rissanen, “Evaluation of parallel coordinates for interactive alarm filtering”, pp. 102–109, 2011.

P. Berkhin, “A survey of clustering data mining techniques.” *Grouping multidimensional data*, vol. 25, p. 71, 2006.

BI-Survey, “The most common problems companies are facing with their big data analytics”, May 2017. En ligne: `https://bi-survey.com/challenges-big-data-analytics`

G. Blanchard, G. Lee, et C. Scott, “Semi-supervised novelty detection”, *Journal of Machine Learning Research*, vol. 11, no. Nov, pp. 2973–3009, 2010.

Canadian Cancer Society, “Breast cancer statistics”, 2017. En ligne: `http://www.cancer.ca/en/cancer-information/cancer-type/breast/statistics/?region=qc`

Canadian Cancer Statistics, “Canadian cancer society’s advisory committee on cancer statistics”, 2017. En ligne: `cancer.ca/Canadian-CancerStatistics-2017-EN.pdf`

V. Chandola, A. Banerjee, et V. Kumar, “Anomaly detection: A survey”, *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

H. Choi, H. Lee, et H. Kim, “Fast detection and visualization of network attacks on parallel coordinates”, *computers & security*, vol. 28, no. 5, pp. 276–288, 2009.

CIMAR-LAB. En ligne: `http://www.polymtl.ca/cimar/`

P. Cunningham et S. J. Delany, “k-nearest neighbour classifiers”, *Multiple Classifier Systems*, vol. 34, pp. 1–17, 2007.

F. De Morsier, D. Tuia, M. Borgeaud, V. Gass, et J.-P. Thiran, “Semi-supervised novelty detection using svm entire solution path”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 4, pp. 1939–1950, 2013.

C. P. Diehl et J. B. Hampshire, “Real-time object classification and novelty detection for collaborative video surveillance”, *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on*, vol. 3, pp. 2620–2625, 2002.

X. Ding, Y. Li, A. Belatreche, et L. P. Maguire, “An experimental evaluation of novelty detection methods”, *Neurocomputing*, vol. 135, pp. 313–327, 2014.

R. O. Duda, P. E. Hart, et D. G. Stork, *Pattern classification*. Wiley, New York, 1973.

R. M. Edsall, “The parallel coordinate plot in action: design and use for geographic visualization”, *Computational Statistics & Data Analysis*, vol. 43, no. 4, pp. 605–619, 2003.

T. Fawcett, “An introduction to roc analysis”, *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

D. M. Hawkins, “Identification of outliers”, vol. 11, 1980.

P. Hayton, S. Utete, D. King, S. King, P. Anuzis, et L. Tarassenko, “Static and dynamic novelty detection methods for jet engine health monitoring”, *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 365, no. 1851, pp. 493–514, 2007.

J. Heinrich et D. Weiskopf, “State of the art of parallel coordinates.” *Eurographics (STARs)*, pp. 95–116, 2013.

R. G. M. Helali, “Data mining based network intrusion detection system: A survey”, *Novel Algorithms and Techniques in Telecommunications and Networking*, pp. 501–505, 2010.

L. Hu, N. Hu, B. Fan, et F. Gu, “Application of novelty detection methods to health monitoring and typical fault diagnosis of a turbopump”, dans *Journal of Physics: Conference Series*, vol. 364, no. 1. IOP Publishing, 2012.

A. Inselberg, “The plane with parallel coordinates”, *The visual computer*, vol. 1, no. 2, pp. 69–91, 1985.

X. Jin et J. Han, “K-medoids clustering”, *Encyclopedia of Machine Learning*, pp. 564–565, 2011.

J. Johansson et C. Forsell, “Evaluation of parallel coordinates: Overview, categorization and guidelines for future research”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 579–588, 2016.

V. Jyothsna, V. R. Prasad, et K. M. Prasad, “A review of anomaly based intrusion detection systems”, *International Journal of Computer Applications*, vol. 28, no. 7, pp. 26–35, 2011.

L. Kaufman et P. Rousseeuw, *Clustering by means of medoids.* North-Holland, 1987.

S. S. Khan et M. G. Madden, “A survey of recent trends in one class classification”, dans *Irish Conference on Artificial Intelligence and Cognitive Science.* Springer, 2009, pp. 188–197.

H. J. Kuijf, P. Moeskops, B. D. de Vos, W. H. Bouvy, J. de Bresser, G. J. Biessels, M. A. Viergever, et K. L. Vincken, “Supervised novelty detection in brain tissue classification with an application to white matter hyperintensities”, dans *Medical Imaging 2016: Image Processing*, vol. 9784, 2016.

M. Lichman, “UCI machine learning repository”, 2013. En ligne: `http://archive.ics.uci.edu/ml`

M. Markou et S. Singh, “Novelty detection: a review—part 1: statistical approaches”, *Signal processing*, vol. 83, no. 12, pp. 2481–2497, 2003.

S. Marsland, “Novelty detection in learning systems”, *Neural computing surveys*, vol. 3, no. 2, pp. 157–195, 2003.

M. M. Moya, M. W. Koch, et L. D. Hostetler, “One-class classifier networks for target recognition applications”, Sandia National Labs., Albuquerque, NM (United States), Rapp. tech., 1993.

J. Owens, A. Hunter, et E. Fletcher, “Novelty detection in video surveillance using hierarchical neural networks”, *Artificial Neural Networks—ICANN 2002*, pp. 140–140, 2002.

R. Pant, “Visual marketing: A picture’s worth 60,000 words”, Jan 2015. En ligne: `http://www.business2community.com/digital-marketing/visual-marketing-pictures-worth-60000-words-01126256#bYUOQ0OZ3WOlexji.97`

M. A. Pimentel, D. A. Clifton, L. Clifton, et L. Tarassenko, “A review of novelty detection”, *Signal Processing*, vol. 99, pp. 215–249, 2014.

P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”, *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, et J. C. Platt, “Support vector method for novelty detection”, *Advances in neural information processing systems*, pp. 582–588, 2000.

D. Sonntag, “Ontologies and adaptivity in dialogue for question answering”, vol. 4, 2010.

C. Surace et K. Worden, “Novelty detection in a changing environment: a negative selection approach”, *Mechanical Systems and Signal Processing*, vol. 24, no. 4, pp. 1114–1128, 2010.

L. Tarassenko, P. Hayton, N. Cerneaz, et M. Brady, “Novelty detection for the identification of masses in mammograms”, *4th International Conference on Artificial Neural Networks*, 1995.

D. M. Tax et R. P. Duin, “Support vector data description”, *Machine learning*, vol. 54, no. 1, pp. 45–66, 2004.

S. Tilouche, S. Bassetto, et V. Partovi Nia, “Parallel coordinate order for high-dimensional data”, *Technical Report G—2017—38*, vol. Cahiers du GERAD, 2017.

K. Worden, H. Sohn, et C. R. Farrar, “Novelty detection in a changing environment: regression and interpolation approaches”, *Journal of Sound and Vibration*, vol. 258, no. 4, pp. 741–761, 2002.

World Cancer Research Fund International, “Breast cancer statistics”, 2012. En ligne: `http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics`

World Health Organization, “Breast cancer: prevention and control”, 2008.

Y. Xu, W. Hong, N. Chen, X. Li, W. Liu, et T. Zhang, “Parallel filter: a visual classifier based on parallel coordinates and multivariate data analysis”, *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence*, pp. 1172–1183, 2007.