分类号 密 级

U D C 学校代码 10500



****

**硕士学位论文**

Computer Sciences - Master

题 目： 针对有异常值的 Naive Bayes 分类的优化鲁棒核密度估计

###### 英文题目：Optimised Robust Kernel Density Estimation for Naive Bayes Classification with Outliers

**学位申请人姓名：BOLI BI IRITIE A-D**

**硕士学位课程：Computer Sciences**

**指导教师姓名：Dr. Chenghao Wei**

二○二 年五月

分类号 密 级

U D C 学校代码 10500



****

**硕士学位论文**

题 目:针对有异常值的 Naive Bayes 分类的优化鲁棒核密度估计

英文题目: Optimized Robust Kernel Density Estimation for Naive Bayes Classification with Outliers

研究生姓名（签名）: BOLI BI IRITIE ARNAUD-DONATIEN

指导教师姓名（签名）: Wei 职 称: Doctor

申请学位学科名称 Computer Science 学科代码:

论文答辩日期 学位授予日期

学院负责人（签名）:Computer School

评阅人姓名 评阅人姓名

年 月 日



学位论文原创性声明和使用授权说明

**Originality Statement**

I solemnly declare that the dissertation submitted is the result of my research work conducted independently under the guidance of my supervisor. Except for what has been marked and cited in the text, this dissertation does not contain any research results that have been published or written by any other individual or collective. Individuals and collectives who have contributed to the research of this paper have been clearly indicated in the text. The legal consequences of this statement will be borne by me.

Signature：                               Date：    　 年  　 月 　  日

**Authorization for the use of copyright in the dissertation**

The author of this dissertation fully understands the regulations of the university concerning the retention and use of the dissertation, that is, the university has the right to retain and send copies of the dissertation and its electronic version to the relevant departments or institutions of the state, and to allow the dissertation to be inspected and borrowed. I hereby authorise Hubei University of Technology to compile all or part of this dissertation into relevant databases for searching, and to preserve and compile this dissertation by means of reproduction such as photocopying, microfilming or scanning.

Signature of the author：          　　　　     Signature of instructor：

Date：     年    月    日         　　　  日期：     年    月     日

# 摘 要

**关键词：**

# Abstract

**Keywords**:

# Table Contents

摘 要 I

Abstract II

Table Contents III

1. Introduction 1

1.1. Background and motivation 1

1.2. Research Objectives 2

1.3. Thesis Structure 2

2. Literature Review 3

2.1. Machine Learning and Intelligence Data Analysis 4

2.2. Kernel Density Estimation (KDE) 5

2.2.1. Gaussian KDE and its Applications 6

2.2.2. Limitations of Traditional KDE 7

2.3. Robust Kernel Density Estimation (RKDE) 7

2.3.1. M-estimation and Hampel Function 8

2.3.2. Iterative Reweighted Least Squares (IRLS) 10

2.4. Naïve Bayes Classifier 11

2.4.1. Baisc Naïve Bayes Classifier 11

2.4.2. Challenges with Noisy and Tough Datasets 12

3. Methodology 13

3.1. Robust RKDE with IRLS 13

3.2. Harris Hawk Optimization (HHO) for bandwidth selection 14

3.2.1. Harris Hawk Optimization 14

3.2.2. Description HHO 15

3.2.2.1. Exploration phase 15

3.2.2.2. Transition from Exploration to Exploitation 16

3.2.2.3. Exploitation Phase 16

3.2.2.4. The Main Steps of HHO 18

3.2.3. Selecting Bandwidth with HHO 19

3.3. Complete Cross-Validation Curve (CCV) 21

3.3.1. Unbiased Cross-Validation (UCV) 21

3.3.2. Biased Cross-Validation (BCV) 21

3.3.3. Bootstrap 21

3.3.4. Importance of using CCV 21

3.4. Integration of Robust KDE and Bandwith selection into Naïve Bayes classifier 21

4. Experimental Setup 21

4.1. Dataset Description 21

4.2. Implementation Details 21

4.3. Performance metrics 21

5. Results and Discussion 21

5.1. Evaluation of Optimized Robust KDE 21

5.2. Impact of Bandwidth selection on classification performance 21

5.3. Comparison with traditional Naïve Bayes Classifier 21

6. Conclusion and Future Work 21

6.1. Summary of findings 21

6.2. Contributions to the field 21

6.3. Limitation and Future direction 21

7. References 21

Appendices 24

A. List of papers published by the author during his degree study 24

B. Python code display 24

# Introduction

# Background and motivation

In recent years, there have been significant developments in the field of machine learning, particularly in the area of intelligence data analysis. Machine learning algorithms have been extensively utilised across many sectors, such as healthcare and finance, due to their ability to process substantial datasets and leverage robust computational resources. These algorithms have proven effective in addressing intricate challenges that were previously deemed intractable. The widespread adoption of machine learning has fundamentally transformed our capacities in the domains of decision-making, automation, and predictive analytics. Nevertheless, the increasing dependence on machine learning methods has led to a heightened awareness of their susceptibilities, particularly in the presence of noisy and complex datasets. Real-world data frequently contains various defects, such as outliers, measurement mistakes, and instances of missing numbers. When left unattended, these flaws can have a substantial impact on the dependability and efficiency of machine learning models, resulting in less than optimal results.

Accurately understanding and modelling data distributions is a fundamental requirement for numerous machine learning techniques. Kernel Density Estimation (KDE) is a crucial technique in this field, providing the capability to represent the underlying data distributions. The KDE method plays a significant role in a wide range of applications, encompassing anomaly detection, classification, and clustering. The effectiveness of machine learning models is heavily reliant on precise density estimate, which forms the basis for essential activities such as categorization, decision-making, and risk evaluation. Nevertheless, conventional KDE techniques frequently encounter difficulties when faced with datasets that include outliers and noisy observations. Aberrant data points, sometimes referred to as outliers, have the potential to exert a disproportionate influence on the process of density estimation. Consequently, this might result in the production of skewed models and inferior outcomes in subsequent tasks. The task of addressing this difficulty and optimising the Kernel Density Estimation (KDE) technique holds significant significance in improving the resilience and dependability of machine learning models.

Outliers can be defined as data points that exhibit substantial deviation from the remaining data. Measurement errors, data corruption, or the occurrence of extraordinary occurrences might give rise to these issues. In numerous practical situations, outliers possess significant informational value and should not be disregarded, but rather be appropriately incorporated into the modelling process. Nevertheless, it is important to acknowledge that these techniques have the potential to alter the fundamental distribution of data, resulting in biassed estimations of density and, consequently, unreliable predictions from the model.

# Research Objectives

The central aim of this thesis is to contribute to the advancement of robust Kernel Density Estimation and its utilisation in the study of intelligence data. Our study aims to accomplish the following specific objectives:

To enhance or maximise the efficiency or effectiveness of a particular process or system. This study aims to develop and enhance robust kernel density estimation (KDE) approaches that can effectively estimate data densities in scenarios where outliers and noisy data points are present. The term "robust" refers to approaches that exhibit reduced vulnerability to the impact of outliers and offer more dependable density estimates.

Maximising the Efficiency of Bandwidth Selection: The utilisation of the Harris Hawk Optimisation (HHO) algorithm is proposed in order to optimise the selection of bandwidth parameters, which play a crucial role in Kernel Density Estimation (KDE). This optimisation aims to boost the accuracy and adaptability of KDE when applied to various datasets. The selection of bandwidth is a crucial and important task in the context of KDE (Kernel Density Estimation), and its optimisation has the potential to enhance the accuracy and precision of density estimates.

This study aims to explore and apply techniques to minimise the influence of outliers in classification prediction. Specifically, the focus will be on investigating strategies to mitigate the negative consequences of outliers in KDE-based models, with a particular emphasis on the Naïve Bayes classifier. Classification tasks are often observed in the field of machine learning, and the capacity to generate reliable predictions even when outliers are present holds substantial practical significance.

# Thesis Structure

This thesis is structured as follows:

**Chapter 2**: The research Review offers a comprehensive examination of the current body of research pertaining to machine learning, Kernel Density Estimation, robust statistics, and the difficulties presented by datasets including noise. We utilize influential literature and contemporary scholarly investigations to ascertain the present understanding in these fields.

**Chapter 3:** The methodology section provides a detailed explanation of the methodologies utilized in this study, encompassing the enhancement of robust KDE techniques, the utilization of the Harris Hawk Optimization algorithm, and the incorporation of these approaches into the Naïve Bayes classifier. In this thesis, we outline the fundamental technical principles that underpin our research.

**Chapter 4**: The Experimental Setup section provides a comprehensive overview of the datasets employed, the technical implementation details, and the performance criteria selected to assess the efficacy of the proposed approaches. The implementation of rigorous experimentation is of utmost importance in order to evaluate the efficacy of our methodologies.

**Chapter 5:** The section titled "Results and Discussion" summarizes the outcomes of the conducted tests, analyses their consequences, and evaluates the influence of bandwidth selection on the performance of categorization. By employing both quantitative analysis and qualitative talks, we thoroughly examine the practical consequences of our research.

**Chapter 6:** In conclusion, this study has provided valuable insights into the topic at hand. Moving forward, there are several avenues for further research that could further enhance our understanding of the subject The present study provides a comprehensive overview of the research's contributions, delineates its inherent limits, and proposes potential directions for further investigation within the domain of robust kernel density estimation (KDE) for machine learning. The present discourse entails a contemplation of the ramifications of our research and its capacity to propel progress in the field of intelligence data analysis.

**Chapter 7:** The references section encompasses an extensive compilation of sources, including influential works that have provided guidance for the present study. We express our gratitude for the valuable contributions made by the wider scientific community in influencing and refining our comprehension of the subject matter.

This thesis aims to make a contribution to the improvement of robust KDE approaches, with the ultimate goal of enhancing the capabilities of machine learning models in effectively managing tough and noisy datasets. The subsequent chapters will explore the complexities of these approaches and their empirical verification, facilitating a more profound comprehension of intelligent data analysis within the realm of machine learning.

# Literature Review

The literature review within this thesis offers a comprehensive examination of the fundamental principles and recent advancements in the domains of machine learning and intelligent data analysis. This chapter establishes the foundation by analysing the fundamental ideas and difficulties linked to these domains, thereby preparing for an in-depth investigation of KDE and its resilient variations.

# Machine Learning and Intelligence Data Analysis

The discipline of machine learning, which falls under the umbrella of artificial intelligence (AI), has experienced significant expansion and broadening in recent decades. Machine learning involves a diverse range of methodologies that facilitate the ability of computers to acquire knowledge from data, identify patterns, and make informed judgements without the need for explicit programming. The progression of machine learning can be succinctly encapsulated as follows:

**Early Developments**: The origins of machine learning may be historically attributed to the mid-20th century, during which influential figures such as Alan Turing and Arthur Samuel established the fundamental principles of rule-based learning and game-playing algorithms. These first advancements laid the foundation for subsequent investigations.

**Statistical Learning**: The field of machine learning has undergone a progressive transformation, transitioning into a statistical discipline that extensively incorporates principles from probability theory and statistical methodologies. Methods such as linear regression and decision trees have gained significant popularity in the field of pattern detection and prediction.

**The Era of Neural Networks**: During the 1980s and 1990s, artificial neural networks emerged as a prominent paradigm within the field of machine learning. Nevertheless, the potential of their capabilities was hindered by the constraints imposed by low computer resources and an inadequate amount of data.

**The Big Data Revolution**: The 21st century has saw a significant transformation in the field of machine learning, chiefly influenced by the abundant availability of extensive datasets and notable progress in computational capabilities. Consequently, the comeback of neural networks in the form of deep learning has emerged as a prevailing force in several domains, encompassing image identification, natural language processing, and autonomous cars.

**Interdisciplinary Impact:** Machine learning has expanded beyond its conventional limitations and has been implemented in other fields, such as healthcare (for purposes of diagnosis and medication development), finance (including algorithmic trading and fraud detection), and natural language processing (including chatbots and language translation).

The study of intelligence data, as it pertains to the field of machine learning, is of significant significance in the extraction of actionable insights, the facilitation of informed decision-making, and the automation of intricate procedures in many industries. The following are few fundamental elements that contribute to its significance:

**Decision Support:** The study of intelligence data offers decision-makers with significant insights derived from extensive databases. Data-driven insights play a pivotal role in enabling organisations across several sectors, including healthcare, finance, and marketing, to make informed decisions that optimise operational efficiency and enhance overall performance.

**Predictive Modeling:** Machine learning methodologies facilitate the creation of prognostic models that anticipate forthcoming occurrences or patterns. This holds significant value in various circumstances, including weather forecasting, stock market prediction, and disease outbreak monitoring.

**Anomaly Detection:** The identification of abnormalities or outliers in data is a critical undertaking that finds utility in various domains such as fraud detection, network security, and quality control. Machine learning algorithms have exceptional proficiency in identifying patterns that differ from the established norm.

**Pattern Recognition:** Machine learning models possess a high level of proficiency in identifying intricate patterns inside datasets. This capability is utilised in the fields of image identification, speech recognition, and natural language comprehension, leading to advancements such as autonomous vehicles and virtual personal assistants.

**Personalization:** The utilisation of intelligence data analysis plays a crucial role in driving recommendation systems, which are designed to offer personalised content and product choices. These systems have the ability to improve user experiences in various domains such as e-commerce, streaming services, and online advertising.

The rising reliance on data-driven decision-making across several industries underscores the growing importance of robust and accurate data analysis tools. The approach of Kernel Density Estimation holds significant importance in the modelling of data distributions, therefore making it a central focus of inquiry in this thesis.

The subsequent sections of this chapter will provide a more comprehensive consideration of the ideas and issues related to KDE, specifically within the framework of robust data analysis.

# Kernel Density Estimation (KDE)

The KDE technique is a widely employed statistical method utilized to estimate the probability density function (PDF) of a continuous random variable. This technique assumes a crucial position in diverse applications, encompassing data smoothing, density estimation, and outlier detection. This section will delve into the Gaussian KDE, which is considered one of the most extensively utilized variations, along with its various applications. Furthermore, we will examine the constraints associated with conventional KDE methodologies.

# Gaussian KDE and its Applications

The Gaussian KDE is a non-parametric method that aims to estimate the probability density function (PDF) of a given dataset. This methodology involves applying a Gaussian (normal) distribution, referred to as a kernel, on each individual data point and aggregating them to form the overall density estimate. The Gaussian kernel function, denoted as , is expressed as:

Where:

* is the kernel function.
* represents the distance from the data point of interest.

The Gaussian KDE estimate for a given data point with a dataset is calculated as follows:

Where:

* stands for the KDE estimate of the PDF at point .
* is the total number of data points in the dataset.
* signifies individual data points within dataset .
* is the bandwidth parameter, responsible for controlling the level of smoothing and is a pivotal factor in KDE as it determines the width of the kernels.

Gaussian KDE has gained widespread application in various domains:

**Probability Density Estimation:** The Gaussian KDE is a highly significant tool in the field of statistical analysis and hypothesis testing due to its ability to give a non-parametric estimation of the PDF .

**Data Smoothing:** The aforementioned methodology is employed to mitigate the impact of noise on data, resulting in the reduction of high-frequency noise and the accentuation of fundamental patterns in temporal data.

**Data Visualization:** The Gaussian KDE method plays a crucial role in producing visually appealing density plots, also known as kernel density plots. These plots are widely employed for the purpose of visualising the distribution of data.

**Outlier Detection:** The Gaussian Kernel Density Estimation (KDE) technique assists in the detection of outliers in datasets by constructing a model of the underlying distribution of the data and identifying data points that exhibit substantial deviations from this model.

# Limitations of Traditional KDE

While Gaussian KDE is a powerful tool, it is not without its limitations:

**Bandwidth Selection:** The selection of bandwidth (ℎ) has a substantial influence on the outcome of the KDE. An excessively limited bandwidth yields an estimate that is prone to overfitting, whereas an excessively broad bandwidth results in an estimate that is excessively smoothed. The process of choosing a suitable bandwidth is a complex undertaking.

**Curse of Dimensionality:** The effectiveness of KDE diminishes as the dimensionality of the data rises. The drop in volume of the kernel exhibits a rapid decline as the dimensionality increases, leading to the emergence of sparsity in spaces with high dimensions.

**Computational Demands**: The computing demands of KDE, particularly in high-dimensional domains, might be substantial, hence restricting its practicality for big datasets.

**Edge Effects:** The estimation of KDE towards the limits of the data domain may exhibit bias due to the absence of data points beyond the boundary, which is necessary for the formation of valid kernels.

**Sensitivity to Data Density:** The performance of KDE can exhibit substantial variations based on the density and distribution of the data. Sparse or multimodal datasets may result in poor performance.

**Boundary Issues:** The presence of boundaries or discontinuities in the data can result in inaccurate estimates due to the potential for undersmoothing or over smoothing.

Given the aforementioned constraints, scholars have undertaken investigations into resilient adaptations of the KDE method and have developed strategies to optimize the selection of bandwidth. In the following portions of this thesis, we will explore robust Kernel Density Estimation techniques and novel strategies to address these limitations.

# Robust Kernel Density Estimation (RKDE)

The KDE technique offers a non-parametric approach to estimate probability density functions based on observed data. The conventional KDE method exhibits a strong susceptibility to outliers and may yield density estimates that are biased when extreme values are present. In order to tackle this issue, the use of RKDE has emerged as a helpful tool, providing resilience against outliers and noise present in the data. It can be expressed using the following formula:

Let be a dataset with n data points, and represents the kernel function. The RKDE estimate of the probability density function at point is given by:

where:

* is the bandwidth parameter, determining the width of the kernel window.
* signifies individual data points within dataset
* is the weight assigned to each data point based on robust M-estimation, such as the Hampel function.

# M-estimation and Hampel Function

M-estimation is a statistical technique that is frequently utilized in the field of robust statistics for the purpose of estimating parameters of a given model. The primary purpose of this study is to identify the optimal parameter values that maximize a designated objective function, commonly known as the "M-estimator." One of the primary obstacles encountered in the field of robust density estimation pertains to the existence of outliers, which possess the potential to exert a substantial impact on the precision of conventional KDE techniques.

In order to tackle this particular difficulty, we propose the utilization of the Hampel function, which is a robust weight function employed in M-estimation techniques. But we have a lot of objective function in M-estimation technique such as *Huber*, *tuTukey’s biweight* and *bisquare* functions. In downweighing outliers while still keeping the influence of inliers, the Huber function is a reguarly used option in robust M-estimation. The formula definition is:

where δ represents the tuning parameter, the Huber function strikes a balance between robustness and sensitivity, making it suitable for mitigating the impact of outliers.

The Huber function is extended by the Hampel function, an-other reliable M-estimator, by adding a parameterized threshold that modifies the estimator’s behaviour in reaction to outliers and its capacity to diminish or "trim" extreme values, so diminishing the impact of outliers in the estimate procedure. As specified as

where x represents the residual (difference between the data point and the estimated density), and δ is a tuning parameter that determines the robustness of the estimator.

For |x| ≤ δ, the function behaves quadratically, penalizing small residuals.

For |x| > δ, the linear behaviour of the function provides a more robust weight to larger residuals, thereby reducing the impact of outliers on the density estimation (17). Although the Tukey’s biweight and bisquare functions, as well as the Cauchy function, offer different approaches for achieving robustness, the Huber and Hampel functions are still popular options. These functions add different levels of non- linearity to the M-estimation process, offering various methods for downweighing outliers and boosting the estimator’s robust- ness. In conclusion, the performance and robustness of KDE approaches in treating outliers are greatly influenced by the M- estimation function selection. The choice should be made in accordance with the features of the data and the desired amount of outlier attenuation because each function introduces a different trade-off between robustness and sensitivity.

The integration of the Hampel function within the KDE framework, the RKDE is defined as:

where:

* is the bandwidth parameter, determining the width of the kernel window.
* signifies individual data points within dataset
* is the weight assigned to each data point based on robust M-estimation, such as the Hampel function.
* is the Hampel function, which assigns weights to data points based on their distance from .

The bandwidth parameter ℎ determines the width of the kernel and plays a crucial role in regulating the smoothness of the density estimate. It requires to be carefully selected to balance bias and variance in the estimate.

The RKDE combines the weighted contributions of all data points, taking into account their proximity to the evaluation point x. This is achieved by the use of a kernel function, which helps to smooth out the contributions. The robustness of RKDE lies in its ability to offer more precise density estimations when confronted with outliers or data distributions that deviate from the standard, in contrast to conventional KDE techniques.

It is important to note that different implementations of RKDE may employ diverse kernel functions and robustness methods, resulting in changes in the specific formulation of the approach.

# Iterative Reweighted Least Squares (IRLS)

The IRLS method is a commonly employed numerical optimization approach in the field of robust statistics, particularly in the context of robust density estimation. It iterative process involves refining parameter estimations by updating them according to the weighted least squares criterion. M-estimation issues, which involve the maximization of a probability function, are particularly well-suited for this method.

In the context of robust kernel density estimation the iteratively reweighted least squares algorithm is utilised to optimise the parameters associated with the Hampel function. The procedure entails iteratively adjusting the tuning parameters of the Hampel function in order to minimise the impact of outliers on the estimation of density. The algorithm operates in the following manner:

1. Initialize the turning parameters.
2. Compute the weights for each data point using the current parameters.
3. Update the parameters based on the weighted least squares criterion.
4. Repeat steps 2 and 3 until convergence is achieved.

The iterative structure of the Iteratively Reweighted Least Squares approach enables it to dynamically adjust to the underlying data distribution. This adaptability allows the method to efficiently mitigate the influence of outliers, while simultaneously maintaining the accuracy of density estimate for non-outlying data points.

To summarize, the technique known as robust Kernel Density Estimation employs M-estimation with the Hampel function and Iterative Reweighted Least Squares to address the impact of outliers and data distributions with heavy tails. The Hampel function, due to its capacity to handle extreme values, presents a convincing approach for obtaining robust density estimation when faced with difficult data features. In the following sections, we investigate the integration of RKDE, with the assistance of these resilient methodologies, into the Naïve Bayes classifier in order to improve its performance in the analysis of intelligence data.

# Naïve Bayes Classifier

The Naïve Bayes Classifier is a fundamental and extensively employed probabilistic classification algorithm within the field of machine learning. The origins of this concept may be traced back to Bayesian probability theory, which is a robust framework used to address uncertainty and make predictions using existing evidence.

# Baisc Naïve Bayes Classifier

The Naïve Bayes Classifier is built upon the foundational principle of Bayes' theorem, which is a fundamental topic within the field of probability theory. The utilisation of prior knowledge of conditions that may be associated with an event enables the calculation of its likelihood. In the domain of NBC, the application of Bayes' theorem for classification is of relevance, particularly in the prediction of the class label () for a given observation based on its features ().

In a mathematical context, Bayes' theorem can be formally represented as:

* is the posterior probability of class given the features
* is the likelihood of observing features given class
* is the prior probability of class
* is the marginal likelihood of the features

The appeal of the Naïve Bayes Classifier rests in its inherent simplicity. The assumption being made is that the features are conditionally independent given the class label , which can be considered "naïve".

Put simply, this assumption posits that the knowledge of one feature's value does not yield any information regarding the value of another feature, given that the class label is already known. The likelihood term is simplified by this approach.

In this context, the variable denotes specific characteristics, while can be approximated using the available training data. Prominent methodologies for estimating encompass Maximum Likelihood estimating and kernel density estimation, which offer non-parametric means to approximate probability densities.

# Challenges with Noisy and Tough Datasets

The Naïve Bayes Classifier has demonstrated its effectiveness across various applications; yet, it does has limits, particularly when confronted with datasets that are noisy or provide challenges. A notable constraint of the method is its susceptibility to outliers and deviations from the premise of independence. In practical scenarios, datasets derived from real-world sources frequently exhibit the presence of noisy or linked features, hence challenging the premise of naïve independence. When this assumption is not valid, the performance of Naïve Bayes Classifier may be negatively impacted.

The objective of our study is to tackle these obstacles by augmenting the Naïve Bayes Classifier with robust Kernel Density Estimation approaches. Our objective is to enhance the robustness and accuracy of the Naïve Bayes Classifier by utilising robust statistical methods such as M-estimation with the Hampel function to improve the estimation of

In the following sections, we will explore the approach employed to strengthen the optimization of RKDE. This will involve the integration of Harris Hawk Optimisation for the purpose of selecting bandwidth, and an assessment of the effects of these changes on the performance of classification.

The objective of this thesis is to present a thorough analysis of the application of RKDE to the Naïve Bayes Classifier. This research aims to enhance the classification capabilities of the Naïve Bayes Classifier, particularly in situations where the assumptions of data quality and independence are called into question.

# Methodology

This section offers a thorough examination of the approaches utilised to enhance the performance of the Naïve Bayes Classifier through the implementation of Robust Kernel Density Estimation in conjunction with Iterative Reweighted Least Squares and Harris Hawk Optimisation for the purpose of selecting the optimal bandwidth. The utilization of these approaches is of utmost importance in improving the resilience and predictive accuracy of the Naive Bayes Classifier when confronted with noisy and complex datasets frequently found in the analysis of intelligence data.

# Robust RKDE with IRLS

The algorithm employed in the adaption of Iteratively Reweighted Least Squares (IRLS) and Robust M-estimation to Robust Kernel Density Estimation (RKDE) is as follows:

|  |
| --- |
| **Algorithm : RKDE with IRLS and Robust M-estimation (Hampel function)** |
| Input: Training dataset D, Bandwidth parameter h  Output: RKDE model |
| 1. Initialization: The process commences by initializing the RKDE model with Gaussian kernel parameters. Additionally, we initialize weights w for each data point in D, which will play a pivotal role in the robustness of the density estimation. |
| 2. Weight Adjustment Iteration: The core of the algorithm consists of iterative weight adjustments. During each iteration, we:  a. Update the weights w: This is achieved using IRLS based on the Hampel function. The Hampel function identifies outliers within the dataset and subsequently down-weights them. This process significantly enhances the robustness of the density estimation.  b. Recalculate the RKDE model: With the updated weights, we recalculate the RKDE model. This model reflects the density distribution while accounting for the identified outliers. |
| 3. Convergence: The iteration continues until convergence is achieved. |
| 4. Return: The final RKDE model is returned, which encapsulates the robust density estimation capable of handling challenging datasets. |

Section 3.1 provides a comprehensive analysis of the computational complexities involved in accurately estimating the Kernel Density utilizing the resilient RKDE approach with the Iteratively Reweighted Least Squares technique. The methodology entails an iterative approach wherein the weights of data points are adjusted using the Hampel function. The Hampel function assumes a crucial role in the identification and subsequent reduction of outliers, hence improving the resilience of the density estimation process.

# Harris Hawk Optimization (HHO) for bandwidth selection

# Harris Hawk Optimization

The swarm intelligence algorithm is a widely used method in the field of computational intelligence and a developing technique in evolutionary computing. Its fundamental principle is to replicate the behaviour of various organisms in nature, such as ants, birds, bees, wolves, bacteria, and others. By leveraging the mechanisms of interaction and information exchange among these groups, swarm intelligence aims to harness collective intelligence to address intricate problems. Extensive research has been conducted by numerous scholars in the field of intelligent algorithms, resulting in the proposal of various innovative algorithms. These algorithms include the grey wolf optimizer (GWO), lightning search algorithm (LSA), marine predators algorithm (MPA), sine cosine algorithm (SCA), salp swarm algorithm (SSA), water cycle algorithm (WCA), whale optimisation algorithm (WOA), cuckoo search (CS), artificial bee colony (ABC), and moth flame optimization (MFO). The Harris Hawks Optimisation (HHO) method is a recently developed swarm intelligence system introduced. The inspiration for this algorithm stemmed from the observation of the pursuit and evasion behaviour exhibited by Harris hawks in their interactions with prey. The algorithm's simplicity in principle, along with its limited number of parameters and robust global search capacity, has garnered significant interest and widespread use across various engineering disciplines since its inception. However, akin to other intelligent optimisation algorithms, the fundamental Harris hawks algorithm is prone to some limitations, such as a diminished level of accuracy in achieving convergence and a tendency to become trapped in local optima while attempting to solve intricate optimisation issues. [33]

In essence, the primary enhancement of the HHO algorithm lies in its capacity to enhance the optimisation capabilities of local exploitation and global exploration by employing diverse optimisation strategies. This, in turn, leads to an improvement in the algorithm's convergence accuracy and overall performance, rendering it suitable for practical engineering applications.

Kernel density estimation is a machine learning technique that can be effectively enhanced by the utilization of a strong tool for parameter optimization.

# Description HHO

The HHO algorithm is a mathematical formulation that emulates the hunting technique of Harris hawks. In this algorithm, individual Harris hawks represent candidate solutions, and the prey is represented by the optimal solution obtained in each iteration. The algorithm consists of two primary stages, specifically exploration and exploitation. The transition between these phases is determined by the size of the prey's escape energy. The following section provides a description of the original Harris hawks optimisation algorithm.

# Exploration phase

The global search phase is mostly influenced by the geographical data pertaining to the population of Harris hawks, and its update approach can be described as follows:

In the given context, 𝑋(𝑡+1) denotes the spatial coordinates of the hawks in the iteration 𝑡+1. 𝑋𝑝𝑟𝑒𝑦(𝑡) represents the spatial coordinates of the prey. 𝑋(𝑡) signifies the spatial position of the hawks in the current generation 𝑡. The variables 𝑟1–𝑟4 and q are randomly generated values within the range of (0,1), and are renewed during each iteration. 𝑈𝐵 and 𝐿𝐵 represent the upper and lower bounds of the population, respectively. 𝑋𝑟𝑎𝑛𝑑(𝑡) refers to a randomly selected hawk from the current population. 𝑋𝑚(𝑡) denotes the average spatial position of individuals in the current population, which is calculated using Equation (9).

Where: represents the positional value of hawk at iteration , whereas n represents the total number of hawks.

# Transition from Exploration to Exploitation

The equation governing the escape of prey is represented by the energy equation.

Where: represent the current iteration count, represent the maximum number of iterations, and symbolise a randomly generated number within the range of (-1,1) that signifies the starting energy state. When the magnitude of the escape energy |𝐸| is greater than or equal to 1, Harris hawks engage in a global exploration phase by searching various locations in order to locate the prey. On the other hand, when the magnitude of |𝐸| is less than 1, Harris hawks engage in a local exploitation phase by exploring nearby solutions.

# Exploitation Phase

During this stage, the Harris hawk will engage in a predatory behaviour towards its target prey upon locating it, drawing from the findings of the preceding phases. Simultaneously, the prey will attempt to evade the pursuit. Four potential tactics are suggested for this portion of the simulation, based on the observed behaviour of Harris's hawk and its prey. The simulated strong beseige and gentle besiege behaviours exhibited by Harris's hawk are replicated by E. The parameter "r" is designated to reflect the outcome of the prey's escape attempt.

1. **Soft Beseige**

In situations when the absolute value of the energy parameter (|𝐸|) is more than or equal to 0.5, and the distance parameter (𝑟) is greater than or equal to 0.5, the prey organism exhibits a behavioural response of attempting to evade the predator by means of jumping. Concurrently, the Harris hawk predator employs a strategy of employing a gentle and steady approach to deplete the prey organism's energy reserves. The behaviour is represented in the following manner:

𝑋(𝑡+1)=Δ𝑋(𝑡)−𝐸|𝐽𝑋𝑝𝑟𝑒𝑦(𝑡)−𝑋(𝑡)| Eq.(11)

Δ𝑋(𝑡)=𝑋𝑝𝑟𝑒𝑦(𝑡)−𝑋(𝑡) Eq.(12)

Where 𝑟5 is used to denote a number that is created randomly from a uniform distribution between 0 and 1. The variable J is incorporated into the model to mimic the characteristics of prey movement, and its value is subject to random variation in each iteration.

1. **Hard Beseige**

When the absolute value of E is less than 0.5 and r is greater than or equal to 0.5, the prey lacks the energy to flee. Consequently, the Harris hawk engages in an aggressive siege approach, employing Equation (13) to modify its present location.

𝑋(𝑡+1)=𝑋𝑝𝑟𝑒𝑦(𝑡)−𝐸|Δ𝑋(𝑡)| Eq.(13)

1. **Soft Beseige with Progressive Rapid Dives**

When the absolute value of 𝐸 is greater than or equal to 0.5 and 𝑟 is less than 0.5, it can be concluded that the prey possesses sufficient energy to successfully evade the predator. The Harris hawks will adjust their locations in accordance with the governing principle outlined in Equation (14):

𝑌=𝑋𝑝𝑟𝑒𝑦(𝑡)−𝐸|𝐽𝑋𝑝𝑟𝑒𝑦(𝑡)−𝑋(𝑡)| Eq.(14)

𝑍=𝑌+𝑆×𝐿𝐹(𝐷) Eq.(15)

In the context of the problem, *D* represents the dimensionality of the system under consideration. *S* denotes a random vector with dimensions 1*×D*. Additionally, *LF* refers to the levy flight function, which may be mathematically expressed as shown in Equation (16).

where u and v are random values within (0,1) and 𝛽 is the default constant, set to 1.5.

Therefore, the implementation of Equation (17) can serve as the ultimate approach for updating the positions of the hawks during the soft siege phase.

where Y and Z are obtained using Equations (11) and (12), respectively.

1. **Hard Beseige with Progressive Rapid Dives**

When the absolute value of E is less than 0.5 and r is less than 0.5, it can be concluded that the prey lacks sufficient energy to successfully escape. In such circumstances, a specific strategy is established to be implemented.

𝑌=𝑋𝑝𝑟𝑒𝑦(𝑡)−𝐸|𝐽𝑋𝑝𝑟𝑒𝑦(𝑡)−𝑋𝑚(𝑡)| *Eq.(19)*

𝑍=𝑌+𝑆×𝐿𝐹(𝐷) Eq.(20)

where is obtained using Equation (9).

# The Main Steps of HHO

The main steps of the overall HHO algorithm are as shown in the following algorithm.

|  |
| --- |
| **Algorithm : Main steps of HHO algorithm** |
| **Input**: Population size N and the maximum number of iterations *T* |
| 1: Initialize the population  2: **while** 𝑡<𝑇 **do**  3: Calculate the fitness of each solution and get the optimal individual  4: **for** i=1:N **do**  5: According to Equation (3) update the escape energy E  6: **if** |𝐸|≥1 **then**  7: According to Equation (1) update the location  8: **else** **if** 𝐭𝐡𝐞𝐧|𝐸|<1  9: **if** |𝐸|≥0.5 and 𝑟≥0.5 **then**  10: According to Equation (4) update the location  11: **else if** 𝐭𝐡𝐞𝐧|𝐸|<0.5 and 𝑟≥0.5  12: According to Equation (6) update the location  13: **else** **if** 𝐭𝐡𝐞𝐧|𝐸|≥0.5 and 𝑟<0.5  14: According to Equation (10) update the location  15: **else if** 𝐭𝐡𝐞𝐧|𝐸|<0.5 and 𝑟<0.5  16: According to Equation (11) update the location  17: **end if**  18: **end if**  19: **end for**  20: 𝑡= 𝑡+1  21:**end while**  22: **return** 𝑋𝑝𝑟𝑒𝑦 |

# Selecting Bandwidth with HHO

The HHO algorithm is a contemporary evolutionary optimisation technique that draws inspiration from the hunting strategies employed by harris hawks. Kernel density estimation is a machine learning technique that can be significantly enhanced with the utilisation of a potent tool for parameter optimisation.

|  |
| --- |
| **Algorithm : Harris Hawk Optimization for Bandwidth Selection** |
| **Input**: RKDE model, Dataset *D*  **Output**: Optimized bandwidth parameter *h* |
| 1. Initialization: HHO begins by initializing a population of hawks with random positions. |
| 2. Fitness Evaluation: The fitness of each hawk is evaluated based on the performance of the RKDE model using Cross-Validation. This fitness assessment guides the search for an optimal bandwidth. |
| 3. Leader Identification: The algorithm identifies the best-performing hawk (leader) in the population, based on fitness. |
| 4. Iteration: The optimization process unfolds over a predefined number of generations. Within each generation, the following steps are executed:  a. Position Update: The positions of hawks are updated using specific position update rules.  b. Fitness Re-evaluation: After position updates, the fitness of each hawk is re-evaluated.  c. Leader Update: The leader hawk is updated based on fitness. This leader guides the exploration of the solution space.  d. Exploration and Exploitation: The algorithm fine-tunes its exploration and exploitation parameters based on the observed fitness landscape. |
| 5. Result: The optimized bandwidth parameter h is determined based on the leader hawk's findings throughout the optimization process. |

In Subsection 3.2, a detailed exposition is presented on the utilisation of Harris Hawk Optimisation for the purpose of optimising the bandwidth parameter in RKDE. The proposed algorithm systematically investigates the solution space in order to determine an ideal bandwidth that greatly improves the resilience and predictive accuracy of the RKDE model.

This section presents the fundamental strategies employed in this study to modify RKDE and Naïve Bayes classifier in order to properly address datasets that are noisy and demanding. The incorporation of these approaches plays a pivotal role in fulfilling the aims of the thesis, which aim to get reliable and precise categorization in the study of intelligence data. The next sections will delve more into the experimental configuration, findings, and ramifications of these approaches.

# Complete Cross-Validation Curve

This section presents the Complete Cross-Validation Curve methodology, which is an essential element of our research that enables the selection of bandwidth in Robust Kernel Density Estimation. The Complete Cross-Validation methodology encompasses three essential components: Unbiased Cross-Validation, Biassed Cross-Validation and Bootstrap. Every individual component has a substantial role in enhancing the resilience and accuracy of bandwidth selection in the context of RKDE.

# Unbiased Cross-Validation

The Unbiased Cross-Validation technique, often referred to as Leave-One-Out Cross-Validation (LOOCV), is the basis of CCV. The technique involves iteratively deleting individual data points and assessing the model's performance against the remaining data in order to evaluate its effectiveness. The mathematical representation of this procedure is as follows:

Where: *n* is the number of data points, *K(·)* is the kernel function, and are data points, *h* is the bandwidth, a parameter to be selected.

The BCV bandwidth selector was originally introduced by Scott and Terrell in 1987. Jones and Kappenman (1992) provided a theoretical analysis comparing the Local Scoring Cross-Validation (LSCV), Biassed Cross-Validation (BCV), and other bandwidth selections.

One of the notable advantages of Unbiased Cross-Validation (UCV) is its ability to provide unbiased estimates of model performance.

Unbiased Evaluation: The use of Unbiased Cross-Validation (UCV) allows for an impartial assessment of the model's performance. This is achieved by deliberately excluding specific data points from the analysis and evaluating the model's predictive ability using the remaining data. This practise guarantees an equitable assessment.

The UCV technique is founded on rigorous mathematical concepts, rendering it a dependable method for evaluating the performance of models.

The method exhibits a low computing cost as it necessitates less calculations in comparison to alternative cross-validation techniques. This characteristic renders it computationally efficient, particularly when dealing with extensive datasets.

Limitations of Unbiased Cross-Validation (UCV):

High volatility: The Unbiased Cross-Validation (UCV) method has the potential to demonstrate a significant amount of volatility, particularly in situations when the dataset is limited in size. The outcomes may exhibit substantial variability based on the exclusion of specific data points in each iteration.

Lack of Robustness to Outliers: The Unweighted Covariance (UCV) method may exhibit sensitivity to outliers due to its equal weighting of all data points during the evaluation process. Outliers has the potential to exert a disproportionate influence on the outcomes.

# Biased Cross-Validation (BCV)

# Bootstrap

# Importance of using CCV

# Integration of Robust KDE and Bandwith selection into Naïve Bayes classifier

# Experimental Setup

# Dataset Description

# Implementation Details

# Performance metrics

# Results and Discussion

# Evaluation of Optimized Robust KDE

# Impact of Bandwidth selection on classification performance

# Comparison with traditional Naïve Bayes Classifier

# Conclusion and Future Work

# Summary of findings

# Contributions to the field

# Limitation and Future direction

# References

[1] Weglarczyk,S.(2018).Kerneldensityestimationanditsapplication.*ITM Web of Conferences*, 23, 00037.

[2] Chevallier, E., Kalunga, E., & Angulo, J. (2017). Kernel Density Estima- tion on Spaces of Gaussian Distributions and Symmetric Positive Definite Matrices. *SIAM Journal on Imaging Sciences*, 10(1), 191-215.

[3] Domingos, P., & Pazzani, M. (1996). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3), 103- 130.

[4] Pe ́rez, A., Larran ̃aga, P., & Inza, I. (2009). Bayesian classifiers based on kernel density estimation: Flexible classifier. *International Journal of Ap- proximate Reasoning*, 50, 341-362.

[5] Scott, D. W. (2015). Multivariate density estimation: *theory, practice, and visualization* (2nd ed.). John Wiley & Sons.

[6] Wand, M. P., & Jones, M. C. (1995). Kernel smoothing (Vol. 60). *CRC press*.

[7] Kim,J.,&Scott,C.(2008).RobustKernelDensityEstimation.In*Proceed- ings of the 22nd International Conference on Neural Information Processing Systems (NIPS)* (pp. 3381-3384).

[8] Maronna,R.A.,Martin,D.,&Yohai,V.J.(2006).RobustStatistics:*The- ory and Methods*. John Wiley Sons.

[9] Huber, P. J. (1981). Robust Statistics. John Wiley Sons.  
[10] Rousseeuw, P. J., & Leroy, A. M. (2005). Robust Regression and Outlier

Detection. John Wiley & Sons.  
[11] Huber,P.J.,&Ronchetti,E.M.(2009).RobustStatistics(2nded.).John

Wiley & Sons.  
[12] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986).

Robust Statistics: The Approach Based on Influence Functions. John Wiley

& Sons.  
[13] Gallego,J.A.,Osorio,J.F.,&Gonzalez,F.A.(2022).FastKernelDen-

sity Estimation with Density Matrices and Random Fourier Features. Mind-

Lab.  
[14] Liu,B.,Tan,P.N.,&Zhou,J.(2022).UnsupervisedAnomalyDetection

by Robust Density Estimation. In *Proceedings of the 36th AAAI Conference*

*on Artificial Intelligence (AAAI-22)* (pp. 4101-4108).  
[15] Humbert, P., Le Bars, B., & Minvielle, L. (2022). Robust kernel density

estimation with median-of-means principle. In *International Conference on*

*Machine Learning* (pp. 9444-9465). PMLR.  
[16] Vandermeulen, R. A., & Scott, C. (2014). Robust kernel density estima-

tion by scaling and projection in Hilbert space. In *Advances in Neural Infor-*

*mation Processing Systems (NIPS)* (pp. 3476-3484).  
[17] Calonico, S., Cattaneo, M. D., & Farrell, M. H. (2019). nprobust: Non-

parametric kernel-based estimation and robust bias-corrected inference.

arXiv preprint arXiv:1906.00198.  
[18] Wang, D., & Li, M. (2017). Robust Stochastic Configuration Networks

with Kernel Density Estimation. arXiv preprint arXiv:1702.04459.  
[19] Wu,X.(2019).Robustlikelihoodcross-validationforkerneldensityesti- mation. *Journal of Business* & *Economic Statistics*, 37(4), 761-770. ———

–  
[20] Wang, H., Mirota, D., & Hager, G. D. (2009). A generalized kernel

consensus-based robust estimator. *IEEE Transactions on Pattern Analysis*

*and Machine Intelligence*, 32(1), 178-184.  
[21] Boedihardjo, A. P., Lu, C. T., & Chen, F. (2015). Fast adaptive kernel

density estimator for data streams. *Knowledge and Information Systems*, 42,

285-317.  
[22] Chen, Z., Fang, Z., Sheng, V., Zhao, J., Fan, W., Edwards, A., & Zhang,

K. (2021). Adaptive robust local online density estimation for streaming data. *International Journal of Machine Learning and Cybernetics*, 12, 1803- 1824.

[23] Seyedali Mirjalili, Andrew Lewis “The Whale Optimization Algorithm” *Advances in Engineering Software* 95 (2016) 51–67.

[24] Zhang, W., Zhang, Z., Chao, H. C., & Tseng, F. H. (2018). Kernel mix- ture model for probability density estimation in Bayesian classifiers. *Data Mining and Knowledge Discovery*, 32, 675-707.

[25] Kim,J.,&Scott,C.(2009).L2KernelClassification.*IEEETransactions on Pattern Analysis and Machine Intelligence*, 32(10), 1822-1831.

[26] Majdara, A., & Nooshabadi, S. (2019). Online density estimation over high-dimensional stationary and non-stationary data streams. *Data* & *Knowledge Engineering*, 123, 101718.

[27] Chen J.-Y. Research on window parameter determination method based on particle swarm optimisation algorithm. *(Doctoral dissertation, Hebei University)*.

[28] Ye, S. L., Chen, X., Xiong, Xiong Xiong, Kan, Y. J., & Wang, Z. P.. (2020). Analysis of surface air temperature in Jiangsu Province based on particle swarm improved adaptive kernel density estimation al- gorithm.*Science, Technology* & *Engineering* , 20(36), 8.

[29] Hayyolalam, V., & Kazem, A. A. P. (2020). Black widow optimization algorithm: a novel meta-heuristic approach for solving engineering opti- mization problems. *Engineering Applications of Artificial Intelligence*, 87, 103249.

[30] Feng, W. T., & Deng, B.. (2021). Study on global convergence analysis and parameter selection of whale optimisation algorithm. *Control Theory and Applications*.

7

[31] Xu, D. G., Wang, Z. Q., Guo, Y. X., & Xing, K. J.. (2023). A review of whale optimisation algorithms. *Computer Applications Research*, 40(2), 328-336

[32] Jia L., Liu X. N., He M., & Song H. C.. (2022). A review of intelligent optimisation algorithms for quantum swarms. *Computer Engineering and Applications*, 58(7), 31-42.

[33] Xu, P.. Improvement of the Harris Hawk optimisation algorithm and its applications. *(Doctoral dissertation, Xi’an University of Technology)*.

[34] Zhan,GuanghanWang,YuhongLiu,Hao.(2022).ImprovedHarrisHawk optimisation algorithm with mixed strategies and its applications. Journal of Sensing Technology, 35(10), 1394-1403.

# Appendices

# List of papers published by the author during his degree study

# Python code display