

湖北工业大学

国际学生硕士学位论文开题报告

Master Thesis Proposal of International Students at HBUT

论文题目(Thesis Title): Naïve Bayes classifier based on
improved RKDE for dataset classification with outliers

研究生姓名(Name of Graduate Student): BOLI BI IRITIE
ARNAUD-DONATIEN

学号(Student ID No.): 2121101101

导师姓名(Name of Supervisor): CHENGHAO WEI

系/所(Department of Student): School of Computer Science

专业(Major): Computer Science and Technology

研究方向(Research Area): Non-parameters estimation for
machine learning

入学时间(Date of Matriculation):_____

毕业时间(Date of Graduation):_____

注：本表可复印，可另加附页

Note: Additional pages can be attached in accordance with the given format.

一、立论依据

I . Proposal Rationale

(论文的研究意义、国内外研究现状分析、附主要参考文献)

(Research Significance, analysis of the current situation of research both at home and abroad, and major bibliography)

Finding outliers in huge data sets is a significant and fascinating topic in knowledge discovery. Almost all current research on outlier detection is based on density-based methods and has been successful because of their simplicity, usability, and capacity to detect a variety of anomalies under different conditions. Examples of these methods include Kernel Density Estimator (KDE), local outlier factor (LOF), and their variants [16]. The most popular nonparametric density estimators for multivariate data are probably KDEs. They are crucial components in researchers' toolkits for statistical data analysis, data mining, and machine learning [17, 18] and serve as the foundation for a variety of approaches for classification, clustering, and level set estimation.

In these varieties of approaches for the same goal of finding outliers, flexible classifiers trying to fit the data closer and with the large sample size and use the principle of winner-takes-all based on the Bayesian Network [1], make the KDE more efficient by employing a kernel approximation produced by random Fourier features and density matrices [3]. Even many typical applications of KDE had been developed as symmetric and asymmetric kernel functions [4], using symmetric positive definite matrices and different spaces [5,6] and for the unlabeled dataset other techniques had been applied like Robust Real-valued non-volume preserving (RobustRealNVP) [7]. In dealing with large datasets the Median-of-Means principle has been combined with KDE [8] to reduce the time of computation and make the robust KDE by optimizing the window's width, Robust Likelihood Cross-validation (RLCV) [9] or Robust Stochastic Configuration Networks (RSCNs) [10] are good examples between many [11, 12,13, 15]. Furthermore, under the right circumstances on the bandwidth going to zero, they are well-known to be consistent density estimators [17, 19]. But these methods have some insufficiencies in the purpose of outlier detection [20].

References

- [1] Aritz Pérez , Pedro Larrañaga, Iñaki Inza, Bayesian classifiers based on kernel density estimation: Flexible classifier, International Journal of Approximate Reasoning 50 (2009) 341-362.
- [2] JooSeuk Kim and Clayton Scott, Robust Kernel Density Estimation, (2008) 3381-3384. (Pas terminé)
- [3] Joseph A. Gallego M, Juan F. Osorio, Fabio A. Gonzalez, Fast Kernel Density Estimation with Density Matrices and Random Fourier Features, MindLab (2022)
- [4] Stanislaw Weglarczyk, Kernel density estimation and its application, ITM Web of Conferences 23, 00037 (2018)
- [5] Emmanuel Chevalliera, Emmanuel Kalungab, Jesu's Anguloa, Kernel Density Estimation on Spaces of Gaussian Distributions and Symmetric Positive Definite Matrices, SIAM Journal on Imaging Sciences, Society for Industrial and Applied Mathematics, 2017, 10 (1), pp.191 - 215.10.1137/15M1053566.hal-01535731
- [6] Vandermeulen, Robert A., and Clayton Scott. "Robust kernel density estimation by scaling and projection in hilbert space." Advances in Neural Information Processing Systems 27 (2014)

- [7] Boyang Liu, Pang-Ning Tan, Jiayu Zhou, Unsupervised Anomaly Detection by Robust Density Estimation, The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22), (2022) 4101-4108
- [8] Humbert, P., Le Bars, B., & Minvielle, L. (2022, June). Robust kernel density estimation with median-of-means principle. In International Conference on Machine Learning (pp. 9444-9465). PMLR (2020)
- [9] Wu, X. (2019). Robust likelihood cross-validation for kernel density estimation. *Journal of Business & Economic Statistics*, 37(4), 761-770.)
- [10] Wang, D., & Li, M. (2017). Robust Stochastic Configuration Networks with Kernel Density Estimation. arXiv preprint arXiv:1702.04459.
- [11] Zhang, X., King, M. L., & Shang, H. L. (2014). A sampling algorithm for bandwidth estimation in a nonparametric regression model with a flexible error density. *Computational Statistics & Data Analysis*, 78, 218-234.
- [12] 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.
- [13] 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.
- [14] Hu, S., Poskitt, D. S., & Zhang, X. (2012). Bayesian adaptive bandwidth kernel density estimation of irregular multivariate distributions. *Computational Statistics & Data Analysis*, 56(3), 732-740.
- [15] Boedihardjo, A. P., Lu, C. T., & Chen, F. (2015). Fast adaptive kernel density estimator for data streams. *Knowledge and Information Systems*, 42, 285-317.
- [16] Zhang, W., Zhang, Z., Chao, H. C., & Tseng, F. H. (2018). Kernel mixture model for probability density estimation in Bayesian classifiers. *Data Mining and Knowledge Discovery*, 32, 675-707.
- [17] Kim, J., & Scott, C. (2009). L_2 Kernel Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10), 1822-1831.
- [18] Calonico, S., Cattaneo, M. D., & Farrell, M. H. (2019). nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference. arXiv preprint arXiv:1906.00198.
- [19] Majdara, A., & Nooshabadi, S. (2019). Online density estimation over high-dimensional stationary and non-stationary data streams. *Data & Knowledge Engineering*, 123, 101718.

二、研究方案

II. Research Program

1. 研究目标、研究内容和拟解决的关键问题

1. Research objectives, research contents and intended key research questions

Facing the problem of dealing with outliers in a large dataset, our research objectives should be cleaning the outliers by using an evolutionary algorithm.

Our research contents are:

- Improve the performance of basic Naïve Bayes classifier with noise
- Optimize the windows size by evolutionary organization algorithm
- Reduce the weight of the outliers to zero

2. 拟采取的研究方法、技术路线、实验方案及可行性分析

2. Research methodology, technical approaches, experimental plans adopted, and feasibility analysis

Evolutionary optimization method: Hares eagle optimization method

Objective function: UCV for KDE

Cost function: p-Huber

3. 本论文的特色与创新之处

3. Characteristics and innovations of this research

Our work tries to use a novel genetic algorithm for searching the optimized window width and we implement the Naïve Bayes classifier based on improved RKDE for dataset classification with the outliers method to tackle the problem of outliers. The principle of RKDE is that the outliers can significantly affect the sample mean, which may reduce the performance of the whole learner, so we need to reduce the weight of outliers until they reach 0. To fulfil our goal, we'll use the Iteratively Re-Weighted Least Square (IRWLS) algorithm for improving the performance of the algorithm with optimized parameters.

Our work contributions are as follows:

- 1- Optimized window width by using a novel genetic algorithm
- 2- Using the P-Huber function to reduce the side bound effect of RKDE with the Iteratively Re-Weighted Least Square (IRWLS)

4. 预期的论文进展和成果

4. Expected schedule and achievements of this research

TIME	ACHIEVEMENT
Sep 2022 – March 2023	Proposal thesis
March 2023 – May 2023	Start writing paper and thesis
Jan 2024	Final defense

三、论文大纲

III. Thesis Outline

1. Introduction
 - 1.1. Research background and significance
 - 1.2. Research status at home and abroad
 - 1.2.1. Research status of two-stage distributed robust optimization and its uncertainty set
 - 1.2.2. Research status of kernel density estimation
 - 1.3. Main contents of this paper
 - 1.4. Main innovations of the research
2. Naïves Bayes Classifier Algorithm and Robust Kernel Density Estimation
 - 2.1. Definition
 - 2.1.1. Why it's called Naïves Bayes
 - 2.1.2. Bayes Theorem
 - 2.1.3. Types of Naïves Bayes Classifier
 - 2.2. Gaussian Kernel density estimation function with variable parameters
 - 2.2.1. Gaussian Kernel density estimation function with variable parameters
 - 2.2.2. Proof of consistency of Gaussian Kernel density estimation function with variable parameters
 - 2.2.3. Selection of smoothing parameters of Gaussian Kernel density estimation function with variable parameter
 - 2.3. Robust Kernel density estimation function
 - 2.3.1. Robust Gaussian Kernel density estimation function
 - 2.3.2. Proof of consistency of Robust Gaussian Kernel density estimation
 - 2.3.3. Weight calculation of robust Gaussian Kernel density estimation function
 - 2.3.4. Performance estimation of robust Gaussian kernel density estimation method
 - 2.4. Improved robust Kernel density estimation function
 - 2.5. Data-driven uncertainty set based on improved robust Gaussian Kernel density estimation function
 - 2.6. Summary of this chapter
3. Simulation experiment
 - 3.1. Effect of different smoothing parameters on kernel density estimation
 - 3.2. Comparison of kernel density estimation and variable parameter kernel density estimation
 - 3.3. Comparison between improved robust kernel density estimation and other kernel density estimation
4. Summary and prospect
5. Reference
- Attachment
 - A. List of papers published by the author during his degree study
 - B. Python code display of this paper
 - C. Dissertation dataset
- To thank

四、研究基础

IV. Research Basis

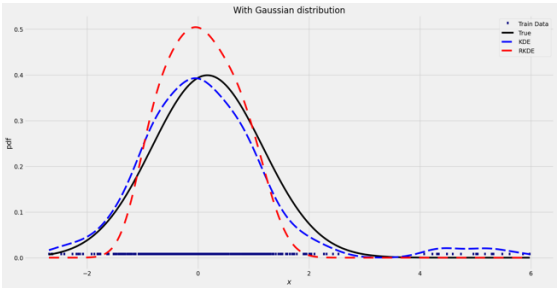
1. 已参加过的有关研究工作和已取得的研究工作成绩

1. Previous research work and research achievements

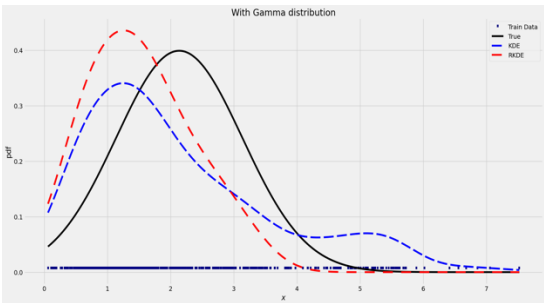
The research work is about implementing a revolutionary algorithm of naives bayes classifier based on improved RKDE for dataset classification with outliers and we achieve reducing the weights of the outliers in the datasets.

Some results:

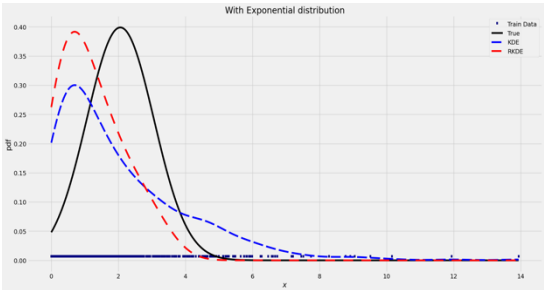
With Gaussian distribution



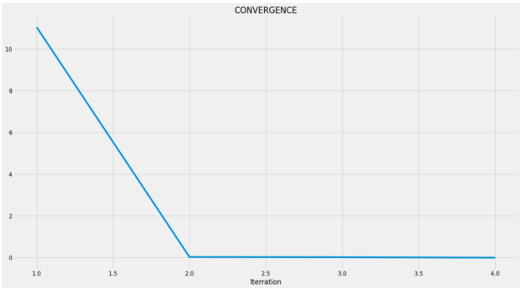
With Gamma distribution



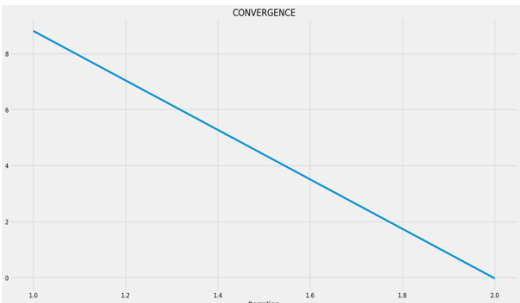
With Exponential distribution



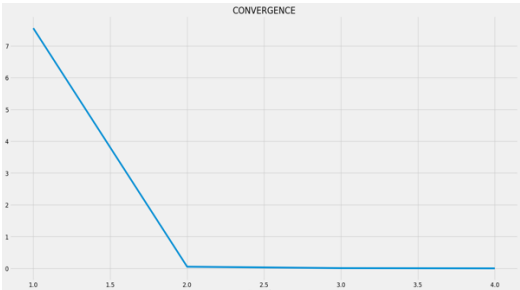
CONVERGENCE



CONVERGENCE



CONVERGENCE



2. 已具备的实验条件，尚缺少的实验条件和拟解决的途径

2. Accessible and inaccessible experimental conditions, and solutions to the difficulties

MacBook Pro (Retina, 25-inch, Late 2013)

Hardware Overview:

Model Name:	MacBook Pro
Model Identifier:	MacBookPro11,3
Processor Name:	Quad-Core Intel Core i7
Processor Speed:	2.3 GHz
Number of Processors:	1
Total Number of Cores:	4
L2 Cache (per Core):	256 KB
L3 Cache:	6 MB
Hyper-Threading Technology:	Enabled
Memory:	16 GB
System Firmware Version:	433.140.2.0.0

五、导师或指导小组意见

V. Comments of the Supervisor or the Committee

导师签名(Signature of Supervisor) :

日期(Date):