Petra Perner (Ed.)

Advances in Data Mining

Applications and Theoretical Aspects

20th Industrial Conference on Data Mining, ICDM 2020
Amsterdam, The Netherlands
July 20 – July 21, 2020

Poster Proceedings

ibai Publishing
www.ibai-publishing.org

Volume Editor

Petra Perner
Institute of Computer Vision and Applied Computer Sciences, IBaI
PF 30 11 14
04251 Leipzig
E-mail: pperner@ibai-institut.de

P-ISSN 1864-9734
E-ISSN 2699-5220
ISBN 978-3-942952-77-4

The German National Library listed this publication in the German National Bibliography.
Detailed bibliographical data can be downloaded from http://dnb.ddb.de.

ibai-publishing
Prof. Dr. Petra Perner
PF 30 11 38
04251 Leipzig, Germany
E-mail: info@ibai-publishing.org
http://www.ibai-publishing.org

P-ISSN 1864-9734
E-ISSN 2699-5220
ISBN 978-3-942952-77-4

20th Industrial Conference on Data Mining ICDM 2020

www.data-mining-forum.de
July 20 – 21, 2020
Amsterdam, The Netherlands

Chair

Prof. Dr. Petra Perner
Institute of Computer Vision and applied Computer Sciences, IBaI

Program Committee

Ajith Abraham	Machine Intelligence Research Labs (MIR Labs), USA
Mohamed, Bourguessa	Universite du Quebec a Montreal - UQAM, Canada
Bernard Chen	University of Central Arkansas, USA
Jeroen de Bruin	University of Applied Sciences JOANNEUM, Austria
Antonio Dourado	University of Coimbra, Portugal
Stefano Ferilli	University of Bari, Italy
Geert Gins	Glaxo Smith Kline, Belgium
Warwick Graco	Australian Tax Office ATO, Australia
Aleksandra Gruca	Silesian University of Technology, Poland
Pedro Isaias	The University of Queensland, Australia
Piotr Jedrzejowicz	Gdynia Maritime University, Poland
Martti Juhola	University of Tampere, Finland
Janusz Kacprzyk	Polish Academy of Sciences, Poland
Mehmed Kantardzic	University of Louisville, USA
Lui Xiaobing	Google Inc., USA
Eduardo F. Morales	National Institute of Astrophysics, Optics, and Electronics, Mexico
Samuel Noriega	Universitat de Barcelona, Spain
Wieslaw Paja	University of Rzeszow, Poland
Juliane Perner	Novartis Institutes for BioMedical Research (NIBR), Switzerland
Rainer Schmidt	University of Rostock, Germany
Moti Schneider	PCCW Global, Greece
Victor Sheng	University of Central Arkansas, USA
Kaoru Shimada	Fukuoka Dental College, Japan
Gero Szepannek	University of Applied Sciences Stralsund, Germany
Joao Miguel Costa Sousa	Technical University of Lisbon, Portugal
Markus Vattulainen	Tampere University, Finnland
Zhu Bing	Sichuan University, China

Preface

The pandemic "Corona" has put us this year before a difficult time. With care we have kept to the hygiene rules not to get an infection with the virus Covid-19. With mask we have got into coaches and trains, have made our purchases or on work worked. Home office was the catchword of these days. The universities and research facilities have maintained only a small emergency company and lectures were held as online lectures. From home we have tried to do our scientific works. In 1-to-1 telephone calls or phone conferences we have organized with our colleagues the work and have discussed important results of the research. Under it the efficiency suffers what is easy to understand.

In the beginning of the pandemic fell the deadline of our conference. Insecurity spread. The figures of the infected persons increased rapidly. The virus spreads out in more and more countries and was further carried by continent to continent. Soon stood the whole world in the spell of Corona. A conference was the last to this in this situation most thought.

In this situation appeared once again which high demands for a scientist are made. It belongs to the job of a scientist that he presents his scientific results in conferences and makes thus his results of a wide public immediately available. A scientist should have well organized his research, should be able to do his scientific tasks and duties in a flexible way, and should have financed his research with suitable financial means. Only those who were meeting these rules could successfully continue in their professional research work.

The best of the best of us are represented with their papers in this volume. They presented themselves personal or in online presentations in the conference. The acceptance rate for the submitted paper of our conference was 33% percent for long paper as well as short papers. Because of many refusals because of missing financial means or other reasons the acceptance rate decreased to few percent. This shows once more the excellent quality of these scientists. Their papers are of most excellent quality and expand the state-of-the-art in an excellent way. The topics of the long papers range from event log file analysis, predictive maintenance, medical application, telecom application, fraud detection to a paper on how we should present the results to stakeholders so that they accept the findings of the data mining methods. The new arising topic we see here is predictive maintenance. All other topics follow the main topics of ICDM but present new excellent results and go over the recent questions to be solved with data mining for the specific applications. The short paper is a fine theoretical paper on optimal kernel density estimation.

The proceedings will be freely accessible as an OPEN-ACCESS Proceedings of a wide public so that, the new acquired knowledge on the different subjects is able to spread around quickly worldwide. You can find the proceedings for long papers and the poster proceedings for short papers at http://www.ibai-publishing.org/html/proceeding2020.php.

In this time, flexibility was a must Because the situation in the USA was still difficult, we have moved the conference to Amsterdam in the Netherlands. Here a variety

of the participants was able to do outward journeys. The ones who could not travel, were online present.

Extended versions of selected papers will appear in the international journal Transactions on Machine Learning and Data Mining (www.ibai-publishing.org/journal/mldm).

We hope to see you in 2021 in New York at the 21th Industrial Conference on Data Mining ICDM (www.data-mining-forum.de) again.

The conference runs under the umbrella of the World Congress on "The Frontiers in Intelligent Data and Signal Analysis, DSA 2021" (www.worldcongressdsa.com), which combines under its roof the following three events: International Conferences Machine Learning and Data Mining MLDM (www.mldm.de) , the Industrial Conference on Data Mining ICDM (www.data-mining-forum.de), and the International Conference on Mass Data Analysis of Signals and Images in Artificial Intelligence and Pattern Recognition with Applications in Medicine, Biotechnology, Chemistry and Food Industry, MDA-AI&PR (www.mda-signals.de).

We will give then the tutorials on Data Mining, Case-Based Reasoning, and Intelligent Image Analysis again (http://www.data-mining-forum.de/tutorials.php) again. The workshops running in connection with ICDM will also be given (http://www.data-mining-forum.de/workshops.php).

We would warmly invite you with pleasure to contribute to this conference. Please come and join us. We are awaiting you.

July, 2020 Petra Perner

Table of Content

Optimal Kernel Density Estimation using FFT based cost function

Kuldeep Jiwani [0000-0002-5637-8833]

Guavus, a Thales company

Abstract. Kernel density estimation (KDE) is an important method in nonparametric learning, but it is highly sensitive to the bandwidth parameter. The existing techniques tend to under smooth or over smooth the density estimation. Especially when data is noisy, which is a common trait of real-world data sources. This paper proposes a fully data driven approach to avoid under smoothness and over smoothness in density estimation. This paper uses a cost function to achieve optimal bandwidth by evaluating a weighted error metric, where the weight function ensures low bias and low variance during learning. The density estimation uses the computationally efficient Fast Fourier Transform (FFT) to estimate the univariate Gaussian kernel density. Thus bringing the computation cost of a single density evaluation from $O(n^2)$ to $O(m \log(m))$, where $m << n$ and m being the grid points of FFT. Based upon simulation results this paper significantly outperforms the de-facto classical methods and the more recent papers over a standard benchmark dataset. The results specially shine apart from the recent and classical approaches when data contains significant noise.

Keywords: Bandwidth estimation, Binning, Fast Fourier Transform, Kernel Density Estimation, Weighted error estimation, Cost function, Optimisation.

1 Introduction

Kernel density estimation (KDE) is an important method in nonparametric learning, used for determining the natural probability density distribution of data. It has high industry value in detecting anomalies and deviations from natural data distributions. It is more useful when the estimated density is close to the true distribution. As for applications like Anomaly detection a density estimate not close to true distribution will result in many false alarms or will miss anomalies. Density estimation techniques are widely used in Exploratory Data Analysis (EDA), Probability Density Function (PDF) estimation and various inference procedures in statistics and Machine Learning.

The Kernel density estimate is highly sensitive to the bandwidth parameter, which controls the smoothness of the density estimate. Interestingly, the discussions about optimal bandwidth selection techniques have been ongoing for nearly four decades. As it is hard to define a perfect analytical solution for all kinds of datasets. Especially when the data is noisy, as in any typical real word scenario the analytical solutions tends to wave off from the true estimate. This is where we need more of a Machine Learning approach where we can model our density estimate based on the data characteristic's at

hand. Where we also have the liberty to control the bias and variance by choosing the right optimisation metric. The only concerning factor in such approaches is the computation time. So to cater to both of these aspects, this paper proposes a cost function with a weighted error metric that guides the optimisation routine to control both the bias and variance. Then for computational efficiency it uses FFT to evaluate density at m grid points obtained from n sampled points, where the algorithmic cost of FFT is O(m log(m)). Where m << n and m could easily be around 2^7, 2^8, 2^9, etc. The search space is further reduced for faster convergence by using some over smooth analytical bandwidth estimation to set an upper bound.

In the following sections, we will first look at the related work in Section 2. Then in Section 3 we will look at definition and derivation of Kernel Density Estimation. Then in Section 4 we will briefly illustrate the existing popular bandwidth selection schemes in practise. Then in Section 5 we will explain how does a FFT based density estimation technique works. Followed by it, in Section 6 we will present the proposed method of a cost function to obtain the optimal bandwidth. Then from Section 6.1 – 6.3 we will present the rationale behind coming to a weighted error metric for optimising the bandwidth. Finally in Section 7 we will present the experiment and results to show how this method achieves improvement over a few of the existing techniques.

2 Related work

There have been many surveys and studies done on existing bandwidth selection techniques like as done by Jones, Marron and Sheather (1996) [3]. There is a list of bandwidth selection techniques classified as first generation methods inspired by works of Silverman [1] like the Rule of thumb (RT), along with other first generation methods like Least Square Cross Validation (LSCV), Biased Cross Validation (BCV). The listed second generation bandwidth selection techniques are claimed to be much superior than the earlier ones. These involve “solve-the-equation plug-in” method and the “smoothed bootstrap” method. Where the “solve-the-equation plug-in” method by Sheather and Jones [2] has been rated as the best performing method. A later survey done by Heidenreich, Schindler and Sperlich (2013) [4] also reports the earlier techniques and a few variants of smoothened cross-validation approaches. The general view is cross-validation criterion tends to under smooth and suffers from high sample variability. At the same time, the plug-in estimates deliver a much more stable estimate but typically over smooths. Another study done by Berwin [5] also placed Sheather and Jones [2] approach to be better than others. But also highlights the drawback that such approaches use unappealing higher order kernels and need big sample sizes. Thus the recommend Fourier transformation based approaches tend to be more suitable.

Then researchers have shown the effectiveness and efficient speed of Fast Fourier Transform in Kernel Density Estimation. Like Raykar, Duraiswami and Zhao (2010) [6] shows the computational cost of FFT based solution comes down from O(nm) to linear O(n + m), where m evaluation points used over n sample points. Then Suhre, Arikan and Cetin (2016) [7] have used the FFT approach along with a cost function for showcasing improvement from Sheather and Jones [2]. While Suhre, Arikan and Cetin (2016) [7] have used a cost function over a linear combination of FFT based estimator and cross-validation estimator, so as to counter the under-smoothness nature of cross

validation approaches. They do show improvements over Sheather and Jones [2] in some type of distributions but not all. This paper takes a step forward, it is totally based upon FFT based density estimate. For ground truth validation a histogram over the same grid points of FFT is used to compute a weighted error metric, to control bias and variance.

3 Kernel Density Estimation

The task of density estimation is to compute an estimate $\hat{f}$ based on n i.i.d. samples $X_1, \dots \dots, X_n \in \mathbb{R}$ drawn from an unknown density f. The nonparametric density estimator is aimed to estimate f, the density of X without assuming any specific form for f.

The simplest method to estimate density f from an i.i.d. sample $X_1, \dots \dots, X_n \in \mathbb{R}$ is the histogram. It aggregates the data and then uses its relative frequency to approximate the density at $x \in [x_0, x_0 + h)$. Given an origin x_0 and a bandwidth $h > 0$, we can compute the bins of histogram $\{B_k := [x_k, x_{k+1}): x_k = x_0 + hk, k \in \mathbb{Z}\}$ by counting the number of sample points inside each of them. The histogram or the naive density estimator at the point x is defined as

$$\hat{f}_H(x; x_0, h) := \frac{1}{nh} \sum_{i=1}^{n} 1_{\{X_i \in B_k : x \in B_k\}} \tag{1}$$

where 1 is the indicator function denoting counts. If we denote the number of points in B_k as v_k, then the density histogram is

$$\hat{f}_H(x; x_0, h) = \frac{v_k}{nh} \tag{2}$$

The shape of this histogram depends upon (x_0, h). The dependency upon x_0 is undesirable, as changing x_0 will notably change the estimation of f for the given data. An alternative to avoid this dependency on x_0 is the moving histogram. The naive density estimator builds a piecewise constant function by considering the relative frequency of $X_1, \dots \dots, X_n$ inside $(x - h, x + h)$:

$$\hat{f}_N(x; h) := \frac{1}{2nh} \sum_{i=1}^{n} 1_{\{x-h<X_i<x+h\}} = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} 1_{\{-1<\frac{x-X_i}{h}<1\}} \tag{3}$$

In the above expression if we define $K(z) = \frac{1}{2} 1_{\{-1<z<1\}}$ then K is the uniform density in (-1,1).

This brings us to one of the most popular nonparametric methods for density estimation that is the Kernel Density Estimator (KDE). We can replace K by an arbitrary density. Then K is known as a kernel: a density with a certain regularity that is symmetric and unimodal at 0. This generalization provides us the definition of KDE:

$$\hat{f}(x; h) := \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{4}$$

In the naive density estimator, we gave the same weight to all data points $X_1, \dots \dots, X_n$. But logically the points closer to x are more important than the one farther away from it. This opens the way for a wider class of density estimators. Let's use the notation

$K_h(z) := \frac{1}{h}K(\frac{z}{h})$. Then we can also bring in the normal kernel (ϕ), where we can have different weights for different points.
Then $K_h(x - X_i) = \phi(x; X_i, h^2) = \phi(x - X_i; 0, h^2)$, the kernel is the density of a normal distribution $\mathcal{N}(X_i, h^2)$. Thus, the bandwidth h can be thought of as the standard deviation of a normal density whose mean is X_i and the KDE as the data-driven mixture of those densities.

Although there are many types of kernels available like {Epanechnikov, Triangle, etc}. But both theory and practise suggest that if a certain smoothness is guaranteed then the choice of the kernel is not crucial to the statistical performance of the method and therefore it is quite reasonable to choose a kernel for computational efficiency [1]. In this paper, we will only focus on using the standard Gaussian density kernel.

4 Bandwidth Selection Schemes

As the performance of KDE is critically dependent on bandwidth so all automatic bandwidth selection techniques, attempt to minimise the estimation error. For evaluation, it is necessary to choose a distance measure between the true density f and the estimated density $\hat{f}_h(x)$. Since L^2 distances have the advantage that they allow an easier analysis than for example the L^1 distances, that's why most of the work was done using L^2 distances. A popular local error criterion for KDE is the Integrated Square Error (ISE):

$$ISE\left[\hat{f}(.;h)\right] := \int \left(\hat{f}(x;h) - f(x)\right)^2 dx \tag{5}$$

The term ISE is a random variable since it directly depends on the sample $X_1, \ldots \ldots, X_n$. So finding an optimal bandwidth is a hard task as it is dependent on the sample and not just on f and n. To avoid this problem the expected value of ISE is computed, stated as the Mean Integrated Squared Error (MISE):

$$MISE\left[\hat{f}(.;h)\right] := \mathbb{E}\left[ISE\left[\hat{f}(.;h)\right]\right] = \int \mathbb{E}\left[\left(\hat{f}(x;h) - f(x)\right)^2\right] dx \tag{6}$$

Our aim then is to find h that minimises MISE, we need an explicit expression for MISE that can be minimised. An asymptotic expansion can be derived when $h \to 0$ and $nh \to \infty$, resulting in AMISE or Asymptotic MISE. So the optimal bandwidth is:

$$h_{AMISE} = \left[\frac{R(K)}{{\mu_2}^2(K)R(f'')n}\right]^{1/5} \tag{7}$$

where $\mu_2(K) := \int z^2 K(z)dz$ and $R(g) := \int g(x)^2 dx$ for some kernel function K.
As the AMISE depends on $R(f'') = \int (f''(x))^2 dx$ that is the second derivative f'', indicating curvature of the unknown density f. So they cannot be readily applied in practise. To overcome this problem the method of "Plug-in selectors" came to rescue, where they estimate $R(f'')$ by assuming that f is the density of a $\mathcal{N}(\mu, \sigma^2)$. When this is combined with a normal kernel (K) for which $\mu_2(K) = 1$ and $R(K) = \frac{1}{2\sqrt{\pi}}$ then we get the famous rule-of-thumb (RT) for estimation of bandwidth:

$$\hat{h}_{RT} = \left(\frac{4}{3}\right)^{1/5} n^{-1/5}\hat{\sigma} \approx 1.06 n^{-1/5}\hat{\sigma} \quad (8)$$

where the estimation $\hat{\sigma}$ can be chosen as the minimum of standard deviation s or the standard interquartile range. This $\hat{h}_{RT}$ is the default bandwidth used in both Python and R packages.

The rule-of-thumb is an example of a zero-stage plug-in selector as $R(f'')$ was estimated by plugging in a parametric estimation at the very first level. A l-stage plug-in selector iterates these steps l-times before plugging in the normal estimate of the unknown $R\left(f^{(2l)}\right)$. Typically 2 stages are considered a good trade-off between bias (mitigated when l increases) and variance (augments with l) of the plug-in selector. This is the method proposed by Sheather and Jones [2] yielding what is called Direct Plug-In (DPI), it is implemented in R via bw.SJ (method="dpi") routine. There are many more methods of bandwidth estimation like Least Squares Cross Validation (LSCV) and Biased Cross Validation (BCV) both of which use a leave-one-out cross validation strategy, interested readers can read survey papers [3, 4].

A nice summary of the merits and demerits of various bandwidth selection strategy is given by Edurado Garcia Portugues [8]:

- RT is quick and simple but tends to give bandwidths too large for non-normal data
- Cross-validation based selectors are better suited for highly non-normal and rough densities, in which plugin-selectors may end up over smoothing
- DPI in theory is expected to better perform than LSCV, BCV, RT and moreover, it has a faster convergence rate. Hence it tends to be the preferred bandwidth selector in literature.

The problem of estimating bandwidth gets harder if we are working on real-world noisy data. So our focus is to estimate the optimal bandwidth and its KDE from random noisy samples to imitate real-world data. This is where the existing techniques like RT, LSCV, BCV, DPI fall short in modelling the underlying f closely. As for non-normal noisy data, tend to either under-smooth it or over smooth it. An under/over smooth estimated $\hat{f}$ leads to wrong probability/p-value estimations.

5 Numerical methods of density estimation

The traditional density estimation techniques based on Eq. (4) have $O(n^2)$ computation cost. Even the optimal bandwidth selection techniques as described in section 4 have $O(n^2)$ computation cost, as they estimate the general integrated squared density derivative. The most commonly used technique to reduce the computation cost of KDE is to use approximation techniques like binning [1]. The main idea behind binning is to subdivide the interval into an equally spaced mesh of m ($m \ll n$) grid points, $x_0, \dots, x_{m-1}$ and replace the data by grid counts $c_0, \dots, c_{m-1}$, where c_j is a weight chosen to represent the amount of data near x_j. Jones [10] proposed linear binning that works better for density estimations. In this method, the weights are no longer integers, as counts are distributed to bins on either side as per linear interpolation in proportion to the distance of the point from the two nearest bins.

By computing the kernel estimates only on the grid points the computational cost is now of the order of $O(m^2)$, where $m \ll n$. Now if we also use the Fast Fourier Transform (FFT) then it can be further reduced to $O(m\ log(m))$ by performing the discrete convolution.

So in the numerical methods of density estimation, the data is first discretized to a very fine grid. Then the Fast Fourier Transform is used to convolve the data with the kernel for obtaining the density estimate [1]. Then Eq. (4) can be re-written in terms of convolution as:

$$\hat{f}(x;h) = K_h(x) * \frac{1}{N}\sum_{i=1}^{N} \delta(x - X_i) \tag{9}$$

The Fourier transform of Eq. (9) is given by:

$$\hat{F}(\omega;h) = K_h(\omega).\frac{1}{N}\sum_{i=1}^{N} e^{-i\omega X_i} = K_h(\omega).\hat{H}(\omega) \tag{10}$$

where $K_h(\omega)$ and $\hat{H}(\omega)$ are the Fourier transform of the kernel K_h and the data. The implementation of Eq. (10) is carried out using the Discrete Fourier Transform (DFT). If the kernel is chosen as the standard normal $\mathcal{N}(0,\sigma^2)$ i.e. $K_h(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$ then its Fourier transform is also a Gaussian function given by:

$$K_h(\omega) = e^{-2\pi^2 h^2 \omega^2} \tag{11}$$

Plugging Eq. (11) back in Eq. (10) we get

$$\hat{F}(\omega;h) = e^{-2\pi^2 h^2 \omega^2}.\hat{H}(\omega) \tag{12}$$

This is also discretized in the DFT implementation as stated by Silverman. Since the DFT imposes a wraparound edge condition on the convolution, it is important to do the calculation on an interval that is somewhat larger than the interval of interest [1]. Then for bandwidth h, the first grid point x_0 can be computed from vector x as:

$$x_0(x;h) = \min(x) - cut * h \tag{13}$$

where Silverman [1] proposes the value of cut to be 3. We observed that if there are high densities around endpoints then a lower value of cut like 3 sometimes results in underfitting. So it is better to have it proportional to grid size. So

$$cut = 0.015m + 3 \tag{14}$$

worked well for us, where m is the grid size. Plugging Eq. (14) back in Eq. (13)

$$x_0(x;h,m) = \min(x) - (0.015m + 3) * h \tag{15}$$

Finally, taking an inverse Fourier transform of Eq. (12) gives the kernel density estimate of data. The estimated density $\hat{\boldsymbol{f}}(\boldsymbol{x};\boldsymbol{h})$ can be computed as following.

Algorithm FFT_DENSITY$(x; \boldsymbol{h})$:

1. Discretize data $x_0, \dots, x_{n-1}$ to equally spaced grid points $x_0, \dots, x_{m-1}$, along with grid counts as weights $c_0, \dots, c_{m-1}$
2. Compute the FFT of the weights to obtain $\widehat{H}(w)$
3. Compute $\hat{F}(\omega; h)$ as per Eq. (12)
4. Find inverse-FFT of $\hat{F}(\omega; h)$ to get $\hat{\boldsymbol{f}}(\boldsymbol{x}; \boldsymbol{h})$

End of FFT_DENSITY

6 The Proposed Density Estimation Method

We would be using a cost function $C(h)$ to compute a weighted error metric as a function of bandwidth h over the data x. Where the sample data with n data points have been mapped to m grids points by binning to get $x_0, \dots, x_{m-1}$ as described in section 5. We first obtain the optimal bandwidth $h_{optimal}$ by minimising the cost function and then using that to compute the optimal kernel density estimate.

$$h_{optimal} = \text{argmin } C(h) \tag{16}$$

$$\hat{f}(x; h)_{optimal} = \text{FFT\_DENSITY}\big(x; h_{optimal}\big) \tag{17}$$

The rationale to choose a weighted error metric has been provided in section 6.1. While it's formula along with the entire cost function is present in section 6.2. Then section 6.3 explains some of the convex properties of weight function to assist in low bias and low variance convergence.

6.1 Choice of the error metric

The most common choice of error metric is Root Mean Square Error (RMSE) because of its nice mathematical properties. But in RMSE since the errors are squared before averaging, so it gives more weight to large deviations such as outliers. As described in section 5 for DFT computation, the grid needs to be extended on either side over the existing range as described in Eq. (15). So, these extensions will lead to high errors, as the naive density estimator will have 0 density for the extended points and the FFT will have some non-zero density value for smooth curve fitting. Hence Mean Absolute Error (MAE) is a better choice for not overweighting the outliers. As we have to also balance out the bias and variance of the density estimation. So, we will use a Weighted Mean Absolute Error (WMAE) metric, where the weights itself will be a function of bandwidth h and x.

6.2 Cost function

The cost function over m grid points is defined as:

$$C(h) = \sum_{i=1}^{m} W(h; x_i) \left| \hat{f}(x; h) - \hat{f}_H(x; h) \right| \tag{18}$$

Where $\hat{f}(x; h)$ is a vector of size m computed via FFT_DENSITY$(x; h)$ routine mentioned in section 5. And $\hat{f}_H(x; h)$ is also a vector of size m obtained from the naive density estimator $\hat{f}_H(x; x_0, h)$ as per Eq. (1) with x_0 substituted from Eq. (15).

Since the density $\hat{f}(x; h)$ is computed on the grid points via FFT, so the naive density estimate $\hat{f}_H(x; x_0, h)$ acts as a reference ground truth evaluated at the same grid points. We can now compute an error metric between the two and use that in a cost function.

The weight function $W(h; x_i)$ used is defined as:

$$W(h; x_i) = (0.1h^2 - 0.2log(h) + 0.95)e^{0.216\, b(x_i)} \tag{19}$$

$$b(x) = \min(abs(\arg\left(\hat{f}_H(x_0, \dots, x_{m-1}; h) > 0\right) - x\ ))/\Delta x \tag{20}$$

where Δx is the bin width between 2 grid points and the arg operator over $(\hat{f}_H(x_0, \dots, x_{m-1}; h) > 0)$ returns the vector x of only those grid points where the naive density estimator has non-zero density. Then $b(x_i)$ is the minimum absolute difference of current value x_i from this vector, normalized by bin width Δx.

We use traditional numerical optimization techniques to optimize scalar cost functions. We set bounds on the optimization routine with a lower limit as 0 and the upper limit as the rounded value of $\hat{h}_{RT}$ (as given in Eq. (8)). As $\hat{h}_{RT}$ already is an over smooth estimate so the optimal value should be either below or around it. So, setting its rounded value as the upper limit reduces the search space.

6.3 Properties of the weight function

The general form of the weight function can be written as

$$W(h; x_i) = (ah^2 - blog(h) + c)e^{d\, b(x_i)} \tag{21}$$

Let us take the case where the naive density estimator is non-zero for all grid points, then $b(x_i)$ will be 0 for all values of x over the grid. Then the expression becomes

$$W(h; .) = ah^2 - blog(h) + c \tag{22}$$

This function has a nice convex shape as depicted in Fig. 1(a). For extremely small values of h this function is dominated by $-log(h)$. As we know bandwidth h can take extremely small values for reducing the bias, but in doing so it increases the variance. So, this weight function will penalize the extremely small values of h more severely than bigger values. Thus, helping in controlling the overall variance.

The derivative of Eq. (22) is $2ah - \frac{b}{h}$ this implies the minima is at $\sqrt{\frac{b}{2a}}$. Thus, if b is set to 2a then minima will always be at 1. Which would be a desirable property as bandwidths above 1 are generally over smooth. So, both very high and very low values of bandwidth will be penalized in proportion with the error to avoid high bias and high variance.

The second term in Eq. (21) $e^{d\,b(x_i)}$ is for exponential penalization of error if no evidence has been seen from the bins of naive density estimator. So, if $\hat{f}(x;h)$ is predicting a density via FFT for a grid point where the naive density estimator has 0 density value. Then this term will rise exponentially, where exponent is proportional to number of bins having 0 values in $\hat{f}_H(x;h)$ until we see a non-zero density grid point. Thus, this component will penalize the high bias scenarios where the smoothing effect of the density estimator is underfitting the underlying density.

The combined effect of both terms is shown in Fig. 1(b).

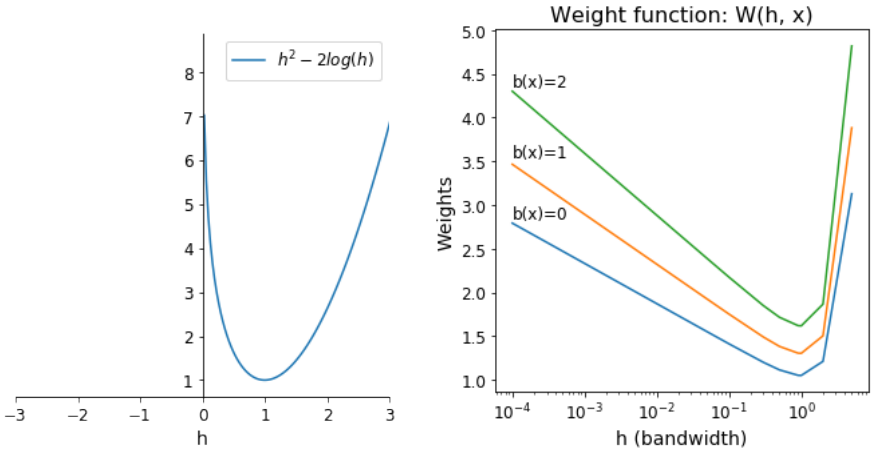


Fig. 1. (a) Eq. (22) with a=1, b=2, c=0; (b) W(h,x): Eq.(19), with varying b(x)

The values of constants in Eq.(19) determine the speed and results of the numerical optimisation routine. The values provided should be good for most cases. They can be quickly re-obtained via grid search if needed.

7 Experiment and results

7.1 Dataset and experiment strategy

To validate the cost function based optimal bandwidth selection strategy, we used the mixture of normal densities provided by Marron and Wand [9]. They have provided 16 densities that are typical representatives of densities likely to be encountered in real-world scenarios. These include PDFs some of which are smooth, some with sharp impulsive peaks, some multi-modal, some mixed multi-model with smooth and sharp peaks. For the test N random samples were drawn from each of the 16 distributions, with N varied from $[2^7, 2^{12}]$. To further imitate real-world scenarios, we added Gaussian noise $\mathcal{N}(0, \sigma^2)$ to the random samples taken from the 16 distributions. For testing purposes σ was varied from $[0.25, 1.5]$. Then we observe the combined effect of increasing both N and σ.

As we stated in the summary of section 4 that out of currently accepted popular methods for computing the bandwidth, the DPI method by Sheather and Jones [2] stands out and performs better in the majority of the cases than RT, LSCV, BCV. So, we will compare the proposed cost-based bandwidth selection strategy with Sheather and Jones's algorithm referred to as SJ henceforth.

7.2 Performance evaluation metric

In the given test setup, since we know the true distribution function of the 16 mixture densities, from which the data was first randomly sampled and then gaussian noise was added to it. So, the objective of the evaluation is to see even after randomness and noise how good are the estimated density as compared to the true density. To measure this for a given sample of data we would compute 3 densities: {True, SJ, Proposed}. We will the compute error between estimated SJ-PDF with True-PDF, then similarly we will compute the error between Proposed-PDF with True-PDF to see which one of {SJ, Proposed} is closer to the True distribution.

As different error metrics capture different aspects so we will use 3 error metrics to quantify the error:

KL-Divergence:
KL-Divergence is a measure of how one probability distribution is different from another reference probability distribution. In simplified terms, it is a measure of surprise. It is given by:

$$D_{KL}(P \parallel Q) = \sum_{x=1}^{n} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \tag{23}$$

Typically P represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution, while Q typically represents a model or approximation of P. The KL-Divergence is large whenever P estimates Q to have mass, but it does not have it. Lower the value of KL-Divergence, the more likely Q is closer to P.

We also measure the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) between the true distribution and the estimated distribution. MAE estimates the median, so it will give error quantification that is unaffected by outliers. While RMSE estimates the mean, so it will include the effect of all large deviations.

Gain:
To compare how the proposed algorithm is doing as compared to SJ. We will use the gain defined as:

$$KL_{gain} = 10log_{10}\left(\frac{D_{KL-SJ}}{D_{KL-Proposed}}\right) \tag{24}$$

A positive value of KL_{gain} indicates that the proposed algorithm is doing better than SJ and negative indicates SJ is doing better.

7.3 Results

Below are the results of the proposed algorithm's output over 3 error metrics: {KL-Divergence, MAE, RMSE} along with KL-Gain over SJ's algorithm for N=256. The numbers below are mean values of 30 trials. The first 4 columns show these 4 metrics for the case when no noise was added and the next 4 columns indicate the same metrics when Gaussian noise with $\sigma = 1$ was added.

Table 1. Error estimation of the proposed algorithm with True-PDF and KL-Gain over SJ

Distri-butions	No Noise KL DIV	No Noise MAE	No Noise RMSE	No Noise KL-Gain	Noise (σ=1) KL DIV	Noise (σ=1) MAE	Noise (σ=1) RMSE	Noise (σ=1) KL-Gain
1	1.029	0.011	0.001	-1.084	4.353	0.027	0.002	1.654
2	1.837	0.014	0.001	-1.178	7.914	0.04	0.003	1.57
3	6.587	0.031	0.003	-0.759	16.268	0.046	0.006	1.612
4	4.715	0.035	0.005	0.381	12.54	0.046	0.007	1.501
5	19.329	0.042	0.005	1.762	66.915	0.133	0.019	1.481
6	1.019	0.013	0.001	0.314	4.45	0.024	0.002	1.654
7	1.119	0.014	0.001	0.998	11.879	0.045	0.003	1.562
8	1.259	0.013	0.001	-0.406	4.627	0.024	0.002	1.583
9	0.898	0.013	0.001	1.235	4.863	0.025	0.002	1.626
10	4.643	0.035	0.003	1.139	8.615	0.042	0.004	1.503
11	1.18	0.014	0.001	0.116	4.416	0.024	0.002	1.692
12	3.263	0.023	0.002	1.566	6.313	0.03	0.002	1.577
13	1.157	0.015	0.001	0.873	4.76	0.026	0.002	1.598
14	5.922	0.03	0.002	1.322	14.743	0.05	0.004	1.291
15	5.97	0.033	0.002	1.137	16.71	0.054	0.004	1.038
16	2.394	0.024	0.003	-0.169	40.316	0.082	0.008	1.504

The graphs in Fig. 2(a) and 2(b) are showcasing the KL-Divergence of SJ with true-distribution compared with KL-Divergence of proposed algorithm with true-distribution. The lower the KL-Divergence the closer is the density estimate to true-distribution. As we can see for No-Noise case in Fig. 2(a) the KL-Divergence of proposed algorithm is majorly below SJ's except for one distribution and marginally above in a couple. But for Gaussian-Noise the proposed algorithm is significantly below SJ for all distributions. If we also look at Table 1, then we can see the net KL-Gain for No-Noise case is 7.3 and 24.5 for the Gaussian-Noise case. Thus the proposed algorithm is closer to the True distribution than SJ in both cases and does much better than SJ when noise is present.

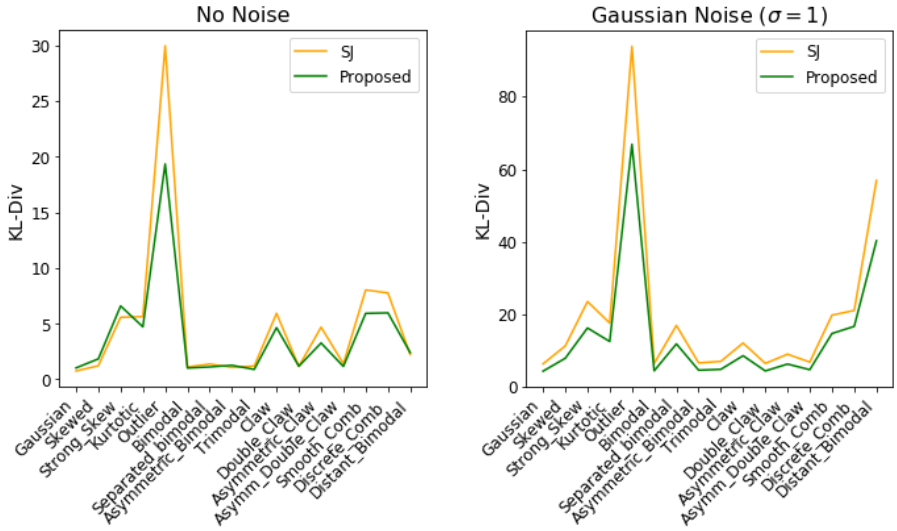


Fig. 2. (a) SJ vs Proposed with No-Noise (b) SJ vs Proposed with Noise

Below are the results of varying N and σ, with effects measured in terms of gain over SJ for all 3 metrics: {KL-gain, MAE-gain, RMSE-gain} as per Eq. (24)

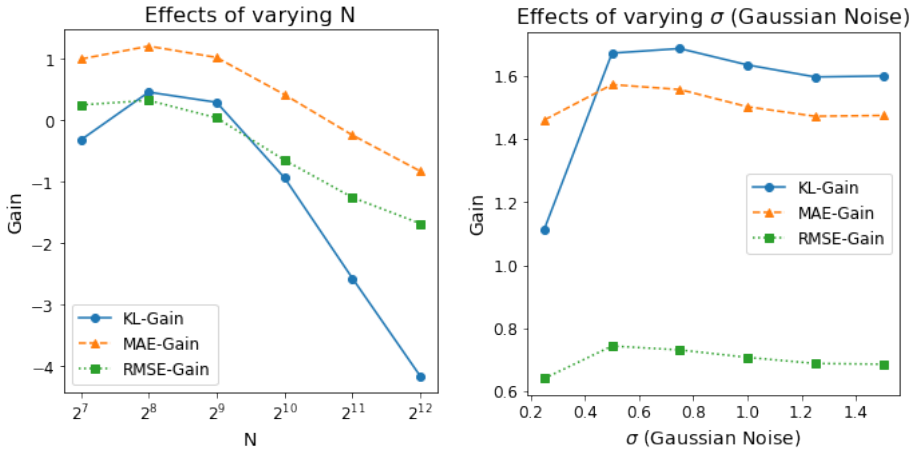


Fig. 3. (a) Gain vs N (b) Gain vs Noise

As we can see in Fig. 3(a) when there is no noise in data and we are varying N. Then for smaller values the proposed algorithm does better, but as data size increases SJ starts doing better. Moreover, when noise is present as in Fig. 3(b) then the proposed algorithm is always better than SJ.

Let's now look at the combined effect of varying N and σ.

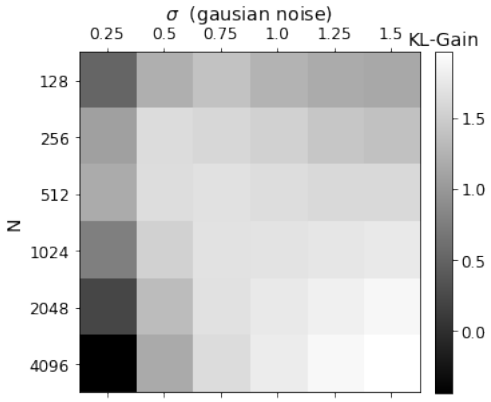


Fig. 4. Effect of varying both N and σ

As we can see in Fig. 4 when noise is low at 0.25 and we increase N then SJ starts doing better, but when we increase both N and σ then the gain increases drastically and the proposed algorithm does much better. So for large noisy data, the proposed algorithm will work better.

If we compare these results with Suhre, Arikan and Cetin [7], then for N=256, with 15 distributions (excluding Gaussian). Their kl-gain was positive for 5/15 distributions with a median value of -0.69. While in the proposed algorithm the kl-gain for No-Noise case as per Table-1 is positive for 11/15 distributions, with a median value of 0.87. For the Gaussian-Noise case 15/15 distributions are positive with a median value of 1.57.

8 Conclusion

From the results, we can conclude that the proposed algorithm does better than the recent paper Suhre, Arikan and Cetin [7] with similar approach. Then for real-world noisy data the proposed algorithm performs significantly better than the widely accepted solution of SJ (Sheather and Jones) [2]. Especially when the noise gets higher, the proposed algorithm gets even better. Hence the proposed cost-based optimal selection technique can find the optimal bandwidth along with keeping both bias and variance in check. Moreover, the computation cost of the proposed algorithm is of the order of O(m log(m)) per cost function evaluation over m grid points, supported by a convex weight function that ensures quick convergence.

References

1. Silverman, Bernhard W. "Algorithm AS 176: Kernel density estimation using the fast Fourier transform." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 31, no. 1 (1982): 93-99.

2. Sheather, Simon J., and Michael C. Jones. "A reliable data-based bandwidth selection method for kernel density estimation." *Journal of the Royal Statistical Society: Series B (Methodological)* 53, no. 3 (1991): 683-690.
3. Jones, M. C., J. S. Marron, and S. J. Sheather. "A Brief Survey of Bandwidth Selection for Density Estimation." *Journal of the American Statistical Association* 91, no. 433 (1996): 401-07
4. Here Nils-Bastian Heidenreich & Anja Schindler & Stefan Sperlich, 2013. "Bandwidth selection for kernel density estimation: a review of fully automatic selectors," AStA Advances in Statistical Analysis, Springer; German Statistical Society, vol. 97(4), pages 403-433, October
5. Turlach, Berwin A. "Bandwidth selection in kernel density estimation: A review." In *CORE and Institut de Statistique*. 1993
6. Raykar, Vikas C., Ramani Duraiswami, and Linda H. Zhao. "Fast computation of kernel estimators." *Journal of Computational and Graphical Statistics* 19, no. 1 (2010): 205-220
7. Suhre, Alexander, Orhan Arikan, and Ahmed Enis Cetin. "Bandwidth selection for kernel density estimation using Fourier domain constraints." *IET signal processing* 10, no. 3 (2016): 280-283.
8. Edurado Garcia Portugues, Book: Predictive Modelling, https://bookdown.org/egarpor/PM-UC3M/npreg-npdens.htm, section 6.1.3
9. Marron, J. S., and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics*, 20 (2), 712–736. [208,213-217]
10. Jones M.C. and Lotwick H.W. A Remark on Algorithm AS176 Kernel Density Estimation using the Fast Fourier transform, Applied Statistics 33, 120-122

Methods for Solving the Challenges Observed In The Multiplatform Setup for Self-Driving Cars

Mirza Mujtaba Baig

University of the Cumberlands, Parsippany, NJ, USA

mujtaba961989@gmail.com

Abstract. Artificial Intelligence (AI) is evolving rapidly and one of the areas which this field has influenced is automation. The automobile, healthcare, education and robotic industries deploy AI technologies constantly and the automation of tasks is beneficial to allow time for knowledge-based tasks and also introducing convenience to everyday human endeavors. The paper reviews the challenges faced with the current implementations of autonomous self-driving cars by exploring the machine learning, robotics and artificial intelligence techniques employed for the development of this innovation. The controversy surrounding the development and deployment of autonomous machines e.g., vehicles begs the need for the exploration of the configuration of the programming modules. This paper seeks to add to the body of knowledge of research assisting researchers in decreasing the inconsistencies in current programming modules. Blockchain is a technology of which applications are mostly found within the domains of financial, pharmaceutical, manufacturing, artificial intelligence. The registering of events in a secured manner as well as applying external algorithms required for the data analytics are especially helpful for integrating, adapting, maintaining, and extending to new domains especially predictive analytics applications.

Keywords: Artificial Intelligence; Automation; Big Data; Self-driving Cars; Machine Learning; Neural Networking Algorithm; Blockchain: Business Intelligence

1 Introduction

The Cloud Computing and Data Science platform, robotics environment and AI are developing rapidly, and it brings its advantages to everyday activities. However, many uncovered gaps are found in the industry [7].

The current challenges faced by the current implementations of the self-driving cars are related to safety, security, privacy, performance, user's experience, reliability, and economic value and a one-to-many relationship between the attribute and its properties. Also, the challenges in the implementations of controllers, deficiency, malfunctions, and inconsistencies that occur during data transmission and the threats detected while performing the implementations for the self-driving vehicles have been discussed.

This paper addresses the lack of sufficiency in data modeling activities for the self-driving cars implemented within the autopilot modes, also discussed are the lazy evaluation mechanisms included for the design of multimodal dialogue systems by combining input modules to generate next-generation autonomous machines as part of Artificial Intelligence and IoT technologies.

The paper discusses about the creation of patterns related to gesture recognition, speech recognition, dialogue system, and conversational knowledge base for enhancing the existing functionality of the autonomous system. These technical advancements can be an enhancement to the field of robotics, healthcare, education system, security access control systems, advertising, driving, aviation, and personal assistance image recognition. The activities associated with machine learning, complex planning, decision-making methods, verification, and guaranteed performance of autonomous driving pipeline have become ever so involved with the presence of the cars in the challenging environments.

The approaches considered for implementing the module of the cars includes sequential planning, behavior-aware planning, end-to-end planning, advanced perception, motion planning, control. Self-driving vehicles which are in the prototypical stage are being put into public traffic and soon will be available to put on display and start being sold. Public awareness and media coverage are already in the stages on unfolding the discussions about self-driving cars from the perspectives of Waymo developed by Google which are developed based on computer simulations and feeds what it learns from those into a smaller real-world fleet, autopilot self-driving vehicle developed by Tesla which are incorporating huge amounts of data [5]. The reports have been published about the situations of accidents caused by autonomous vehicles. The programming of the software embedded within the Electronic Control Units (ECUs) is playing a key role in the modern vehicles as well as the self-driving vehicles. The discussion has already been in place to describe about the software engineering activities as well as the production and manufacturing activities encompassed within the electrical and software architecture of the car.

The challenge of the current automation construction for self-driving cars lies within handling the integration between technical areas and integrated approach required for the autonomous driving related to decision-making and real-time control of the driving situation. The components of the models can include creativity, emotional knowledge, self-awareness along with the incorporation of the neural networks, fuzzy systems, evolutionary computation, and computational models. The current model with the help of AI is able to produce intelligent decisions in variable, real-time traffic conditions. Also, the simulation-based learning situations along with neural networks and artificial intelligence systems are able to address properly to the emergency situations in real life events. Combination of robotics, automation, AI, machine learning, data science and Big Data benefits to the automobile industry and transportation safety. The advances made in the sensor capabilities of the self-driving cars system was able to improve the event-based vision, which was involved with the broad-set of visual pattern recognition. In contrast to traditional machines, smart machinery is being employed by the system to handle the behavior, operations, circumstances, and feedback of the autonomous machines.

Taking into consideration, the devices handling wireless sensors as well as universal connectivity enables the flow of data to be monitored while being sent to the autonomous machines; onboarding processors can handle the operations on the data as well as provide appropriate responses and as such it is required for new software architecture to be developed. The capacities of this model can include creativity, emotional knowledge, self-awareness incorporating the technologies from computational intelligence related to neural networks, fuzzy systems, evolutionary computation, computational models. This level of intelligence can be extended to handle tasks with the proper level of rationality to supersede humans thinking capability. Autonomous vehicles have been provided with proper sensors, cameras for enabling recognition of 3D environment as well as providing the capability to make intelligent decisions in variable, real-traffic conditions. Artificial Intelligence platform when combined with virtualized environments have the ability to respond to the situations in real-time taking into consideration voice recognition, facial recognition, text analytics and natural language processing, virtual assistants, robotic process automation, biometric recognition, hardware optimization. By combining robotics and automation, it has been possible to merge traditional programming and learning modules with developing applications that are more robust, efficient, and provide safe automation. Mobile robotics include the platforms for localization, mapping, planning, image recognition, reinforcement learning modules, focus on prior knowledge about the structures of tasks and environments. Within the framework of autonomous computing, deep RL-methods for end-to-end computing has been enabled to provide optimal adversarial policy (i.e., mapping of states to actions) for autonomous motion planning. By employing optimal motion-planning strategy it is possible to avoid collisions within the trajectory surface for the self-driving vehicles.

With a help of suitable choices of optimality criteria as well as employing a reward function, it will be capable to enable the adoption of powerful state-space exploration and policy optimization techniques developed using deep Reinforcement Learning applications. Furthermore, we argue that both the training and test-time procedures of adversarial policies provide quantitative measures of reliability, which can be used for benchmarking of behaviors in worst-case scenarios. In this proposal, the design is based on the technical combinations of machine learning techniques including: deep learning, neural networks, sensor networks, computer vision, natural language processing, and, additional forms of processing the data including databases, networking, security, memory, and infrastructure capabilities. The produced outputs are presented in the form of graphical displays, navigation systems, motions, and control devices. The different test case scenarios will be conducted by setting the appropriate measures for the test performance.

The proposed model shows the architecture flow of Big Data, Data Science, Robotics, Blockchain, AI technologies along with programming modules. In the long run, the benefit will be provided to the transportation system by decreasing the number of incidents and improving the driving conditions. Using the technical combinations in cloud computing, data science and analytics, robotics and Artificial Intelligence platforms, the paper aims to uncover the challenges faced by the implementations of these algorithms for the model in the real-time environment. As a matter of fact, it also focuses

to provide solutions to some of these challenges. The information stored can be combined using knowledge processors and information stored on the blockchain networks. The approach uses blockchain as a decentralized ledger for storing information about robots and coupled sensors, using cameras and smart-contracts for defining the logic of control, complexity, computing power. The majority of implementations require either Proof-of-Work (PoW) or Proof-of-Stake (PoS) as consensus algorithms. Business Intelligence mostly focusses on the areas of making decisions as well as controlling company values on a national as well as international scale. The main properties include decision tree classification, measuring relative performance concerning the different KPIs ingested on the platform. Also, the data attributes can consist of structured, semi-structured, and unstructured attributes onto the platform. By enabling AWS EC2 instances and having mounted the docker containers on the platform and the data ingestion using S3 data pipelines, the data flow management can be made possible. It is possible to create repositories of data ingested within the Big Data platform. By utilizing these scenario's with machine learning algorithms, including decision trees, classification, clustering, neural network modeling can lead to generative prediction models. Also, increase the return of investment for the organization's. In addition, the python libraries in the form of Keras, Tensorflow, Pandas can be implemented within the model.

The paper is organized in 4 sections. Section 2 is related to the literature review and the previous works related to the industry. The Section 3 describes the methods and technologies used in developing the model, and the results and benefits from the produced module. The Section 4 provides the conclusion on the innovation of the current method and advantages for the automobile industry.

2 Literature Review

IoT platform collects diverse real-time valuable information of objects' concerns in objective world, forms a giant network through Internet and realizes interconnection among massive sensing devices in order to make co-fusion between data world and physical world. It has been highlighted in the context of autonomous vehicles that there are emerging trends and challenges related to performing operations in the self-driving vehicles. Trust requirements at the early stage for developing autonomous cars taking into consideration the factors of safety, security, privacy, performance, user's experience, reliability and economic value and a one-to-many relationship between the attribute and its properties is established. A safety-critical control function, such as steering, brake assist, self-parking and other functions without a direct driver input. A framework for integrating mutual trust computation with a standard human - robot interaction. The system consists of both functional and non-functional requirements. Functional requirements can be related to calculation, technical details or other specific functionality that define what a system is supposed to accomplish. Non-functional requirements for autonomous car describes properties, such as the look and feel of safety, security and privacy, which are critical to the product's success based on the user's expectations and demand.

Trust requirements can be related to i) Customer's satisfaction with the technology provided, ii) Willingness of learning and using the automated features in the current autonomous technology in a car, and iii) Preferable method for learning to use the automated features. Security relates to an attribute that protects the digital information and data from any danger or threat from any malicious activity. Examples of the properties that relate to this attribute are data sharing, global positioning system (GPS) and visible vehicle identification number, smart key, remote keyless entry system, and remote panic features. Safety features can be related to automatic steering, intelligent pairing assist system, blind spot system, adaptive light control, and camera sensor technology. Privacy is related to the location of information, cloud storage, crash data retrieval, event data retrieval and cabin monitoring system. Reliability attribute is measured based on the rating system for head protection technology, blind spot technology, advanced safety assist technology and national highway traffic safety administration. Performance attributes are related to engine smart cities, weather-sensitive tires, tires sensor, active suspension control system and electronic transmission control. User's experience are the push start button, screen touch control, keyless entry, garage door system, design dashboard and on-board maps and navigation. Economic value are trusted brand, selling price, trend, product features and car warranty. The measures of error of distance and error of angle from camera images, and then applying fuzzy logic to fuzzify them into a combined error degree. Autonomous driving application has features composed of low speed, short connection, fixed routes, less complicated traffic. The path tracking controller can be divided into two categories: localization-based path tracking and vision-based path tracking. Localization-based path tracking can consist of high-precision GPS or LiDAR could provide absolute global position. And one of the path tracking controllers is pure pursuit for the robotic modeling using vehicle kinematic bicycle method.

Control theory controllers are also based on precise localization, using the approaches of Linear Quadratic Regulator (LQR) and Model Predictive Control (MPC). Vision-based path tracking task is defined as guiding the motion of a robot with respect to a target path based on the feedback obtained through a vision system. The following measures can be used to calculate the distance error and direction error at preview distance in the image, however they require high-precision error values. Environment factors such as daylight and whether severely affect visual detecting, path tracking. Visual Path Tracking consists of the following paths related to Inverse Perspective Mapping, Track Line Detection, Fuzzification. A range of AI techniques can reduce investigators' cognitive load and support decision-making, including: planning the assessment of the scene; ongoing evaluation and updating of risks; control of autonomous vehicles for collecting images and sensor data; reviewing images/videos for items of interest; identification of anomalies; and retrieval of relevant documentation. With the utilization of robots, associated with Micro Unmanned Aerial Vehicles (MUAV) for carrying out remote sensing in the current hazardous environments. Threats can be possibly detected using responders as well as multi-robot reconnaissance. The RAVs can be able to operate as multi-agent robot swarm for dividing the work as well as relaying information from the sensors as well as cameras to a Central Hub. The Image Analysis model can

be used by utilizing a Deep Neural Network algorithm for detecting as well as identifying the objects in images taken by the RAV cameras. It is also possible to perform pixel-level semantic annotation of the terrain by using the network algorithm to support subsequent route-planning for Robotic Ground-Based Vehicles (RGVs).

The Probabilistic Reasoning module assesses the likelihood of occurrence of different threats, for incoming information as well as monitoring the images and sensor readings. By deploying Information Retrieval TF-IDF module is possible to simulate with respect to the incident. The communication protocol can be developed using JSON-based real-world RAVs using RESTful APIs, cameras, sensor systems. In this way, the system can be maintained loosely coupled as well as testing can be taken into consideration the real-world scenarios. The routing algorithms employed can include hyper-heuristics which can be integrated with machine learning algorithms for optimizing the routes as well as handling the components of failure and battery usage [3]. The documentation can be based on standard operating procedures and guidance documents from the knowledge base based on the implementations done using Elastic Search implementations. There are major contributions that can be seen in the form of machine learning techniques including: deep learning, reinforcement learning, Markov decision processes, robotics, psychology etc. [2]. Starting with the machine learning techniques, the necessary factors were taken into consideration to cleanse the data that can be handled across different Big Data sources and subsequently generate intelligence results and neural network forms related to enhancing perpetual intelligence and eliminate the necessary feature engineering options. Advances were made for improving computer vision, which was involved with the broad set of visual pattern recognition. According to a recent research conducted by the team of Modular and Integrated approach, an algorithm was developed using the technologies which include Sensing, perception and Decision making where, in addition, the operating system and cloud platform deliberately are designed in accordance with HD mapping of the sensor data. Sensors like LiDAR, GPS, IMU, radars and GNSS helped in determining the reflection time based on the distance and enabling in creating High Definition (HD) maps, real-time localization, as well as map projection and process update, respectively. Action prediction and path planning mechanisms are integrated with the autonomous systems to generate an effective action plan in real time.

The driving conditions are challenging in the complex traffic environments. Irrespective of an action plan based on predictions the stochastic model of the reachable position sets of the other traffic participants and associate within the reachable sets. Searching all possible paths is the best approach within the given dataset and, a cost function can be utilized to identify the best approach. However, this requires enormous computational resources which may make it incapable of delivering real-time navigation plans. To ensure that the processing pipeline is effective, sensor data generates faster results, and, if a part of the system fails, it needs to be robust enough to recover from the failure and the system needs to perform all the computations under energy and resource constraints. Reinforcement learning defined as having the capability to support artificial intelligence by enabling dynamic programming to train the model using the forms of pass or failure for the situations encountered while performing technical evaluations for the model and provide mechanisms for developing mobile robots. Deep

Learning, a form of artificial intelligence function is capable of processing the data and creating patterns for use in decision making, and, capable of learning unsupervised forms of data from data that is unstructured in nature. Both the technologies have contributed significantly to the autonomous systems including pedestrian, car, cyclist, traffic sign detection, semantic segmentation, and other tasks. While these systems heavily rely on learning: localization, reasoning, and planning modules that often continue to be the domain of secured rules and programs, designing and developing geometric priors and intuitions. This design usually operates with expert knowledge and repeated iterations between testing – in simulation as well as on the platform for handling and refining heuristics data as well as to reduce the manual effort and focus on the automation tasks relevant to learning decision patterns.

The successful application usually requires detailed domain knowledge, systems engineering, and demands significant time for data collection and curation, experimental setup and safety arrangements. One of the AI applications, namely, Advanced Driver-Assistance Systems have been developed and consequently its full potential can be realized by the implementing self-driving vehicles. The paper describes the state of the art technologies used to implement machine learning including deep reinforcement learning (RL) for optimal regulation and tracking of single and multiagent systems [6]. The proposal draws on the effectiveness of uncertainty-based information-theoretic approach for performing optimal searches within the data. The algorithm related to iteration-based Q-learning is guaranteed to excel in the implementations of critical-only structures which requires the neural networks to be combined with the Q-function for analytical purposes. The theory is proved using Lyapunov techniques for the algorithmic implementations. Subsequently, the RADP algorithms are proposed to transform the nonlinear optimal control problems with closed loop systems to be stable. There are three groups based on extended Kalman filters, particle filters, and graph optimization paradigms for the algorithmic implementations. The global features are able to extract information in the form of raw pixels, shape signatures, color information. Sequences of image can be matched based on deep convolutional neural networks modeling. On the other hand, local features utilize a detector to locate points of interest (e.g., corners) in an image and a descriptor to capture local information of a patch centered at each interest point.

The Bag-of-Words (BoW) model is often used as a quantization technique for local features to construct a feature vector in place recognition applications. The image-to-image matching, which localize the most similar individual image that best matches the current frame obtained by the robot. Combining pair wise similarity scoring, nearest neighbor search, and sparse optimization provides reliable solutions. This has been demonstrated by being able to integrate information from a sequence of frames that can significantly improve place recognition accuracy and decrease the effect of perceptual aliasing. This method uses rasterized vehicle context (including the high-definition map and other actors) as a model input to predict actor's movement in a dynamic environment. As the vehicle is approaching the intersection the multimodal model (where we set the number of modes to 2) estimates that going straight has slightly less probability than a right turn. Actors motion prediction methodologies consist of computing object's future motion based on the time progressed by using Kalman filter as well as utilizing

short-term predictions using the map method (Perez, Deligianni, Ravi, Yang, 2018). The factors considered include possible paths computation, lane connectivity, vehicle's current state estimate. Machine learning models including Hidden Markov Model, Bayesian networks, or Gaussian Processes, Inverse Reinforcement Learning approach are utilized mostly in the autonomous systems design.

One line of research follows the recent success of recurrent neural networks (RNN), namely Long Short-Term Memory (LSTM), for sequence prediction tasks. Authors applied LSTM to predict pedestrians' future trajectories with social interactions and long-term dependency problem. The LSTM network has the ability to remove or add information for the state in the form of gates which have the ability to let the information pass through. This layer consists of sigmoid neural net layer and point wise multiplication operation. At the same time, the gate layers are able to look at the cell state. To find the location of the device, the past trajectory data can be employed. In another recurrent network variant called gated recurrent unit (GRU) combined with conditional variation auto-encoder (CVAE) was used to predict vehicle trajectory. Convolutional neural networks also encompassed visual glimpses. Mixture Density Networks (MDNs) are conventional neural networks which solve multimodal regression tasks by learning parameters of a Gaussian mixture model. Contrast to the neuro-fuzzy is proposed to simulate the propagation model to predict RSS and compare the performance with empirical models of channels. RSS from the aerial platform is calculated using Hata model. However, Kaiser integrated the combination of received signal strength (RSS) and variance fractal dimensions as an input to ANN for prediction of location features. ANN was used to predict channel propagation of multi-user transmission under Rayleigh fading to maximize the efficiencies [1]. Conditional Variational Atuo-Encoder was considered to be nonparameterized prediction method to be employed for a diverse set of prediction hypotheses to capture the multimodality of the space of plausible futures [8]. Emergent behavior includes induced lane changes and changes in velocity at intersections and highway segments. The behavioral models are learned by maximum-entropy IRL from demonstrations of different social acceptability [4]. The maximum-entropy deep IRL framework exploits the expressive capacity of deep fully convolutional neural networks to represent the cost model underlying driving behaviors [4]. In general, deep fully convolutional neural networks, as robust, flexible, high-capacity function approximates are able to model the complex relationship between sensory input and reward structure very well [4].

In fact, AI works on the principle of costly hardware, as well as massive datasets to process the data, machine learning phase of deep learning, reinforcement learning, inverse reinforcement learning principles plays a major part in the software development. The computations performed are very intensive in nature and cost-effective computing resources. Object recognition AI systems can include facial recognition on the phones as well as systems. A national approach is followed for maximizing the potential capabilities which are critical factors utilizing coordination and mobilizing key agencies to make the necessary organizational changes to take advantage of AI to the next level (Sharifzadeh, Chiotellis, Triebel, Cremers, 2016). Cognitive scientists explore mental ability of human beings through observation on aspects such as language, perception, memory, attention, reasoning and emotion. Brain-like computing aims to enable the

computers to understand and cognize the objective world from the perspective of human thinking. Communication field emphasizes on transmission of information, while computer realm emphasizes on utilization of information. The data can be in the form of structured and unstructured data. The automotive industry is facing the challenges of operational inefficiencies and security issues leading to cyber-attacks, unnecessary casualties, incidents, losses, costs and inflated prices for parts and services. The activities involved are not only limited to data analytics, but also extendable to applications of scientific principles and methodologies, identifying complex patterns in data and extraction, feedback loop algorithms for predicting of data, simulating the patterns of neurons for the perception activity, neural network is made of input units, hidden units and output units [9]. The risks can include malicious server sign the past transactions, denial of service attacks, maintenance tasks, software problems, eavesdropping, 51% attack affecting data integrity and data availability, lack of scalability, high energy consumption, low performance, interoperability risks or privacy issues, usability, cryptocurrencies volatility.

Business Intelligence application activities can be in the form of analyzing customer behavior for particular products bought, social network activity data, statistical analysis. Other criteria for implementing business intelligence solutions include determining sales for a particular region by considering holidays, geographical regions based on which the organization can prosper. For the platform, data migration activities can be carried out using the scripts implemented in JSON, python, java, scala languages. With the data extracts coming in the form of CSV, text files, salesforce data, Oracle, and Postgres database. Sentiment analysis and neural network implementations are useful for predicting the behavior of customers and sales. By utilizing Djikstra's algorithm, the shortest path can be determined between locations, products, warehouses, and, sales. The workflows can be designed using the algorithm of Support Vector Machine, Decision tree modeling of data for handling the activities of data pre-processing, data processing, and model performance attributes. The mentioned criteria can be related to warehouse product availability, promotions, customer data, social media sentiment analysis, catalog list of products, and items categories. Finally, with the capability of integrating machine learning algorithms from the architectural perspective of artificial intelligence and business intelligence, prediction models related to predicting future sales can be determined based on the criteria of performance score.

3 Discussion. Technical Implementation

The techniques of perception, data planning, decision-making, interactive planning and endto-end learning has been proved valuable in deriving the insights for the future generation of cars. Improvements in functional capabilities have been proven to be useful for the prototypes developed as well as there is a requirement still focused on the performance and safety of the vehicles being driven under different driving conditions. The concepts of autonomous vehicles, decision-making, motion planning, artificial intelligence, verification, fleet management have proved useful for the communication to be distributed across. With the development of network-enabled sensors and artificial

intelligence algorithms, various human-centered smart systems are proposed to provide services with higher quality, such as smart health care, affective interaction, and autonomous driving. Cognitive computing consists of the domains including data science, discovery, cognitive science, and big data. The technologies of networking, cloud computing, analytics can be implemented in this domain. The applications of this technology are found in human-centered cognitive computing, including robot technology, emotional communication system, and medical cognitive system. The human-centered cognitive cycle includes machine, cyberspace and human. Also, autonomous driving can improve the quality, performance, productivity, reliability, efficiency of the driving vehicles. The level of automation derived from human-operated vehicle to self-driving autonomic vehicles is varying in nature. The advancements made in the technical implementations of the self-driving cars have supported the drivers in making decisions for taking actions related to driving and maintaining speed, to be in a lane, performing car-driver handover. The experiments performed in this area were conducted based on real-time environment simulations. Driving in dynamic environments, needs to handle the outcomes to be dealt in an unpredictable manner for situations related to human-level reliability, and reacting safely in complex urban situations. The navigation system for the autonomous vehicles can be developed by taking into consideration the factors of perception, decision-making, control. Machine learning techniques and complex planning and decision-making methods, verification and guaranteed performance of autonomous driving pipeline have become ever so involved with the activities of challenging environments.

Robotic Programming includes mechanisms related to kinematic modeling consisting of scenarios related to reference path, proportional-integral-derivative control, feedback linearization, model-predictive control, nonlinear model, model predictive control, feedback-feed forward control (Khan, 2018). Operating procedures at high speeds or aggressive maneuvers employ full dynamic modeling of the vehicle which includes tier forces. Simulation modelling of the cars has been developed taking into consideration the factors of steering angle and projection trajectories. Advances made for fast nonlinear optimizers is to optimize simultaneously over steering angle and velocity or throttle input for achieving minimal intervention. Firstly, is the input space discretization with collision checking, secondly is randomized planning, third is the constrained optimization and receding horizon control and can also compute collision-free trajectories for employing a navigating device in the vehicle. The main advantage of this vehicle is in the constrained optimization for the smoothness of trajectories and direct encoding of the vehicle model utilized in the trajectory planning. The rules included can be based on the motion-planner, maximizing visibility may reduce risk, questions of rules violation arises, rules related to logic functions as well as utilizing automatic control synthesis methods can be demonstrated (Perez, Deligianni, Ravi, Yang, 2018). Also, by utilizing the cost function, traditional motion planning methods can be employed to determine the trajectory of the lowest cost. Minimum-violation routing and single-trip scenarios to effectively monitor the flow of data. Integrated perception and planning can be done to generate the control input of the vehicle based on sensory information as well as machine learning optimizations. Perception can be based on the following techniques recognition, reconstruction, motion estimation, tracking, scene

understanding, and end-to-end learning. Maps have been constructed as part of handling navigational features in the autonomous engines. Object detections are done by employing bounding-box detection, semantic segmentation, deep neural networks have played a significant role in this area, effective neural network have played a significant role in the dataset's in production. Bayesian deep learning incorporates the intersection between deep learning and Bayesian probability theory, providing uncertainty estimates within deep architectures.

Sensory information and actuation can be achieved in the network by enabling path proposals, perception and planning module, neural network schemas, supervised and unsupervised learning modules, GPU-computing capabilities computed using efficient learning of convolutional neural networks, improved performance of end-to-end driving capability. In comparison to research development work done at NVIDIA, the organization's project trained a deep convolutional neural network to map raw images from a front-facing camera directly to steering commands and were able to handle challenging scenarios such as driving on a gravel road, passing through roadwork, and driving during the night in poorly lit environments. During training, random shifts and rotations are applied to the original input image and virtual human interventions are simulated to artificially increase the number of training samples that require corrective control actions. By observing which regions of the input image contributed most to the output of the network. A large-scale driving video data set to train an end-to-end fully convolutional long short-term memory network to predict both multimodal discrete behaviors on a task-based level and continuous driving behaviors. The architecture for time-series prediction essentially fuses a long short-term memory temporal encoder with a fully convolutional visual encoder. Once the information quantity of various experience become large, he or she may possess human's big data thinking, which is hierarchical as deep learning. One differences between big data analysis and cognitive computing is data size. There are various degrees of processes such as cognition, memory, learning, thinking, and problem solving.

Cloud computing virtualizes the computing, storage, and band width. Information related to the literal information and the pictorial information correspond to natural language processing and machine vision. Cloud computing and IoT provide cognitive computing with software and hardware basis, while big data analysis provides methods and thinking for discovering and recognizing new opportunity and new value in data. Traditional supervised learning and unsupervised learning are based on closed training with data input. The end-to-end motion planning has been also applied to robotics—for example, to learn a navigation policy in simulation from an expert operator, with a 2-D laser range finder and relative goal position as inputs. It is then feasible to transfer the knowledge gained from training to unseen real-world environments to perform target-oriented navigation and collision avoidance. Socially aware collision avoidance with deep reinforcement learning was introduced to explain and induce socially aware behaviors capable of learning directly from multiagent scenarios by developing a symmetrical neural network structure. Robots that use learned perceptual models in the real world must be able to safely handle cases where they are forced to make decisions in scenarios that are unlike any of their training examples. Automated driving with hu-

manlike driving behavior requires interactive and cooperative decision-making, socially compliant motion planning. Probabilistic approaches can be defined to include lower cost for the functions involving maximum likelihood, or, maximum a-posteriori resulting in a receding-horizon planner. Increased complexity in this model is another challenge. Decision tree grows exponentially with the number of agents such as Monte Carlo tree search algorithm. Markov decision process can be applied for scenarios related to making decisions in the highway platform. Factors related to maximum acceleration, desired acceleration, desired velocity, minimum distance, and desired time gap can be considered. The task of the probabilistic graphical model is to generate an intention estimation with maximum probability, given observed information.

Individual Gaussian processes are coupled through an interaction potential that model cooperation between different agents' trajectories. Terms for affordance, for progress, and to penalize close distances to other agents can also be included in their joint cost function. The interaction model simply consists of a constant braking action triggered if a time to collision falls below a threshold. Planning on abstractions rather than detailed trajectories can lower planning complexity significantly. Solving the POMDP in a conventional way or by domain knowledge and specific simplifications, is to employ nonparametric reinforcement learning, to immediately receive an approximately optimal policy without optimization. However, generalization to arbitrary environments remains a challenge. Support vector machine for lane-change decisionmaking with features composed of relative position and relative velocity. If a lane-change desire is triggered, a lane-change reference trajectory is executed by a model predictive controller with the objective from minimal deviation to the reference subject to a set of safety constraints. Data being handled by low-cost sensors. Data will be able to flow from information blind spots to augment and improve decision making. With focus on sensor networks, especially wireless sensor networks in which an object having a sensor on it comprises of heterogeneous computing environment.

Within the ecosystem, the system is integrated with hardware, software, connectivity, and information flows linked with decision making capability. Objects in a SOA architecture can consist of Internet Protocol address and sends data about its state and immediate environment and also can receive data linked to actions. The individual tasks become end-to-end process consisting of a workflow managed in real-time data environment. Trustable data is the primary advantage of Blockchain technology, can be used in applications that can seek to improve the situations where low levels of trust exist. The strategic advantage of IOT strategy is that to handle emerging changes and integrate IT innovations such as Big Data and Artificial Intelligence based on which new skills will emerge for making jobs to be digitally displaceable as well as new jobs arising. We also need to see new tasks being undertaken such as monitoring the state and health of embedded sensors within the system on a construction site and beyond. Internet Transport Protocols including TCP/IP, UDP, OSI model of packaging the messages as well as sending and receiving the packets of data. The data from user-requests on end-user devices could be brought in by encapsulating data from sensors which are involved in sending more information-rich pages back to the client. The other infrastructure related components are related to cloud and virtualization. SOA architecture is being evolving into DevOps, Virtualization, Serverless Architecture, Operational

Technology, Information Technology convergence. Process control systems consisting of equipment, process flows, sensors and actuators. IoT solution is consisting of network of computing resources example processors, volatile and persistent memory/storage, networking software, applications, analytics algorithms and more. Embedded Sensors, Smart Sensors, embedded and autonomous computing Machine Learning model similar to Linear regression model can be used for predicting the temperature of radiation waves present with the components in the platform that can include integration of hardware, operating system, virtualization, IoT enablement related to MQTT, DDS, Kafka, Cassandra, Tensorflow tools. Data running through an API as well as data virtualization layer should be monitored for security analysis. Machine Learning and Neural Networks techniques related to Convolutional Neural Network, Recursive Neural Network Security and Privacy Preserving algorithms including man-in-the-middle attack, Public-Private keys, RSA algorithm, Format Preserving Encryption, Blockchain are all the security methods through which the data in the pipeline can be preserved. W

With the advancement in technology by utilizing sensors, big data techniques, improvements in connectivity and computational power, emerging new machine learning approaches, developing new computer paradigms, human-machine interface, IIOT enhancements, using robotics and 3D/4D printing systems, these capabilities have enabled a wide range of features and services with threats of malicious attacks or risks within the cybersecurity infrastructure. In this aspect, automotive industry can include blockchain technology for providing additional capabilities of decentralized platforms containing information about insurance, proof of ownership, patents, repairs, maintenance, tangible/intangible assets that can be securely recorded, tracked, managed. The ability to accessing real-time data provides a range of opportunities and business models capable of handling automation of processes through Internet of Things and smart contracts, advancements in predictive maintenance and forensics, smart charging services for electric vehicles, peer-to-peer lending, leasing, financing, collaborative mobility. Blockchain can provide multiple security benefits, as well as cyber-resilient applications including decentralization, cryptographic security, transparency, immutability. Within the blockchain, the chains might be linked to various different timestamp of events. Data recording and storing activities can be performed using synchronous communication among the nodes through open-source sharing protocols. The software code can contain bugs and vulnerabilities, in which a full blockchain node can contain information about the whole blockchain.

Within the system, it would be possible to inculcate security and privacy. Encryption algorithm implementations can be observed in the form of TLS, RSA, ECC, ECDHE on the platform, and, as such users can be identified using public key or hash. Hash functions are designed using SHA-256d, SHA-256, Scrypt that are being adopted by cryptocurrencies handling privacy and security issues. Also, security can be enhanced by enabling cryptographic membership authentication scheme to support blockchain-based identity management. Also, blockchain technology prevents from IP spoofing and forgery attacks, also enable certificate authorities, data modifications from unauthorized users, information alteration using hard forks, data integrity is essential, data distribution among peers. Smart contracts can overcome the challenges of data distribution, enabling decentralized code for managing physical or digital elements.

Strong smart contracts involve high revocation and modification costs, and, requires methods for adding modifications that are legally required. Some blockchain-based applications are more complex and involve the use of smart contract oracles and decentralized autonomous organization, blockchain applications that are controlled by a set of immutable and incorruptible rules embedded in its source code.

Within the automotive industry, the applications can be extended to record keeping and transactions, target stakeholders who can be impacted by blockchain deployment. Benefits of this approach can be applicable to reducing errors, improving real-time access to data, supporting natural workflows around creation, modification, detection of data elements, auditing. Also, operational benefits possible are transactions can be transformed in real-time, overheads and cost-intermediaries can be removed, the risks of tampering, fraud, cyber-attacks can be minimized, services related to credit letters, financing, leasing, cross-border import and export systems can be improved, financial and logistics operations can be coupled with IoT devices. IoT devices when coupled with blockchain can help in tracking process and exchange transactions within the devices. The data from the different stakeholders are stored in centralized databases, which implies costly and unreliable business processes, records are tamper-proof, non-repudiation and immutability provide unique data view shared among the stakeholders, data transparency provides global access to the blockchain and can have trusted information for the big data analytics. Blockchain also can reduce the information asymmetry, preventing fraud, reducing systemic risk, data tampering and delayed communication, reducing verification processes for ensuring overall conformity and delivery, guarantee of the provenance and authenticity of components involved in the infrastructure, costs reductions. By encompassing the properties of blockchain technology, i.e., immutability, decentralization, irreversibility, accessibility, timestamp of transactions, non-repudiation, anonymity the linked chain of hashes become more secure. The characteristics of the data movement are mostly related to formal verification of smart contracts, self-amending property in which changes are allowed for voting on chain without needing the changes, consensus algorithm where energy and time needed to validate transactions are more efficient than PoW. This way it was possible to store data as well as ensuring security of data, inserting smart-contracts data with different logics as well as taking actions into account for blockchain transactions using either permissioned or permission-less blockchain mechanisms. Also, data storage and control logic can be linked to smart contracts as well over the network infrastructure [9].

The blockchains act as the ledger that securely stores the data across the network and external processes can be performed for integrating artificial intelligence with the blockchain, monitoring controlled stations on the blockchain network. The benefits presented are the possible use of global information within a secured and validated manner for faster and change adaptable behavior within the network infrastructure leading to higher productivity and easier maintenance. Business Intelligence application activities can be in the form of analyzing customer behavior for particular products bought, social network activity data, statistical analysis. Other criteria for implementing business intelligence solutions include determining sales for a particular region by considering holidays, geographical regions based on which the organization can prosper. For the platform, data migration activities can be carried out using the scripts implemented in

JSON, python, java, and scala languages. With the data extracts coming in the form of CSV, text files, salesforce data, Oracle, and Postgres database. Sentiment analysis and neural network implementations are useful for predicting the behavior of customers and sales. By utilizing Djikstra's algorithm, the shortest path can be determined between locations, products, warehouses, and, sales. The workflows can be designed using the algorithm of Support Vector Machine, Decision tree modeling of data for handling the activities of data pre-processing, data processing, and model performance attributes.

The mentioned criteria can be related to warehouse product availability, promotions, customer data, social media sentiment analysis, catalog list of products, and items categories. Finally, with the capability of integrating machine learning algorithms from the architectural perspective of artificial intelligence and business intelligence, prediction models related to predicting future sales can be determined based on the criteria of performance score. Within the context of industrial applications, it is possible to use the combination of different tools within the business intelligence environment as well as different data mining algorithms useful for predicting sales, and, managing customer calls and sales inventory. Preditice Modeling logistics can be implemented using Djikstra's and Web Mining algorithms. Tools that can implement these types of situations include R, and, Weka. Data logging can be performed using ElasticSearch, Splunk, and, monitoring using Control-M and Airflow. Within the business purpose of satisfying the strategic marketing initiatives, the information infrastructure can change. The data repositories storage can be in Hive, Cassandra, HBase, Impala, and execution modes of YARN, Tez, MapReduce for processing the data. Neural network modeling can be utilized with convolutional neural networks, recurrent neural networks for the processing of gross domestic and international warehouse solutions. It is also possible to include forecasting methods for predicting future periods based on the multi-attribute properties associated with the geographical areas.

The data can be expressed in the form of mean, standard deviation, median and interquartile range, count. The comparisons among groups can be identified by taking into consideration the chi-square test, one-way analysis of variance (ANOVA). Also, correlations between variables can be determined taking into consideration the Spearman’s rank correlation test characteristic. The feasibility of the variables when considered can include discriminating different subjects that can be analyzed by plotting receiver operating characteristic curve and calculating the area under ROC curve and specificity and sensitivity of data with the mean values plotted for the variables on the graph. Software’s utilized for performing this kind of analysis includes R, Excel, SPSS, where, if the p value $< .05$, it is considered statistically significant. The linear models developed within the environment can be compared using ANOVA the line of fit that can improve the replication timing for the products. This procedure can be applied to enhance the replication timing at each of the development point. Changes observed in residual sum of squares can be calculated by subtracting RSS from the first linear model without replication from the RSS from the second linear model with replication timing included. A negative change included in RSS means that RSS has decreased with the addition of replication timing for the predictive model and the model would better explain variations in the density functions. Comparisons of data performance can be done by using quantitative methods against categorical variables of data, and, the chi-square

test can be used for comparing categorical data versus categorical variables and Spearman's correlation coefficient can be used for comparing quantitative methods and qualitative methods and provided comparable results. Stratification can be performed using applied classification and regression trees. By utilizing the K-Means clustering algorithm's it is possible to divide the data into grades with their corresponding parameters.

A subset of results is obtained taking into consideration the clustering algorithm when the effects are optimal. The Spearman's rank correlation coefficient can describe the correlations among data that do not obey normal distributions which is a nonparametric method which can be utilized for getting the operational cost and attributes within the data.

4 Conclusion

The current automation industry that uses the Big Data technologies has a gap in neural network algorithm development. The current implementation and development shows the advantages of overcoming the shortcoming within the self-driving car model and its benefits to the industry.

The paper has been involved in identifying the challenges associated by implementing the self-driving car model in order to handle the driving conditions safely while inculcating driving modes by having additional layers of security encompassed into the model to prevent accidents. Also, future research in this area can be conducted to include additional metrics of security, privacy, safety, decision-making, behavioral-driven, networking layers to handle the movement of objects based on the principles of virtual reality and being able to control it.

Further research will also be needed to explore new approaches to building DAOs with the appropriate standardization and interoperability. In the future, by combining artificial intelligence with business intelligence, it is possible to include forecasting methods for predicting future periods based on the multi-attribute properties associated with the geographical areas.

References

1. Alsamhi, S. H., Ma, O., & Ansari, M. S. (2018). Predictive Estimation of the Optimal Signal Strength from Unmanned Aerial Vehicle over Internet of Things Using ANN. arXiv preprint arXiv:1805.07614.
2. Kuutti, S., Fallah, S., Katsaros, K., Dianati, M., Mccullough, F., & Mouzakitis, A. (2018). A Survey of the State-of-the-Art Localization Techniques and Their Potentials for Autonomous Vehicle Applications. IEEE Internet of Things Journal, 5(2), 829-846.
3. Perez, J. A., Deligianni, F., Ravi, D., & Yang, G. Z. (2018). Artificial Intelligence and Robotics. arXiv preprint arXiv:1803.10813.
4. Schwarting, W., Alonso-Mora, J., & Rus, D. (2018). Planning and decision-making for autonomous vehicles. Annual Review of Control, Robotics, and Autonomous Systems.
5. Uhlemann, E. (2016). Connected-vehicles applications are emerging [connected vehicles]. IEEE Vehicular Technology Magazine, 11(1), 25-96.

6. Wulfmeier, M. (2018). On Machine Learning and Structure for Mobile Robots. arXiv preprint arXiv:1806.06003.
7. Yang, J., & Coughlin, J. F. (2014). In-vehicle technology for self-driving cars: Advantages and challenges for aging drivers. International Journal of Automotive Technology, 15(2), 333-340.
8. Yin, C. (2018). Policy learning for task-oriented dialogue systems via reinforcement learning techniques.
9. Fraga-Lamas, P., & Fernández-Caramés, T. M. (2019). A review on blockchain technologies for an advanced and cyber-resilient automotive industry. IEEE Access, 7, 17578-17598.

Author´s Index

Announcement

World Congress DSA 2021

The Frontiers in Intelligent Data and Signal Analysis July 12 - 23, 2021, New York, USA

www.worldcongressdsa.com

We are inviting you to our fourth World congress on the Frontiers of Signal and Image Analysis DSA 2021 to New York, Germany.

This congress will feature three events:

- the 17th International Conference on Machine Learning and Data Mining MLDM (www.mldm.de),
- the 21th Industrial Conference on Data Mining ICDM (www.data-mining-forum.de),
- and the 16th International Conference on Mass Data Analysis of Signals and Images in Artificial Intelligence&Pattern Recognition MDA-AI&PR (www.mda-signals.de).

Workshops and Tutorial will also be given.

Come to join us to the most exciting event on Intelligent Data and Signal Analysis.

Sincerely your,
Prof. Dr. Petra Perner

MLDM www.mldm.de

icdm www.data-mining-forum.de

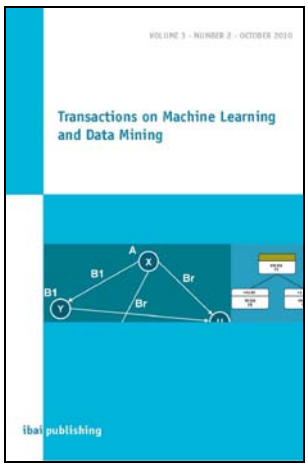
mda www.mda-signals.de

Journals by ibai-publishing

The journals are free on-line journals but having in parallel hardcopies of the journals. The free on-line access to the content of the paper should ensure fast and easy access to new research developments for researchers all over the world. The hardcopy of the journal can be purchased by individuals, companies, and libraries.

Transactions on Machine Learning and Data Mining

P-ISSN: 1865-6781 E-ISSN: 2509-9337



The International Journal "Transactions on Machine Learning and Data Mining" is a periodical appearing twice a year. The journal focuses on novel theoretical work for particular topics in Data Mining and applications on Data Mining.

Net Price (per issue): EURO 100
Germany (per issue): EURO 107 (incl. 7% VAT)

Submission for the journal should be send to:
info@ibai-publishing.org

For more information visited: www.ibai-publishing.org/journal/mldm/about.html

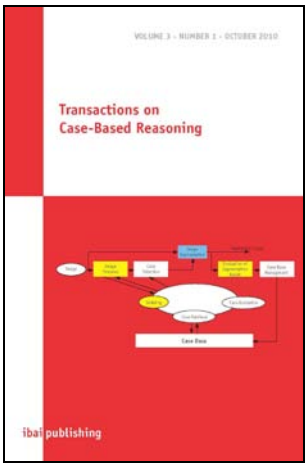
Transactions on Case-Based Reasoning

P-ISSN: 1867-366X E-ISSN: 2509-9345

The International Journal “Transactions on Case-Based Reasoning" is a periodical appearing once a year.

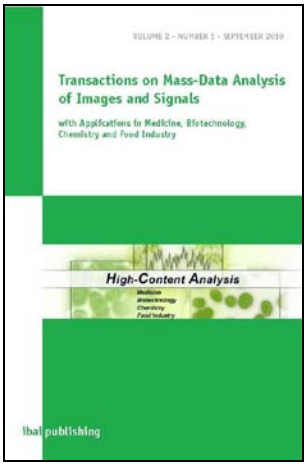
Net Price (per issue): EURO 100
Germany (per issue): EURO 107 (incl. 7% VAT)

Submission for the journal should be send to:
info@ibai-publishing.org



For more information visited: www.ibai-publishing.org/journal/cbr/about.html

Transactions on Mass-Data Analysis of Images and Signals ISSN: 1868-6451 E-ISSN: 2509-9353



The International Journal "Transactions on Mass-Data Analysis of Images and Signals" is a periodical appearing once a year.

The automatic analysis of images and signals in medicine, biotechnology, and chemistry is a challenging and demanding field. Signal-producing procedures by microscopes, spectrometers and other sensors have found their way into wide fields of medicine, biotechnology, economy and environmental analysis. With this arises the problem of the automatic mass analysis of signal information. Signal-interpreting systems which generate automatically the desired target statements from the signals are therefore of compelling necessity. The continuation of mass analyses on the basis of the classical procedures leads to investments of proportions that are not feasible. New procedures and system architectures are therefore required.

Net Price (per issue): EURO 100
Germany (per issue): EURO 107 (incl. 7% VAT)

Submission for the journal should be send to:
info@ibai-publishing.org

For more information visited: www.ibai-publishing.org/journal/massdata/about.php