ISBN - 978-989-9121-08-9
43rd PATTAYA International Conference on “Advances in Science, Engineering & Information Technology” (PASEIT-22)
Pattaya (Thailand) Aug. 25-26, 2022

BigData and Machine Learning Model for Building Efficient Marine Navigation System

Dr. Shabeen Taj G A[1], Dr. Mahadevan G[2], Dr. Sridhar C S[3] and Venkataraman R[4]

[1]Assistant professor, Departm, Computer Science and Engineering, Government Engineering College, Ramanagar, India.

[2]Principal, Annai college of Engineering and Technology, India.

[3]Professor Electronics and communication Engineering, Annai college of Engineering and Technology, India.

[4]Senior Manager, EY Bangalore, India.

Abstract**:** *with the development of diversity of marine data acquisition techniques, marine data grow exponentially in last decade, which forms marine big data. As an innovation, marine big data is a double-edged sword. On the one hand, there are many potential and highly useful values hidden in the huge volume of marine data, which is widely used in marine-related fields, such as tsunami and red-tide warning, prevention, and forecasting, disaster inversion, and visualization modeling after disasters. These fields requires efficient processing algorithm to analyze marine data with minimal time. Thus, to meet research challenges the manuscript firstly builds a fast and accurate classification algorithm using high dimensional data. Secondly, develop a novel distributed feature selection algorithm for analyzing various high dimensional data. Thirdly, for reducing classification time adopts Map Reduce and cloud computing framework.*

Keywords: *Bigdata, classification algorithm, Feature selection, High dimensional data, marine safety*

1. Introduction

Recently, the data volume all over the world is growing at an overwhelming speed, which is acquired by various devices with regard to Internet of Things and Social Networks. In this context, big data emerges and has been investigated extensively so far. In terms of marine field, countries around the world have launched several observing projects, for example, Argo [1], NEPTUNE-Canada [2], GOOS [3], OOI [4], IOOS [5], and so forth, and numerous marine observation satellites [6], [7]. Acquiringmarine data by various observing techniques leads to a sharp increase in data volume. For example, Argo [1] has set up four data centers and deployed up to 10231 buoys all over the world, for real-time acquiring marine data like temperature, salinity, acidity, density, and carbon dioxide. Even one data center alone has to process 21954 profile data with 657 active buoys over the whole of last year [8], [9]. The different data collection devices result in various data as well as their format. We denote the diverse data provisions. A marine observation satellite emitted by NASA, named as Aquarius [6], records all the element of ocean circulation, temperature, and ingredient and sea surface height every 7 days. Statistically, the data volume collected by Aquarius within every 2 months amounts to that collected by survey ships and buoys in 125 years [6]. By the end of year 2012, the annual data volume had been up to 30 PB (1 PB = 1024 $*$ 1024GB) maintained by NOAA and over 3.5 billion observational files would be gathered together from satellites, ships, aircrafts, buoys, and other sensors each day [7]. As all round marine observation systems and multiple observing techniques are widely put into service, data volume sharply increases, data type is greatly diversified, and data value is highly delivered, which forms marine big data.

Marine big data contains great values and embodies giant academic appeal, which can be transformed into a rich set of information for people to learn, exploit, and maintain the marine. For example, after analyzing the

https://doi.org/10.17758/DIRPUB12.DIR0822132

Argo data, it is found that the earth is seeking an intensification of global hydrological cycle [10]. Communities and species distribution can be determined by analysis of acoustic remote sensing data, which works as powerful scientific supportingevidence to maintain the marine ecological balance [11]. In addition, researches on forecasting and warning of undersea earthquake and tsunami can be successfully preceding, by analyzing observation data concerning seismic activity, faulting activity and midoceanic ridges acquired by Neptune project [12], [13]. In summary, marine big data supports forecasting and warning potential problems in the field of ecology, climate, and disasters and helps decision making.

In order to maximally exploit the value in marine data, it is of great realistic and theoretical significance to study on the management of marine big data concerning data storage, data analysis, quality control, and data security.At present, almost all the existing researches concentrate on solving general issues about big data management. As a kind of typical big data,marine big data features massiveness, diverse data provisions, high-dimension besides temporality, and spatiality, which brings exceptional challenges and problems. In terms of data storage, there are problems like weak scalability in storage system and dissatisfaction on timeliness. In terms of data analysis, there are still problems like slow processing speed and failure in real-time response.

Slower adoption in the maritime industry:

Even though big data has significantly benefitted industries such as finance, media, telecom and healthcare, its uptake by the maritime industry has been slow. According to a report by Ericsson, the maritime industry lags behind other transport industries in terms of its use of information and communications technology. Only a handful of marine companies currently leverage big data.

There are several benefits that the industry can derive through the use of big data. The industry generates roughly 100-120 million data points every day, from different sources such as ports and vessel movements.Companies can analyze these data points to identify efficiencies such as quicker routes or preferred ports. Ultimately resulting in an extra 5 to 10% increase in performance. Thus, this work present an efficient High dimensional marine BigData and Machine Learning Model for building efficient marine navigation system.

Research Contributionare as follows:

- The proposed system will aid in reducing processing time and memory overhead in classifying large high dimensional data.
- The proposed framework combines distributed feature selection methods and socio-eco safety models for efficient socio-eco safety analysis, which can reveal the valuable insights from the low-quality, high dimensionality, and huge-volume large socio-eco safety data.
- The proposed clustering based distributed feature selection and classification algorithm to select and identify the important features of data aid in reducing processing time.
- Further, the distributed feature selection and classification algorithm is implemented on distributed platform for real-time analysis of socio-eco safety high dimensional data analysis.
- Along with, we conducted correlative and collaborative analysis simultaneously to explore the direct and indirect relations among between socio-eco safety and its response indicators based on the identified socio-eco safety features by varying dimension size.

The rest of the paper is organized as follows. In section II the literature survey of various feature section and classification model to analyze high dimensional data is presented. The proposed distributed feature selection and classification model is presented in section III. In section IV the result attained by proposed distributed feature selection and classification model over existing model is discussed. The conclusion and future research direction of proposed research is discussed.

2. Literature Survey

Extensive survey is carried out in this section to enhance the performance of classification model for classifying high dimensional data. Feature selection aims to process multidimensional data by detecting the relevant features and discarding the irrelevant ones. Effective feature selection can lead to reduction of measurement costs yet generate a better understanding of the original domain [14], [15]. With respect to different selection strategies, feature selection algorithms can be categorized into four groups, namely the filter, wrapper, embedded, and hybrid methods. The filter methods present the feature selection process independent of any classifier and evaluate the relevance of a feature by studying the characteristics of training data using certain statistical criteria. The correlation-based feature selection, consistency-based filter, information gain, relief [16], fisher score [17], and minimum redundancy maximum relevance are the most representative filter techniques.

The wrapper methods integrate a classifier, such as SVM, KNN, and LDA, to select a set of features that have the most discriminative power. Representative wrapper feature selection methods include: wrapper C4.5, wrapper SVM, FSSEM, and ℓ1SVM. Other examples of the wrapper method could be any combination of a preferred search strategy and given classifiers. The embedded methods perform feature selection in the process of training and achieve model fitting to a given learning mechanism simultaneously. For example, SVMRFE trains the current features of the given data set by a SVM classifier and removes the least important features indicated by the SVM iteratively to achieve feature selection. Other embedded methods include FS-P, BlogReg and SBMLR. In summary, the filter methods, independent of any classifier, have lower computational complexity than wrapper methods yet with favorable generalization ability. Unlike filters, the wrapper methods are superior to filters in terms of classification accuracy, whereas they take more time due to the cost of expensive computation. The embedded methods, with lower computational cost than wrappers, are also integrated with classifiers, leading the risk of over-fitting.

Due to the shortcomings in each method, the hybrid methods [18], [19], [20], [21], [22], [23], and [24] are proposed to bridge the gaps between them. However, the existing feature selection methods are incapable of being adapted to safety and economic analysis. Since they analyze the data through its inherent knowledge characteristics, they cannot identify the feature cointegration and intrinsic association between safety and economic indicators. Besides, the low-quality and huge-volume characteristics of these big data present great challenges when the existing feature selection methods are directly applied to process inductive analysis.

In [25], [26], [27], [28], and [29] presented a model to bring a good performance tradeoff issues such as accuracy, memory and processing time. In [26], [27] addressed memory overhead issues considering high dimensional data. In [28], [29] overcome the processing time and accuracy issues but failed to address memory overhead issues in analyzing high dimensional data. In [30] they addressed processing time improvement and memory overhead issues by presenting a novel nearest neighbor classification algorithm. However, they considered only single dimension data. As a result, it cannot be applied to agro data which is multi-dimension in nature.

3. Distributed feature selection and classification model for building efficient marine navigation system

This work aims to overcome the above mentioned limitation of existing model and present a novel fast and accurate classification model of high dimensional data. This work aims to reduce the potentially huge set of candidate attributes produced by the preprocess layer to a small set of possible attributes, which are diverse and similar to the attributes in the original data set. However, there is no universal method for all problem settings, so we design a novel, systematic attribute selection approach for marine analysis. Our objectives of such an ideal approach are twofold such as parallel distributed clustering is generalized to select important attributes, and the attribute coordination based parallel clustering is designed to identify divergence ones. Thus, we can make full use of the divergence factors and their related important factors to mine the direct and indirect effects on marine development.

3.1. System model:

The proposed model explore the hidden relations between safety (of marine, environment) and its response indicators from a new angle and extract the meaningful knowledge from large high dimension data in order to derive right insights and conclusions based on an innovative distributed feature selection framework that integrates advanced feature selection techniques and socio-eco safety methods.

- First, in order to reduce the noise yet promote the data quality, we propose to use usability preprocessing, relative annual catastrophic computation, accident and growth rate computation and normalization techniques to clean and transform the collected large high dimensional data.
- Then, to distill the features related to socio-eco safety development from high-dimensional socio-eco safety data, distributed feature selection methods are proposed to quickly partition the importance of given socio-eco safety indicators.
- After that, the relations between response indicators and economic growth can be established by conducting correlative and collaborative analysis.
- Further, for enhancing speed of classification we adopt MapReduce framework such as HadoopMapReduce framework or any other open source Map Reduce framework.

3.2. Attribute selection and classification method:

For marine analysis, some records may be related to other records and some indicators can be represented by the combination of other indicators. Therefore, by approaching correlation analysis on economic data, the important and representative records and indicators can be identified. Distributed feature selection and classification (DFSC) algorithm, a density-based clustering algorithm, is a favorable method to investigate the correlations between data samples. It assumes that each data point is a potential cluster center and calculates a measure of the likelihood based on the density of surrounding data points. In this way, it can construct the relationships among all the data points. When decomposing the relationships to a same attribute, the contribution of the attribute to preserve the relationships can be achieved. According to this idea, we use DFSC to identify the important indicators for marine analysis.

The aim of this work is to establish the analytical models for marine life development so that the hidden patterns of marine and the correlations between marine and its response indicators can be captured. In this work, we describe the construction of marine life models in details based on the important and representative attributes identified by the proposed feature selection method at first. Then the relationships between marine life growth and its response indicators are discussed.

The architecture of proposed model is shown in below figure 1.

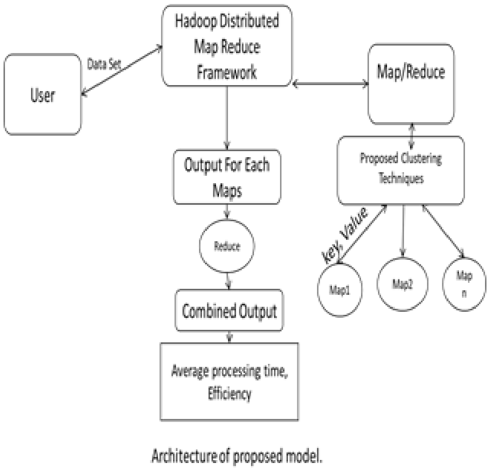
Hadoop Distributed Map Reduce Framework
Data Set
User
Map/Reduce
Proposed Clustering Techniques
Output For Each Maps
Reduce
key, Value
Map1
Map2
Map n
Combined Output
Average processing time, Efficiency
Architecture of proposed model.

Fig. 1. Architecture of proposed real-time efficient marine navigation system

The architecture is composed of module such as Hadoop Map Reduce (HMR) framework module: Hadoop is a distributed computing platform that perform feature selection and classification task in distributed manner. It takes input dataset (marine) from user for performing classification. The HMR is composed of Map and Reduce phase. In Map phase it read all input data and divide it into chunks of small data and perform execution parallel across different virtual machine. Post completion of Map phase Reduce phase is initialized. In this phase it reads the output of map phase and aggregate the classification output and store it in Hadoop distributed file system. Detail of HadoopMap Reduce execution can be obtained from [12]. The algorithm to perform distributed classification is shown in Algorithm 1.

Algorithm 1: Building Distributed feature selection and classification model using MR framework on HadoopHDInsight cluster

```
Input: Marine data, KeyValPair
Output:    $ConstructKeyPair(E$
Step 1:    $j \leftarrow MapRed\_functio$
Step 2:  read chunk of the marine data   with respect to function   using HDFS.
Step 3: construct key in parallel manner on each computing nodes with marine data   and keyValPair
Step 4:    $MapRed\_Cumulat$ // Synchronize all computing nodes.
```

The proposed DFSC method is built by dividing the data points at each stages into diverging influence of feature selection unique area using k-mean clustering. Post clustering, the same method is iteratively applied to the data points in a location area. The iterative computation is terminated when number of data points of an area is lesser than diverging influence of feature selection. The proposed feature selection and classification model is presented in Algorithm 2.

Algorithm 2: Distributed feature selection and classification model

```
Input: Marine dataset feature, diverging influence of feature selection, maximum iteration, center selection strategy to be applied.
Output: Clustering tree. (Classification output)
Step 1:if marine dataset is a set of diverging influence of feature selection then
Step 2:    build terminal node with feature points in marine dataset feature
Step 3:else
Step 4: choose diverging influence of feature selection data points from marine dataset feature using center selection strategy.
Step 5:    Set Converged to false
Step 6:    Set Iterations to Zero
Step 7:while converged = false&& iteration<maximum iteration do
Step 8:  cluster the feature points in marine dataset feature around closest centers
Step 9:    $Q_I$ averages of clusters in cluster the feature points in marine dataset feature around closest centers.
Step 10:if    $Q =$ then
Step 11:        Converged is set to true
Step 12:end if
Step 13:    $Q \leftarrow$
Step 14:    iterations iteration + 1
Step 15:end while
Step 16:for each cluster $D_j \in$ do
Step 17:    build non-terminal node with center
Step 18:    Continuously apply clustering method to the feature points in
Step 19:end for
Step 20:end if
```

The number of cluster to be considered for dividing the data at each node is a feature/attribute of the algorithm, known as the diverging influence and selecting number of cluster is significant for attaining good classification outcome. Another parameter of proposed DFSC algorithm is maximum iteration, which depict the maximum iteration to perform clustering process. Considering smaller iteration aid in reducing clustering time at the cost of accuracy. However, the proposed clustering will aid in attaining good convergence with minimal time which is experimentally shown below.

4. Experiment Result and Analysis

This section conducted experiment to evaluate the performance of DFSC model over existing classification model [30] in terms of computation time and memory overhead for performing analysis on raw unstructured high dimensional data into understandable and useful form. The result obtained is shown in Table I. The result shows ANN attain better performance than Random classification model. As a result, we compare proposed outcome performance improvement over ANN classification model [30]. The DFSC-Local classification model reduce computation/processing time and Memory overhead by 32.78% and 49.27% over ANN based classification method, respectively. Similarly, DFSC-Hadoop classification model reduce processing time and Memory overhead by 95.82% and 65.21% over ANN based classification method, respectively, and attain speedup performance of 16. Further, experiment are conducted to evaluate the effect of dimension size on classification performance which is shown in Fig. 2. We have varied the size of dimension as 4, 6, 8, and 10and evaluated the classification outcome in terms of processing time and Memory overhead. The experiment outcome shows when dimension size is increased the processing time and memory overhead increases. The overall result achieved shows scalable performance of DFSC model compared with existing classification model.

TABLE I: Proposed feature selection and classification algorithm performance evaluation over existing algorithm

Algorithm	Computation/ Processing time (s)	Memory overhead	Speedup
ANN classification model [30]	52.5	0.069	14
DFSC-local classification model	35.29	0.35	-
DFSC-Hadoop classification model	2.19	0.24	16

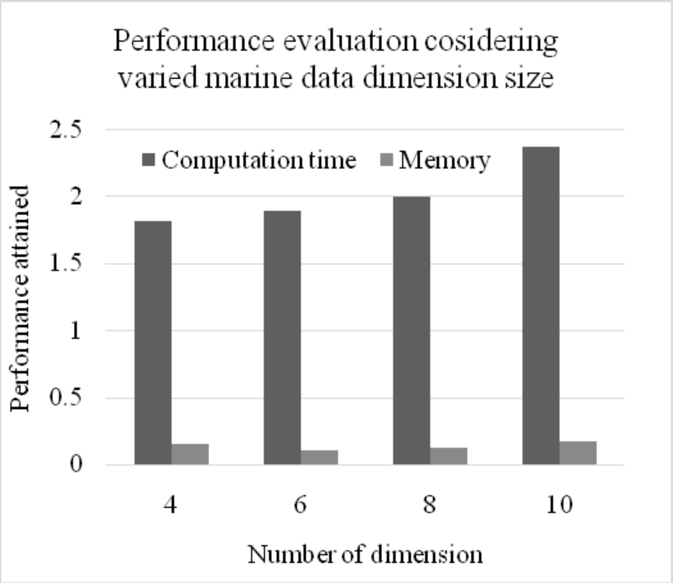

Fig. 2. Perforamnce evaluation of considering varine marine dta dimension size.

5. Conclusion

In this work we have proposed a novel feature selection based framework, aiming at effective and efficient analyzing the socio-eco safety marine data. In particular, it tries to learn the important features from the high-dimensionality, huge-volume, and low-quality socio-eco safety data for socio-eco safety marine model construction. Experiment is conducted on marine dataset and evaluate the performance of proposed DFSC model. The DFSC model attain better performance than in terms of reducing processing time and memory overhead when compared with existing model. The DFSC model reduces processing time by 32.78%, and memory overhead by 49.27% on local computing node over existing ANN classification model. Similarly, the DFSC model reduces processing time by95.82%, memory overhead by 65.21% over existing ANN classification model on parallel computing framework namely HadoopMapReduce framework. Along with, attain a speedup of 16. The overall result attain shows efficiency of proposed model. The future work would consider performance evaluation considering more dynamic dataset. Along with consider enhancing classification model and also consider different parallel computing model.

References

[1] "The International ARGO Project," http://www.argo.net/.

[2] Ocean Networks Canada, http://www.neptunecanada.ca/.

[3] "Ocean Observatories Initiative," http://oceanobservatories.org/.

[4] The Global Ocean Observing System, http://www.ioc-goos.org/.

[5] Intergrated Ocean Observing System, http://www.ioos.noaa.gov.

[6] National Aeronautics and Space Administration, http://www.nasa.gov/.

[7] National Oceanic and Atmospheric Administration, http://www.noaa.gov/.

[8] China Argo Data Center, http://www.argo.gov.cn/.

[9] "News in brief of Argo," China Argo Real-time Data Center, China, 2014, (Chinese), http://www.argo.org.cn/.

[10] P. J. Durack, S. E. Wijffels, and R. J. Matear, "Ocean salinities reveal strong global water cycle intensification during 1950 to 2000," Science, vol. 336, no. 6080, pp. 455–458, 2012.
https://doi.org/10.1126/science.1212222

[11] C. J. Brown, S. J. Smith, P. Lawton, and J. T. Anderson, “Benthic habitatmapping: a review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques,” Estuarine, Coastal and Shelf Science, vol. 92, no. 3, pp. 502–520, 2011.

https://doi.org/10.1016/j.ecss.2011.02.007

[12] G. C. Rogers, R. Meldrum, R. Baldwin et al., “The NEPTUNE Canada seismograph network,” Seismological Research Letters, vol. 81, no. 2, p. 369, 2009.

[13] A. B. Rabinovich, R. E. Thomson, and I. V. Fine, “The 2010 chilean tsunami off the west coast of Canada and the northwest coast of theUnitedStates,” Pure andAppliedGeophysics, vol. 170, no. 9-10, pp. 1529–1565, 2013.

https://doi.org/10.1007/s00024-012-0541-1

[14] S. Alelyani, J. Tang and H. Liu, ”Feature Selection for Clustering: A Review,” Data Clustering: Algorithms and Applications, vol.29, 2013.

https://doi.org/10.1201/9781315373515-2

[15] J. Liang, F. Wang, C. Dang and Y. Qian, ”A Group Incremental Approach to Feature Selection Applying Rough Set Technique,” IEEE Transactions on Knowledge and Data Engineering, vol.26, no.2, pp.294-308, 2014.

https://doi.org/10.1109/TKDE.2012.146

[16] L. Beretta and A. Santaniello,”Implementing ReliefF Filters to Extract Meaningful Features from Genetic Lifetime Datasets,” Journal of Biomedical Informatics, vol.44, no.2, pp.361-369, 2011.

https://doi.org/10.1016/j.jbi.2010.12.003

[17] Q. Gu, Z. Li and J. Han, ”Generalized Fisher Score for Feature Selection,” arXiv preprint arXiv:1202.3725, 2012.

[18] P. Ghamisi, M. S. Couceiro and J. A. Benediktsson, ”A Novel Feature Selection Approach Based on FODPSO and SVM,” IEEE Transactions on Geoscience and Remote Sensing, vol.53, no.5, pp.2935- 2947, 2015.

https://doi.org/10.1109/TGRS.2014.2367010

[19] M. Jamjoom, and K. E. Hindi, ”Partial Instance Reduction for Noise Elimination,” Pattern Recognition Letters, vol.74, no.4, pp.30- 37, 2016.

https://doi.org/10.1016/j.patrec.2016.01.021

[20] B. Xue, M. Zhang and W. N. Browne, ”Particle Swarm Optimization for Feature Selection in Classification: A Multi-objective Approach,” IEEE Transactions on Cybernetics, vol.43, no.6, pp.1656-1671, 2013

https://doi.org/10.1109/TSMCB.2012.2227469

[21] G. Casalino, N. Del Buono and C. Mencar, ”Subtractive Clustering for Seeding Non-negative Matrix Factorizations,” Information Sciences, vol.257, no.2, pp.369-387, 2014.

https://doi.org/10.1016/j.ins.2013.05.038

[22] J. Liu and Z. Jiang, ”Innovation-driven and Investment-supportive Strategies for China’s Economic Transformation based on Collaborative Theory,” Science of Science and Management of S.&T., vol.36, no.2, pp.25-33, 2015.

[23] E. P. Xing, Q. Ho, W. Dai, J. K. Kim, J. Wei, S. Lee, X. Zheng, P. Xie, A. Kumar and Y. Yu, ”Petuum: A New Platform for Distributed Machine Learning on Big Data,” IEEE Transactions on Big Data,vol.1, no.2, pp.49-67, 2015.

https://doi.org/10.1109/TBDATA.2015.2472014

[24] X. Hu, L. Tang and H. Liu, ”Embracing Information Explosion without Choking: Clustering and Labeling in Microblogging,” IEEE Transactions on Big Data, vol.1, no.1, pp.35-46, 2015.

https://doi.org/10.1109/TBDATA.2015.2451635

[25] L. Kuang, L. T. Yang, J. Chen, F. Hao and C. Luo, "A Holistic Approach for Distributed Dimensionality Reduction of Big Data," in IEEE Transactions on Cloud Computing, vol. 6, no. 2, pp. 506-518, 2018.

https://doi.org/10.1109/TCC.2015.2449855

[26] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronin, “Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2916–2929, 2013.

https://doi.org/10.1109/TPAMI.2012.193

[27] T. Ge, K. He, Q. Ke, and J. Sun, “Optimized product quantization,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 4, pp. 744–755, april 2014.

https://doi.org/10.1109/TPAMI.2013.240

[28] L. Bao, Q. Yang, and H. Jin, “Fast edge-preserving patchmatch for large displacement optical flow,” IEEE Transactions on Image Processing, vol. 23, no. 12, pp. 4996–5006, 2014.

https://doi.org/10.1109/TIP.2014.2359374

[29] D.Cozzolino, G.Poggi, and L.Verdoliva, “Efficient dense-field copymove forgery detection,” IEEE Transactions on Information Forensics and Security, vol. 10, no. 11, pp. 2284–2297, 2015.

https://doi.org/10.1109/TIFS.2015.2455334

[30] L. Verdoliva, D. Cozzolino and G. Poggi, "A Reliable Order-Statistics-Based Approximate Nearest Neighbor Search Algorithm," in IEEE Transactions on Image Processing, vol. 26, no. 1, pp. 237-250, 2017.

https://doi.org/10.1109/TIP.2016.2624141