International Journal of Future Generation Communication and Networking
Vol. 14, No. 1, (2021), pp. 1641–1648

Data Fusion methods and challenges in Machine Learning

Reena G.Bhati
Department of Computer Science
Tilak Maharashtra Vidyapeeth
Pune,India
reena4bhati@gmail.com

Abstract

Data fusion is a prevalent way to deal with incomplete raw data in order to collect reliable, usable and accurate information. A machine-learning framework that automatically learns from past experiences without programming specifically compares a variety of conventional probabilistic data fusion techniques, unintentionally renovates the fusion process via high computer and predictive abilities. However the literature also lacks a comprehensive analysis of the recent developments in data fusion machine learning. It is also useful to study and summaries the state of the art in order to gain a deeper insight into how machine learning can support and optimize data fusion. In this paper, we include a detailed survey of data fusion approaches based on machine learning. First, we give a comprehensive introduction to the context of data fusion and machine learning in terms of concepts, implementations, architectures, processes and typical techniques. Through the literature review, study and comparison, we finally come up with a range of open issues and suggest future directions for research in this area.

Keywords— *Data fusion, Fusion, Luo & Kay, machine learning*

I. INTRODUCTION

In an age of knowledge explosion, massive quantities of data are produced, collected and processed. In order to check for world rules to discover the essence and to obtain useful knowledge from data. We're more likely and more confident to draw a conclusion or to take a decision based on real-world facts, rather than to believe in experiences or intuition. Big data, on account of its "5V" characteristics, follow problems and challenges in the provision of data-driven services: amount, variance, speed, veracity, value. Obviously, in the modern age of big data, conventional literature data processing methods are difficult to satisfy. One of the most relevant research issues today is the way to collect reliable, useful and accurate information in large data.

The cyber world gives us a lot of data to dispose of. However the raw data obtained from different environments are heterogeneous, dynamic, incomplete, and of a vast scale, which brings us many challenges in translating them into useful knowledge. All sorts of data processing technologies, including but not limited to pre-processing, data storage, data transfer, data fusion, data collection, information retrieval, and so on are crucial to resolving these issues and arising from a range of processing theories. We're concentrating on data fusion in this article. It is a technology that combines data to obtain more reliable, insightful and accurate information than the original raw data, which are often ambiguous, imprecise, contradictory, conflicting and similar. Various approaches to data fusion in different fields of application have been developed. Generally speaking, fusion of information is widely used in networking wireless sensors, image processing, radar systems, object tracking, aim and recognition detection, intrusion detection, situation assessment, etc.[1].

In this paper, we are conducting a serious survey of data fusion techniques with machine learning. First, we present detailed basic definitions and context information of machine learning and data fusion. We then set out the crucial problems of data fusion and suggest a range of conditions for data fusion. We provide an in-depth overview of data fusion techniques based on machine learning by reflecting on the results of each job under analysis with the support and use of the parameters.

II. OVERVIEW OF DATA FUSION

Authors in[2] described data fusion in the book "Data Fusion Lexicon" as a process" Addressing the interaction, correlation and combination of data and knowledge from single and various sources for the achievement of refined role and identity estimates, and detailed and timely evaluation and

ISSN: 2233-7857 IJFGCN

relevance of situations and risks. The method is characterized by continuous refining of its estimates and evaluations and an evaluation of the need for additional sources or a shift in the process itself in order to obtain improved results." Authors in [1] thought that "knowledge fusion is an analysis of the process." Powerful methods for the automatic or semi-automatic transformation of information from various sources and points in time into a representation that offers efficient support for human or automated decision-making." For ease of understanding, we present the most critical elements of data fusion:

A. *Data sources: Single or multiple data sources from various places and at different times are involved in data fusion.*
B. *Operation: The operation of data combination and refining needs Knowledge that can be described as "transforming"*
C. *Purpose: To obtain better information with less potential for error in detection or prediction and higher reliability as a fusion objective. Examples of actual implementations are decision-making, object recognition, condition prediction, and so on.*

The dominance brought about by the fusion of multi-source data is very clear. Even in the as tatic single source method, combining sampling with replication will result in more accurate observation. On the other hand, distributed data fusion, particularly in wireless sensor networks, reduces the redundancy of data, which reduces the time and resource consumption and the frequency of data collisions in the data transport process. What is more in all data fusion applications, data is converted into a more value-added and higher-quality modality that allows a data fusion device to re-emerge the full view of the observed phenomenon. For example, data greatly increases its coverage in both the dimension of time and space. Appropriate handling of redundant data in other models may help to obtain improved, correct and reliable information with little imperfection.

An excellent and succinct architecture can also encourage interactions between researchers and developers, which promote research growth. In this context, we present a range of narrowly extended architectures of data fusion such as the Jointed Laboratory Directors (JDL) [2], architecture of Luo and Kay [3] and the Architects of Dasarathy [4].

Jdl

In 1986 the US Defense Department (DoD) first suggested that JDL be used for military use in particular. But it can also be modified easily for non-military use. A lot of updated or planned versions of JDL data fusion models were later released for extensive use of the architecture, which makes it suitable for many applications.

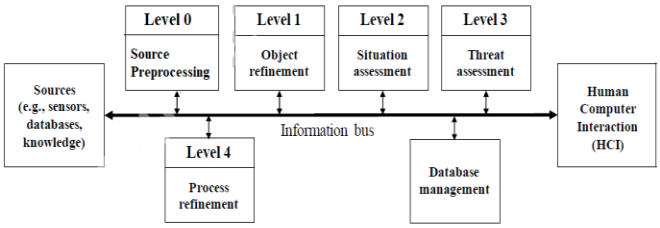

Figure 1. Joint Directors of Laboratories (JDL) architecture [15]

Multi-sensor integration and fusion were studied by Luo and Kay [3, 5]. They given an abstract level of integrated data based on the current general multi-sensor integration architecture shown in Figure 2.

ISSN: 2233-7857 IJFGCN

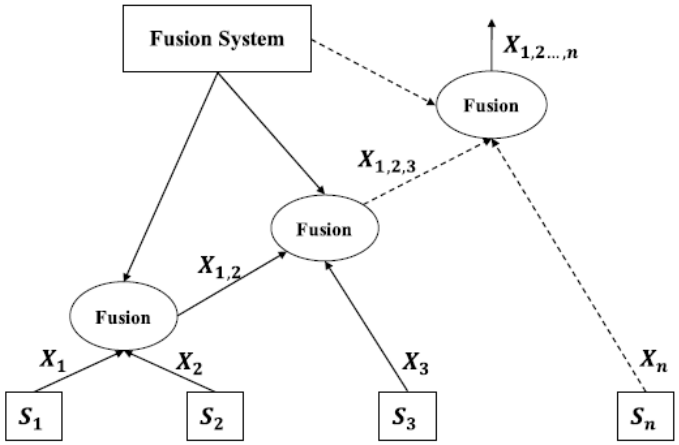

Figure 2. Luo and Kay's architecture [3]

Based on the three-layer fusion architecture of Luo & Kay, Dasarathy expanded it to five I/O characterization fusion processes in 1997[5]. In 1997, this fusion process was considered. Dasarathy felt that the demand for a more specific description led to some unclear conditions within the three-layer architecture.

III. DATA FUSION CHALLENGES

The The fusion of data also faces a variety of challenges to increase its effectiveness advantages[6] while in several realistic applications different model data fusion have been proposed to adapt to particular demands. The difficulty of the application environments in which sensors are situated, the diversity of data that should be integrated and so on are the key reasons for these challenges. We list some of them as follows in this subsection.

(1) Data imperfection: It is a common problem and a key issue to fix all data fusion approaches. There is often unclear, unreliable, ambiguous, unknown knowledge collected by sensors. Usually we will enhance data quality through the simulation of its imperfection and other skills and effective mathematical methods. Fusion performance can be imperfected when fusion cannot extract accurate and functional data. Fusion of data can be adverse.

(2) Data inconsistency: These uncertainties are caused by inherent noises in the measurements, sensors and environments. These noises result in a contour or a disorder, which is commonly called data incoherence. Data incoherence tends to have rather poor consequences on data fusion if a model of fusion cannot distinguish between noise causes.

(3) Data type heterogeneity: In different settings, data are gathered by sensors. They can be really different shapes as well. Often sensors are like people's eyes, nose, and mouth for various functions. Data fusion methods should be capable of integrating many types of data to determine the whole status of an entity.

(4) Fusion location: It is also a major problem for networks of wireless sensors and other fusion distributions. It is possible to merge data in a central or local node. Bandwidth and time are required first. We may later reduce the burden of collaboration but, because of loss of knowledge due to local mergers, we undoubtedly need to abandon the data quality. It is a difficult problem to balance fusion costs and merger performance.

(5) Dynamic fusion: Not only the data type and data recovery environment are causing the complexity of data merger but also its timeliness. Data could only be important for estimating system status, especially in the case of a time-varying system, within a limited period. In a real-time application setting, this issue should be managed well. It should be possible to distinguish between the right data order and its validation by the fusion node.

IV. LITERATURE REVIEW

ISSN: 2233-7857 IJFGCN

Traditional data fusion techniques include probabilistic fusion (e.g. bayesian fusion), belief reasoning proof, reasoning fusion, etc. [7]. Many surveys with various emphases on data fusion have been published in recent years. The IoT data fusion literature review was concluded [8] with statistical approaches including probabilistic methods, artificial intelligence and theory of IoT beliefs. Concentrating on IoT reduces analysis, while learning machinery covers a wide spectrum of data fusion. In contextual systems, the data fusion models used was based on [9]. The state-of-the-art sensor fusion techniques in mobile devices are summed up by [10]. The methods used to integrate data protection information were explored by [11]. [12] focused on the use of smart transport systems data fusion models. Computer security information merger processes were tested by authors [13]. The study of web knowledge fusion and integration [14] was given. In the internet of things, Authors in [15] investigated data fusion methods primarily aimed at secure, data security. In the above studies, we can see that our analysis shows different concentrations

Some works on machine learning provide an overview in some particular applications, especially in environments associated with large-scale data processing. Liao et al.[16] surveyed implementations and accomplishments of machine learning in the past ten years, for example (2000-2011). The success of machine learning in the real world issues of science and society has been reviewed by Rudin and Wagstaff[16]. Qiu et al. [17] also researched Large Data processing machine learning. In a literature review, they suggested five critical problems in the learning of big data. Zhang et al. [18] analyzed the profound learning representative work in broad data.
Banerjee et al.[19] suggested a multi-sensor data fusion, SVM, Short-Term Fourier (STFT) and time-based observer model for the hybrid method of failure detection. The system classifies a system into three types: good, degraded and ineffective.

Challa et al. [20] optimises the Bayesian approach to SVM data fusion as a compression technique. It minimises the objective function of SVM to translate inputs to a small collection of signals called vectors for help, the approximating function of which is defined. Further non-support vectors are ignored as there is no useful knowledge in associated signals. A kernel SVM dictionary is provided to change the model to achieve sound efficiency in various practical applications. This model has been validated by a density assessment framework that demonstrates excellent data compression efficiency. So, it's Performs in efficiency and extensibility of fusion as well. On the other hand, it acquires a number of training samples for its strength, so its strength and stability are not very effective. We think that fusion efficiency can be further improved by experimental results review.

Tong et al. [21] suggested the boiler drum level measurement information fusion model. Precise water level calculation in drums is important, since the difference between boiler load and water supply leads to serious consequences. A simple and effective method is used for calculating differential pressure level. However, the robustness of disruptions does not comply well. The neural network Radial Base Function (RBF) model is designed to map highly non-linear models. Fuse attributes such as pressure of operation, operating temperature, water supply, etc. More problems like local optimization are solved by the RBF networks compared to BP-NN. The RBF neural network can change the drum level measurement error well with a higher gradient descent algorithm. Simulation results show that, with a two-step training process, both the number of performance errors and the training time are reduced very rapidly, indicating the high efficiency and consistency of this fusion model.

Wang et al.[22] have proposed a hierarchical clustering algorithm based on the K-means the multi-target tracking process. In this paper, specifics have been established for target tracking problems with targets detected by multiple radars. The target path, for example, is erratic; the radar tracks are inconsistent or lack a common range. A hierarchical clustering model was developed to solve these problems. After pre-processing, Hausdorff distance, which describes the similar level between tracking data sets, was described and measured. Xiao and Liu [23] supported an Un-even clustering routing protocol and a simulated annealing algorithm. Compared to the classic protocol LEACH (Low Energy Adaptive Clustering Hierarchy), two noticeable differences are inconsistent initial clustering and complex time interval for cluster head re-selection. Fessi et al.[24] have suggested a data fusion

ISSN: 2233-7857 IJFGCN

model focused on clustering for intrusion. Detection to remedy the limitation of some of the current cluster literature, such as the lack of ability to detect composite attacks and positive attacks, the lack of productivity and a great deal of human interference.

Authors in [46] proposes a system for supporting data mining services, including data fusion, data cleaning, and other data preparation and pre-treatment services, based on massive data generated by smart grids. Targeted automated selection and fusion was carried out based on pre-processed power big data, based on the standard application scenarios of the power industry. A universal and reliable cleaning solution on the basis of widely used specifications for data mining is proposed for data cleaning. A [47] significant number of data are collected by several heterogeneous medical devices from health applications. These heterogeneous devices generate data in various formats. In general, an accurate judgment cannot be made in clinical decisions through a source of knowledge. Authors provide various viewpoints of fusion of data to test medical applications on this basis. In the evaluation of applications in the field of healthcare, these different proposed perceptual classifications are also implemented. The authors highlight the complexities of data fusion in the field of healthcare and explore future research opportunities in the healthcare field.

The methods of signal level data are among all the studies examined in this section. The merger is obviously overwhelming, with almost half of all the papers checked [25, 26-32, 33]. Some work merged features derived from raw data to achieve better fusion efficiency [34, 35, 36-38, 39, 40]. In [41-44], researchers extracted knowledge and blended high-level decisions. During the survey, we find that the data fusion application landscape with machine learning is varied. Representative merger scenarios involve, but are not limited to, WSN[29, 33, 38, 40, 41], radar monitoring and remote systems[31, 36, 45], detection of intrusion [25], generation of credibility [35], mechanical engineering scenarios [27-28, 34], and so on. More and more machine-based fusion is required in all kinds of fields. Much of the works examined resolved the "data imperfection" Data fusion crisis. In addition, some of the works used in distributed systems and WSN have established the position fusion problems with SVM and K-Means [29, 38]. We are of the opinion that machine learning methods cannot address all the problems of data fusion, such as data conflict due to the limitation caused by its existence. Data fusion models are based on a variety of typical machine learning approaches. Supervised learning approaches such as SVM [26, 27, and 28] & NN [35, 30] have been commonly used. Correspondingly, clustering models [42, 43, and 44] and K-Means [25, 22, 38, and 45] were also adopted to increase the efficiency and performance of mergers. SVM is good at handling high-dimensional data, while NN is good. More adept at learning from incomplete and ambiguous data, or when a system is difficult to explain with a linear formula. There is no clear connection between the forms of fusion and the methods of machine learning. Typically, machine learning approaches are good at dealing with the classification problem during the fusion process.

V. MACHINE LEARNING FOR DATA FUSION

The lowest data fusion stage is signal fusion in accordance with the Luo & Kay architecture. Data outputs with high precision, reliability and few noises are reported with raw data inputs obtained from sensors. Or function outputs are extracted to represent an element of observation directly. In signal fusion, image fusion (also called pixel fusion) and others related scenarios, signal level models are also applied.

SVM provides the proper fusion feature at the signal level as a representative supervised machine learning algorithm. Authors in [48] proposed a hybrid approach based on SVM, Short Term Fourier Transform (STFT), multi-sensor data fusion and a time-based observer model for the detection of defects. The system divides a system into three types: good, degraded and broken. Figure explains the basic scheme for the SVM based fault classification.

ISSN: 2233-7857 IJFGCN

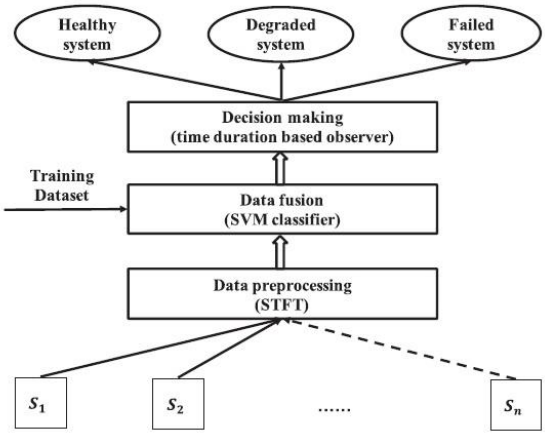

Figure 3. The model structure of [48]

VI. Open Issues & Future Research Direction

First, simplex are the machine learning methods used to fuse results. Most of the listed models of machine learning for data fusions are based on classical methods and simple neural networks, SVM and clustering. SVM and clustering processes also strive for high accuracy classification. NN is ideal to characterize dynamic structures that are unpredictable. The strength of machine learning methods can however be far more than that. For example, in ten years to come, deeper learning will be considered an important area of research in artificial intelligence. Deep learning defines methods simulating complex human neural networks. The framework would have greater precision and learning efficiency in comparison to the simple neural networks, more hidden layers introduced into the network. The absence of profound data fusion learning approaches leads one to look at new concepts.

Based on the indicated open issues few potential research directions is to explore more application scenarios for machine learning based data fusion. It is gratifying to see a number of models being applied to various scenarios after the great growth of machine learning for data fusion over decades, such as intrusion detection, target identifier and monitoring for military and non-military use, human-computer interaction, navigation, and Geographic Use etc. Moreover, there are several other scenarios for applications that can use data fusion approaches based on machine learning. The powerful ability to learn machines in nonlinear mapping offers additional data fusion possibilities.

VII. CONCLUSION

This study carried out a systematic analysis of the computer literature Learning to merge info. Firstly we present basic context information on data fusion and machine learning. A variety of parameters is suggested for assessing the works reviewed in this paper with a view to remarkably commenting on their pros and cons. The recent literature is carefully reviewed based on the degree of fusion found and the form of machine learning, and then used a table to summarize the key findings of our study. On the basis of our survey, we went ahead to define a range of open issues and suggested some potential directions for study that warrant further investigation. This study provides a succinct and detailed guide for researchers and practitioners in the field of data fusion machine learning.

REFERENCES

1. D. L. Hall and J. Llinas, An introduction to multisensor data fusion, Proceedings of the IEEE, 85 (1) (1997) 6-23.
2. F.E. White, Data Fusion Lexicon, (1991).
3. R. Luo and M. Kay, Multisensor integration and fusion: issues and approaches, SPIE Sensor Fusion, 931 (1988) 42-49.
4. B. Dasarathy, Sensor fusion potential exploitation-innovative architectures and illustrative applications, Proceedings of IEEE, 85 (1) (1997) 24-38.

ISSN: 2233-7857 IJFGCN

5. Ren C. Luo, C. C. Yih, K. L. Su, Multisensor fusion and integration: approaches, applications, andfuture research directions, IEEE Sensors Journal, 2 (2) (2002) 107-119.
6. B. Khaleghi, et al., Multisensor data fusion: a review of the state-of-the-art, Information Fusion, 14 (1)(2013) 28-44.
7. C. Federico, A review of data fusion techniques, The Scientific World Journal, (2013) 1-19.
8. F. Alam, R. Mehmood, I. Katib, N. N. Albogami and A. Albeshri, Data fusion and IoT for smartubiquitous environments: a survey, IEEE Access, 5 (2018) 9533-9554.
9. S. Gite and H. Agrawal, On context awareness for multisensor data fusion in IoT, Springer India, 381(2016) 85-93.
10. I. M. Pires, N. M. Garcia, N. Pombo, and F. Flórez-Revuelta, From data acquisition to data fusion: a comprehensive review and a roadmap for the identification of activities of daily living using mobile devices, Sensors, 16 (2) (2016) 184.
11. G. Navarro-Arribas and V. Torra, Information fusion in data privacy: a survey, Information Fusion, 13(4) (2012) 235-244.
12. N. Faouzi, H. Leung and A. Kurian, Data fusion in intelligent transportation systems: progress and challenges – A survey, Information Fusion, 12 (1) (2012) 4-10.
13. I. Corona, G. Giacinto, C. Mazzariello, F. Roli and C. Sansone, Information fusion for computer security: state of the art and open issues, Information Fusion, 10 (4) (2009) 274-284.
14. J. Yao, V. Raghavan and Z. Wu, Web information fusion: a review of the state of the art, Information Fusion, 9 (4) (2008) 446-449.
15. W. X. Ding, X. Y. Jing, Z. Yan, L. T. Yang, A survey on data fusion in Internet of Things: towards secure and privacy-preserving fusion, Information Fusion, 51 (2019) 129-144.
16. S. Liao, et al. Data mining techniques and applications – a decade review from 2000 to 2011, Expert Systems with Applications, 39 (12) (2012).
17. J. Qiu et al., A survey of machine learning for big data processing, EURASIP Journal on Advances in Signal Processing, 2016 (1) (2016).
18. Q. Zhang et al., A survey on deep learning for big data, Information Fusion, 42 (2018) 146-157.
19. T.P. Banerjee and S. Das, Multi-sensor data fusion using support vector machine for motor fault detection, Information Sciences, 217 (24) (2012) 96-107.
20. S. Challa, M. Palaniswami and A. Shilton, Distributed data fusion using support vector machines, International Conference on Information Fusion, 2 (6) (2013) 881-885.
21. [34] W. Tong, B. Li, X. Jin, Y. Yang and Q. Zhang, A study on model of multisensor information fusion and its application, in Proceedings of International Conference on Machine Learning and Cybernetics, 2006, pp. 3073-3077.
22. H. Wang et al., An algorithm based on hierarchical clustering for multi-target tracking of multi-sensor data fusion, In Proceedings of Control Conference. IEEE, 2016, pp. 5106-5111.
23. L. Xiao and Q. Liu, A data fusion using un-even clustering for WSN, in Proceedings of International Conference on Advanced Intelligence and Awareness Internet, 2012, pp. 216-219.
24. B.A. Fessi, S. BenAbdallah, Y. Djemaiel and N. boudriga, A clustering data fusion method for intrusion detection system, in Proceedings of 11th IEEE International Conference on Computer and Information Technology, 2011, pp. 539-545.
25. X. Guo, D. Wang and F. Chen, An anomaly detection based on data fusion algorithm in wireless sensor networks, International Journal of Distributed Sensor Networks, 2015 (2015) 1-10.
26. T.P. Banerjee and S. Das, Multi-sensor data fusion using support vector machine for motor fault detection, Information Sciences, 217 (24) (2012) 96-107.
27. S. Challa, M. Palaniswami and A. Shilton, Distributed data fusion using support vector machines, International Conference on Information Fusion, 2 (6) (2013) 881-885.
28. M. S. Fahmy et al., Biometric fusion using enhanced SVM classification, in Proceedings of International Conference on Intelligent Information Hiding and Multimedia Signal Processing IEEE, 2008, pp. 1043-1048.

ISSN: 2233-7857 IJFGCN

29. H. Shu, Y. Wang and J. Jiang, Multi-radar data fusion algorithm based on K-central clustering, in Proceedings of International Conference on Fuzzy Systems and Knowledge Discovery, 2007, pp. 617-621.
30. K. Kolanowski, A. Swietlika, R. Kapela, J. Pochmara and A. Rybarczyk, Multisensor data fusion using Elman neural networks, Applied Mathematics & Computation, 319 (2017) 236-244.
31. L. Xiao and Q. Liu, A data fusion using un-even clustering for WSN, in Proceedings of International Conference on Advanced Intelligence and Awareness Internet, 2012, pp. 216-219.
32. W. Tong, B. Li, X. Jin, Y. Yang and Q. Zhang, A study on model of multisensor information fusion and its application, in Proceedings of International Conference on Machine Learning and Cybernetics, 2006, pp. 3073-3077.
33. H. Shu, The application of cell-based clustering algorithm dealing with radar data fusion, in Proceedings of 2008 Congress on Image and Signal Processing, 2008.
34. S. Xiao, Y. Zhang, X. Liu, and J. Gao, Alert fusion based on cluster and correlation analysis, in Proceedings of International Conference on Convergence and Hybrid Information Technology, 2008, pp. 163-168.
35. N. Ghosh et al., Estimation of tool wear during CNC milling using neural network-based sensor fusion, Mechanical Systems & Signal Processing, 21 (1) (2017) 466-479.
36. Y. Cao, T. J. Huang and Y. H. Tian, A ranking SVM based fusion model for cross-media meta-searchengine, Frontiers of Information Technology & Electronic Engineering, 11 (11) (2011) 903-910.
37. A. Starzacher and B. Rinner, Embedded realtime feature fusion based on ANN, SVM and NBC, in Proceedings of IEEE International Conference on Information Fusion, 2009, pp. 482-489.
38. R. Pouteau and S. Benoît, SVM selective fusion (self) for multi-source classification of structurally complex tropical rainforest, IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 5 (4) (2012) 1203-1212.
39. Z. He, Accelerometer based gesture recognition using fusion features and SVM, Journal of Software, 6 (6) (2011) 1042-1049.
40. D. Qiu et al., The study of self-organizing clustering neural networks and applications in data fusion, in Proceedings of IEEE World Congress on Intelligent Control & Automation, 2008.
41. B. Bigdeli, F. Samadzadegan and P. Reinartz, A decision fusion method based on multiple support vector machine system for fusion of hyperspectral and LIDAR data, International Journal of Image & Data Fusion, 5 (3) (2014) 196-209.
42. C.L. Bowman and M.S. Murphy, Description of the VERAC NSource tracker/correlator, Naval Res Lab. Report R-01O-80, (1980).
43. R. Luo and M. Kay, Multisensor integration and fusion: issues and approaches, SPIE Sensor Fusion, 931 (1988) 42-49.
44. X. Jing, Z. Yan, and P. Witold, security data collection and data analytics in the Internet: a survey. IEEE Communications Surveys & Tutorials, 21 (1) (2018) 586-618.
45. G. Giacinto, R. Perdisc, Del Rio M and F. Roli, Intrusion detection in computer networks by a modular ensemble of one-class classifiers, Information Fusion, 9 (1) (2008) 69-82.
46. Z. Lv, W. Deng, Z. Zhang, N. Guo and G. Yan, "A Data Fusion and Data Cleaning System for Smart Grids Big Data," *2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking*
47. S. Kalamkar and G. Mary A., "Clinical Data Fusion and Machine Learning Techniques for Smart Healthcare," *2020 International Conference on Industry 4.0 Technology (I4Tech)*, Pune, India, 2020, pp. 211-216
48. T.P. Banerjee , S. Das , Multi-sensor data fusion using support vector machine for motor fault detection, Inf. Sci. 217 (24) (2012) 96–107 .

ISSN: 2233-7857 IJFGCN