Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 4276 - 4283
Received 05 March 2021; Accepted 01 April 2021.

Duplication Avoided Data Fusion Technique for Internet of Things

D.Balakrishnan[1], T.Dhiliphan Rajkumar[2], S.Dhanasekaran[3], B.S.Murugan[4]

[1]Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil-626126, Tamilnadu, India

[2]Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil-626126, Tamilnadu, India

[3]Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil-626126, Tamilnadu, India

[4]Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil-626126, Tamilnadu, India

ABSTRACT

Growing technologies and popularities increased number of technology users which tends to generate more volume of data. Handling these data will be more difficult in practical which needs to be fine tuned before giving it for any processing. This is performed in our previous research work by introducing the method namely Semantic based Hierarchical Data Fusion Technique (SHDFT). However computational time of the previous work will be more due to presence of more duplicate data. This is resolved in this research work by introducing the method namely Duplication avoided Data Fusion Technique (DDFT). In this research work, before data fusion, duplication avoidance is done in order to reduce the computational burden of handling repeated data's. In this research work, initially context based feature extraction is done for extracting the useful features and repeated features. After feature extraction, duplication avoidance is done with the concern of similar concept. After duplication avoidance data fusion is done using hierarchical data fusion technique as done in previous work SHDFT. Finally performance of the data fusion technique is tested using Improved Convolutional neural network. The overall evaluation of the research work is done in the matlab against existing work in terms of accuracy, precision, recall and f-measure.

Keywords: Contextual similarity, computational burden, duplication avoidance, data fusion, feature extraction, deep learning technique

Introduction

World has turning into more digitized which leads organizations and industries to produce more volume of data [1]. Handling this large volume of generated data will more challenging task which needs to be handled with more concern [2]. Organization and industries working on same application might produce more similar data which would increase computational burden of handling these data [3]. As more repeated data's are present this would also contain more useful information. Finding of repeated data with the concern of useful information is more difficult task which needs to be done with more concern [4].

Data mining techniques are most widely used to discover the conceptual similarity based features from the input data [5]. There are many researches has been conducted earlier to discover the useful information and also avoid the repeated information. Data fusion is the technique which is used for integrating the data's together, thus the size of data can be reduced without losing the concept [6]. Data fusion process will be affected by various factors such as repeated information, noise information and so on [7]. This factors needs to be taken in mind before performing data fusion to improvise the fusion performance and accuracy.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 4276 - 4283
Received 05 March 2021; Accepted 01 April 2021.

When all is said in done, all undertakings that request any sort of boundary assessment from various sources can profit by the utilization of information/data fusion strategies [8]. The terms data fusion and information fusion are commonly utilized as equivalents; yet in certain situations, the term information fusion is utilized for crude information (got straightforwardly from the sensors) and the term data fusion is utilized to characterize effectively handled information [9]. In this sense, the term data fusion infers a higher semantic level than information fusion [10]. Different terms related with information fusion that commonly show up in the writing incorporate choice fusion, information blend, information conglomeration, multisensor information fusion, and sensor fusion.

Duplication is the presence of repeated information in the gathered data which would lead to more memory consumption and increased computational burden [11]. Duplicated data needs to be avoided in order to improvise the performance of the data fusion task. And also it is required to concentrate on useful information when avoiding the duplicated content from the input data. Duplication avoidance needs to be performed with more concern by eliminating the repeated terms from the input data [12].

The main goal of this research work is to introduce the technique that can perform the data fusion with the concern of duplication. It is done by avoiding the repeated terms from the input data before performing the data fusion. Here repeated data's are avoided by extracting the features based on contexual similarity. Based on contextual similarity duplication is avoided. After duplication avoidance, data fusion is performed. Finally improved Convolutional neural network is introduced for testing the performance of the classifier.

The overall organization of the research work is given as follows: In this section, introduction about data fusion and need of duplication is given. In section 2, comparison analysis of various research methodologies has been given. In section 3, discussion of the proposed research work is given with the suitable examples and explanation. In section 4, comparison analysis of the research work is given in terms of obtained results. Finally in section 5, overall conclusion of the research work is given based on obtained numerical outcome.

Literature Review

Sivaram et al [13] performed data fusion over gathered online data by introducing the crossover based Tabu Genetic algorithm. This research work attempted to fine tune the input data by applying the tenure point operator. This research method ensures the optimal and reliable information retrieval outcome by extracting the more useful information.

Losada et al [14] introduced rank fusion model for ensuring the accurate and efficient data fusion outcome. Here score data features will be calculated based on data characteristics which will leads to accurate and reliable data fusion outcome. Based on evaluated score, this research work will assign the rank values, thus the accurate outcome can be ensured.

Majumder et al [15] performed data fusion for the data gathered from the multiple sensors by adapting the clustering approach. Here fuzzy clustering technique and predictive tools are utilized for the efficient data fusion outcome. This research work ensures both compactness and distinctness of the data at the time of data fusion process.

Ahmad et al [16]attempted to perform social media data fusion by adapting the CNN and GAN technique. The evaluation of this research work ensures that this technique guarantees increased average mean and precision values for increased cut off values.

Simonetta et al [17] tried to process music information by adapting the multi model architecture. In this work fusion process is carried out to handle the music information more politely which will lead to increased computational efficiency.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 4276 - 4283
Received 05 March 2021; Accepted 01 April 2021.

Chen et al [18] introduced land cover classification accuracy by combining the RS features through data fusion task. Here performance analysis of the work is carried out over Landsat operational land imager dataset. The overall analysis of the research work proves that the proposed work ensure the accuracy prediction outcome with optimal and reliable data fusion process.

Zhang et al [19] adapting the deep learning model for the efficient and accurate multi sensor data fusion outcome. Here performance analysis has been carried out on 7 datasets which involved 7 degradation stages and 9 working conditions. Here more characteristics of input data is concentrated to improvise the data fusion outcome.

Duplication Avoided Data Fusion

In this research work, before data fusion, duplication avoidance is done in order to reduce the computational burden of handling repeated data's. In this research work, initially context based feature extraction is done for extracting the useful features and repeated features. After feature extraction, duplication avoidance is done with the concern of similar concept. After duplication avoidance data fusion is done using hierarchical data fusion technique as done in previous work SHDFT. Finally performance of the data fusion technique is tested using Improved Convolutional neural network. The overall flow of research work is shown in the following figure 1.

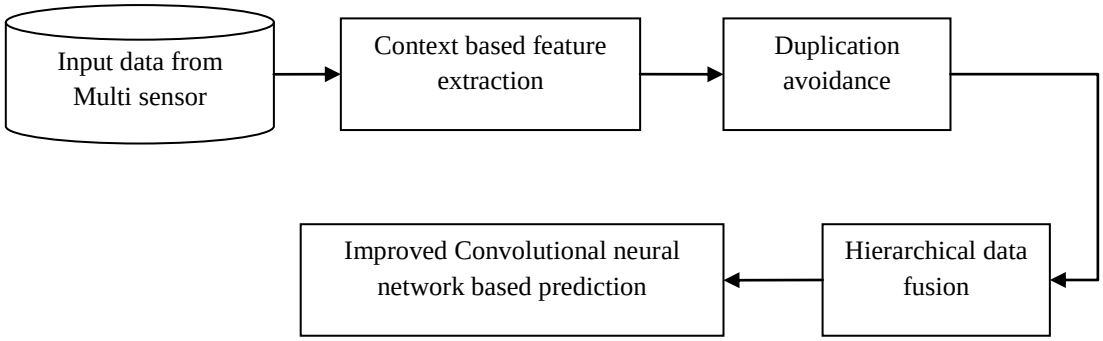

Figure 1. Processing flow of data fusion process

Context Based Feature Extraction

In this work context based feature extraction is done to improvise the data fusion processing outcome. This is attained by concentrating on the context features which tends to provide the conceptual similarity between different kinds of texts. The features that are extracted to find the conceptual similarity are explained below:

Context Feature: In this work context feature is extracted from the input samples which will provide the location and meaning of the input data.

ADR lexicon feature: ADR lexicon feature represents the combination of the ADR list and the list of UML ID. This feature is used to differentiate the source information from where the data is obtained. This feature is represented by using binary function values.

Parts-of-Speech feature: This feature is used to represent the different segments that define the verbal meaning of the content. This feature is extracted from the input data with the help of Stanfordparser tool.

Negation feature: Negation feature is used represent the negative contract terms present in the input documents. This feature is represented by using the negation words such as not, no, and so on. Negation feature will reflects the different terms that are different to each other from the document.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 4276 - 4283
Received 05 March 2021; Accepted 01 April 2021.

Embedding Cluster feature: Embedding cluster feature is used to represent the different number of features which represents the similar concepts. Here, unlabelled instances will be extracted from the input data which will be grouped together.

The above features will be extracted from the input document in order to ensure the accurate and reliable data fusion outcome. Based on this extracted features duplication avoidance will be done, thus the accurate and efficient data fusion is done.

Duplication Avoidance

In this work duplicate data removal is done using two levels. In this work data redundancy is done by adapting the fuzzy technique. Here initially input data will be grouped based on the contextual similarity which is found in terms of features extracted in the previous step. After grouping, duplication avoidance is done by adapting the fuzzy region which will find the more similar data. Here fuzzy classification algorithm ensures the matching and removal of the similar and repeated data from the input data. Thus the data fusion can be performed accurately. The fuzzy classification algorithm is explained below:

Algorithm: Fuzzy based duplication avoidance

Input: Input data samples

Output: Duplication avoided input samples

1: Randomly create the membership matrix U

2: Find the error values based on similarity level

3: Repeat the process until threshold value obtained

4: If membership function value not reached threshold

Find the membership for new set of data and repeat the task

5: Else

Output the training set

Hierarchical Data Fusion

In this work, data fusion is performed in the hierarchical manner. That is data fusion is performed in three levels as done in our previous research work. In our previous research work conceptual similarity is calculated between different terms based on which data fusion will be carried out. The concept similarity between the two data's can be calculated by using the following formulae:

$$\text{Sim}(w_1, w_2) = \max_{i = 1,..,m, j = 1,..,n} \text{Sim}(S_{1i}, S_{2j})$$

Where w_1, w_2 → English words

S → concepts1

By using the concept similarity measured above, semantic similarity can be computed as like below:

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 4276 - 4283
Received 05 March 2021; Accepted 01 April 2021.

$$\text{Sim}\,(p_1, p_2) = \frac{\alpha}{d + \alpha}$$

Where p_1 and p_2→ mean values

α→ adjustable parameter

d → positive integer

By using the above formulae, semantic similarity between the data items can be calculated based on which final prediction can be made.The above process will be resultant with data fusion outcome accurately.

Improved Convolutional Neural Network

In this work Improved Convolutional Neural Network (ICNN) is introduced for data fusion performance analysis. In this work duplication avoided features termed as DF are given input to the ICNN. Those DF features will be convolved with three masks X1, X2 and X3. In this work logistic sigmoid activation function and hadamard product is used for the process the input in the hidden layer. The hidden neuron resultant with X_1, X_2, and X_3 along with weight matrix U_j^k where j = 1, 2, 3 and k = 1, 2, 3.

Lets assume DF is the features which is represented as $M_F \times N_F$. Here 2D filter is applied on input features of matrix $M_W \times N_W$. The outcome of convolution operation performed over input features are represented as in equation 2.

$C_j = F * M_j$

* → 2D convolution

The 2D data matrix C_j has size $(M_X + M_W - 1) \times (N_X + N_Y - 1)$ with (m, n)th entry.

$$C_j(m, n) = \sum_{a=1}^{M_W} \sum_{b=1}^{N_W} X(a - m, b - n) W_j(a, b)$$

Finally logistic sigmoid function s is applied on the output matrices obtained from the hidden layer which is represented as in equation 4.

$$Z_j(m, n) = s\left(c_j(m, n)\right) = \frac{1}{1 + \exp\left(-C_j(m, n)\right)}$$

If the network consists of K number of output neurons then weight matrix U_j^k will be multiplied with the j number of hidden neurons Z_j element wise. The calculation procedure of kth output neuron using softmax activation is given in the equation 5.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 4276 - 4283
Received 05 March 2021; Accepted 01 April 2021.

$$a_k^t = \frac{\exp\left(\sum_{j=1}^{J} e^T Z_j \odot U_j^k e\right)}{\sum_{k1=1}^{K} \exp\left(\sum_{j=1}^{J} e^T Z_j \odot U_j^{k1} e\right)}$$

Where

$\odot$→ Hadamard product

e→ vector of length $(M_X + M_W - 1)(N_X + N_W - 1)$

Results and Discussion

In this section, numerical evaluation of the proposed research methodology is done in terms of various performance measures to analyze the performance improvement of the proposed and existing research methodologies. The matlab simulation environment is used to implement the proposed research methodology. The performance measures considered in this work are listed as follows: "Accuracy, Precision, Recall and F-Measure". The performance metrics values are given in the following table 1.

Table 1. Performance metric values

Metrics	Methods		
	DHDFA	SHDFT	DDFT
Accuracy	65	87	91
Precision	74	98	99
Recall	79	98.6	99.2
F-Measure	87	99	99.4

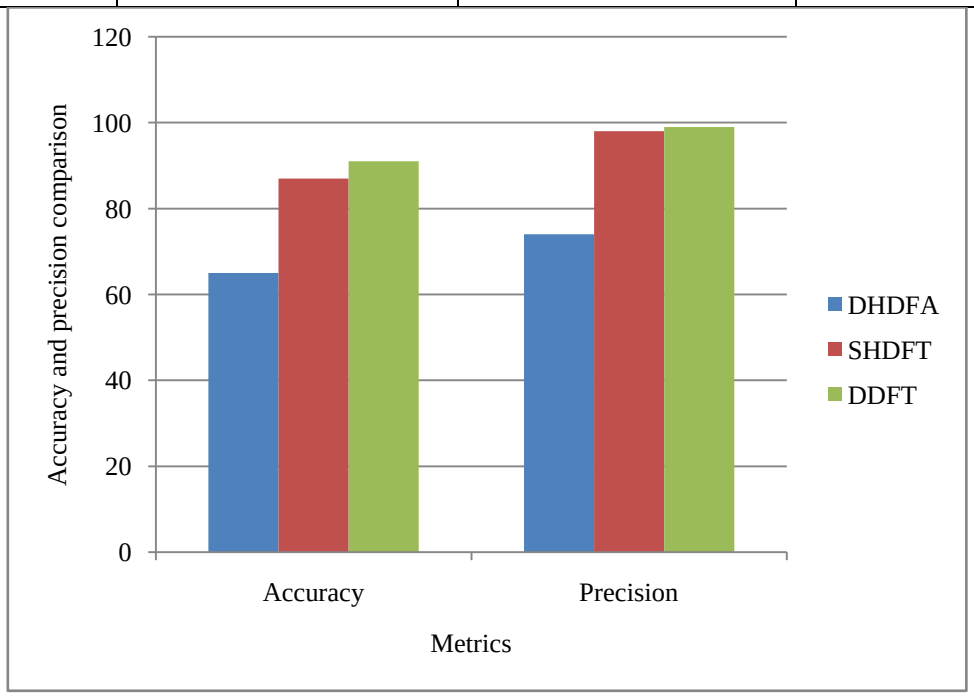

Figure 3. Accuracy and precision comparison

In figure 3, comparison analysis of the proposed method and the existing method namely DHDFA and SHDFT is given. From this analysis it is proved that the proposed shows better

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 4276 - 4283
Received 05 March 2021; Accepted 01 April 2021.

performance than the existing technique. Proposed DDFT attains 26% increased accuracy than DHDFA and 4% increased accuracy than SHDFT. In terms of precision, 25% increased precision than the existing DHDFA and 1% increased precision than SHDFT.

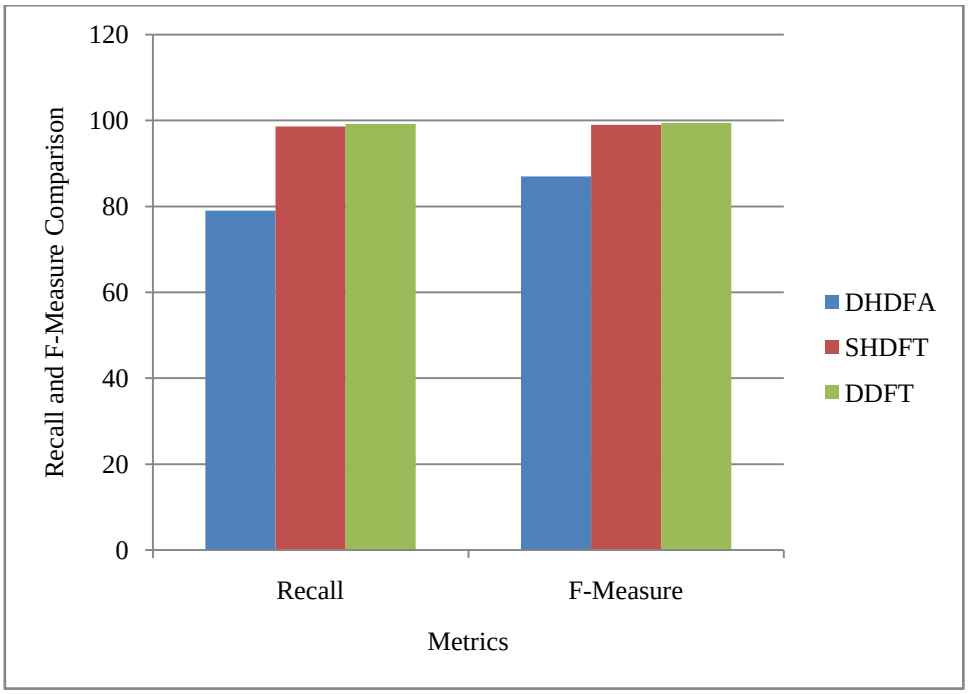

Figure 4. Recall and F-Measure comparison

In figure 4, comparison analysis of the proposed method and the existing method namely DHDFA and SHDFT is given. From this analysis it is proved that the proposed shows better performance than the existing technique. Proposed DDFT attains shows 0.6% increased recall than DHDFA and 10.2% increased recall than SHDFT. In terms of F-Measure 12.4% increased F-Measure than DHDFA and 0.4% increased F-Measure than SHDFT.

Conclusion

In this research work, before data fusion, duplication avoidance is done in order to reduce the computational burden of handling repeated data's. In this research work, initially context based feature extraction is done for extracting the useful features and repeated features. After feature extraction, duplication avoidance is done with the concern of similar concept. After duplication avoidance data fusion is done using hierarchical data fusion technique as done in previous work SHDFT. Finally performance of the data fusion technique is tested using Improved Convolutional neural network. The overall evaluation of the research work is done in the matlab against existing work in terms of accuracy, precision, recall and f-measure.

REFERENCE

[1] Wanka, A., & Gallistl, V. (2018). Doing age in a digitized world—a material praxeology of aging with technology. *Frontiers in Sociology*, *3*, 6.

[2] Papadokostaki, K., Mastorakis, G., Panagiotakis, S., Mavromoustakis, C. X., Dobre, C., & Batalla, J. M. (2017). Handling big data in the era of internet of things (IoT). In *Advances in Mobile Cloud Computing and Big Data in the 5G Era* (pp. 3-22). Springer, Cham.

[3] Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods*, *21*(3), 525-547.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 4276 - 4283
Received 05 March 2021; Accepted 01 April 2021.

[4] Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. *International journal of information management*, *39*, 156-168.

[5] Cambria, E., Poria, S., Hazarika, D., & Kwok, K. (2018, April). SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

[6] Alam, F., Mehmood, R., Katib, I., Albogami, N. N., & Albeshri, A. (2017). Data fusion and IoT for smart ubiquitous environments: A survey. *IEEE Access*, *5*, 9533-9554.

[7] Yokoya, N., Grohnfeldt, C., & Chanussot, J. (2017). Hyperspectral and multispectral data fusion: A comparative review of the recent literature. *IEEE Geoscience and Remote Sensing Magazine*, *5*(2), 29-56.

[8] Del Vecchio, P., Di Minin, A., Petruzzelli, A. M., Panniello, U., & Pirri, S. (2018). Big data for open innovation in SMEs and large corporations: Trends, opportunities, and challenges. *Creativity and Innovation Management*, *27*(1), 6-22.

[9] Gravina, R., Alinia, P., Ghasemzadeh, H., & Fortino, G. (2017). Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges. *Information Fusion*, *35*, 68-80.

[10] Himeur, Y., Alsalemi, A., Al-Kababji, A., Bensaali, F., & Amira, A. (2020). Data fusion strategies for energy efficiency in buildings: Overview, challenges and novel orientations. *Information Fusion*, *64*, 99-120.

[11] Mohanrao, M., & Karthik, S. (2017, September). Intelligent data mining principles with privacy preserving procedures. In *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)* (pp. 2513-2516). IEEE.

[12] Farrow, E., Moore, J., & Gašević, D. (2019, March). Analysing discussion forum data: a replication study avoiding data contamination. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 170-179).

[13] Sivaram, M., Batri, K., Mohammed, A. S., Porkodi, V., & Kousik, N. V. (2020). Data fusion using Tabu crossover genetic algorithm in information retrieval. *Journal of Intelligent & Fuzzy Systems*, (Preprint), 1-10.

[14] Losada, D. E., Parapar, J., & Barreiro, A. (2018). A rank fusion approach based on score distributions for prioritizing relevance assessments in information retrieval evaluation. *Information Fusion*, *39*, 56-71.

[15] Majumder, S., & Pratihar, D. K. (2018). Multi-sensors data fusion through fuzzy clustering and predictive tools. *Expert Systems with Applications*, *107*, 165-172.

[16] Ahmad, K., Pogorelov, K., Riegler, M., Conci, N., & Halvorsen, P. (2017, September). CNN and GAN Based Satellite and Social Media Data Fusion for Disaster Detection. In *MediaEval*.

[17] Simonetta, F., Ntalampiras, S., & Avanzini, F. (2019, January). Multimodal music information processing and retrieval: survey and future challenges. In *2019 International Workshop on Multilayer Music Representation and Processing (MMRP)* (pp. 10-18). IEEE.

[18] Chen, B., Huang, B., & Xu, B. (2017). Multi-source remotely sensed data fusion for improving land cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *124*, 27-39.

[19] Zhang, L., Gao, H., Wen, J., Li, S., & Liu, Q. (2017). A deep learning-based recognition method for degradation monitoring of ball screw with multi-sensor data fusion. *Microelectronics Reliability*, *75*, 215-222.