Vilnius university
Institute of Data Science and Digital Technologies
L I T H U A N I A

INFORMATICS ENGINEERING (T007)

DATA FUSION FOR BETTER DECISION MAKING

Jaroslava Arsenjeva

October 2019

Technical Report DMSTI-DS-T007-19-12

VU Institute of Data Science and Digital Technologies, Akademijos str. 4, Vilnius
LT-08412, Lithuania
www.mii.lt

Abstract

Data fusion is the process of combining data from multiple sensors into one framework to provide better data analysis and improve decision making. It is a rapidly evolving trend among other ones such as IoT, Industry 4.0 and Big Data. And like every method, data fusion has its own difficulties. From having to deal with heterogeneous data and noise, different sampling rates and improper weight assignment for the raw data to receiving inferences that are contradictory – all of these are issues that can be assessed with the help of Kalman filter or Bayesian/ Demster-Shafer methods or some other well known algorithms.

But there are other complexities that have been known for more than 20 years and a lot of attempts were made to solve them however the ideal solution is yet to be found. The issues arising in data fusion implementation include the fact, that there is no ideal algorithm for any situation, a faulty sensor cannot be "replaced" by a complex framework, there is not enough sufficient training data due to changing environmental conditions and the value of output is hardly quantifiable. These problems possess a great interest in multiple fields and some solutions have been proposed however the determination of the best way to tackle them is yet to be determined.

Keywords: data fusion, data mining, data analysis, Kalman filter

Contents

1 Introduction

According to the official definition, data fusion is “the process of getting data from multiple sources in order to build more sophisticated models and understand more about a project. It often means getting combined data on a single subject and combining it for central analysis” [1].

Data fusion was first mentioned in 1985 by Joint Directors Laboratories and was implemented in a 5level model. The initial description of data fusion was “A process dealing with the association, correlation, and combination of data and information from single and multiple sources to achieve refined position and identity estimates, and complete and timely assessments of situations and threats, and their significance. The process is characterized by continuous refinements of its estimates and assessments, and the evaluation of the need for additional sources, or modification of the process itself, to achieve improved results”. The model proposed is still used for data visualization however it has its flaws and was several times modified [2].

With the recent advancement of IoT, Industry 4.0 and Big Data, data fusion has gained some popularity among information scientists. Being mainly a militaristic methodology in the beginning, data fusion is now being applied in many fields such as automotive, agricultural, medical and others. The main tasks data fusion tackles are target acquisition [3], health diagnosis estimation [4], image combining [5], risk analysis [6], fingerprint and/or object recognition [7] [8] [9], sensor information processing in wireless networks [10] [11] [12].

2 Data fusion: structure and problems

The are several data fusion advantages over using a single sensor. If identical sensors are used that would provide a statistical advantage in gaining additional data from another source and would result in a more accurate inference. This can be compared to simply having more observations from a single sensor. Adding another sensor (radar) can tremendously help in target detection/ acquisition because the position could be now calculated using triangulation and two observation angles would minimize the blind sport in collected information making a better performance than each sensor would individually. Further motivation to include data fusion into any framework is the continuous reliability of the decision in case of a technical sensor failure and faster data acquisition since several sensors are working simultaneously [13] [14].

There are generally three levels where the fusion can be performed: the direct raw data fusion, the feature vector representing the data fusion and inference fusion after data has been processes and vectors combined.

If the sensors are mining homogeneous data (like two cameras or sound sensors) the initially collected data can be fused into one bigger data set however the dispersion of information must be taken into account for data further normalization. The easiest intuitive way to combine data from homogeneous sensors is to simply take the weighted average however a better method for raw data processing is Kalman Filter. It will take into account the given data and predict the estimate of the observable object. Since Kalman Filter is a recursive algorithm, it can work with noisy data. In case the algorithm is non-linear, the extended Kalman Filter can be used. Although the data is homogeneous there is a need for some adjustments: different sampling rates must be synchronized.

If the obtained data from sensors is not of the same kind, the fusion must happen at feature extraction or inference levels. The feature level calls for significant feature extraction (for example, how an architect can combine other building features to create a new one). Since it is complicated to work with the whole concatenated feature set there arises a need to select the most important features. This way the feature world can be grouped into various regions by feature similarity or, in other words, clustered. The common unsupervised algorithms are the self-organizing map (that not only produces a map of the data samples given but also aids in dimensionality reduction [15]) and other neural network algorithms, K-means clustering (with creating K centroids among data points and assigning each data point to the nearest cluster) or other clusterisation algorithms [14].

Inference fusion level combines all preliminary decisions made and selects the most likely option. For the inference level the commonly used algorithms are Bayesian inference and Dempster-Shafer. The Bayesian theory provides us with the probability of a hypothesis with updating information option and the Demster-Shafer method is a more general form of Bayesian since it can calculate the likelihood for multiple evidences that are not known (therefore the sum of all possible events does not necessarily need to be equal to 1 and the prior state is not necessary) [16].

Although many methods for data fusion are being currently implemented there are still various problems that arise. In [17] multiple problems were listed such as:

- There is no substitute for a good sensor;
- Downstream processing cannot make up for errors (or failures) in upstream processing;
- Sensor fusion can result in poor performance if incorrect information about sensor performance is used;
- There is no such thing as a magic or golden data fusion algorithm;
- There will never be enough training data;
- It is difficult to quantify the value of a data fusion system;

- Fusion is not a static process.

These problems were knows almost 20 years ago and still most of them remain unsolved. If there is no way to accurately observe something, no amount of other sensors can solve this issue. This becomes especially challenging when the observation objects shift from physical target to the human based targets and even the observation phenomena determination is a challenge.

Data processing still must be precise and accurate at every level. Failure to do so cannot be covered by more complex algorithms and techniques, because the data and therefore the decision will be contradictory to the expected result.

The failure in assigning correct accuracy and weights to the sensor data will lead to biased fused vectors and errors in the final estimation. This way the obtained fused decision might be worse than the decision based on the processing of data from the best sensor.

There are many sophisticated algorithms that were stated in the past couple of years but there is no universal method for all of the situations. Every time the fusing methods must be modified to match the current situation.

The not enough data is only thing that can be argued about with the Big Data trend providing tremendous amounts of data however not any data set is ideal for training and pattern recognition due to condition change. A combination of sample data and artificially created data (for example, for galaxy/star detection artificial galaxy/star images can be made) can be the solution.

The evaluation of the fused inference is still based on a probability and data fusion process still has to be continuously updated with new information for better results [13].

2.1 Application examples

Data fusion can be applied in many fields most revolutionary of them being the medical one. In [18] authors propose to compare a part of a medical data stream (measuring body temperature, blood pressure etc.) to the same data stream in the past and tries to find the most similar sequence. This way, if any person has monitoring sensors present it would be easy to predict a critical situation.

In [19] the patients are proposed to wear body sensors for condition monitoring. Since these sensors have the capability to measure the position and the acceleration/ velocity of the person it is possible to predicts his actions: if he is doing something physically demanding, sleeping; it is even possible to distinguish if the person fell on the ground and need assistance.

In [20] a way to estimate the position of a wheelchair is proposed. In [21] kernel Random Forest (see Figure 1) is used for heart disease prediction. Daily activity data is obtained from a collection of sensors which measure number of steps taken in a day,

calories burned, cholesterol level, sleeping hours, sleep quality, calorie intake and others. The fused data is later used to form a unified classifier. The data is store in an online cloud and transmitted over 5G. For classification the random forest method is used which proved to be the most accurate in [21] with tree depth 15 and accuracy 98%.

In [28] an overview of other military and non-military applications of data fusion can be found (Table 1).

Table 1. Data fusion application fields [28]

Application specifics	**Inference obtained**	**Primary data observed**	**Surveillance volume**	**Sensor platform**
Ocean surveillance	Detection; tracking; identification of targets	EM signal; Acoustic signals; Nuclear related	Hundreds of nautical miles; air/subsurface	Ships; aircraft; submarines (ocean or ground-based)
Air defense	Detection; tracking; identification of aircraft	EM radiation	Hundreds of km if strategic; km if tactical	Ground based; Aircraft; Ships
Target acquisition	Detection and identification of ground targets	EM radiation	Battlefield (10 to 100 km)	Ground-based; Aircraft
Strategic warning and defense	Detection of ballistic missiles and warheads	EM radiation; Nuclear related	Global	Satellite; Aircraft; Ground-based
Condition maintenance	Detection of faults; Recommendation for maintenance	EM signal; Acoustic signals; Magnetic signals; Temperature; X-rays; Vibration	Microscopic to decimeter	Ships; Aircraft (any industry)
Robotics	Location; Obstacle identification; Manipulation of objects need	TV signal; Acoustic signals; EM signals; X-rays	Microscopic to decimeter	Robot body
Environmental monitoring	Identification and location of natural phenomena	SAR; Seismic; EM radiation; Core samples; Bio data	From km to hundreds of km	Satellites; Aircraft; Ground-based

Require: Dataset D in the form of (X, Y) pairs, number of trees P, $f \in \{1, \ldots, P\}$, $\tau_n \in \{1, \ldots, n\}$, $x \in [0, 1]^P$
Ensure: Prediction of the random forest at x

for each $j \in M$ **do**
  Choose τ_n points from D
  For all $l \in \tau_n$, set $\rho_l = \phi$, $\rho_0 = [0, 1]^P$ ▷ root partition
  Set $\eta_v = 1$, $\psi = 0$ ▷ η: number of vertices; ψ: level
  while $\eta_v < \tau_n$ **do**
    if $\psi \neq \phi$ **then**
      $\rho \leftarrow$ point x
      if $\sum \rho = 1$ **then**
        $\rho_\psi \leftarrow \rho_\psi \bigcup \rho$
      else
        Generate and split the set ρ into ρ_A, ρ_B
        $\rho_{\psi+1} \leftarrow \rho_{\psi+1} \bigcup \rho_A \bigcup \rho_B$ ▷ ρ updated as a result of split into ρ_A and ρ_B
        $\eta_v \leftarrow \eta_v + 1$
      end if
    else
      $\psi \leftarrow \psi + 1$
    end if
  end while
  Compute $f(x, \Delta_j, D)$ for x ▷ local prediction for x
end for
Compute $f_{Pn}(x, \Delta_1, \ldots, \Delta_P, D)$ ▷ global prediction for x
return

Figure 1. Pseudocode for Random Forest.

2.2 Most common algorithms

As was mentioned before, data fusion can be divided into 3 levels: raw data fusion, feature fusion and inference fusion. In case of homogeneous data, such algorithms can be used:

- k-Nearest neighbors (kNN, a supervised machine learning algorithm used for classification and/or regression problems);
- Probabilistic Data Association Filter (PDAF, used for plot association in a target tracking algorithm);
- Joint Probabilistic Data Association Filter (JPDAF, similar to PDAF but works with multiple targets).

If the data is heterogeneous, each data set must form a feature vector, which could be later combined with the help of:

- Maximum Likelihood Estimator (MLE, a method of estimating the parameters of a statistical model given observations, by finding the parameter values that maximize the likelihood of making the observations given the parameters) [22];
- Maximum A Posteriori Estimation (MAP, which tackles the same problem as MLE but needs less data for producing the same results) [23];
- Kalman Filter (a tool that based on a series of observations produces a prediction about the future state of an object in a linear system with Gaussian noise);

- Particle Filter (similar to Kalman filter, but more likely to produce better results in a non-linear system) [24];
- Self-organziing map (SOM, type of neural network, that does unsupervised learning to produce a low-dimensional discrete representation of the training samples in a map form) [25].

After the common vectors are extracted it is necessary to provide inferences for each vector or their combination and come to a final decision. In order to deal with contradictory inferences at the inference fusion level such methods are used:

- Bayes theorem (a method used for calculating conditional probabilities) [26];
- Demster-Shafer theory (A generalization of the Bayesian theory. Whereas the Bayesian theory requires probabilities for each question of interest, belief functions allow us to base degrees of belief for one question on probabilities for a related question) [27].

3 Conclusions

Unfortunately, there is still space for improvement in the data fusion field. Although fusing algorithms become more complex and in many fields sensor monitoring can lift the workload from humans the final decision in most cases still must be made by a human user due to uncertainties in inference fusing process.

4 References

1. Article on Data Fusion, weblink: https://www.techopedia.com/definition/32735/data-fusion
2. Liggins, Martin E.; Hall, David L.; Llinas, James (2008). Multisensor Data Fusion, Second Edition: Theory and Practice (Multisensor Data Fusion)
3. Shen X., P.K. Varshney, Sensor selection based on generalized information gain for target tracking in large sensor networks, IEEE Trans. Signal Process., 62 (2) (2014), pp. 363-375
4. Dong M., He D., Hidden semi-markov model-based methodology for multi-sensor equipment health diagnosis and prognosis, Eur. J. Oper. Res., 178 (3) (2007), pp. 858-878
5. Yang F., Wei H., Fusion of infrared polarization and intensity images using support value transform and fuzzy combination rules, Infrared Phys. Technol., 60 (2013), pp. 235-243
6. Zhang L., Wu X., Zhu H., S.M. AbouRizk, Perceiving safety risk of buildings adjacent to tunneling excavation: an information fusion approach, Autom. Constr., 73 (2017), pp. 88-101
7. D. Peralta, I. Triguero, S. García, Y. Saeys, J.M. Benitez, F. Herrera, Distributed incremental fingerprint identification with reduced database penetration rate using a hierarchical classification based on feature fusion and selection, Knowl. Based Syst., 126 (2017), pp. 91-103
8. D. Peralta, I. Triguero, S. García, F. Herrera, J.M. Benitez, DPD-DFF: a dual phase distributed scheme with double fingerprint fusion for fast and accurate identification in large databases, Inf. Fusion, 32 (2016), pp. 40-51
9. G. Fortino, S. Galzarano, R. Gravina, Li W., A framework for collaborative computing and multi-sensor data fusion in body sensor networks, Inf. Fusion, 22 (2015), pp. 50-70
10. Zhang Z., Zhang W., Chao H.-C., Lai C.-F., Toward belief function-based cooperative sensing for interference resistant industrial wireless sensor networks, IEEE Trans. Ind. Inf., 12 (6) (2016), pp. 2115-2126
11. Zhang Z., Hao Z., S. Zeadally, Zhang J., Han B., Chao H.-C., Multiple attributes decision fusion for wireless sensor networks based on intuitionistic fuzzy set, IEEE Access, 5 (2017), pp. 12798-12809
12. Duan Y., Fu X., Li W., Zhang Y., G. Fortino, Evolution of scale-free wireless sensor networks with feature of small-world networks, Complexity, 2017 (3) (2017), pp. 1-15
13. Martin Liggins II, David Hall, James Llinas, Handbook of Multisensor Data Fusion: Theory and Practice, Second Edition, (2017)
14. R. C. Luo, Chih-Chen Yih and Kuo Lan Su, "Multisensor fusion and integration: approaches, applications, and future research directions," in IEEE Sensors Journal, vol. 2, no. 2, pp. 107-119, (2002)

15. Article by A. Ralhan “Self Organizing Maps”, weblink: https://towardsdatascience.com/self-organizing-maps-ff5853a118d4

16. James C. Hoffman, Robin R. Murphy, "Comparison of Bayesian and Dempster-Shafer theory for sensing: a practitioner's approach," Proc. SPIE 2032, Neural and Stochastic Methods in Image and Signal Processing II, (29 October 1993)

17. Hall, David & Steinberg, Alan. (2001). Dirty Secrets in Multisensor Data Fusion

18. Jolita Bernataviciene, Gintautas Dzemyda, Gediminas Bazilevicius, Viktor Medvedev, Virginijus Marcinkevicius, Povilas Treigys, “Method for Visual Detection of Similarities in Medical Streaming Data”, in INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL ISSN 1841-9836, 10(1):8-21, February, (2015)

19. Rachel C. King, Emma Villeneuve, Ruth J. White, R. Simon Sherratt, William Holderbaum, William S. Harwin, “Application of data fusion techniques and technologies for wearable health monitoring”, in Medical Engineering & Physics, vol. 42, pp. 1-12, (2017)

20. Nada, D., Bousbia-Salah, M. & Bettayeb, M. Int. J. Autom. Comput. (2018) 15: 207. https://doi.org/10.1007/s11633-017-1065-z

21. Muhammad Muzammal, Romana Talat, Ali Hassan Sodhro, Sandeep Pirbhulal, “A multi-sensor data fusion enabled ensemble approach for medical data from body sensor networks”, in Information Fusion, vol. 52, pp. 155 – 164, (2020)

22. Article by Jonathan Balanban, “A Gentle Introduction to Maximum Likelihood Estimation”, weblink: https://towardsdatascience.com/a-gentle-introduction-to-maximum-likelihood-estimation-9fbff27ea12f

23. Article by Shota Horii, “A Gentle Introduction to Maximum Likelihood Estimation and Maximum A Posteriori Estimation”, weblink: https://towardsdatascience.com/a-gentle-introduction-to-maximum-likelihood-estimation-and-maximum-a-posteriori-estimation-d7c318f9d22d

24. Article by S. Srini, “Particle Filter : A hero in the world of Non-Linearity and Non-Gaussian”, weblink: https://towardsdatascience.com/particle-filter-a-hero-in-the-world-of-non-linearity-and-non-gaussian-6d8947f4a3dc

25. T. Kohonen, "The self-organizing map," in Proceedings of the IEEE, vol. 78, no. 9, pp. 1464-1480, Sept. 1990

26. Artcile on Bayes theorem, weblink: https://stanford.library.sydney.edu.au/entries/bayes-theorem/

27. Dempster, A.P. (1968). A generalization of Bayesian inference. Journal of the Royal Statistical Society, Series B 30 205-247

28. David l. Hall, James LLinas, “An introduction to multisensor data fusion”, proceedings of the IEEE, vol. 85, no. 1, January 1997.