Distributed Decision Fusion for Large Scale IoT-Ecosystem

15th IEEE International Symposium
on
Embedded Multicore/Many-core Systems-on-Chip (MCSoC-2022)
Universiti Sains Malaysia, Penang, Malaysia

presented by

Ashwin Raut

Indian Institute of Information Technology, Allahabad, India

December 15, 2022

Distributed Decision Fusion for Large Scale IoT-Ecosystem

15th IEEE International Symposium
on
Embedded Multicore/Many-core Systems-on-Chip (MCSoC-2022)
Universiti Sains Malaysia, Penang, Malaysia

presented by

Ashwin Raut

Indian Institute of Information Technology, Allahabad, India

December 15, 2022

Introduction

- Internet of Things (IoT) is an emerging paradigm that has improved the traditional standard of living to smart and high-tech lifestyles of humans worldwide through great connectivity of devices, big data, and analytics.
- These applications have generated an enormous amount of data that learn new insights and gain information, which can further help to improve the performance of end-user and IoT applications.
- In the IoT paradigm, data fusion plays an important role in making information more intelligent, decisive, sensible, and precise.
- Data fusion techniques combine data from heterogeneous IoT devices and related information from associated data storage to achieve improved accuracies and more specific inferences could be achieved by using a single IoT ecosystem.
- There are multiple IoT applications built on several rules and protocols, but when it comes to building a large IoT ecosystem, working together is a very challenging and non-trivial task.

Contd...

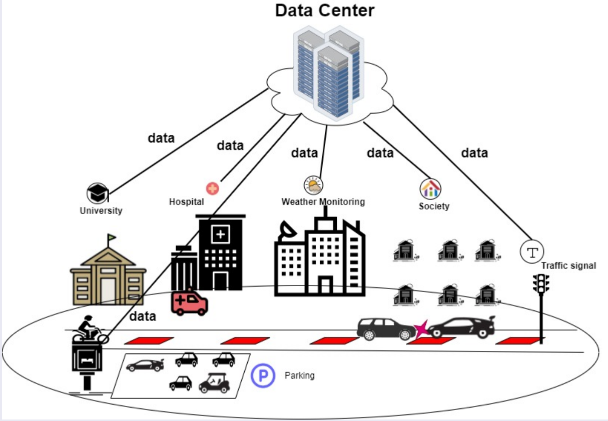

Figure: Data from small IoT ecosystem

Contd...

Literature Review

Title	Problem Addressed	Achievements
Data Fusion and IoT for Smart Ubiquitous Environments: A Survey Alam et al. (2017)	Proposed and implemented a solution for ingesting, analyzing and correlating heterogeneous data streams.	The opportunities and challenges for each of the mathematical methods and environments are discussed.
Real-time probabilistic data fusion for large-scale IoT applications Akbar et al. (2018)	Proposed a two-layer architecture to analyze IoT data.	Demonstrated the solution on real-world use-case in the domain of intelligent transportation system (ITS) and analyzed tra c, weather and social media data streams from Madrid city in order to predict the probability of congestion in real-time with an F-measure accuracy of 80%.

Contd...

Processing IoT Data:From Cloud to Fog—It's Time to Be Down to Earth Pramanik et al. (2018)	Reviewed the usage of cloud computing for centralized data processing.	Discussed important challenges. in centralized data processing.
A survey of decision fusion and feature fusion strategies for pattern classification Mangai et al. (2010)	Presents a review of the di erent techniques and algorithms used in decision fusion and feature fusion strategies, for the task of pattern classification.	A novel framework has been proposed by us, combining both the concepts of decision fusion and feature fusion to increase the performance of classification.

Definition

- The large IoT ecosystem L is connected with small ecosystems $M = \{M_1; M_2; ::::M_n\}$. Every small ecosystem M is collecting data from their attached sensors S_n, where M has a decision-making capability.
- If any inference problem has occurred at L; then L solve the problem by centralize data processing.But centralized data processing mechanism faced following problem which are determine from literature:
  - Heterogeneity of data sources
  - Unnecessary data processing
  - Large computation time
  - Uncertainty in decision making
  - Data biasness in decision making
- The data is collected from various small ecosystems hospital, university, weather monitoring department, society, tra c department, etc, which forms the large ecosystem environment L.

Methodology

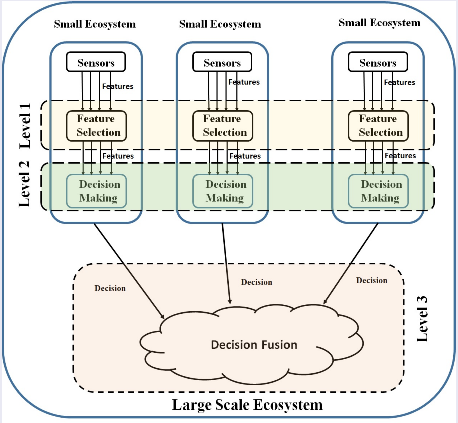

Figure: Process flow of distributed decision fusion architecture

First Level: Feature Selection

- The primary objectives of the feature selection mechanism:
  - reduce computation time,
  - maximize learning performance,
  - minimize the number of features of the decision-making classifier.
- The following formula is used to find out the Pearson correlation for feature selection:

$$C_r(f_1; f_2) = \frac{\sum (f_1 - \mu_{f_1})(f_2 - \mu_{f_2})}{\sqrt{\sum (f_1 - \mu_{f_1})^2 \sum (f_2 - \mu_{f_2})^2}} = \frac{E(f_1; f_2)}{\sigma_{f_1} \sigma_{f_2}} \quad (1)$$

  where $C_r(f_1; f_2)$ is the Pearson Correlation Coefficient, $E(f_1; f_2)$ is the cross-correlation between feature f_1 and f_2, and σ_{f_1}, σ_{f_2} are the variances of the respective features.
- The closer the value of $C_r(f_1; f_2)$ to 1, the stronger the correlation between the two features. For the proposed Correlation Analysis, weak relation: 0 to 0.19, weak: 0.2 to 0.39, moderate: 0.4 to 0.59, strong: 0.6 to 0.79, and very strong: 0.8 to 1[18].

Second Level : Decision Making

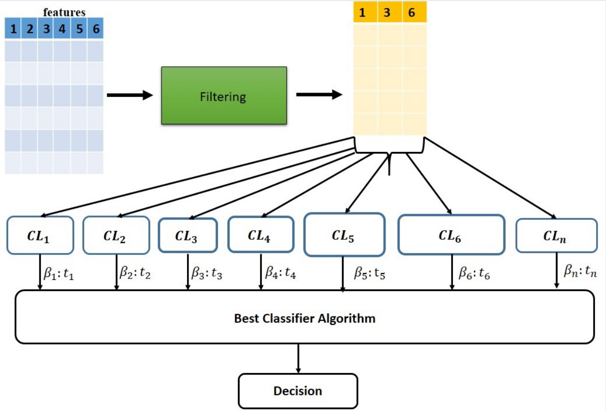

Figure: Decision making of small IoT ecosystem

Third Level: Decision Fusion

- At this level, the proposed distributed decision fusion approach collects the decision from each small IoT ecosystem and processes using ensemble methods like Majority voting, Weighted majority voting, and distributed Naive Bayes, according to the di erent situations.
- *Majority Voting:* In Majority voting, the ensemble chooses a class when:
  - Unanimous voting
  - Simple majority
  - Plurality voting
- The ensemble decision for the Majority voting can be described as follows: choose class $_c$ if,

$$\sum_{q=1} d_{q;c} = \max_{j=1}^{c} \sum_{q=1} d_{q;j} \tag{2}$$

where, the individual decision of the q^{th} classifier of M^{th} small ecosystem is defined as $d_{q;c}$ 2 0,1 where 0 indicates: the decision is missing, and 1 indicates the decision is collected at a centralized level, q = f1,2,....., Zg (number classifier) and c =f1,2....Kg (number of decision class), $_c$ = f $_1$; $_2$; :::; $_K$g is the set of decision class labels.

Contd...

Table: Di erent ensemble methods with conditions, where M = Small Ecosystem, C = Class Label

Ensemble Methods	M_1	M_2	M_3	M_4	M_n
Unanimous majority voting	C_1	C_1	C_1	C_1	C_1
Simple majority	C_1	C_1	C_1	C_1	C_2
Plurality voting	C_1	C_1	C_2	C_3	C_4
Weighted majority voting	C_1	C_1	C_2	C_2	C_3

Contd...

- *Weighted Majority Voting:* In Weighted majority voting the decision class label is evaluated as,

$$g_j(x) = \sum_{q=1} r_q d_{q;c} \quad (3)$$

where r_q is the weight of the decision of q^{th} classifier on M^{th} small ecosystem.

- r_q is calculated using the lookup table which is based on the historical classification accuracy of every individual ecosystem from second level. The q^{th} classifier of the M^{th} system gives the highest accuracy that is assigned as the highest weight to r_q.

Contd...

Steps for decision making in Weighted majority voting is as follows:

- Find the conflicting class labels among k number of class labels.
- Determine the corresponding small ecosystems for each conflicting class label.
- Now using the lookup table (from the second level), we take a summation of the validation accuracy of the small ecosystem for each class label.
- Among all conflicting class labels, the highest summation of the validation accuracy is given as the final decision for output.

Naive Bayes

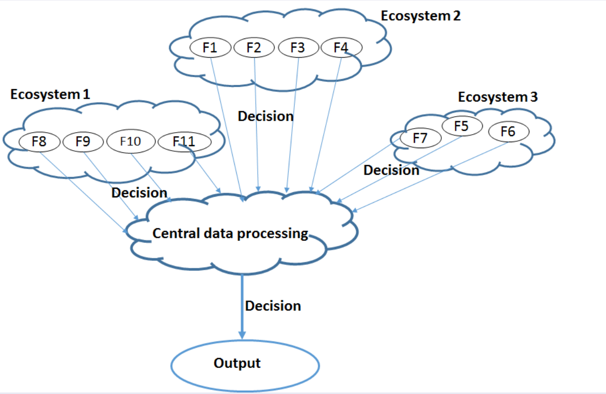

Figure: Bayesian Network of large IoT ecosytem

Contd...

- The proposed distributed Naive Bayes is also an ensemble technique, but before processing a decision fusion in distributed environment, it assumes that the decision of every small ecosystem is mutually independent.
- The conditional independence of the decisions is represented by Naive Bayes as follow:

$$P(q|\omega_c) = P(q_1, q_2, \ldots, q_Z|\omega_c) \; \prod_{i=1}^{Z} P(q_i|\omega_c) \tag{4}$$

where, $P(q_j)$ is the probability that classifier D_j classifies a sample X in class ω_j.

Dataset

- We used the US-Accidents dataset, which covers 49 states in the United States, to validate the proposed hypothesis.
- This dataset currently contains approximately 1.5 million accident records. The dataset includes streaming data from five IoT networks: Weather, POI, Environment, Location and Tra c IoT networks, which have a total of 45 attributes.
- In this, we predict the severity level, which ranges from 1 to 4, where severity level 1 refers to minor accidents and severity level 4 refers to major accidents which can cause death or, in this case, immediate medical assistance is needed.

Contd...

Table: Dataset

IoT Ecosystem	Paramaters
Weather	Weather_Timestamp, Temperature(F), Wind_Chill(F), Humidity(%), Pressure(in), Visibility(mi), Wind_Direction, Wind_Speed(mph), Precipitation(in), Weather_Condition
Environment	Sunrise Sunset, Civil Twilight, Nautical Twilight, Astronomical Twilight
Tra c	Tra c Calming, Tra c Signal, Turning Loop
Point-Of-Interest (POI)	Amenity, Bump, Crossing, Give-way, Junction, No-exit, Railway, Roundabout, Station, Stop
Location	Distance(mi), Street, City, County, State, Zipcode, Timezone, Airport Code

First Level: Feature Selection

Figure: Correlation analysis of Weather Ecosystem

Contd...

Figure: Correlation analysis of Environment Ecosystem

Contd...

Figure: Correlation analysis of Træ Ecosystem

Contd...

Figure: Correlation analysis of POI Ecosystem

Contd...

Figure: Correlation analysis of Location Ecosystem

Contd...

IoT Ecosystem	Input Features	Selected Features
Weather	Weather_Timestamp, Temperature(F), Wind_Chill(F), Humidity(%), Pressure(in), Wind_Direction, Visibility(mi), Wind_Speed(mph), Precipitation(in), Weather_Condition	Temperature(F), Humidity(%), Pressure(in), Visibility(mi), Wind_Direction, Wind_Speed(mph), Precipitation(in)
Environment	Sunrise_Sunset, Civil_Twilight, Nautical_Twilight, Astronomical_Twilight	Sunrise_Sunset, Nautical_Twilight
Traffic	Traffic_Calming, Traffic_Signal, Turning Loop	Traffic_Calming, Traffic_Signal
POI	Amenity, Bump, Crossing, Give-way, Junction, No-exit, Railway, Roundabout, Station, Stop	Amenity, Bump, Crossing, Give-way, Junction, Railway, Roundabout
Location	Distance(mi), Street, City, County, State, Zipcode, Timezone, Airport_Code	Distance(mi), Street, City, County, State, Zipcode, Timezone, Airport_Code

Figure: Overall correlation analysis results of every individual small ecosystem

Second Level : Decision Making

- In second level, set of features from first level is fed as input to di erent classifiers for appropriate decision-making.
- We utilized,
  - Support Vector Machine (SVM),
  - Decision Tree (DT),
  - Gradient Boosting Tree (GBT),
  - Random Forest (RF),
  - XGBoost (XGB),
  - Multi-layer Perceptron (M-l-P) and,
  - Logistics Regression (LR) classifier algorithm to evaluate the decision.

Contd...

Table: Overall results of best classifier.

IoT Ecosystem	Best Classifeir Algorithm	Accuracy	Precision	Recall	F1-score	Time
Weather	Random Forest	88.7188	88.3985	88.9188	88.3846	1.059513
POI	Decision Tree	86.0647	88.0066	86.0647	79.6189	0.453624
Environment	Decision Tree	86.0647	88.0066	86.0647	79.6189	0.443254
Location	Logistics Regression	86.0853	85.4436	86.0853	79.9508	0.479508
Tra c	Decision Tree	86.0647	88.0066	86.0647	79.6189	0.462809

Third Level-Decision Fusion

- We observed that 89.73 % samples are of Unanimous voting, 10.27 % samples are of a Simple majority.
- In Unanimous voting instances, the accuracy is 94.8 %, whereas, in Simple majority voting situations, the accuracy is 10.8 %.
- In our data-set, there are no plurality voting and Weighted majority voting samples. However in future instances related to these conditions might be observed.

Contd...

For Weighted majority voting, let's consider the scenario:

Table: Di erent ensemble methods with conditions, where Unanimous majority voting (UMV), Simple majority (SM), Majority Voting (MV), Weighted majority voting (WME).

Method	Weather	Environment	Tra c	POI	Location
UME	1	1	1	1	1
SM	1	1	1	2	2
MV	1	1	2	3	4
WME	1	1	2	2	3

Contd...

- The conflicting class labels are 1 and 2.
- Small Ecosystems corresponding to class label 1 is Weather, Environment, and Tra c, POI corresponds to class label 2.
- Now, using the lookup table (from second level), summation of validation accuracy of class label 1 : 88.71 + 86.06 = 174.77 and, class label 2 : 86.06 + 86.08 = 172.14.
- Among class labels 1 and 2, summations of validation accuracy of the class label 1 is highest, so the assigned decision class label is 1.

Centralized Data Processing

Table: Result of centralized data processing

Classifier Algorithm	Accuracy	Precision	Recall	F1-score	Time in secs
XGBoost	89.3567	89.6829	89.3567	89.3423	1.635
RF	86.4327	86.8956	86.4327	86.3397	0.832
GBT	84.9123	85.1991	84.9123	84.8916	2.649
DT	81.7544	81.7668	81.7544	81.7399	1.062
LR	66.3158	65.5357	66.3158	65.5947	0.578
MLP	64.3275	63.2479	64.3275	63.337	11.720
SVM	32.1637	67.47	32.1637	23.4249	2.897

Centralized Data Processing

Table: Naive Bayes result for distributed decision fusion (DFF) and centralized data processing (CDP)

Naive Bayes	Accuracy	Precision	Recall	F1-score	Time (sec)
CDP	84.106	84.1	84.1	84.1	0.02
DDF	94.162	94.2	94.2	94.2	0.01

The table shows the accuracy attained for decision fusion of severity level for DFF and CDP using the Naive Bayes approach.

Conclusion

- In large IoT ecosystem, centralized data processing faces several challenges like changing the architecture to ingest data from multiple sources, unnecessary data processing, large computation time, uncertainty in decision making, etc.
- In comparison to distributed decision fusion, the experimental results of centralized data processing show a slight improvement in the context of accuracy and other measured parameters, but in the context of time, the performance is considerably low.
- Using Naive Bayes approach higher accuracy is achieved in distributed decision fusion as compared to Centralized data processing, as in DDF decisions from each local ecosystem are fed into the classifier as input.
- During experiments, we observed that the decision class labels are imbalanced; decision class 2 has 86.06% instances, while decision class 1, 3, and 4 have 2.82%, 6.23%, 4.87% instances, respectively.
- The successful implementation of an advanced modified model in decision fusion field could lead to development of intelligent systems for transport planning in future smart, sustainable cities.

References

[1] S. Kumar, P. Tiwari, and M. Zymbler, "Internet of things is a revolutionary approach for future technology enhancement: a review," Journal of Big data, vol. 6, no. 1, pp. 1–21, 2019.

[2] A. Akbar, G. Kousiouris, H. Pervaiz, J. Sancho, P. Ta-Shma, F. Carrez, and K. Moessner, "Real-time probabilistic data fusion for large-scale iot applications," Ieee Access, vol. 6, pp. 10 015–10 027, 2018.

[3] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classifi- cation and clustering," IEEE Transactions on knowledge and data engineering, vol. 17, no. 4, pp. 491–502, 2005.

[4] F. Alam, R. Mehmood, I. Katib, N. N. Albogami, and A. Albeshri, "Data fusion and IoT for smart ubiquitous environments: A survey," Ieee Access, vol. 5, pp.9533–9554, 2017.

Contd...

[5] P. Ferrer-Cid, J. M. Barcelo-Ordinas, J. Garcia-Vidal, A. Ripoll, and M. Viana, “Multisensor data fusion calibration in iot air pollution platforms,” IEEE Internet of Things Journal, vol. 7, no. 4, pp. 3124–3132, 2020.

[6] N. U. Okafor, Y. Alghorani, and D. T. Delaney, “Improving data quality of low-cost iot sensors in environmental monitoring networks using data fusion and machine learning approach,” ICT Express, vol. 6, no. 3, pp. 220–228, 2020.

[7] M. A. Al-Jarrah, M. A. Yaseen, A. Al-Dweik, O. A. Dobre, and E. Alsusa, “Decision fusion for iot-based wireless sensor networks,” ieee internet of things journal, vol. 7, no. 2, pp. 1313–1326, 2019.

[8] H. E. Kiziloz, “Classifier ensemble methods in feature selection,” Neurocomputing, vol. 419, pp. 97–107, 2021.

Contd...

[9] P. K. D. Pramanik, S. Pal, A. Brahmachari, and P. Choudhury, "Processing IoT daata: From cloud to fog—it's time to be down to earth," in Applications of Security, Mobile, Analytic, and Cloud (SMAC) Technologies for E ective Information Processing and Management. IGI Global, 2018, pp. 124–148.

[10] J. Manyika, "The internet of things," 2015.

[11] S. Moosavi, M. H. Samavatian, S. Parthasarathy, and R. Ramnath, "A coun- trywide tra c accident dataset," arXiv preprint arXiv:1906.05409, 2019.

[12] A. M. Kowshalya, R. Madhumathi, and N. Gopika, "Correlation based feature selection algorithms for varying datasets of di erent dimensionality," Wireless Personal Communications, vol. 108, no. 3, pp. 1977–1993, 2019.

Contd...

[13] U. G. Mangai, S. Samanta, S. Das, and P. R. Chowdhury, “A survey of decision fusion and feature fusion strategies for pattern classification,” IETE Technical review, vol. 27, no. 4, pp. 293–307, 2010.

[14] S. Moosavi, M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and R. Ram- nath, “Accident risk prediction based on heterogeneous sparse data: New dataset and insights,” in Proceedings of the 27th ACM SIGSPATIAL Inter- national Conference on Advances in Geographic Information Systems, 2019, pp. 33–42.

[15] L. Atzori, A. Iera, and G. Morabito, “The Internet of Things: A survey,” Comput. Netw., vol. 54, no. 15, p. 2787–2805, Oct. 2010.

[16]L. I. Kuncheva, J. C. Bezdek, and R. P. W. Duin, “Decision templates for multiple classifier fusion: an experimental comparison,” Pattern Recognition, vol. 34, pp. 299- 314, February 2001.

Contd...

[17] T.K. Ho, J. J. Hull, and S. N. Srihari, “Decision in multiple classifier systems,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 16, pp. 66-75, January 1994.

[18] M. Dash and H. Liu, “Feature selection for classification,” Intelligent Data Analysis, vol. 1, pp. 131-156, August 1997.

[19] S. Manikandan and V. Rajamani, “A Mathematical Approach for Feature Selection and Image Retrieval of Ultra Sound Kidney Image Databases,” European Journal of Scientific Research, vol. 24, pp. 163-171, December 2008.

Thank You!