RocksDB Festival

Failure Prediction

Supported by IITP, StarLab.

Aug 23, 2021

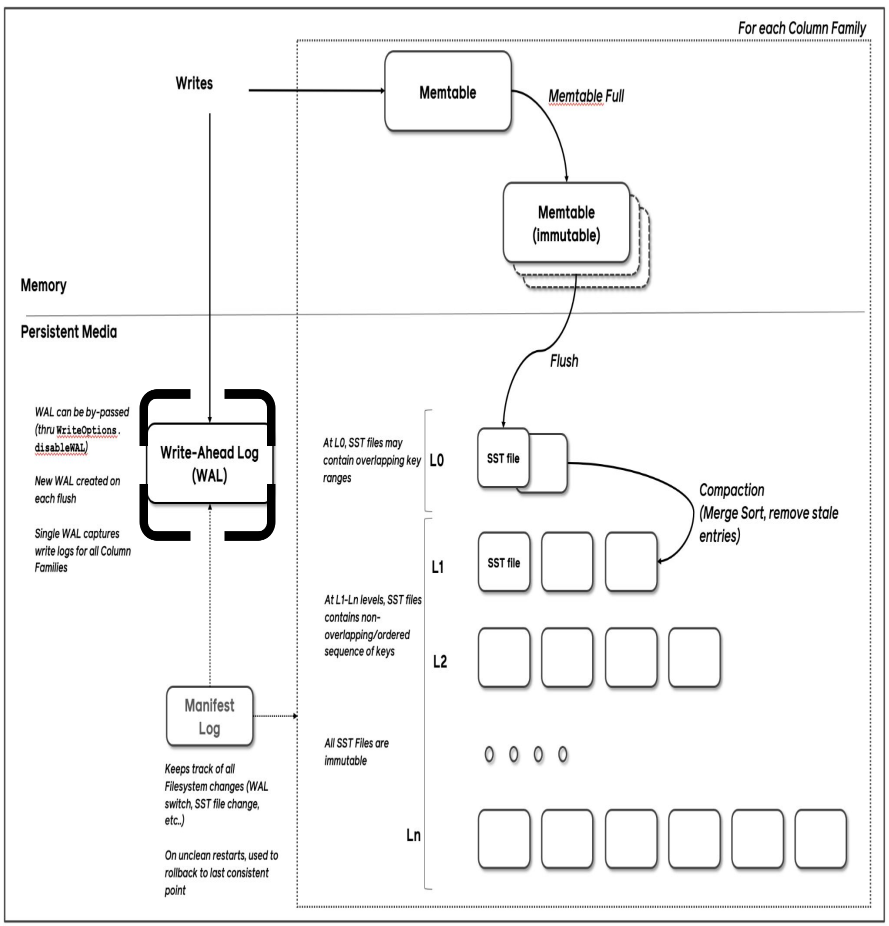
김민준, 이빈
alswnssl0528@naver.com,
32183118@dankook.ac.kr

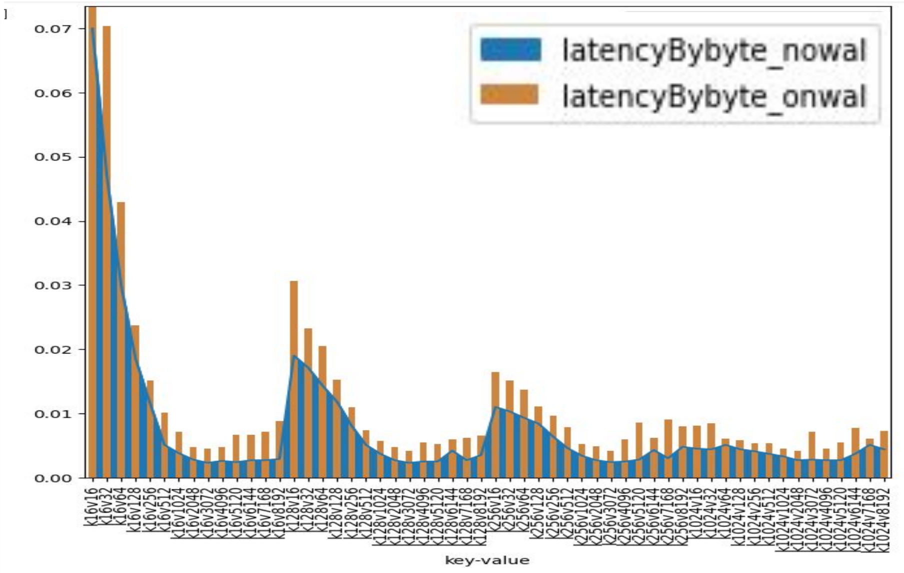
RocksDB Festival

- Contents
  - ✔ Wal performance improvement – failure prediction
    - Introduction
    - Failure prediction (paper reading)
    - Discussion
  - ✔ Next Experiment

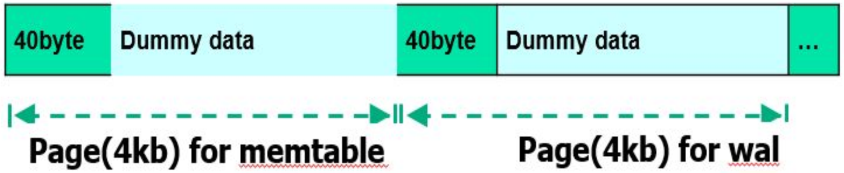
RocksDB Festival

- background
  - ✔ WAL needs to additional write to work
  - ✔ Accordingly, latency increases and space amplification causes





latency increase (latency per byte)



space amplification

RocksDB Festival : Failure prediction

- proposal
  - ✔ The system failures cause **data loss**.
  - ✔ RocksDB solution : **reactive** data protection (recovery)
    - post-failure recovery will not scalable when the storage demand keep increase.
    - recovery process can require unexpected amounts of time and resources.
  - ✔ **Proactive** actions can be taken in advance to improve service reliability.

RocksDB Festival

- proposal
  - ✔ To reduce WAL overhead, instead of always insert WAL in RocksDB, only apply **when system crash predict**
  - ✔ Using failure prediction techniques provides early warnings for potential failures RocksDB
  - ✔ Finally, By reducing WAL, rocksdb has **better performance**

RocksDB Festival : Failure prediction

- Principle
  - ✔ Log analysis by using machine learning algorithms
    - 1. Collect log files
    - 2. Preprocessing log files
    - 3. Machine learning algorithms
    - 4. Analysis

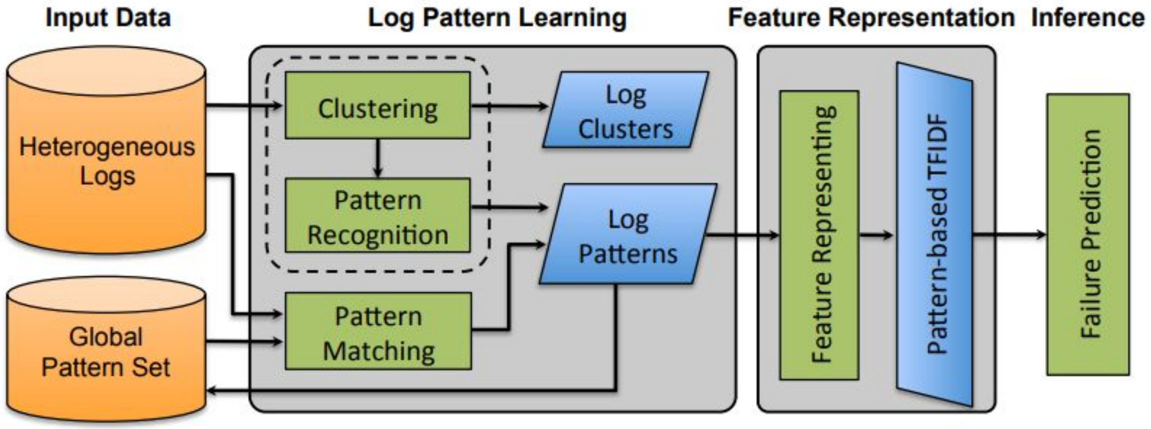


Figure 1. The framework of our solution

example from “Automated IT System Failure Prediction: A Deep Learning Approach”

RocksDB Festival : failure prediction

- A Machine Learning Approach to Database Failure Prediction
  - ✔ 1. Collect log files
    - Data consist of logs that are collected from several different servers **hourly** during 170 days. (38,184 hourly observations in total)
    - Data from log files are collected hourly from **Oracle database systems**.
  - ✔ 2. Preprocessing log files
    - labeled with two classes; normal or abnormal.

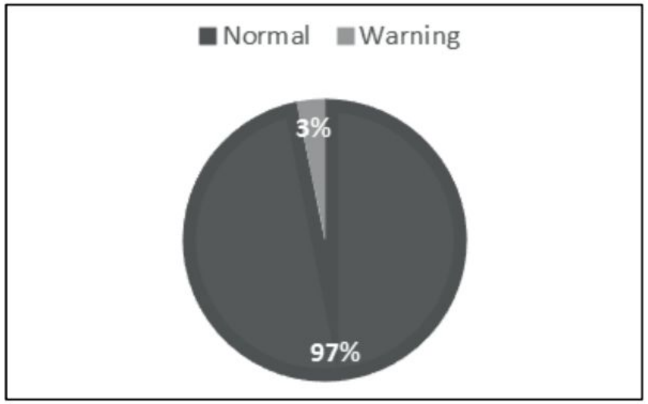


Figure 1: Class Distribution in our dataset

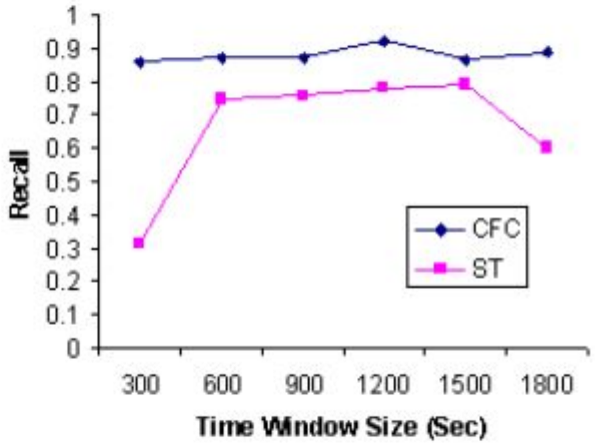
Warnings that lead to failures.
Hourly logs that produce these warnings are labeled as abnormal.

RocksDB Festival : failure prediction

- A Machine Learning Approach to Database Failure Prediction
  - ✔ 3. Machine learning algorithms
    - **Normalization** methods found in preprocessing library of sklearn (scikit-learn) library in python. The values of attributes between 0 and 1.
    - In the **feature selection** part, Shannon's Information Theory is used in order to calculate information gain (IG) values.
  - ✔ 4. Analysis
    - Evaluation metrics : Accuracy, F-measure (F1), Precision, and Recall.
    - The Random Forest algorithm, with a relatively satisfactory Recall (75.7%) and Precision (84.9%) which is visibly higher than the other classifiers.

RocksDB Festival : failure prediction

- System Log Pre-processing to Improve Failure Prediction
  - ✔ Preprocessing log files
    - (1) event **categorization** to uniformly classify system events and identify fatal events;
    - (2) event **filtering** to remove temporal and spatial redundant records, while also preserving necessary failure patterns for failure analysis;
    - (3) **causality-related** filtering to combine correlated events for filtering through Apriori Association Rule Mining.



(b)

With respect to the improvement on failure prediction, recall can be dramatically boosted from 0.3 to 0.7

CFC : our preprocessing methods
ST : existing spatial and temporal filtering method

RocksDB Festival : failure prediction

- Automated IT System Failure Prediction: A Deep Learning Approach
  - ✔ 1. Collect log files
    - Console Logs(Heterogeneous formats)
    - Dataset has been collected from a web server cluster (WSC) and a mailer server cluster (MSC)

Table I
TWO LOG DATASET FROM A WEB SERVER CLUSTER (WSC) AND MAIL SERVER CLUSTER (MSC).

Dataset	WSC	MSC
Collection Periods	[2013-02-24 - 2014-08-29]	[2014-03-22 - 2014-08-29]
# Logs	2,316,081	4,690,583

RocksDB Festival : failure prediction

- Automated IT System Failure Prediction: A Deep Learning Approach
  - ✔ 2. Preprocessing log files
    - Pattern recognition
      - Log tokenize and Time Stamp Standardization:
      - Log Clustering(Log Clustering Tree)
      - Pattern Recognition
      - Pattern Matching

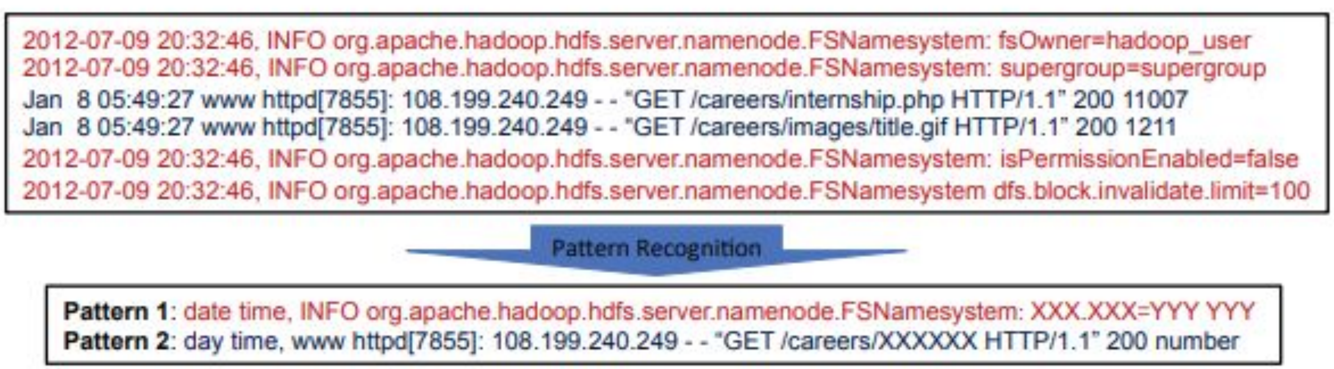


Figure 3. Logs are first clustered into groups and a regular expression based pattern is extracted for each group of logs.

RocksDB Festival : failure prediction

- Automated IT System Failure Prediction: A Deep Learning Approach
  - ✔ Pattern recognition
    - Log tokenize and Time Stamp Standardization:
      - Different types of time stamp formats make the following log clustering and pattern recognition difficult.
      - Detect all the time stamps within the logs and transform them into a standard format (YYYY / MM / DD HH : MM : SS . mss).

2012-07-09 20:32:46,
2012-07-09 20:32:46,
Jan 8 05:49:27
Jan 8 05:49:27

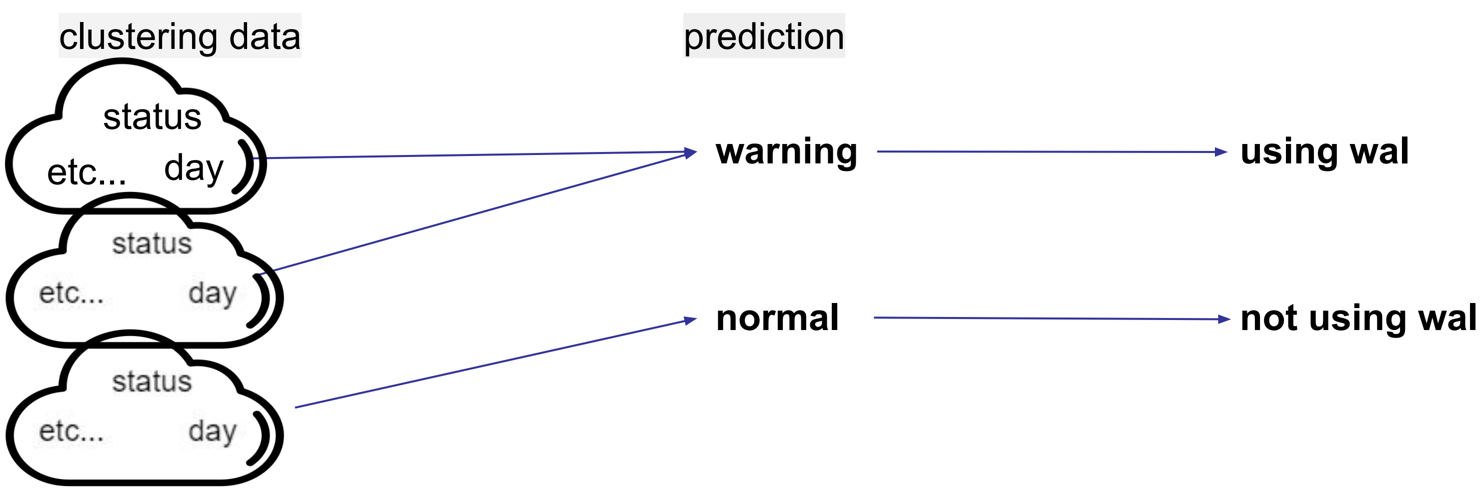
using regular express

Log tokenize →

(YYYY / MM / DD HH : MM : SS . mss)
(2012 / 07 / 09 20 : 32 : 46)
(2012 / 01 / 08 05 : 49 : 27)

RocksDB Festival : failure prediction

- Automated IT System Failure Prediction: A Deep Learning Approach
  - ✔ Pattern recognition
    - Log Clustering(Log Clustering Tree)
      - categorize data purely based on their intrinsic properties and relations.
      - so as to obtain an initial “view” of the data and hierarchical structure.
    - Pattern Recognition
      - obtain more detailed patterns within each cluster.
    - Pattern Matching



RocksDB Festival : failure prediction

- Automated IT System Failure Prediction: A Deep Learning Approach
  - ✔ 3. Machine learning algorithms
    - Feature Representation
      - pattern-based TF-IDF Features Extraction
    - LSTM : Recurrent Neural Network (RNN) architecture designed to improve storing and accessing information compared to classical RNNs

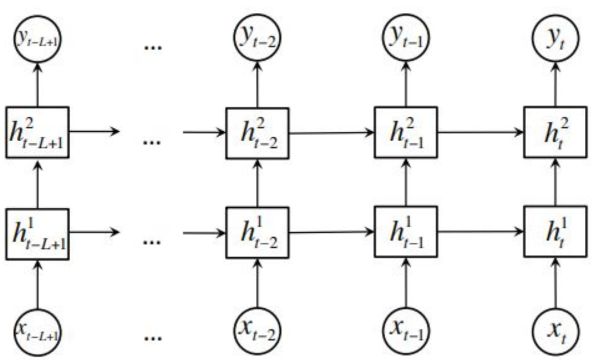


Figure 4. A many-to-many deep recurrent neural network prediction architecture. The rectangles represent the hidden layers, and the circles at the bottom and on the top represent the input layer and output layer, separately. The solid lines represent weighted connections.

Long Short-Term Memory Network

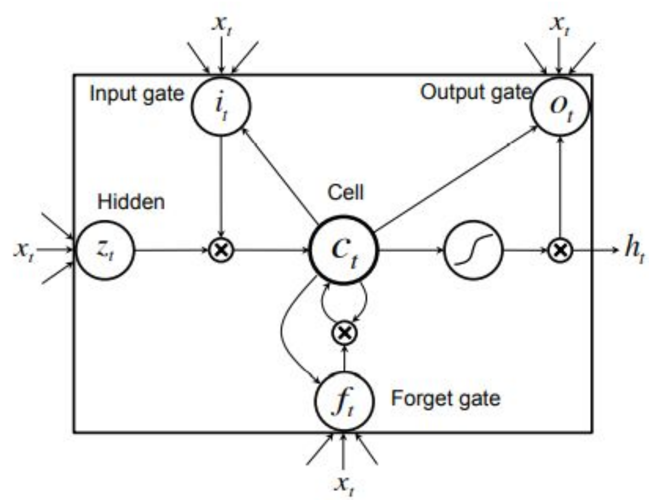


Figure 5. Long short-term memory cell

RocksDB Festival : failure prediction

- Automated IT System Failure Prediction: A Deep Learning Approach
  - ✔ 4. Analysis (three performance metrics)
    - PR-AUC: the Area Under the Curve of Precision-Recall (PR-AUC)
      - Precision = True Positive / (True Positive + False Positive)
      - Recall = True Positive / (All positives)
    - Predictable Interval
      - the time difference between the earliest reported warning and the starting time of the failure
    - Predictable Frequency
      - the fraction of epochs during the predictive period that are predicted and reported as alarms

Table III
PREDICTABLE INTERVAL AND FREQUENCY WITH AT LEAST 70.0% PRECISION. LSTM CAN PREDICT EARLIER ON AVERAGE AND PROVIDE MORE CONFIDENT EARLY ALERTS.

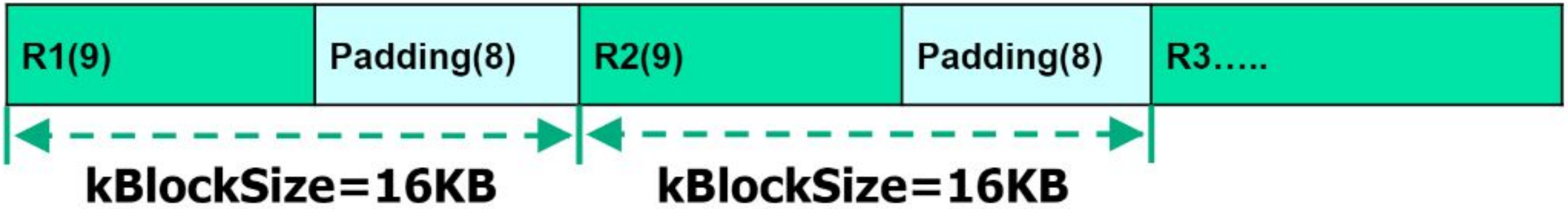
Dataset		SVM	Random Forest	LSTM
WSC	Recall	72.7%	63.6%	90.9%
	Interval (mins)	64.3	45.5	73.0
	Frequency	51.3%	36.1%	66.2%
MSC	Recall	–	60.0%	80.0%
	Interval (mins)	–	22.7	22.0
-	Frequency	–	5.4%	30.4%

RocksDB Festival : failure prediction

- Discussion
  - ✔ To reduce wal overhead, instead of always insert wal in rocksdb, wal apply when system crash predict
  - ✔ We don’t know this technology help increasing performance because prediction overhead may be bigger than wal overhead
  - ✔ So it might be **ineffective**

RocksDB Festival : next experiment

- Next experiment
  - ✔ WAL padding



  - ✔ block_size for packing

block_size -- RocksDB packs user data in blocks. When reading a key-value pair from a table file, an entire block is loaded into memory. Block size is 4KB by default. Each table file contains an index that lists offsets of all blocks. Increasing block_size means that the index contains fewer entries (since there are fewer blocks per file) and is thus smaller. Increasing block_size decreases memory usage and space amplification, but increases read amplification.

Discussion

RocksDB Festival : failure prediction

- Reference

✔ İ. Karakurt, S. Özer, T. Ulusinan and M. C. Ganiz, "A machine learning approach to database failure prediction," 2017 International Conference on Computer Science and Engineering (UBMK), 2017, pp. 1030-1035, doi: 10.1109/UBMK.2017.8093426.

✔ Z. Zheng, Z. Lan, B. H. Park and A. Geist, "System log pre-processing to improve failure prediction," 2009 IEEE/IFIP International Conference on Dependable Systems & Networks, 2009, pp. 572-577, doi: 10.1109/DSN.2009.5270289.

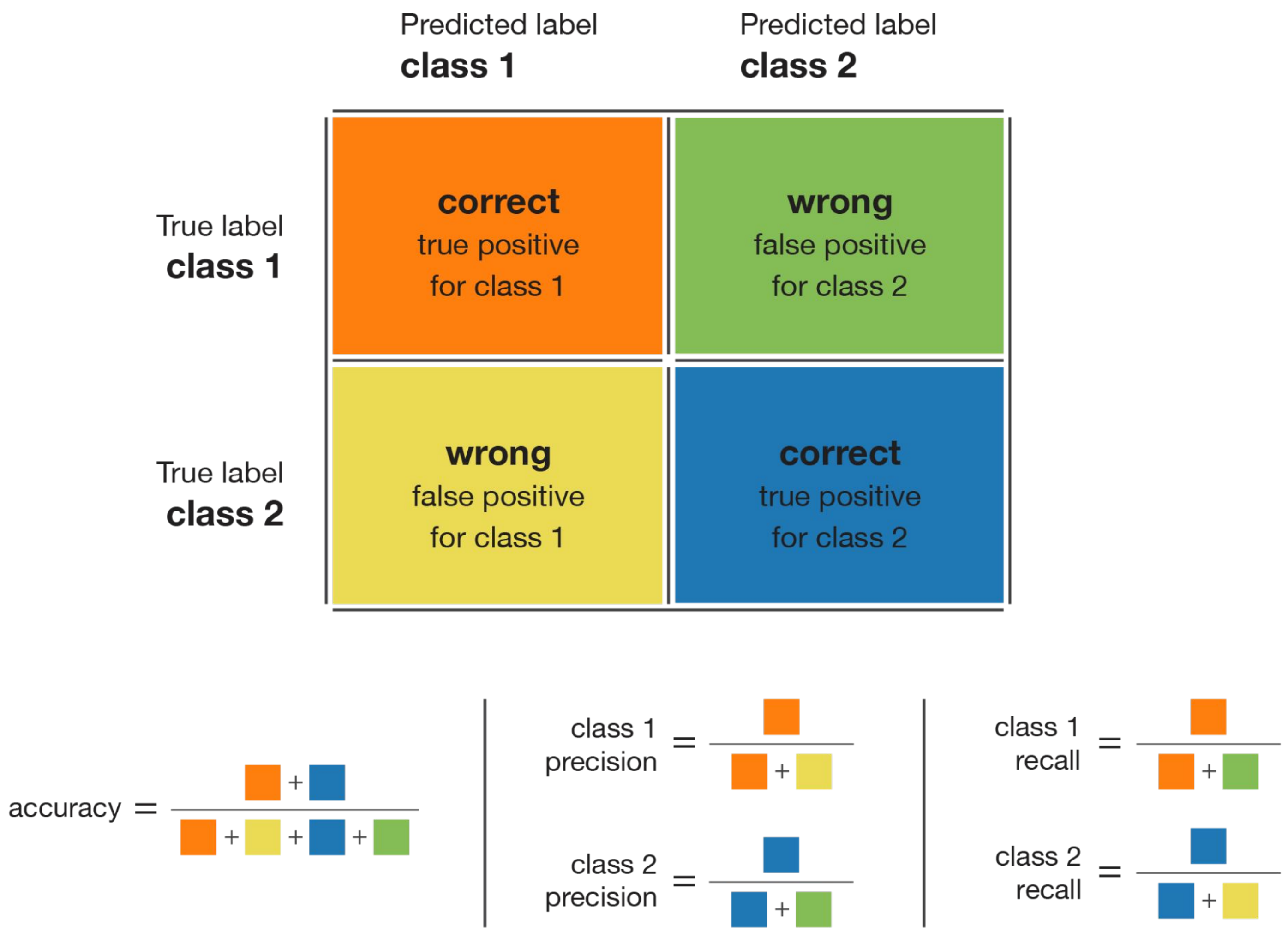
✔ K. Zhang, J. Xu, M. R. Min, G. Jiang, K. Pelechrinis and H. Zhang, "Automated IT system failure prediction: A deep learning approach," 2016 IEEE International Conference on Big Data (Big Data), 2016, pp. 1291-1300, doi: 10.1109/BigData.2016.7840733.

RocksDB Festival : failure prediction

- Reference

  - ✔ Chuan Luo, Pu Zhao, Bo Qiao, Youjiang Wu, Hongyu Zhang, Wei Wu, Weihai Lu, Yingnong Dang, Saravanakumar Rajmohan, Qingwei Lin, and Dongmei Zhang. 2021. NTAM: Neighborhood-Temporal Attention Model for Disk Failure Prediction in Cloud Platforms. In <i>Proceedings of the Web Conference 2021</i> (<i>WWW '21</i>). Association for Computing Machinery, New York, NY, USA, 1181–1191. DOI:https://doi.org/10.1145/3442381.3449867
  - ✔ Z. Qiao, J. Hochstetler, S. Liang, S. Fu, H. Chen and B. Settlemyer, "Incorporate Proactive Data Protection in ZFS Towards Reliable Storage Systems," 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), 2018, pp. 904-911, doi: 10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00-10.
  - ✔ S. K. Yang, "A condition-based failure-prediction and processing-scheme for preventive maintenance," in IEEE Transactions on Reliability, vol. 52, no. 3, pp. 373-383, Sept. 2003, doi: 10.1109/TR.2003.816402.
  - ✔ Lu, Sidi, et al. "Making disk failure predictions smarter!." 18th {USENIX} Conference on File and Storage Technologies ({FAST} 20). 2020.
  - ✔ Fronza, Ilenia, et al. "Failure prediction based on log files using random indexing and support vector machines." Journal of Systems and Software 86.1 (2013): 2-11.
  - ✔ Karel Beneš, Basics in Machine Learning(International Summer School in IT Brno, July 2021)

Appendix : Analysis (performance metrics)



Appendix : Analysis (performance metrics)

- Precision(=Positive Predictive Value(PPV)) : 정밀도
  - ✔ True로 예측한 값 중에 실제 정답(true)을 맞춘 비율

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Appendix : Analysis (performance metrics)

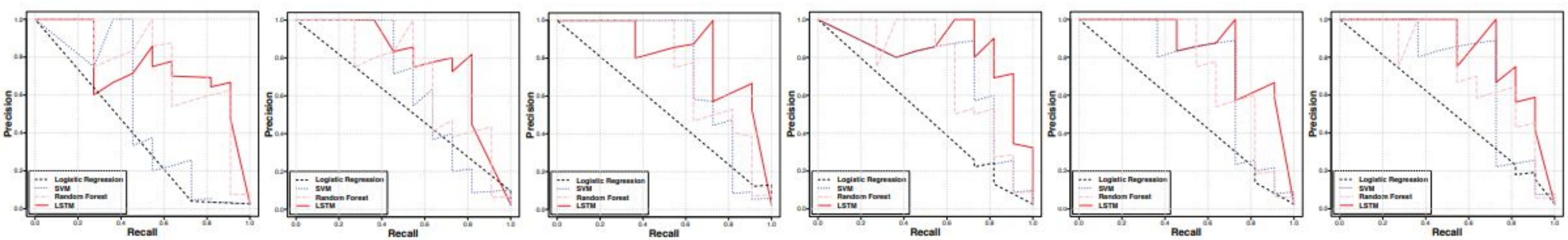
- Recall(=Sensitivity, hit rate, True Positive Rate(TRR))
  - ✔ 정답을 맞춘 것 중에 True로 예측한 것의 비율

$$Precision = \frac{TruePositive}{TruePositive + FalseNegative}$$

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Appendix : Analysis (performance metrics)

- Precision-Recall Curve
  - ✔ PR-AUC: the Area Under the Curve of Precision-Recall (PR-AUC)
  - ✔ Ideally, both high precision as well as high recall. This means that the part of the curves that are <u>closer to the upper right corner are desirable</u>.



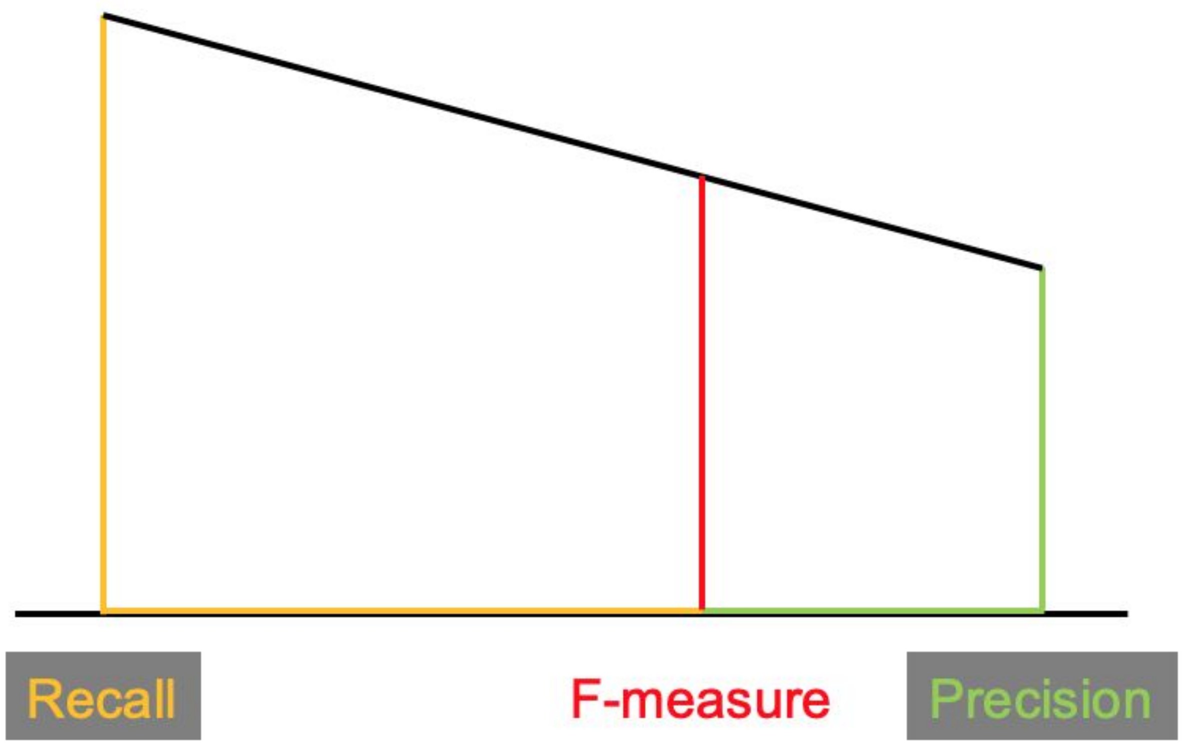
(a) $L = 1$ (b) $L = 2$ (c) $L = 3$ (d) $L = 4$ (e) $L = 5$ (f) $L = 6$

Figure 6. The curve of Precision-Recall for WSC dataset with regard to different sequence length L. The result for MSC dataset is eliminated here.

Appendix : Analysis (performance metrics)

- F-measure(F1 score)
  - ✔ Harmonic mean of the Precision and Recall

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



latency increase