Goal-Driven On-Line Imbalanced Streaming Data Preprocessing

***Abstract*— For the upcoming era of Internet of things (IoT), plethora of data needs to be processed on a real-time basis. However, most prior data streaming related researches ignored the imbalanced data and its semantic meaning, thus often leading to unexpected or even unacceptable results. To address the above issue, this study proposed goal-driven on-line data preprocessing for the analytics of IoT-enabled streaming data. The proposed ontology-based approach can enhance the abilities of semantic understanding for real-time IoT applications. With the above enhancement, the experiment results show its potential in guaranteeing that the model does not go astray from the learning goal due to data imbalance.**

***Index Terms*—Data stream analytics, goal-driven machine learning, IoT, semantic web, data imbalance**

I. Introduction

Internet of things (IoT) has changed our lives with new technologies and unprecedented convenience. Plethora of data will be produced in IoT-enabled environments, and the data needs to be processed rapidly to improve system responsiveness through data stream analytics. In this regard, models with the ability to timely adjust parameters have been attracting more attention. For example, Pratama et al. [1] used Extreme Learning Machine (ELM), Mirza et al. [2] used Online Sequential Extreme Learning Machine (OsELM), and Babu et al. [3] used Neural-Network (NN) for real-time data analytics.

Applying machine learning to on-line data stream analytics did not always go well as expected in the beginning. For example, the initial version of beauty AI (http://beauty.ai) turned out to be a racist because its model was trained using imbalanced on-line input data from users (i.e., data with imbalanced race distribution). It is difficult to control the quality of the on-line input streaming data, and most existing machine learning researches did not consider the semantic appropriateness of input data.

To address the above issue, this study empowers a smart machine with the ability of pre-learning goal planning from human, and proposes its counterpart as goal-driven on-line data preprocessing. In other words, this study leverages ontology to acquire the semantic meaning of input data for checking if an input data conflicts with the learning goal.

II. The Proposed System

Before implementing the proposed goal-driven on-line data preprocessing, a user needs to set a learning goal for model training. To achieve the desired goal, this study focuses on utilizing feature balancing or filtering via sampling technique to exclude those data which may deviate the model from learning unexpected input data. In the system interface, the user chooses those features in the training data which may be related to the sensitive issues such that the system can examine the data of the features on a real-time basis to exclude undesired data.

To enable the system to examine sensitive data, this study made uses of Wiki APIs to acquire the semantic meaning for a chosen feature. The relation of a data w.r.t. its semantic meaning is represented using ontology. During on-line data analytics, the system can continuously construct or update the knowledgebase, represented using ontology trees, and estimate the degree of similarity between a data and a sensitive issue of interest.

In this study, an example of the URL to query the wiki APIs is "https://en.wikipedia.org/w/api.php?action=query&prop=linkshere&format=json&lhlimit=500&titles=topic_to_query." The key argument is titles=topic_to_query, which is used to assign the topic of interest for the query (i.e., a sensitive issue). The query result enumerates the pageID, name space (ns), and other information associated with the topic.

Since our study needs to search related information regarding a sensitive issue from different fields on the Wiki, we choose the "linkshere" property on a Wiki page. For example, to know if there is any relationship between "black" and "racism," the system can first query all hyperlinks that link to web pages of "black" and "racism" topics independently. Next, the system can examine if there exist common hyperlinks among these two topics. If yes, it means these two topics may be relative because both topics are discussed at the same time on Wiki.

To determine the degree of similarity between two topics on Wiki, an index is proposed based on [4], which was originally used to compute the frequency of two words used together to determine the similarity of the two words. The proposed index of the topic similarity is shown below:

$$S(\mathrm{x},\mathrm{y}) = \frac{\|L_x \cap L_y\|}{\min(\|L_x\|,\|L_y\|)} \qquad (1)$$

where L_x and L_y mean the numbers of "linkshere" results for topic x and y on Wiki pages. We calculate the number of common "linkshere" between L_x and L_y as the numerator. The denominator is the less popular topic between L_x and L_y for normalization.

Intuitively, a query on a popular topic will return more results but each with less relevant information. On the contrary, less popular topic often has stronger related information but each result has less amount of "linkshere." This topic similarity index will be categorized into strongly-related, related, weakly-related and not-related. A generalized ontology tree is shown in Fig. 1, which can represent the similarity degree between a data and the sensitive issue. This representation enables the system to find the potential relationship between two chosen

This work was supported by the Ministry of Science and Technology, Taiwan under Grants MOST 106-2221-E-011-029.

Ching-Hu Lu, Chun-Hsien Yu, and Chang-Ru Chen are with the Department of Electrical Engineering, Taiwan Tech, Taipei 106, Taiwan (e-mail: jhluh@ieee.org, tpps88404@gmail.com, and st91012st@gmail.com).
Shih-Shinh Huang is with National Kaohsiung First University of Science and Technology, Kaohsiung, Taiwan (powwhuang@gmail.com)

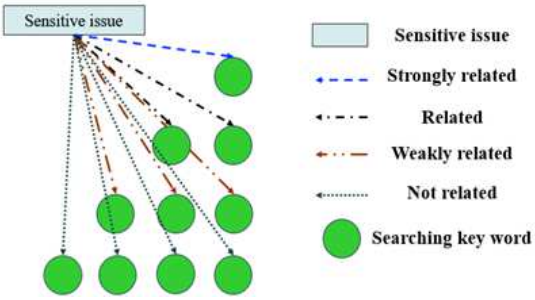

Fig. 1. The issue tree represented using an ontology tree

topics. To avoid the model from training using unbalanced data, this study will keep track of the number of each category of the training data related to the sensitive issue is close to each other using random resampling.

III. Experiment Result

Since there is no open dataset that is suitable for this study, we used the adult dataset [5] shared by UCI to evaluate our system. This dataset contained nearly 50,000 adult data with 14 features including job, nation, skin color, age, gender, etc. The output label is whether an adult earns over 50,000 USD per year or not.

Since this dataset is not real streaming data, this study therefore implemented a service recommendation system yet used the dataset to simulate an IoT-enabled smart-living scenario. In the scenario, a very crowded airport equipped with many smart cameras at the customs to speed up the immigration process. As soon as a passenger's passport is scanned, the system will obtain the basic profile of this passenger like age, job, gender, nation, etc. in the dataset. The cameras connected to the electronic billboards will first recognize the passenger's face as well as the skin color.

This busy airport will soon accumulate many data, which enables the billboards to estimate this passenger's income for later recommending travel and commercial information. This race/income pattern may last for a while when there is no concept drift problem occurs. However, the country suddenly encourages business and technology immigration from all Asian countries. The immigrants from these countries were commonly regarded having lower income, but the business and technology immigrants often have higher income. Without the proposed mechanism of this study, the system will always keep the old income distribution for all Asian countries, thus leading to potential racism issue. Such a problem also hinders the system to adjust its estimation model to ensure the accuracy of an advertisement.

With the proposed approach, the system can keep track of the numbers of each race in the dataset to avoid racial bias. This dataset contains five racial categories, including Asian, Black, Indian, White, and others. With the proposed approach, the system now will not always keep its stereotype regarding race vs. income pattern. When there is any new nation/race appears in the feature, our systems will be able to query Wiki to calculate the similarity degree for this new race.

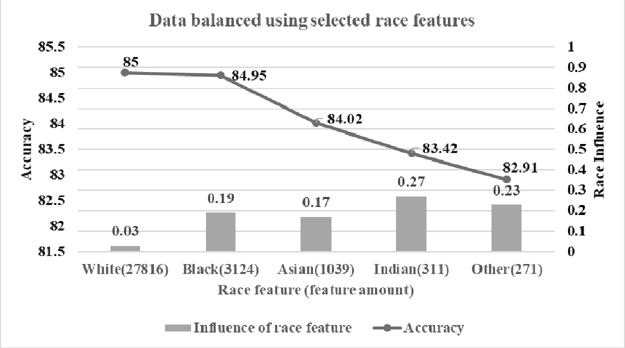

Fig. 2. Number of feature and simularity degree of the adult dataset

In our experiment, we used the Adaptive Network-Based Fuzzy Inference System (ANFIS) [6] as the data model. Fig. 2 shows the accuracy of using testing data to verify the system, where the data was balanced by under-sampling those race with more data than the selected baseline race. The accuracy of the system is the highest because the White race has the highest number of data samples (i.e., complete training dataset was used). However, under-sampling data using other race features (other than the white race) led to a decline in accuracy due to forsaking some data sampling during model training. After balancing the dataset, the race feature becomes more influential. The influence of each feature is primarily represented by the variance parameter of a Gaussian distribution for the ANFIS models. The higher variance represents increasing influence of a feature. Without the under-sampling, an important feature of interest may always stay insignificant due to data imbalance. This experiment result verifies that the proposed approach can improve the influence of a selected feature under the situation of data imbalance, which in turn prevents the model from going astray from the preset learning goal.

IV. Conclusion and Future Work

The goal-driven on-line data preprocessing for IoT-enabled streaming data analytics is proposed to ensure the contextual model not going astray from the preset learning goal. The experiment result has shown that the proposed approach can help balance imbalanced data at the cost of decreased accuracy. The system will let the user decide the trade-off between accuracy and data balance. This study has preliminarily evaluated the feasibility but further enhancement is required. One key improvement is to enhance the feature balance method using oversampling to avoid the loss of data.

References

[1] M. Pratama, G. Zhang, M. J. Er, and S. Anavatti, "An Incremental Type-2 Meta-Cognitive Extreme Learning Machine," *IEEE Transactions on Cybernetics,* vol. PP, pp. 1-15, 2016.

[2] B. Mirza and Z. Lin, "Meta-cognitive online sequential extreme learning machine for imbalanced and concept-drifting data classification," *Neural Netw,* vol. 80, pp. 79-94, Aug 2016.

[3] G. Sateesh Babu and S. Suresh, "Meta-cognitive Neural Network for classification problems in a sequential learning framework," *Neurocomputing,* vol. 81, pp. 86-96, 2012.

[4] Y. Wang, T. Peng, and W. Zuo, "Hyponymy Graph Model for Word Semantic Similarity Measurement," *Chinese Journal of Electronics,* vol. 24, pp. 96-101, 2015.

[5] I. University of California, "Adult Data Set," 2013.

[6] J. S. R. Jang, "ANFIS: adaptive-network-based fuzzy inference system," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 23, pp. 665-685, 1993.