Imbalanced Classification for Fault Detection in Monitored Critical Infrastructures

Yan-Fu Li[1]

[1]*Department of Industrial Engineering, Tsinghua University, Beijing, 100084, China*
liyanfu@tsinghua.edu.cn

ABSTRACT

Safety and reliability are among the most crucial factors for the critical infrastructures (CIs). For this reason, they are typically closely monitored and large amounts of data have been collected. Due to their importance, CIs are designed to be highly reliable such that fault cases are rare in the Big-data set. This renders the fault detection an imbalanced binary classification task. In this work, we developed accurate data mining classifier to tackle this problem. The imbalance ratio of the data can be more than 200.

1. INTRODUCTION

The prosperity of human society largely depends on safe and stable operation of several key industrial systems, e.g. power network, railways, aviation, etc. Due to their importance, these systems are often referred to as critical infrastructures (CIs). To ensure the safety operation of CIs, many sensors are installed to monitor the systems' states, and control their efficient and safe operation. The large amount of real-time data recorded by the sensors can be utilized in data-driven approaches for fault detection of such systems, which is a binary classification problem that amounts to discriminating the data as belonging to the fault class or the normal operation class. However, not all recorded data are useful for such objective and pre-processing, including data cleaning and dimensionality reduction (DR), is essential to extract the features relevant for the detection task. A typical complication for fault detection in CIs is that these systems are highly reliable, so that only a small fraction of the available data is relevant to its failures. This makes the fault detection task a classification problem with highly imbalanced data, where the data size of the class of interest (failure of CI) is much smaller than that of the other class (normal operation of CI).

For fault detection in practice, an explainable model is more acceptable by engineers and operators than a black-box model, and a probabilistic model is needed to handle the uncertainties in the detection model, due to measurements errors and incomplete datasets. A probabilistic explainable model can provide insights into the mechanisms of failures, along with the failure predictions. By analysis and comparison of four representative data mining methods, Bayesian Network (BN) has been chosen for this work. Cost-sensitive learning is integrated as the objective for training the BN model. A symmetric uncertainty (SU)-based feature selection method, i.e. Correlation-based Feature Selection (CFS) is combined with BN for DR. Sensitivity analysis is performed empirically with respect to the Imbalance Ratio (IR).

The rest of this paper is organized as follows. In section 2, the selection of dimensionality reduction and machine learning method is presented. Section 3 describes the implementation method and empirical results. Section 4 concludes this work.

2. DIMENSIONALITY REDUCTION & MACHINE LEARNING

The following section outlines the selection of dimensionality reduction method and data mining method for the imbalanced classification problem.

2.1. Dimensionality Reduction Methods

Given the large number of features at least four reasons call for a reduction on the number of features, for practical purposes. Dimensionality Reduction (DR) is the term used for the task of reducing the number of features, while representing the original data at sufficient level of accuracy. Both Feature Extraction (FE) and Feature Selection (FS) methods can perform a DR.

FE techniques map the initial n-dimensional data into an m-dimensional space, where $m<n$ (Dash and Liu, 1997). All n measurements are used to obtain the m-dimensional data, which are expected to non-redundantly contain all relevant information of the original data, so that the subsequent machine learning activities are performed on this reduced representation.

Differently, FS techniques aim at selecting a subset of features, which can efficiently represent the original data while reducing effects from noise or irrelevant variables and still provide good classification results (Guyon and

Elisseeff, 2003). To remove an irrelevant feature, a FS criterion, which can measure the relevance of each feature in determining the output, is required. Once a FS criterion is determined, a search procedure is developed to find the subset of useful features, which most satisfy the criterion. Several methods have been proposed for this task. They are grouped into two main categories: filter and wrapper (Kohavi and John, 1997). Filter methods act as preprocessing to select features by evaluating certain preset criteria independent of the accuracy of the classifier; they discard irrelevant and/or redundant features *a priori* of the construction of the classifier (Dash and Liu, 1997). On the contrary, wrapper methods convolve with the specific learning algorithms to generate the optimal feature subset, i.e. that group of features upon which one can construct a classifier with the highest possible accuracy (Kohavi and John, 1997).

To identify the most appropriate FS technique for this work, the following evaluation criteria are considered: representativeness, adaptability and efficiency. For FE, it will create new features, which might not have clear physical meaning and eventually hamper the explainability of the classifier. For wrapper, it might consume considerable amount of computation time to find a feature subset, which still might contain redundant features. Filter is preferred because by properly choosing the selection criterion and search strategy we are able to efficiently exclude the redundant or irrelevant features for classifier building in the next stage.

The SU based feature subset selector, correlation-based feature selection (CFS) (Hall, 1999), has been, then, utilized. CFS evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. The subsets of features that are highly correlated with the class while having low inter-correlation are preferred in this method.

2.2. Data Mining Methods

There are numerous data mining methods that have been developed in the literature. However, not all of them are well suitable for knowledge extraction. For example, the well-known artificial neural network (ANN) (Misra and Saha, 2010) can produce accurate predictions on many classification and regression problems, but the trained ANN model is rather difficult to explain or comprehend and remains a black-box to the practitioners (Saravanan and Ramachandran, 2010). In applications where the knowledge extraction from the database is of the main interest, such black-box cannot be accepted, as the knowledge is sought in certain format, e.g. logic rules, which can be understood, analyzed and modified for the reuse by the practitioners.

In the data mining literature, there are four established knowledge extraction techniques: association rule learning, Bayesian networks, neuro-fuzzy inference systems, case-based reasoning. The following five criteria are used to select appropriate data mining method.

1) *Explainability*: A transparent model giving explainable results is always preferred, to gain insights into the problem itself.
2) *Probabilistic:* In order to have the capability to predict future failures, it is more reasonable to produce a probability distribution.
3) *Efficiency of model building:* Given the number of features and amount of events, it is important to consider the training time for each method.
4) *Capability of using expert knowledge:* There can be knowledge provided by domain experts: it is a plus for the method to be able to incorporate that knowledge.
5) *Adaptability:* More data may be available for knowledge extraction and the method is expected to be able of being adaptively updated.

From Table 1, it is seen that BN is well suited for the work. As to association rule learning (ARL), it is not very straightforward to handle adaptively the time sequence data. For neuro-fuzzy system, it is non-probabilistic and relatively difficult to train and explain. For case based reasoning (CBR), it is a good alternative because it can be built more quickly than the others, but it is purely data-driven and, thus, not capable of using the expert knowledge.

Table 1. Comparison of Different Knowledge Extraction Methods with Respect to the Selection Criteria

Characteristics	ARL	BN	Neuro-fuzzy system	CBR
Explainability	Explainable through statistical relations	Explainable through probability inference	ANN is a black box. It is not very straightforward to explain the parameters obtained for the fuzzy rules	Explainable through analogical inference
Probabilistic	Yes	Yes	No	Yes
Efficiency of model building	Exponential, but there are polynomial algorithms	Exponential, but there are polynomial algorithms	Include two parts: for fuzzy rule generation it is exponential but polynomial algorithms exist; for parameter update the ANN has to be trained	Linear complexity
Capability of using expert knowledge	Yes, able to deal with logic rules	Yes, via Bayesian updating	Yes, able to deal with linguistic rules	No
Adaptability	Yes, but need special technique to extract temporal rules	Yes, via dynamic Bayesian network	Yes	Yes

3. MODEL BUILDING AND RESULTS

Prior to applying any FS method to the dataset available, due to the highly imbalanced class distribution in the dataset we have produced a random sample of the original dataset using sampling with replacement, in order to bias the class distribution toward a uniform distribution. The resulting dataset contains 146014 event records, among which 73042 are non-failures and 72972 are failures, randomly replicated from the original 308 failures. This technique, named balanced minority repeat (BMR), is frequently used to preprocess imbalanced class distributions in order to avoid selecting the features that may favor majority class examples but are of very little values for predicting the minority failure events. After the BMR re-sampling, CFS is used to rank all the features.

There are 3 major phases for model construction: 1) discretization of the numeric (continuous) features; 2) feature selection; 3) learning the Bayesian network model. During this process, there are 2 parameters that need to be optimized: the number of selected features and the parameter α in cost-sensitive learning. Cost-sensitive learning is one promising approach to deal with highly imbalanced class distribution without over or under sampling the original dataset. Under this framework, the objective of the trained BN model for the imbalanced data are defined as minimizing the total cost expressed as

$$Cost_{all} = \alpha * FN + FP \tag{1}$$

with FN and FP being respectively the number of misclassified events in minority (positive) and majority (negative) class and α being a parameter larger than one showing that the accuracy on the positive class is of more interest.

10-fold cross-validation scheme is chosen to evaluate the quality of different parameter values. After the training process, the best parameters will be selected to construct BN model for final testing. The number of selected features ranges from 1 to 43 and α ranges from 1 to 50. In total, 43*50 = 2150 parameter combinations are investigated. The accuracy metrics are the following:

$$Precision_{+} = \frac{TP}{TP+FP} \tag{2}$$

$$Recall_{+} = \frac{TP}{TP+FN} \tag{3}$$

$$F - measure_{+} = \frac{2 * Precision_{+} * Recall_{+}}{recision_{+} + Recall_{+}} \tag{4}$$

where TP is the number of correctly classified minority events. Typically the BN model maximizing the precision could result to a relatively low recall rate. To remedy this issue, $F - measure_{+}$ in (4), a composite metric taking into account both recall and precision on the positive class, is used for model selection. The accuracy metrics values in (2)-(4) are shown in Table 2. The number of selected features is 32 and FP penalty α equals to 4.

Table 2 Accuracy Metrics of the Model with Maximal F-Measure on the Positive Class.

Class	Recall	Precision	F-measure
Normal (negative)	0.995	0.999	0.998
Failure (positive)	0.893	0.542	0.674

The selected features cover 11 out of the 12 features proposed by the expertise of our industrial partner. This shows that the proposed method can extract reasonably well the useful knowledge for fault detection of braking system in a high speed train.

4. CONCLUSION

In this paper, an imbalanced classifier is proposed to extract knowledge for fault detection of CIs. It integrates the CFS and BN model. The objective of a trained BN model is to minimize the total cost on the dataset while a larger cost is assigned to the classification error on the minority (failure) class. Experiment shows that the proposed method can achieve good results. The future work will focus on the test of the proposed method on other CIs and the detection of different types of failures in the CIs.

ACKNOWLEDGEMENT

The author would like to thank Dr. Jie Liu and Prof. Enrico Zio at CentraleSupélec, Paris for their insightful comments and suggestions on this manuscript and the industrial partners for providing empirical example and the data.

REFERENCES

Dash, M., & Liu, H. (1997). Feature selection for classification. Intelligent data analysis, 1(3), 131-156.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(3), 1157-1182.

Hall, M. A. (1999). Correlation-based feature subset selection for machine learning. PhD Dissertation, University of Waikato.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1-2), 273-324.

Misra, J., & Saha, I. (2010). Artificial neural networks in hardware: A survey of two decades of progress. Neurocomputing, 74(1), 239-255.

Saravanan, N., & Ramachandran, K. I. (2010). Incipient gear box fault diagnosis using discrete wavelet transform (DWT) for feature extraction and classification using artificial neural network (ANN). Expert Systems with Applications, 37(6), 4168-4181.