Journal of Manufacturing Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Manufacturing Systems

journal homepage: www.elsevier.com/locate/jmansys

Imbalanced data fault diagnosis of rotating machinery using synthetic oversampling and feature learning

Yuyan Zhang[a], Xinyu Li[a], Liang Gao[a,*], Lihui Wang[b], Long Wen[a]

[a] State Key Laboratory of Digital Manufacturing Equipment and Technology, Huazhong University of Science and Technology, Wuhan, China
[b] Department of Production Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

ARTICLE INFO

Keywords:
Rotating machinery fault diagnosis
Weighted minority oversampling
Feature learning
Imbalanced data fault diagnosis

ABSTRACT

Imbalanced data problems are prevalent in the real rotating machinery applications. Traditional data-driven diagnosis methods fail to identify the fault condition effectively for lack of enough fault samples. Therefore, this study proposes an effective three-stage fault diagnosis method towards imbalanced data. First, a new synthetic oversampling approach called weighted minority oversampling (WMO) is devised to balance the data distribution. It adopts a new data synthesis strategy to avoid generating incorrect or unnecessary samples. Second, to select useful features automatically, an enhanced deep auto-encoder (DA) approach is adopted. DA is improved in two aspects: 1) a new cost function based on maximum correntropy and sparse penalty is designed to learn sparse robust features; 2) a fine-tuning operation with a self-adaptive learning rate is developed to ensure the good convergence performance. Finally, the C4.5 decision tree identifies the learned features. The proposed method named WMODA is evaluated on 25 benchmark imbalanced datasets. It achieves better results than five well-known imbalanced data learning methods. It is also evaluated on a real engineering dataset. The experimental results show that WMODA can detect more fault samples than the traditional data-driven methods.

1. Introduction

The fault diagnosis of rotating machinery is of utmost importance, because early detection of emerging fault helps to significantly improve the operational continuity and equipment safety [1]. Generally, existing diagnosis methods are categorized into model-based methods, data-driven methods and hybrid methods [2]. Data-driven methods are especially powerful for the complex industrial processes where the explicit system models are challenging to derive [3]. However, in real applications, data are imbalanced because there are an insufficient number of fault samples, and this causes the traditional data-driven methods to fail in identifying the fault condition. Considering its importance, rotating machinery fault diagnosis with imbalanced data has attracted widespread research interest.

Synthetic oversampling is one of the main methods to handle the imbalanced data in rotating machinery fault diagnosis. The synthetic minority oversampling technique (SMOTE) [4] is one of the well-established approaches. A combination of SMOTE and logistic regression was proposed for breakage detection of multi-tooth tools [5]. SMOTE and Bagging techniques were combined for wind turbine fault diagnosis [6]. Borderline-SMOTE was put forward to optimize the data synthesis strategy of SMOTE and had been applied to the bearing fault diagnosis [7]. Adaptive synthetic oversampling (ADASYN) [8] adaptively determined the synthetic number for the border samples. ADASYN was integrated with particle swarm optimization to perform fault diagnosis of power transform systems [9]. However, these synthetic oversampling approaches may generate incorrect or unnecessary samples in some scenarios (see section 3.1.1).

Therefore, to address the imbalanced data distribution, a new synthetic oversampling approach called weighted minority oversampling (WMO) is devised. The main steps of WMO include: 1) find informative samples; 2) partition them into subspaces using a clustering algorithm and determine the synthetic sample number adaptively; 3) adopt a new data synthesis strategy to avoid generating incorrect or unnecessary samples. It should be noted that the function of WMO is only to balance the data distribution. How to select useful features and identify them effectively remains to be solved.

Deep auto-encoder (DA) is an unsupervised learning approach that has the ability in selecting the useful low-dimensional features automatically [10,11]. As DA is not dependent on statistical analysis

* Corresponding author.
E-mail addresses: gaoliang@mail.hust.edu.cn, 439360237@qq.com (L. Gao).

https://doi.org/10.1016/j.jmsy.2018.04.005
Received 26 August 2017; Received in revised form 9 April 2018; Accepted 9 April 2018
0278-6125/

knowledge, it has been widely used for feature selection. Jia [12] employed DA to select features from a high-dimensional frequency spectrum. Qi [13] utilized DA to select features from time-frequency data. In our previous work, DA was used to select features from frequency spectrum [14]. DA has shown the good performance in automatic feature selection, but it still has two drawbacks: 1) the standard cost function is not robust for noise data and also not suitable for redundant data [15,16]; 2) when fine-tuning DA, the learning rate is usually fixed and this results in slow convergence or error oscillation [12]. To overcome these drawbacks, DA is improved in two aspects. First, a new cost function based on maximum correntropy and sparse penalty is designed. Maximum correntropy is insensitive to noise data [17] and the sparse penalty is suitable for handling redundant data. Second, a fine-tuning operation with a self-adaptive learning rate is developed to ensure the good convergence performance.

Once the useful low-dimensional features have been obtained, the C4.5 decision tree is employed to classify them. Overall, the proposed method, named WMODA, is a three-stage method. It is evaluated on 25 UCI (University of California Irvine) benchmark imbalanced datasets and achieves better results than five well-known imbalanced data learning methods. It is evaluated on a real engineering dataset as well. Extensive experimental results show that WMODA can detect more fault samples than the traditional data-driven methods.

The remainder of this paper is organized as follows. Section 2 briefly describes the basic methods. Section 3 details the proposed WMODA method. In section 4, WMODA is evaluated on the UCI datasets. In section 5, we apply WMODA on a real engineering dataset, and then present the conclusions of the study.

2. Brief introduction of several basic methods

Synthetic oversampling approaches are widely used to balance the data distribution. DA can select the useful features automatically. This section provides a brief introduction of several representative synthetic oversampling approaches along with the principle of DA.

2.1. Overview of synthetic oversampling approaches

In this subsection, several representative synthetic oversampling approaches are reviewed. It should be noted that this study focuses on the binary classification problems. For clarity, the class with ownership of many samples is called the majority class and the other is called the minority class.

Synthetic oversampling approaches generate new samples to alleviate the imbalance level [18]. SMOTE is one of the best-known approaches, whose main idea is to interpolate a new data point between a selected data point and one of its nearest neighbors [4]. Based on SMOTE, several variants were proposed to optimize the selection procedures and the generation procedures. Borderline-SMOTE [19] generated new samples only for the borderline samples. Modified-SMOTE [20] grouped the minority class samples into security samples, border samples and noise samples. Different generation strategies were employed on them separately. ADASYN [8] determined the synthetic sample number adaptively and used SMOTE to generate new samples. Safe-level-SMOTE [21] assigned each minority class sample a so-called safe level. It adopted SMOTE to generate new samples only in the safe region. Ramentol [22] combined a rough set approach with SMOTE to improve the generation strategy. Lim [18] adopted three clustering algorithms to partition the minority class dataspace and used SMOTE to generate new samples. However, the above reviewed approaches may generate incorrect or unnecessary samples in some scenarios (see section 3.1.1). Therefore, we propose a new synthetic oversampling approach and improve the generation strategy.

2.2. Deep auto-encoder

DA introduced by Hinton [10] is widely applied to learn low-dimensional features from the high-dimensional data automatically [23]. DA is constructed by stacking several auto-encoders and the stacking details can refer to literature [10]. Each auto-encoder (AE) consists of an input layer, a hidden layer and an output layer. Suppose the inputs are $\{\mathbf{x}_i\}_{i=1}^{K}$, where $\mathbf{x}_i \in \mathrm{R}^{n\times 1}$ is a sample, K is the sample size. The hidden values and the output values are computed by

$$\mathbf{h} = s(\mathrm{W}\mathbf{x} + \mathbf{b}) \tag{1}$$

$$\hat{\mathbf{x}} = s(\mathbf{W}'\mathbf{h} + \mathbf{b}') \tag{2}$$

where $\mathbf{W} \in \mathrm{R}^{m\times n}$ and $\mathbf{W}' \in \mathrm{R}^{n\times m}$ are connection weights, $\mathbf{b} \in \mathrm{R}^{m\times 1}$ and $\mathbf{b}' \in \mathrm{R}^{n\times 1}$ are biases, and $s(\cdot)$ is the sigmoid activation function. Training an auto-encoder is to minimize the cost function. The standard cost function is the mean square error (MSE)

$$J = \frac{1}{2K}\sum_{i=1}^{K} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||^2 \tag{3}$$

3. The proposed WMODA method

This section details the proposed WMODA method. It involves three parts: WMO for balancing the data distribution, enhanced DA for learning sparse robust feature and the overall framework of the proposed method.

3.1. Weighted minority oversampling

3.1.1. Motivation

Initially, we point out several potential problems in the reviewed

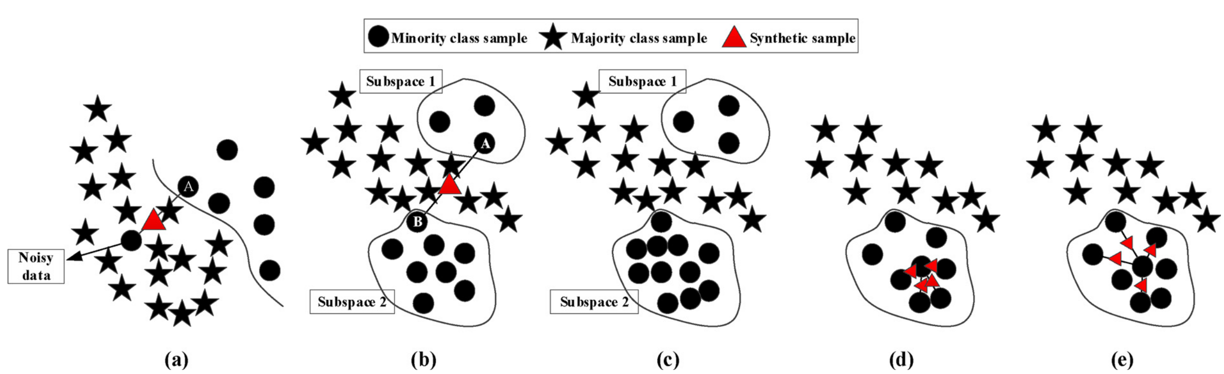

Fig. 1. Illustrations of potential problems in the data generation.