Correction of Atmospheric Model Through Data Mining With Historical Data of Two-Line Element

XUE BAI1, CHUAN LIAO2, MING XU1, AND YARU ZHENG1

1School of Astronautics, Beihang University, Beijing 100191, China
2No. 10 Research Institute, China Electronics Technology Group Corporation, Chengdu 610036, China

Corresponding author: Ming Xu (xuming@buaa.edu.cn)

This work of Ming Xu was supported in part by the National Natural Science Foundation of China under Grant 11772024 and Grant 11432001.

ABSTRACT The existing atmospheric mass density models (AMDM) would produce considerable errors in orbital prediction for Low Earth Orbit (LEO) satellites. In order to reduce these errors and correct the AMDM, this paper presents methods based on data mining with historical data of two-line element (TLE). Starting from a typical LEO satellite, TIANHUI, two orbital dynamical models are firstly proposed as the simulation environment to generate training data. The historical TLE data are regarded as actual space environment and used to generate application data. Secondly, three data mining methods, Random Forest (RF), Artificial Neural Network (ANN) and Support Vector Machine (SVM), are combined with the training data to investigate their feasibility in recovering the known deviation of AMDM under simulation environment. Training results show that RF displays the best performance and achieves the accuracy of 99.99%, while the other two methods only achieve 86.83% and 71.90% respectively. Thirdly, under the actual space environment, this paper uses new training and application data to research the ability of the three methods in recovering the unknown deviation of the AMDM and improve the accuracy of orbital prediction. Numerical results are evidential to the accuracy of the proposed methods based on data mining. It is concluded that the capabilities of the data mining for correction for the atmospheric model are very promising, with great potential to advance practical applications on on-orbit propagation.

INDEX TERMS Data mining, atmospheric mass density model, random forest, artificial neural network, support vector machine, two-line element.

I. INTRODUCTION

Data mining, also known as Knowledge Discovery in Database (KDD), was first emerged in the 1980s and got rapid development in the 1990s. It provides a method to discover the potential pattern in a big database. Feasibility, usefulness, validity and scalability are the four major problems in this technology [1]. In the past few years, some industries and researchers have applied KDD to improve their services and products. The financial securities industry use KDD to predict the fluctuation and tendency to share price, and analyze the investment habit of investors [2]. Fu [3] gave a comprehensive revision on the existing time series data mining research which is categorized into representation and indexing, similarity measure, segmentation, visualization, and mining. He thought that according to the unique behavior of the time series data, existing research is still inadequate.

The associate editor coordinating the review of this manuscript and approving it for publication was Hailing Chen1.
of onboard data mining and believed that large networks of remote sensors would become more common in improving technology and declining costs. Immediate management of the collected data on board would contribute to increasing the network communication capabilities and cutting down expenses. Sánchez Sánchez et al. [11] used Deep Neural Network (DNN) to solve the Hamilton-Jacobi-Belman equation and provided a near-optimal control strategy for spacecraft. Hennes et al. [12] focused on the problem of optimally transferring a spacecraft from an original asteroid to the targeted one and studied in detailed approximations for the phasing value, the maximum initial mass and the arrival mass. Li et al. [13] analyzed the connotation and extension of Spatial Data Mining and Knowledge Discovery (SDMKD), and used it to reveal potential patterns from space databases automatically or semi-automatically. Gong et al. [14] focused on the lack of effective technologies of data mining for volume remote sensing data and designed a technological framework of the remote sensing data mining and knowledge discovering system. Peng and Bai [15] concentrated on the lack of area-to-mass ratio of a resident space object in most space catalogs and proposed an improved method of satellite orbital prediction. They used Random Forest (RF) to learn the connection between the consistency error and area-to-mass ratio, and Support Vector Machine (SVM) to learn the errors of orbital prediction [16], [17]. However, their work is just based on their simulation environment and did not apply TLE data or any other actual on-orbit data. Hu et al. [18] analyzed the working and fault characteristic of liquid–propellant rocket engine, and used data mining techniques to detect and diagnose the fault of liquid-propellant rocket engines. However, they did not use any actual data either. In order to rapidly diagnose the satellite’s fault, Xu and Pi [19] proposed a method to mine abnormal patterns in satellites based on Prefix Span. With the same purpose, Zhao and Li [20] proposed a new fault diagnosis method of spacecraft based on time series data mining. We have also used data mining techniques to detect the orbital maneuvers of satellites at different scales in a previous study [21].

North American Air Defense Command (NOARD) has accumulated a large amount of historical TLE data of the LEO satellite. TLE data are “mean” values obtained by removing periodic variations in a particular way. Therefore, they are using an obsolete true equator and mean equinox (TEME) coordinate system. To predict satellite position and velocity under the inertial coordinate system, NOARD has released the simplified general perturbations (SGP4) propagator to reconstruct periodic variations [22]. It is noticeable that the derived orbital elements are fundamentally decided by the accuracy of original TLE data. The TLE uncertainties are originated from semi-analytic SGP4 orbit model for orbit propagation, radar tracking and processing. For LEO satellites, the TLE data within 5-day propagation provide appropriately accurate orbital information with errors remaining bounded within roughly 0.5 km [23], [24]. In this paper, the TLE data and SGP4 are used to calculate the satellite’s historical on-orbit states.

The satellites would be influenced by various effects of perturbations, including Earth’s gravity, third-body perturbation, solar radiation pressure and atmosphere drag, etc. Especially for LEO satellites, atmosphere drag would cause considerable orbit decay and it is the main error source in orbital prediction. Most empirical atmospheric density models [25], such as Jacchia [26], NRLMSISE [27], DTM [28], Jacchia–Bowman 2006/2008 [29], [30] and Russian GOST [31] combine theoretical computation with drag measurements of satellites. Their accuracy is much better than the traditional exponential one, but still no better than 10%~30%. It is worth mentioning that indirect solar activity has a great influence on AMDM by thermosphere and causes cycle, seasonal and geomagnetic effects, which result in errors [32]. Generally, there are two main methods to increase the accuracy of atmospheric density model [33]. One is to derive the AMDM directly using measurements from a high-accuracy onboard accelerometer [34]. The other is to calibrate the AMDM using large numbers of orbital measurements with lower accuracy [35]. Chen et al studied the performance of the AMDM coefficient calibration method, the Nazarenko method, and the parameterization method in improving orbital prediction of LEO objects [36]. The central idea of this paper is to use the second method based on historical TLE data to correct the AMDM and realize the orbital prediction. Meanwhile, data mining technology is good at dealing with a large amount of data to excavate those hidden, short-term but potentially useful, providing new techniques to improve the accuracy of the AMDM instead of replacing them and benefit the orbital prediction of LEO satellite.

Based on the historical TLE data of TIANHUI satellite, this paper investigates the feasibility of data mining technology in recovering the deviation of the existing AMDM. Two orbital dynamical models, the accurate model and the simplified one with deviation, are built as the simulation environment of TIANHUI to verify the feasibility of data mining technology. The historical TLE data are used as application data to correct the atmosphere model errors. Three data mining methods, including RF, ANN and SVM, are used to classify and correct the AMDM in orbital prediction. This paper is mainly carried out from two aspects: methodology and practical application, shown as the flowchart in Figure 1. In the stage of methodology, the differences between the accurate model and the simplified one with known deviation generate the training and application data. The three data mining methods learn on the training data to obtain three classifiers and then apply them on the application data to get model errors. The training results verify the feasibility of data mining technology in correcting AMDM errors. In the stage of practical application, the application data are generated from the differences of the accurate model and historical TLE data with unknown errors. After retraining with training data, the classifiers import the application data to recover the real AMDM errors. The errors
are then added to the accurate AMDM to get the improved accurate model which exactly indicates the correction results.

We make four contributions in this paper listed as following: 1) It is proved that data mining provides a practical technique to refine valuable information of orbital elements as well as atmospheric density model from historical TLE data. 2) The feasibility of data mining in predicting the AMDM errors is verified in simulation environment provided by the accurate orbital dynamical model and the simplified model with deviation. 3) The proposed methods are demonstrated to have a good performance on recovering the unknown deviation of the AMDM through application data derived from historical TLE data. 4) Orbital prediction is promoted by the improved accurate model with error correction of the classification methods according to the numerical results. Furthermore, data mining methods are verified that with the accessible data based on the point-to-point historical TLE data, they can provide an accurate calculation of satellite’s on-orbit state for ground station and design of satellite’s maintenance control, which are important for orbital prediction.

## II. BASIC KNOWLEDGE

### A. ORBIT DYNAMICS

Accurate orbital prediction needs accurate modeling of the space environment. According to Reference [37], Table 1 gives an overview of the magnitude of the major perturbations. It deals with a LEO satellite with an orbital altitude of 500 km. The absolute values of acceleration are computed as mean values over the time interval of the integration of one day. In order of importance, Earth’s gravity, solid Earth tides, solar attraction perturbation, solar radiation pressure and atmospheric drag are taken into consideration in the following orbit dynamics. Due to other perturbations, such as direct radiation pressure and connected relativistic factors, in small orders of magnitude compared with the TLE data errors themselves, they are not included in the accurate orbit dynamic and not listed in the following contents. The long-term perturbation, such as 11-year solar activity, is regarded as constant because of the short-term TLE data used in this paper.

This section presents the mathematical models of each perturbation above. According to Reference [38], Earth gravitational potential can be written as the following form:

$$ U = \frac{GM_e}{R_e} \sum_{n=0}^{\infty} \sum_{m=0}^{n} (C_{nm} V_{nm} + S_{nm} W_{nm}) $$

(1)

where

$$ V_{nm} = \left( \frac{R_e}{r} \right)^{n+1} P_{nm} (\sin \varphi) \cos m \lambda $$

$$ W_{nm} = \left( \frac{R_e}{r} \right)^{n+1} P_{nm} (\sin \varphi) \sin m \lambda. $$

(2)

$GM_e$ is the Earth gravitational constant, $r$ is the distance from the Earth’s center, $\varphi$ is the geocentric latitude and $\lambda$ is the longitude of the satellite, $R_e$ is the equatorial radius of the Earth, $P_{nm}$ is the associated Legendre polynomial of degree $n$ and order $m$, $C_{nm}$ and $S_{nm}$ are the coefficients which describe the dependence on the Earth’s internal mass distribution. Besides, this paper uses EGM96 model. The detailed values can be found in Reference [39].

Therefore, the acceleration $\ddot{x}$ in the Earth-center fixed coordinate system can be calculated by $V_{nm}$ and $W_{nm}$ listed in Equation (2) as

$$ \ddot{x} = \sum_{n,m} \ddot{x}_{nm}, \quad \ddot{y} = \sum_{n,m} \ddot{y}_{nm}, \quad \ddot{z} = \sum_{n,m} \ddot{z}_{nm} $$

(3)

where

$$ \ddot{x}_{nm} = \begin{cases} \frac{GM_e}{R_e^2} (-C_{n0} V_{n+1,1}) & m = 0 \\ \frac{GM_e}{2R_e^2} \left( \frac{(n-m+2)!}{(n-m)!} \right) \left( C_{nm} V_{n+1,m-1} + S_{nm} W_{n+1,m-1} \right) & m > 0 \end{cases} $$

$$ \ddot{y}_{nm} = \begin{cases} \frac{GM_e}{R_e^2} (-C_{n0} W_{n+1,1}) & m = 0 \\ \frac{GM_e}{2R_e^2} \left( \frac{(n-m+2)!}{(n-m)!} \right) \left( C_{nm} W_{n+1,m-1} + S_{nm} V_{n+1,m-1} \right) & m > 0 \end{cases} $$

### B. ACCURATE ORBITAL PREDICTION

| Perturbation                                | Acceleration ($m/s^2$) |
|---------------------------------------------|------------------------|
| $1/r^2$-Term                                | 8.42                   |
| Atmospheres                                 | $1.5 \times 10^4$       |
| Higher Terms of the Earth’s Gravity Field   | $2.5 \times 10^4$       |
| Solar Attraction                            | $5.0 \times 10^3$       |
| Direct Radiation Pressure                   | $9.7 \times 10^4$       |
| Solid Earth Tides                           | $1.1 \times 10^7$       |

*TABLE 1. Accelerations acting on LEO with an orbital altitude of 500 km.*
Atmosphere drag is the major non-conservative force for the LEO satellite. The acceleration \( \ddot{r}_a \) caused by the atmosphere in the inertial coordinate system is

\[
\ddot{r}_a = -\frac{1}{2} C_D \frac{A}{m} \rho v^2 \varepsilon_v \quad (8)
\]

where \( C_D \) is the drag coefficient, \( A \) is the windward area, \( m \) is mass of the satellite, \( \rho \) is the atmospheric density at the position of the satellite, \( v \) is the velocity relative to the atmosphere \( \varepsilon_v \) is the unit vector of relative velocity, i.e., \( \varepsilon_v = v_r / |v_r| \). It can be seen that the parameters in Equation (8) except \( v \) can cause certain errors and they are required to refine through the correction derived from the data mining technique. MSISE-86 atmospheric density model, the most common model, is applied to calculate the atmospheric density \( \rho \). The traditional modeled atmospheric density can predict long-term behaviors just from the physical factors, but on the other hand, it is weak in refining more accurate short-term predictions [41]. In the latter case, some historical data can contribute to refining the short-term prediction, which is reported in this paper. The proposed data mining technique can excavate those hidden, short-term but potentially useful information from the historical TLE data.

What’s more, from the point of mission design, there are a lot of other complicated noise sources during the process of state transformation, such as the attitude and orbit control, guidance system and transform in the reference frames, which have developed algorithms to deal with the noise in published references. Thus, these noises are not the points discussed in this paper.

### B. ACCURATE MODEL AND SIMPLIFIED MODEL WITH DEVIATION

Section II.A gives the accurate mathematical models of each perturbation in space. These models provide the basic data in orbit prediction of the satellite. In order to generate training and application data for classifiers, this section gives two orbital dynamical models: the accurate model which is the reference of orbit prediction and simulates an “actual” space environment, and the simplified one with deviation which uses simplified mathematical models and adds known deviation value \( \varepsilon \) to the AMDM. From the perspective of engineering, the product of parameters \( C_D \frac{A}{m} \rho \) in Equation (8) causes certain errors and is usually regarded as one parameter \( C_\delta \) during satellite design. Therefore, the goal of this paper is to correct \( C_\delta \). As a result, Equation (8) has been changed in the simplified model with deviation \( \varepsilon \) to the following form:

\[
\ddot{r}_a = -\frac{1}{2} (1 + \varepsilon) C_D \frac{A}{m} \rho v^2 \varepsilon_v = -\frac{1}{2} (1 + \varepsilon) C_\delta \varepsilon_v \quad (9)
\]

The orbit data are generated by these two models and the differences between them are regarded as training data. Specific parameter settings are shown in Table 2. The accuracy of the models with these perturbations here is enough to use as the training data because the errors predicted by these models are beneficial for convergence of the classification and make it easy to get the classifiers.
TABLE 2. Parameter setting of two models.

| Model                  | Accurate Model | Simplified Model with Deviation |
|------------------------|----------------|---------------------------------|
| Earth Shape            | WGS84          | WGS84                           |
| Gravity Field          | 70 × 70        | 5 × 5                           |
| Third-body Perturbation| Sun            | ~                               |
| Solar Radiation Pressure| Accurate Eclipse Prediction | ~                              |
| Atmospheric Model      | MSISE-86       | MSISE-86(1+ε)                   |

Here the orbital states derived from point-to-point TLE data are regarded as the near-real ones. Compared with them, both the accurate model and the simplified one cause certain errors $\Delta_{at}$ and $\Delta_{as}$ in orbital prediction, as shown in Figure 2. It uncovers that the accurate model contains an unknown deviation value $\varepsilon_a$ of the AMDM. Meanwhile, the simplified model with deviation contains known deviation value $\varepsilon_s$ which is added to the AMDM artificially. The purpose of this paper is to recover the deviation value $\varepsilon$ of the AMDM to reduce the certain errors $\Delta$ in orbital prediction and taking these results to improve orbital prediction.

C. DATA MINING METHODS

Data mining techniques can extract valuable information from volume disordered data. But the process is relatively complicated. Considering the characteristics of the research content and data in this paper, this section introduces three typical classification methods used to deal with orbit data and meanwhile validates their accuracy.

RF is a popular algorithm that combines multiple trees through a certain bridge. Figure 3 shows the visualization of a random forest model making a prediction. RF chooses $n$ instances from all instances and $m$ classes from all classes, randomly. It concludes the best partition properties to build Decision Tree (DT). It consists of a large number of individual DTs that operate as an ensemble. After repeating the above procedures $m$ times, $m$ trees could be obtained and each tree in the random forest represents a class prediction. Through voting mechanism, the class with the most votes becomes the final prediction.

ANN is a mathematical model similar to the structure of synaptic connection in the human brain. From an early viewpoint, the human brain is made up of nerve cells named “Neuron” which are connected through axon and dendrite. Similarly, ANN is made up of a group of nodes and directed chains. Figure 4 plots the sketch map of ANN Perceptron, where the perceptron is the simplest model of ANN, including multiple input nodes and one output node. These nodes are “Neuron” in ANN. Each input node is connected to the output node through a weighted chain. These weighted chains are “Synapse” in ANN. Same as the human brain, training a perceptron is to change the weight constantly till the perceptron could fit the connection of the input and the output.

This paper applies a feed-forward backpropagation network in orbital prediction. Feedforward represents that the neurons in each layer are only connected to the previous layer and output to the next layer, there is no feedback between the layers, and no adjustment is made to the parameters of the network. Backpropagation means that the error will be propagated back to adjust the weight and threshold. Backpropagation is a conventional training method for the feed-forward neural network during which the neurons adapt their weights to acquire new knowledge. A multi-layer neural network,
which has at least one hidden layer, could fit any target functions theoretically. However, ANN is much sensitive to noise and relatively time-consuming.

SVM is based on statistical learning theory, and the original intention of SVM is to build a hyperplane to segment multi-classes. Figure 5 is the sketch map of SVM. It shows that during its building process, SVM would try to maximize the interval between multi-classes.

![Figure 5. Classification algorithm of SVM: the circles and pentagrams are divided by SVM.](image)

There are three key points in SVM as max interval, kernel function and Lagrange dual theory. This paper uses rational quadratic kernel, which is a kind of radial basis function, as the kernel function after trying different kernel functions and it can be written as

$$k(x, y) = 1 - \frac{\|x - y\|^2}{\|x - y\|^2 + c}. \quad (10)$$

where $k(\cdot, \cdot)$ denotes the kernel function, $x, y$ are sample data vectors and $c$ is a constant. The rational quadratic kernel appears as a substitute for the Gaussian kernel, which has a better anti-interference ability for noise in the data, with less time consumption meanwhile. Besides, the comparison of different kernel functions is tabulated in Table 3 using the training data and accuracy definition from Section III.

### III. CLASSIFIER TRAINING AND EVALUATION

In this section of methodology, the two orbital dynamical models in Section II are used to conduct orbital prediction and generate training data starting from the same specific epoch. The accurate model gives a high-order description of the space environment with the perturbations and provides an environment without observation and measurement errors compared with TLE data. As for real application in the satellite control center, the low-order model is usually used to propagate the orbital information due to the existing initial errors. Although the processing procedure in this section is based on simulation, the data accuracy is ensured, higher than other observation data. So, in this simulation environment, the accurate model is assumed as the “actual” space condition and the simplified model is served to predict and propagate. It is meaningful and practical. Then the data are imported to the three classification methods. After setting proper parameters, three classifiers can be obtained and then their performance is compared.

TIANHUI is a Chinese remote sensing satellite that is successfully launched on August 24th, 2010. Its orbital altitude is around 500 km and atmosphere drag would cause considerable orbital decay. Therefore, this section chooses the real on-orbit state of this satellite as the starting point of orbital prediction.

### A. GENERATION OF TRAINING DATA AND PARAMETER SETTING

All the simulations are implemented by Matlab2017a on a PC with Intel® CORE i5-4590 processor and 8G RAM. The starting epoch of simulation is set at 21:07:05, May 31th, 2016 (UTC). The real on-orbit state of this satellite at the starting epoch is shown in Table 4, where $a$ is semi-major axis, $e$ is eccentricity, $i$ is orbit inclination, $\Omega$ is the right ascension of ascending node, $\omega$ is the argument of perigee, $M$ is mean anomaly, $r = [r_x, r_y, r_z]$ is the satellite position vector in inertial coordinate system, $v = [v_x, v_y, v_z]$ is the satellite velocity vector in inertial coordinate system.

| $a$ (km) | $e$ | $i$ (°) | $\Omega$ (°) | $\omega$ (°) | $M$ (°) |
|----------|-----|---------|--------------|-------------|---------|
| 6871.2093 | 0.0025 | 97.5464 | 285.7813 | 95.2957 | 21.1083 |
| $r_x$ (km) | $r_y$ (km) | $r_z$ (km) | $v_x$ (km/s) | $v_y$ (km/s) | $v_z$ (km/s) |
| -1611.4252 | 2731.7174 | 6681.6417 | -1.4261 | 6.6905 | -3.3722 |

The sample interval between two adjacent TLE data points of TIANHUI is less than 1 day. Since the TLE propagation accuracy is with respect to the propagation intervals and a five-day or less propagation is likely used [23], the propagation intervals in this paper are set as following: the minimal propagation interval of this point, $\Delta t_j$ is the prediction interval of this point, $\Delta$ means

$$X_j = \begin{bmatrix} \Delta t_j & a_j & e_j & i_j & \Omega_j & \omega_j & M_j & r_{xj} & r_{yj} & r_{zj} & v_{xj} & v_{yj} & v_{zj} \end{bmatrix}$$

(11)
the differences between the simplified model with deviation and the accurate model, and the subscript $j$ means the ordinal of this state. In conclusion, 181440 pairs of satellite on-orbit states can be obtained, and each state is corresponding to one known deviation value $\varepsilon_j$ in the AMDM, i.e. $[X_j, \varepsilon_j]$. The form of partial training data is shown in Table 5.

Table 5 indicates that the differences between orbit states obtained from different models are relatively small. Therefore, it is hard for a traditional statistical method to find out the relationship between satellite on-orbit state $X_j$ and known deviation value $\varepsilon_j$. This is the main reason why this paper applies the typical methods in data mining to process and classify these large amounts of data.

**TABLE 5. Partial training data examples.**

| $\varepsilon_j$ (day) | $\Delta \varepsilon$ (km) | $\varepsilon$ | $\omega$ | $\Omega$ | $\omega$ | $M$ |
|----------------------|---------------------------|-------------|--------|--------|--------|------|
| -50%                 | 0.301                     | 6871.102    | 0.025  | 97.565 | 286.288| 94.707| 239.007|
| -40%                 | 0.600                     | 6871.051    | 0.025  | 97.563 | 286.395| 93.301| 72.578 |
| -30%                 | 0.300                     | 6871.102    | 0.025  | 97.565 | 286.088| 94.701| 227.745|
| -20%                 | 0.600                     | 6871.048    | 0.025  | 97.563 | 286.393| 93.299| 73.216 |
| -10%                 | 0.500                     | 6871.142    | 0.025  | 97.561 | 286.293| 94.698| 241.763|
| 0%                   | 0.900                     | 6871.174    | 0.025  | 97.563 | 286.699| 93.709| 278.117|
| 10%                  | 0.800                     | 6871.061    | 0.025  | 97.565 | 286.597| 93.410| 87.899 |
| 20%                  | 0.200                     | 6871.072    | 0.025  | 97.564 | 285.986| 95.667| 39.952 |
| 30%                  | 0.100                     | 6871.096    | 0.025  | 97.563 | 286.688| 94.703| 228.378|
| 40%                  | 0.500                     | 6871.115    | 0.025  | 97.561 | 286.293| 94.097| 241.766|
| 50%                  | 0.600                     | 6871.836    | 0.025  | 97.563 | 286.395| 93.298| 73.220 |

In this section, the procedure for the application of data mining methods to recover the deviation of AMDM is as following: Firstly, four-fifths satellite on-orbit states are firstly regarded as the training data to obtain the classifiers in the guarantee of the data consistency. Then, the remaining one-fifth data are included to compare and evaluate the training results. It can reveal the generalization ability of the proposed methods. The classification model accuracy is defined respectively: in terms of RF and SVM, it is considered as a correct prediction when the classification prediction is identical to the artificial ones and the accuracy metric is scaled by the percentage of correct prediction; as for ANN, the accuracy metric is formulated by $|\varepsilon_p - \varepsilon_j| < 1 \times 10^{-3}$, where $\varepsilon_p$ is deviation value predicted by ANN and $1 \times 10^{-3}$ is the threshold due to $1 \times 10^{-1}$ as a minimum step.

**TABLE 6. Training parameter setting of RF.**

| Classification Model | Number of DT |
|---------------------|--------------|
| RF                  | 50           |

Three classification methods have been introduced in Section III, and the training data have been obtained from the above content. Some other parameter settings are tabulated in Tables 6-8 to complete the initialization of each classification method. The selection of parameters includes two aspects: 1) rough selection: according to the number of variables and the size of the data set, a reasonable parameter interval is determined through experience; 2) fine selection: by using the control variable method and constantly changing the single parameter value within its interval, the final choice is the value of the parameter with the highest model accuracy.

**TABLE 7. Training parameter setting of ANN.**

| Classification Model | Hidden Layers | Number of Neurons* | Training Function |
|---------------------|--------------|--------------------|------------------|
| Feed-forward Network | 5            | 20                 | TRAINLM          |
| Adaption Learning Function | Max Fail | Min Gradient | Performance Function |
| LERANGDM             | 10           | $1 \times 10^{-9}$ | Mean Square Error (MSE) |

*: Except the last layer which is output layer, other 4 layers each have 20 neurons.

**TABLE 8. Training parameter setting of SVM.**

| Classification Model | Kernel Function | Kernel Scale | Box Constraint Level | Multiclass Method |
|---------------------|-----------------|--------------|----------------------|-------------------|
| SVM                 | Quadratic       | Automatic    | 1                    | One-vs-One        |

![FIGURE 6. Results of RF: ideal results and classified ones are shown; a confusion matrix; b voting results of RF.](image-url)

**B. COMPARISON AND EVALUATION OF TRAINING RESULTS**

The results of RF are shown in Figure 6, where x-axis means the serial number of each on-orbit state, and y-axis means the
deviation value of this state predicted through RF. Because of the voting mechanism, the result in Figure 6a is obtained through choosing the value which has the maximum probability in Figure 6a. It is obviously indicated from Figure 6a that the color is deeper, the probability is larger. Beneficial from the randomness and the voting mechanism of RF, the final accuracy of the classification is up to 99.99%. It should be noted that the accuracy is obtained by calculating the ratio of correct classification value.

Figure 7 gives the training process and results of ANN. Figure 7a shows that during 7107 iterations, ANN continuously adjusts the weight to reduce the value of performance function, i.e. MSE, gradually. The minimal value of MSE is $9.2948 \times 10^{-7}$. Figure 7b shows the evolution rules of the gradient, mu (in TRAINLM) and validation fail during 7107 iterations. It also indicates that the validation fails in achieving its maximum value in the end and ANN stops its training process. Figure 7c presents the regression results where the $x$-axis means the known deviation value, while the $y$-axis means the predicted deviation value. If the small circles in Figure 7c are closer to the diagonal, the training results are better. Four subfigures, displaying “Training Set”, “Validation Set”, “Test Set”, and “ALL Data” respectively, show different results of different data partitions.

Similar to Figure 6, Figure 8 shows the results of ANN. The accuracy of the results has better reduced compared with RF. It should be noted that after training, ANN can generate new classes for data (i.e. new deviation values which are different from the known values), which is more obvious when the prediction time is small. While the results of RF only show the original classes. This ability of ANN is called generalization ability referred to as the networks producing reasonable “new” outputs for inputs during training. As shown in Table 5, the training data are discrete. But the results of ANN are nearly linearly varying. Moreover, Figure 9 demonstrates the errors between ideal results and the ones of ANN which indicate the accuracy of ANN. It shows more clearly that early on-orbit states with shorter prediction time possess larger
error. With the increase of the prediction time, the prediction errors reduce gradually.

Figure 10 shows the classification results of SVM. The accuracy of SVM is only 71.90%. The similarity of ANN and SVM is that when the prediction time is small, the differences between different models are small, and the results may have larger errors. Moreover, the results of ANN only fluctuate within a certain range, while the results of SVM fluctuate within the whole domain discretely. Moreover, SVM and RF will not generate new classes which demonstrates that ANN possesses the best generalization ability.

The comparison of the three methods is tabulated in Table 9. It indicates clearly that RF owns the shortest training time and the highest accuracy. Considering the number of DT is 50, properly reducing the number may shorten training time and meanwhile not reduce the accuracy. On the other hand, the results of ANN and SVM are not ideal. As for these two methods, the training data of on-orbit states and atmospheric deviation values used in this paper are hard to train. Notice that the training time of ANN and SVM is far beyond 200 times more than RF. Especially for ANN, this training method will occur overfitting and failure easily. This section chooses $1 \times 10^{-3}$ as a calculation standard. When the calculation standard is lowered, such as $1 \times 10^{-2}$, the accuracy of ANN may achieve more than 95%.

IV. APPLICATION OF CLASSIFIERS IN IMPROVING THE ACCURACY OF ORBITAL PREDICTION

From the perspective of methodology, Section III investigates the feasibility of the three methods in recovering the known deviation $\epsilon_j$ in the AMDM under the simulation environment. However, Section III is conducted under a pure simulation environment. In this section of practical application, new training data and application data are firstly generated by using the two dynamical models and historical TLE data. Three classifiers are retrained and used to process the application data. After recovering the unknown deviation, the results are implemented to improve the AMDM in accurate dynamical model in terms of different classification methods respectively.

A. GENERATION OF NEW TRAINING AND APPLICATION DATA

For better understanding, this section uses the change of the semi-major axis to show the process of data generation, shown in Figure 11. The blue points mean the orbit states...
FIGURE 12. Results of RF: ideal results and classified ones are shown; a confusion matrix; b voting results of RF.

calculated from historical TLE data. 704 TLE data points are chosen as the benchmark of simulation and there is no orbital maneuver among these points. The red points mean the orbit states calculated from the accurate dynamical model. The green points mean the orbit states calculated from the simplified model with deviation. It should be noted that the deviation value $\varepsilon_s$ in Figure 11 is equal to 300%. In order to calculate the differences between historical TLE data and different dynamical models to generate new training and application data, the prediction interval is the interval of adjacent TLE data points.

Figure 11 indicates that there are considerable prediction errors between historical TLE data and the dynamical models. Similar to Section III, three types of orbit states are recorded at every epoch corresponding to the accurate model $X_a$, the simplified model with deviation $X_s$, and historical TLE data $X_t$. The differences between $X_a$ and $X_s$ can be regarded as new training data, while the differences between $X_s$ and $X_t$ can be regarded as application data. The forms of partial training and application data are shown in Tables 10 and 11, respectively. Based on the actual engineering experience, the range of known deviation values in AMDM is extended to $[-300\%, 300\%]$ in this section. It is worth mentioning that the deviation values which are smaller than $-100\%$ mean that atmosphere causes a larger lift force than the drag force due to the satellite’s attitude.

B. RETRAINING RESULTS OF NEW CLASSIFIERS

Three methods are used to learn new training data and generate new classifiers. Some different settings are listed as follows to adapt to new training data and reduce computational burden: 1) the number of DT in RF reduces to 25; 2) the number of neurons in each layer increases to 30. The other parameter settings are in accordance with Tables 6-8. Retraining results are shown in Figures 12-14. It can be seen that both RF and SVM give results with high accuracy, while there are some discrete points in ANN results.
FIGURE 14. Retraining Results of SVM: ideal results and classified ones are shown.

TABLE 10. Partial new training data examples.

| $e_j$ (day) | $\Delta t$ (day) | $a$ (km) | $e$ | $i^\circ$ | $\Omega^\circ$ | $\omega^\circ$ | $M^\circ$ |
|------------|-----------------|----------|-----|----------|--------------|--------------|-----------|
| 100        | 0.3225          | 6871.1835| 0.0025 | 97.5640 | 286.1095 | 94.6685 | 349.1779 |
| 270        | 0.3325          | 6871.1808| 0.0025 | 97.5640 | 286.1095 | 94.6686 | 349.1784 |
| 230        | 0.2628          | 6871.2225| 0.0025 | 97.5585 | 286.3770 | 93.420 | 350.8553 |

TABLE 11. Partial application data examples.

| $\Delta t$ (day) | $a$ (km) | $e$ | $i^\circ$ | $\Omega^\circ$ | $\omega^\circ$ | $M^\circ$ |
|-----------------|----------|-----|----------|--------------|--------------|-----------|
| 0.3225          | 6871.1953| 0.0025 | 97.5639 | 286.1093 | 94.9526 | 348.8854 |
| 0.3028          | 6871.2059| 0.0025 | 97.5588 | 286.3771 | 92.9178 | 351.5796 |
| 0.4091          | 6871.1165| 0.0024 | 97.5593 | 286.7760 | 94.4643 | 22.6408 |
| 0.3821          | 6871.0855| 0.0025 | 97.5639 | 287.1625 | 93.3412 | 317.7268 |
| 0.5352          | 6871.1322| 0.0025 | 97.5611 | 287.8362 | 92.2625 | 21.6711 |
| 0.5243          | 6871.1661| 0.0025 | 97.5623 | 288.1631 | 91.8699 | 359.1118 |
| 0.3237          | 6871.1045| 0.0022 | 97.5618 | 288.4908 | 89.4844 | 335.2834 |
| 0.5375          | 6871.1314| 0.0024 | 97.5573 | 288.8264 | 91.2313 | 23.3758 |
| 0.3972          | 6871.1615| 0.0025 | 97.5623 | 289.1588 | 90.1748 | 28.9400 |
| 0.3366          | 6871.1368| 0.0025 | 97.5612 | 289.4950 | 87.9460 | 42.5073 |

TABLE 12. Statistic error values of each model and method.

| Statistic Values | Training Model | Accurate Model | Improved Accurate Model |
|-----------------|----------------|----------------|------------------------|
| $a$ (km)        | 0.0328         | 0.0478         | 0.0395                 |
| $e$ ($10^{-5}$)| 1.2575         | 2.7649         | 2.7642                 |
| $i^\circ$       | 0.0004         | 0.0027         | 0.0027                 |
| $\Omega^\circ$  | 0.0003         | 0.0030         | 0.0030                 |
| $\omega^\circ$  | 0.8020         | 1.7629         | 1.7635                 |
| $M^\circ$       | 0.8030         | 1.7602         | 1.7600                 |

The errors of ANN are also increased compared with the ones in Section III. One of the reasons is the larger data set and ANN possesses the generalization ability. Obviously, RF still possesses the highest accuracy.

C. APPLICATION ANALYSIS OF CLASSIFIERS

Three retrained classifiers are used to process the application data, respectively. Figures 15-17 show the application results and comparison of the application results of different methods, respectively. It is obvious that the predicted deviation values between the accurate model and historical TLE data range from $-300\%$ to $300\%$. Owing to the high accuracy, RF can recover relatively small deviation, while ANN and SVM can only recover partial values. Different from RF and ANN, the results of SVM present a type of stepped distribution.

The validity and accuracy of each method are not intuitive in Figures 15-17 which only present the recovering results of deviation values. It is necessary to prove the effects of...
the improvement in a realistic orbit prediction application. Therefore, the deviation errors in terms of different classification methods are added to the accurate model to improve the

AMDM respectively, namely the improved accurate model. Figures 18a-20a are the orbital prediction results using the improved accurate model corresponding to each method. Figures 18b-20b show the errors caused by the accurate models and improved accurate ones based on historical TLE data. Because of the uneven recording intervals and the small sample size of the TLE data used here, i.e. 2-4 TLE data samples a day released by NOARD, it can reveal the seasonal short-term variation instead of diurnal instantaneous one.

In order to qualify the improvement in orbit prediction of the three methods, the statistic error values of each model and method are given in Table 12. Standard deviations of the orbital element errors are regarded to scale the relative dispersion of AMDM errors. The second column is denoted the training model errors, i.e. the differences between $X_a$ and $X_s$. The third is the statistic value of application data. The others are the ones of improved accurate model errors using different classification methods. What’s more, this paper chooses a performance indicator, Difference of Mean Absolute Error (DMAE) which can be calculated through the
following Equation (12):

\[
DMAE = \text{mean}(|a_a - a_t|) - \text{mean}(|a_{imp} - a_t|)
\]  

(12)

where \(\text{mean}(\cdot)\) is the function for calculating the average value. \(a_a\) is the semi-major axis calculated by the original accurate model. \(a_t\) is the semi-major axis calculated by historical TLE data. \(a_{imp}\) is the semi-major axis calculated by the improved accurate model. As shown in Tables 12 and 13, the three classification methods in this paper can improve part of atmospheric density models inside the accessible data. Among the three improved accurate models, RF shows the greatest accuracy, which is analogous to the DMAE results.

V. CONCLUSION

This paper focuses on the orbital prediction of TIANHUI and data mining technology. Firstly, the mathematic models of the space environment are built, and two orbital dynamical models are proposed based on the detailed perturbations. The accurate model is used to simulate the “actual” space environment, while the simplified one with the known deviation is regarded as the basis of training data. Secondly, this paper introduces three classification methods, including their theories and characteristics. After the above two parts, this paper unfolds from two major aspects: methodology and practical application. In the stage of methodology, the two orbital dynamical models are used to generate training data starting from one specific epoch. And the three classification methods are used to train the data. The training results show that data mining technology can predict the known deviation values in AMDM. Especially for RF, the accuracy of RF can achieve 99.99\%, while the ones of ANN and SVM are only 86.83\% and 71.90\% respectively. In the stage of practical application, new training data and application data are generated based on historical TLE data. In terms of the new training data, three classifiers are retrained and used to figure out the errors. The application results show that all the methods can recover the unknown deviation of AMDM in the accurate model. Using the improved accurate model to conduct orbital prediction, the three methods show certain capabilities of improving orbital prediction, of which RF still possesses the best performance which maximum DMAE.

This paper provides novel thoughts and data mining methods for orbital prediction. The proposed methods improve part of atmospheric density models inside the accessible data based on the point-to-point historical TLE data, instead of replacing the traditional models. Its dependence on data makes it more updated. Its correction of empirical atmospheric density model offers a more reliable and accurate way of orbital propagation and extends the validity of atmospheric density model. Combining with popular data mining technology, this paper explores the application of the data mining method in traditional aerospace engineering and provides favorable reference. In the future work, based on more accurate orbital information, such as GPS post-processing data, a systematic two-level correction will be proposed: the first level is to improve the atmospheric density on specified heights from TLE data of limited LEO and the second one is to mine all the height span of atmospheric density from the results achieved in the first level. The correction of full atmospheric density model varying instantaneously as well as with more detailed perturbations are worth investigating.

REFERENCES

[1] I. H. Witten, E. Frank, and M. A. Hall, “What’s it all about,” in Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Cambridge, MA, USA: Morgan Kaufmann, 2011, p. 338.
[2] L. Self, “Use of data mining on satellite data bases for knowledge extraction,” presented at the 13th Int. Florida Artif. Intell. Res. Soc. Conf., Orlando, FL, USA, May 2000.
[3] T.-C. Fu, “A review on time series data mining,” Eng. Appl. Artif. Intell., vol. 24, no. 1, pp. 164–181, Feb. 2011, doi: 10.1016/J.ENGAPPAI.2010.09.007.
[4] F. M. Bianchi, A. Rizzi, A. Sadeghian, and C. Moiso, “Identifying user habits through data mining on call data records,” Eng. Appl. Artif. Intell., vol. 54, pp. 49–61, Sep. 2016, doi: 10.1016/j.engappai.2016.05.007.
[5] N. Pillay, R. Qu, D. Srinivasan, B. Hammer, and K. Sorensen, “Automated design of machine learning and search algorithms [guest editorial],” IEEE Comput. Intell. Mag., vol. 13, no. 2, pp. 16–17, May 2018, doi: 10.1109/MCI.2018.2806988.
MING XU received the B.S. and Ph.D. degrees in aerospace engineering from Beihang University, Beijing, China, in 2003 and 2008, respectively. He was an Engineer in orbital design and operation with DFH Satellite Company Ltd., China Academy of Space Technology, Beijing, in 2010. He joined Beihang University, as an Assistant Professor, where he was promoted as an Associate Professor, in 2012. He has published 50 publications in journals, books, and proceedings. His current research interests include applications of dynamical systems theory into astrodynamics and orbital control. He received the National Top 100 Excellent Doctoral Dissertation Award Nomination, in 2010, and the Third Class Prizes of the National Defense Technology Invention Award, in 2016. He serves as an Associate Editor for the journals of Astrodynamics and Advances in Aircraft and Spacecraft Science.

YARU ZHENG received the M.S. degree in aerospace engineering from Beihang University, Beijing, China, in 2020. Her research interests include vehicle trajectory control and inertial navigation.