Usage of the machine learning to organize time series and find anomalies

A.S. Kopyrin[1,*], *E.V.* Vidishcheva[1], and *Yu.I.* Dreizis[1]

[1]Sochi State University, 94, Plastunskaya str., Sochi, 354000, Russia

Abstract. The subject of the study is the process of collecting, preparing, and searching for anomalies on data from heterogeneous sources. Economic information is naturally heterogeneous and semi-structured or unstructured. This makes pre-processing of input dynamic data an important prerequisite for the detection of significant patterns and knowledge in the subject area, so the topic of research is relevant. Pre-processing of data is several unique problems that have led to the emergence of various algorithms and heuristic methods for solving such pre-processing problems as merging and cleaning and identifying variables. In this work, an algorithm for preprocessing and searching for anomalies using LSTM is formulated, which allows you to consolidate into a single database and structure information by time series from different sources, as well as search for anomalies in an automated mode. A key modification of the preprocessing method proposed by the authors is the technology of automated data integration. The technology proposed by the authors involves the joint use of methods for building a fuzzy time series and machine lexical matching on a thesaurus network, as well as the use of a universal database built using the MIVAR concept. The preprocessing algorithm forms a single data model with the possibility of transforming the periodicity and semantics of the data set and integrating into a single information bank data that can come from various sources.

1 Introduction

Nowadays, data analysis and decision-making based on it are critical in many economic and social areas. Besides, as more data is generated, collected, and analyzed on an ever-increasing scale, there is an increasing need for methods to purify source information and detect knowledge based on it. Poor data quality is a serious problem as it is often created automatically, entered manually, or integrated from disparate and heterogeneous sources. Thus, before data is analyzed, it must be pre-processed to correct errors, typos in raw data, and convert raw data into homogeneous data [1], making it usable. This process is both time-consuming and tedious.

According to a study [2] more than half of analysts believe that data cleaning and preparation take more than 60% of the total analysis time. The quality of preprocessing results affects the result of pattern detection and analysis [3].

Economic information is heterogeneous and semi-structured or unstructured in nature.

* Corresponding author: kopyrin_a@mail.ru

Due to the heterogeneity of the primary documents and the human factor, the original statistical data may contain a large amount of noise, as well as records, the automatic processing of which can be very difficult. This makes pre-processing of dynamic input data an important precondition for discovering meaningful patterns and knowledge in the domain.

The purpose of preprocessing is to transform raw data from various sources into a set of fuzzy time series that have an ordered structure. Data preprocessing represents several unique tasks that have led to the emergence of various algorithms and heuristic methods for solving preprocessing tasks such as merging and clearing, variable identification, etc.[4].

In this paper, we formulate a preprocessing algorithm that allows combining into a single database and structure information on time series from different sources.

2 Materials and methods

Modern data processing methodologies involve extensive use of machine learning. Between machine learning and traditional statistical methods, there are significant differences. First, machine learning focuses on the task of "forecasting", using universal learning algorithms to find patterns in variable and volumetric data [5]. In contrast, statistical methods are mainly focused on confirming hypotheses based on inferences, which are achieved by defining and approximating probability functions for a particular model [6].

Second, most machine learning methods do not contain hypotheses, since their goal is to detect non-obvious associations in the data, whereas traditional statistics usually rely on certain assumptions and hypotheses, often those that derive from the model that generated the data [7].

Third, the toolsets used to evaluate errors in building a machine learning model, are usually different from those of traditional statistical methods, mainly rely on the calculation of p values for the zero hypotheses [6, 8].

Fourth, traditional statistical modeling typically provides an easy-to-understand model, that represents a result that is easy to understand and interpret.

However, the factors of the real economy are usually not independent of each other, and their relationships can be nonlinear. Machine learning approaches, in turn, consider all possible interactions between variables following multidimensional nonlinear patterns, regardless of the degree of complexity, while trying to capture as many informative and potential features and correlations as possible. This leads to the fact that the model created using these methods is not easy to understand or interpret. The fundamental goal of machine learning is a generalization that goes beyond the examples in the training sample. Generalization is possible because models are based on a much larger data set, then checked in an independent data set, and then configured to achieve the best performance [9].

Historically the concept of detecting useful patterns in data has been given many names such as data mining, knowledge extraction, data archaeology, and pattern processing, etc.

Many data processing tasks assume that data corresponds to ideal distributions without any (or if there is a very slight amount) missing, incorrect, or inconsistent values. However, this happens quite rarely. Most existing data cleaning methods use relational models, which assume that data is in a structured table format. However, structured datasets (or organized datasets, we use these terms interchangeably) do not always have a relational format. Many datasets, such as XML, can be described using markup languages and templates. This necessitates the development of a more general theory capable of handling structured datasets, whether the dataset is a table or other structured format [10].

Raw data is usually not in a structure that is convenient for researchers to work with and is not enough organized.

The term "data preprocessing" refers to any data transformation before the training algorithm is applied. Data preprocessing addresses a variety of data quality issues associated with outliers and noise. The main purpose of removing outliers is to find data objects that are irrelevant or only weakly relevant to the basic key relationship analysis. The focus at this stage is to remove objects that prevent data analysis. Cluster analysis is similar to detecting outliers since both methods deal with eliminating poorly relevant or irrelevant objects.

This includes searching for examples and correcting inconsistencies; replacing missing values; identifying, deleting, or replacing outliers; discretizing numeric data or generating numeric dummy variables for categorical data; reducing dimensionality; extracting/selecting features; and scaling features (normalization, standardization, or Box-Cox conversion) [11]. It is also assumed that scaling objects through standardization (or normalization of the Z-score) is an important step in preprocessing for many machine learning algorithms.

Predictor variables with ranges of different orders of magnitude can have a disproportionate impact on results. In other words, predictor variables with a large scaling range may dominate in the context of the algorithm. Scaling feature values implicitly provides equal weights for all features in their representation and should be a preprocessing approach used in machine learning algorithms such as decision tree, linear regression, SVM, and others [8].

3 Results

As shown in Figure 1, the general process includes data cleanup, integration and unification, transformation to a single form, and converting and finding outliers for further processing by intelligent methods.

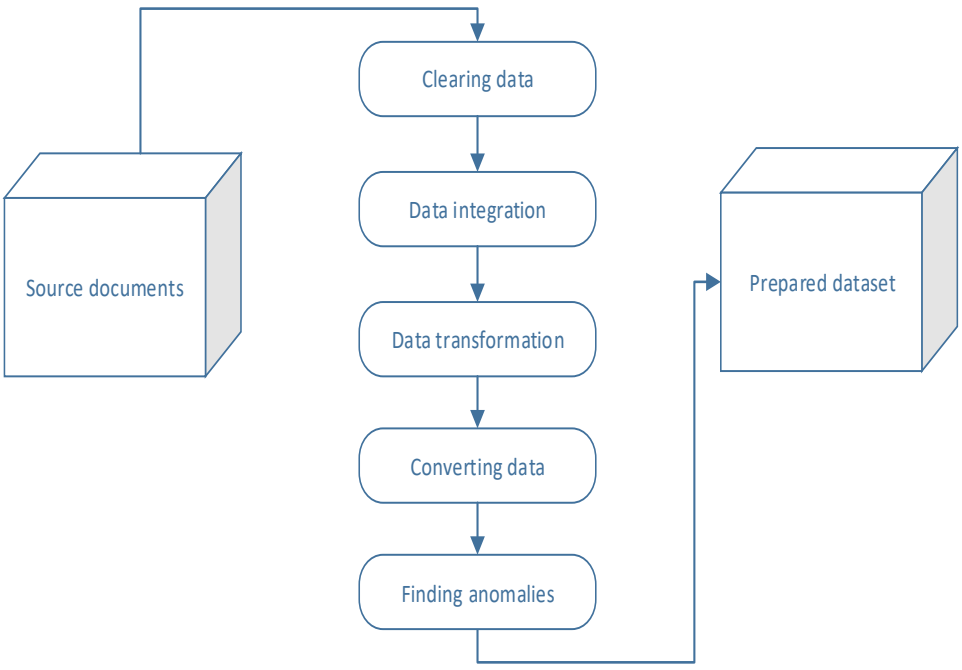

Fig. 1. General scheme of the preprocessing algorithm.

Let's look at each of these stages in more detail.

3.1 Clearing data

Data from primary documents (economic indicators, currency rates, etc.) which are incomplete, noisy, or inconsistent should be improved by filling in the default values, smoothing out the noise, and correcting data inconsistencies.

a. Eliminating incompleteness of the time series. When data is collected, some data attributes may be lost due to human factors and/or a failure of the information collection and recording system. There are several ways to get around this problem. Missing values can be ignored, filled in with some values, given by default, or use the average attribute values to fill in. You can also insert the most likely values in the missing records using a statistical correlation model.

If, when the missing value has a large impact on the treatment process, the missing data is usually ignored. For example, when processing information about a currency transaction, if the currency name and identifier are lost, the data must be ignored; however, if the transaction time is skipped and there is a date, the data cannot be ignored. If the data set is relatively small, the default values can be filled in manually. However, when working with large sets with a large number of values, this method is not advisable, since it is laborious and expensive.

In case data distribution is uniform, the default values can be filled in with average values of the attributes. Besides, you can use machine learning methods to determine the optimal value for default data, including regression, Bayesian formal methods, and decision tree induction. Although the forecast may show a relatively large deviation in extreme cases, these methods are still better able to recover missing data. Thus, the user must define a method (or a combination of both) to recover missing data before downloading information from the source.

b. Noise in this context refers to a deliberately incorrect attribute value in the data source that goes beyond the set of acceptable values for this property. For example, such data can be attributed to the negative consumption of utility services or sales of excess inventory and current production. Processing noisy data includes binning [12], regression search, and outlier analysis.

Binning methods smooth out ordered data values by examining the values around the data. The key parameter of binning methods is the size of the nested field. The regression method consists of changing the noise value by configuring the functional model corresponding to the value of the data attribute. Outlier analysis consists of constructing clusters using the clustering method. The attributes of data within the same cluster are similar, but the values of data point attributes between different clusters may have a significant deviation [13].

c. Bringing data to a consistent view. In different sources, data about the same attribute may contain inconsistencies, for example, in units of measurement or the frequency of values. Data inconsistencies can be corrected by analyzing the correlation between the data and automatically applying data translation functions from one periodicity to another.

3.2 Data integration

At this stage, data stored in different sources have to be combined, and the task is to unify heterogeneous data and resolve possible redundancy. By integrating data, you can improve the accuracy and speed of data mining.

a. Unification of heterogeneous data. Since economic data can be collected from different sources, the overall set will have problems with the heterogeneity of content. The specified problem is mainly represented by inconsistencies in data attributes, such as attribute names

and units of measurement. For example, the cost indicator can be called "Costs", "Expenses", and the unit of measurement can be rubles, thousand rubles. USD, etc.

b. Processing redundant data. In simplified terms, if an attribute can be obtained from other attributes, then it is redundant and should be cleared. Redundancy is mainly reflected in duplicate data attribute entries or inconsistencies in the way attributes are expressed. Most redundant data can be found using correlation analysis. For numerical data commonly used method of analysis metrics chi-square test.

The key modification of the preprocessing method proposed by the authors is the technology of automated data integration.

For the automated solution of these problems, the authors propose the combined use of methods for constructing a fuzzy time series and machine lexical comparison on the thesaurus network, as well as using a universal database built using the MIVAR concept [14].

According to [15] a fuzzy time series can be defined as follows:

The fuzzy set A of the universal set U, U={u1, u2,..., un}, has the form

$$A = f\,A(u1)/u1 + f\,A(u2)/u2 + \cdots + f\,A(un)/un \quad (1)$$

where fA is the membership function of the fuzzy set A, fA: U→[0, 1], fA(ui) represents the degree to which Ui corresponds to the fuzzy set A, and $1 \le i \le n$.

Let Y(t)(t=..., 0, 1, 2, ...), a subset of R (real time), be the universal set on which fuzzy sets fi(t)(i =1, 2, ...) are defined and let F(t) consists of f1(t), f2(t),

Then F(t) is called a fuzzy time series defined on Y (t) (t=..., 0, 1, 2, ...)

There are two types of fuzzy time series models: time-varying and time-invariant. If for any time t R (t, t-1) does not depend on t, then F (t) is called a fuzzy time series invariant in time. Otherwise, it is called a time-varying fuzzy time series [16].

Imagine a non - integrated time series as a time-varying fuzzy time series, where, on the one hand, the distribution of time intervals of observations changes (since data from different sources vary), and on the other hand, different values of the series are more or less related to the phenomenon under study (the semantics of variable names differs). The integration problem is proposed to be solved using machine learning as follows: the graph semantic network Wordnet using fuzzy algorithms determines the function of belonging of each variable to the reference value [17]. This function is defined as the degree of proximity of the semantics of the studied variable to the reference one, where 1 – equivalent concepts, 0 – unrelated categories.

All obtained variables are written into the knowledge base developed by the authors with a structure (object, property, relations) [18], forming a data set with set primary relationships. With further implementing the algorithm, a confidence coefficient in the interval [0, 1] is requested from the user, and for subsequent transformation, an integrated fuzzy time series is also extracted from the knowledge base, the membership function of the values of which is higher than the specified coefficient.

3.3 Data transformation

Based on the premise of data integrity, a data transformation involving data reduction can reduce the size of the data set, which will increase the usability and efficiency of intellectual data mining. In China, a large amount of economic activity data will be generated every day. Given the circumstances, data reduction is necessary to improve the efficiency of the analysis. Data reduction methods include reducing the size of the data set, reducing the number of observations, and data compression.

The data set size reduction method usually controls the size by reducing the number of random variables or attributes. It suggests the use of wavelet transform I and the method of

analysis and principal component analysis, which project the original data in a smaller set of variables.

3.4 Converting data

Data conversion involves converting a data set into a single form suitable for data mining. Data conversion methods include secondary noise smoothing, data aggregation, and data normalization. In accordance with the direction and purpose of data mining, the data transformation method filters and summarizes data, stored in the time series knowledge base. Data analysis can be more effective if there is a directed, purposeful aggregation of data.

To avoid data attributes depending on units of measurement, the data must be normalized so that the values have a single range of acceptable values, for example, from [0, 1]. There is linear and nonlinear normalization. The application of these techniques is particularly important if further analysis involves the use of neural network algorithms or classification algorithms based on distance measurements (such as the k -nearest neighbor method).

3.5 Finding anomalies

After bringing a fuzzy time series to a stationary appearance, machine methods for finding anomalies can be applied to it, for example, using a Long Short-Term Memory (LSTM) neural network, which is a modified version of a recurring neural network (RNN) that facilitates storing past data in memory [19]. In such networks, there is no problem with the disappearing gradient. LSTM is well suited for classifying, processing, and predicting time series based on time delays of unknown duration. LSTM networks, like conventional RNN, are trained using reverse propagation. when analyzing time series, the LSTM network will efficiently process unstructured statistical information, making it suitable for big data analysis. LSTM networks have been used in similar tasks, for example, in [20].

4 Conclusion

Thus, the task of pre-processing data before subsequent machine learning is very relevant. The specified processing or preprocessing should aim at collecting and integrating data from various sources into a single database.

The practical differences encountered by the authors in conducting empirical research are as follows:

- Different input data formats.
- Different semantic content – different names of variables.
- Different periodicity of the data (day, week, month, etc.).

The authors developed and proposed a modified technique for data integration in the process of pre-processing and data unification.

The proposed algorithm for preprocessing makes it possible to form a unified data model with the possibility of transforming the periodicity and semantics of the data set.

The following processing steps are suggested:

- Connecting various data sources (csv, xlsx, odbc, ado drivers, etc.).
- Forming a loadable time series based on a sample (name- period-value – period type).

Recording of the generated series into the universal knowledge base of the initial time series (records are formed in the following frames: FrObject, FrProperty, FrDescription, for more details about the structure of the knowledge base, see [18]).

Fuzzy time series processing and writing to the knowledge base:

4.1. If there are differences in the semantics of a variable (object name or properties), relationships with current records are established using the thesaurus and linguistic graphs (Wordnet).

4.2 Selection of a translation algorithm with a different periodicity for converting a fuzzy time series to an invariant form.

5. Summarize a single time series and record it into the knowledge base.

The use of this algorithm makes it possible to improve the quality of analyzed data in comparison with traditional statistical methods and reduce the labor intensity of the preliminary stage of intellectual analysis.

6. The LSTM anomaly search method [21, 22] applies to an ordered data set

As a starting point, we consider a network in which there is an input layer with 1 input, a hidden layer with 4 blocks or LSTM neurons, and an output layer that predicts one value. For LSTM units, use the sigmoid activation function. The network will be trained for 1000 eras. The error will be calculated by the standard method, and the ADAM algorithm is used as the gradient descent optimization method. In the future, network parameters can be adapted to solve more specific and specific problems.

The obtained trained model can be used to generate forecasts for samples, but this is not the only possible approach, since, given the stochastic essence of the problem, in some situations, some methods are better suited, and in some - others.

We may also consider using several recent samples to predict the next time step, that is, applying the sliding window method to the task by selecting the optimal window type and size for the input data.

The reported study was funded by the Russian Foundation for Basic Research (RFBR), according to the research project No. 19-01-00370

References

1. Rahm E, Do H 2000 *IEEE Data Eng Bull* DOI: 10.1145/1317331.1317341
2. Han Q, Gao X, Wu W 2008 *9th International Conference on Computer-Aided Industrial Design and Conceptual Design: Multicultural Creation and Design - CAIDCD* DOI: 10.1109/CAIDCD.2008.4730759
3. Anand S, Aggarwal R R *Int J Comput Appl* **48**
4. Gama J, Ganguly A, Omitaomu O, et al 2009 *Int Data Anal* DOI: 10.3233/IDA-2009-0372
5. Kim K-J, Tagkopoulos I 2019 *Korean J Intern Med* **34** 708
6. Bzdok D, Altman N, Krzywinski M 2018 *Nat Methods* **15** 233
7. Waljee A K, Higgins P D R 2010 *Am J Gastroenterol* **105** 1224
8. Kuhn M, Johnson K 2013 *Appl predictive model* DOI: 10.1007/978-1-4614-6849-3
9. Handelman G S, Kok H K, Chandra R V, et al 2018 *J Intern Med* **284** 603
10. Khedri R, Chiang F, Sabri K E 2013 *Procedia Computer Science* 50
11. Kotsiantis S B, Kanellopoulos D, Pintelas P E *World Acad Sci Eng Technol Int J Comput Electr Autom Control Inf Eng*.
12. Xiong X, Dubin J A 2010 *Stat Med. Epub ahead of print* DOI: 10.1002/sim.3953
13. Vidishcheva E V, Kopyrin A S, Vasilenko M S 2019 *J of Altai Acad of Ec and Law* **6-1** 41

14. Ivanchenko N O *MIVAR technologies modelling of enterprise's technical and technological potential. Actual Probl Econ*
15. Bose M, Mali K 2019 *Int J Approx Reason* DOI: 10.1016/j.ijar.2019.05.002
16. Song Q, Chissom B S 1993 *Fuzzy Sets Syst.* Epub ahead of print DOI: 10.1016/0165-0114(93)90372-O
17. Zhu X, Yang X, Huang Y, et al 2020 *Knowl Inf Syst.* DOI: 10.1007/s10115-019-01387-6
18. Kopyrin A, Vidishcheva E, Makarova I 2020 *Advances in Intelligent Systems and Computing* DOI: 10.1007/978-3-030-37919-3_82
19. Hochreiter S, Schmidhuber J J 1997 *Mem Neural Comput*
20. Malhotra P, Vig L, Shroff G, et al 2015 *23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN*
21. Vasilenko M S, Kopyrin A S 2019 *Model Artif Intell* **2** 13
22. Dawoud A, Shahristani S, Raun C 2019 *Proceedings - International Conference on Machine Learning and Data Engineering, iCMLDE 2018* DOI: 10.1109/iCMLDE.2018.00035