Detection Method of Duplicate Data in Software Testing Based on BP Neural Network

Chang'an Pan and Xiaozhou Chen
Quanzhou University of Information Engineering, Quanzhou, China

Keywords: BP Neural Network, Software Testing, Repetition Testing Data.

Abstract: Most of the conventional duplicate data detection methods are designed based on the principle of proximity sorting algorithm, which has strong limitations in use. In software testing, in the face of massive data, the detection recall rate of this method is low, and it cannot comprehensively detect duplicate data. Based on this, BP neural network is introduced to carry out the research on the detection method of duplicate data in software testing based on BP neural network. First, the similarity of software test repeated data is matched to determine whether two records are similar to each other. Secondly, a BP neural network model is constructed to preprocess duplicate data through iterative training of the model to reduce the impact of data redundancy interference. On this basis, the software testing process, similar or duplicate data records for all-round detection. The experimental test results show that after the application of the new method, the recall rate of duplicate data detection in software testing has reached more than 98% when the amount of data to be detected has gradually increased.

1 INTRODUCTION

In the current trend of collaborative development of Internet information technology, computer technology and various intelligent technologies, a large amount of electronic data has been generated, and the role of data in various fields has become more prominent, especially in the field of software development and testing (Valstar N, 2021). Based on the analysis from a broad perspective, software testing refers to the comprehensive operation of various programs under the specified conditions, and then discover the abnormal problems and errors in the program operation, so as to achieve the goal of software development quality assessment (Albayrak O S, 2021). Software testing is the sum of software quality and software characteristic testing, not only to verify the correctness of software operation process, but also to find out more problems in software operation as far as possible, analyze and track the problems, and ensure that the software can meet the needs of users (Rocha T M, 2021). Since most software is composed of documents, data and programs, the main object of software testing is not only the software program itself, but also the documents, development data and running programs (Alberto Gascón, 2022)covering the entire software formation process. Among them, in software testing, there are a large number of test duplicate data affected by the test hardware environment conditions and software environment conditions. If the duplicate data is not detected and cleaned in time, it is easy to interfere with the software test results and reduce the credibility of the test results (B Q Z A, 2021).

Based on this, scientific and reasonable software testing duplicate data detection methods are needed. With the continuous development of science and technology, through the efforts of many scholars, duplicate data detection methods are gradually mature and can quickly detect some duplicate data (Kremezi M, 2021). However, the traditional duplicate data detection method has strong limitations, relatively speaking, it is not suitable for the detection of duplicate data in software testing (Zhou Y, 2021). The main reason is that massive data will be generated in the process of software testing. The detection level of traditional methods is limited. When facing massive data, its detection recall rate is low, and it cannot detect duplicate data in multiple dimensions, which cannot provide strong data support for software testing (Zhang Y, 2021). As a combination of universal model and error correction function, BP neural network can improve the shortcomings of traditional duplicate data detection methods (Cecilia A, 2021). Through the

Pan, C. and Chen, X.
Detection Method of Duplicate Data in Software Testing Based on BP Neural Network.
DOI: 10.5220/0012285900003807

In *Proceedings of the 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (ANIT 2023)*, pages 458-463
ISBN: 978-989-758-677-4

establishment of BP neural network, the results obtained from each training are comprehensively analyzed, compared with the expected results, and then the training weights and thresholds are constantly modified, so as to obtain a model that can output consistent with the expected results, providing an important reference basis and support for the detection of duplicate data in software testing (Zhong J, 2022). Based on this, this paper introduces BP neural network on the basis of the current traditional duplicate data detection methods, and carries out an all-round research on the duplicate data detection methods of software testing based on BP neural network, which makes certain contributions to promoting the rapid development of the software testing field.

2 THE RESEARCH OF SOFTWARE TESTING DUPLICATE DATA DETECTION METHOD BASED ON BP NEURAL NETWORK

2.1 Software Test Repeated Data Similarity Matching

In the software test duplicate data detection method based on BP neural network designed in this paper, first, it needs to match the similarity of software test duplicate data, take one duplicate record data as a whole, and then judge whether two records are similar to each other by calculating the overall similarity of two duplicate record data (Liu C, 2021). Based on the characteristics of data changes during software testing, this paper uses a weighted algorithm based on word frequency and inverse document to match the similarity of repeated data of software testing (Zhang R, 2022). The similarity matching algorithm flow designed in this paper is shown in Table 1.

Table 1: Weighting algorithm flow based on word frequency inverse document.

No	Step
1	Segmenting the content of the fields that need to be matched to obtain mutually independent software test words.
2	Assign weights to each software test word, record the frequency of word occurrences, and divide the total number of records by the number of records containing independent words, which is the inverse document frequency.
3	Convert the duplicate data fields of the software test to be matched into vectors A and B.
4	Calculate the cosine similarity of the transformed vector.
5	The closer the vector cosine similarity value is to 1, the higher the similarity. Compare the vector cosine similarity value with the given threshold to determine the similarity of duplicate data records in software testing.

As shown in Table 1, it is the algorithm flow of software test repeated data similarity matching, where the vector cosine similarity calculation formula of step 4 is:

$$COS\theta = \frac{\sum_{i=1}^{n}(A_i \cdot B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

Among them, A_i、B_i represent the transformed vectors respectively A and vector B; n indicates the frequency of words that are independent of each other; i indicates the assigned weight of independent words. Through calculation, the cosine similarity of the vector after the repeated data field of software test is obtained. The closer the calculated value is to 1, the higher the similarity of software test repeated data is. If the value is greater than the judgment threshold, it will be judged as a duplicate record, and vice versa (Lakhmiri D, 2022).

2.2 Preprocessing Duplicate Data Based on BP Neural Network

After completing the similarity matching of repeated data in software testing, the next step is to build a neural network model based on the principle of BP neural network, preprocess the repeated data in software testing, and provide strong support and guarantee for subsequent testing. The quality of duplicate data detection results depends not only on the detection algorithm, but also on the software test data used. Reasonable pre-processing of the data can avoid misleading the subsequent detection caused by redundant data and missing data (Kiersztyn A, 2021).

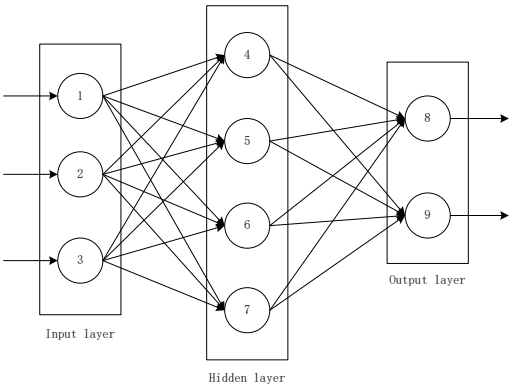

Figure 1: Basic structure of BP neural network model.

Because BP neural network has good nonlinear expression ability and generalization ability, it can better solve the problems of difficult selection of data set features and poor adaptability. Therefore, this paper constructs a BP neural network model, whose basic structure is shown in Figure 1.

As shown in Figure 1, the BP neural network model built in this paper belongs to a fully connected neural network structure, which is composed of three hierarchies: Input layer, Hidden layer, and Output layer. The Input layer is mainly responsible for receiving nonlinear input messages in software testing; As the connection of the other two hierarchies, the hidden layer is composed of one or more layers, which is responsible for undertaking the other two hierarchies and has strong linear simulation capability; The output layer is mainly responsible for transmission, analysis, trade off and result generation (Lv F, 2021). The function of any neuron in the neural network model is to receive the weighted input from other neurons, then combine its own threshold (bias), and finally get the output result (Song D, 2021) through the processing of nonlinear function. In order to improve the quality of iterative training of BP neural network model, this paper adds nonlinear factors to the model, increases the nonlinearity of the model, maps the repeated data of software test into the logical space, and the mapped data is relatively more helpful for repeated data preprocessing(Fellah A, 2021). On this basis, based on the BP neural network model, the software test duplicate data is preprocessed to remove unqualified dirty data, and generate and output qualified similar duplicate data. The main methods of data pre-processing include data cleaning, filling in missing values, smoothing noise data, identifying or deleting outliers, and solving inconsistent problems (Rani S, 2021). Data integration, which combines data from various sources into a unified format (Jf A, 2022). Data transformation, which converts the data that does not meet the data type requirements to the corresponding data type (Dehnen G, 2021).

2.3 Software Test Similar Duplicate Data Record Detection

After the above repeated data preprocessing based on BP neural network is completed, on this basis, some similar or duplicate data records are comprehensively detected during the software testing process, and similar duplicate data records are cleared, so as to achieve the goal of improving the quality of software test data. The software test designed in this paper is similar to the duplicate data record detection process, as shown in Figure 2.

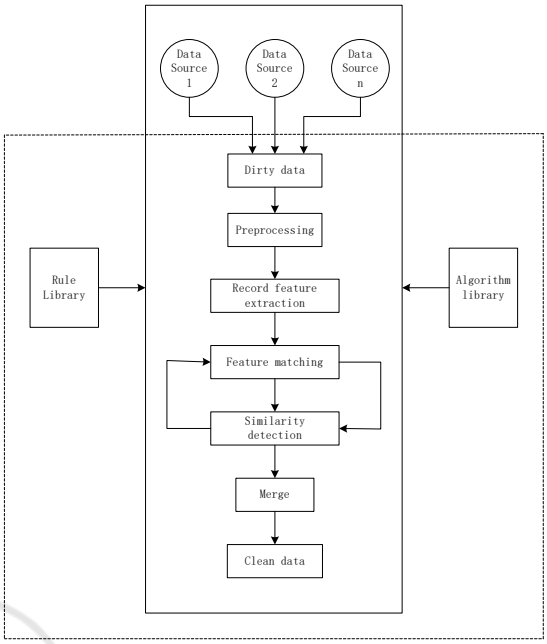

Figure 2: Software Test Similar Duplicate Data Record Detection Process.

As shown in Figure 2, first, input the software test data source 1, data source 2 and data source n respectively, extract the dirty data in the input data source, preprocess the dirty data under different data sources through certain rules and algorithms, and upload them to the storage system (Staowska P, 2022). Secondly, based on the software testing requirements, select the feature extraction rules with high matching degree with the software testing, establish the corresponding feature extraction model, and then conduct feature extraction (Ahlawat A, 2022)on the dirty data records after the above preprocessing through the iterative training of the model. It mainly extracts attribute keywords from similar duplicate records, calculates the weight of keywords, and sorts the weight values to provide basic support for subsequent similar duplicate data detection (Wulkan R W, 2021). On this basis, use sorting rules or algorithms to sort the records' keywords, and call relevant algorithms to detect the similarity and repeatability of records in the nearest neighbor range according to the sorting results of the records' keywords, so as to realize the calculation of the similarity between records (Fels-Klerx H J V D, 2021). Finally, the similarity repeatability of records is determined according to the predefined similarity detection rules. In order to improve the detection accuracy, it is sometimes necessary to carry out multiple rounds of sorting comparison based on

different keywords. In order to improve the detection efficiency, it is sometimes necessary to design a corresponding record retrieval strategy. According to the application requirements and specific rules, the similar duplicate records detected are merged or cleared to complete the processing of similar duplicate records. Finally, set the evaluation criteria for similar duplicate data records in software testing, and determine whether the duplicate data detection results meet the corresponding standards and specifications according to the criteria. Considering the actual situation of software testing and the characteristics of duplicate data, this paper selects the accuracy of duplicate data detection P as the evaluation standard, the calculation expression is:

$$P = \frac{n}{N} \times 100\% \quad (2)$$

Among them, n indicates the number of similar duplicate data records in the detected software test similar duplicate data records; N indicates the number of similar duplicate data records of software test detected. Evaluate and analyze the above duplicate data detection results through the evaluation criteria. If the detection results meet the evaluation criteria, output the software test duplicate data detection results to complete the detection task; If the test results do not meet the evaluation criteria, repeat the above test process steps until the evaluation criteria are met, and improve the quality of duplicate data detection in an all-round and multi-dimensional way.

3 APPLICATION TEST

3.1 Test Reparation

Table 2: Specific Description of Experimental Data Set.

Data set	Project	Explain
Movies Dataset	Data volume	The initial total includes 3480 XML pieces of data.
	Storage Content	Movie Information
	Access	Dirty X ML Generator (ttps://hpi.de/naumann/projects/completed-projects/dirtyxml.htm) External library.
	Duplicate Object	150
FreeCD Dataset	Data volume	10000 target data
	Storage Content	Singer Record Information
	Access	htp://www. feedb.org/ website.
	Duplicate Object	350

The above content is the whole process of the software testing duplicate data detection method designed by using BP neural network. Before the proposed data detection method is put into actual software testing application, its feasibility and detection effect need to be objectively tested and analyzed to ensure its duplicate data detection effect before it can be put into practical application. First of all, select the data set required for this experimental test, so that the experiment is conducted on two different data sets. The two selected data sets are Movies data set and FreeCD data set with certain similarity. The description of the two data sets is shown in Table 2.

As shown in Table 2, this is a detailed description of the data set used in this experimental test. The data mode of the two data sets is shown in Figure 3.

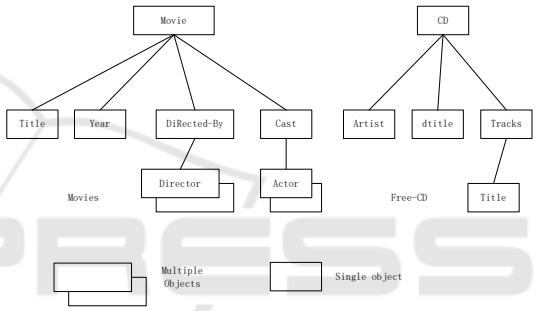

Figure 3: Schematic diagram of data mode of experimental test data set.

As can be seen from Figure 3, the Movies dataset contains a single object and multiple objects, while the FreeCD dataset contains only a single object. This paper sets the repetition threshold as 0.7. Only when the repetition probability of two objects is greater than or equal to 0.7, can two objects be considered as duplicates. All the values in the two data sets used in the experiment are of text string type. Set the environment for this experiment test, including hardware environment and software environment. Hardware environment: Intel Xeon (R) CPU 3.1GHz, memory size 4.00GB, hard disk size 500GB. Software environment: Windows10, MySQL 5.1, Eclipse.

3.2 Test Results

In order to present the results of this application test in a more intuitive and clear form, this paper introduces the method of comparative analysis, sets the above software test duplicate data detection method based on BP neural network as the

experimental group, and sets the traditional duplicate data detection method based on Dogmat iX as the control group, and analyzes and compares the detection effects of the two methods respectively. Set the recall rate of duplicate data detection in software test as the evaluation index of this experiment, and its calculation expression is:

$$R_m = \frac{T}{D} \times 100 \quad (3)$$

Among them, R_m represents the recall rate of duplicate data detection in software testing; T indicates similar duplicate data record of software test identified by detection; D indicates that there are similar duplicate data records of software test in the dataset. Through calculation, the recall ratio of the two detection methods is obtained. The higher the recall ratio, the more data the software test duplicate data in the data set is correctly detected and recognized, the better the corresponding detection method is, and the higher the detection quality is, and vice versa. The experiment was conducted 10 times in total. Each time, 500, 1000, 1500, 2000, 2500 and 3000 pieces of data were randomly selected from the Movies dataset and the FreeCD dataset. In 10 experimental tests, repeated data were continuously put in artificially, thus increasing the effect of experimental tests. Record the time required for duplicate data detection experiment test, measure the duplicate data values detected by the two methods after each experiment test, summarize and calculate the recall rate of duplicate data of software test corresponding to the two methods, and draw the experimental test results into a comparison chart as shown in Figure 4.

It can be seen from the comparison results in Figure 4 that after the application of the software test duplicate data detection method proposed in this paper based on BP neural network and the traditional duplicate data detection method based on Dogmat iX, the evaluation indicators show different effects. Among them, after the application of the duplicate data detection method proposed in this paper, the recall rate of software test duplicate data detection is significantly higher than that of traditional methods, reaching more than 98%, while the recall rate of traditional detection methods is not more than 94%. From the comparison results, it is easy to see that the software test duplicate data detection method proposed in this paper based on BP neural network has high feasibility, significant advantages in detection effect, and can effectively deal with similar duplicate data problems. With the increase of the data set to be detected, the proportion of potential duplicate data records increases. The proposed detection method can also meet the requirements of duplicate data detection.

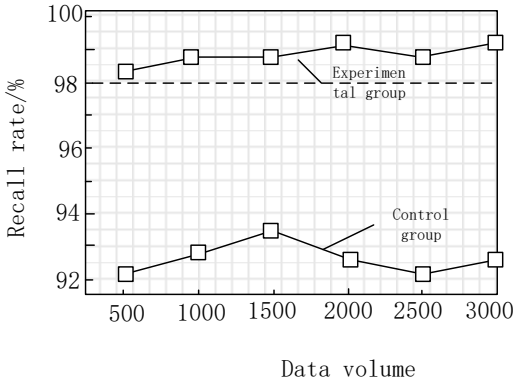

Figure 4: Comparison results of the recall ratio of duplicate data detection in two software testing methods.

4 CONCLUSION

In the process of software testing, affected by the hardware environment and software environment conditions, multiple duplicate data may appear in the test, which makes some structured test data difficult to meet the needs of software for rapid information exchange, reduces the scalability, flexibility and self description of data, and cannot provide basic guarantee for improving software service quality. In order to improve this problem, based on the current traditional duplicate data detection methods, this paper introduces BP neural network and proposes a duplicate data detection method for software testing based on BP neural network. Through the research in this paper, unnecessary detection steps are effectively removed, and the quality and efficiency of duplicate data detection are improved. According to the comparison results of the experimental test evaluation indicators, after the application of the software proposed in this paper to test the duplicate data detection method, the recall rate of duplicate data detection has reached more than 98%, which optimizes the algorithm performance, provides an important reference for the research in the field of similar duplicate data detection, and has good application prospects.

REFERENCES

Valstar N, Frasincar F, Brauwers G. APFA: Automated Product Feature Alignment for Duplicate Detection (J). *Expert Systems with Applications*, 2021:114759. https://doi.org/10.1016/j.eswa.2021.114759.

Albayrak O S, Aytekin T, Kalayc T A .Duplicate product record detection engine for e-commerce platforms (J).*Expert Systems with Applications*, 2022, 193:116420-. https://doi.org/10.1016/j.eswa.2021.116420.

Rocha T M, Carvalho A L D C .Siamese QAT: A Semantic Context-Based Duplicate Bug Report Detection Using Replicated Cluster Information (J).*IEEE Access*, 2021, PP(99):1-1.https://doi.org/10.1109/ACCESS.2021.3066283.

Alberto.Gascón, Casas R, Buldain D,et al. Providing Fault Detection from Sensor Data in Complex Machines That Build the Smart City (J).*Sensors* (Basel, Switzerland), 2022, 22(2). https://doi.org/10.3390/s22020586.

B Q Z A , A X L , C Q W .Interpretable duplicate question detection models based on attention mechanism(J).Information Sciences, 2021, 543:259-272. https://doi.org/10.1016/j.ins.2020.07.048.

Kremezi M, Kristollari V, Karathanassi V, et al. Pansharpening PRISMA Data for Marine Plastic Litter Detection Using Plastic Indexes(J). *IEEE Access*, 2021, PP(99):1-1.https://doi.org/10.1109/ACCESS.2021.3073903.

Zhou Y, Ren H, Li Z,et al.An anomaly detection framework for time series data: An interval-based approach(J).*Knowledge-Based Systems*, 2021, 228:107153-. https://doi.org/10.1016/j.knosys.2021.107153.

Zhang Y, Liu H, Qiao L. Context-sensitive Data Race Detection for Concurrent Programs (J). *IEEE Access*, 2021, PP(99):1-1. https://doi.org/10.1109/ACCESS.2021.3055831.

Cecilia A , Sahoo S, Dragicevic T ,et al.Detection and Mitigation of False Data in Cooperative DC Microgrids With Unknown Constant Power Loads(J).*IEEE Transactions on Power Electronics*, 2021, 36(8):9565-9577. https://doi.org/10.1109/TPEL.2021.3053845.

Zhong J , Zhang Y , Wang J ,et al.Unmanned Aerial Vehicle Flight Data Anomaly Detection and Recovery Prediction Based on Spatio-Temporal Correlation(J).*IEEE Transactions on Reliability*, 2022(1):71. https://doi.org/10.1109/TR.2021.3134369.

Liu C, Su X, Li C .Edge Computing for Data Anomaly Detection of Multi-Sensors in Underground Mining (J).Electronics, 2021, 10(3):302. https://doi.org/10.3390/electronics10030302.

Zhang R, Mei Y, Shi J .Robust change detection for large-scale data streams (J).*Sequential analysis*, 2022(1):41.

Lakhmiri D , Alimo R ,Sébastien Le Digabel.Anomaly detection for data accountability of Mars telemetry data(J).*Expert Systems with* Applications, 2022, 189:116060-.https://doi.org/10.1016/j.eswa.2021.116060.

Kiersztyn A , Karczmarek P , Kiersztyn K ,et al.Detection and Classification of Anomalies in Large Data Sets on the Basis of Information Granules(J).*IEEE Transactions on Fuzzy Systems*, 2021, PP(99):1-1. https://doi.org/10.1109/TFUZZ.2021.3076265.

Lv F , Zhao J , Liang T ,et al. Latent Gaussian process for anomaly detection in categorical data(J).*Knowledge-Based Systems*, 2021(12):106896. https://doi.org/10.1016/j.knosys.2021.106896.

Song D , Xu J , Pang J ,et al. Classifier-adaptation knowledge distillation framework for relation extraction and event detection with imbalanced data(J).*Information Sciences*, 2021, 573(3). https://doi.org/10.1016/j.ins.2021.05.045.

Fellah A .Near-Optimal and Domain-Independent Algorithms for Near-Duplicate Detection (J).*Array*, 2021, 11(3881):100070. https://doi.org/10.1016/j.array.2021.100070.

Rani S, Kumar A, Kumar N .Hybrid Deep Neural Model for Duplicate Question Detection in Trans-Literated Bi-Lingual Data (J).*Recent advances in computer science and communications*, 2021(3):14. https://doi.org/10.2174/2213275912666190710152709.

Jf A , Zx B, Hc A, et al. Anomaly detection of diabetes data based on hierarchical clustering and CNN(J). *Procedia Computer Science*, 2022, 199:71-78. https://doi.org/10.1016/j.procs.2022.01.010.

Dehnen G , Kehl M S , Darcher A ,et al.Duplicate Detection of Spike Events: A Relevant Problem in Human Single-Unit Recordings(J).*Brain Sciences*, 2021, 11(6):761. https://doi.org/10.3390/brainsci11060761.

Staowska P,Suchocki C. TLS data for cracks detection in building walls(J). *Data in brief*, 2022, 42:108247. https://doi.org/10.1016/j.dib.2022.108247.

Ahlawat A, Sagar K .Automating Duplicate Detection for Lexical Heterogeneous Web Databases (J).*Recent advances in computer science and communications*, 2022(4):15. https://doi.org/10.2174/2666255813999200904170035.

Wulkan R W, Horst M V D .Detection and correction of incomplete duplicate 24-hour urine collections – theory and practical evidence (J).*Biochemia Medica*, 2021, 31(1). https://doi.org/10.11613/BM.2021.010706.

Fels-Klerx H J V D , Smits N G E , Bremer M G E G ,et al.Detection of gluten in duplicate portions to determine gluten intake of coeliac disease patients on a gluten-free diet(J).*The British journal of nutrition*, 2021, 125(9):1051-1057. https://doi.org/10.1017/S0007114520002974.