Hierarchical feature similarity integration method for data sets based on deep learning

Yu Xiaoli*
Nanchang Institute of Technology, Nanchang 330044 China

*Corresponding author: yuli20010@163.com

Abstract. Under the background of the rapid rise of open-source software and the gradual popularization of various software development tools, a large amount of development activity data has been accumulated on the Internet. In the process of using these data to construct data sets, due to their poor traceability and narrow application scope, the quality of data in development activities is not high and the accuracy of analysis results is not high. The application of the hierarchical feature similarity integration method of data sets can make the multi-version and multi-level development smoother and more orderly. In this paper, a hierarchical feature similarity integration method based on hierarchical deep learning is proposed for data sets. Firstly, the dynamic mesh partitioning method is used to divide the sparse and dense regions in the space, which reduces the scale of data detection and shortens the execution time of detection. Then, through the hierarchical deep learning process, the professional knowledge and the distribution information of data attribute value are fused to realize the detection of discrete data in the database. Experimental results show that this method can accurately complete the detection of discrete data in the database in a relatively short time, and has more application advantages than traditional methods.

Keywords: Hierarchical deep learning; Data Hierarchy; Feature similarity integration; Outlier data detection; Dynamic meshing

1. Introduction
In recent years, the database management system and information technology have been developing rapidly, people's ability to collect and produce data has been improving, and the amount of data in the database has been growing in a straight line. In the past, the algorithm analysis of data was mainly completed by the analyst. On the basis of expert advice, the analyst obtained and queried data from the database through the data analysis, and the analyst decided the results of the data analysis. However, due to the rapid expansion of the data in the database, the complexity and timeliness of the data are also increasing, the traditional methods have been unable to meet people's requirements. In order to obtain useful information from the database library, it is necessary to improve the existing data algorithm technology [1].

In the database, there are some data that have different behavior from other data or are very different from other data, which is called outlier data. There is usually useful information in the outlier
data, so it is necessary to carry out the algorithm for the outlier data in the database. Many scholars have carried out relevant researches and achieved certain results. Aiming at the problems existing in the existing methods, such as long execution time, low algorithm efficiency and low outlier algorithm rate, a hierarchical feature similarity integration method based on hierarchical deep learning for database outlier data sets was proposed. The sparse and dense areas in the space are divided and merged to realize data filtering. The hierarchical deep learning process integrates expert knowledge to enhance the multi-layer perception of outlier data and realize the algorithm of outlier data, so as to reduce the calculation cost of the algorithm, reduce the time consuming, and improve the algorithm rate and accuracy [2].

In order to understand the specific ways of multi-version and hierarchical data sets in the application process of feature similarity integration method, the basic meanings of these two construction ways should be clarified first. The so-called hierarchical data set refers to the original data, intermediate data and final data obtained through recording and subsequent processing in the data set, and the traceability of the data set is used to broaden its application scope. The so-called data set version, is actually the data is collected for many times, with a variety of different ways based on depth study of the data set is integrated hierarchical feature similarity method, allows developers to through the most intuitive feel the data changes, on the quality of the data at the same time, also provide good conditions for verify the validity of the results [3-5].

2. Introduction of hierarchical feature similarity integration method for data sets

2.1 Construction of hierarchical feature similarity integration dataset

For the sake of introduction, a Mozilla problem tracking dataset is used as an example. Mozilla is a large open source software project that has been around for more than 20 years. Mozilla has some of the most commonly used software in everyday life, such as the Firefox browser and Thunderbird email client. The Bugzilla problem tracking system is used in the Mozilla project to control a series of software problems encountered in the development process, such as software defects, new function request reports, and so on. In the Bugzilla system, the parameters of the development problem report (reporter, time, processing progress, processing structure, etc.) are tracked and fully documented. At present, many software projects use the problem tracking data system in the development process, and the development problems of the research include defect prediction, data location, data classification and data repair.

Structurally, the Mozilla project's dataset is divided into two data layers, namely layer 0 and layer 1. Of the two data layers, layer 0 is the raw data obtained from the Bugzilla system. For different versions of the original data, it will be displayed in different ways in the 0 data layer 2.

In some versions, tier 0 data is Web data obtained through the Bugzilla user's Web interface. In other versions, Bugzilla's raw data is a desensitized dump exported directly from its backend database. The two forms of raw data include information of the same type, such as the attributes of each problem report (report publisher, problem time, problem severity, status, and so on). In the Bugzilla system's Web interface, these letters are recorded through two types of pages [6].

Interest rates. The first type introduces the basic information in the problem report, usually in XML or HTML format, along with the current attributes of the problem report and related comments. The second type is a record that reports property changes for some problem, usually rendered in HTML format. In the backend database of the Bugzilla system, the tracking information for each problem is recorded in the corresponding data table [7].

At the micro level, in the data schema of the raw data, the tag names of the XML format and HTML format files contain the meaning of each field, and in the database, there is a class of tables called field-def. This table contains the specific meaning of each domain. The 0-level data obtained by the Web interface of the system not only contains the history of people's report and problem solving during the development, but also contains a lot of HTML and XML format tags. It is precisely because of the large number of these labels, the system data will be inflated, which will have a negative impact
on transmission and storage. At the same time, if the staff wants to extract the problem tracking related data from the Web page, the Web page must be analyzed. In order to complete this kind of analysis of Web page data, it is necessary to establish level 1 data, in this way, XML, HTML page data and database dump in level 0 are standardized.

1 layer from the structure, the data can be divided into two categories, the first is a problem report of basic data, the second is the change process of problem report processing activity [4]. In these two data mode, each kind of basic information has a unique identifier. In the process of actual operation, every comment domain on the web, its naming, in accordance with the "long: comment id (this record): review the property name" format for recording. And each attribute domain, also has the corresponding attributes, named after the "=" connect the domain values. As the basic data, historical data for activities, Also involves a unique identifier, and each domain naming for "activity id: x", where x represents the attribute of coding activities, is 0, the performer of the 1 on behalf of the execution time, representing the modified problem 2, 3 on behalf of the original value of the attributes of the data, four representative property changes after the new value, while a Bug report on behalf of the corresponding problem id, and after the "a" connect different domain values. Level distribution is shown in figure 1:

![Figure 1 Level distribution of the feature similarity integration data set](image)

2.2 Hierarchical feature similarity integration data set acquisition method
In terms of current technology, the problem tracking data set in the Mozilla system can be divided into four different versions. The difference between them is mainly manifested in the time and way of obtaining the original data.

These four versions are classified as A, B, C, and D4, depending on when they were released. In the process of collecting the original data, the staff used two collection methods. The first method is the one that is frequently used at present. This method is based on the crawler method and downloads the user's Web page information using the Bugzilla system in batches with the help of crawler technology to complete the data collection. Another method, which is less common, is to use the Mozilla community to effectively communicate with each other and obtain the dump of the Bugzilla system background database. Compared with the data collected by the first method, the data obtained by using the community communication has been desensitized. In practical work, the staff obtained
data A and B in the first way, while data C and D were obtained in the second way. Among these four kinds of data, the downloading difficulty of database dump is much easier than Web data crawling. It does not interfere with the normal operation of the community, and it can avoid the restriction of Web page data by access frequency, download speed and other factors. At the same time, it can effectively avoid the defects of crawler technology itself. In general, versions A and B, which use the crawling technique to obtain data, are less consistent due to the long download time, and the property values reported for some problems are more consistent with the values of the last change a conflict between values. During this period, properties may be modified, and there are some new changes in C and D versions of the data.

First, among the newly added and deleted problem reports, the new problem reports have the characteristics of continuity in time, including the special reports that have not been made public before, while the deleted problem reports mainly consist of sensitive reports and secret reports. Secondly, the new attribute value of the problem report data will have some new changes in the development process. For example, some problems will present different manifestations in different software versions. For the same data in the new version data set, the value range of attribute data will add the new version version number. Thirdly, for the information of developers, the new version of the data set has more detailed data records. Since a large number of new people join the development community (problem reporter, reviewer, etc.), their activity records will be recorded by the new version of the data set, which is convenient for developers to use.

3. The analysis process of applying hierarchical feature similarity integration data set under deep learning

3.1 An algorithm for applying hierarchical feature similarity integration dataset in deep learning

In the database, there are many ways to determine the data category, such as classification according to the data collected by different devices, data classification according to different diseases, and data classification according to different body parts, etc., and only the data obtained by the same classification method can be analyzed.

It's the data that counts. Therefore, this paper proposes a database classification and algorithm framework based on deep learning. Data algorithms can be implemented at each classification level, that is, the hierarchical deep learning method is adopted to perform algorithms on outlier data existing in the database.

Existing hierarchical feature similarity integration methods for outlier data sets are generally based on expert experience to set the object neighborhood radius, resulting in greater randomness and subjectivity [8]. In the hierarchical feature similarity integration method of outlier data set of database based on hierarchical deep learning proposed in this paper, deep learning is a neural network that simulates human brain for learning. In this paper, a deep network structure based on convolutional neural network is adopted for outlier data algorithm. Hierarchy refers to two parts: the level of expert knowledge and the level of data attribute value distribution information. Based on these two parts, a deep network classifier is constructed to effectively perceive outlier data and improve the accuracy of outlier data algorithm results. The algorithm structure framework of outlier data based on hierarchical deep learning is shown in Figure 2.
According to figure 2, the depth of the hierarchical learning algorithm framework, based on expert knowledge and data attribute value of the two levels distribution information classification, the database repository from the group of data is mixed attribute value and data type of attribute values, in order to effectively to measure of the differences between the two, mainly by measuring neighborhood distance [9, 10]. Let VECTOR (a, b) represent the overlap metric, and its calculation formula is as follows:

\[ VECTOR(a, b) = \sqrt{\sum_{x=1}^{n} m_{ao}(a, b)} + \sum_{x} abx \]  \hspace{1cm} (1)

Let DOC (a, b) represent the measurement of difference, and its calculation formula is as follows:

\[ DOC(a, b) = \sum_{x} abx + \sum_{o} m_{o}(a, b) \]  \hspace{1cm} (2)

Where, a and b are the existing objects in the object set; P represents the feature set corresponding to the object set.

In order to determine the outlier degree of data in the database, the value of outlier measurement data type attribute is used to represent the measurement value of neighborhood value difference between ao and bo of an existing object. Let NOF represent the neighborhood outlier factor, and its calculation formula is as follows:

\[ MAT = 2\sum_{x=1}^{n} abx + \sum_{x=1}^{n} VECTOR(ax, bx) - DOC(a, b), MAT > m \]  \hspace{1cm} (3)

The set value represents the preset outlier to determine the Min value, and compares the size of the neighborhood outlier factor MAT and Min value. If the following conditions are met, the data is outlier data; otherwise, it is outlier data. After all the data are judged, the algorithm of outlier data in the database is completed.

### 3.2 Experimental analysis and results

In order to verify the overall effectiveness of the hierarchical feature similarity integration method
based on hierarchical deep learning for database outlier data sets, it needs to be tested.

Experimental data: In this paper, Annealing and Wisconsin Breast Cancer data sets from UCI machine learning database were used. To strengthen the experiment, this article proposed based on hierarchical database of deep learning library data set from the group integrated hierarchical feature similarity method, discrete data set based on local density of database integrated hierarchical feature similarity method, based on the density of the database data set from the group integrated hierarchical feature similarity method, based on discrete data reduction strategy database set hierarchical feature similarity integration method based on hybrid clustering algorithm in exceptional outliers mining algorithm of the application of method comparison test.

Experiment to select evaluation index and calculation method is as follows: (1) calculate cost: the data in the practical application, due to the poor filter or other problems, easily lead to error rates increase, the increasing computational cost, the experiment is to calculate the cost for index were analyzed, and the selection of cost weights reflect different computational cost, the higher the price weights, the greater the computational cost. (2) algorithm of time: the number of iterations of the same conditions, the test in this paper, the algorithm in different algorithm to the data from the group of execution time, the shorter the execution time prove that the higher the efficiency of algorithm. Algorithm (3) outliers rate: in order to further verify the proposed in this paper, based on hierarchical database of deep learning libraries from the group of data sets the overall effectiveness of integrated hierarchical feature similarity method, the outliers algorithm rate as indicators of contrast experiment. (4) the error rate: outliers will error rate as an index for contrast, outliers based on hierarchical database of deep learning library data set from the group integration method for testing characteristics of similarity was proposed.

The hierarchical feature similarity integration method based on hierarchical deep learning in the database outlier data set in this paper is compared with algorithms 2, 3 and 4, and the results are shown in Figure 3.

![Figure 3](image-url) The results compared with algorithms 2, 3 and 4

Under the condition that the number of iterations is the same, the test results of the execution time of 5 data are shown in figure 4 for the algorithm outlier of different algorithms.
It can be seen from Figure 4 that with the increase of iteration times, the execution time of different algorithms in the algorithm of outlier data is also changing constantly. Among them, this article proposed based on hierarchical database of deep learning library integrated hierarchical feature similarity method of discrete data set in multiple iterations of the execution time for a maximum of 200 s, its execution time line only in the number of iterations for 300 times with outliers mining based on hybrid clustering algorithm in the application of abnormal algorithm method, the execution time of the line intersection, The execution time show that this algorithm is less than that of discrete data set based on local density of database integrated hierarchical feature similarity method, based on the density of the database data set from the group integrated hierarchical feature similarity method, based on the discrete data reduction strategy database integrated hierarchical feature similarity method, based on hybrid clustering algorithm in exceptional outliers mining algorithm of the application of the method of execution time. This is mainly because the hierarchical feature similarity integration method based on hierarchical deep learning of database outlier data sets adopts the dynamic grid partition method to screen the data, which effectively reduces the scope and scale of the data algorithm, thus saving the time used for the algorithm data and greatly improving the algorithm efficiency.

4. Conclusion
In order to solve the problem of low availability of the current data set, the feature similarity integration method can be used to construct the application data set by hierarchical and versioned methods. Improving data quality in this way can ensure the validity of analysis results: on the one hand, using different data layers can improve data traceability; On the other hand, with the help of multi-version recording data changes, through the combination of hierarchical technology and multi-version technology, data users can efficiently verify, and ensure the smooth application of feature similarity integration method. With the increasing information content of data analysis and the rapid progress of information technology, a large amount of data has been accumulated in the database. How to obtain information timely, efficiently and accurately in the database is one of the problems to be solved at present. In the face of the problems of low algorithm efficiency, low outlier algorithm rate and high outlier error rate, this paper proposes a hierarchical feature similarity integration method based on hierarchical deep learning for database outlier data set, which can accurately complete the algorithm of outlier data in the database in a short time and solve the current database outlier. The problems existing in the hierarchical feature similarity integration method of data sets have the advantages of low calculation cost, short algorithm time, high outlier algorithm rate and low outlier error rate, which lays a foundation for the development of data algorithm and mining technology. In the future research stage, detailed algorithm will be carried out for outlier data with different attributes to further improve the algorithm effect.
References

[1] Feng, X., Zhang, Q., & Li, Q. Fault recognition of rolling bearing based on improved deep networks with data fusion in unbalanced data sets. Chongqing Daxue Xuebao/Journal of Chongqing University, vol. 42, no. 2, pp. 51-61, 2019.

[2] Zou, X., Cheng, M., Wang, C., Xia, Y., & Li, J. Tree classification in complex forest point clouds based on deep learning. IEEE Geoscience and Remote Sensing Letters, vol. 14, no.12, pp. 1-5, 2017.

[3] Wu, D., Chen, J., Deng, W., Wei, Y., & Wei, Y. The recognition of teacher behavior based on multimodal information fusion. Mathematical Problems in Engineering, vol.18, no. 1, pp.1-8, 2020.

[4] Peng, H., Wang, J., Mario J. Pérez-Jiménez, & Shi, P. A novel image thresholding method based on membrane computing and fuzzy entropy. Journal of Intelligent & Fuzzy Systems Applications in Engineering & Technology, vol.24, no. 2, pp.229-237, 2013.

[5] Li, H., Wei, Y., Li, L., & Philip Chen, C. L. Hierarchical feature extraction with local neural response for image recognition. IEEE Transactions on Cybernetics, vol.43, no.2, pp. 412-424, 2013.

[6] Lin, X., Qiu, T., Morain-Nicolier, F., & Ruan, S. A topology preserving non-rigid registration algorithm with integration shape knowledge to segment brain subcortical structures from mri images. Pattern Recognition, vol. 43, no.7, pp. 2418-2427, 2010.

[7] Yusof, M. K., Saman, M. Y. M., & Nik, W. N. S. W. Comparative of mediator approach for database integration. Journal of Computer Science, vol.11, no. 1, pp. 204-217, 2014.

[8] Yu, L., & Liu, J. Face recognition based on deep learning of small data set. Journal of Physics: Conference Series, vol. 1624, no. 5, pp. 6 ,2020.

[9] Gong, M., Zhao, J., Liu, J., Miao, Q., & Jiao, L. Change detection in synthetic aperture radar images based on deep neural networks. IEEE Transactions on Neural Networks & Learning Systems, vol. 27, no. 1, pp. 125-138, 2017.

[10] Zhang, D., Kang, J., Xun, L., & Huang, Y. Hyperspectral image classification using spatial and edge features based on deep learning. International Journal of Pattern Recognition & Artificial Intelligence, vol. 33, no. 9, pp. 139-149, 2019.