Convolutional Neural Network Based Structured Data Field Granularity Classification Method

Pin Wu, Quan Zhou, Qiu Wei, Zhidan Lei and Xiaoqiang Li
School of Computer Engineering and Science, Shanghai University, No. 99 Shangda Road, Baoshan District, Shanghai
jzhou8763@shu.edu.cn

Abstract. The data warehouse provides data support for enterprise decision-making and online analysis. In the process of building a data warehouse, many heterogeneous source system data needs to be integrated. In the integration process, these heterogeneous data needs to be classified and put into different topics. The diversity of the source systems of large and medium-sized enterprises poses difficulties for granularity at the field-level automated classification. However, the accuracy of previous methods cannot satisfy users. Therefore, this paper proposes a neural network-based classification technology to classify the data in the granularity field. This method adopts a sampling method to construct the characteristics of the field and innovates a novel classification framework based on the database field on the basis of the CNN network. Accurately achieve 89% by testing the data in the TPC-DS's dimension tables, and achieve 93% accuracy in real-world data testing. This method was validated in the actual environment of the three banks in China and achieved satisfactory results.

1. Introduction
Decision-making plays a crucial role in the success or failure of a business. The data warehouse is a structured data environment for decision support systems (DSS) and online analytical application data sources [1]. The quality of data warehouses is particularly important. Mouncey P [2] believed that the premise of improving the quality of data warehouse and business information is that the quality of information stored in the database is as measurable as the quality of the assembly line. So our efforts are to improve the quality of data warehouses.

Improving data quality requires the correct classification of data. Data warehouses are characterized by integration, topic-oriented, stability, and time-varying [3]. On the one hand, the data warehouse is integrated, and the data in the data warehouse comes from scattered operational data. The required data is extracted from the original data, processed and integrated, and unified and integrated before loading into the data warehouse. The data in the data warehouse is processed, summarized, and organized on the basis of the extraction and cleaning of the original scattered database data. The inconsistency in the source data must be eliminated to ensure that the information in the data warehouse is about the entire enterprise. On the other hand, data warehouses are topic-oriented. The data organization of the operational database is oriented to the transaction processing task, and the data in the data warehouse is organized according to a certain subject domain. Topics are the key areas of concern for users when making decisions using data warehouses. A topic is often associated with multiple operational information systems.
Building a data warehouse generally includes three key stages. They are data acquisition, data storage, and data access. There are many problems to be solved in the data acquisition phase for the characteristics of the data warehouse. Facing the task of high quality data, we have to classify heterogeneous data properly.

| id | name | age | address                      |
|----|------|-----|------------------------------|
| 1  | Jack | 25  | Dong Xin Road, Shanghai, China|
| 2  | Wendy| 22  | Oxford Street, London, United Kingdom|
| 3  | Iris | 12  | Fifth Avenue, New York, United States|
| 4  | Bella| 13  | Champs-Elysees, Paris, France |
| 5  | Cherry| 18 | Shinjuku Street, Tokyo, Japan |

Figure 1. Topic classification at field granularity

As shown in Figure 1, our task is to map each field in the table to the one of the topics or subclasses in the data warehouse. The name field is mapped to the ACCOUNT topic, and the age and address fields are mapped to the PERSONAL INFORMATION topic. The data types in the fields are diverse, including character, numeric, and so on. This brings with it the difficulty of classification tasks.

The previous technology could not meet people's requirements. Now it takes a lot of labor to classify it. Therefore, this article reviews the previous technologies and finds their flaws. Then put forward our approach. Then describe in detail the architecture and implementation of the method. Finally, verify through experiments.

2. Related Works

People have been working hard to build data warehouses efficiently and continuously improving the quality of data warehouses. One of the important goals is to achieve automated classification.

2.1. Data Warehouse

The father of data warehouse Bill Inmon proposed the concept of data warehouse in 1990 [4]. Data warehousing is a group of decision-based technologies designed to enable knowledge workers to make better and faster decisions. In recent years, there has been explosive growth both in the number of products and services provided and in the industry's acceptance of these technologies. The first stage of building a data warehouse requires a variety of data extraction and cleaning tools. One of the most important tasks to use these tools before loading is to sort the data and load it into the corresponding theme, as shown in the blue box in Figure 2. Currently available tools and methods cannot achieve high-precision classification results.

Golfarelli M et al. [5] proposed a general method framework for data warehouse design based on dimension fact model (DFM). After analyzing the existing information system and collecting user requirements, the concept design is semi-automatically started from the operation of the database solution. Fang Y [6] described the data quality problems in the data warehouse environment and the importance of ensuring data quality for data quality, and finally provided a data quality maturity model (DQMM) and method for controlling data quality of data warehouses. Chen Z Q [7] proposed a new data warehouse model based on topic graph technology, which automatically generates and merges thematic maps. Batra V S et al. [8] proposed the use of DAGs to generate ETL scripts that can be used to populate newly proposed data warehouses based on existing data in existing schemas.
2.2. **Sampling Methods**

Due to the large amount of source data for large and medium-sized companies, analyzing the source data in full amount will lead to inefficiency. Therefore, we try to find a good sampling method to better characterize the source data. People have done a lot of work in the sampling field. Random sampling requires strict adherence to the principle of probability. Each sampling unit has the same probability of being extracted and can be reproduced. Random sampling is often used when the overall number is small, and its main feature is to extract one by one from the population. Random sampling can be divided into simple random sampling, systematic sampling, stratified sampling and cluster sampling.

Olken F et al. [9] analyzed the basic operations of large-scale database auditing and statistical analysis, and discussed sampling on this basis. A method of obtaining a sample from the result of a relational query without first executing the query is given. Specifically, check for simple random sampling of selections, projections, connections, joints and intersections. Kadilar C et al. [10] proposed ratio estimators by adjusting Ray and Singer's estimator types to find conditions that make each proposed estimator more effective than other estimators. The advantage of simple random sampling is that it is simple to operate, but it is not suitable for oversized data.

Earlier, Neyman J. et al. [11] proposed stratified sampling. The core idea is to divide the population into non-intersecting layers, and then to extract a certain number of individuals from each level according to a certain proportion, and to combine the individuals taken from each layer together as a method of sample. The smaller the variation within the layer, the better, and the greater the variation between the layers, the better. Tong C et al. [12] described some refinement strategies for layered experimental design, such as Latin hypercube, orthogonal array, and factorial design to solve the adequacy of a given sample to give acceptable statistical estimates. The problem identified. The advantage of this method is that it reduces the sampling error, the sampling method is flexible and the different layers can be analyzed independently. The disadvantage is that if the stratification variables are not properly selected, the intra-layer variability is large and the inter-level mean numbers are similar, stratified sampling will lose its significance.

2.3. **Machine Learning Classification Methods**

Once we have good characterization data, we will consider what method to classify these characterization data. With the rapid development of computing power, storage space, and networks, the amount of data that humans accumulate is rapidly growing. Classification is a very important task
in data mining and is currently used most commercially. Classification algorithms are used to solve classification problems. The method is an important research area in machine learning. Classification is very common in our real life. Classification is one of the most useful technologies. Its application covers all areas of society. Common applications include: user portraits, credit assessment, recommendation systems, trend prediction, image classification, and text classification.

Breiman L. [13] proposed that random forests are a combination of tree predictors, so each tree relies on the value of an independent sampled random vector, and the distribution of all trees in the forest is the same. Zhu J [14] developed a new algorithm that directly extends the AdaBoost algorithm to multiple classes without reducing it to multiple types of problems. We have proved that the proposed multi-class AdaBoost algorithm is equivalent to a forward-stage additive modeling algorithm, which can minimize the novel exponential loss of multi-class classification. Landgrebe D [15] introduced the current decision tree classifier (DTC) design methodology and investigation of various existing problems. Decision tree is an algorithm model that is more widely used in the industry. The advantage is that the computational complexity is not high, and the reuse is high; the output result is easy to understand; the intermediate value is insensitive; the irrelevant feature data can be processed; and the continuous features can be processed well. The disadvantage is that it is prone to overfitting problems; it is more sensitive to noise data. Friedman J H [16] proposed that the gradient promotion of the regression tree provides a competitive, highly robust and interpretable procedure for regression and classification, and is particularly suitable for mining less than clean data.

2.4. Neural Network Classification Methods

The classification model of machine learning still has the defect of insufficient accuracy for large-scale data. In 2006, Hinton used the pre-training method to alleviate the problem of local optimal solution, pushed the hidden layer to the 7th layer [17], and the neural network had a “depth” in real sense, thus opening up a boom of deep learning. In the structure of the fully connected DNN, the underlying neurons and all upper neurons can form connections. The potential problem is the expansion of the number of parameters. For CNN, not all upper and lower neurons can be directly connected, but through the "convolution kernel" as an intermediary. It is precisely because the CNN model limits the number of parameters and taps this feature of the local structure. In the recurrent neural network RNN, the output of the neuron can be directly applied to itself at the next timestamp, ie, the input of the i-th layer neuron at time m, in addition to the output of the (i−1) layer neuron at this moment, Its own output at (m−1). The RNN can be seen as a neural network that passes in time. For t, the gradient it produces disappears after it has spread several layers to history on the time axis. In order to solve the disappearance of the gradient in time, the field of machine learning develops. The long-term memory unit LSTM realizes the memory function in time through the switch of the door and prevents the gradient from disappearing. The CNN and RNN are often connected to the full-connection layer before they are output by the upper layer, so they can complete the classification task.

Yoon Kim proposed in 2014 that TextCNN is an algorithm that uses convolutional neural networks to classify text. The algorithm transforms sentences into similar images through word vectors, and then obtains local information through convolutions to obtain word-based feature maps. It has achieved good results in sentiment analysis and question-answering classification tasks. FastText is a fast text classifier developed by Facebook that provides simple and efficient text classification and representation of learning [18]. The input to the FastText model is a sequence of words (a piece of text or a sentence), and the output is the probability that the sequence of words belongs to a different category. The words and phrases in the sequence form a feature vector. The feature vector is mapped to the middle layer through a linear transformation and then mapped to the label by the middle layer. FastText uses a nonlinear activation function when predicting labels, but does not use nonlinear activation functions in the middle layer. Pengfei Liu and Xipeng Qiu et al. [19] use a multitasking learning framework to learn multiple related tasks together. Based on recursive neural network, three different information sharing mechanisms are proposed to simulate texts with task-specific and shared layers, thus completing a very good text classification task.
3. Our method

3.1. Data Flow of Classification
The model we implemented was before the source data was loaded into the data warehouse, that is, in the blue box in Figure 3. It is possible to achieve highly automated loading into data warehouses through our highly accurate and efficient models.

![Figure 3. Data Flow of Classification](image)

In the real world, large and medium-sized enterprises have a large number of heterogeneous source systems. For example, in China, a medium-sized bank has more than 300 source systems. Here, we use A, B, and C as the representative. The core of this paper is the processing in the blue box, which will be described in detail in the next section.

3.2. Architecture of Our Method
The method we propose can be used to classify data on the one hand, as shown in Figure 4. On the other hand, since there are a large number of fact tables in the real world, the data of its fields is foreign key ID data that has no practical significance. Therefore, for such data, we design the method as shown in Figure 5.

![Figure 4. Architecture of Our Method for Data-Based](image)
Since our task is to classify the data of the granularity at field-level to topic classification, we first select the data in one field of a table in the source system. Secondly, these data are sampled in a hierarchical manner and 100 samples are extracted. The 100 samples are then sorted lexicographically. The sorting process allows the data to have certain contextual information. Then, the words in the sample are mapped into a 300-dimensional vector space. Next, it is convolved with a convolution with coefficient k of 2, 3, 4, 5. The 128-dimensional feature map generated by this process is concatenated into a 512-dimensional feature map. Maxfielding and dropout (0.5 in Training, 1 in Testing) are performed on the feature map and input to the softmax function. Finally, the classification result is obtained. That is, we need to obtain the topic classification.

![Figure 5. Architecture of Our Method for Name-Based](image)

The name-based approach differs from the data-based approach in that the input to the neural network is not the data from the field, but the table's metadata or structure, the main active system name, the table name, and the field name. We enter the entire field names in order. We introduce a focus mechanism to reinforce each field, and when this is categorized, the red box in Figure 5 is used.

3.3. Algorithm of Our Method
To better illustrate our approach, we list our algorithm in Table 1.

### Table 1. Algorithm of Sampling and Sorting.

| Input          | Column Data, Topic Dictionary |
|----------------|-------------------------------|
| Output         | Topic Name                    |
| 1              | $\text{DataFrame} = \text{GetColumnData}$(Column Data) |
| 2              | $\text{DataFrame} = \text{Sort}$(Sampling(DataFrame)) |
| 3              | $\text{EmbeddingData} = \text{Embedding}$(Lookup(DataFrame)) |
| 4              | For i in enum(kernels):       |
|                | $\text{FeatureMap}.\text{append}(\text{Conv2d} \ (\text{EmbeddingData}, \ i))$ |
| 5              | $\text{FeatureMap}$.concat(2) |
| 6              | $\text{TopicName} = \text{LookupDicts}$(Dropout(Softmax(Maxpooling()))) |

4. Experiments and Discusses
4.1 Datasets and Environments
The configuration of our main equipment is a machine with a 4-core, 64G memory and a GTX-1080 GPU 8G memory. Our experimental data selects two types, one is the test data generated by the TPC-DS tool. TPC-DS is a set of decision support system test benchmarks, mainly for the retail industry. Test data is highly similar to real-world data and has various business models. See http://www.tpc.org. The data size of the test can be customized. The data has a fact table and a dimension table, so it is very consistent with our experimental tasks. We generated 1TB of test data in this experiment. We select 90% as the training set and 10% as the test set.

The other one selects data from three Chinese banks. Since the table names and field names given by TPC-DS are very regulated, if we put them into our method, the model would get 100% accuracy very quickly. However, in reality, the heterogeneity of the source system results in very diverse table names and field names. Therefore, we choose real-world data to complete the classification task by data-based method.

4.2 Experimental Result
We evaluate our method with accuracy, recall rate and F1 value. The results of the test are shown in Table 2.

| Algorithm Name          | Accuracy | Recall | F1 Value |
|-------------------------|----------|--------|----------|
| RandomForestClassifier  | 0.68     | 0.53   | 0.57     |
| AdaBoostClassifier      | 0.73     | 0.73   | 0.58     |
| DecisionTreeClassifier  | 0.85     | 0.85   | 0.85     |
| GradientBoostingClassifier | 0.89   | 0.85   | 0.86     |
| Our Methods             | 0.93     | 0.91   | 0.92     |

4.3 Discusses
It is not difficult to find from the experimental results that our method obtains relatively high performance indicators. To get good results, you should choose the right way to process data and choose the right algorithm to learn the data according to different data structures and different tasks. As we discussed in this article is a classification task. Different feature extraction methods are selected according to the different characteristics of the fact table and the dimension table. For the fact table, we need to extract the characteristics of the source system data, such as the source system name, table name, field name, field type and field length to complete the fact table classification task. For the dimension table, we select the data in the source system data to complete the classification task. For data that has semantic information and context, such as table names and field names, using CNN can achieve better classification results. For a large number of practically meaningful data, the choice of ensemble learning in machine learning can achieve very good classification results and rapid acquisition of model features.

5. Conclusions
The data loading stage of the data warehouse requires high-accuracy classification tools to accomplish the task of topic classification and secondary classification. The method proposed by us has high-performance classification of data with a granularity field. Accurately achieve 89% by testing the data in the TPC-DS's dimension table, and achieve 93% accuracy in real-world data testing. The preliminary results can be drawn from the experimental results. Selecting the appropriate feature extraction method for the field data in the database and using the computational framework of the convolutional neural network can obtain the ideal classification effect. At the same time, the great potential of using neural networks to solve traditional problems can be seen. This research result can also be applied to the application of automatic ER map construction and other databases. The method we propose can save a lot of manpower for large and medium-sized enterprises that use data warehouses, and can contribute to obtaining higher quality data. The future work is to adopt a better
field characterization method to further improve the neural network classification accuracy, achieve an accuracy rate that exceeds the manual classification, so as to achieve full automation of data warehouse loading.

Acknowledgments
Thanks to the professors and students of Shanghai University for their valuable suggestions. Thanks for the support of the hardware and software environment provided by Beyondsoft (Shanghai) Co., Ltd. Thanks to three Chinese banks for providing demand analysis and real-world data for verification.

References
[1] Widom J. Review - An Overview of Data Warehousing and OLAP Technology[J]. 1999.
[2] Mouncey P. Improving Data Warehouse and Business Information Quality[J]. Interactive Marketing, 2001, 2(3):295-297.
[3] Mathur N, Mathur N, Mathur N. An Overview of Data Warehouse and OLAP Technology Used in Banking System[J]. International Journal of Emerging Trends & Technology in Computer Science, 2016, 33(2):76-79.
[4] Inmon W H. Using Oracle to Build Decision Support Systems[M]. John Wiley & Sons, Inc. 1990.
[5] Golfarelli M, Rizzi S. A methodological framework for data warehouse design[J]. 1998, 2:3-9.
[6] Fang Y. Data Quality Managements in Data Warehouse[J]. Computer Engineering & Applications, 2003.
[7] Chen Z Q, Wang Z Y, Xia L X. Data Warehouse Model Based on Topic Map[J]. Information Science, 2008, 26(11):1689-1693.
[8] Batra V S, Bhide M A, Mohania M K, et al. Data warehouse data model adapters[J]. 2017.
[9] Olken F, Rotem D. Simple Random Sampling from Relational Databases.[C]/ Vldb’86 Twelfth International Conference on Very Large Data Bases, August 25-28, 1986, Kyoto, Japan, Proceedings. DBLP, 1986:160-169.
[10] Kadilar C, Cingi H. Ratio estimators in simple random sampling[J]. Applied Mathematics & Computation, 2004, 151(3):893-902.
[11] Neyman J. On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection[J]. Journal of the Royal Statistical Society, 1934, 97(4):558-625.
[12] Tong C. Refinement strategies for stratified sampling methods ☆[J]. Reliability Engineering & System Safety, 2017, 91(10):1257-1265.
[13] Breiman L. RANDOM FORESTS--RANDOM FEATURES[J]. Machine Learning, 1999, 45(1):5-32..
[14] Zhu J, Zou H, Rosset S, et al. Multi-class AdaBoost[J]. Statistics & Its Interface, 2009, 2(3):349-360.
[15] Landgrebe D. A survey of decision tree classifier methodology[J]. IEEE Transactions on Systems Man & Cybernetics, 2002, 21(3):660-674.
[16] Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine[J]. Annals of Statistics, 2001, 29(5):1189-1232.
[17] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786):504-507.
[18] Joulin A, Grave E, Bojanowski P, et al. Bag of Tricks for Efficient Text Classification[J]. 2016:427-431.
[19] Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning[C]/ International Joint Conference on Artificial Intelligence. AAAI Press, 2016:2873-2879.