Research and Regression Analysis of Enterprise Attendance Data Based on Big Data Technology

Lei Yao, Yuchuan Bian, Xiaoming Ji, Mingyu Guo
ylcasc@163.com, bianyuchuan@163.com, jixiaoming@163.com, guomingyucast@163.com

Beijing Institute of Spacecraft Environment Engineering Beijing, China

Abstract—The traditional database technology can hardly complete full data analysis on mass accumulated attendance data for the enterprise over the years, while big data technology is regarded as a better way to achieve it. After the study on a data-driven attendance management system, this paper proposed a decision tree regression analysis method based on the Spark platform, analyzing mass attendance data in a short time and generating a decision tree to show what affects employee attendance. The decision tree has certain reference significance for the monitoring and warning of employee attendance and employee management decision. This paper detailed the process of data-parallel processing and the regression analysis method of employee attendance by the Classification and Regression Tree (CART) algorithm and compared the efficiency of traditional database technology and parallel data processing for employee attendance data analysis, which verified the effectiveness of the method.

Keywords- Big data technology; attendance data; regression analysis; decision tree

1 INTRODUCTION

To achieve informatization in a digital era, enterprise attendance management mainly demands to focus on core data, respond promptly, analyze matrix reports, strengthen labor discipline, and promote enterprise development. By doing so, the enterprise attendance management system will become more intelligent with cutting-edge information technology integrated into it. Therefore, it is imperative to set up an attendance management system suitable to the development of enterprise as it is one of the key points for enterprise informatization development [1].

2 STUDY ON DATA-DRIVEN ATTENDANCE MANAGEMENT SYSTEM

With the cloud computing technology and mobile Internet technology booming, the current attendance management system has realized decentralization in contrast to traditional human resource management tools by the good combination of advanced cloud and mobile technology

ICICA 2022, December 02-04, Chongqing, People's Republic of China

DOI 10.4108/eai.2-12-2022.2328007

and by more intelligent and convenient management tools. Under the development of artificial intelligence and big data, the user interactive mode based on iris recognition and intelligent robots is changing the traditional identification ways. Advanced enterprises have already adopted methods such as iris recognition for their attendance management, which not only improves the efficiency of the management but also ensures the fairness of attendance data. Some enterprises build a large number of data models to form a complex mapping knowledge domain through big data technology and massive historical data before applying them to fields such as attendance and making decisions with the help of artificial intelligence technology.

According to the current status and future development direction of the attendance management system, this paper presented a new attendance management system architecture, as shown in Figure 1, which formed a data-driven system based on new technologies and the internal demand core of enterprise informatization.

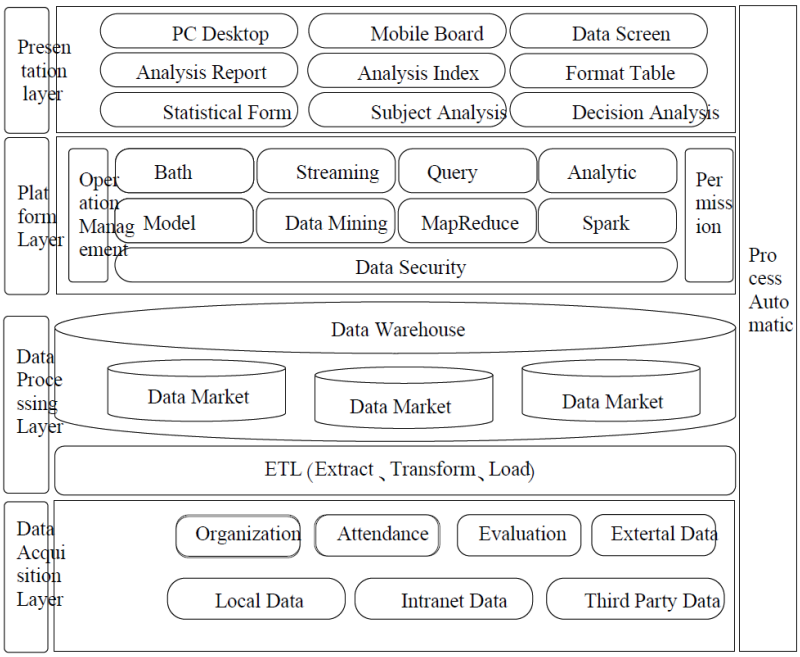

Figure 1 Architecture Diagram of Data-driven Attendance Management System

3 THE ALGORITHM AND COMPUTING ENVIRONMENT OF ATTENDANCE DATA

3.1 Analysis of the Platform Layers

Data mining technology was used for future data analysis to meet the need of some high-level data analysis [2]. The platform provided reliable analysis components for the analysis and utilization of attendance management data, mainly including the data model generation by

multiple data mining algorithms and visual display function. Clustering algorithms and classification algorithms are the two main types. Clustering algorithm includes hierarchical clustering algorithm, segmentation clustering algorithm, clustering algorithm based on constraints, clustering algorithm of machine learning, and clustering algorithms for high-dimensional data [3]; classification algorithm mainly includes decision tree classification algorithm, Bayes classification algorithm, classification algorithm based on association rules, and classification algorithm based on database [4]. Besides, the two major big data processing engines of Spark and MapReduce were also offered to extract important values from huge amounts of data with low value density [5]. The two engines were set to deal with different data processing needs, Spark running faster in memory, and MapReduce brings its superiority into full play by processing large amounts of data on disk. As shown in Table 1, according to the different business needs of the attendance management system and various requirements for data analysis and processing modes and data analysis components, this paper mainly divides the data analysis modes of attendance management into the following four types.

Table 1 Comparative Table for Calculation Models

No.	Calculation Models	Application Scenarios	Computing Framework
1	Bath Computing	Attendance time generation in batch, Commute time screening in batch	MapReduce, Spark
2	Streaming Computing	Absenteeism and other abnormal behavior analysis	Storm, Flink Spark Streaming
3	Online Query	Staff basic information search, staff attendance record search	Elasticsearch Redis
4	OLAP (On-Line Analytical Processing) analysis	To evaluate the labor discipline of the department according to the cumulative attendance time of the employees searched in the department within a specific period.	Impala, Kylin, ClickHouse, AnalyticDB

3.2 Spark computing environment

As a fast and universal distributed parallel computing platform, Spark runs much faster than the traditional serial computing mode and the MapReduce model based on Hadoop by extending the MapReduce computing model and using memory computing mode. In the Spark platform, the Spark Structured Query Language (SQL) was selected to process structured and semi-structured data, which can operate relational databases easily and perform queries and statistics efficiently [6]. And a special kind of Resilient Distributed Dataset (RDD) called Dataframe is included in Spark SQL to store Row objects and contains abundant data operations such as Map, Flatmap, Filter, ReduceByKey, and others. In this paper, attendance data and related data were read by Spark SQL, which connects relational database and memory parallel computing process of attendance data were completed by Dataframe data set and related operations. As a result, the computing efficiency was greatly improved.

3.3 Parallelization of Classification and Regression Tree Algorithm

Cart recursively divides the input space and defines a local model in the result area of each input space. In each region, there is a separate model to predict the target variables. It determines which region it belongs to by starting from the root node at the top of the tree and moving down the path to a specific leaf node according to the decision conditions of each node.

The parallelization of the CART algorithm was completed through the Spark platform. Firstly, the attendance sample data was read into the distributed dataset through Spark SQL and spread into each worknode in the cluster. After being stored in the memory of the work nodes, these samples would be calculated in the memory if they are within the limit of the maximum amount of memory space. Otherwise, they would be stored in local space. Then each node performed operations of the CART algorithm on allocated data through a series of mapping and aggregation and sent the results to the shared data structure. Finally, the decision tree model training was completed. This method is more efficient than the serial calculation as a large number of calculations were completed in the memory of each work node, and a small number of global calculations through network shared data.

Algorithm 1 Construction Algorithm of Regression Tree

INPUT: Training sample set D

OUTPUT: Regression tree f (x)

Divide the data space of D recursively into two sub-regions and set the output value on each region:

1. Select the optimal splitting attribute j and the attribute value of segmentation point s, and find

$$\min_{(j,s)}[\min_{c1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2]$$

Traverse variable j, scan the segmentation point s of fixed split attribute j, and select (j, s) to satisfy the formula above.

2. Divide the data space with the selected (j, s) and set the output value:

$$R_1(j,s)=\{x\mid x^{(j)}\le s\},\ R_2(j,s)=\{x\mid x^{(j)}>s\}$$

$$c_m=\frac{1}{N_m}\sum_{x_i\in R_m(j,s)}y_i,\ x\in R_m,\ m=1,2$$

3. Apply Steps 1 and 2 on the two sub-regions until the stop condition is met (the parameter is less than the threshold or the data in the data region belongs to the same category)

4. Divide the input data space D into M regions R1, R2, ...RM, to generate a decision tree

$$f(x)=\sum_{m-1}^{M}c_m I(x\in R_m)$$

4 ATTENDANCE DATA PROCESSING METHOD AND REGRESSION ANALYSIS

4.1 Cleaning and Aggregation of Attendance Data

Because of the application method, management mode, or even the system itself, and null values in some fields or repeated records, there is no guarantee that every record in the database of the attendance management system is accurate and in line with the standard. Therefore, data cleaning should proceed first in case it affects the later data processing and analysis. The problem of null field values can be solved by setting default values or shielding records and duplicate records by deleting records after the comparison.

Employee attendance data was stored in the form of a two-dimensional table, and its main fields include ID number, department name, attendance value, recording time, and others. A record in the table represents the attendance of an employee in a department on a particular occasion, where the meanings of the attendance values are: 1 attendance, 2 sick leave, 3 personal leave, 4 absenteeism, 5 late, 6 leave early, 7 special circumstances. To analyze data for attendance rate, we need to aggregate the attendance values into attendance rate and collect the values in the form of attendance times/record times while the ID number and department name are used as classification fields. After aggregation, the attendance table was formed, and only the ID number, department name, and attendance rate were kept in the data.

4.2 Association and Conversion of Attendance Data

The employee attendance database also stores other relevant information, such as work arrangement, overtime hours, and employee information, which may have a certain relationship with employee attendance, but remain unknown before data analysis. Retaining too many invalid fields will affect the efficiency of data analysis, so it is necessary to filter the fields in these tables and then associate them with the attendance data. Preliminary screening was proceeded in numerous fields to exclude the fields that are obviously irrelevant to the attendance rate until the fields such as workdays, working hours, department attributes, required hours, assessment methods, and employee categories were screened out, and then the results of the previous step of aggregation were associated with them.

Database fields that are not numeric types require code conversions. The conversion method is to use the Map ( ) method of distributed data set to convert non-numerical data into real data. For example, the workday field values from Monday to Friday are 1-5, the working hours are 1-8, and the department attribute 1 represents "leadership," 2 represents "department," and 3 represents "research office." Staff category 1 represents "management personnel," 2 represents "technical personnel," and 3 represents "skilled personnel." After the above data processing, data formats that meet the requirements of the decision tree algorithm were obtained, as shown in Table 2, where Label is the target variable of decision tree regression analysis, and Feature is the data attributeto be classified [7].

Table 2 Data Format Table After Processing

Label	Feature				
Attendance Rate	Workday	Working Hours	Required Hours	Department Attributes	Staff Category
0.96	1	10	9	2	1
0.87	5	8	8	3	3

4.3 Processing and Sample Selection of imbalanced Attendance Data

According to the global attendance statistics, full attendance was found to be 82% in the database. That is to say, the records of attendance for 100% are far more than that of non-full attendance, which is called the class imbalance problem. If the decision tree is built without data processing, most branches will output the same results. And it will be meaningless to set up a decision tree. The main method to solve this problem is conducting data resampling, which is divided into over-sampling and under-sampling. The former is to increase data for the classification with a small number of samples. As additional system overhead is required to increase data, this method is suitable for the case with a small total data amount [8]. This paper adopted Under-Sampling to compare the proportion of non-full attendance records and carry out random Sampling on full attendance records to solve the imbalanced data problem.

4.4 Regression Analysis of Attendance Data

In this paper, 300,000 attendance records in the database were selected as the original data, and the decision tree algorithm was applied for regression analysis. The purity measurement method was set as Variance Impurity= "Variance," and the maximum depth of the decision tree was MaxDepth=5. The set value is a reasonable setting based on a large number of experiments, which can ensure that the calculated results are reliable and effective in quantity. The decision tree was built with five layers and 21 nodes through the calculation, as shown in Figure 2. If the default parameter value is set for operation, over-fitting will occur, and a large number of nodes in the decision tree will be generated. Here the test samples were used to test the decision tree to gain an RMSE (root mean square error) value of 0.164, indicating a successful prediction on the decision tree.

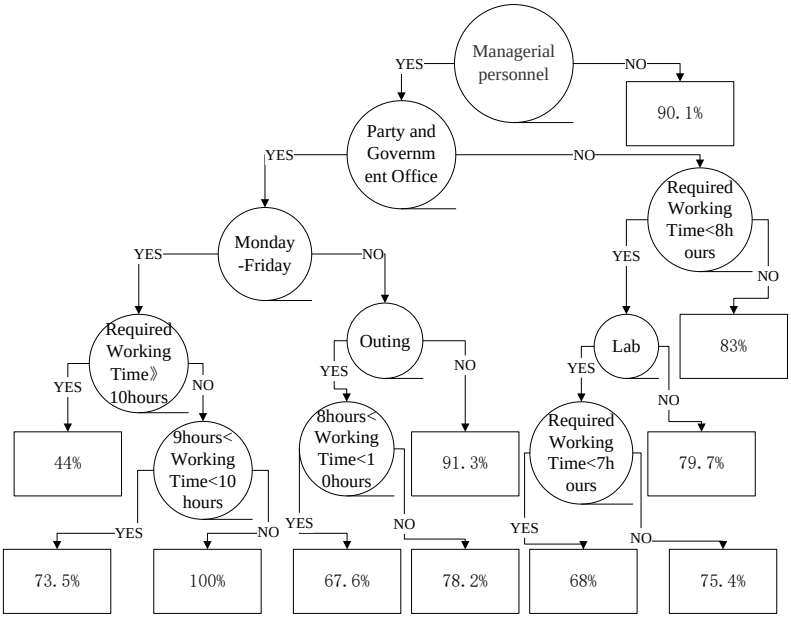

Figure 2 Decision Tree for Employee Attendance

According to Figure 2, the factors affecting employee attendance in the selected data are listed in descending order by influence degree: staff category, department attribute, workday, required time, and working time. The meaning of the decision tree is simple and clear, but the interpretation of its meaning is not unique. From the perspective of employee management, the meaning of the decision tree is summarized as follows: firstly, the attendance rate of management personnel and technical personnel is relatively stable and higher than that of skilled personnel; secondly, the attendance rate of the party and government offices is better than that of the research office; thirdly, the attendance on Friday is worse than other weekdays.

Advice could be taken from the data analysis for employee management. As staff attendance is the foundation of good working effect, we can work from the following points of view to improve the attendance. From the perspective of the staff category, attention should be paid to the attendance rate of skilled workers; from the perspective of department attributes, the attendance rate of staff in the laboratory; from the perspective of workday, the attendance rate on Friday. For the scheduled work, an early warning should be given according to this model before they go to work.

4.5 The Efficiency Analysis of Data Processing

Spark, adopting parallel memory computing, is more efficient than serial computing and traditional database technologies when processing large-scale data. The following is a computing task in data processing, namely, attendance data aggregation, where the data was processed by traditional database technology and Spark, respectively, and the calculation time of different data volumes were compared, as shown in Table 3.

Table 3 Comparative Tables for Time Consuming of Data Processing

Number of Records (piece)	Traditional SQL (ms)	Spark (ms)
100	3	4,387
1,000	11	4,682
10,000	79	4,708
100,000	2,179	4,765
300,000	7,198	5,233

A computer with the same configuration was used for the test, whose CPU is i5-4200U and memory is 4G. The traditional SQL technology is MySQL5.0, and the aggregation was completed by SQL statements. Spark-2.2.0-bin-hadoop2.7 was selected to run in Standalone mode, and Spark SQL was adopted to complete the aggregation. The test results show that the traditional SQL technology takes less time than Spark when the amount of data is small, and Spark has an advantage when the amount of data is more than 300,000. In addition, there is little difference in the time when the amount of data processed on Spark is less than 100,000, indicating that the overhead of processing 100 to 100,000 records on Spark is similar in job processing, but less in task calculation. The practice has proved that Spark is efficient as expected during the attendance data processing and that the traditional SQL technology can complete small-scale statistical analysis, but not the large-scale analysis as it will be time-consuming or crushing the computing task.

5 CONCLUSION

This paper adopted Spark and the decision tree algorithm to process and analyze the attendance data of employees, revealing the factors affecting employee attendance and their relationship. To start a global data analysis, the data from a relational database can be processed by cleaning, aggregating, associating, and transforming, and the problems of imbalanced data and sample selection should also be addressed. The practice shows that the attendance data processing and analysis method proposed in this paper is reasonable, feasible, and efficient. Further research and improvement are required in data analysis methods and algorithm optimization.

REFERENCES

[1] He L. (2013) Risk management of enterprise information technology outsourcing projects. Spacecraft Environment Engineering, 30 (04): 446-451.

[2] Zhang, Y, Chen M, Liao X F. (2013) Status and prospects of big data applications. Journal of Computer Research and Development, 50 (S2): 216-233.

[3] He L, Wu L D, Cai Y C. (2007) Clustering algorithms in data mining. Application Research of Computers, {4} (01): 10-13.

[4] Liu H Y, Chen J, Chen G Q. (2002) A review of data classification algorithms in data mining. Journal of Tsinghua University (Science and Technology), {4} (06): 727-730.

[5] Li X L, Gong H G. (2015) A review of big data systems. Science China Information Science, 45(01): 1-44.
[6] Wang, X N. (2016) The study on Spark-based parallelism of classification regression tree algorithm. Chongqing: Chongqing University.
[7] WipawanBuathong. (2016) TLiSVM (Triple Linear SVM Weight) for dimensionality reduction. Computer Science & Education(ICCSE), 2016 11th International Conference, Nagoya, Japan.
[8] Ping G, Yang Y. (2017) An oversampling algorithm for minority class subdivision in imbalanced data sets. Journal of Computer Engineering.43 (2), 241~247.