Large Imbalance Data Classification Based on MapReduce for Traffic Accident Prediction

Seoung-hun Park
Department of Computer Science and Engineering
Konkuk University
Seoul, Republic of Korea
wolfire@konkuk.ac.kr

Young-guk Ha*
Department of Computer Science and Engineering
Konkuk University
Seoul, Republic of Korea
ygha@konkuk.ac.kr

Abstract— **In modern society, our everyday life has a close connection with traffic issues. One of the most burning issues is about predicting traffic accidents. Predicting accidents on the road can be achieved by classification analysis, a data mining procedure requiring enough data to build a learning model. Regarding building such a predicting system, there are several problems. It requires lots of hardware resources to collect traffic data and analyze it for predicting traffic accidents since the data is very huge. Furthermore, data related to traffic accidents is few comparing with data which is not related to them. The numbers of two types of data are imbalanced. The purpose of this paper is to build a predicting model that can resolve all these problems. This paper suggests using Hadoop framework to process and analyze big traffic data efficiently and a sampling method to resolve the problem of data imbalance. Based on this, the predicting system, first of all, preprocess traffic big data and analyzes it to create data for the learning system. The imbalance of created data is corrected by a sampling method. To improve predicting accuracy, corrected data is classified into several groups, to which classification analysis is applied. These analysis steps are processed by Hadoop framework.**

Keywords-component; Imbalance data; Accident prediction; Big-data inference; MapReduce; Classification;

I. Introduction

The prediction through data analysis is composed of mainly the classification analysis through learning for data of the past in data mining. The classification analysis learns the training data set and creates a standard predicting model for prediction result. Based on these, the model predicts the result to imitate existing content.

Making a new model has some problems. The first problem is about imbalance data. Imbalance data means data that have a huge difference between the observed sizes from one data set. To solve this problem, sampling techniques can be used. There are two kinds of sampling techniques, under-sampling and over sampling [1]. Under-sampling is to use all of observation value in a small class and to use part of observation value in a big class. And over-sampling is to use all of observation value in a big class and to increase size of observation value in a small class and use this value. Under-sampling calculates to use a lost data. Such sampling techniques can speed up processing data but cannot help losing the reliability of data. On the other hand, over-sampling utilizes all data but requests more resources for processing additional data [2].

The second problem is about data processing to create training data set. The training data set has a set of multiple features that can affect the prediction result. Therefore, processing the dataset takes a lot of time for the many different types of data in it and makes more resources needed depending on the data size.

We want to solve these two problems by using MapReduce algorithm. And this paper presents processing steps of a parallel classification based on MapReduce and shows the validity of these steps.

MapReduce is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster, which processes big data efficiently [3]. Efficient processing of big data facilitates the analysis of big data. So, many researches for big text data processing used MapReduce platforms [4]. Using MapReduce, processing performance for generating training data set and conducting classification has been improved, which can resolve problems about a high processing overhead and low processing speed [5].

Our proposed processing steps consist of five steps. After preprocessing target data sets, it combines the data sets and creates training data sets. Over-sampling technique is used to solve imbalance data problem in the training dataset. And then, for generating learning model, it runs classification based on training datasets. Finally, it verifies the accuracy of the result for processing.

We have implemented proposed processing steps. We selected accidents and non-accidents data related to traffic in a lot of imbalance data. This data is related to the actually highway, and the size of the data is about 300GB.

The outline of this paper is organized as follows: Section 2 explains highway traffic data used in the experiment and section 3 explains proposed processing steps. In Section 4, we are showing experiments. Finally, Section 5 will give conclusions and future works.

[*] Corresponding Author

978-1-4799-4331-9/14 $31.00 © 2014 IEEE
DOI 10.1109/IMIS.2014.6

II. Highway Traffic Data

The data that I used for the experiment is from the Korea Highway Corporation. The data are text files which contain traffic data created between Jan. 1st, 2011 and Jun. 30th, 2013 on the Gyeongbu line which connects Seoul with Busan. I will explain the structure of the data.

A. *Traffic data*

The traffic data are created by VDS, Vehicle Detection System, which measure, every 30 seconds, speeds of cars and record the number of cars that run on it. Table 1 shows the format of the data.

TABLE I. Format of Traffic data

Variable	Format
Time	YYYYMMDDhhmmss
VDS ID	ID, String
Number of line	Integer
Traffic volume	Integer
Traffic density	Real number
Average speed	Real number

Time means when the data was recorded. VDS ID means each loop coil's ID. Line number means which line the information came from. The volume of traffic means the number of cars for 30 seconds. The occupation rate means the number of cars for 30 seconds depending on the length of the section. The average speed means the average speed of cars that pass for 30 seconds. The sizes of the whole data in each year are shown in Table 2.

TABLE II. Data sizes on Year and Way

Year	Way	Size(GB)
2011	Seoul	62.3
	Pusan	58
2012	Seoul	62.6
	Pusan	56.6
2013(~Jun.)	Seoul	31.1
	Pusan	28.8

B. *Accident & Non-accident data*

The accident data are recorded by the police firsthand, and the format of it is shown in Table 3. The data are written to the minute. What day and where each accident happened are also recorded in the data. And the death toll, the number of casualties and information about people related to each accident are recorded in the data. The weather condition and shapes of lines are in the data in the form of categories' ID. The number of accidents in the whole data is 1,888. 848 accidents happened in 2011, 741 accidents in 2012, and 299 accidents in 2013.

TABLE III. Format of Accident & Non-accident data

Variable	Format
Time	YYYYMMDDhhmmss
Day	Category
Position	Real number (Km)
Death	Integer
Casualty	Integer
People#1	Information
People#2	Information
Weather	Category
Accident type	Category
Road shape	Category
Road alignment	Category

III. Propose Method

The classification analysis process of imbalance data prediction has five steps, and the whole process is shown as Figure 1. The overall system based on Hadoop [6], and with implementing data pre-process, learning data creation, over sampling by Hive [7]. The cluster and classification analysis operated by Mahout [8]. In this paper, a novel method to determine each operation step for classification analysis of traffic accident prediction will be explained in detail.

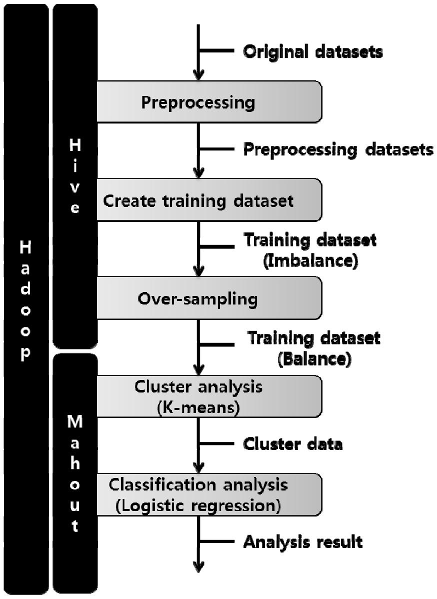

Figure 1. Steps of proposed Classification analysis process.

A. *Preprocessing*

A data pre-process was required to process the variable selection of data for training and each ideal value of the variables. The variables of data were selected by according to each variable characteristic and omitted variable in the total data. The ideal value process was replaced to omitted or initial values by considering each value of the variables. Also, the formatting was necessary to match the data.

B. Create training dataset & Over-sampling

To build a learning data, the pre-processed data were combined, and formed a unit learning data model then, the data were modified to implement the final learning data. The features which affected to target variables were always considered when the learning data created. Therefore, Identification (ID) or index value was applied to use the variables of data effectively when the learning data were created.

Unfortunately, these learning data certainly contained the data unbalance. To overcome this problem, an over-sampling operation was processed to repair the data.

C. Cluster analysis

Because of unique characteristics of data, a cluster classification analysis surely brought high accuracy in process because it was operated by partitioning several clusters and processed by clusters rather than individual classification analysis method [9]. In that, influences of each cluster characteristic were obviously recognized when the data clustered [10].

When the cluster analysis operated, the numbers of cluster have decided, and the data ratio of cluster have confirmed. And also, finding each representative value of the clusters have processed.

D. Classification analysis

A classification analysis algorithm which selected the created learning data built learning model. It was same as a criterion of categorizing data which based on learning data, by using this model, the results could be predicted with choosing estimated target variable value when the data which formed with estimated variables inputted [11].

The classification analysis was processed with several clusters which created by cluster analysis. Each cluster was applied different logistic regression method [12], and the results were collected to present a contingency table.

E. Prediction accuracy

We used the total, and target precision as reference to determine the performance. As Table 4 showed, the contingency table was utilized, and the equations for the results were calculated and recorded below.

TABLE IV. CONTINGENCY TABLE

		Prediction	
		True	False
Actually	True	TP	FN
	False	FP	TN

$$\frac{(TN+TP)}{(TN+TP+FN+FP)} \tag{1}$$

$$\frac{(TP)}{(TP+FN)} \tag{2}$$

$$\frac{(FP)}{(FP+TN)} \tag{3}$$

The (1) were showed accuracy, (2), and (3) referred "true positive rate", and "false positive rate" respectively.

IV. Experiment

We experiment our method based on MapReduce for traffic prediction using actually highway traffic data. The system was built on Hadoop and increased processing performance using multi-node. Our system's environment is shown below Table 5.

TABLE V. SYSTEM ENVIRONMENT

Hadoop Environment			
Hadoop version	Hadoop 1.0.4		
Nodes	NameNode	2ndNameNode	DataNode
	1	1	30
Hive version	Hive 0.12.0		
Mahout version	Mahout 0.8		
Node Environment			
CPU	3.10 GHz quad-core		
Memory	16 GB		
OS	Ubuntu Server 12.04 LTS (64-bit)		

A. Preprocessing

I input the accident and traffic data to Hive and process them with Hadoop. I removed duplicated and unclear data in 1,888 pieces of the whole data, and 1,421 pieces of the data are left.

B. Create training dataset & Over-sampling

I also used Hadoop and Hive to create learning data and sorted preprocessed data with Hive query to fit them for Hadoop. Table 6 shows the model of training. The number of created accident and non-accident data is 1,023,120, which is shown in Table 7. In Hadoop process, 1,161 mapper were created, and reducers 300. It took 768 seconds to finish the process.

The accident data accounted for 0.14% and the non-accident data accounted for 99.86%, which shows the imbalance of the data. To solve this problem, as increasing the number of the accident data, I checked results of classification analysis. The biggest change in the results was observed when the rate of the accident data was between 28% and 30%. I stopped duplicating the data at 28.6%. The details of the result are shown in Table 7.

TABLE VI. SCHEMA OF TRAINING DATASET

Column	Type	Variable
adate	SMALLINT	Month
atime	SMALLINT	Hour
aday	SMALLINT	Day
shape	SMALLINT	Road shape
linear	SMALLINT	Road alignment
weather	SMALLINT	Weather
line	SMALLINT	Road number
accident	BOOLEAN	Accident happen
No[1~4]_30	FLOAT	Velocity 30sec ago
No[1~4]_60	FLOAT	Velocity 60sec ago
No[1~4]_90	FLOAT	Velocity 90sec ago
No[1~4]_120	FLOAT	Velocity 120sec ago
No[1~4]_150	FLOAT	Velocity 150sec ago
No[1~4]_180	FLOAT	Velocity 180sec ago

TABLE VII. NUMBER OF TRAINING DATASET BEFORE & AFTER SAMPLING

	Before Sampling	After Sampling
Accident	1,421(0.14%)	409,821(28.6%)
Non-accident	1,023,120(99.86%)	1,023,120(71.4%)
Total	1,024,541(100%)	1,432,941(100%)

C. K-means cluster analysis

TABLE VIII. DISTRIBUTION OF DATA IN CLUSTER

Clusters	Number of all	Number of accident
Cluster #0	177,143	55,384(31.27%)
Cluster #1	134,356	39,142(29.13%)
Cluster #2	96,405	25,637(26.59%)
Cluster #3	111,535	28,656(25.69%)
Cluster #4	110,681	32,419(29.29%)
Cluster #5	125,904	36,722(29.17%)
Cluster #6	117,493	31,695(26.98%)
Cluster #7	140,505	35,564(25.31%)
Cluster #8	117,027	36,957(31.58%)
Cluster #9	107,522	28,833(26.82%)
Cluster #10	92,193	25,808(27.99%)
Cluster #11	102,177	33,004(32.30%)
Total	1,432,941	409,821(28.6%)

When the data were divided into 12 groups with K-means algorithm, the result were the most efficient. The accident data were distributed to each group. Table 8 shows the distribution and the rate of the accident data. The rate ranges from 32.30% to 25.31%, which shows the distribution was achieved properly. The result was more accurate when classification analysis was achieved with a set of groups that represent each feature.

D. Logistic regression analysis & Prediction accuracy

Lastly, I conducted logistic regression on the 12 groups, the accuracy of which is shown in Table 9. Based on this result, I got the 80.56% whole prediction accuracy. And the positive accuracy is 42.71%. This accuracy is what I tried to get with over-sampling technique.

TABLE IX. CONTINGENCY TABLE OF LOGISTIC REGRESSION RESULT

		Prediction		
		Accident	Non-accident	Total
Actuality	Accident	175,015	234,806	409,821
	Non-accident	43,782	979,338	1,023,120
	Total	218,797	1,214,144	1,432,941

We compared other experiments. First experiment, it is classification for score of categories and products review in ACOM (Arabic Corpus for Opinion Mining) using SVM algorithm and under-sampling [13]. And second experiment is classification for continuous insurance contract using logistic regression analysis and over-sampling. The comparison result is shown in Table 10. Our experiment result showed high accuracy than first experiment and showed low accuracy than second experiment. But our experiment result showed 4% high aimed accuracy than second experiment. Our dataset is more imbalances and the dataset size is bigger than other experiments. Our method can be better than other experiments depending on the data.

TABLE X. COMPARISON WITH OTHER EXPERIMENTS

	First experiment	Second experiment	Our experiment
Sampling	Under	Over	Over
Classification	SVM	Logistic regression	Logistic regression
Number of dataset	1,846	40,000	1,024,541
Positive	1,702 (92.2%)	1,557 (3.9%)	1,421 (0.14%)
Negative	145 (7.8%)	38,443 (96.1%)	1,023,120 (99.86%)
Accuracy	73.63%	84.77%	80.56%
Aimed accuracy	-	39.4%	42.71%

V. Conclusion

This article suggested a data mining process for classification of imbalance data based on MapReduce, and by applying the process, a classification analysis has accomplished to predict traffic accidents with the highway traffic data. As a result, the performance of data mining process was tested by total, and target precision. The imbalance data problem has modified with over-sampling method to prevent the biased result in classification analysis of the imbalance data, and by using Hadoop with MapReduce, the big traffic data was efficiently processed therefore, learning data creation, cluster analysis, and classification analysis were performed.

According to the experiment, the total, and target precision was 80.56% and 42.71% respectively, which showed almost no difference with other papers. Further, Hadoop was efficient to process the big data, and to create the learning data which could be a solution for continuously increased data in the future. The precision can be improved by adding the locations of traffic accidents, and information of the highway in the data. Additionally, the boosting technique which enhances the precision of classification analysis can be also applied to extract better result.

There are few developments need to be addressed in the future works. First, the efficient over-sampling standard should be established. Second, the process time and the accuracy of analyzed results should be compared when Hadoop processed several classification analyses. Third, other testing methods will be needed to evaluate the analyzed results. Finally, real time data learning and prediction cannot be applied with Hadoop, since it is impossible to process real time data. Therefore, other researches are required to find a new method to solve the real time process problem.

Acknowledgment

This research was supported by MSIP(Ministry of Science, ICT&Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2013-(H0301-13-1012)) supervised by the NIPA(National IT Industry Promotion Agency)

References

[1] Kamei, Y., Monden, A., Matsumoto, S., Kakimoto, T., Matsumoto, K.-i.,"The Effects of Over and Under Sampling on Fault-prone Module Detection", Empirical Software Engineering and Measurement, 2007. ESEM 2007. First International Symposium. Madrid, Sept. 2007, pp. 196-204.

[2] Gothenberg, A., Tenhunen, H., "Performance analysis of low oversampling ratio sigma-delta noise shapers for RF applications", 『Circuits and Systems, 1998. ISCAS '98. Proceedings of the 1998 IEEE International Symposium』, 1998.

[3] Dean, J., Ghemawat, S., "MapReduce: simplified data processing on large clusters", Communications of the ACM. New York, vol. 51, Jan. 2008, pp. 107-113.

[4] T, Lee., H, Kim., K-H, Rhee., S-U, Shin., "Implementation and Performance of Distributed Text Processing System Using Hadoop for e-Discovery Cloud Service", Innovative Information Science & Technology Research Group (ISYOU), Oct. 2013, pp. 12-24.

[5] Fan Zhang, Sakr, M., "Dataset Scaling and MapReduce Performance", Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2013 IEEE 27th International. Cambridge, May 2013, pp. 1683-1690.

[6] Apache™, "Apache™ Hadoop", http://hadoop.apacahe.org.

[7] Apache™, "Apache™ Hive", http://hive.apacahe.org.

[8] Apache™, "Apache™ Mahout", http://mahout.apacahe.org.

[9] Vapnik, V. N., "Statistical Learning Theory", Wiley, 1998.

[10] Kukar, M., "Transduction and typicalness for quality assessment of individual classifications in machine learning and data mining", Data Mining, 2004. ICDM '04. Fourth IEEE International Conference, Nov. 2004, pp. 146-153.

[11] Raghavendra,P.S., Chowdhury, S.R., Kameswari, S.V., "Comparative study of neural networks and k-means classification in web usage mining", Internet Technology and Secured Transactions (ICITST), 2010 International Conference. London, Nov. 2010, pp. 1-7.

[12] Rahayu, S.P., Purnami, S.W., Embong, A., "Applying Kernel Logistic Regression in data mining to classify credit risk", Information Technology, 2008. ITSim 2008. International Symposium. Kuala Lumpur, Aug. 2008, pp. 1-6.

[13] Mountassir, A., Benbrahim, H., Berrada, I., "An empirical study to address the problem of Unbalanced Data Sets in sentiment classification", Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference. Seoul, Oct. 2012, pp. 3298-3303.