Life Expectancy and Mortality by Race Using Random Subspace Method

Yong Gyu Jung, Dong Kyu Nam and Hee Wan Kim

Abstract: In recent years, massive data is pouring out in the information society. As collecting and processing of such vast amount of data is becoming an important issue, it is widely used with various fields of data mining techniques for extracting information based on data. In this paper, we analyze the causes of the difference between the expected life expectancy and the number of deaths by using the data of the expected life expectancy and the number of the deaths. To analyze the data effectively, we will use the REP-tree for a small and simple size problem using the Random subspace method, which is composed of random subspaces. The performance of the REP-tree algorithm was analyzed and evaluated for statistical data.

Index Terms: Random subspace, REP-tree, racism, life expectancy, random subspaces

I. INTRODUCTION

In the wake of the westernization boom in the United States in the early 1900s, the cultural atmosphere of the West and the East began to vary greatly. Discrimination against blacks who were slaves in the western territories became a major problem in the United States. Blacks in the 1900s went to school under the guards of soldiers to be protected from the threats of white people, were economically very poor, and suffered extreme stress from racial discrimination. With the efforts of President Martin Luther King and President Johnson in 1965, blacks have completely restored their legal rights, resulting in a significant reduction in racial discrimination, but so far white racism against blacks still exists. In order to examine the effects of racial discrimination on human life expectancy and mortality, it is analyzed the life expectancy and mortality statistics of black and white men in the United States from 1900 to 2013. It is important to identify race-by-race comparative analyzes to analyze race-by-race life expectancy and deaths. Therefore, this study aims to analyze the statistical situation using data mining. To effectively analyze the data, we will use the REP-tree for a small and simple size problem using the Random subspace method, which is composed of random subspaces. We also analyze and evaluate the performance of REP-tree algorithm for statistical data

II. RELATED RESEARCH

A. *Random Subspace method (RSM)*

In machine learning the random subspace method, also called attribute bagging or feature bagging, is an ensemble learning method that attempts to reduce the correlation between estimators in an ensemble by training them on random samples of features instead of the entire feature set. In ensemble learning one tries to combine the models produced by several learners into an ensemble that performs better than the original learners. One way of combining learners is bootstrap aggregating or bagging, which shows each learner a randomly sampled subset of the training points so that the learners will produce different models that can be sensibly averaged. In bagging, one samples training points with replacement from the full training set.

The random subspace method is similar to bagging except that the features are randomly sampled, with replacement, for each learner. Informally, this causes individual learners to not over-focus on features that appear highly predictive /descriptive in the training set, but fail to be as predictive for points outside that set. For this reason, random subspaces are an attractive choice for problems where the number of features is much larger than the number of training points, such as learning from fMRI data or gene expression data.

The random subspace method has been used for decision trees; when combined with "ordinary" bagging of decision trees, the resulting models are called random forests. It has also been applied to linear classifiers, support vector machines, nearest neighbours and other types of classifiers. This method is also applicable to one-class classifiers. Recently, the random subspace method has been used in a portfolio selection problem showing its superiority to the conventional resampled portfolio essentially based on Bagging.

RSM uses the method of random subspaces as the configuration / aggregation classifier. Therefore, when the data has rich features, it has a better classifier function than the original feature space classifier. Random extraction requires more modification to the learning algorithm than bagging, but it can be applied to more learning devices. In addition, the classifiers learned through the use of bugs are quite accurate individually, but their low diversity can compromise the accuracy of the overall ensemble. At this time, if the ensemble members can be variously and accurately made using the RSM technique, a smaller ensemble will be used, which will result in a great gain in computation. Of course, RSM can be integrated with the bagging technique to introduce randomization into the learning process in terms of both instances and attributes.

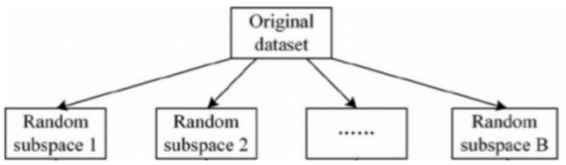


Fig. 1 Random subspace random extraction

Revised Manuscript Received on March 02, 2019.

Yong Gyu Jung, Dept. of Medical IT, Eulji University, Korea
Dong Kyu Nam, CEO, Project Jung, Co ltd., Korea.
Hee Wan Kim, Division of Computer-Mechatronics, Shamyook University, Korea

The RSM processes the training data in a specific space and the RSM procedure is as follows.

1. Repeat b = 1, 2, ..., B:

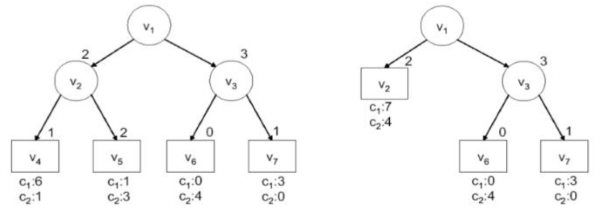
p-demensional feature Selects a random subspace of r-dimensional from space

Construct a classifier. (Crystal boundary = *0*

2. Combine the variables categorized by the major majority decision of the final decision rule. Is the kronecker symbol and y *{-1, 1}* is the decision class label of the classifier.

B. REP-tree Algorithm

REP-tree can be used only in a numeric tree. It constructs a decision tree or a regression tree using information acquisition / variance reduction as a quick decision tree learner, and truncates it using a reduced error pruning.



$$E(T_{v_2}) = 3, E(v_2) = 2, E(T_{v_3}) = 1, E(v_3) = 3.$$

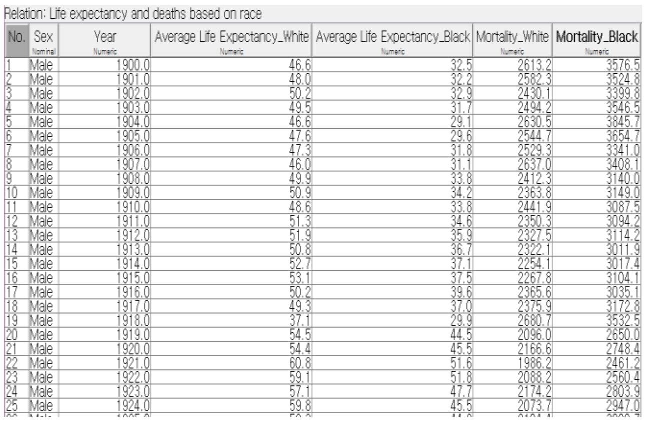
Fig. 2 REP-tree

The missing values are handled using C4.5, a way to deal with fragmented instances, and REP-tree, which is optimized for speed, divides the instances into fragments to handle missing values. The user can set the minimum number of instances per leaf node, the maximum tree depth, the training set dispersion LCHLTH ratio, and the number of layers for pruning, and provides UCI repository and confusion metrics.

III. EXPERIMENT

A. Experimental data

We used WEKA developed by Waikato University as a tool for the experiment. The data used are statistical data on life expectancy and mortality for men in races in the United States, Life expectancy and deaths based on race.



Relation: Life expectancy and deaths based on race

No.	Sex Nominal	Year Numeric	Average Life Expectancy_White Numeric	Average Life Expectancy_Black Numeric	Mortality_White Numeric	Mortality_Black Numeric
1	Male	1900.0	46.6	32.5	2613.2	3576.5
2	Male	1901.0	48.0	32.2	2582.3	3524.8
3	Male	1902.0	50.2	32.9	2430.1	3399.8
4	Male	1903.0	49.5	31.7	2494.2	3546.5
5	Male	1904.0	46.6	29.1	2630.5	3845.7
6	Male	1905.0	47.6	29.6	2544.7	3654.7
7	Male	1906.0	47.3	31.8	2529.3	3341.0
8	Male	1907.0	46.0	31.1	2637.0	3408.1
9	Male	1908.0	49.9	33.8	2412.3	3140.0
10	Male	1909.0	50.9	34.2	2363.8	3149.0
11	Male	1910.0	48.6	33.8	2441.9	3087.5
12	Male	1911.0	51.3	34.6	2350.3	3094.2
13	Male	1912.0	51.9	35.9	2327.5	3114.2
14	Male	1913.0	50.8	36.7	2322.1	3011.9
15	Male	1914.0	52.7	37.1	2254.1	3017.4
16	Male	1915.0	53.1	37.5	2267.8	3104.1
17	Male	1916.0	50.2	39.6	2365.6	3035.1
18	Male	1917.0	49.3	37.0	2375.9	3172.8
19	Male	1918.0	37.1	29.9	2680.7	3532.5
20	Male	1919.0	54.5	44.5	2096.0	2650.0
21	Male	1920.0	54.4	45.5	2166.6	2748.4
22	Male	1921.0	60.8	51.6	1986.2	2461.2
23	Male	1922.0	59.1	51.8	2088.2	2560.4
24	Male	1923.0	57.1	47.7	2174.2	2803.9
25	Male	1924.0	59.8	45.5	2073.7	2947.0

Fig. 3 Experimental data

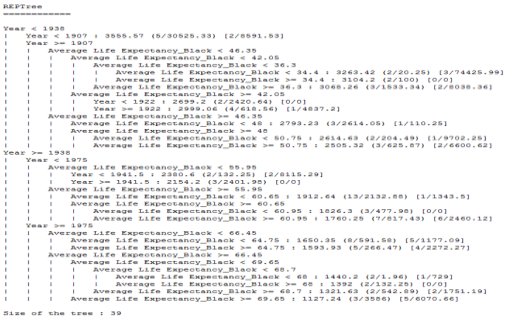
A total of 114 data were used for the experiment. This experimental data shows that six numeric attributes such as year, average life expectancy_white and sex representing sex are fixed as one constant 'male' as a property of life expectancy and number of deaths of male according to race in the United States symbloing nominal property for representing. Details of each property are shown in Tables.

Table 1 Experiment Data Attributes

Attribute	type	value
Sex	nominal	{male}
Year	numeric	continuous from 1900 to 2013
average life expectancy_white	numeric	continuous from 37.1 to 76.7
average life expectancy_black	numeric	continuous from 29.1 to 72.3
mortality_white	numeric	continuous from 859.2 to 2680.7
mortality_black	numeric	continuous from 1052.8 to 3845.7

B. Experimental Results

Experiment was based on the data of Life expectancy and deaths based on race.csv. We used REPTree as the detailed algorithm in the Random Subspace which selected the value attribute. We set debug to False, numlteration to 10, seed = 1, fords = 10, and subSpaceSize = 0.5.



```
REPTree
============

Year < 1938
|   Year < 1907 : 3555.57 (5/30525.33) [2/8591.53]
|   Year >= 1907
|   |   Average Life Expectancy_Black < 46.35
|   |   |   Average Life Expectancy_Black < 42.05
|   |   |   |   Average Life Expectancy_Black < 36.3
|   |   |   |   |   Average Life Expectancy_Black < 34.4 : 3263.42 (2/20.25) [3/74425.99]
|   |   |   |   |   Average Life Expectancy_Black >= 34.4 : 3104.2 (2/100) [0/0]
|   |   |   |   Average Life Expectancy_Black >= 36.3 : 3068.26 (3/1533.34) [2/8038.36]
|   |   |   Average Life Expectancy_Black >= 42.05
|   |   |   |   Year < 1922 : 2699.2 (2/2420.64) [0/0]
|   |   |   |   Year >= 1922 : 2999.06 (4/618.56) [1/4837.2]
|   |   Average Life Expectancy_Black >= 46.35
|   |   |   Average Life Expectancy_Black < 48 : 2793.23 (3/2614.05) [1/110.25]
|   |   |   Average Life Expectancy_Black >= 48
|   |   |   |   Average Life Expectancy_Black < 50.75 : 2614.63 (2/204.49) [1/9702.25]
|   |   |   |   Average Life Expectancy_Black >= 50.75 : 2505.32 (3/625.87) [2/6600.62]
Year >= 1938
|   Year < 1975
|   |   Average Life Expectancy_Black < 55.95
|   |   |   Year < 1941.5 : 2380.6 (2/132.25) [2/8115.29]
|   |   |   Year >= 1941.5 : 2154.2 (3/2401.98) [0/0]
|   |   Average Life Expectancy_Black >= 55.95
|   |   |   Average Life Expectancy_Black < 60.65 : 1912.64 (13/2132.88) [1/1343.5]
|   |   |   Average Life Expectancy_Black >= 60.65
|   |   |   |   Average Life Expectancy_Black < 60.95 : 1826.3 (3/477.98) [0/0]
|   |   |   |   Average Life Expectancy_Black >= 60.95 : 1760.25 (7/817.43) [6/2460.12]
|   Year >= 1975
|   |   Average Life Expectancy_Black < 66.45
|   |   |   Average Life Expectancy_Black < 64.75 : 1650.35 (8/591.58) [5/1177.09]
|   |   |   Average Life Expectancy_Black >= 64.75 : 1593.93 (5/266.47) [4/2272.27]
|   |   Average Life Expectancy_Black >= 66.45
|   |   |   Average Life Expectancy_Black < 69.65
|   |   |   |   Average Life Expectancy_Black < 68.7
|   |   |   |   |   Average Life Expectancy_Black < 68 : 1440.2 (2/1.96) [1/729]
|   |   |   |   |   Average Life Expectancy_Black >= 68 : 1392 (2/132.25) [0/0]
|   |   |   |   Average Life Expectancy_Black >= 68.7 : 1321.63 (2/542.89) [2/1751.19]
|   |   |   Average Life Expectancy_Black >= 69.65 : 1127.24 (3/3586) [5/6070.66]

Size of the tree : 39
```

Fig. 4 Random subspace execution screen –REP-tree

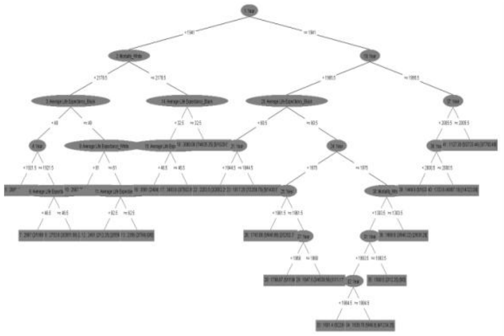


Fig. 5 Tree visualizer execution screen

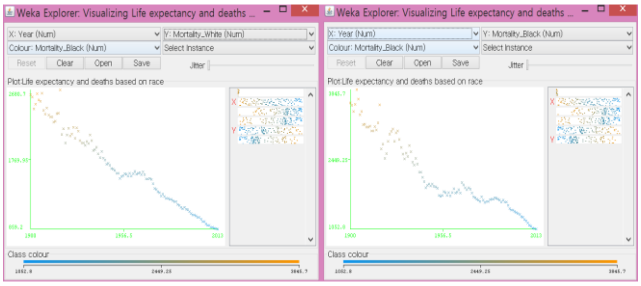
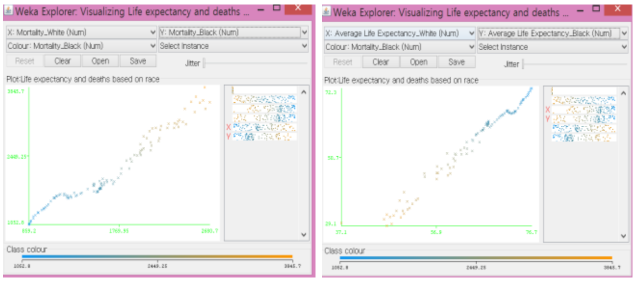

Weka Explorer: Visualizing Life expectancy and deaths ...
X: Year (Num)
Y: Mortality_White (Num)
Colour: Mortality_Black (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:Life expectancy and deaths based on race
Class colour
Weka Explorer: Visualizing Life expectancy and deaths ...
X: Year (Num)
Y: Mortality_Black (Num)
Colour: Mortality_Black (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:Life expectancy and deaths based on race
Class colour

Fig. 6 White death rate by year, black death rate by year

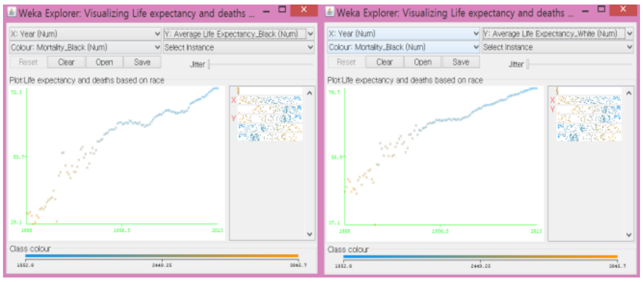
Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication



Weka Explorer: Visualizing Life expectancy and deaths ...
X: Mortality_White (Num) | Y: Mortality_Black (Num)
Colour: Mortality_Black (Num) | Select Instance
Plot:Life expectancy and deaths based on race

Weka Explorer: Visualizing Life expectancy and deaths ...
X: Average Life Expectancy_White (Num) | Y: Average Life Expectancy_Black (Num)
Colour: Mortality_Black (Num) | Select Instance
Plot:Life expectancy and deaths based on race

Fig. 7 Mortality and life expectancy of blacks and whites



Weka Explorer: Visualizing Life expectancy and deaths ...
X: Year (Num) | Y: Average Life Expectancy_Black (Num)
Colour: Mortality_Black (Num) | Select Instance
Plot:Life expectancy and deaths based on race

Weka Explorer: Visualizing Life expectancy and deaths ...
X: Year (Num) | Y: Average Life Expectancy_White (Num)
Colour: Mortality_Black (Num) | Select Instance
Plot:Life expectancy and deaths based on race

Fig. 6 Life expectancy of white people by year, life expectancy of black people by year

IV. EXPERIMENT RESULTS

Experimental results show that before 1965, there were 2574.066 deaths in black, 2016.634 in white death, and 558 death in black. At the same time, the life expectancy of a white person is estimated to be about 10 years longer than that of a white person. Starting in 1965, when black people had fully reached their legal rights, 1,545,133 black deaths and 1,117,24 white deaths were identified.

Table 2 Summary of analysis results

	Death before 1965	Death since 1965	Life expectancy prior to 1965	Life expectancy since 1965
black	2574.066	1545.133	48.15231	65.57755
White	2016.634	1197.243	58.92615	72.35714

As a result of the analysis, the number of deaths of black people decreased and the life expectancy increased, but at the same time the same phenomenon appeared to white people, but the number of white people and black people is six times as shown in Table 4, The death rate of black people is about 5.7% higher than that of black people. In addition, given the steady increase in the expected life expectancy and number of deaths among black men and white men in the United States, it is presumed that there is a combined effect of exogenous variables such as economic growth, improved quality of life.

Table 3 US population by race in 2010

Subject	Number 2010	Percent
White	223,553,265	72.4
Black or African American	38,929,319	12.6
American Indian and Alaska Native	2,932,248	0.9

V. CONCLUSION

Recently, the field of data mining has attracted attention in order to extract information that can be practically used based on a vast amount of data in various fields.Therefore, in this paper, we analyze the performance of REP-tree algorithm which can be used as a search algorithm in Random subspace, one of the data mining techniques. By performing Random subspace search, it is possible to solve the problem of small size and simple size composed by random subspace and to combine by various random spaces instead of one training set, and to obtain very good conclusion. In order to analyze more precisely in future, it is necessary to control exogenous variables, to construct environment factors similar to reality, to evaluate performance, to identify the relationship between algorithms using attribute data and reliable data Algorithm models will be studied.

ACKNOWLEDGMENT

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. 2016M3A9B694241).

REFERENCES

1. Frank, Eibe, Chang Chui, and Ian H. Witten. "Text categorization using compression models." (2000)
2. Witten, Ian H., and Eibe Frank. "WEKA-Waikato Environment for Knowledge Analysis." Internet: http://www. cs. waikato. ac. nz/ml/weka,[Mar. 2, 2008] (2000).
3. Kuncheva, Ludmila I., Marina Skurichina, and Robert PW Duin. "An experimental study on diversity for bagging and boosting with linear classifiers." Information fusion 3.4 (2002): 245-258.
4. Skurichina, Marina, and Robert PW Duin. "Bagging, boosting and the random subspace method for linear classifiers." Pattern Analysis & Applications 5.2 (2002): 121-135.
5. Kalmegh, Sushilkumar. "Analysis of WEKA data mining algorithm REPTree, Simple CART and RandomTree for classification of Indian news." Int. J. Innov. Sci. Eng. Technol 2.2 (2015): 438-446.
6. Suryanto, T., Haseeb, M., & Hartani, N. H. (2018). The Correlates of Developing Green Supply Chain Management Practices: Firms Level Analysis in Malaysia. Int. J Sup. Chain. Mgt Vol, 7(5), 316.
7. Polikar, Robi, et al. "Multimodal EEG, MRI and PET data fusion for Alzheimer's disease diagnosis." Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE. IEEE, 2010
8. Patel, Tejash, et al. "EEG and MRI data fusion for early diagnosis of Alzheimer's disease." Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE. IEEE, 2008.
9. Lary, David John, et al. "Holistics 3.0 for health." ISPRS International Journal of Geo-Information 3.3 (2014): 1023-1038.
10. Agrawal, Ankit, et al. "Five year life expectancy calculator for older adults." Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on. IEEE, 2016.
11. Nanni, Loris, Alessandra Lumini, and Sheryl Brahnam. "Survey on LBP based texture descriptors for image classification." Expert Systems with Applications 39.3 (2012): 3634-3641.

12. Lee, Jung-Min. "Validity of consumer-based physical activity monitors and calibration of smartphone for prediction of physical activity energy expenditure." (2013).
13. Summoogum, Jena Parameswari, and Benjamin Chan Yin Fah. "A Comparative Study Analysing the Demographic and Economic Factors Affecting Life Expectancy among Developed and Developing Countries in Asia." Asian Development Policy Review 4. 4 (2016): 100-110.
14. Prasad, B. Hari. "A Study on Commensal Mortality Rate of a Typical Three Species Syn-Eco-System with Unlimited Resources for Commensal." Review of Information Engineering and Applications 1. 2 (2014): 55-65.
15. Lai, Carmen, Marcel JT Reinders, and Lodewyk Wessels. "Random subspace method for multivariate feature selection." Pattern recognition letters 27.10 (2006): 1067-1076.
16. Kotsiantis, Sotiris. "Combining bagging, boosting, rotation forest and random subspace methods." Artificial Intelligence Review 35.3 (2011): 223-240.