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.
Life Expectancy And Mortality By Race Using Random Subspace Method

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

1. Repeat \( b = 1, 2, \ldots, B \):
   - \( p \)-dimensional 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.

![Fig. 2 REP-tree](image)

\[
E(T_{v_2}) = 3, \ E(v_2) = 2, \ E(T_{v_1}) = 1, \ E(v_1) = 3.
\]

**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.

![Fig. 3 Experimental data](image)

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 symboing nominal property for representing. Details of each property are shown in Tables.

**Table 1 Experiment Data Attributes**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>type</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>nominal</td>
<td>{male}</td>
</tr>
<tr>
<td>Year</td>
<td>numeric</td>
<td>continuous from 1900 to 2013</td>
</tr>
<tr>
<td>average life expectancy_white</td>
<td>numeric</td>
<td>continuous from 37.1 to 76.7</td>
</tr>
<tr>
<td>average life expectancy_black</td>
<td>numeric</td>
<td>continuous from 29.1 to 72.3</td>
</tr>
<tr>
<td>mortality_white</td>
<td>numeric</td>
<td>continuous from 859.2 to 2680.7</td>
</tr>
<tr>
<td>mortality_black</td>
<td>numeric</td>
<td>continuous from 1052.8 to 3845.7</td>
</tr>
</tbody>
</table>

**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.

![Fig. 4 Random subspace execution screen –REP-tree](image)

![Fig. 5 Tree visualizer execution screen](image)

![Fig. 6 White death rate by year, black death rate by year](image)
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


