Logistic Regression Models for Classifying Attributes

Contents

Classifying Categorical Attributes

As users specify two groups for comparison, a logistic regression algorithm classifies each attribute into highly similar attributes, highly different attributes and other attributes. If the two groups have similar distributions in an attribute, the attribute is classified as similar; if the two groups have different distributions in an attribute, the attribute is classified as different; if the two groups' distributions are not highly similar or different in an attribute, the attribute is classified as an other attribute. Details of how the classification is done is provided by Law et al. [1].

The following is the logistic regression model for classifying categorical attributes into the three classes. The model first computes the Bhattacharyya coefficient Bh [2] for an attribute, and uses the following formula to compute the probability that it is a highly similar attribute $P(S)$, a highly different attribute $P(D)$ or an other attribute $P(O)$. Bhattacharyya coefficient Bh [2] is a measure of the degree of overlapping between two distributions. The final class is the class to which the highest probability is assigned.

$$P(D) = \frac{\exp(-52.0298 * Bh + 46.2363)}{1 + \exp(-52.0298 * Bh - 46.2363) + \exp(-33.6953 * Bh + 31.0733)}$$

$$P(O) = \frac{\exp(-33.6953 * Bh + 31.0733)}{1 + \exp(-52.0298 * Bh - 46.2363) + \exp(-33.6953 * Bh + 31.0733)}$$

$$P(S) = 1 - P(D) - P(O)$$

The model is trained using the training set "bhClass.csv" in the "Classifying Attributes" folder. The training set was derived from the distribution pairs collected by Law et al. [1] from the R datasets. The above logistic regression model for classifying categorical attributes has a cross-validation accuracy of 78.2692%.

Reference

[1] Po-Ming Law, Rahul C Basole, and Yanhong Wu. 2018. Duet: Helping Data Analysis Novices Conduct Pairwise Comparisons by Minimal Specification. IEEE Transactions on Visualization and Computer Graphics (2018). https://doi.org/10.1109/tvcg.2018.2864526.

[2] Wikipedia. Bhattacharyya distance - Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Bhattacharyya_distance.

Classifying Numerical Attributes
(One-to-One and One-to-Multiple Comparisons)

The following is the logistic regression model for classifying numerical attributes into the three classes when the pairwise comparison being conducted is one-to-one or one-to-multiple. For one-to-one comparisons, the model computes the difference in attribute values and for one-to-multiple comparisons, it computes the difference between the attribute value of the one-object group and the mean for the multiple-object group. The difference ($difference$) is used as a predictor variable for determining whether an attribute should be classified as similar, different, or other. The model computes the probability that a numerical attribute is a highly similar attribute $P(S)$, a highly different attribute $P(D)$ or an other attribute $P(O)$ for one-to-one and one-to-multiple comparisons as follows:

$$P(D) = \frac{\exp(45.6985 * difference - 5.4464)}{1 + \exp(45.6985 * difference - 5.4464) + \exp(23.9107 * difference - 1.9907)}$$

$$P(O) = \frac{\exp(23.9107 * difference - 1.9907)}{1 + \exp(45.6985 * difference - 5.4464) + \exp(23.9107 * difference - 1.9907)}$$

$$P(S) = 1 - P(D) - P(O)$$

The model is trained using the training set "bhMeanDiffClass.csv" in the "Classifying Attributes" folder. Only the "meanDiff" and "class" columns were used in the training. The above logistic regression model for classifying categorical attributes has a cross-validation accuracy of 75.5789%.

Classifying Numerical Attributes (Multiple-to-Multiple Comparisons)

The following is the logistic regression model for classifying numerical attributes into the three classes for multiple-to-multiple pairwise comparisons. There are two predictor variables. First, difference in the means of the two groups ($difference$) is computed. Second, the Bhattacharyya coefficient (Bh) is computed for the attribute to approximate the degree of overlapping between the two attributes.

Using $difference$ and Bh, the model computes the probability that a numerical attribute is a highly similar attribute $P(S)$, a highly different attribute $P(D)$ or an other attribute $P(O)$ for one-to-one and one-to-multiple comparisons as follows:

$$P(D) = \frac{\exp(-50.6583 * Bh + 36.204 * difference + 40.0443)}{1 + \exp(-50.6583 * Bh + 36.204 * difference + 40.0443) + \exp(-30.0439 * Bh + 17.9018 * difference + 26.2568)}$$

$$P(O) = \frac{\exp(-30.0439 * Bh + 17.9018 * difference + 26.2568)}{1 + \exp(-50.6583 * Bh + 36.204 * difference + 40.0443) + \exp(-30.0439 * Bh + 17.9018 * difference + 26.2568)}$$

$$P(S) = 1 - P(D) - P(O)$$

The model is trained using the training set "bhMeanDiffClass.csv" in the "Classifying Attributes" folder. The above logistic regression model for classifying categorical attributes has a cross-validation accuracy of 85.6842%.

The original model proposed by Law et al. [1] uses only Bhattacharyya coefficient as a predictor variable and has a cross-validation accuracy of ~78%. By adding mean difference as an extra predictor variable, the above model achieved a higher cross-validation accuracy of 85.6842%.

Reference

[1] Po-Ming Law, Rahul C Basole, and Yanhong Wu. 2018. Duet: Helping Data Analysis Novices Conduct Pairwise Comparisons by Minimal Specification. IEEE Transactions on Visualization and Computer Graphics (2018). https://doi.org/10.1109/tvcg.2018.2864526.

Files in the "Classifying Attributes" Folder

Here, we describe the files in the "Classifying Attributes" Folder.

1. ***bhClass.csv***

It is the training data for the logistic regression model for classifying categorical attributes (see P.2).

2. ***bhMeanDiffClass.csv***

It contains the training data for two logistic regression models: 1) the model for classifying numerical attributes when the pairwise comparison is one-to-one and one-to-multiple (see P.3), and 2) the model for classifying numerical attributes when the pairwise comparison is multiple-to-multiple (see P.4).

3. ***"data" folder***

This folder contains the raw data collected by Law et al. [1] from the R datasets. It is for computing the "bh" column in the "bhClass.csv" file and the "bh" and "meanDiff" columns in the "bhMeanDiffClass.csv" file. Please refer to [1] for the details.

4. ***labels.csv***

It is for generating the "class" columns in both "bhClass.csv" and "bhMeanDiffClass.csv". Please refer to [1] for the details.

Reference

[1] Po-Ming Law, Rahul C Basole, and Yanhong Wu. 2018. Duet: Helping Data Analysis Novices Conduct Pairwise Comparisons by Minimal Specification. IEEE Transactions on Visualization and Computer Graphics (2018). https://doi.org/10.1109/tvcg.2018.2864526.