An Improved Naive Bayes Classifier on Imbalanced Attributes

S Geetha, Government Arts College (A), Kumbakonam, India
R Maniyosai, Government Arts College (A), Kumbakonam, India

ABSTRACT

Data plays a major and prominent role in this modern information era. Classification is a data mining task to discover the hidden information from large amounts of data stored in the repository. This process becomes extremely challenging in case of highly imbalanced dataset. Prediction from imbalanced attributes cannot be done accurately in the following case: During the training phase, the categorical variable is not observed but the test phase encounters the categorical variable and hence it assigns zero probability which leads to false prediction. To overcome this scenario, this article proposes a novel smoothing technique called optimized laplace smoothing estimation. This technique adds a bias value function to improve the accuracy of imbalanced attributes. For example, a child dataset has more attributes and the classification model is used to predict the child weight. Some of the attribute values may not be present in the child dataset due to which Naive Bayes assigns a zero for incomplete and an empty attribute. This leads to inaccurate prediction. In such cases, Naive Bayes can be further tuned by adding some new parameters as well as altering the existing optimization method. Experimental analysis shows that this novel smoothing technique enhances the classification accuracy by means of accurate predictions for imbalanced attributes.

KEYWORDS

Child Dataset, Classification, Imbalanced Attributes, Naive bayes

1. INTRODUCTION

Classification is a supervised (Caruana & Niculescu-Mizil, 2006; John & Langley, 1995) learning algorithm where the attribute labels are known. But in some cases (Nguyen et al., 2011), the attribute label is missing or the attribute label is incomplete which leads to an imbalance in the attributes. This dataset imbalance affects the prediction accuracy. Some of the applications which face this attributes imbalance problem are intrusion detection in networks, medical diagnosis, credit card fraud detection, document classification etc (Gonzalez-Abril et al., 2008). The main objective of classification algorithm is that the model constructed using the training data should improve the prediction accuracy. The performance of the classification algorithms achieves better classification accuracy since it does not consider the imbalance between the attributes in the dataset while constructing the classification model.

DOI: 10.4018/IJOCI.2019040101

Copyright © 2019, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
An imbalanced attributes problem arises when training data has not observed any categorical variable during the training phase but it exist in the test data during the testing phase. In such cases zero probability is given by the model which leads to false prediction in the attribute label (He & Garcia, 2009; Fran & Asuncion, 2010). If the child dataset have this imbalance problem, then predicting child’s health condition is difficult. To overcome this attributes imbalance problem, different approaches are considered namely, rebalancing the dataset, considering the error during construction of classification model and merging the results of different classifiers trained on different data values (Batuwita & Palade, 2013). dataset imbalance problem can be handled by naive bayes classifier. The reason for choosing the naive bayes is that it considers only some (Narasimha & Susheela, 2011; Chawla et al., 2002) from the whole population as in probability theory. Here instead of taking the whole dataset, only a subset of training data is considered. Moreover naive bayes classifier reduces the error during model construction and biased towards the target attribute, when the attributes imbalance is more.

If the Naive Bayes classifier deals with the imbalanced attributes, then its performance can be improved through several techniques which can be applied either during data pre-processing or post-processing stage. In the former case, the attributes is rebalanced using the sampling (Nguyen et al., 2011) method and in the latter case, the classification model constructed using naive bayes will minimize the bias towards the attribute (Lopez et al., 2013; Sun et al., 2009). Hence a new method of estimating the threshold or bias is recommended in Naive Bayes classifier. The proportion of different attribute labels in the dataset is taken into account by the bias which adjusts the actual function trained by the naive bayes classifier. This is done to increase the performance of naive bayes classifier.

In this research article, section 2 gives a summary of different techniques used to enhance the performance of Naive Bayes classifier with attributes imbalance problem. Section 3 gives the pre-processing and post-processing methods used in the classifier. Section 4 describes the proposed framework of calculating the new bias. The experimental analysis is given in section 5 and finally the conclusion is given in section 6.

2. RELATED WORKS

In this section, various existing wroks have been discussed.

2.1. Gaussian Naive bayes

Gaussian naive Bayes (GNB) classification (Cohen et al., 2006; Joseph et al., 2016) is a supervised learning algorithm that uses Bayes’ theorem as a framework (Chang & Ramachandran, 2016) for classifying observations into one of a pre-defined set of classes based on information provided by predictor variables. GNB classifiers estimate the conditional probabilities that an observation belongs to a particular class given the values of the predictor variables under the assumption that the predictor variables are class-conditionally independent, and thus (naively) do not take into account the covariance among the predictor variables. Thus, the posterior probability that an observation Y has class index k given the values of predictor variables X_1, \ldots, X_p is modeled according to

\[
\hat{P}(Y = k | X_1, \ldots, X_p) = \frac{\pi(Y = k) \prod_{j=1}^{p} P(X_j | Y = k)}{\sum_{k=1}^{K} \pi(Y = k) \prod_{j=1}^{p} P(X_j | Y = k)}
\]

(1)

When where \( \pi(Y = k) \) is the prior probability that the class index is k.For each predictor \( X_1, \ldots, X_p \), the algorithm estimates a separate Gaussian distribution for each class, and observations are assigned to the class with the maximum posterior probability given the predictor values (Vapnik, 1999).
A Theoretical Framework for Estimating Swarm Success Probability Using Scouts

Particle Swarm Optimization Algorithm in Electromagnetics- Case Studies: Reconfigurable Radiators and Cancer Detection