Hybrid of Naive Bayes and Gaussian Naive Bayes for Classification: A Map Reduce Approach

Shikha Agarwal, Balmukumd Jha, Tisu Kumar, Manish Kumar, Prabhat Ranjan

Abstract: ***Naive Bayes classifier is well known machine learning algorithm which has shown virtues in many fields. In this work big data analysis platforms like Hadoop distributed computing and map reduce programming is used with Naive Bayes and Gaussian Naive Bayes for classification. Naive Bayes is manily popular for classification of discrete data sets while Gaussian is used to classify data that has continuous attributes. Experimental results show that Hybrid of Naive Bayes and Gaussian Naive Bayes MapReduce model shows the better performance in terms of classification accuracy on adult data set which has many continuous attributes.***

Index Terms: ***Naive Bayes, Gaussian Naive Bayes, Map Reduce, Classification.***

I. INTRODUCTION

The group of classifiers which are build upon Bayes Theorem is called Naive Bayes classifiers, also called simple Bayes and independence Bayes. All the classifiers of this category share the common principle of classification. The reason behind calling it Naive Bayes is its assumption which says that all the attributes of a datasets have no correlation means each attribute is independent of other [1]. Naive Bayes classifiers could be scaled easily. It requires linear parameters in the number of features in a classification problem. Training of Naive Bayes could be done using Maximum-likelihood. In these type of classifiers training is fast because iterative approximation is not done in Naïve Bayes classifiers. Naive Bayes is a simple technique for classification. It models a classifier which assigns a class labels to instances in the testing datasets. Each instance is represented by as a vector of feature values, where the class labels are drawn from training set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle. All Naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For some types of probability models, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for Naive Bayes models uses the method of maximum likelihood. Naive Bayes model could be used without accepting Bayesian probability or using any Bayesian methods. Despite their naive design and simple assumptions, Naive Bayes classifiers have worked quite well in many complex situations. An analysis of the Bayesian classification problem showed that there are sound theoretical reasons behind the apparently implausible efficacy of types of classifiers[5]. A comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests [6]. An advantage of Naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification. The fundamental property of Naive Bayes is that it works on discrete value. If attributes values are continuous it is recommended to use Gaussian Naive Bayes classifier. Here, in this work Naive Bayes and Gaussian Naive Bayes both are applied on a data set. The Data set has mix types of attributes. Some attributes are discrete and some are continuous. Therefore, to check which algorithm would perform well a experiment is conducted. The experimental results show that mix type of approach is better classifying the data which has mix types of attributes.

II. NAIVE BAYES CLASSIFIER

Naive Bayes is a group of supervised machine learning techniques which are used for classification. The crux of this classification method is Bayes Theorem. It predicts membership probabilities for each class in the dataset such as the probability that given data point belongs to a particular class. The class with the highest membership probability is considered as the most likely class of data point. Classifier identifies in which category a new observation belongs, on the basis of a training set of data. This classifier is a simple probabilistic classifier based on applying Bayes theorem as follows.

$$P(A \mid B) = \frac{P(B \mid A)P(B)}{P(B)} \quad (1)$$

Naive Bayes classification has an assumption that attribute probabilities $P(x_i|c_j)$ are independent given the class c_j, where x_i is i^{th} attribute of the data instance. Denominator part for entries of the datasets will not change. Hence the equation can be rewritten as:

$$p(y \mid x_1,...,x_n)\alpha P(y)\prod_{i=1}^{n} P(x_i \mid y) \quad (2)$$

Revised Manuscript Received on December 22, 2018.

Shikha Agarwal, Department of Computer Science, Central University of South Bihar, Gaya, India.
Balmukund Jha, Department of Computer Science, Central University of South Bihar, Gaya, India.
Tishu Kumar, Department of Computer Science, Central University of South Bihar, Gaya, India.
Manish Kumar, Department of Computer Science, Central University of South Bihar, Gaya, India.
Prabhat Ranjan, Department of Computer Science, Central University of South Bihar, Gaya, India.

Classify the data point to the class for which value of (2) is maximum. Hence class level of data point is given by (2).

III. GAUSSIAN NAÏVE BAYES

If in a data set most of the attributes are continues then Gaussian Naive Bayes is used. It is assumed in this algorithm that predictor values are samples from Gaussian distribution. Hence, Formula for conditional Probability becomes:-

$$P(x_i \mid y) = \frac{1}{\sqrt{2\Pi\sigma^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \quad (3)$$

Here μ_y and σ_y are mean and variance of predictor distribution.

IV. HADOOP

Hadoop is a framework of java that allows us to first store Big Data in a distributed environment, so that, we can process it parallel. There are basically two components in Hadoop, first one is HDFS(Hadoop Distributed File System) for storage, that allows to store data of various formats across a cluster and the second one is mapreduce for processing [2–4].

V. NAIVE BAYES CLASSIFIER AND MAPREDUCE CONCEPT

First of all record reader reads each line and coverts each line into key value pair for the input into mapper. In Map Reduce each key value pair again converted into new key value pair, where the key is a unique string which is a combination of the class, attribute and its attribute value, Ai ai ci . The value is 1. The input key of the mapper is the offset of the beginning of the line from the beginning of the file. Input value to the mapper is the string content of the line. Output key is the attribute name attribute value class. For example if a attribute name is age, its value in the sample is 2 and its class is <= 50k income. Then output key for this class is age 2 <= 50k. In reducer all the values with the same keys are merged and there values are added up. And a unique key value pair is generated. Value is the count of occurrence of such string combination [5–8].

VI. EXPERIMENT AND RESULTS

We used map reduce concept on data (Adult.data) in which Naive Bayes classifier is used to classify two types of class [9]. First class is person with less than 50k income and second is of person with more than or equal to 50k income. The adult dataset has 48842 instances (records) and 14 attributes. Among 14 attributes six are continuous variables and eight are categorical variables. Here in this experiment two form of Naive Bayes are used ; Naive Bayes, and Hybrid of Naive and Gaussian Naive Bayes. To apply Naive Bayes, continuous variables in the datasets are discretized using Binning method. The data is divided into training (70%) and testing (30%) set. In hybrid method posterior probability of the continuous variables are estimated using Gaussian Naive Bayes rule. Over the training dataset 6 folds cross validation is performed [10]. Table I shows the classification accuracy in six runs of the Naive Bayes and Hybrid of Naive and Gaussian Naive Bayes classifiers. The results of Naive Bayes are obtained from [7]. From Table I, we can see that the accuracy of each run is very close with very less standard deviation for both the methods. The Accuracy of Gaussian variant of Naive Bayes is comparable to the classification accuracy of Naive Bayes. Since Gaussian variant is gracefully handling the continuous attributes of the data. The two-by-two contingency table of the Hybrid model is shown in Table II. Contingency table of Naive Bayes is not shown because accuracy, precision and recall for Naive Bayes is obtained from [7]. Table III Shows the accuracy, precision and recall values of both model.

Table I: Classification Accuracy obtained in Naive Bayes classifier and proposed hybrid classifier (Naive Bayes-Gaussian Naïve Bayes map reduce classifiers) in multiple runs.

Run	Naive Bayes	Hybrid Naive and Gaussian Naive Bayes
1	0.762563	0.79425
2	0.763433	0.79429
3	0.763625	0.79417
4	0.762413	0.79460
5	0.762104	0.79376
6	0.76257	0.79419

Table II: Contingency Table of Naive Bayes-Gaussian Naïve Bayes map reduce classifiers.

	Actual annual Income >50 k	Actual annual income <=50 K
Predicted to "> 50 k"	tp=198978	fp=14320
Predicted to "<=50 k"	fn=45986	tn=33768

Table III: Accuracy, Precision and Recall of Naive Bayes and Hybrid Model

	Accuracy	Precision	Recall
Naive Bayes Classifier	0.763062	0.505306	0.86623
Naïve-Gaussian Bayes MapReduce Classifier	0.794213	0.812274	0.93286

VII. CONCLUSION

Naive Bayes group of classifiers have same basic principle of classification which is based on the Bayes Theorem. Naive Bayes works well when data attributes are discrete. Gaussian Naive Bayes shows the effective results when attributes are derived from the Gaussion distribution. Adult data has some discrete and some continuous attributes. Hence application of Naive Bayes for discrete

attributes and implementation of Gaussion for continuous attributes increases the classification performance of the model compared to the individual implementation of both the variants.

REFERENCES:

[1] S. B. Kim, K. S. Han, H. C. Rim and S. H. Myaeng, "Some effective techniques for Naive Bayes text classification," *IEEE T Knowl Data En, vol.* 18(11), 2006, pp. 1457-1466.
[2] "Hadoop Wiki - Partitioning your job into maps and reduces" [Online]. Available: http://wiki.apache.org/hadoop/HowManyMapsAndReduces.
[3] T. White, *Hadoop: The definitive guide*. Yahoo Press, 2010.
[4] "Hadoop Wiki - Sequence File" [Online]. Available: http://wiki.apache.org/hadoop/SequenceFile.
[5] S. Agarwal, and P. Ranjan, "MR-TP-QFPSO: Map Reduce Two Phase Quantum Fuzzy PSO for Feature Selection," *Int J Syst Assur Eng Manag*, vol. 9, 2017, pp. 888-900.
[6] S. Agarwal and P. Ranjan, "Map Reduce Fuzzy Ternary Particle Swarm Optimization for Feature Selection," *Journal of Statistics and Management Systems*, vol. 20(4), 2017, pp. 601-609.
[7] S. Zheng, "Naïve Bayes Classifier-A MapReduce Approach," 2014.
[8] S. Ghemawat, H. Gobioff and S. Leung, "The Google file system," *Proceedings of the nineteenth ACM symposium on Operating systems principles*, 2003, pp. 29-43.
[9] "UCI Machine Learning Repository" [Online]. Available: http://archive.ics.uci.edu/ml/.
[10] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics surveys, vol.* 4, 2010, pp. 40-79.

AUTHORS PROFILE

Shikha Agarwal is working as Assistant Professor in Department of Computer Science at Central University of South Bihar, Gaya, India. She has done her Ph.D. from Central University of South Bihar. He has done her M.Tech from Indian Institute of Information Technology–Allahabad. She has published many reputed research articles. She is the member of Indian Science Congress. She has awarded with Senior Research Fellowship from Council of Scientific and Industrial Research India in 2015. Her research areas are Machine Learning, Big Data, Bioinspired Computing and Block Chain.

Balmukund Jha is M.Sc. final Semester student in Department of Computer Science at Central University of South Bihar, Gaya, India. He completed his Bachelor in Computer Applications from Magadh University, India. His research areas are Big Data and Deep Learning.

Tisu Kumar is M.Sc. final Semester student in Department of Computer Science at Central University of South Bihar, Gaya, India. His research areas are Big Data and Data Science.

Manish Kumar is M.Sc. final Semester student in Department of Computer Science at Central University of South Bihar, Gaya, India. He has completed Bachelor in Computer Applications from Patna University, India. His research interests is in fields of Networking, AI and Machine learning.

Prabhat Ranjan is currently head incharge of Department of Computer Science, Central University of South Bihar, Gaya, India. He has done his M.Tech and Ph.D from Motilal Nehru National Institute of Technology-Allahabad. He has published twenty Scopus/SCI research articles. He is editor of "Recent Advances in Mathematics, Statistics and Computer Science" and "". His research areas are Software Engineering and Big Data.