Journal of Information Hiding and Multimedia Signal Processing ©2021 ISSN 2073-4212
Ubiquitous International Volume **12**, Number **3**, September **2021**

An Improving Using MapReduce Model in Predicting Learning Ability of Pupils Based on Bayes Classification Algorithm

Trung Nguyen Tu*

Department of Information Technology, Thuyloi University
175 Tay Son, Hanoi, Vietnam

Received December 2020; revised June 2021

Abstract. *Learning ability assessment is an important issue in assessing high school students. The assessment is based on a student's subject scores throughout the learning process. For a long time, machine learning algorithms in general and Bayes classification algorithm in particular have been applied to solve classification and prediction problems effectively. In addition, A huge amount of data arises from the construction of centralized student management applications for the whole provinces, the cities as well as the whole country. Currently, the MapReduce model is being used effectively in big data analysis. This paper uses the Bayes algorithm and MapReduce model in predicting students' academic ability to support the management and assessment of students in high school.*
Keywords: Learning ability, medium score, Bayes, prediction, MapReduce.

1. **Introduction.** Forecasting is a science and the art of predicting things that will happen in the future, based on scientific analysis of the collected data. The forecasting process should be based on collecting and processing the data in the past and at present to determine the movement trend of future phenomena thanks to a number of (quantitative) mathematical models. However, forecasting is also a subjective or intuitive prediction (qualitative) of the future; and, people try to eliminate the predictors subjectivity in order that the qualitative prediction is more accurate. There are many different forecasting methods. Currently, the use of machine learning methods applied for predictive problems has become very popular. In particular, the forecast by using Bayes classification is widely applied ... For example, forecasting the prices of all types of goods, forecasting population growth rate ... when knowing the past information and given conditions. ... The Bayes classification is also used in a way in text subject classification [7]. In [13], the authors used Deep learning to classify text topics. One of the most common applications of the Bayes classification is spam classification. In [1], Awad presented the evaluation and the comparison of among some machine learning methods as Bayesian classification, k-NN, ANNs, SVMs... for spam filtering. Jialin et al discussed and evaluated the method of filtering SMS spam using SVM and MTM [3]. In [5], Phan Huu Tiep and his colleagues presented the process of filtering Vietnamese spam based on Bayes algorithm and the processing of Vietnamese sentence separation. Tianda et al presented the comparison between the spam classifier using only the Bayes technique and the spam classifier using the technical spam classifier and association rules [6]. In [4], the authors evaluated several approaches of calculating the token's SPAM probability in spam classification.

Bayes classification is used in identifying correctness data scheme in wireless sensor network and resource scheduling [15][17]. Currently, with the development of information technology, the Industrial Revolution 4.0 has led to the explosion of data (Big Data). Big data and its analysis play an important role in the IT world with the applications of Cloud Technology, Data Mining, Hadoop and MapReduce [10]. Traditional technologies only apply to structured data while big data includes both structured, semi-structured and unstructured data. How to effectively handle big data has become a big challenge in the new time and new processing methods are needed. MapReduce is a highly efficient distributed data processing model that has been widely used in big data processing [2]. MapReduce is used in Hierarchical PSO Clustering and Social Network Privacy Protection [14][16]. Conduct and learning ability are two very important factors of each student when studying at school. In particular, the result of learning ability will be used to evaluate and consider students for rewarding and moving up to the next grade [8]. Based on the medium scores of subjects of the semester and the whole year, the ranking of learning ability is divided into 5 categories: Good, Middling, Medium, Weak, Poor.Therefore, the assessment of learning ability is performed strictly. At present, due to the need of connecting, sharing and centralized management, the data of schools and educational levels are stored on the servers of a province or a country. This will give rise to a huge amount of data. Therefore, the methods of exploitation and calculation on traditional data will be difficult and ineffective. If new models of computation can be applied to this data, it will be extremely effective. In this paper, we propose a solution that applies the Bayes algorithm and MapReduce model to predict learning ability of students based on their subject scores.

2. **Related Work.**

2.1. **Overview of MapReduce model.** MapReduce is a model of parallel and distributed computing model that is proposed by google (Figure 2). It includes two basic functions: Map and Reduce which are defined by the user [4]. Through the MapReduce library, the program fragments the input data file. Machines include: master and worker. The master machine coordinates the operation of the MapReduce implementation process on the worker machines, the worker machines perform the Map and Reduce tasks with the data it receives. Data is structured in the form of key and value.

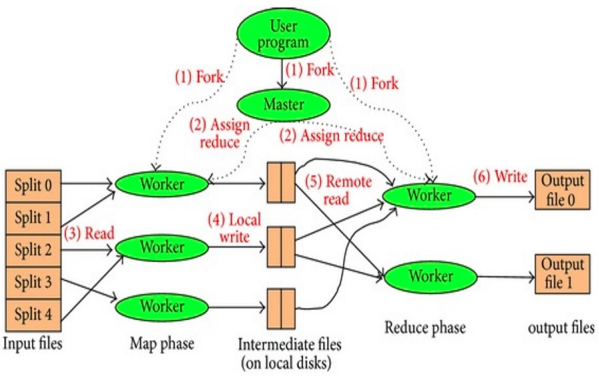

FIGURE 1. Flowchart of MapReduce model [2].

The formal representation of MapReduce model: According to [6] [12], we have the formal representation of the MapReduce model as follows:

- map: (K1 k1, V1 v1) → list(K2 k2, V2 v2)

- reduce: (K2 k2, list(V2 v2)) → list(K3 k3, V3 v3)

Where:

- K1, V1 are the input key and value types of the map function; k1, v1 are the corresponding objects with the types K1, V1.
- K2, V2 are the output key and value types of map function and still are the input key and value types of reduce function; k2, v2 are the the corresponding objects with the types K2, V2.
- K3, V3 are the output key and value types of the reduce function; k3, v3 are the the corresponding objects with the types K3, V3.

In other words, we can see:

- If k1, v1, k2, v2 are identified, we have the input and output of map function. Commonly, with text data, k1 is offset value of a data row, v1 is the content of a data row.
- If k2, v2, k3, v3 are identified, we have the input, and output of reduce function.

The formal Representation may be rewritten only with k1, v1, k2, v2, k3, v3 as follows:

$$map : (k1, v1) \rightarrow list(k2, v2) \tag{1}$$

$$reduce : (k2, list(v2)) \rightarrow list(k3, v3) \tag{2}$$

Figure 3 illustrates the diagram of the MapReduce job execution and data conversion from types (K1, V1) to types (K2, V2) and types (K2, V2) to types (K2, V3).

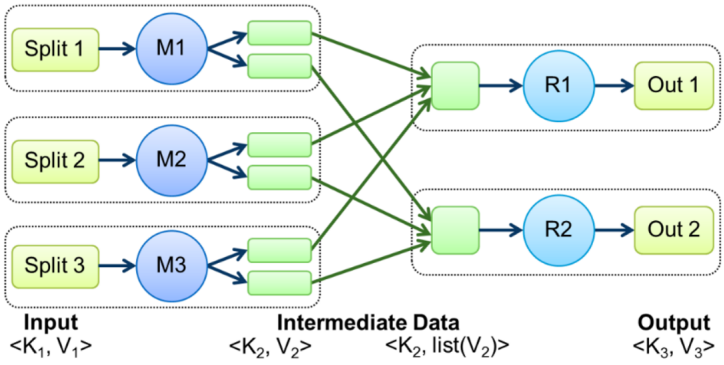

FIGURE 2. Flowchart of MapReduce model [12].

2.2. **The algorithm Nave Bayes.** According to [9], the problem can be described as follows: Data is needed:

- D: The training data set vectorized as $\vec{x}(x_1, x_2, ..., x_n)$.
- C_i: The set of documents of D that belong to C_i with i=$\{1, 2, 3...\}$.
- The components $x_1, x_2, ..., x_n$ independent probability of a double together.

The algorithm Nave Bayes as follows:

- Step 1: Training Nave Bayes is based on the training data set, as illustrated in figure 3, that includes calculate probabilities $P(C_i)$ and $P(x_k|C_i)$
- Step 2: Classify X_{new}
  - Calculate $F(X_{new}, C_i) = P(C_i) \prod_{k=1}^{n} P(x_k|C_i)$
  - X_{new} belongs to C_q so that $F(X_{new}, C_q) = max(F(X_{new}, C_i))$

$P(x_i|C_i)$ is calculated as follows: $P(x_i|C_i) = \frac{C_{i,D}\{c_k\}}{|C_{i,D}|}$ Where:

- $|C_{i,D}|$: sample number of training data set D that belong to C_i

- $C_{i,D}\,\{c_k\}$: sample number of $C_{i,D}$ whose value is x_k

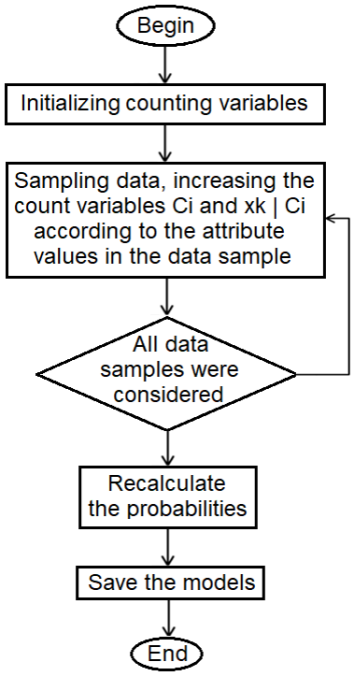

FIGURE 3. Training flow chart of the Bayes algorithm.

2.3. **Overview of MapReduce model.** In [13], the authors proposed the algorithm of predicting students' learning ability based on algorithms Nave Bayes.

2.3.1. *The model of predicting learning ability based on Bayes Training phase.* Input: The list of records that includes the core information of students' subjects: Math (Mat), Physics (Phy), Chemistry (Che), Biology (Bio), Informatics (Inf), Literature (Lit), History (His), Geography (Geo), English (Eng), Civic Education (Civ), Technology (Tec), Defense Education (Def) and information about learning ability as shown in figure 4. Output: Predicting model of learning ability.

Mat	Phy	Che	Bio	Inf	Lit	His	Geo	Eng	Civ	Tec	Def	LeAb
3.9	5.4	5.1	7.7	6.8	6.1	7.6	6.9	5.5	6.7	7.2	7.4	Tb
5.6	5.4	4.5	5.4	6.6	5.7	8.4	5.1	4.9	5.5	6.3	7.4	Tb
5.9	6.6	4.5	7.2	7.5	5.6	6.8	5.7	5.4	6.6	6.4	7.7	Tb
4.2	6.6	6.6	6	6.2	5.8	8.3	7.1	4.6	6.5	6.2	6.7	Tb
8.7	7.1	7.5	7.2	7.8	6.7	7.9	7.5	5.5	6.1	7	6.9	K
4.7	6.8	5.7	5.3	6.9	5.6	8.1	6.4	4.5	5.1	7.8	7.3	Tb
8.8	7.6	7.4	7	7.7	7.1	8.6	7.1	7.5	6.6	8.1	8.3	K
4.2	5.3	4.4	5.3	7.3	4.5	4.4	7	3.7	5.2	6.6	5.9	Y
7.7	7.1	6.5	6.6	7	5.5	7.2	6.3	5.8	6.4	5.6	7	K
8.8	8.3	7.4	7.6	6.5	6.3	8	7.4	8	7.2	6.6	8.3	K
8.7	6.7	6.1	7.4	7.2	6.6	7.9	6.6	5.5	6	7.6	8.1	K
5.8	6.2	6.5	7.2	6.7	6	8.2	8.3	5.4	7.3	7.4	8.3	Tb
5	6	5.3	6.6	6.3	5.8	8.5	6.8	6.2	6.4	6.1	7.3	Tb
5.3	5.6	6.1	7	5.3	4.2	7.6	7.1	2.8	4.1	5.1	6.4	Y
4.5	5.9	5.3	5.4	5.7	4.8	6.7	7.4	4	5.3	6.1	6.9	Y

FIGURE 4. Example of input data in the training phase.

To be able to use the method Bayes, the labels C_i and the data samples are be determined as follows:

- The labels C_i are: Good, Little good, Medium, Weak, Poor.

- The data samples $\vec{x}$ is a vector whose components $x_1, x_2, , x_n$ are a the subject cores of student.

There is a problem that arises from formula (6) as follows: if each subject's score is directly used as the value of $x_1, x_2, , x_n$ then in the case of a core value xk included in the test data but not in the training data, value of $P(x_k|C_i)$ is 0 $\forall i$. Therefore, $F(x_{new}, C_i) = 0\ \forall i$. This means that we will not select a valid learning ability label.

To overcome the above problem, we use one of the two data smoothing methods as follows:

- Method 1: Do not use subject score values directly as a component of a vector $\vec{x}$. We propose 3 data conversion techniques as follows:
  - Technique 1: Convert score into one of the levels: G_S (Good score), Mi_S (Middling score), Me_S (Medium score), W_S (Weak score), P_S (Poor score):
    - If $score >= 8$ then $x_k = G_S$
    - If $score >= 6.5$ and $score < 8$ then $x_k = Mi_S$
    - If $score >= 5.0$ and $score < 6.5$ then $x_k = Me_S$
    - If $score >= 3.5$ and $score < 5.0$ then $x_k = W_S$
    - If $score < 3.5$ then $x_k = P_S$
  - Technique 2: Convert the score into one of the levels: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10:
    - If $score >= i$ and $score < i + 1$ then $x_k k = i$ (with i = 1..10)
  - Technique 3: Convert the score into one of the levels: 0A, 0B, 1A, 1B, 2A, 2B, 3A, 3B, 4A, 4B, 5A, 5B, 6A, 6B, 7A, 7B, 8A, 8B, 9A, 9B, 10.
    - If $score >= i$ and $score < i + 0.5$ then $x_k = iA$ (with i = 1..10)
    - If $score >= i + 0.5$ and $score < i + 1$ then $x_k = iB$ (with i = 1..10)
- Method 2: Using smoothing formula Laplace as follows:

$$P(x_k|C_i) = \frac{C_{i,D}\{x_k\} + 1}{|C_{i,D}| + r} \tag{3}$$

Wherein, r is the discrete value of the attributes. The training algorithm for the prediction model of the learning ability is illustrated by the flowchart in figure 5. Accordingly, the algorithm is start with the initialization of variables for counting the labels like Learning ability C_i and Score-Learning ability $x_k|C_i$. For each data sample, the score values will be smoothed according to one of the above methods and techniques. Depending on the appearance of the labels like Learning ability C_i and Score-Learning ability $x_k|C_i$ in the which is being considered, the counting variables of the corresponding labels are increased. After all data samples are considered, the probabilities $P(C_i)$ and $P(x_k|C_i)$ will be calculated and saved to the model file to finish the training algorithm.

Classifying phase:

Input: The data sample is the scoring information for any student.

Output: Forecasting information about learning ability: Good, Middling, Medium, Weak, Poor.

The learning ability classification algorithm is illustrated by the diagram in figure 6. Accordingly, for each data sample of student score, the score values will be smoothed. Smoothed data along with the model data that was generated after training, is used to calculate and produce appropriate result of learning ability based on the Bayes classification as described in section 2.2.

2.3.2. *The rules of learning ability decision based on Bayes.* Based on the Bayes forecasting model of learning ability as presented in section 3.1, the rule of deciding learning ability based on Bayes is formulated as follows:

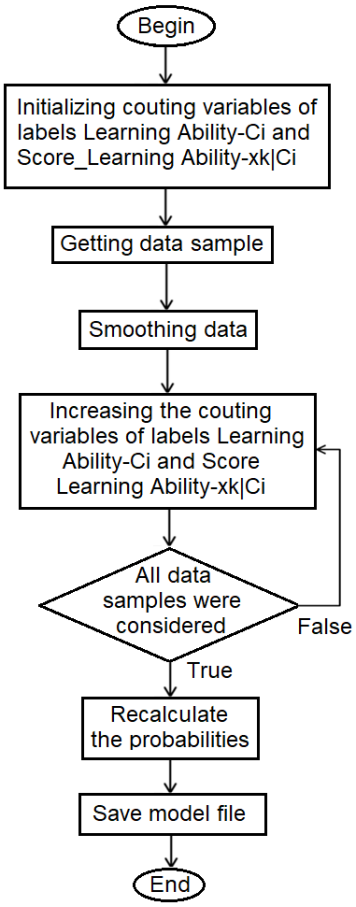

FIGURE 5. Flowchart of training algorithm for the model of predicting learning ability based on Bayes.

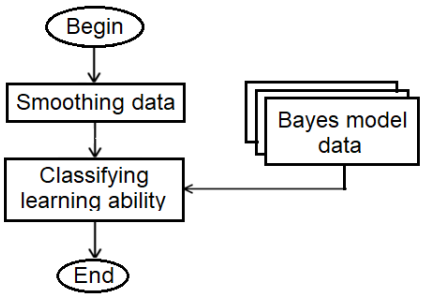

FIGURE 6. The flowchart of the algorithm of classifying learning ability based on Bayes.

- Rule 1: deciding Good kind:
  - $F(X_{new}, Good) = max(F(X_{new}, C_i))$
  - Average score of 1 in 2 subjects as either Mathematics or Literature is from 8.0 or more.
  - No subject whose average score is under 6.5.
  - Scores of the subjects that are assessed by the comments are D.
- Rule 2: deciding Middling kind
  - $F(X_{new}, Middling) = max(F(X_{new}, C_i))$ or not Good kind.

  - Average score of 1 in 2 subjects as either Mathematics or Literature is from 6.4 or more.
  - No subject whose average score is under 5.0.
  - Scores of the subjects that are assessed by the comments are D.
- Rule 3: deciding Medium kind
  - $F\left(X_{new}, Medium\right) = max\left(F\left(X_{new}, C_i\right)\right)$ or not Good, Middling kind.
  - Average score of 1 in 2 subjects as either Mathematics or Literature is from 5.0 or more.
  - No subject whose average score is under 3.5.
  - Scores of the subjects that are assessed by the comments are D.
- Rule 4: deciding Weak kind:
  - $F\left(X_{new}, Weak\right) = max\left(F\left(X_{new}, C_i\right)\right)$ or not Good, Middling, Middling kind.
  - No subject whose average score is under 2.0.
- Rule 5: deciding Poor kind:
  - $F\left(X_{new}, Poor\right) = max\left(F\left(X_{new}, C_i\right)\right)$ or not Good, Middling, Middling, Weak kind.

The algorithm of predicting learning ability is based on the rules of learning ability decision that we call DBHL_Bayes as illustrated in the flowchart in figure 7.

Input: The data sample is the scoring information for any student.

Output: Forecasting information about learning ability: Good, Middling, Medium, Weak, Poor.

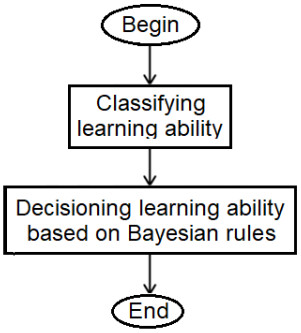

FIGURE 7. The algorithm for forecasting students' learning ability based on the rules of Bayes decision.

3. Improving the Bayesian model training algorithm for learning ability prediction using MapReduce model.

3.1. Analysis of Bayesian model training algorithm for forecasting learning ability.

We have some comments as follows:

- In the trainning process, if the amount of data is too big, there will be problems such as lack of memory, long execution time.
- The majority of Bayes training time is devoted to counting the number of occurrences of the labels Learning ability C_i or the Scores—Learning $x_k|C_i$.
- About caculating the probability, counting the number of occurrences of each label are independent. Therefore, it is possible to divide the data into multi small parts and execute these parts in parallel.

3.2. **Improving the Bayesian model training algorithm for learning ability prediction using MapReduce model.** Idea as follows:

- Training data is the list of student's score-learning ability information which is divided into multi small parts by the system that intergrated MapReduce library.
- Map function: Accumulating 1 for each occurrence time of each label Learning ability Ci and label Score— Learning ability $x_k|C_i$ (with attached label)
- System groups numbers 1 that have label Learning ability C_i or label Score—Learning ability $x_k|C_i$ automatically.
- Reduce function: Calculating the sum of the numbers 1 by each label Learning ability C_i or label Score—Learning ability $x_k|C_i$
- Caculating the probabilities of the labels Learning ability $P(C_i)$ and label Score—Learning ability $P(x_k|C_i)$
- Saving the model file that contains probability information for labels Learning ability $P(C_i)$ and label Score—Learning ability $P(x_k|C_i)$

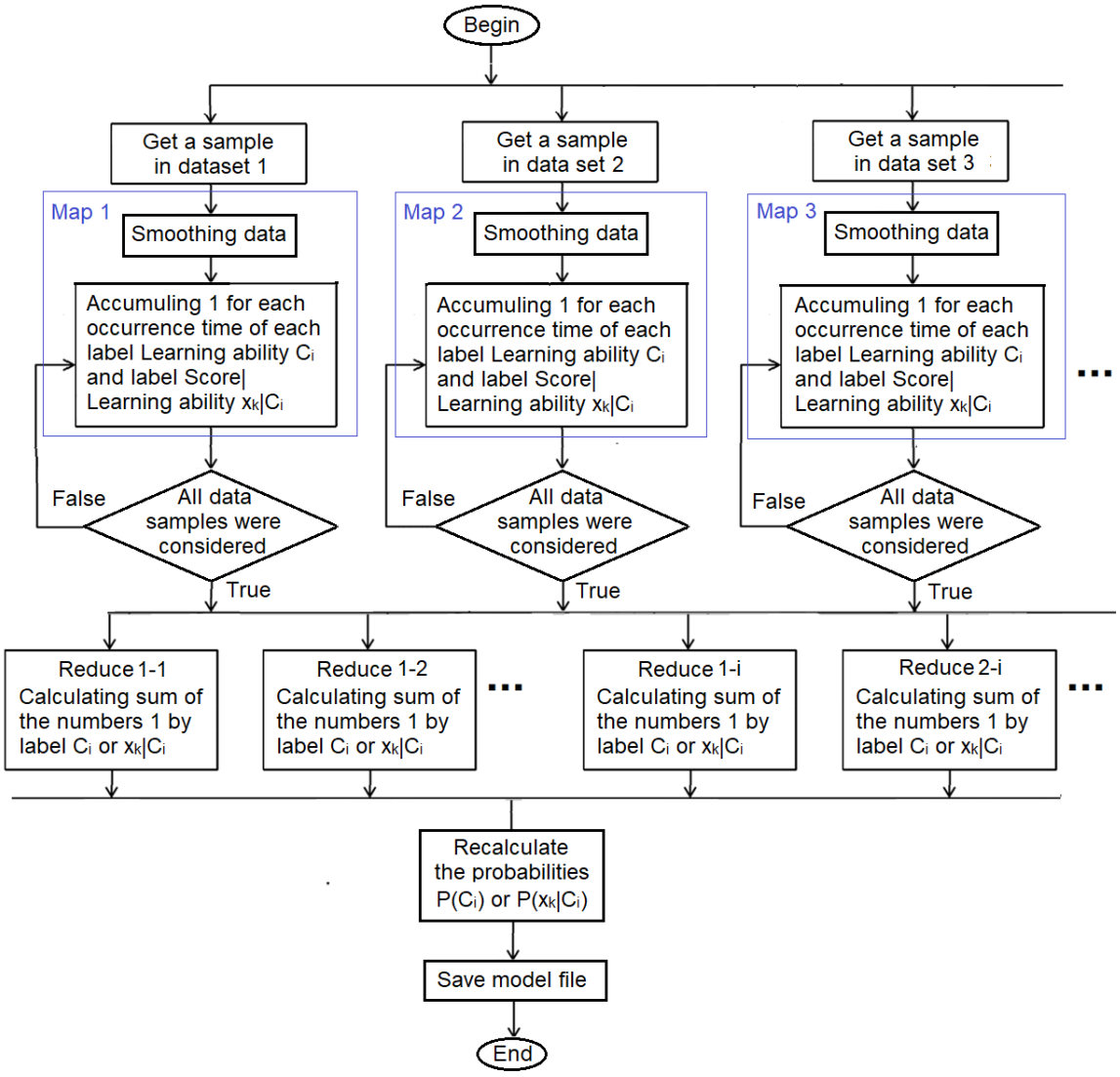

FIGURE 8. The flowchart of the improved Bayes-based learning forecast model algorithm using MapReduce model

The learning ability prediction algorithm is based on the rule of Bayes decision and the MapReduce model called DBHL_MapReduce_Bayes has the same steps as the DBHL_Bayes algorithm and only replaces the model file as the result of the training algorithm as described in figure 8.

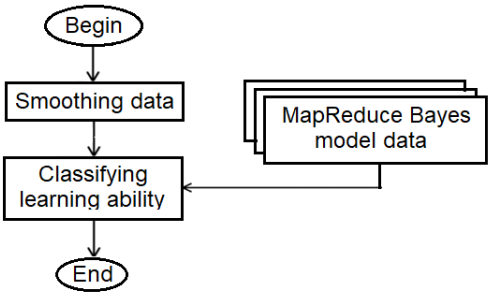

FIGURE 9. The flowchart of the algorithm of classifying learning ability based on MapReduce_Bayes

3.3. **Formal representation for the Map and Reduce functions in the model of predicting learning ability using MapReduce.** Input: Each data row rowi is a set of the list of subject scores and Learning ability: $(list\,(x_k)), C_i)$.

Output: The pairs of the label of Learning ability C_i or Score—Learning ability $x_k|C_i$ and the total number of occurrences of the corresponding label: $list\,(Or\,(C_i, x_k|C_i)\,, count)$. Therefore, the pairs (k1, v1) and (k3, v3) are determined as follows:

- k1 is offset, v1 is the content of data row: $(list\,(x_k)), C_i)$
- k3 is the label of Learning ability C_i or Score—Learning ability $x_k|C_i$, v3 is the total number of occurrences of the corresponding label which is stored by k3

The Map funtion accumulates 1 for each occurrence time of each label Learning ability C_i and the label Score— Learning ability $x_k|C_i$ (with attached label) so the pair (k2, v2) is determined as follows: k2 is label of Learning ability C_i or Score—Learning abiliy $x_k|C_i$, v2 is 1.

In this case, the formal representation of Map and Reduce procedures is as follows:

$$map2D : (offset, row_i) \rightarrow list\,(Or\,(C_i, x_k|C_i)\,, 1) \tag{4}$$

$$reduce2D : (Or(C_i, x_k|C_i), list(1)) \rightarrow list(Or(C_i, x_k|C_i), sum(list(1))) \tag{5}$$

3.4. **The algorithm of the procedure map_Bayes.** This part presents the algorithm for the procedure map_DBHL.

This algorithm is responsible for separating the input data into the labels of the subject scores and learning ability. Next, the procedure accumulates 1 for each occurrence time of each label Learning ability C_i and label Score— Learning ability $x_k|C_i$ (with attached label).

This algorithm is described as follows:

- Input: key k1 is offset, value v1 is the content of data row_i: $(list(x_k), C_i)$
- Output: The list lstk2v2 includes the pairs (k2,v2): k2 is label Learning ability C_i or Score—Learning ability $x_k|C_i$, v2 is 1
- B1: Separating the labels Score x_k and Learning ability C_i from v1
- B2: Initializing the list lstk2v2 to store the pairs as (k2,v2)
- B3: Adding the pair (C_i,1) into the list lstk2v2
- B4: Browse each score label x_k

– B4.1: Adding the pair ($x_k|C_i$,1) into the list l lstk2v2

3.5. **The algorithm of the procedure reduce_Bayes.** This part presents the algorithm for the procedure reduce_DBHL.

This algorithm is responsible for summing the values of 1 for the input label as Learning ability C_i or Score—Learning abiliy $x_k|C_i$.

This algorithm is described as follows:

- Input: key is $Or(C_i, x_k|C_i)$, value is the list of numbers 1 with the label stored in the key, it means $list(1)$
- Output: The pair (k3,v3): k3 is $Or(C_i, x_k|C_i)$, v3 is the sum of the elements of the list(1)
- B1: Initializing sum = 0
- B2: Foreach list(1)
  - B2.1: Increasing sum = sum + 1
- B3: Assigning $k3 = Or(C_i, x_k|C_i)$
- B4: Assigning v3 = sum
- B5: Return the pair (k3,v3)

3.6. **Proving that the accuracy of the DBHL_Bayes and DBHL_MapReduce_Bayes algorithms for leaning ability prediction are the same.** From the above sections, we have some comments as follows:

- Comment 1: The differences between the two model training algorithms for predicting learning ability based on Bayes is presented in Figure 5 and Figure 8:
- Comment 2: According to comment 1, the number of occurrences of each label Learning ability C_i or Score—Learning abiliy $x_k|C_i$ is done in 2 different ways in 2 training algorithms so the results are the same.
- Comment 3: From comment 2, the value for the probabilities of labels Learning ability C_i or Score—Learning abiliy $x_k|C_i$ are same for both training algorithms.
- Comment 4: From comment 3, when classifying, the value of the function $F(X_{new}, C_i)$ is the same for both classification algorithms. Thus, the results of the rules of learning ability decision are the same for both algorithms DBHL_Bayes and DBHL_MapReduce_Bayes. That is thing which must be proved.

4. **Experiments.** The test data set which is the subject scores and the learning ability information of students in some high schools (permission is not shared for security reasons). This data is collected on the internet. Training data is stored in the excel file which includes 7962 records. Test data set: is stored in the excel file which includes 1162 records. Table 3 shows the training time for the model and the forecasting accuracy of each specific method and technique with the algorithm DBHL_Bayes.

Table 1. The algorithm DBHL_Bayes

Method(M)/Technique (T)	M1-T1	M1-T2	M1-T3	M2
Training time	188 ms	332 ms	322 ms	285 ms
Forecasting accuracy	99.14%	100%	100%	99.48%

Table 4 shows the model training time and forecasting accuracy for each specific method and technique with the algorithm DBHL_MapReduce_Bayes.

From the results in Tables 3 and 4, we see:

- About training time:

TABLE 2. The algorithm DBHL_MapReduce_Bayes

Method(M)/Technique (T)	M1-T1	M1-T2	M1-T3	M2
Training time	100 ms	119 ms	111 ms	112 ms
Forecasting accuracy	99.14%	100%	100%	99.48%

  - The training time of Method 1-Technique 1 is minimal and Method 1-Technique 2 is the longest. With all the methods and techniques used, the training speeds are very fast. This will be very convenient if retraining is required to improve accuracy in case of the training data size changed.
  - Training time of the algorithm DBHL_MapReduce_Bayes is much smaller than the algorithm DBHL_Bayes. This demonstrates the advantage of the parallel and distributed model MapReduce.
- About accuracy:
  - The accuracy of Method 1-Technique 1 is minimal with 99.14% and Method 1-Technique 2 and 3 is greatest with 100% accuracy. This shows that the use of Bayesian machine learning method is very suitable for predicting learning ability
  - The forecasting accuracy using the algorithms DBHL_Bayes and DBHL_MapReduce_Bayes is the same. This is a justification for the proof in section 4.6.

5. **Conclusions.** In this paper, the authors have proposed the method of learning ability prediction using Bayes classification algorithm and an improvement of this algorithm using MapReduce model to speed up the execution of the algorithm. In our improvement, we also propose some techniques to refine raw data (initial score value) before training or classifying Bayes. The testing results show that the training speed is very fast and the accuracy is very high, exceeding 99 percent with all 4 methods of techniques. Especially, the training speed of the algorithm DBHL_MapReduce_Bayes is much faster than the algorithm DBHL_Bayes without sacrificing the accuracy compared to the algorithm DBHL_Bayes. In the next study, the authors plan to continue applying the MapReduce model to other algorithms to enhance the effectiveness of algorithms in the context of increasingly big and complex data.

Acknowledgment. This work is partially supported by Department of Information Technology, Thuyloi university and Institute of Information Technology, Vietnamese Academy of Science and Technology. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Awad W.A. and ELseuofi S.M., Machine learning methods for spam e-mail classification, *International Journal of Computer Science & Information Technology (IJCSIT)*, vol. 3, no. 1, pp. 173-184, Feb 2011,.

[2] Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, *USENIX Association OSDI '04: 6th Symposium on Operating Systems Design and Implementation*, 2004.

[3] Jialin Ma, Yongjun Zhang, Jinling Liu, Intelligent SMS spam filtering using topic model, *ieee international conference on intelligent networking and collaborative systems (incos)*, pp. 380-383, 2016.

[4] Nguyen Tu Trung, Nguyen Ngoc Hung, Pham Thanh Giang, Assess some methods of calculating spam probability of tokens applied in spam email classification, *Journal of Science and Technology on Information and Communications*, no. 3, pp. 27-32, 2018.

[5] Phan Huu Tiep et al., Vietnamese spam filtering method based on compound words and user tracking, National Conference *Vietnamese Information and Communication Technology*, Can Tho, pp. 463-473, 2011.

[6] Tianda Yang, Kai Qian, Dan Chia-Tien Lo, Spam filtering using Association Rules and Nave Bayes Classifier, *IEEE International Conference on Progress in Informatics and Computing (PIC)*, pp. 638-642, 2015.
[7] http://viet.jnlp.org/kien-thuc-co-ban-ve-xu-ly-ngon-ngu-tu-nhien/machine-learning-trong-nlp/phan-loai-van-ban-bang-dinh-ly-bayes
[8] https://thuvienphapluat.vn/van-ban/giao-duc/Thong-tu-58-2011-TT-BGDDT-Quy-che-danh-gia-xep-loai-hoc-sinh-trung-hoc-co-so-133268.aspx
[9] https://towardsdatascience.com/introduction-to-naive-bayes-classification-4cffabb1ae54
[10] Nandhini.P, A Research on Big Data Analytics Security and Privacy in Cloud, Data Mining, Hadoop and Mapreduce, *Int. Journal of Engineering Research and Application*, 2018.
[11] Herodotos Herodotou, Business Intelligence and Analytics: Big Systems for Big Data, Cyprus University of Technology, 2016.
[12] Tom White, Hadoop: The Definitive Guide : The Definitive Guide, 2009.
[13] Dao Duc Anh, Nguyen Tu Trung, Vu Van Thoa, Using bayesian classification in predicting learning ability of high school students, *Journal of Science and Technology on Information and Communications*, no. 1, pp. 46-49, 2020.
[14] Ei Nyein Chan Wai, Pei-wei Tsai, Jeng-Shyang Pan, Hierarchical PSO Clustering on MapReduce for Scalable Privacy Preservation in Big Data, *ICGEC 2016*, pp. 36-44, 2016.
[15] Shu-Chuan Chu, Thi-Kien Dao, Jeng-Shyang Pan, Trong-The Nguyen, Identifying correctness data scheme for aggregating data in cluster heads of wireless sensor network based on naive Bayes classification, *EURASIP Journal on Wireless Communications and Networking*, 2020:52 https://doi.org/10.1186/s13638-020-01671-y.
[16] Shou-Lin Yin and Jie Liu, A K-means Approach for Map-Reduce Model and Social Network Privacy Protection, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 7, no. 6, pp. 1215-1221, November 2016.
[17] Juan Wang, Xiuyan Sun, Wenmin Song and Linlin Tang, Resource Scheduling Method Based on Bayes for Cloud Computing, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 9, no. 6, pp. 1444-1451, November 2018.