# NEW RECOMMENDER SYSTEM USING NAIVE BAYES FOR E-LEARNING

Mehmet ÖZCAN
Anadolu University, Turkey

Furkan GÖZ
Kocaeli University, Turkey

Tansu TEMEL
Anadolu University, Turkey

**ABSTRACT**: Coming into prominence at the present time, e-learning is a great opportunity for learners. It provides tremendous assets most valuable of which is distance free learning. Besides, there is a great deal of e-learning resources on the web that causes information overload. Accordingly, it turns into a requisite that you ask for recommendation so as to find the resource you surely need. There are readily available recommendation services arranged for that purpose. Such systems have various rating systems; furthermore, users tend to rate the materials in different manners. Our goal with this paper is to generate confidential referrals thanks to Naive Bayesian algorithm for e-learning materials rated multifariously by learners. We also researched the effects of several data preprocessing techniques on achieving this goal.

**Key words:** Naïve Bayesian Classifier, Data Preprocessing, E-learning, Recommendation Systems

## INTRODUCTION

Rapid advances in technology have led to be used the new methods in education as in other areas. One improvement is that people learn something using mobile phone or web based systems. It was named as e-learning which supports traditional education yet began to replace it because in this way they feel comfortable and so prefer e-learning platforms because of independent time or place. In other words, people, both learn something whatever they want without going to school in a specific time easily and take a certificate or diploma in this way so demand has increased to e-learning day by day.

On the other hand, people, can present any educational documents or videos effortlessly. If you have a camera and internet connection, you can publish any educational materials without control mechanism. Whereas some of them are useful for learning, others cannot be beneficial for a learner. The problem is that given the increasing number of e-learning platforms and materials, learners are frequently overwhelmed with the large amount of learning resources available online (Souali, El Afia, Faizi, & Chiheb, 2011). Therefore, having a right material in right time is also difficult. People have tried to find most suitable resources themselves by asking someone who used this to solve the problem but not enough. It is almost impossible to select appropriate materials because of reaching limited users. In order to overcome this issue, there is need for a system that recommends the correct materials extracted automatically from preferences of similar users.

Computer-based recommender systems are the most appropriate methods in order to recommend materials for people. The main purpose of a recommender system is to generate meaningful recommendations to users which expect suggestion for items or products that might interest those (Melville & Sindhwani, 2011). Recommender systems have a wide usage area in our daily life such as movies, music, books, food and healthcare.

Our goal in this paper is to implement Recommender System with Naïve Bayes algorithm for e-learning materials rating from learners with different ways. Several data preprocessing operations are applied before applying Naïve Bayes Classifier. The vestigial of this paper is regulated as follows. Section 2 presents related works. Section 3 exhibits proposed architecture. Section 4 includes experimental results. Section 5 gives a short conclusion and future works.

## RELATED WORKS

In this section, we present some of the research literature related with e-learning recommender systems. Bayesian Network is utilized in order to detect learner's learning style and discover their preferences (Carmona, Castillo & Millán, 2007; García, Amandi, Schiaffino & Campo, 2007). Ueno and Toshio (2007) created learner model via

Bayesian Network. Using the learner model, learner's final status (Failed, Abandon, Successful, Excellent) is predicted. Next, active learner's learning processes are compared with excellent learners' learning processes, and appropriate messages to the learner are generated. Colace and De Santo (2010) studied on the role of ontologies in the context of e-learning. A novel algorithm for ontology building with Bayesian Networks is presented in their work. Analyzing students' learning performances, their proposed method can analyze the courses' ontology and propose corrective actions. Thus, teachers better understand the requirements of their students and can redesign their courses appropriately. Moreover, an ontological basis is provided to determine learning paths to personalize learning. Chang, Kao, Chu and Chiu (2009) proposed a learning style classification mechanism to classify and identify students' learning styles. The proposed method improves k-nearest neighbor classification and combines it with genetic algorithms. The proposed method is implemented on an open-learning management system. García Amandi and Schiaffino (2008) detected a student's learning style automatically from the student's actions in an e-learning system using Bayesian Networks. E-teacher uses the information contained in the student profile to proactively assist the student by suggesting him/her personalized courses of action that will help him/her during the learning process. Özpolat and Akar (2009) addressed the problem of extracting the learner model based on Felder–Silverman. Using Naïve Bayesian Tree in conjunction with Binary Relevance classifier, the learners are classified according to their interests. Learners' learning styles are defined using these classification results.

## PROPOSED ARCHITECTURE

In a conventional e-learning system, instructors procure some teaching documents or materials to the e-learning system for learners. Learners using the system can utilize these materials comfortably via the web in their education. These learners also leave ratings for the materials they use according to their interest. These collected ratings are stored by the e-learning system. In the case that a new learner enters into the system he can search and use any material he wants. He can also ask for a material recommendation from the system but, he must have rated a certain number of materials before. By means of the user ratings, preference of that user is extracted first. The system tries to recommend the most appropriate material by combining that preference with the previous learners' rating feedbacks.

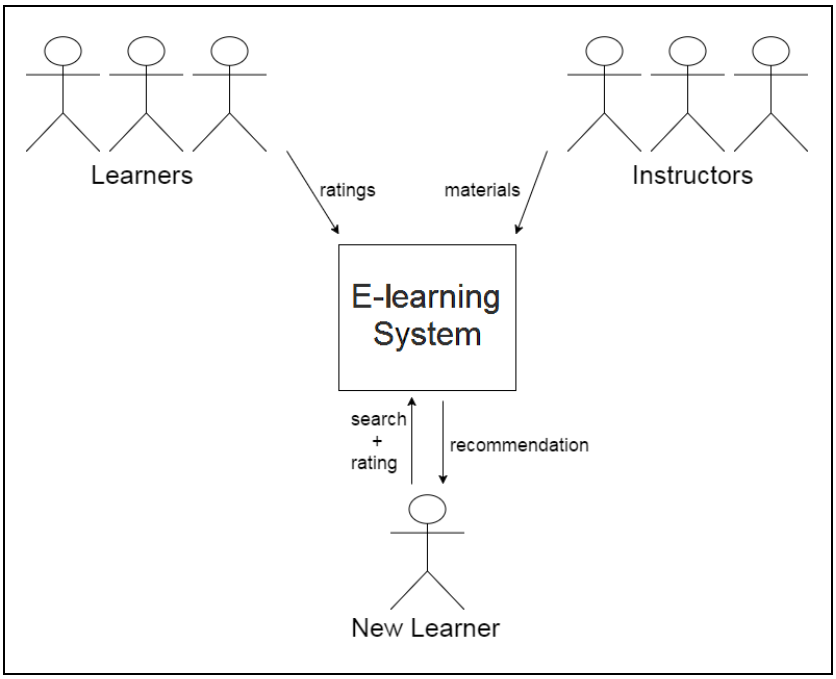


**Figure 1.** E-learning system scenario

Our proposed architecture basically aims to produce effective predictions with Naïve Bayesian Classifier for e-learning systems. Figure 1 shows the framework model for e-learning recommendation. It is necessary to have a quality dataset in order to get efficient predictions. Thus, we take the advantage of some data preprocessing operation (missing values problem). We also slog on maintaining their studies at generating predictions from binary data because of not having a binary dataset with great amount of data. After preprocessing, conversion step is dataset into binary dataset which is includes all values 0 and 1. This data is trained with leave-one out method and eventually the system recommends to user with Naïve Bayesian Classifier.

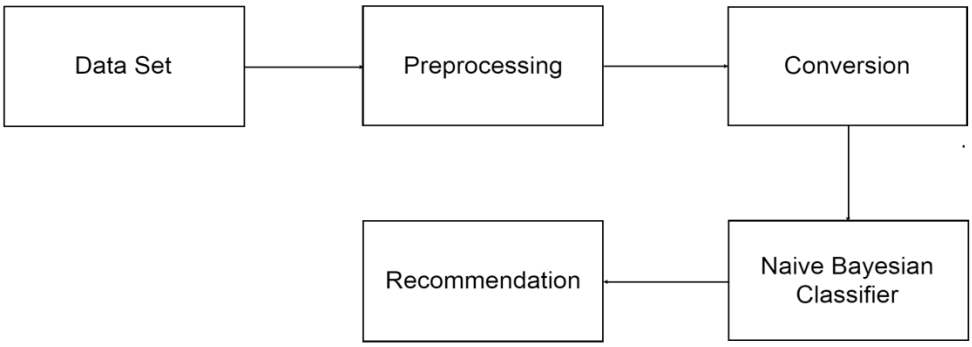


**Figure 2.** The framework of the recommendation model for e-learning

## EXPERIMENTAL WORK

### Dataset

In our approach, we need e-learning dense data set and we wait particularly learner information, educational material and ratings of learners to materials. When used this dense data set, system can suggest most suitable material to learners with a high probability. A small sample of e-learning data set which is expected is below for this study:

**Table 1.** A small sample of e-learning data set

| Learner | Education Material | Rating |
|---|---|---|
| User 1 | Material 2 | 5 |
| User 2 | Material 1 | 4 |
| User 2 | Material 3 | 2 |
| User 3 | Material 2 | 4 |

Firstly, we need to select suitable data set because we do not have a real e-learning data set with continuous values. After researches a lot of recommender data sets have been used in academic works due to efficiency such as MovieLens, Jester, Amazon and Book Crossing.

In this study, Jester Data Set are selected. This data set includes 4.1 million continuous ratings of 100 jokes from 73,496 users which are -10 to +10. Ratings are real values and null values are 99 (Goldberg, 2016). Compared to number of jokes and users this data set is dense so it is suitable to our e-learning recommendation study.

### Experimental Design

We select a subset containing 200, 1000 and 2000 users each rated 100 items. We get rid of the missing values by filling them with the mean of overall ratings. Then we selected five items randomly for each user and produced predictions for them with leave-one-out technique.

We are constrained to convert continuous or discrete dataset into binary dataset. In these conversions, researchers make some assumptions to decide the rating scales to be converted into 'true' and 'false'. If we notate the possible minimum and maximum ratings as $R_{min}$ and $R_{max}$ respectively, common technique is selecting a threshold value $t$ as $(R_{min}+R_{max})/2$ then converting the ratings greater than $t$ as 1 and less than $t$ as 0. In a 1~5 rating scenario, converting 1, 2, 3 into 0 and 4, 5 into 1; is another frequently used technique. We proposed some new approaches to convert continuous data into binary data in the hope of creating more accurate predictions.

After the conversion process, predictions for the selected items are generated via Naïve Bayesian Classifier algorithm. This is a probabilistic classification method based on Bayes Theorem on the work of Thomas Bayes. According to this theorem probability of P(a|b) can be expressed as:

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)} \quad (1)$$

In our converted dataset, binary values 1 and 0 are referred to as particular classes. In Naïve Bayesian Classifier attributes have independent distributions.

$$P(a|b_j) = P(a_1|b_j) * P(a_2|b_j) * P(a_3|b_j) * \dots * P(a_n|b_j) \quad (2)$$

Applying equation (2), class of the target item is determined which can be expressed as the prediction result.

**4.3 Evaluation**

We present our solution with accuracy, specificity, precision, recall, f-score and g-mean metrics using confusion matrix.

**Table 2.** Confusion Matrix for Evaluation

| | **Recommended by System** | **Not Recommended by System** |
|---|---|---|
| **Expected** | True Positive (TP) | False Negative (FN) |
| **Not Expected** | False Positive (FP) | True Negative (TN) |

Formulas are as below:

$$Accuracy = \frac{TP + TN}{TN + TP + FN + FP} \quad (3)$$

$$Specifity = \frac{TN}{FP + TN} \quad (4)$$

$$Precision\ (P) = \frac{TP}{TP + FP} \quad (5)$$

$$Recall\ (R) = \frac{TP}{TP + FN} \quad (6)$$

$$F\ Score = \frac{2\ x\ P\ x\ R}{P + R} \quad (7)$$

$$G\ Mean = \sqrt{P\ x\ R} \quad (8)$$

**Performance Analysis**

We conducted 3 groups of experiments which have different concepts to convert discrete ratings into binary.

In the first group; we make the conversion with respect to quartiles of the rating domain. The quartiles are selected as thresholds and higher and lower values are converted to 1 and 0 respectively. For the Jester dataset, values of Quartile 1 (Q1), Quartile 2 (Q2) and Quartile 3 (Q3) are -5, 0 and 5 respectively because values of this data set are -10 to 10. According to threshold of Quartile 1, ratings smaller than -5 is taken 0, bigger than -5 is taken 1. This implementation is similarly applied to Quartile 2 for threshold 0 and Quartile 3 for threshold 5. Threshold of Quartile 2 is the most widespread approach in applications.

In the second group; conversion is made considering the ratings in the dataset. Consecutively, thresholds are defined for each user and each item separately so the conversions are made separately as well. Besides, threshold value is assigned to overall mean of the ratings and conversion is made with that threshold for all ratings. For User Mean Method (UM), ratings of each user are collected and mean value of each user is calculated. This mean value is a threshold and each user has a different threshold value in this method. For Item Mean Method (IM), ratings of each item are collected and mean value of each item is also calculated. Each item has a different threshold value in IM. For Overall Mean Method (OM), mean value is calculated by using all ratings.

Third group methods are similar to second group techniques for conversion. If mean of ratings is negative, this value is multiplied by -1. If positive, methods are same as second group. Respectively Absolute User Mean (AUM), Absolute Item Mean and Absolute Overall Mean are similar to UM, IM and AOM.

Test results for 200 users and 100 items are shown in Table 3.

**Table 3.** 200 users - Performance of Quartiles and Means

| | **Accuracy** | **Specificity** | **Precision** | **Recall** | **F-Score** | **G-Mean** |
|---|---|---|---|---|---|---|
| **Q1** | 0.8460 | 0.6019 | 0.9503 | 0.8740 | 0.9106 | 0.7253 |
| **Q2** | 0.7850 | 0.8168 | 0.6519 | 0.7182 | 0.6833 | 0.7659 |
| **Q3** | 0.8500 | 0.8681 | 0.4471 | 0.7265 | 0.5535 | 0.7942 |
| **UM** | 0.7490 | 0.7723 | 0.7032 | 0.7179 | 0.7104 | 0.7446 |
| **IM** | 0.7700 | 0.8049 | 0.6701 | 0.7075 | 0.6883 | 0.7547 |
| **OM** | 0.7890 | 0.8214 | 0.6639 | 0.7226 | 0.6919 | 0.7704 |
| **AUM** | 0.8100 | 0.8393 | 0.5019 | 0.6859 | 0.5796 | 0.7587 |
| **AIM** | 0.8130 | 0.8319 | 0.4963 | 0.7297 | 0.5908 | 0.7791 |
| **AOM** | 0.8120 | 0.8351 | 0.5092 | 0.71502 | 0.5948 | 0.7727 |

Test results for 1000 users and 100 items are shown in Table 4.

**Table 4.** 1000 users - Performance of Quartiles and Means

| | **Accuracy** | **Specificity** | **Precision** | **Recall** | **F-Score** | **G-Mean** |
|---|---|---|---|---|---|---|
| **Q1** | 0.8412 | 0.5132 | 0.9382 | 0.8803 | 0.9083 | 0.6721 |
| **Q2** | 0.7642 | 0.8071 | 0.6375 | 0.6785 | 0.6574 | 0.7400 |
| **Q3** | 0.8210 | 0.8532 | 0.3869 | 0.6091 | 0.4732 | 0.7209 |
| **UM** | 0.7552 | 0.7897 | 0.6648 | 0.6975 | 0.6808 | 0.7422 |
| **IM** | 0.7650 | 0.8092 | 0.6946 | 0.6942 | 0.6944 | 0.7495 |
| **OM** | 0.7732 | 0.8177 | 0.6469 | 0.6823 | 0.6641 | 0.7469 |
| **AUM** | 0.8030 | 0.8356 | 0.4567 | 0.6499 | 0.5365 | 0.7369 |
| **AIM** | 0.7954 | 0.8312 | 0.4894 | 0.6513 | 0.5589 | 0.7358 |
| **AOM** | 0.8014 | 0.8387 | 0.4938 | 0.6479 | 0.5604 | 0.7371 |

Test results for 2000 users and 100 items are shown in Table 5.

**Table 5.** 2000 users - Performance of Quartiles and Means

| | **Accuracy** | **Specificity** | **Precision** | **Recall** | **F-Score** | **G-Mean** |
|---|---|---|---|---|---|---|
| **Q1** | 0.8477 | 0.6086 | 0.9502 | 0.8758 | 0.9114 | 0.7300 |
| **Q2** | 0.7784 | 0.8154 | 0.6558 | 0.7042 | 0.6791 | 0.7577 |
| **Q3** | 0.8306 | 0.8683 | 0.4432 | 0.6109 | 0.5138 | 0.7283 |
| **UM** | 0.7564 | 0.7642 | 0.7090 | 0.7461 | 0.7270 | 0.7552 |
| **IM** | 0.7708 | 0.8096 | 0.6889 | 0.7058 | 0.6972 | 0.7560 |
| **OM** | 0.7807 | 0.8169 | 0.6683 | 0.7108 | 0.6889 | 0.7620 |
| **AUM** | 0.7872 | 0.8246 | 0.4880 | 0.6432 | 0.5550 | 0.7283 |
| **AIM** | 0.7972 | 0.8364 | 0.4883 | 0.6372 | 0.5530 | 0.7300 |
| **AOM** | 0.7930 | 0.8347 | 0.5071 | 0.6366 | 0.5646 | 0.7290 |

According to the experimental results, F-score of the Q1 results the best among all techniques. Here Q1 can be thought as an outlier because of the characteristics of the dataset. As we can see from the tables, techniques used in the second group of experiments are all resulted in better f-scores than Q2 while techniques in third group of remain deficient according to f-scores. In the view of g-means, techniques in second group can be selected as the bests. Specificity remains stable in the third group of techniques which may be used for different purposes. Q3 has the best accuracy except Q1 but, there is a critical fall in f-score. Hence we may select the techniques used in the third group for accuracy concerns. Thereupon changes in the user counts do not cause a remarkable change in the specified measures, all of the applied conversion techniques can be approved as scalable.

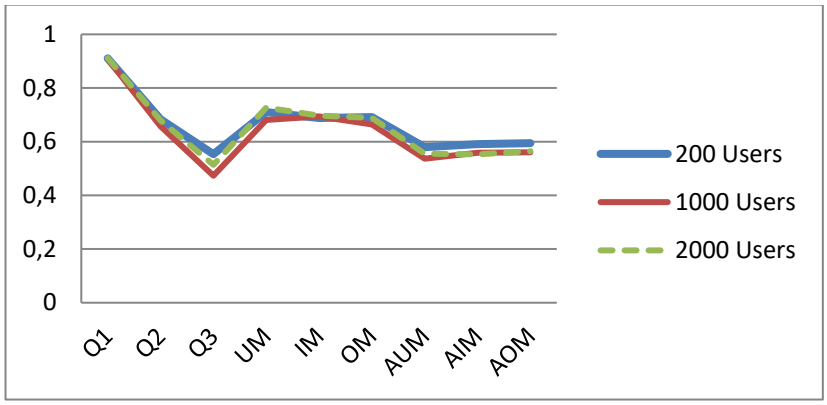


**Figure 3:** F-Score Metric

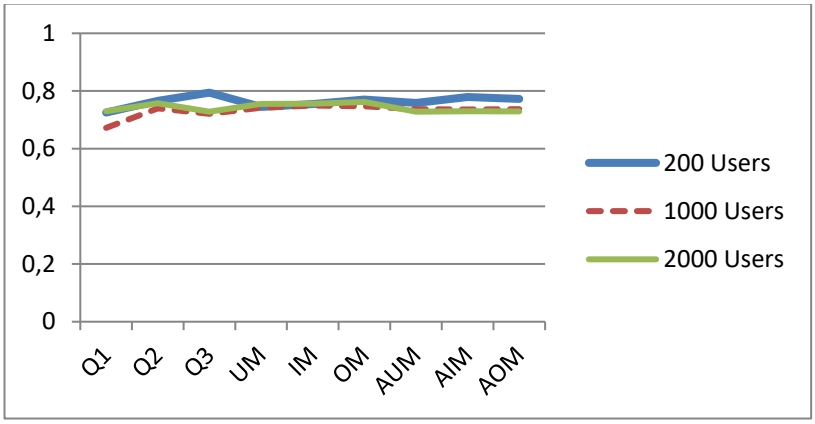


**Figure 4:** G-Score Metric

## CONCLUSION

E-learning is a leading practice for every kind of learners with its tremendous opportunities. On account of the fact that there is a huge amount of e-learning resource on the web, it is inevitable to benefit from a recommender system in order that one can determine the right material to study. We proposed to take advantage of Naïve Bayesian algorithm to achieve this goal. Our study includes the evaluations of several data preprocessing operations applied in continuous to binary conversion step. It is inferred from the results that preprocessing techniques considering the rating means are the best regarding f-measure. The other preprocessing techniques can be preferred to apply through different aspects.

Experiments in this study are held on a different kind of dataset instead of a real e-learning one. In future work, we desire to use real e-learning data set with continuous and discrete values, and improve our approach in this way.

## REFERENCES

Carmona, C., Castillo, G., & Millán, E. (2007). Discovering student preferences in e-learning. In *Proceedings of the international workshop on applying data mining in e-learning* (pp. 33-42).

Chang, Y. C., Kao, W. Y., Chu, C. P., & Chiu, C. H. (2009). A learning style classification mechanism for e-learning. *Computers & Education*, *53*(2), 273-285.

Colace, F., & De Santo, M. (2010). Ontology for E-learning: A Bayesian approach. *Education, IEEE Transactions on*, *53*(2), 223-233.

García, P., Amandi, A., Schiaffino, S., & Campo, M. (2007). Evaluating Bayesian networks' precision for detecting students' learning styles. *Computers & Education*, *49*(3), 794-808.

Melville, P., & Sindhwani, V. (2011). Recommender systems. In *Encyclopedia of machine learning* (pp. 829-838). Springer US.

Özpolat, E., & Akar, G. B. (2009). Automatic detection of learning styles for an e-learning system. *Computers & Education*, *53*(2), 355-367.

Schiaffino, S., Garcia, P., & Amandi, A. (2008). eTeacher: Providing personalized assistance to e-learning students. *Computers & Education*, *51*(4), 1744-1754.

Souali, K., El Afia, A., Faizi, R., & Chiheb, R. (2011, April). A new recommender system for e-learning environments. In *Multimedia Computing and Systems (ICMCS), 2011 International Conference on* (pp. 1-4). IEEE.

Ueno, M., & Okamoto, T. (2007, July). Bayesian agent in e-learning. In*Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on* (pp. 282-284). IEEE.

Ken Goldberg , Jester Data Set, *http://goldberg.berkeley.edu/jester-data/* (Access Date: October 2016)