ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

Personalized Course Recommendation in Formal Learning Based on Logistic Regression

Ms Hanh Nguyen Tran Diem

Lecturer, Faculty of Technology Engineering, Tra Vinh University, Tra Vinh, Vietnam

Abstract: The term Technology Enhanced Learning (TEL) is gradually familiar to both researchers in E-Learning and learners. This development aims to facilitate learners in searching for suitable learning resources such as courses, learning contents which can satisfy those learner's needs or interests. In general, the current techniques of recommender system (RS) play a major role in developing such education applications. There are currently two remarkable trends in building a recommendation system, including collaborative based RS and content based RS. Particularly, each approach employs some different algorithms for implementation depending on applied domains. In this paper, the logistic regression classification is analyzed to design a collaborative filtering (CF) recommendation system for courses in formal training programs where students could be advised to choose some suitable courses to their preferences in an upcoming semester, basing on ratings from previous students who finish the same training program. In addition, the problem of missing values is discussed in detailed. Generally, the purpose of this study is to propose a suitable method for building a CF recommendation system in course domain.

Keywords: Course recommendation systems, Collaborative Filtering, Gaussian Distribution, Missing value, Mean imputation, multivariate Gaussian distribution. Logistic regression.

I. INTRODUCTION

Nowadays, with the popularity of online training courses, learners can improve their knowledge and skills by taking some preferred courses. In reality, the number of distance and online courses have been significantly increasing along with the explosion of internet resources. Therefore, searching for some certain courses which can satisfy to new learner's preferences and their prior knowledge can be complicated and time consuming. Thus, some intelligent programs facilitating course searches in the Internet are necessary in this case. Likewise, in formal education programs, students often follow their advisor's suggestions for choosing the most suitable courses among many optional courses at the beginning of each semester. If we think the problem in another aspect that previous student's preferences can play a special role in helping new students to choose the suitable courses, those students might have more similar suggestions without asking their teachers. In fact, student's ratings on courses taken are relatively easy for us to get through websites of the schools. Generally, this course suggestion designed in formal education programs can be applied in a large number of courses in the Internet.

Regarding to the problem of course recommender system, there are currently several methods such as RS based on learning objects like in [5], course RS based on ontology in [3]. In those preliminary studies, they considered the Collaborative Filtering Recommendation system because of several reasons. Firstly, CF recommendation systems probably prove their strength in generating relevant suggestions especially for the cold start problems. A large number of CF recommendation systems used rating results from previous users in some certain domains and were very successful in prototyping phases. The domains might be very various such as music, tourism, books, products, etc. Secondly, using ratings on the courses of previous students is rather appropriate in both theory and practice.

In addition, the problems of missing values frequently happen in real survey tables because of various reasons. Proposing different methods to fill in those blanks is concerned in many previous researches for many different domains. In this paper, we examine Expectation Maximization (EM) algorithm on Gaussian distribution in estimating missing values for course domain particularly. In addition, we also make a comparison between the EM and Mean Average in predicting suitable values for missing places in terms of the accuracy and physical resource usage. Finally, the Logistic regression classification algorithms is used. One of the advantages of using classification over similarity methods such as Cosine Similarity or Pearson Similarity is efficiency feature in execution. Specifically, while logistic regression can run several times to train the data and apply for many users, the similarity techniques have to go over each data case for searching similar interests. In practice, the latter is not suitable if the number of users and data training samples are really large. Therefore, using logistic regression in course RS is probably much more effective than the Similarity techniques.

Through this study, we find out some good features in recommendation system with five scale rating. Firstly, experiments in this study show that there are not many big differences among the proposed methods of missing value with respect to classification results. Secondly, one minor drawback is that EM method takes more memory and CPU usage than Mean Average Imputation. Therefore, with some large scale applications with a big number of training samples and users, designers should take more consideration in choosing a right approach.

The rest of the paper is organized as follows. In Section 2 we present some previous researches in recommendation systems for education domain specifically. In this section, the logistic regression algorithm for incomplete data is also analyzed. Section 3 is background approach which is needed for experiments in course recommendation system in section 4. Experiments and results of course recommendation system is presented in section 4. Discussion for the experiment is presented in section 5. Finally, the conclusion and future work are discussed in section 6.

II. RELATED WORKS

Curriculum layout**:** In order to prepare new students for good training in a certain university, the training curriculum should be available long before the first semester. That university designed and published his own training curriculum frames of different majors to learners by their school websites. Particularly, the curriculum layout also displays schema of syllabuses, course credit and course description. In general, there are some close correlations between courses in different semesters in the program. Each course has some requisite courses which students are required to finish first. However, there are also some courses with no requisite courses. Curriculum layout tells learners about optional courses and compulsory courses in each semester separately as well. Basing on the layout, students can choose to follow some optional courses which are suitable to their preferences and capabilities at the beginning of each semester.

Learning outcomes: Each course includes inertly several learning outcomes, which was designed and composed by the teacher who will conduct the course. The learning outcomes describe major knowledge and skills which the teacher should follow in his/her teaching. Similarly, the students must be required to gain at a certain basic level of the knowledge and skills after finishing the course. Learning outcomes in a course might relate to other learning outcomes in other courses in the whole program. Therefore, if learners have not finished one of the learning outcomes in a conducted course, those students would go back to that course to study again. The next semester syllabus can be known by reading in the course diagrams. Learning outcomes could be used as in meaningful input for a recommendation system and were studied in some researches. However, in this study we just focus on student's preferences on courses in a general manner for simplicity.

A general layered model for course recommendation system is presented in **[5].** In this model, one knowledge base for learning recourses like courses and course outlines is used to make a recommendation. The calculation for TF and IDF in learning documents are also mentioned in the study to identify the most relevant resources for recommendation. Both technique and model in **[2]** for designing and building an application of recommendation for lifelong learning are analyzed in detail. The authors of that paper used user's demographics information for reasoning user preferences over learning resources. In addition, the techniques like neuron network, classification with probability like Bayesian network and latent semantic analysis are exploited in their recommendation systems.

Workflow for course recommendation is presented in **[10].** In addition to available relational operators, the authors also develop some new operators which can operate data in database for different results of recommendation. Authors in **[1]** also believe that the method of CF approach can not provide appropriate advices to specific users in the domain of learning resource presented in ontology.

Logistic regression can be applied in supervised learning where the outcome of each sample are already designed in the training set. While it is rather easy to classify with binary outcomes, there is rather problematic to deal with multiple outcomes in learning the logistic model for a data set. The implementation in 3 outcome labels A, B and C for a training set is explained clearly in **[8].**

Missing values is very common in survey scheme when responders let blank in some particular questions. The reason for that problem might be the lack of experience of responders in providing suitable answers to asked questions. In this situation, a good recommendation system must contain a procedure to solve the missing values before implementing classification algorithms. There are 3 main types of missing values. Those are Missing Completely at Random (MCAR) where probability of records with missing values not depending on observed features. Missing at Radom (MAR) is categorized when features having missing values depend on observed values. For the Not Missing at Random (NMAR) features, designers need to examine the data cases first because those features must have their values inputted by responders .

Regarding to missing problems, there are many algorithms for dealing with missing values proposed in the last few years. Mean methods, the imputation approach, estimating missing values by depending on observed values, reducing features with missing values are discussed in many studies as well. Using a certain approach depends on training samples.

Incomplete data for logistic regression is proposed in **[9]** where the authors investigated methods to predict missing values thank to the observed data and Gaussian distribution. The Gaussian Mixture Model (GMM) using Expectation Maximization and Variance Bayesian Expectation Maximization plays the most important role in predicting missing real values **[7].** Similarly, mixtures of Bernoulli is used for predicting discrete missing values.

 DOI 10.17148/IJARCCE.2015.410119

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

Those authors also presented 2 core steps in building an EM system, comprising E step and M step.

Multiple imputation in solving the missing value problems is a main content in **[12]** in which filling a missing value *x(i,j)* in the data training samples with an average value of probabilities from a set *S=(1,2,3…S)* is done first. Particularly, combining multiple imputation with cross validation before building the imputation can increase the efficiency. In general, the authors also mentioned that this imputation could solve many clinical analysis with missing values. However, they also pointed out several reasons of not using imputation in **[12]** such as too much work to do when the data samples have a large number of data cases and the wrong inferences may occur.

Replacing missing values with the Mean is implemented in training cases having numeric missing values. In study **[13]**, the use of this method in C4.5 and CN2 algorithms can lower the error rate in some data sets. K- nearest neighbor imputation estimates values at missing positions basing on similar data rows in the same cluster. Manhattan distance and Euclidean can be introduced to K nearest neighbor imputation method (Math Works inc 2013) for generating the separate clusters. Weighted Mean can be combined with K Nearest neighbor to generate Expectation for missing values (also in Math Works inc, 2013) as well.

Reducing method deals with removing data cases having missing values or features with missing values. This method may reduce the data content and affect the accuracy in the final classification results. However, if the number of missing values is small while we have a very large dataset, we could benefit from this method.

III. BACKGROUND

This section is divided into two parts which are logistic regression and dealing with missing values. The first section represents data representations, logistic hypothesis function and method of reducing the cost between the hypothesis and results. The second section tells us about some proposed approaches to estimate appropriate values for missing places. Specifically, Mean average and EM algorithm are analyzed in estimating values for missing places.

A. Logistic regression

Firstly, we define some notations that could be used later

n = *<number of courses in our training data samples>*

m = *<number of responds from students according to the list of courses>*

$x^{(i)}$ = *<values of responds from student i^{th} in the training sample>*

$x^{(i)}_{(j)}$ = *<value of respond of student i^{th} on the course j^{th} >*

y = *<class labels >*

X = *<Training data samples>*

Then, the hypothesis model for linear regression with multiple features $\mathbf{x_i}$, $\mathbf{i \subset N}$ is defined as following:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots \theta_n x_n \quad \textbf{(1)}$$

We assume that $x_0=1$

$$h_\theta(x) = g(\theta^T x) = \mathrm{g(z)} \quad \textbf{(2)}$$

$$where\ z = \theta^T \mathrm{x}$$

Let call $h_\theta(x)$ is a hypothesis for our classification model

$$\mathrm{h}_\theta(\mathrm{x}) = \mathrm{g}(\theta^T \mathrm{x}) = \frac{1}{1+e^{-\theta^T x}} \quad \textbf{(3)}$$

Where the logistic function (*sigmoid function*) is

$$g(z) = \frac{1}{1+e^{-z}}$$

In this case *g(z)* and *h(x)* are bounded between *0 and 1*

$$\theta^T x = \theta_0 + \sum_{j=1}^{n} \theta_j x_j \quad \textbf{(4)}$$

The cost function for presenting the difference between our hypothesis corresponding to the input value x is $h_\theta(x)$ and the real value of class *y*.

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & if\ y = 1 \\ -(1 - \log(h_\theta(x))) & if\ y = 0 \end{cases} \quad \textbf{(5)}$$

Finally, we can use the update algorithm named Gradient Descent. Let α be the learning rate for the following algorithm convexes.

Repeat:{

$$\theta_j := \theta_j - \alpha \frac{\delta}{\delta \theta_j} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}) \quad \textbf{(6)}$$

$$Simultaneously\ update\ \ \theta_j$$

} **[13]**

B. Dealing with the problem of Missing Value

Missing value is a common problem in survey results because responders would not like to enter their answers to some questions in our survey tables. Missing values could make the data analysis more complicated and cause classification results biased. Therefore, it is essential to resolve the missing value problem before applying any classification algorithm to our dataset should be considered carefully. Most of the techniques dealing with missing values are replacing all missing values with appropriate estimated values. These are imputation with mean average and estimated values basing on data distribution of responding values in the dataset.

Mean Imputation: [12]

For the mean imputation, this is rather straight forward and easy to implement. Likewise, the method is very useful when dealing with missing at random case. **[8].** Suppose we have X_i =[x_1, x_2,…x_n] is a data case with some missing values at j^{th} positions in the *X* training data. We impute those missing values in row i^{th} with the mean calculation as following:

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

Then $X[i,j] = \frac{1}{n}\sum_{k=1}^{n} X[k,j]$, where $X[k,j]$ is not a missing value

Multiple Imputation [12]

This method is completely different from the mean imputation in terms of the estimated values replaced for missing values. While mean imputation is simply replaced with the mean value in the rows having missing values, the multiple imputation calculates the mean posterior probability of outcome y given the data case x as $P(y|x)$ where $x=[x^o,x^m]$, x^o and x^m are observed data and missing data respectively. The algorithm repeatedly calculates P by replacing x^m with some given values s; s={1,2,3…S} **[8].** Then

$$P(y|x_*^o) = \frac{1}{S}\sum_{s=1}^{S} P^s(y|x_{*s})$$

Multiple Imputation can produce better results when the training data samples are divided into cross validation data cases. However, this method requires much work on calculation the posterior probability.

Expectation Maximization (EM) in Gaussian Distribution [13]

In Normal Gaussian Distribution, let consider a vector $X=\{X_1, X_2,\ldots X_n\}$ as a set of real values in Gaussian Distribution $N(\mu,\sigma)$. Where the mean μ and squared deviation σ^2 of X are calculated as following:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i\,; \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

Then the probability $P(x_i, \mu, \sigma^2)$ for the element x_i in that vector X according to above Gaussian distribution μ, σ as following:

$$P(x_i, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

In Multivariate Gaussian Distribution, it is similar to the Gaussian Distribution. Except the fact that multivariate Gaussian Distribution calculates Mean and Sigma all in one. Specifically, probability of $x \in \mathcal{R}^n$ with $\mu \in \mathcal{R}^n$ and $\Sigma \in \mathcal{R}^{n\times n}$ is calculated like following:

$$P(x;\mu;\Sigma) = \frac{1}{(2\pi)^{n/2}\,|\Sigma|^{\frac{1}{2}}} exp^{\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)}$$

Expectation Maximization [7]

Let $X=\{x_1,x_2,\ldots x_n\}$ is a set of observed data samples

$Z=\{z_1,z_2,\ldots.z_m\}$ is a set of missing data

$\theta=\{\theta_1, \theta_2,\ldots \theta_n\}$ is a set of unknown parameters for Multivariate Gaussian distributions. Where $\theta_i = (\mu_i,\delta_i^2)$

The likelihood function $L(\theta;X,Z) = p(X,Z|\theta)$. The maximum likelihood is defined by the marginal likelihood like this

$$L(\theta|X) = p(X|\theta) = \sum_{p} p(X,Z|\theta)$$

The EM algorithm to find MLE of the marginal likelihood by applying the E and M steps repeatedly.

Expectation step (E step): calculate the expected value of the log likelihood function:

$$Q(\theta|\theta^t) = E_{z|x,\theta^t}[\log L(\theta;X,Z)]$$

Maximization step (M step): Find the parameter that maximizes the above value

$$\theta^{(t+1)} = \underset{\theta}{\arg\max}\, Q(\theta|\theta^t)$$

Update value Mean and Sigma in Gaussian Distribution

$$\mu^{i+1} = \frac{\sum P(i)*X}{\sum P(i)};$$

$$\sigma^{i+1} = \frac{\sum P(i)(X-\mu^i)(X-\mu^i)^T}{\sum P(i)}$$

Case deletion

The method called *complete case analysis* deletes all the whole data cases with missing data. This method is available in some current statistical programs like math work or R. Another method called *discarding instances or attributes* deletes features with missing values. These two methods might result in some negative effects for the classification results because in some cases, a large amount of valuable information can be deleted in our training data samples. Therefore, prior researches in those methods recommend checking correlations between instances or attributes and observed data samples to increase the accuracy in the classification systems.

IV.EXPERIMENT AND RESULT

In this section, we present the results of the experiment in the course dataset in computer science domain. This dataset is for students who are studying in 4th semester and those students would like to have course suggestion for their upcoming 5th semester. The dataset stimulates the ratings on courses from semester 1st to semester 4th. The ratings are done by students who already finished the computer science program previously.

The structure of this course dataset consists of 16 features which are courses from the 1st semester until the 4th semester. Each feature has integer value from 1 to 5 where 1 is for less likely and 5 is for strong likely according to 5 scale ratings. The classification labels are from 1 to 5. These course labels represent suggestive optional courses for students to choose in their 5th semester.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

We collect about 107 data cases rated from previous students who already finished the computer science training program. There are 6 optional courses with missing data because the students did not take some of the courses in their learning duration. We do not have any missing value in class labels because students must present the chosen courses in their 5th semester. Likewise, there are 10 compulsory courses rated by all interviewed students without having any missing value. If there is any missing in those features, the corresponding data cases should be deleted and left out from our data training samples.

In this study, we use mean imputation, Expectation Maximization with Gaussian Distribution to fit the training data before applying logistic classification.

A. Course interpretation

TABLE 1. LIST OF COMPULSORY COURSES

#	Code	Course Description
1	CTT003	Introduction to Programming
2	CTT008	Advanced programming
3	CTT006	Object Oriented programming
4	CTT005	Mobile Programming
5	CTT104	Assembler
6	CTT101	Data Structure and Algorithm
7	CTT102	Database Design
8	CTT103	Operating System
9	CTT105	Networking
10	CTT303	Introduction to Artificial Intelligent

The courses in columns from 1st to 10th must have values inside because those are compulsory courses that the students need to study from the semester 1st to semester 4th. The following courses in this below table are optional and students can leave blank in some courses if they did not take those courses. For this reason, students could choose from 2 until 4 courses in the whole program.

TABLE 2. LIST OF OPTIONAL COURSES

#	Code	Course titles
1	CTT304	Computer Graphics
2	CTT301	Automata
3	CTT310	Image Processing
4	CTT305	Data Mining
5	CTT308	Complexity Calculation of Algorithms
6	CTT323	Embedded programming

The below courses are used in recommendation. Students could choose from 2 until 5 courses below. For this prototype, we just use 5 courses for demonstration. However, this list of courses could be longer in real applications.

TABLE 3. LIST OF TARGET COURSES

#	Code	Course title	Requisite courses
Class 1	CTT302	Knowledge Engineering	CTT303
Class 2	CTT306	Machine Learning	CTT303
Class 3	CTT322	Artificial Intelligence	CTT303
Class 4	CTT335	Web programming	
Class 5	CTT334	Advanced Data Mining	CTT305

B. The Probability Table

After running the logistic classification on the mentioned data samples, we have the following probability table. The first row in table 4 indicates the targeted class. The remain rows display the probability value according to the targeted class. If a certain new student provides their interest in 16 courses, we can suggest that student a course by multiplying interest values with the probabilities in each column. Finally, we just rank the results for recommendation.

TABLE 4. COLUMN EXPECTATION MAXIMIZATION AND GAUSSIAN DISTRIBUTION PROBABILITY TABLE

#	Class 1	Class 2	Class 3	Class 4	Class 5
0	0.537824	0.246455	-5.78773	-0.33851	-0.14991
1	0.012856	-0.68068	-0.265	-0.18333	0.72941
2	-0.67223	0.584491	0.327506	0.148456	-0.23931
3	-0.47026	0.881645	-0.17188	-0.23739	0.47106
4	0.91533	-0.60513	0.129919	-0.34848	-0.2725
5	0.510647	-0.13232	0.018791	-0.32456	-0.0947
6	0.02825	0.593448	0.368505	0.142751	-0.55072
7	-0.33009	0.198065	-0.8472	0.253803	0.320257
8	-0.74537	-0.16346	0.04073	0.383817	0.239285
9	0.464628	-0.51707	0.287898	0.120196	-0.23662
10	-0.70715	-0.35619	0.909447	0.228482	0.143244
11	0.085495	0.100616	-0.88999	0.502613	-0.40325
12	-0.62979	-0.27211	0.778584	0.198741	-0.51828
13	0.491025	0.193069	1.079327	-0.47432	-0.93434
14	-1.16678	-0.46293	0.001476	-0.2648	1.218516
15	1.048182	-0.05492	-0.42583	-0.56374	-0.3399
16	0.608436	-0.44624	-0.13468	-0.25523	-0.11703

Table 4 is one of typical result when we use the Expectation Maximization and Gaussian Distribution. The other results are calculated similarly.

C. Histograms of the data

The two following histograms show the distribution of chosen courses from students in our data training samples and the distribution of student's ratings after solving missing values.

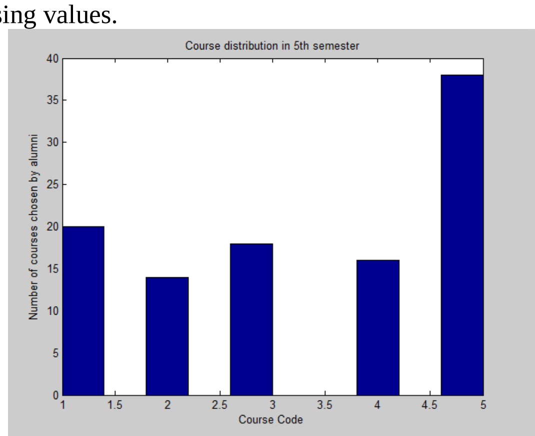

Fig1. Training Data Samples

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

The histograms could show initial intuition to us. Firstly, students who show less interest in most courses tend to choose to study course coded class 1. In contrast, students who evaluate highly in most courses tend to take course coded 5. Interestingly, courses are chosen by students is 3 for those who rate 3 in most courses.

The above figure presents the histograms of our experiment after we use those five methods for filling missing values. We can see from the figure that the distributions are not very different between the first two methods which are Row Mean Average and Column Mean Average. In contrast, there is a rather big difference between applying EM in Columns and EM in Rows respectively. However, the prediction accuracy of the top two methods is not deviated considerately. This trend is similar to the last two methods.

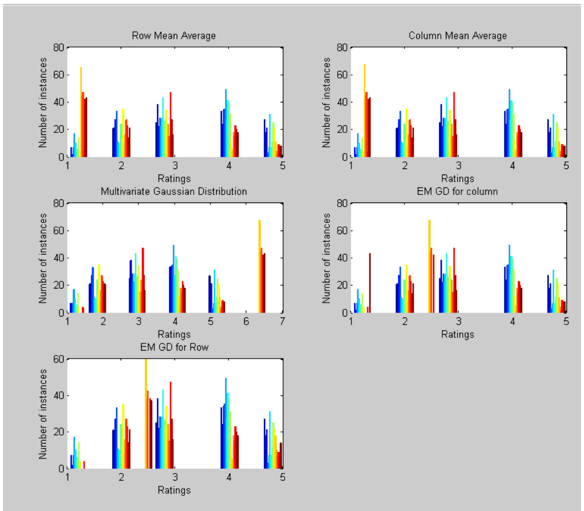

Fig2. Training Data Samples after fixing missing value

D. Scenarios

Suppose, we have a new student who shows his interest in compulsory courses from 1 to 10 and optional courses from 11 to 16 like the table 5 below. Our program needs to suggest him two courses in an upcoming semester. The two courses must be in this list {'Knowledge Engineering'; 'Machine Learning'; 'Artificial Intelligence'; 'Web Programming'; 'Advanced Data Mining'}

TABLE 5. COLUMN EXPECTATION MAXIMIZATION AND GAUSSIAN DISTRIBUTION PROBABILITY TABLE

#	Code	Course Description	Ratings
1	CTT003	Introduction to Programming	4
2	CTT008	Advanced programming	3
3	CTT006	Object Oriented programming	4
4	CTT005	Mobile Programming	5
5	CTT104	Assembler	4
6	CTT101	Data Structure and Algorithm	3
7	CTT102	Database Design	2
8	CTT103	Operating System	1
9	CTT105	Networking	5
10	CTT303	Introduction to Artificial Intelligent	4
11	CTT304	Computer Graphics	2
12	CTT301	Automata	1
13	CTT310	Image Processing	4
14	CTT305	Data Mining	2
15	CTT308	DPTTT	3
16	CTT323	Embedded programming	3

Table 6 represents the results of running 5 fixing missing value algorithms. We could see from the table that algorithm 5 shows a higher prediction probabilities than remain algorithms. So, we can choose the courses like Knowledge Engineering and Machine learning as our recommendation for the given student.

TABLE 6. RECOMMENDATION RESULTS

Course titles	Alg1	Alg2	Alg3	Alg4	Alg5
Knowledge Engineering	0.42	0.22	0.38	0.14	0.52
Machine Learning	0.13	0.41			0.57
Artificial Intelligence			0.20		
Web programming					
Advanced Data Mining				0.36	

Alg1: Row Mean Average Fixing
Alg2: Column Mean Average Fixing
Alg3: Multivariate Gaussian Distribution Fixing
Alg4: Gaussian Distribution for columns Fixing
Alg5: Gaussian Distribution for rows Fixing

V. DISCUSSION

For the Mean Imputation, this can be seen as a very promising method for filling missing values because the method leads to classification results with higher accuracy. Particularly, Column Mean Imputation outperforms Row Mean Imputation as different courses have different difficult levels. It is clearly that some training courses are difficult and students often get lower scores in those courses. As a consequence, majority of students might rate lower in the courses when the formers are asked to fill in a course survey. This results in mean values of the corresponding columns be smaller than those of the row. Therefore, it is more appropriate to apply the mean column in the case of similar course recommendation systems.

Replacing missing values with zeros can also recommend high relevant courses. However, one limitation of this method is under-estimating rated values in taken courses in prior semesters. If we consider those ratings generally, they could tell some information about the course difficult levels and student's preferences in those courses. One advantage of the zeros imputation is that performance efficiency as it does not require too much calculation and memory utilization. In general, the above results show that EM and Mean Imputation can be applied to fit missing value in training data samples first. Then the logistic regression is executed later for providing more relevant recommendations in the course domain.

Case deletion algorithm is inappropriate in the course classification problems because this method might delete

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

all data cases in our training samples. The reason is that most ratings from user i^{th} , X_i contain some missing values as not many students can take all provided optional courses. Therefore, deleting a large number of data cases might happen. Consequently, when training samples is narrowed down considerately, the accuracy of classification results might become lower.

VI. CONCLUSION AND FUTURE WORK

Our main contribution in this study is dealing with missing problems and classification with logistic regression in recommending optional courses to students in undergraduate education context. Specifically, the problem of missing at random is very common in course domain. Therefore, tackling this problem is the first step in building a course recommendation system. For this purpose, EM algorithm in estimating the maximum rating values for replacing with missing values is used appropriately. The techniques of imputation and Expectation maximization for collaboration filtering recommendation systems are also discussed and prototyped in this study.

Future works would be studies in seeking the coefficient among features of courses together. Normally, courses rated belong to a certain skills' group which might be meaningful in CF Recommendation System. Group recommendation should be considered in finding matched recommendations because there are definitely many students following the same certain branch in computer science. Therefore, similar preferences among alumni students can be analyzed for providing more relevant recommendations. Identifying more features which are important in the final recommendation results should be considered in detail so that the accuracy of the consultation results can be improved.

ACKNOWLEDGEMENT

The study was done in Computer Science Department of Gerona University, Spain in 2015. This work is completed thank to the Techno Project belonging to Erasmus Mundus Fund. We also express my gratitude to **Dr Ramon Fagregat** and **Dr Silvia Margarita Baldiris Navarro** in Gerona University for their useful comments.

REFERENCES

[1] Nikos Manouselis, Hendrik Drachsler, Riina Vuorikari, Hans Hummel, Rob Koper, "Recommender Systems in Technology Enhanced Learning" Greece, Nertherland and Belgium, in 2010.
[2] Hendrik Drachsler, Hans G.K. Hummel and Rob Koper, "Personal recommender systems for learners in lifelong learning networks: the requirements, techniques and model", in International Journal of Learning Technology, Volume 3 Issue 4, Pages 404-423, July 2008
[3] Liping Shen and Rui-min Shen, Learning Content Recommendation Service Based on Sequencing Specification, Lecture notes in Computer Science, page 363 – 370, Volume 3143, Springer, 2004.
[4] Kun Hua Tsai, Ti Kai Chiu, Ming Che Lee, Tzone I Wang, "A learning Object Recommendation Model based on Preference and ontological approach", Sixth International Conference on Advanced Learning Technologies , 5-7 July 2006.
[5] Xiao Hong Tan and Ruimin Shen, "Personalized course Generation on Layered Recommendation systems", E. Popescu et al. (Eds): ICWL 2014, LNCS 8613, pp 166-172, 2014, @ Springer Publisher in Switzerland 2014.
[6] Crawley, E.F., Malmqvist, J., Östlund, S., Brodeur, D.R. Rethinking Engineering Education, the CDIO approach, Springer, 2014.
[7] Zoubin Ghahramani and Michael I. Jordan, "Supervised learning from Incomplete data via an EM approach", Advances in Neuro Information processing System 6, Morgan Kaufmann Publishers, San Francisco, CA, 1994,
[8] David W. Hosmer JR, Stanley Lemeshow and Roxney X. Sturdivant, Applied Logistic Regression, Third Edition, Willey Publisher, 2013.
[9] David Williams, Xuejun Liao, Ya Xue and Lawrence Carin, "Incomplete Data using Logistic Regression", Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005.
[10] Georgia, Benjamin Bercovitz, and Hector Garcia Molina, "FlexRecs: Expressing and Combining Flexible Recommendations", SIGMOD 2009, Providence, Rhode Island.
[11] N. Manouselis, R. Vuorikari† & F. Van Assche, "Collaborative recommendation of e-learning resources: an experimental investigation", Journal of Computer Assisted Learning (2010), 26, 227-242, Blackwell Publishing.
[12] Gustavo E. A. P. A. Batista and Maria Carolina Monad, "An analysis of four missing data treatment methods for supervised learning", Applied Artificial Intelligence, Vol. 17, 5-6, pp 519-533, 2003.
[13] CS229 Materials, Machine Learning course materials, Available: http://cs229.stanford.edu/materials.html.