# Effective Active Course Learning Method for Education Data Using Feature Base Classification Model

[#]Mrs.R.Nirmala Devi MSc.,M.Phil., [*]Ms.Anupriya Sharma MCA.,(M.Phil).,

[#](*Asst. Professor/Department of Computer Science*)
[*](*Department of Computer science*)

***Abstract:***

*Educational researchers commonly use a variety of methods such as classroom observation, content analysis, surveys and interviews to collect data and analyze teachers' reflective thinking. This paper describes the different methods and reveals in depth meanings within the unstructured data, but is time-consuming and cannot be implemented in a large scale. The objectives from these paper studies may be subject to the subjective impression of educational researchers. Moreover, the results obtained from these methods are lagging behind and cannot help teacher trainers make timely intervention policies. On the other hand, the data-driven approaches such as educational data mining and learning analytics are able to analyze mass of relative data and visualize results. The proposed approaches, however, are mainly used for the analysis of structured data, including learning behavior data, performance data and administrative data recorded in effective course review management systems (ECMS) or online learning environments (OLE) such as Moodle. In this proposed system analysis the inductive content analysis and a common classification method for analyzing review content.*

**Keywords***: Active Education Data Mining, Teacher Reflection Data, TF-IDF Classification, Machine Learning***.**

## I. INTRODUCTION

Educational data mining develops computational and psychological methods and techniques for understanding how students learn by collecting and analyzing student data, discovering learning patterns and trends, and making new discoveries about how student learn. Learners' online discussion data has been collected and analyzed to generate specific knowledge of performance, learning behaviors and experiences. A representative set of classification algorithms has been used for predicting whether students will pass or fail the course review on the basis of data about their online discussion forums usage. Reliable and applicable proxy variables that reflect theoretical and empirical evidence and a prediction model were constructed, and results indicated that the predication model was highly accurate and early detection and timely interventions were possible. Similar studies have been conducted for understanding learners' complex problem solving knowledge building. Educational data mining researchers use a variety of methods and applies techniques from statistics, machine learning and data mining, including prediction, classification, clustering, relationship mining, distillation for human judgment, and discovery with models.

Among these methods and applies techniques, course review classification is a very mature research field and most relevant to this study.

Popular classification algorithms have been used in educational data mining and machine learning domain, including logistic Regression, Naive Bayes, Support Vector Machine (SVM), Decision Tree, Boosting, etc. Based on the number of classes that each data point falls into, there are single-label and multi-label classification. Each data point can merely fall into one class in single-label classification and all classes in single-label classification are mutually exclusive. The single-label classification system includes binary classification and multi-class classification approaches. There are only two classes in binary classification, while more than two classes in multi-class classification. In multi-label classification, however, each data point can fall into several classes in the meantime. In this study, a single-label classification model was built based on inductive content analysis and we allowed each online discussion data point to fall into one category in the meantime. This classification model was then applied on a large-scale and unexplored online discussion course review data set for understanding teachers' reflective thinking.

## II.RELATED WORKS

"Mining Social Media Data For Understanding Students' Learning Experiences" by Mihaela Vorvoreanu and Krishna Madhavan [1] study is

beneficial to researchers in learning analytics, educational data mining, and learning technologies. It provides a workflow for analyzing social media data for educational purposes that overcomes the major limitations of both manual qualitative analysis and large scale computational analysis of user-generated course reviewual content. Our study can inform educational administrators, practitioners and other relevant decision makers to aim further understanding of engineering students' college experiences. As an initial attempt to instrument the uncontrolled social media space, we propose many possible directions for future work for researchers who are interested in this area.

.

"Representation And Communication: Challenges In Interpreting Large Social Media Datasets" [2] by Mattias Rost, Louise Barkhuus, Henriette Cramer and Barry Brown describe the online services provide a range of opportunities for understanding human behavior through the large aggregate data sets that their operation collects. These analyses rest upon the assumption that large-scale data sets are representative, in some quantifiable way, of real world behavior. Extensive work has been done to understand the relationships between these databases and human behavior, such as inferring location from humorous posts, or extracting 'real' from 'fake' online reviews. As we argue in this paper, however, this focus on the representativeness of such data inadvertently neglects the communicative features of social network data sets. In this paper contribution here does not centre on the issue of deciding what these large datasets can be used for, but rather about exploring emerging genres of communication. In conclusion we would argue more broadly for the role of check-in services like Foursquare not as location-based services, but rather as particular communication genre, with particular developing forms and style

"Using Video to Develop Skills in Reflection in Teacher Education Students" [3] by Anne M. Coffey describe the results of this research confirm the use of video in developing skills in critical reflection. Whilst the students in this research were engaged in a peer teaching situation video has been used in a variety of other ways. Given the portability of digital video cameras and the rich opportunity that is provided by students seeing they teach the continued exploration of ways to incorporate such experiences into teacher education programs is a worthwhile undertaking. In graduate teacher education programs, which demand that students develop their teaching skills in a much shorter time period, the use of video and analysis tools such as Critique would seem to be particularly useful. As noted by it is not sufficient to simply provide practical opportunities for students to develop their reflection skills. These opportunities need to be very purposeful in order to facilitate the development of these skills. The use of video would seem to be a very powerful ally for use in developing these important skills in critical reflection in graduate entry teacher education programs.

"Validity in Quantitative Content Analysis" [4] by Liam Rourke and Terry Anderson describe the article divided into three phase. The first phase describe with the observation that QCA is a form of testing and measurement but notes that the procedures of test development codified in the psychometric literature are given meager consideration in QCA research. The second phase describes the process of constructing a coding protocol that is theoretically valid, or as Sheppard says, reasonable. This phase draws on presentation of essential steps in test construction. The third phase follows discussion of several types of empirical studies that can be conducted to establish the validity of inferences derived from a testing procedure.

'Incorporating Online Discussion In Face To Face Classroom Learning: A New Blended Learning Approach" [5] by Wenli Chen and Chee-Kit Looi discusses an innovative blended learning strategy which incorporates online discussion in both in-class face to face, and off-classroom settings. Online discussion in a face to face class is compared with its two counterparts, off-class online discussion as well as in-class, face to face oral discussion, to examine the advantages and disadvantages of the proposed strategy. However, the lack of face to face interactions and the need for sufficient time to do online postings pose challenges in implementing online discussion for face to face classroom learning.

In this study, all the learners were adults who were highly motivated to participate because the discussions were closed related to their work. This online group showed good participation even though the discussions were not graded. Due to the specific concourse review for this study, generalization from the findings may be limited. The finding that in-class online discussion is a useful strategy may be concourse review dependent. Therefore, further research is needed to explore the effectiveness or weakness of in-class online discussion for blended learning in broader concourse reviews.

## III. MACHINECLASSIFICATION EARNING MODEL

This is a task performed to generalize known structure in data mining to apply to new course review data. It is also the categorization of data for its most effective and efficient use. There are numerous data mining classification algorithms being studied and implemented in different domains. Some of the most popular and common are adapted and presented herein, based on their capabilities simplicity and robustness.

### *A. K-Nearest Neighbor (k-NN)*

The principle behind this method is to find predefined numbers of training course review samples closest in the distance to the new point and predict label from these. The number of samples can be a user defined constant or varied based on the local density of points. The distance can be any metric measure. There are distance measures implemented in the k-NN, Euclidean, Chebysheb, Manhattan and Edit Distance, but the Euclidean distance measure is the most common choice. Despite its simplicity it is successful in large number of classification problems.

### *B. J.48*

J4.8 decision trees algorithm is an open source Java implementation of the C4.5. It grows a tree and uses divide-and-conquer algorithm. It is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. To classify a new course review item, it creates a decision tree based on the attribute values of the training course review data. When it encounters a set of course review items in a training course review dataset, it identifies the attribute that discriminates. It uses information gain to tell us most about the course review data instances so that it can classify them the best.

### *C. Naïve Bayes (NBC)*

This classifier is based on the Bayes rule of conditional probability. It uses all of the attributes contained in the course review data, and analyses them individually as though they are equally important and independent of each other. The Naïve Bayes classifier works on a simple, but comparatively intuitive concept. It makes use of the variables contained in the course review data sample, by observing them individually, independent of each other. It considers each of the attributes separately when classifying a new instance. It assumes that one attribute works independently of the other attributes contained by the course review sample.

### *D. Multi Layer Perceptron (MLP)*

MLP is a feed forward artificial neural network model that maps sets of input course review data onto a set of appropriate outputs. It consists of multiple layers of nodes, with each layer fully connected to the next one. Each node is a neuron with a nonlinear activation function. It uses a learning technique called back propagation for training the network.

### *E. Linear Regression (LR)*

It is a statistical measure that can be used to determine the strength of the relationship between one dependent variable and a series of other changing variables known as independent variables (regular attributes). If independent variable contains multiple input attributes like in our research (rainfall, sunshine hours, humidity, pH etc), then it is termed as multiple linear regressions. Linear regression provides a model for the relationship between a scalar variable and one or more explanatory variables.

### *F. Artificial Neural Network (ANN)*

One widely used artificial neural network, back propagation neural network (BPNN), was applied to predict rice yield because of its simplicity in structure and robustness in simulation of nonlinear systems. A typical three-layer BPNN comprising one input layer, a hidden layer, and an output layer were used in the current study (Fig.3.1). The neurons of adjacent layers are connected by the nodes' weights. There are two weights in the three-layer BPNN, which are vij between input and hidden layers and ujk between hidden and output layers.

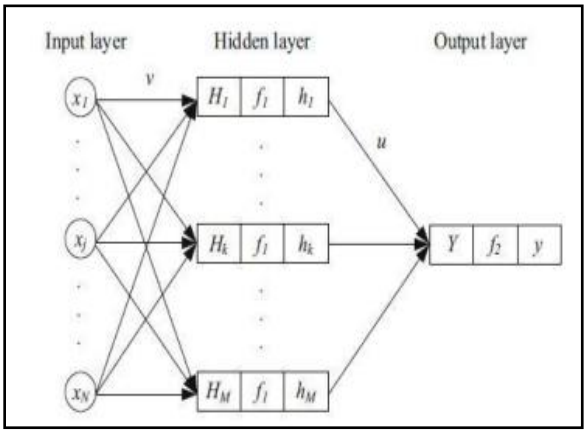


**Fig 3.1 Three-Layer Artificial Neural Network Model**

The aim of BPNN is to constantly modify the weights of connections between contiguous layers based on the deviation between actual values and outputs until the accuracy of the model meets the requirement of forecasting. The BPNN model can be used to forecast with new data when the weights are determined after numerous modifications. The ANN-Marquardt algorithm [18] combined with Newtonian gradient descent algorithm was used to adjust the connection weights and biases to minimize the error. The number of neurons in the hidden layer was usually determined by trial and error.

### *G. Support Vector Machine (SVM)*

The current study investigated the applicability of support vector machines (SVMs) in determining the relative importance of climate factors (mean temperature, rainfall, relative humidity, sunshine hours, daily temperature range, and rainy days) to yield variation of paddy rice in India.[4] Support vector machine (SVM) which was

originally developed by Vapnik (1998) has been widely applied to many different fields, such as signal process and time series analysis. Based on the statistical learning theory and Structural risk minimization principle, SVM is less Vulnerable to over fitting problem and it uses a hypothesis space of linear functions in a higher dimensional feature space . Studies have demonstrated that SVMs are superior to traditional artificial neural networks in solving classification and regression problems due to their good generalization ability.

### *H. Decision Trees (DT)*

A decision tree represents a structure with two types of components:
Leaf nodes that assign class labels to observations
Internal nodes that specify tests on individual attributes with one branch and sub tree for each outcome of the test. The tree classifies observations in a top-down manner, starting from the root and working one's way down according to the outcomes of the tests at the internal nodes, until a leaf node has been reached and a class label has been assigned. The tree is then constructed by means of recursive partitioning until the current leaf nodes contain only instances of a single class or until no test offers any improvement. However, since most course review data sets are noisy, and since in most cases the attributes have limited predictive power, this tree growing strategy often results in a complex tree with many internal nodes that over-fits the course review data.

### *I. Random forest (RF)*

The RF models were trained to predict course review yield using multiple biophysical variables as predictors. Environmental variables included climate, soil, photoperiod, water, and fertilization data. The same data were used for training MLR models for benchmarking purposes. The RF algorithm intrinsically set aside partial data for its own internal validation, called out-of-bag (OOB) data. However, to ensure a fair and conservative comparisons between RF and MLR, we used only a random half of each dataset (i.e., wheat, maize grain, potato, silage maize) for training (‘training dataset’) both RF and MLR models. The other half that was not used for training was then used as the ‘test dataset’ to validate and compare performances between the RF and MLR models. This process ensured that identical data points were available for training and independent testing using the data points not included in training.

## IV. FEATURE CLASSIFICATION MODEL

Feature selection plays an important role in data mining analytical model. It computes an optimal subset of predictive features measured in the original data. It enables to achieve maximum classification performance by reducing the number of features used in classification while maintaining acceptable classification accuracy. Subsets of the original features which retain adequate information to discriminate well among classes are selected. Several search algorithms have been used for feature selection. This work implements Principal Component Analysis, Information gain and Relief-f Attribute Evaluator.

### *A. Principal Component Analysis (PCA)*

Principal component analysis performs a linear mapping of the course review data to a lower dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. PCA, a non-parametric method builds a set of features by selecting those axes which maximize course review data variance.

PCA can be used to reduce a complex course review data set to a lower dimensionality, to reveal the structures or the dominant types of variations in both the observations and the variables. It is a quantitatively rigorous method that generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information. The principal components as a whole form an orthogonal basis for the space of the dataset.

### *B. Information Gain*

The information gain of an attribute tells the amount of information an attribute provides with respect to the classification target. Information gain (IG) measures the amount of course review information about the class prediction, if the only information available is the presence of a feature and the corresponding class distribution. In machine learning, information gain can be used to help ranking the features. A feature with high information gain should be ranked higher than other features because it has stronger power in classifying the course review data. Shannon entropy is the common measure for the information. Information gain is the reduction in the entropy that is archived by learning a variable. Concretely, it measures the expected reduction in entropy.

$IG= H(Y)-H(Y/X)$
$H(Y)= -\sum P(Y)LOG(P(Y))$
$H(Y/X)= -\sum P(X)\sum P(Y/X)LOG(P(Y/X))$

Where $P(Y)$ is the marginal probability density function for the random variable Y and $P(Y|X)$ is the conditional probability of Y given X.

### *C. Relief-F Attribute Evaluator*

A key idea of the original Relief algorithm (Kira & Rendell, 1992b), is to estimate the quality of attributes according to how well their values distinguish between instances that are near to each other. Given a randomly selected instance Ri, Relief searches for its

two nearest neighbors, one from the same class, called nearest hit H, and the other from the different class, called nearest miss M . It updates the quality estimation for all attributes. The ReliefF (Relief-F) algorithm Kononenko, 1994) is not limited to two class problems, is more robust and can deal with incomplete and noisy data. Similarly to Relief, ReliefF randomly selects an instance Ri, but then searches for k of its nearest neighbors from the same class, called nearest hits Hj, and also k nearest neighbors from each of the different classes, called nearest misses Mj (C). It updates the quality estimation WA for all attributes A depending on their values for Ri, hits H.

The selected features from each technique are used for training three classifiers SVM, Naive Bayes and Decision trees. The machine learning algorithms are tested using 10-fold cross validation to obtain better classification result.

## V. CONCLUSION

An inductive review content analysis on samples taken from 2000 posts was implemented and the categories of teachers' reflective thinking were obtained. Based on inductive course review content analysis results, a single-label course review classification algorithm is implemented to classify sample data. Then, the trained classification model on a large-scale and unexplored online discussion course review data set is applied and two visualizations types of results were provided. It's capable of distinctive co-regulated classification of course review document whose average expression is strongly related to the sample classes. The known document classified could contribute to revealing underlying category structures, providing a useful gizmo for the explorative analysis of biological information. In future, besides the fast algorithm enforced in information set here, additional planning should be made to implement also in numerical information set. Additional plan should be made to compare with algorithm like naïve bayes, K-Nearest neighbor, and SVM so on

## REFERENCES


[1] X. Chen, M. Vorvoreanu, and K. Madhavan, "Mining social media data for understanding students' learning experiences," IEEE Trans. Learning Technol., vol. 7, no. 3, pp. 246–259, Jul.-Sep. 2014.

[2] M. Rost, L. Barkhuus, H. Cramer, and B. Brown, "Representation and communication: Challenges in interpreting large social media datasets," in Proc. Conf. Comput. Supported Cooperative Work, 2013, pp. 357–362.

[3] S. Cherrington and J. Loveridge, "Using video to promote early childhood teachers' thinking and reflection," Teaching Teacher Educ., vol. 41, pp. 42–51, 2014.

[4] L. Rourke and T. Anderson, "Validity in quantitative content ana-lysis," Educ. Technol. Res. Development, vol. 52, no. 1, pp. 5–18, 2004.

[5] W. L. Chen, and C. K. Looi, "Incorporating online discussion in face to face classroom learning: A new blended learning approach," Australasian J. Educ. Technol., vol. 23, no. 3, pp. 308– 327, 2007

[6] L. van den Bergh, A. Ros, and D. Beijaard, "Teacher learning in the concourse review of a continuing professional development program: A case study," Teaching Teacher Educ., vol. 47, pp. 142–150, 2015.

[7] C. Tsai, "Understanding social nature of an online community of practice for learning to teach," Educ. Technol. Soc., vol. 15, no. 2,271–285, 2012.

[8] Dawson, D. Gasevic, G. Siemens, and S. Joksimovic, "Current state and future trends: A citation network analysis of the learning analytics field," in Proc. 4th Int. Conf. Learning Analytics Knowl., Mar. 2014, pp. 231–240.

[9] M. Liu, R. A. Calvo, A. Pardo, and A. Martin, "Measuring and visualizing students' behavioral engagement in writing activities," IEEE Trans. Learning Technol., vol. 8, no. 2, pp. 215–224, Apr.-Jun. 2015

[10] M. Van Manen, "Linking ways of knowing with ways of being practical," Curriculum Inquiry, vol. 6, no. 3, pp. 205–228, 1977.