Centroid-based Clustering for Student Models in Computer-based Multiple Language Tutoring

Maria Virvou, Efthymios Alepis and Christos Troussas
Department of Informatics, University of Piraeus, 80, Karaoli and Dimitriou str., Piraeus, Greece

Keywords: User Modelling, User Clustering, Multiple Language Learning, Intelligent Tutoring Systems, K-means Algorithm.

Abstract: This paper proposes an approach for the initialization and the construction of student models in an intelligent tutoring system that teaches multiple foreign languages. The basic concept for the construction of the initial user models is to assign each new student to a model with similar characteristics. As it is quite easy to understand that a tutoring system has rather little information about its new users, our effort is to provide as much information as possible for each specific user relying on the user's initial data. To this end, a machine learning algorithm, namely k-means, is responsible for creating clusters relying on the system's pre-entered past data and as a next step, each new entry is assigned to the nearest centroid.

1 INTRODUCTION

In the past few years, there has been an increasing focus on the use of Internet, which allows greater flexibility in all aspects of modern life, especially with the spread of unmetered high-speed connections. Educational material at all levels is available from Internet. Regarding e-learning, it has never been easier for people to access educational information at any level from any place. The low cost and nearly instantaneous sharing of ideas, knowledge and skills has rendered the distant learning process feasible for people with less spare time. By these means, not only can a group cheaply communicate and share ideas but the wide reach of the Internet allows such groups to form in an easy and efficient way. Hence, the development of web-based applications has become common place. Moreover, all the emerging needs of modern life accentuate the importance of learning foreign languages (Virvou and Troussas, 2011). Taking into account the scientific area of Intelligent Tutoring Systems (ITSs), there is an increasing interest in the use of computer-assisted foreign language instruction. Especially, when these systems offer the possibility of multiple language learning at the same time, the students may further benefit from this educational process (Virvou et al., 2000).

An issue of great importance in e-learning is the personalization of users, since it is quite difficult to monitor users' learning patterns (Licchelli et al., 2004). Specifically, it is performed through student modeling, which consists of the analysis of students' behavior and prediction of their future behavior and learning performance. A solution to this problem is the exploitation of automatic tools for the generation and discovery of user profiles in order to obtain an effective student model based on his/her learning performance and preferences, that in turn allows to create a personalized education environment. Adaptive personalized e-learning systems could accelerate the educational process by revealing the strengths and weaknesses of each student. Most student models are concerned with representing the student's ability on portions of the domain (Beck and Woolf, 2000). However, the way of mapping the low-level knowledge to higher level teaching actions is not always obvious.

In view of the above, in this paper we propose a machine learning architecture which permits the initialization of students' models. Our framework uses an innovative combination of stereotypes and the k-means clustering algorithm in order to partition multiple observations into a number of k clusters in which each observation belongs to the cluster with the nearest mean. Each cluster is represented by a single mean vector. In particular, a student is first assigned to a stereotype category on the basis of his/her background knowledge level in the instruction of multiple foreign languages. This is

Virvou M., Alepis E. and Troussas C..
Centroid-based Clustering for Student Models in Computer-based Multiple Language Tutoring.
DOI: 10.5220/0004128201980203
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications and Wireless Information Networks and Systems* (SIGMAP-2012), pages 198-203
ISBN: 978-989-8565-25-9

conducted based on the students' performance on a preliminary test posed to the student at the first time of his/her interaction with the system. Then, the k-means algorithm takes as input multiple students' characteristics, which are described below and serves as means for the initialization of the new-student-model based on recognized similarities between the new student and past students who belong to the same stereotype category.

This paper is organized as follows. First, we present the related scientific work. In sections 3 and 4, we discuss our system's architecture, namely the machine learning in student modelling and the k-means clustering algorithm. Finally, in section 5, we come up with a discussion about the usability of centroid-based clustering for user models and we present our next plans.

2 RELATED WORK

Teaching languages through computer-assisted approaches is a quite significant field in language learning. User modeling has already been applied in a wide variety of scientific areas, including educational software for language instruction. Machine learning techniques have been applied to user modeling problems for acquiring models of users. In this section, we try to imprint the speckle of the scientific progress of student modeling concerning Machine Learning and CALL (Computer Assisted Language Learning).

Basile et al (2011) proposed the exploitation of machine learning techniques to improve and adapt the set of user model stereotypes by making use of user log interactions with the system. To do this, a clustering technique is exploited to create a set of user models prototypes; then, an induction module is run on these aggregated classes in order to improve a set of rules aimed as classifying new and unseen users. Their approach exploited the knowledge extracted by the analysis of log interaction data without requiring an explicit feedback from the user. Nino (2009) presented a snapshot of what has been investigated in terms of the relationship between machine translation (MT) and foreign language (FL) teaching and learning. Moreover, the author outlined some of the implications of the use of MT and of free online MT for FL learning. Friaz-Martinez et al (2007) investigated which human factors are responsible for the behavior and the stereotypes of digital libraries users so that these human factors can be justified to be considered for personalization. To achieve this aim, the authors have studied if there is a statistical significance between the stereotypes created by robust clustering and each human factor, including cognitive styles, levels of expertise and gender differences. Virvou and Chrysafiadi (2006) described a web-based educational application for individualized instruction on the domain of programming and algorithms. Their system incorporates a user model, which relies on stereotypes, the determination of which is based on the knowledge level of the learner. Liccheli et al (2004) focused on machine learning approaches for inducing student profiles, based on Inductive Logic Programming and on methods using numeric algorithms, to be exploited in this environment. Moreover, an experimental session has been carried out from the authors, comparing the effectiveness of these methods along with an evaluation of their efficiency in order to decide how to best exploit them in the induction of student profiles. Tsiriga and Virvou (2004) introduced the ISM framework for the initialization of the student model in Web-based ITSs, which is a methodology that uses an innovative combination of stereotypes and the distance weighted k-nearest neighbor algorithm to set initial values for all aspects of the student model.

SignMT was implemented by Ditcharoen et al (2010) to translate sentences/phrases from different sources in four steps, which are word transformation, word constraint, word addiction and word ordering. Finally, Virvou and Troussas (2011) described a ubiquitous e-learning tutoring system for multiple language learning, called CAMELL (Computer-Assisted Multilingual E-Language Learning). It is a post-desktop model of human-computer interaction in which students "naturally" interact with the system in order to get used to electronically supported learning. Their system presents advances in user modeling, error proneness and user interface design.

However, after a thorough investigation in the related scientific literature, we came up with the result that there was no implementation of multilingual educational systems that combine student modeling and machine learning. Hence, we implemented a prototype system, which incorporates intelligence in its diagnostic component, offers proneness to students' errors provides error diagnosis and advice based on students' needs.

3 MACHINE LEARNING IN USER MODELING

Student modeling can undoubtedly benefit from

machine learning, given that machine learning consists of the induction of knowledge, normally leading to improvements in classifying objects in a specific domain. Thus, our system's student model can extend or compile its background knowledge, namely its bug library, so that the resulting student model could be more accurate and efficient. User modelers that deal with simple students' behaviors have the ability to collect a set of behaviors from which to induce a student model. The task of constructing the student model from a multiple behaviors set can be regarded as an inductive learning task and therefore machine learning techniques can be used to address this task. Constructing student models in multiple language learning environments is quite complex, since student behaviors are likely to be inconsistent and incomplete, which can be due to any of the following reasons (Virvou and Troussas, 2011):

I. Accidental sips.

II. Quick elimination of old knowledge errors.

III. Recurrence of old knowledge errors.

IV. Sudden appearance of new knowledge errors.

Except for accidental slips, all the above error categories that our student model can predict may change over time in unforeseen ways. This causes problems in the generation of student models because the predictions become less accurate. The degree of precise prediction conducted by the user models can be ameliorated by the use of unsupervised inductive learning techniques.

For the incorporation of the algorithm into the resulting multilingual system (Fig. 3) we may observe the following basic steps:

i. For the initialization of the system, the k-means algorithm receives as input, pre-stored data or data from empirical studies. It uses several fundamental characteristics which tend to influence the educational procedure:

a.the age of students,

b. their level of knowledge in one of the foreign language taught,

c.the degree of carefulness when answering questions and

d. the error proneness of the student in each concept of the domain knowledge.

These characteristics have been found quite significant in past language learning applications (Tsiriga and Virvou, 2004).

ii. Machine learning techniques are used as a next step in order to describe efficiently the cognitive processes that underlie the student's actions along with the student's behavioral patterns and preferences.

iii. Based on the aforementioned characteristics, the system creates clusters of the already existing students. These clusters contain valuable information about their members, considering their behavior, their preferences and generally their interaction with the system.

Our system uses this model to support students while studying the theory and solving exercises. In particular, based on the information that emanates from the knowledge level of the student in each concept of the domain knowledge, the system provides personalized help and support when s/he navigates through the curricula. The error proneness of the student supported by the student modeler is used for error diagnosis. In particular, this information is used in cases where the system has to disambiguate between competing hypotheses that concern the cause of students' mistakes (Tsiriga and Virvou, 2004).

4 K-MEANS CLUSTERING ALGORITHM

K-means clustering is a well known machine algorithm that is widely used to classify or to group objects based on attributes/features into a number of k groups/sets. "K" is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus the purpose of K-mean clustering is to classify the data. Each object represented by one attribute point is an example to the algorithm and it is assigned automatically to one of the cluster. This consists of unsupervised learning as the algorithm classifies the object automatically only based on the criteria of minimum distance to the centroid. The learning process depends on the training examples with witch the algorithm is fed. There are two choices in this learning process:

i. Infinite training. Each data that feed to the algorithm will automatically consider as the training examples.

ii. Finite training. After the training is considered as finished, the algorithm is started to work by classifying the cluster of new points. This is conducted simply by assigning the point to the nearest centroid without recalculate the new centroid. Thus after the training finished, the centroid are fixed points.

The basic steps of k-means clustering are simple. In

the beginning we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects in sequence can also serve as the initial centroids.

Then the K means algorithm will do the three steps below until convergence is reached, namely there is no object move in groups, as illustrated in Figure 1:

i. Determine the centroid coordinate randomly from the data set.

ii. Determine the distance of each object to the centroids and creation of k clusters.

iii. Group the object based on minimum distance and determine the new means as the centroid of each one of the k clusters.

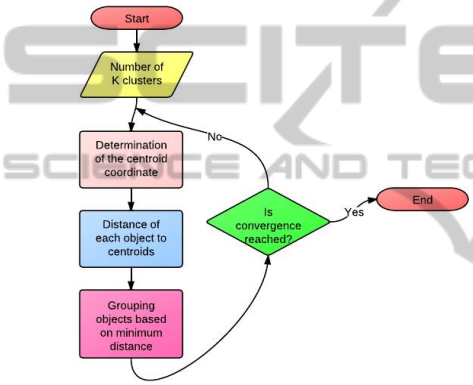

Figure 1: Steps of K-means algorithm.

More specifically, the above steps can be summarized as follows (Teknomo, 2006):

i. Step 1. Begin with a decision on the value of k as the number of clusters.

ii. Step 2. Put any initial partition that classifies the data into k clusters. Assign of the training samples randomly, or systematically as the following:

a. Take the first k training sample as single-element clusters.

b. Assign each of the remaining (N-k) training sample to the cluster with the nearest centroid. After each assignment, recomputed the centroid of the gaining cluster.

iii. Step 3. Take each sample in sequence and compute its distance from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.

iv. Step 4. Repeat step 3 until convergence is achieved, namely until a pass through the training sample causes no new assignments.

The key idea of k means is simple and is described as follows: In the initialization phase, the number of clusters k is determined. Then the algorithm assumes the centroids or centers of these k clusters. These centroids can be randomly selected or designed deliberately. If the number of data is less than the number of clusters, then each data is assigned as the centroid of the cluster. Each centroid will have a cluster number. If the number of data is greater than the number of clusters, the algorithm computes the Euclidean distance between each object and all centroids to get the minimum distance. This data is belongs to the cluster that has minimum distance from itself. Given that the location of the real centroid is unknown during the process, the algorithm needs to revise the centroid location with regard to the updated information (i.e., minimum distance between new objects and the centroids). After updating the values of the centroids, all the objects are reallocated to the k clusters. The process is repeated until the assignment of objects to clusters ceases to change much, or when the centroids move by negligible distances in successive iterations. Mathematically the iteration can be proved to be convergent.

Since the location of the centroid cannot be fixed or prearranged, the centroid location is adjusted, based on the current updated data. Then all the data is assigned to this new centroid. This process is repeated until no data is moving to another cluster anymore. Mathematically, this loop can be proved to be convergent. The convergence will always occur if the following condition satisfied:

i. Each switch in step 2 the sum of distances from each training sample to that training sample's group centroid is decreased.

ii. There are only finitely many partitions of the training examples into k clusters.

In order to better clarify the clustering algorithmic process, we are providing the pseudo-code of k-means algorithm:

```
   Input: A dataset D, a user specified
number k
   Output: k clusters
   Initialize cluster centroids (randomly);
   While not convergent
    For each object o in D do
        Find the cluster c whose centroid is
most close to o;
        Allocate o to c;
    End
    For each cluster c do
        Recalculate the centroid of c based
on the objects allocated to c;
    End
   End
```

The two key features of k-means which make it efficient are often regarded as its biggest drawbacks:

i. Euclidean distance is used as a metric and variance is used as a measure of cluster scatter.

ii. The number of clusters k is an input parameter: an inappropriate choice of k may yield poor results. That is why, when performing k-means, it is important to run diagnostic checks for determining the number of clusters in the data set. The correct choice of k is often ambiguous, with interpretations depending on the shape and scale of the distribution of points in a data set and the desired clustering resolution of the user. In addition, increasing k without penalty will always reduce the amount of error in the resulting clustering, to the extreme case of zero error if each data point is considered its own cluster (i.e., when k equals the number of data points, n). Intuitively then, the optimal choice of k will strike a balance between maximum compression of the data using a single cluster, and maximum accuracy by assigning each data point to its own cluster. If an appropriate value of k is not apparent from prior knowledge of the properties of the data set, it must be chosen somehow. There are several categories of methods for making this decision. One simple principle, that we incorporated in the implementation of k-means algorithm, sets the number to (Mardia et al., 1979):

$$K \approx \sqrt{n/2} \quad (1)$$

with n as the number of objects (data points). In our case, given that the data points, which are an outcome from empirical studies, are 32 we come up with the conclusion that k=4. Figure 2 illustrates a snapshot of our system and specifically a report of k-means, the initial user data, the resulting k-mean vectors, the number and members of means.

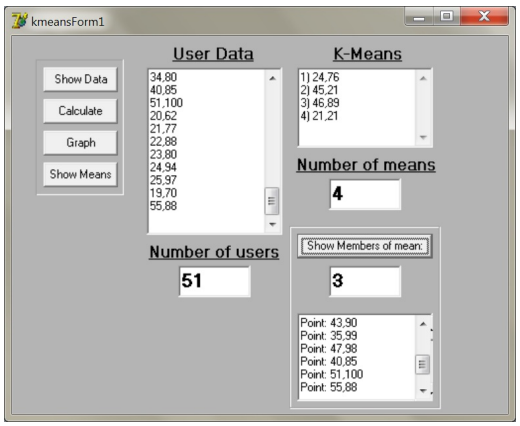

Figure 2: Snapshot of k-means algorithm.

5 CONCLUSIONS

In this paper we have presented our approach for improving student models in the initialization phase of an educational system. We have already our own implementation of the k-means machine learning algorithmic approach, as well as a tutoring system for multiple language learning. After processing user personal data we come up with more sophisticated user models containing stereotypic information that is based on similarities with other user groups seen as clusters. We believe that this approach will produce good results since it uses well known techniques, already implemented in other similar scientific areas with quite promising reports. As a next step, it is in our near future plans to give our resulting system to real students to use it supplementarily for their language learning courses in order to evaluate it and test its usefulness as an educational tool.

REFERENCES

Niño, A., 2009. Machine translation in foreign language learning: Language learners and tutors perceptions of its advantages and disadvantages. In *ReCALL*. Vol. 21, pp. 241-258.

Friaz-Martinez, E., Chen, S. Y., Macredie, R. D., Liu, X., 2007. The role of human factors in stereotyping behavior and perception of digital library users: a robust clustering approach. In *User Modelling and User-Adapted Interaction*. Vol. 13, pp. 305-337.

Webb, G. I., Pazzani, M. J., Billsus, D., 2001. Machine Learning for User Modeling. In *User Modelling and User-Adapted Interaction*. Vol. 11, pp. 19-29.

Virvou, M., Troussas, C., 2011. CAMELL: Towards a ubiquitous multilingual e-learning system. In *CSEDU 2011 - Proceedings of the 3rd International Conference on Computer Supported Education*. Vol. 2, pp. 509-513.

Virvou, M., Troussas, C., 2011. Web-based student modeling for learning multiple languages. In *International Conference on Information Society, i-Society 2011*. Article number 5978484, pp. 423-428, 2011.

Virvou, M., D. Maras, D., Tsiriga, V., 2000. Student modelling in an intelligent tutoring system for the passive voice of english language. In *Educational Technology and Society*.

Virvou, M., Chrysafiadi, K., 2006. A web-based educational application for teaching of programming: Student modeling via stereotypes. In *Proceedings - Sixth International Conference on Advanced Learning Technologies, ICALT 2006.* Vol. 2006, pp. 117-119.

Ditcharoen, N., Naruedomkul, K., Cercone, N., 2010. SignMT: An alternative language learning tool. In

Computers and Education, Vol. 55, pp. 118-130.
Licchelli, O., Basile, T. M. A., Di Mauro, N., Esposito, F., Semeraro, G., Ferilli. S., 2004. Machine Learning Approaches for inducing Student Models. In *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*. Vol. 3029, pp. 935-944.
Basile, T., Esposito, F., Ferilli, S. 2011. Improving User Stereotypes through Machine Learning. In *Communications in Computer and Information Science*. Vol. 249, pp. 38-48.
Tsiriga, V., Virvou, M., 2004. A framework for the initialization of student models in web-based intelligent tutoring systems. In *User Modelling and User-Adapted Interaction*. Vol. 14, pp. 289-315.
Mardia, K. et al.,1979. Multivariate Analysis, In *Academic Press*.
Teknomo, K., 2006. K-Means Clustering Tutorials.