Toward a Framework for Learner Segmentation

BAHAREH AZARNOUSH
School of Computing, Informatics, and Decision Systems Engineering,
Arizona State University,
Tempe, AZ USA
bazarnou@asu.edu.

JENNIFER M. BEKKI
Department of Engineering & Computing Systems,
Arizona State University
Mesa, AZ USA
jennifer.bekki@asu.edu.

GEORGE C. RUNGER
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University, AZ USA
runger@asu.edu

BIANCA L. BERNSTEIN
Counseling & Counseling Psychology
Arizona State University, AZ USA
bbernstein@asu.edu.

ROBERT K. ATKINSON
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University, AZ USA
robert.atkinson@asu.edu.

Effectively grouping learners in an online environment is a highly useful task. However, datasets used in this task often have large numbers of attributes of disparate types and different scales, which traditional clustering approaches cannot handle effectively. Here, the use of a dissimilarity measure based on the random forest, which handles the stated drawbacks of more traditional clustering approaches, is presented for this task. Additionally, the application of a rule-based method is proposed for interpreting the resulting learner segmentations. The approach was implemented on a real dataset of users of the *Career*WISE online educational environment, designed to provide resilience training for women STEM doctoral students, and was shown to find stable and meaningful groups of users.

This research was supported by NSF.

1. INTRODUCTION

The educational landscape of the future includes an array of online experiences for learners, ranging from those coupled to formal course work (e.g., course management systems like Blackboard or Moodle) to more informal learning resources (e.g., MOOCS or the Khan Academy), which provide free, high quality instructional materials for students to use in augmenting their other educational experiences. The online format provides a wide variety of mechanisms both by which instructional content can be presented (e.g., video, text, images, etc.) and through which learners can interact according to their particular learning styles and objectives [Narciss et al. 2007]. It also makes the educational resources scalable and helps to overcome geographic, temporal, and

financial boundaries, representing and reflecting the importance of the successful use of technology for instructing the current and future generations [Hines et al. 2009].

Accompanying the online environments themselves, the data generated by these resources can include learner profile data as well as dates and times for learner downloads of information, student postings to discussion boards, learner activity /paths through an online learning environment, measures of student learning, etc., making available an incredibly rich, learner centered set of data that can be analyzed and interpreted in an effort to improve educational experiences and access for individuals. Recent years have seen the rise of intellectual communities that are working to develop approaches for taking advantage of the unique types of data that are generated in educational settings with goals of improving student learning, optimizing learner experiences, and increasing the understanding of educational phenomena [Ferguson 2012; Johnson *et al*. 2012; Romero and Ventura 2010].

One such approach is to use the data available to segment learners into meaningful groups so that decisions/interventions can be made accordingly based on group membership. The notion of grouping learners is not new; in fact, there have been many applications of such analyses in the educational sciences. In early work, for example, Shavelson [1979] demonstrated the applicability of cluster analysis to educational research in higher education, teaching, curriculum and cognitive structure. Meece and Holt [1993] clustered students based on their mastery, ego, and work-avoidant goal orientations, and, later, Talavera and Gaudioso [2004] used cluster analysis to discover patterns reflecting user behavior in unstructured collaboration spaces. Merceron and Yacef [2004] suggested that cluster analysis could be used to identify types of students so that different remedial instructional support methods could be applied accordingly. More recently, Valle and Duffy [2009] used cluster analysis to discover learner characteristics and approaches to managing learning, while Perera, et al. [2009] applied clustering to assess if data attributes reflected characteristics of a predefined grouping and Hershkovitz and Nachmias [2010] used decision trees for grouping students in Web-supported courses.

However, there are challenges with clustering educational data that have not yet been adequately addressed by previous work. For example, the attributes used to form the clusters are likely to be of different types. Categorical attributes such as demographic indicators, for example, may be coupled with numerical attributes related to learning measures. Furthermore, it is not uncommon for educational data to contain variables that are recorded in different scales. Although transformation can be applied in such cases (e.g., Valle and Duffy [2009]), such transformations may actually collapse true structure in the data [Hastie et al. 2009], motivating the need for a better approach. Additionally complicating the clustering of educational data is the fact that most

standard clustering techniques are inadequate in situations where there are a large number of potential attributes on which clustering is to be performed. With the large quantity of data available to describe online learners, the attributes available for grouping learners can be particularly large in number. Previous work handles this situation through an ad hoc search for relevant attributes prior to performing the cluster analysis. Talavera and Gaudioso [2004], for example, remove attributes that show uniform behavior across all users, while other work simply identifies and then clusters based on a small number of defined variables judged as relevant [Valle and Duffy 2009; Perera, et al. 2009]. These approaches, however, lack general applicability.

The first contribution of this paper is advancing educational cluster analysis by proposing and demonstrating an approach that can better accommodate the complex datasets of high dimension and size, with disparate attribute types of different scales that often arise in data generated from online learning environments. The approach utilizes the Random Forest dissimilarity measure in forming the clusters and does not require transformation of the original attributes. The second contribution is to suggest an improved approach for tackling the critical step of describing and interpreting the generated clusters; without this step, relevant actions based on cluster membership cannot be taken. Toward this end, Duran and Odell [1974] introduced the idea of cluster description. Later, Diday et al. [1976] presented two algorithms that use a point in the feature space of the formed clusters as the representation of the clusters. Michalski et al. [1981] used conjunctive statements derived from classification trees for describing characteristics of the rows in each cluster, while Liu et al. [2000] used tree-based methods and Zenko et al. [2006] used rule-based methods to generate and describe clusters.

Also, specifically within the field of education, there have been several approaches suggested for describing clusters. Talavera and Gaudioso [2004], for example, use a measure of the attributes' degree of discrimination between the clusters and the probability of each attribute value in a given cluster, while Merceron and Yacef [2004] use graphical visualization. Valle and Duffy [2009] as well as Meece and Holt [1993] use statistical tests, means and standard deviations to derive meaning from the clusters, and Hershkovitz and Nachmias [2010] characterized each cluster by the hierarchical conditions derived from a decision tree. While all these approaches are able to provide some insight into the characteristics of the members of clusters, a more descriptive and exhaustive approach is proposed. Specifically, the adoption of an exhaustive rule-based method to mine cluster describing rules is proposed. These rules clearly describe subgroups within each cluster and their exhaustiveness prevents interesting rules from being left undiscovered.

The remainder of the paper is organized as follows. The next section provides a brief background on the statistical elements that are utilized in the proposed framework. A discussion of the elements of the framework, including mechanisms for validating the results, is then described in Section 3. Sections 4 and 5 first provide a case study demonstrating the implementation of the methodology using a dataset generated from users of the *Career*WISE (http://careerwise.asu.edu) online learning environment, which provides psychological education in areas linked to persistence in women doctoral students in STEM fields, and then provide additional experimentation results comparing the proposed clustering approach with more traditional approaches. Finally, Sections 6 concludes with a brief discussion and plans for future work.

2. BACKGROUND

For the remainder of the paper, datasets are presented as N rows (observations) of M dimensional vectors. Each dimension represents an attribute (variable or a feature). We denote x_{ij} as the value of the j^{th} attribute on the i^{th} row. For example, each row could represent a user and each attribute an item from a demographic survey.

We start with a short description of clustering, which segments rows of a dataset into groups such that rows within each group are similar. Clustering methods may be classified into two main categories: hierarchical and partitional [Tan et al. 2006]. A partitional clustering is a division of the rows into non-overlapping clusters such that each row is assigned to one cluster. *K*-means and *K*-medoids are two popular partitional clustering methods. Given the order of the model, both these methods generate clusters by finding a representative object for each cluster and then assign the rows to the cluster with the closest representative [Singh and Chauhan, 2011]. In contrast to partitional approaches, hierarchical approaches form clusters at each level of the hierarchy by merging clusters at the next lower level. Within hierarchical clustering, there are two basic strategies for generating clusters: agglomerative and divisive. Agglomerative strategies start with each point as a cluster and at each stage merge the closest clusters. Divisive strategies, in contrast, start with a single cluster and split one of the existing clusters at each stage. Additionally, within the agglomerative strategy, there are different approaches that depend on the definition of proximity between clusters. Some examples are the single linkage, complete linkage, group average, and Ward's method. Kaufman and Rousseeuw [1990], Hastie et al. [2009] and Tan et al. [2006] provided descriptions of popular clustering methods, and Jain, et al. [1999] provided a survey of clustering methods.

Ultimately the choice of the clustering method is driven by the dataset and the objectives of the cluster analysis. The complexity of the algorithms becomes an important factor in the case of a dataset of large size. A

general hierarchical clustering algorithm using a dissimilarity-update formula forms clusters from the two closest clusters, which is of complexity $O(N^2)$ and redefines the dissimilarities between the new clusters, which is carried out in $O(N)$ time [Murtagh,1983]. For larger datasets, other methods such as the *K*-means may be used. While more efficient, this approach is appropriate when squared Euclidean distance is taken to be the dissimilarity measure. The Euclidean distance requires all attributes to be of the numeric type (or be transformed to numeric). Furthermore, this procedure lacks robustness against outliers. The *K*-medoids approach generalizes the *K*-means for use of arbitrarily defined dissimilarity measures. It is applicable when the data is described only through the dissimilarity matrix and is more robust than *K*-means in the presence of outliers [Hastie et al., 2009]. Partitioning Around Medoids (PAM) and Clustering LARge Applications (CLARA) are two popular *K*-medoids algorithms [Kaufman and Rousseeuw, 1990]. To start, PAM begins with an arbitrary selection of *k* objects. Then, at each iteration a swap between a selected object and a non-selected object is made, as long as such a swap would result in an improvement of the clustering. The complexity of each iteration of PAM, used in this paper, is $O(K\ (N-K)^2)$, where *N* is the number of rows. CLARA reduces the complexity to $O(K^3+NK)$ by drawing samples of the data set and applying PAM on the samples [Ng and Han,1994].

In addition to determining which clustering method to employ (e.g., partitional vs. hierarchical), the application of clustering also requires a decision on which dissimilarity measure to use for generating the clusters. Euclidean-based distances are commonly used for this purpose. However, such dissimilarity measures are challenged in the high-dimensional datasets used to describe learners because of the sparsity of the datasets, which causes distances to become less meaningful. In fact, in very high-dimensional spaces, points become nearly equidistant from each other, rendering traditional similarity measures to be useless [Parson et al. 2004]. This phenomenon is referred to as the curse of dimensionality [Bellman 1966]. Consequently, in the context of learner segmentation, in which the number of attributes on which clusters could be formed might be very large, an approach for developing clusters that is robust to the curse of dimensionality is preferred. This paper adopts a tree-based method for this purpose.

A decision tree [Breiman, Friedman, Olshen and Stone, 1984] is a popular tree-based method. Decision trees have many attractive properties such as the ability to capture nonlinear relationships, handle missing values, be invariant to attribute units, and be robust to outliers. However, one of their major drawbacks is their instability and high variance. Averaging results across a number of decision trees alleviates this problem, and one such approach is the random forest.

A random forest (RF) is an ensemble of a large collection of decision trees [Breiman, 2001]. Each of the individual decision trees is constructed using a randomly selected sample of rows (with replacement) of the same size as the original dataset, *N*. A subset of *m* candidate attributes from the set of *M* input attributes is selected for consideration at each node. We note that $m<M$ and that $m \approx \sqrt{M}$ are the default values for classification. The candidate attributes are each evaluated, and the attribute that results in the most homogenous nodes is selected. At each tree, the process continues until a specified minimum node size, n_{min}, is reached. Each tree predicts a class, and the final classification of a particular row of data is based on the majority vote from all of the trees.

In addition to the attractive properties discussed for decision trees, RFs offer additional benefits. They have high accuracy and provide estimates of both variable importance and generalization error. The variable importance measure is useful for determining which attributes from a potentially large set are most significant in the classification, and the generalization error estimate gives a measure of prediction capability on independent test data. The complexity of the random forest is $O(ntree\sqrt{M}\,N\log(N))$, where *ntree* is the number of trees in the forest, *M* is the number of attributes, and *N* is the number of rows [Breiman, 2002]. Furthermore, RFs have the ability to compute pair-wise proximities between the rows in the dataset. The method implemented in this paper exploits these proximities for clustering learners and is explained further in Section 3.

3. CLUSTER GENERATION, VALIDATION & DESCRIPTION

This section provides a discussion of the elements in the proposed framework, highlighting important considerations for cluster generation, validation, and description. Additionally, algorithms for the calculation of the RF-based dissimilarity measure and for calculating a measure of cluster stability for use in validation are provided.

3.1 Calculating the RF Dissimilarity Measure

To perform cluster analysis, a measure of dissimilarity (or similarity) between the data rows to be clustered must be used. RF's measure of proximity between rows of the dataset, which is a form of similarity between rows of the dataset and can be used for clustering, scaling and outlier detection [Breimen and Cutler 2003], is one such approach that is of particular interest in the segmentation of learners. This is because it can effectively handle the high-dimensional feature spaces that are difficult for Euclidean-distance based measures and can also handle disparate attribute types (i.e., categorical, ordinal, and numerical) of different scales. This measure has been

applied to many unsupervised learning problems including genomic sequence data [Allen et al. 2003], and tumor marker data [Shi et al. 2005; Seligson et al. 2005; Shi and Horvath, 2006].

The RF dissimilarity measure is developed through the transformation of the clustering problem, an unsupervised problem, to one of supervised learning. To do so, a class attribute, *y*, is defined by labeling the rows i=1: N in the original data as class "0" and creating a second, synthetic dataset of the same size labeled with y = 1 (representing class "1"). Each row in class 1 is created by randomly sampling from the product of the marginal distributions of the attributes in the original data. That is, the value of attribute j for any row in class 1 is created by randomly sampling from the N values $x_{i,j}$, i=1: N. This breaks down the dependencies between the attributes, as the dependence structure has not been used in creating the assigned members of class 1 [Breimen and Cutler, 2003].

Once the problem has been structured as a supervised learning problem (through the assignment of classes), the RF is constructed based on the defined y attribute. The proximity between rows x_i and $x_{i'}$, $RFprox\ (x_i,x_{i'})$, is the proportion of trees in which both rows appear in the same terminal node, and the RF dissimilarity between rows x_i and $x_{i'}$ is defined as $\sqrt{1-RFprox(x_i, x_{i'})}$.

Algorithm 1 provides the details of generating the RF dissimilarity measures (sometimes referred to as RF distance). It is important to note that the RF proximities can vary according to the realization of the synthetic data generated, so averaging the results across multiple RFs to calculate the final proximities (used towards calculating the RF dissimilarity) is recommended. Shi and Horvath [2006] provide recommendations on the number of forests and trees to construct. Moreover, the computation of the RF dissimilarity can be parallelized by performing both the tree and forest computations at the same time.

Algorithm 1

CalcRFDiss
input: D denoting the dataset, *nforest* denoting the number of forests, *ntree* the number of trees
output: *RFdis* denoting the RF dissimilarity matrix for rows in D

y←Rcomb(0_N, 1_N)	define class attribute y as a column of N 0s, and N 1s
for j←1 to *nforest* **do**	
D_{1j}←Syn (D)	generate synthetic data D_{1j}
D_j←RComb(D_{1j}, D)	append the rows of D_{1j} and D to form D_j
D_j←CComb(D_j, y)	append class attribute
F_j←BuildRF(D_j, *ntree*)	build random forest of *ntree* trees on D_j
$RFprox_j$←Prox (F_j, D):	calculate $RFprox_j$, the proximity matrix for rows in D using forest F_j
end for	
$RFprox$← Avj ($RFprox_j$)	average the proximities across the j forests
return $RFdis$←$\sqrt{1-RFprox}$	calculate the RF dissimilarity matrix denoted by *RFdis*

3.2. Cluster Generation and Validation

Model assessment of a learning method is an important step, as it provides insight into the quality of the model and guides the model selection [Hastie et al. 2009]. There are different methods for model assessment in supervised learning problems, as the class attribute dictates a measure of success and guides the assessment. In unsupervised problems, however, a measure of success is less clear, as there is no class attribute. Nonetheless, there remains a need for a quantitative evaluation of clustering solutions. The stability of the clustering solution is one such approach [Breckenridge, 2000; Tibshirani and Walther, 2005; Lange, et al, 2004] in which agreement between clustering solutions obtained from different data sets generated from the same source is measured.

Model selection in clustering consists of choosing the distance measure to be used as the input to a clustering algorithm as well as selecting the order of the model to determine how many clusters exist in the data. However, due to the absence of a formal objective in clustering, the first problem of selecting an appropriate distance measure and clustering algorithm is ill posed [Lange, et al, 2004]. Different distance measures and algorithms accentuate different facets of the data; therefore, at best, the distance measure and clustering algorithm deemed most appropriate for a given problem can be selected, but a universal "best" choice does not exist. Nevertheless, cluster stability has been shown to be a useful for this purpose [Breckenridge, 2000; Lange, et al, 2004]. A clustering solution that exhibits higher stability is likely to be more appropriate for the data under study. The notion of stability is also useful for choosing the order of the model; that is, a partitioning with an appropriate order should be reproducible using different datasets from the same source.

In order to establish the notion of stability of a clustering solution, rather than considering clustering to be partitioning of the rows of a single dataset, we consider it to be the more general task of partitioning the feature space that defines not only the rows of the dataset at hand, but possibly an infinite collection of other rows from the same source. The dataset at hand, then, merely represents a realization from this source provided to partition the feature space and to assign each row of the source to a different cluster. Unlike supervised learning problems where the class attribute dictates the partitions, the relationships of the rows (defined through the dissimilarity measure) dictate these partitions. This view allows for the clustering of a new dataset based on the clustering solution obtained from another dataset. The idea of stability, then, can be understood as finding a stable partitioning of the feature space that is reproducible using different realizations of data from the same source. We note that the stability of the clustering solution is important for the problem under study, as the

partitioning of users of an online learning environment can be used to classify future site users who are assumed to be generated from the same population. The steps required in order to assess the stability of a clustering solution are:

1. **Data splitting**: The dataset is split into test and training datasets, D_{tst} and D_{trn} , in order to imitate having access to different realizations of the data from the same source. It is recommended that the dataset is split into two disjoint and equally sized datasets, $N_{tst} = N_{trn} = N/2$. The equal size assures that the same structure is observable in both datasets, and the disjoint requirement avoids an overly optimistic estimate of stability caused by data overlap [Lange, et al, 2004].

2. Clustering: The test and training datasets, D_{tst} and D_{trn} , are clustered separately by a selected clustering algorithm denoted as clustering algorithm A, distance matrix *d*, and model order *K*. Each row in D_{tst} and D_{trn} is assigned a cluster membership, denoted by $clusterID_{0\text{-}tst}$, $clusterID_{0\text{-}trn.}$

3. Classifying: Each row in D_{tst} is assigned a second cluster membership, denoted by $clusterID_{1\text{-}tst}$, based on the feature space boundaries obtained from clustering D_{trn} in Step 2. A supervised learner may be used for this purpose. Toward this end, each row of D_{trn} is augmented with its assigned $clusterID_{0\text{-}trn.}$, obtained from Step 2, and is used as the class attribute in training the classifier. Then, the *n* nearest neighbor method may be used for the classification of the test rows to a training cluster [Breckenridge, 2000; Lange, et al, 2004]. To implement such an approach, a distance measure must be used to define the nearest neighbors. As with clustering, the presence of non-numeric attributes calls for the use of a distance measure that handles disparate attributes, or indicator variables that transform the non-numeric attributes. Our approach is described further below.

To assign a test row to a training cluster, $clusterID_{1\text{-}tst}$, the distance matrix *d*, used in the clustering in Step 2, of the test row to all the trainings rows is considered and the most frequent cluster membership, $ClusterID_{0\text{-}trn}$, of the *n* closest training rows is taken to be the test row's train cluster. This results in two different partitions of the test rows due to $ClusterID_{0\text{-}tst}$ and $ClusterID_{1\text{-}tst}$.

4. Stability Calculation: The agreement between the two partitions $ClusterID_{0\text{-}tst}$ and $clusterID_{1\text{-}tst}$ of the test dataset is measured to determine cluster stability. Towards this end, the Adjusted Rand Index (ARI) by Hubert, and Arabie [1985] to measure the agreement has been adopted. This measure has been shown to be appropriate for the comparison of partitions both theoretically in Hubert, and Arabie [1985] and empirically in Milligan and Cooper [1986]. In the context of cluster stability, this measure has been used in Breckenridge [2000]. It is available in an R package.

To define the ARI, consider the two data partitions of the test rows obtained through Steps 2 and 3. Then, summarize the rows in the test dataset using a contingency table such as Table 1 according to these partitions. The ARI is given by

$$\text{ARI}=\frac{\binom{N_{tst}}{2}(a+d)-\left[(a+b)(a+c)+(c+d)(b+d)\right]}{\binom{N_{tst}}{2}^{2}-\left[(a+b)(a+c)+(c+d)(b+d)\right]}$$

Table 1*:* A contingency table summarizing the result of two partitions $ClusterID_{0\text{-}tst}$ and $clusterID_{1\text{-}tst}$. In this table *a*, *b*, *c*, and *d* represent the total number of pairs meeting the requirements described in the contingency table and each pair represent pairs of rows from D_{tst}..

	Pair in same group $ClusterID_{0\text{-}tst}$	Pair in different groups $ClusterID_{1\text{-}tst}$
Pair in the same group *U*	*A*	*B*
Pair in different groups *U*	*C*	*D*

Algorithm 2 details the steps for determining the cluster stability of a particular solution using the ARI. Generally clustering solutions with higher ARI are preferred. For example, Breckenridge [2000] used ARI of around 0.7 as an indication of high agreement. The distance measure, algorithm and model order of the stable clustering solution is then be applied to the full dataset (without splitting) to form the final cluster.

3.3. Cluster Description

Once clusters have been generated, the proposed approach aims to describe them via rules. These rules should be small in number to allow understandability, have high coverage, and exhibit significant difference with respect to cluster membership compared to the entire population [Lavrac et al, 2004].

In this context, the application of subgroup discovery [Wrobel, 1997 & 2001; Klosgen, 1996] to uncover subgroups within each cluster, in turn highlighting key characteristics of each cluster is appropriate. Unlike rule-based classifiers, the goal of subgroup discovery is not to maximize classification accuracy; instead, the goal is to discover simple and interesting individual rules [Lavrac et al, 2004.].

Algorithm 2

CalcARI

input: D denoting the dataset, δ percentage used in splitting the data into test and train datasets, d the distance matrix, A the clustering algorithm, K the model order, n the number of nearest neighbors
output: ARI_K denoting the Adjusted Rand Index for model order K

$D_{tst}, D_{trn} \leftarrow$ Split (D, δ) split the dataset into test D_{tst} and training D_{trn} with $N_{tst} = (1-\delta)N$, $N_{trn} = \delta N$
$clusterID_{0\text{-}tst} \leftarrow \text{AssignClusterID}_0(D_{tst}, d, \text{A}, K)$ perform clustering on D_{tst} according to specified inputs d, A, K
$clusterID_{0\text{-}trn} \leftarrow \text{AssignClusterID}_0(D_{trn}, d, \text{A}, K)$
$clusterID_{1\text{-}tst} \leftarrow \text{LinkClusterID}_1(D_{tst}, D_{trn}, d, clusterID_{0\_trn}, n)$ as defined below
return $ARI_K \leftarrow \text{CalcARI}(clusterID_{0\_tst}, clusterID_{1\_tst})$ calculate the Adjusted Rand Index

LinkClusterID$_1$
input: $D_{tst}, D_{trn}, d, clusterID_{0\text{-}trn}, n$
output: $ClusterID_{1\text{-}tst}$

for each $x_i \in D_{tst}$ **do**
$D_{trn\text{-}n} \leftarrow \text{NearestNeighbor}(x_i, D_{trn}, d, n)$ select the set of n nearest rows of x_i to D_{trn} using d
$clusterID_{1\text{-}}x_i \leftarrow \underset{x_{i'} \in D_{trn-n}}{\text{Mode}} [clusterID_{0\text{-}}x_{i'}]$ take the most frequent cluster membership among $D_{trn\text{-}n}$
end for
return $clusterID_{1\text{-}tst:} \leftarrow \{clusterID_{1\text{-}xi}, x_i \in D_{tst}\}$

Previous work has studied the adaption of rule learning approaches for the task of subgroup discovery. For example, Lavrac et al [2004] adapted the CN2 rule learning algorithm for subgroup discovery, and Atzmueller and Puppe [2006] used the FP-growth algorithm [Han, Pei, Yin, 2000], originally developed for association rule mining, for this task. In the latter approach, the rules are based on conditional association rules in which the consequent of the rules is restricted to the target attribute. A quality function is used to reduce the number of the generated rules in the post-processing step. Compared to association rules that measure the support and confidence of rules, subgroup discovery uses a quality function to measure interestingness. Such functions are usually monotonically increasing with respect to the subgroup size and the frequency of the target concept in the subgroup. The Piatetsky-Shapiro (PS) function, $q_{ps}=n_s\,(p\text{-}p_0)$, is an example of such a quality function. Here n_s denotes the subgroup size, p the relative frequency of the target attribute in the subgroup, and p_0 the relative frequency of the target attribute in the total population.

Towards generating the cluster describing rules, each row is assigned to a cluster denoted by $clusterID_0$ derived through clustering of the data. The $clusterID_0$ is treated as the class attribute. The goal is to find population subgroups that are "most interesting", are as large as possible and have the most unusual characteristic with respect to the property of interest [Wrobel, 1997& 2001; Klosgen, 1996]. The property of

interest in our application is the cluster membership of the individual. Specifically, we define the task of describing the clusters as follows: Given the *K* formed clusters, generate rules for each cluster *k, k=1:K*, in the form of r_{gk}: (condition$_{gk}$)→$clusterID_0 = k$ where r_{gk} is the g^{th} rule for cluster *k*, and the condition is summarized as a conjunction of attribute, value pairs.

Finally, it should be noted that within the educational literature there has been some work that use rules to characterize predefined or external groups. For example, Perera, et al. [2009] use sequential pattern mining to discover patterns that characterize a predefined grouping of users based on the users' performance and Romero, et al. [2009] apply subgroup discovery to uncover rules that describe relationships between the students' usage activity in an e-learning system and the final marks obtained. Although our approach is similar to these, the distinction lies in the fact that the groupings in our work were not predefined and were obtained from the cluster analysis itself.

3.4. Learner Segmentation Analysis Framework

Having explained the building blocks of the proposed framework, Algorithm 3 summarizes the suggested framework for grouping users of an online learning environment based on key learner attributes.

Algorithm 3

LearnerSegmentation
input: *D* denoting the datasets, *nforest* the number of forests, *ntree* the number of trees, , δ percentage used in splitting the data into test and train datasets , *d* the distance , A the clustering algorithm, $\mathcal{K}$ a set of model order values, *n* the number of nearest neighbors, and ρ criteria for model order and *T* the desired number of rules
output: $clusterID_0$ denoting cluster memberships and *R* the set of cluster describing rules

RFdis←CalcRFDiss(*D, nforest, ntree*) calculate RF dissimilarity matrix as in Algorithm 1
for each $\kappa \in \mathcal{K}$
ARI_κ ←CalcARI(*D*, δ ,*d*, A, κ, *n*) calculate the Adjusted Rand Index for different model orders as in Algorithm 2
$ARI_{\mathcal{K}}$ ←{ ARI_κ, $\kappa \in \mathcal{K}$ }
end for
K←SelectOrder($ARI_{K,}$ ρ) select model order using solution with an ARI within ρ of the peak value
$clusterID_0$←AssignClusterID$_0$(*D, d,* A, *K*) perform clustering on *D* according to specified inputs *d*, A, *K*
R←SD-MAP(*D*, $clusterID_{0,}$ *T*) generate top *T* rules with respect to the PS quality function using SD-MAP
return $clusterID_{0,}$ *R*

4. CASE STUDY DATASET

The approach described in Section 3 was applied to datasets collected during a study previously performed by the *Career*WISE (CW) research program. The CW research program is a large interdisciplinary research program housed at Arizona State University and supported by the National Science Foundation. One of the major components of the program is the development of an online, psycho-educational resilience training program designed to improve knowledge and skills that are associated with the persistence of women in

Science, Technology, Engineering, and Math (STEM) programs [Bernstein, 2011]. The CW website users are able to interact with content that focuses on the application of a problem-solving model to manage personal and interpersonal difficulties that may be impeding their progress toward completing their PhD program. The CW website has almost 250 unique pages, each classified in several ways including by problem-solving step (CW teaches a four-step problem-solving approach applicable to personal and interpersonal issues) and by type of problem commonly experienced by women in STEM doctoral programs (difficulties with advisors, undesirable trade-offs between academic and personal responsibilities, unfriendly environments, and impediments to research and timely progress).

A Randomized Controlled Trial (RCT) was performed to formally evaluate the effectiveness of the resource in affecting the variables it was designed to influence [Bekki et al., under review]. During the RCT, 133 women doctoral students in STEM programs from 27 universities around the United States completed an assessment instrument, known as the context assessment, prior to gaining access to the actual CW resilience training site. The data from this instrument was used in the clustering analysis framework described in this paper. The context assessment measured the elements in the proximal environment of the participant that provide tangible support or represent external barriers in accomplishing academic and career goals. In the assessment, 85 items were organized under five topics: characteristics of the program or department, workplace, institution, relationship with advisor, and consideration of home and family life. Responses to four sections were arrayed on a five-point Likert scale; responses to a fifth scale were numeric. Items from this instrument to which too few participants responded in the study were excluded for the case study presented here. Ultimately, 69 items (each item represents an attribute in the dataset) were used in the cluster analysis (referred to as CW1 dataset). Examples of the items in the CW context assessment include those below. For the complete assessment instrument, please contact the authors.

- "My program is appropriate to my abilities and interests"
- "Support your advisor provides for your personal career goals"
- "There are not enough women faculty and researchers in my degree program."
- "I feel isolated from other students in my program."
- "I believe it would be almost impossible to finish my degree and raise a family".
- "I consider my adviser to be my mentor".

5. CASE STUDY

The problem definition is as follows: first, cluster users based on the CW1 dataset such that cluster membership reflects on the learning need of the users. Second, describe the learning need of each cluster so that targeted material may be presented accordingly. These objectives are carried out in Sections 5.1 and 5.2. Section 5.3 provides additional empirical investigations of the RF-based clustering approach, comparing the stability of clustering solutions generated using RF-based approaches to those using more traditional, Euclidean, based approaches.

5.1. CW1 Cluster Generation and Validation

To identify clustering solution with high stability, two clustering approaches were examined and compared based on their ARI. First, the RF dissimilarity was used as input to the PAM clustering algorithm (referred to as RF-PAM), and second, the Euclidean distance was used as input to PAM (E-PAM). The PAM clustering algorithm was selected because the RF only provides distance measures, and a clustering algorithm that can operate with only this input was needed. Additionally, PAM is robust to outliers and can easily be extended for use on larger datasets by implementing the related CLARA algorithm.

To produce the RF dissimilarity measures, the CW1 dataset was used as the input to the procedure outlined in Algorithm 1. The randomForest package in the R system for statistical computing was used to construct the RFs. In order to include different realizations of the synthetic data in constructing the unsupervised RFs, results are averaged across several forests each generated through different synthetic data. The proximities were averaged across 100 forests each consisting of 100 trees. A subset of $m \approx \sqrt{M}$ candidate attributes were selected for consideration at each node of each tree, and each tree was grown to the maximum possible size.

In addition to comparing clustering approaches, models orders of K from 2 to 6 were investigated. To assess the stability, the procedure in Algorithm 2 was replicated10 times for each of the two clustering solutions (RF-PAM and E-PAM), yielding 10 ARI values for each model order. Each replicate splits the data differently and applies clustering accordingly. This yields 10 ARIs for each model order. Figure 1 summarizes the results for model order of K from 2 to 6 based on a 10 nearest neighbor classifier in Step 3.2 of Algorithm 2. It should be noted that the relative ARI of the two solutions was not sensitive to the number of nearest neighbors (experiments considered n= 5, 10, 15 and showed results for 10 in Figure1). Figure 1 demonstrates that the RF-PAM clustering solution exhibits higher stability compared to E-PAM, highlighting the benefits of the RF distance measure for this data. Using the recommendation in Breckenridge [2000] of estimating the model order

using the clustering solution with an ARI within 0.05 of the peak value, the results also show that the estimated model order is *K=2*.

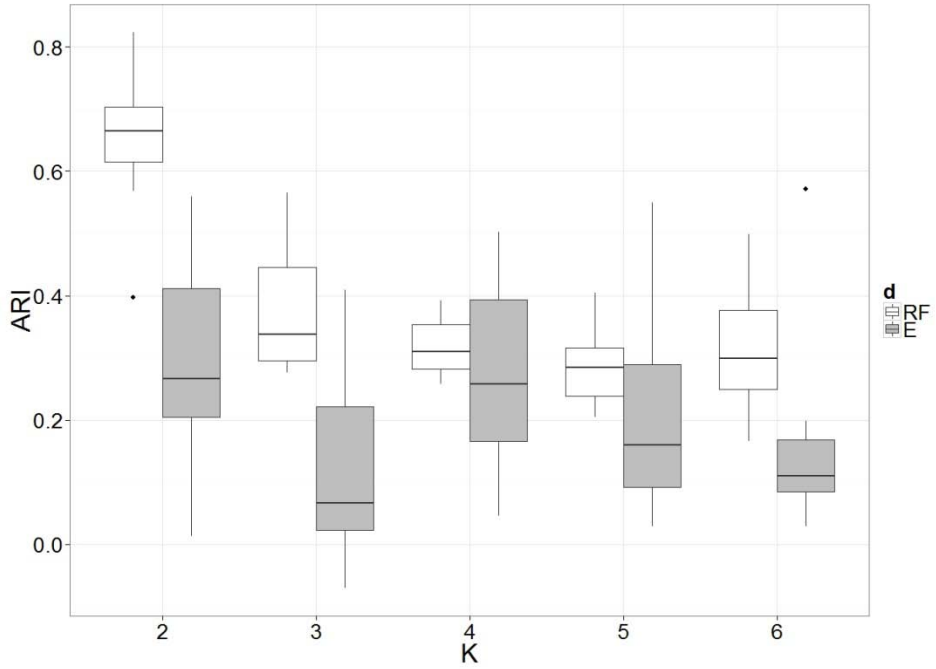

Figure 1: The ARI for the RF-PAM and E-PAM to the PAM for the CW1 dataset. The RF-PAM clustering solution of order two clearly exhibits the highest ARI.

5.2. CW1 Data Cluster Description

Following cluster generation using the RF-PAM approach on the entire CW1 data set, rules were generated to describe the resultant two clusters (*K*=2 with RF-PAM is the solution that shows the greatest stability in Figure 1). This clustering solution partitioned the rows into clusters of size 57 and 74 .

The subgroup discovery approach was then used to mine rules that describe subgroups of each cluster. The 69 items of the CW1 dataset were used as the input attributes, and the $ClusterID_0$ as the class attribute. Because there are only two clusters, this transformed the unsupervised problem into a binary target problem. For rule generation, the SD-MAP subgroup discovery algorithm [Atzmueller and Puppe, 2006] was implemented in the Vikamine software [Atzmueller and Lemmerich, 2012]. This algorithm adopts the FP-growth algorithm [Han, Pei, Yin, 2000], originally developed for mining association rules, for to the task of subgroup discovery. It is important to note that, in contrast to heuristic methods, this exhaustive approach guarantees the discovery of all interesting patterns.

The Piatetsky-Shapiro (PS) quality function was used as the quality function for ranking the discovered rules. We considered the top 20 rules (subgroups) in each cluster. Tables 2 and 3 present some examples of the

generated rules for each of the two formed clusters. Tables 2 and 3 also provide the lift values for each rule as well as the number of true and false positives.

Table 2: Examples of the cluster describing rules for the first cluster. The PS quality measure, lift and true and false positives (TP, FP) are also shown.

$ClusterID_0$=1	Quality function	Lift	TP	FP
d3: "Opportunities to collaborate with your adviser": Somewhat Dissatisfied AND f7: "Chances that you will transfer to a Master's degree program": High	10.733	2.298	19	0
d14: "The frequency of feedback your adviser provides": Somewhat Dissatisfied	14.947	2.145	28	2
c8: "My work is criticized or overlooked more than it should be": Agree	10.298	2.183	19	1

Table 3: Examples of the cluster describing rules for the second cluster. The PS quality measure, lift and true and false positives (TP, FP) are also show.

$clusterID_0=2$	Quality function	Lift	TP	FP
d6: "Clear expectation established by your adviser": Somewhat Satisfied OR Very Satisfied AND b5: "I believe that I can both finish my program and have a satisfying personal life": Agree OR Strongly Agree AND d1: "The overall relationship between you and your adviser": Somewhat Satisfied OR Very Satisfied	26.282	1.716	63	2
d7: "Extend to which your adviser serves as a role model for you": Somewhat Satisfied OR Very Satisfied AND d14: "The frequency of feedback your adviser provides": Somewhat Satisfied OR Very Satisfied	25.458	1.644	65	5

The objective of the rules is to discover traits of the formed clusters and, hence, the property of interest is derived by the data itself through clustering as opposed to be known beforehand as in the usual subgroup discovery problem. Since the data under study describes the learning needs of the users, the subgroup characteristics of each cluster allow insight about the learning needs of the users in each cluster. The first cluster is generally composed of subgroups that are dissatisfied with some aspects of their study. For example, a rule reflects on the dissatisfaction with the frequency of feedback from the adviser, and another rule describes users who are both unsatisfied with the opportunities to collaborate with their advisor and who report a high chance of not completing their PhD. On the other hand, the rules for the second cluster generally describe users who are satisfied with their PhD study. As an example, one rule for the second cluster describes users who are satisfied with the clear expectations established by their adviser and their overall relationship with their adviser who also believe that they can finish their studies and have a satisfying personal life.

Following the definition of actionability in Silberschatz and Tuzhilin [1995], a pattern is said to be interesting if "the user can react to it to his or her advantage". In this direction, the rules discovered for the first

cluster are especially interesting as they give insight on the learning needs of the users who may benefit the most from the CW website.

The generated rules were submitted to a content expert for interpretation and validation. The domain expert's interpretation of the clusters focused on whether patterns of user needs could be detected. The review of the rules revealed clear differentiation on the dimension of satisfaction across the clusters. Moreover, the cluster analysis provided strong support for the notion that the doctoral students' working relationship with the advisor is an important factor in forming similar groups of students. The derived CW clusters have important implications for the eventual development of a more personalized design for the online learning environment. The site is intended to assist a wide audience of interested students in improving their personal and interpersonal skills and knowledge. The application of the clustering approach described here enables the CW researchers to begin the process of identifying those students who may benefit the most from the CW resource and then directing them to the most relevant materials on the site.

It is also worth mentioning that RFs produce a measure of variable importance that can be used to determine which attributes were most important in defining the RF dissimilarities and forming the clusters. For each attribute, the information gain (based on the Gini index) is totaled over all splits in all trees in the forest where that attribute is the splitter [Breiman 2001]. Figure 2 shows the top 30 important attributes. As depicted in the figure, attributes with the highest variable importance measure pertained to adviser issues (section d of the context assessment instrument), providing further support for the interpretation of the final clustering solution.

We briefly summarizes our grouping of the CW users based on key learner attributes through Algorithm 3 where we have used the RF distance as input to the PAM algorithm (A=PAM, *d*=RFdis). We averaged the results over 100 forests of 100 trees (*ntree*=100, *nforest*=100), split the dataset in to two equal sized sets, $N_{tst} = N_{trn}=N/2$ (δ=0.5) and used 10 nearest neighbors (n=10). Model order was selected from the set $\mathcal{K}$ =2:6 choosing the model order of the clustering solution with an ARI of ρ=0.05 of the peak value. Finally, the top T =20 rules with respect to the PS quality function were considered as the set of cluster describing rules.

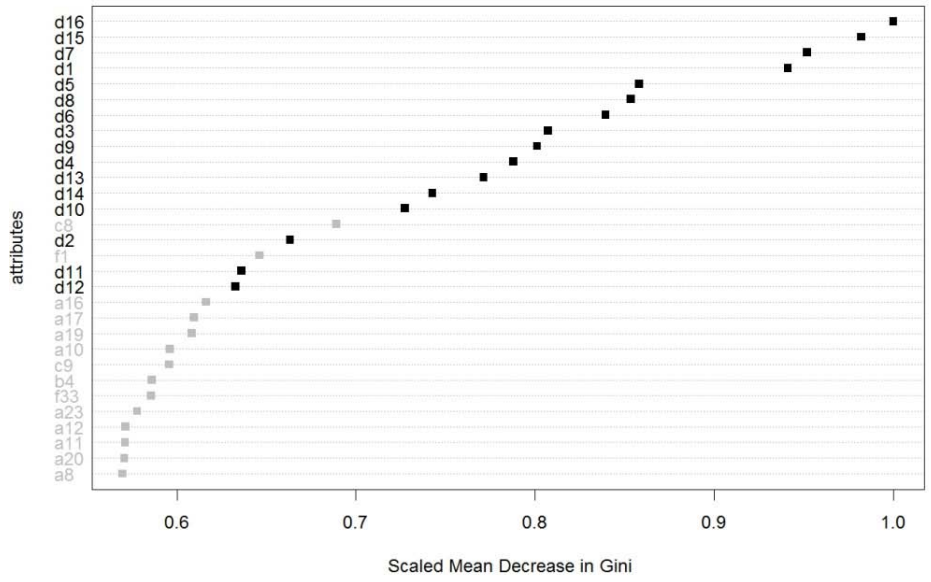

Figure 2*:* RF variable importance measure reflecting the high importance of adviser related attributes (section d of the context assessment instrument noted by items starting with a d shown in black). The information gain (based on the Gini index) is averaged over all trees across all forests and is scaled.

5.3. Additional Empirical Investigations of the RF-based Clustering Approach

To further investigate the effectiveness of the RF-based clustering approach in comparison to the more traditional, Euclidean distance based approaches, three additional experiments were conducted. Of note is that the complete learner segmentation methodology is not applied in these experiments; their purpose is simply to provide additional, empirical comparisons between the RF and more traditional Euclidean-based distance measures.

The first of the additional experiments utilizes a datasets that came from a later study, also performed by the CW research team, but with different participants and meant to evaluate a companion online learning environment dedicated exclusively to interpersonal communication skill training. The dataset used for the cluster analysis here is based on a background survey that was used as part of the study to obtain standard demographic information from the participants, as well as information about their academic status, career goals, and disciplinary research experiences. This dataset (referred to as CW2) was collected for 336 students and includes 15 attributes, eight of which are categorical and seven of which are numeric. As was done with the CW1 dataset, the stability of clustering solutions obtained from RF-PAM and E-PAM and for various model orders ($\mathcal{K}$ = 2:6) is compared. Due to the presence of categorical attributes, transformation of the categorical

attributes to numeric was required for calculating Euclidean distance. This was done through creating indicator variables. Figure 3 shows the ARI of the clustering solutions. As depicted by the plot, based on the ARI, the clustering solutions obtained from the RF dissimilarity are superior to those obtained from the Euclidean distance.

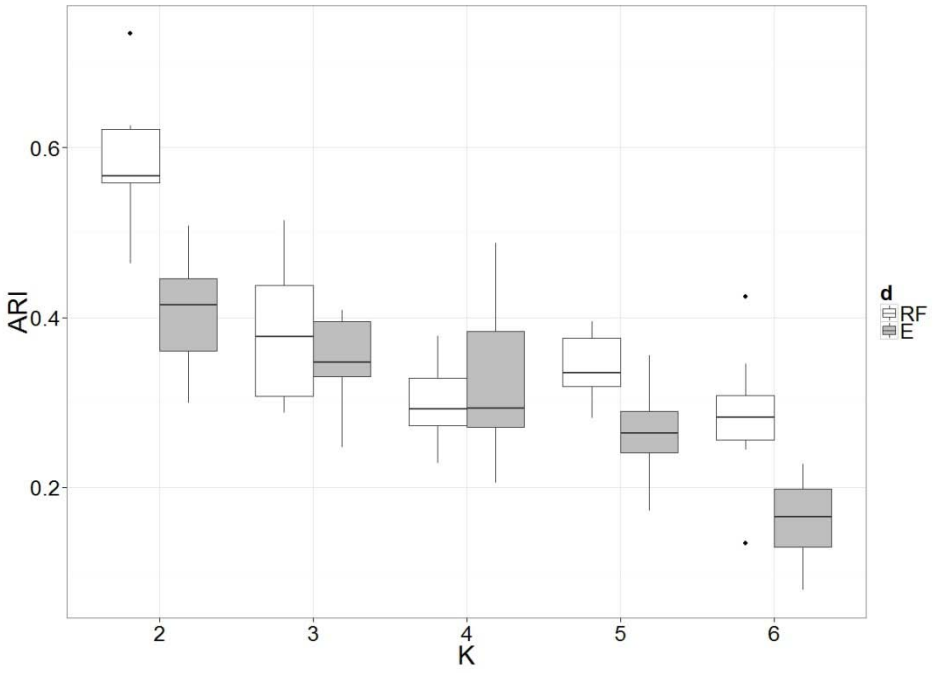

Figure 3: The ARI for the RF-PAM and E-PAM to the PAM for the CW2 dataset. The RF-PAM clustering solution exhibits the highest ARI.

Additionally, the stability of clustering solutions generated from noise was investigated. In this direction, a new dataset obtained through randomly sampling from the product of the marginal distributions of the attributes in the CW1 dataset is created. The random data is not expected to possess natural clusters and, therefore, clustering based on each partition of the dataset should result in a drastically different partition of the feature space. Consequently, we expect to have low agreement between the partitions. The procedure in Algorithm 2 was replicated 10 times on this random dataset. Figure 4 depicts the results. The considerable decline of the ARI compared to Figure 1 and Figure 3 is worth noting.

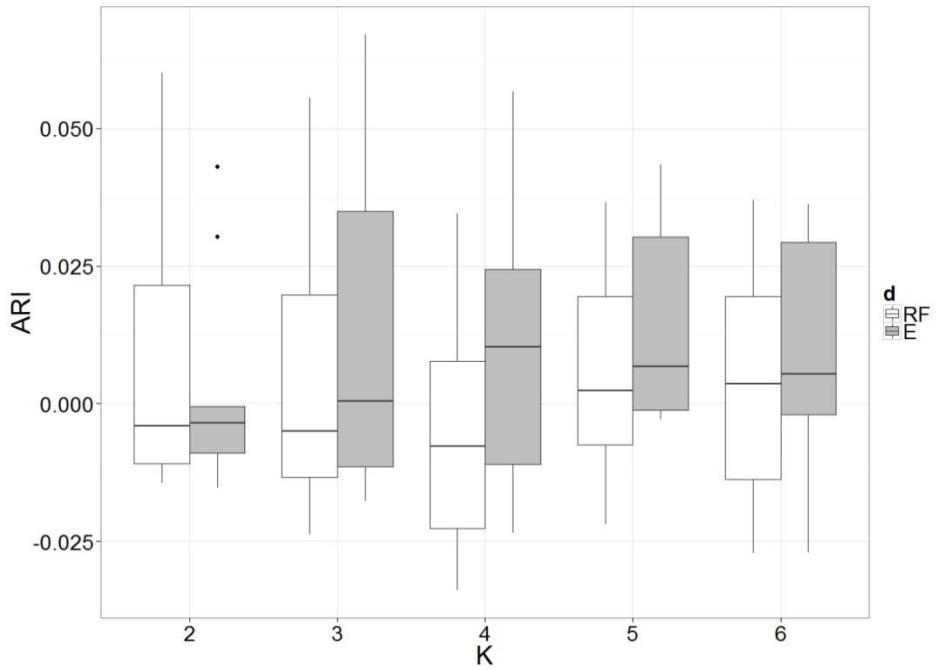

Figure 4: The ARI of RF-PAM and E-PAM for the random dataset. The clustering solutions exhibit low ARI due to the absence of natural clusters in the random data.

Finally, the removal of attributes that exhibit low standard deviation is considered to compare the proposed approach to one using simple feature selection. To do so, attributes that exhibit low standard deviation are first removed. Figure 5 shows the standard deviation, expressed in the same units as the attributes, of the CW 1 dataset. Then, Figure 6 gives the resulting ARI stability measure after removing the eight attributes with the smallest standard deviation and applying clustering with both Euclidean and RF distance measures. As shown in Figure 6, even after removal of the attributes with low standard deviation, the higher stability of the RF clustering solution remains. It is worth noting that the choice of removing the first eight attributes with smallest standard deviation is driven by Figure 5, but in order to apply such an approach with a traditional clustering approach, a more systematic approach may be required. The RF approach, however, uses an intrinsic attribute selection approach, bypassing the need for this step.

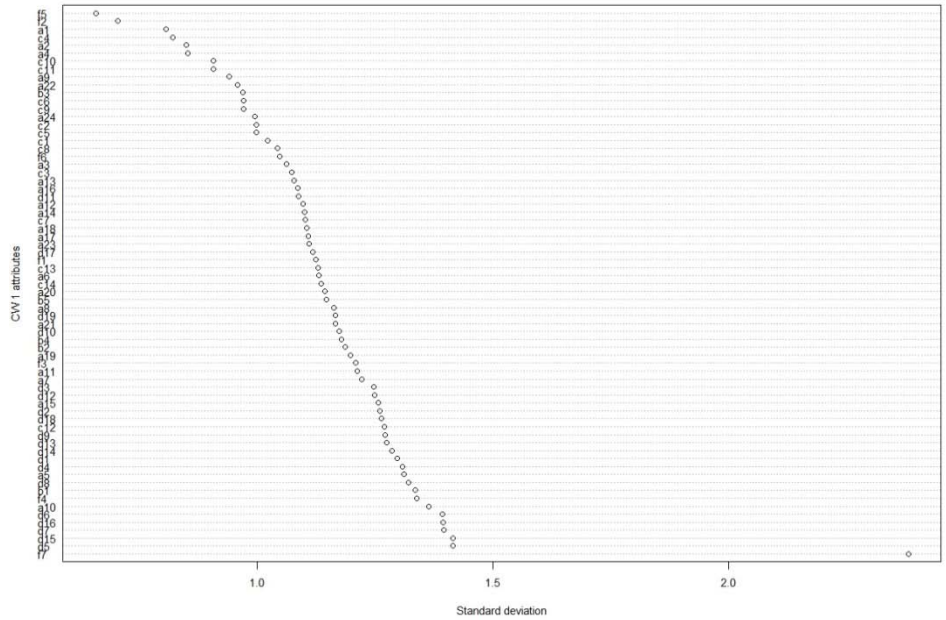

Figure 5: Standard deviations of CW1 attributes.

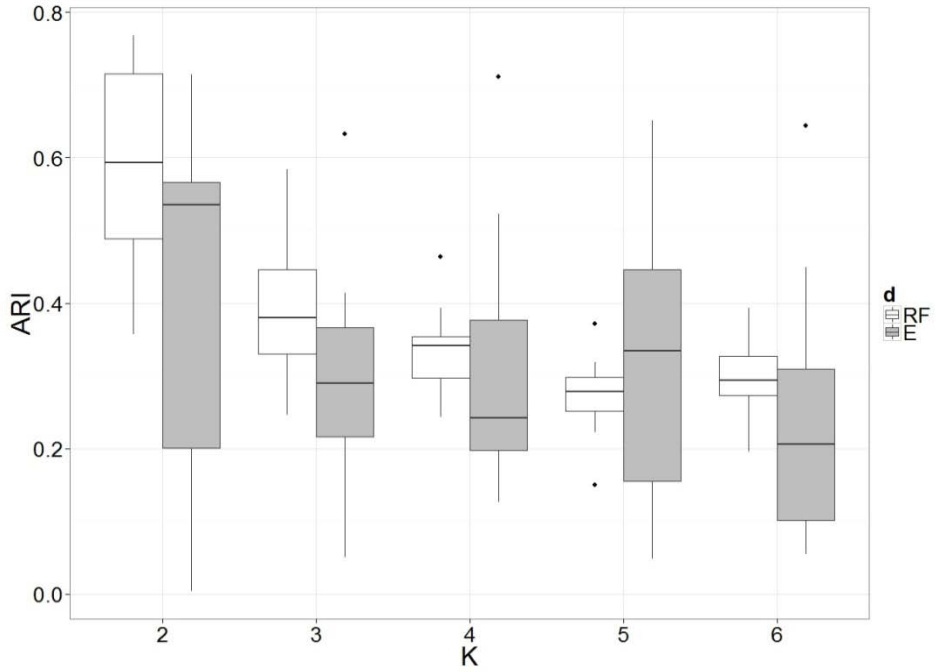

Figure 6: ARI after removing attributes with smaller standard deviation from the CW1 dataset.

6. CONCLUSION

This paper presents the application of a novel clustering approach for the problem of learner segmentation. We note that while the approach could be used for segmenting any objects, it is well suited for the learners of online environments. The applied approach handles many of the challenges common to educational data, such as high dimensional data with attributes of disparate type and scale, which has not been adequately addressed in the literature. An exhaustive subgroup discovery approach is adopted for uncovering subgroups within each cluster, in turn highlighting key characteristics of each cluster.

The approach was applied successfully to the data associated with the CW online learning environment, which provided a rich test bed for the application of the approach. When submitted to a domain expert for interpretation, the generated clusters were determined to represent meaningful and useful groupings. Additional experimentation with the approach also showed that it produces clusters with greater stability than those using more traditional, Euclidean distance based approaches.

It should be noted that although the CW dataset has served as an example in order to demonstrate the detailed steps of the proposed framework, this dataset does not include all the complexities that are expected to arise from larger-scale environments. Furthermore, while this clustering work represents an important first step toward the goal of developing the statistical approaches required for the personalization of online learning environments, there is still much to do. In future work, we plan to investigate mechanisms for determining the appropriate content to display to the clustered groups. This involves investigating user paths within the environment and mapping them to appropriate learning outcomes.

ACKNOWLEDGEMENTS

This study was supported by the National Science Foundation (NSF) grants 0634519 and 0910384. Any opinions, findings, and conclusions and recommendations expressed in this report are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

ALLEN, E., HORVATH, S., KRAFT, P., TONG, F., SPITERI, E., RIGGS, A., AND MARAHRENS, Y. 2003. High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes. *Proceedings of the National Academy of Sciences*, 100 .17, 9940–9945.

ATZMUELLER, M. AND LEMMERICH, F. 2012. VIKAMINE--Open-Source Subgroup Discovery, Pattern Mining, and Analytics, *Machine Learning and Knowledge Discovery in Databases*, 842--845

ATZMUELLER, M., AND PUPPE, F. 2006. SD–MAP–A Fast Algorithm for Exhaustive Subgroup Discovery. *In Proce. 10th European Conference on Principles and Practices of Knowledge Discovery in Databases (PKDD 2006),* 4213 in LNAI, 6-17. Berlin: Springer Verlag.

BEKKI, J.M., SMITH, M.L., BERNSTEIN, B.L., AND HARRISON, C.J. 2012. under review. Effects of an Online Personal Resilience Training Program for Women in STEM Doctoral Programs.

BELLMAN, R.E. 1966. *Adaptive control processes: a guided tour*. New Jersey: Princeton University Press.

BERNSTEIN, B. L. 2011. Managing barriers and building supports in science and engineering doctoral programs: Conceptual underpinnings for a new online training program for women. *Journal of Women and Minorities in Science and Engineering.* 17.1, 29-50.

BERNSTEIN, B. L. AND RUSSO, N. F. 2008. Explaining too few women in academic science and engineering careers: A psychosocial perspective. In *The psychology of women at work: Challenges and solutions for our female workforce* , M. Paludi, Ed., Praeger, Westport, 1 – 33 .

Breckenridge, James N, 2000. Validating cluster analysis: Consistent replication and symmetry, *Multivariate Behavioral Research*, 35.2, 261-285.

BREIMAN, L. 2001.Random forests. *Machine Learning*, 45 .1, 5-32.

BREIMAN, L. 2002. RFtools–two-eyed algorithms. Invited talk at *SIAM International Conference on Data Mining* . Available at*:* http://oz.berkeley.edu/users/breiman/siamtalk2003.pdf).

BREIMAN, L., AND CUTLER, A. 2003.Random forest manual v4.0. Technical report.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., AND STONE, C.J. 1984. *Classification and Regression Trees*. Wadsworth, California.

DIDAY, E., AND SIMON, J. C. 1976. Clustering Analysis, *Digital Pattern Recognition*, 10, 47–94.

DURAN, B.S., AND ODELL, P. L .1974. *Cluster Analysis: A survey*. Springer, New York.

FERGUSON, R. 2012. The state of learning analytics in 2012: A review and future challenges. *Technical Report KMI-12-01, Knowledge Media Institute,* The Open University, UK. http://kmi.open.ac.uk/publications/techreport/kmi-12-01, accessed June, 2012.

First European Conference on Principles of Data Mining and Knowledge Discovery, Springer, 78–87.

HAN, J., PEI, J., YIN, Y. 2000. Mining frequent patterns without candidate generation. In Chen, W., Naughton, J., Bernstein, P.A., eds: *2000 ACM SIGMOID Intl. Conference on Management of data*, ACM Press.

HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York.

HERHSKOVITZ, A., AND NACHMIAS, R. 2010. Online Persistence in Higher Education Web-supported Courses. *The Internet and Higher Education*, 14.2, 98-106.

HINES, P.J., JANSY, B.R., AND MERVIS, J .2009. Adding a T to the Three R's. *Science*, 323 .5910, 53-89.

HUBERT, L., AND ARABIE, P. 1985. Comparing partitions, *Journal of classification*,2.1,193-218.

JAIN, A.K., MURTY, M. N., AND FLYNN .P.J. 1999. Data clustering: a review. *ACM Computing Surveys .CSUR.,* 31 .3, 264-323.

JOHNSON, L., ADAMS, S., AND CUMMINS, M. 2012. *The NMC Horizon Report: 2012 Higher Education Edition*, The New Media Consortium. Austin.

KAUFMAN, L., AND ROUSSEEUW, P. J. 1990. Finding groups in data: A introduction to cluster analysis, New York: Wiley.

KLOSGEN, W. 1996. Explora: A multipattern and multistrategy discovery assistant. *Advances in Knowledge Discovery and Data Mining*, MIT Press, 249–271.

LANGE,T. ,ROTH,V., BRAUN, M.L., AND BUHMANN, J.M. 2004, Stability-based validation of clustering solutions, *Neural computation* ,16.6,1299-1323.

LAVRAC, N., KAVSEK, B., FLACH, P., AND TODOROVSKI, L. 2004. Subgroup Discovery with CN2-SD. *Journal of Machine Learning Research 5*, 153–188.

LIU, B., XIA, Y., AND YU, P., S. 2000. Clustering Trough Decision Tree Construction. *In Proceedings of the ninth international conference on information and knowledge management* , 20–29.

MEECE, J., L., AND HOLT, K. 1993. A pattern analysis of students' achievement goals. *Journal of Educational Psychology*, 85.4, 582-590.

MERCERON, A. AND YACEF, K. 2003. A web-based tutoring tool with mining facilities to improve learning and teaching. In *Proceedings of the 11th International Conference on Artificial Intelligence in Education*, 201– 208.

MERCERON, A., AND YACEF, K. 2004. Clustering Students to Help Evaluate Learning. In *Technology Enhanced Learning* , J.P. COURTIAT, C. DAVARAKIS, AND T. VILLEMUR .Eds. Kluwer, Toulouse, 31-42.

MICHALSKI, R., STEPP, R. E., AND DIDAY, E. 1981. A recent advance in data analysis: clustering objects into classes characterized by conjunctive concepts. In *Progress in Pattern Recognition*, L. N. KANAL AND A. ROSENFIELD, Eds., New York, 33-56.

MILLIGAN, G.W., AND COOPER, M. 1986: A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis, Multivariate Behavioral Research, 21.4, 441-458.

MURTAGH, F. 1983. A survey of recent advances in hierarchical clustering algorithms}, *The Computer Journal*, 26.4, 354-359.

NARCISS, S., PROSKE, A., AND KOERNDLE, H .2007. Promoting self-regulated learning in web-based environments. *Computers in Human Behavior*, 23.3, 1126 – 1144.

NG, R. T. AND HAN, J. 1994. Efficient and effective clustering methods for spatial data mining. In Proceedings of the Twentieth International Conference on Very Large Data Bases,144–154.

PARSON, L., HAQUE, E., AND LIU, H .2004. Subspace clustering for high dimensional data: A Review. *ACM SIGKDD Explorations Newsletter*, 6.1, 90-105.

PERERA, D., KAY, J., KOPRINSKA, I., YACEF, K., ZAÏANE, O. R. 2009. Clustering and Sequential Pattern Mining of Online Collaborative Learning Data. *In IEEE Transaction on Knowledge and Data Engineering*, 21.6, 759-772.

R Development Core Team .2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.

ROMERO, C. AND VENTURA, S. 2007. Educational data mining: A survey from 1995 – 2005. *Expert Systems with Applications*, 30, 135-146.

ROMERO, C. AND VENTURA, S. 2010. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetcis – Part C: Applications and Reviews*, 40.6, 601-618.

ROMERO, C., GONZALEZ, P., VENTURA, S., DEL JESUS, M.J., HERRERA, F.2009. Evolutionary algorithms for subgroup discovery in e-learning: A practical application using Moodle data. In *Expert System with Application Journal*, 36, 1632-1644.

ROMERO, C., VENTURA, S., AND GARCÍA, E. 2008. Data mining in course management systems: Moodle case study and tutorial. *Computers AND Education*, 51, 368-384.

SELIGSON, D.B., HORVATH, S., SHI, T., YU, H., TZE. S., GRUNSTEIN, M., AND KURDISTANI, S.K. 2005. Global histone modification pattern predict risk of prostate cancer recurrence. *Nature*, 435, 1262-1266.

SHAVELSON, R.J .1979. Application of cluster analysis in educational research: looking for a needle in haystack. *British Educational Research Journal*, 5.1, 45-53.

SHI, T., AND HORVATH, S. 2006. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15.1, 118-138.

SHI, T., SELIGSON, D., BELLDEGRUN, A.S., PALOTIE, A., AND HORVATH, S. 2005. Tumor classification by tissue microarray profiling: Random forest clustering applied to renal cell carcinoma. *Modern Pathology*, 18, 547-557.

SILBERSCHATZ, A., TUZHILIN, A. 1995. On Subjective Measures of Interestingness in Knowledge Discovery, *Proc. of the First Int'l Conference on Knowledge Discovery and Data Mining* , Montreal, Canada.

SINGH, S., S., AND CHAUHAN, N.,C. 2011, K-Means v/s K-Medoids: A Comparative Study, *National Conference on Recent Trends inEngineering And Technology*.

TALAVERA, L. AND GAUDIOSO, E. 2004. Mining student data to characterize similar behavior groups in unstructured collaboration spaces. *Proceedings of the Artificial Intelligence in Computer Supported Collaborative Learning Workshop at the ECAI 2004*, 17-23.

TAN, P.N., STEINBACH, M., AND KUMAR, V .2006. *Introduction to data mining*, Pearson Addison Wesley, Massachusetts.

TIBSHIRANI, R., AND WALTHER, G. 2005. Cluster validation by prediction strength, *Journal of Computational and Graphical Statistics*, 14.3, 511-528.

VALLE, R., AND DUFFY, M. 2009. Online learning: Learner characteristics and their approaches to managing learning, *Instructional Science*, 37, 129-149.

WROBEL, S. 1997. An algorithm for multi-relational discovery of subgroups. *In Proceedings of the*

WROBEL, S. 2001. Inductive logic programming for knowledge discovery in databases. *Relational Data Mining*, Springer, 74–101.

ZENKO, B., DZEROSKI, S., AND STRUYF, J. 2006. Learning predictive clustering rules. *Proceedings of the 4th International Workshop on Knowledge Discovery in Inductive Databases* , F. BONCHI AND J.F. BOULICAUT .Eds, Springer, Berlin, 234–250.