Fitness-Based Student Clustering Combining Clustering Algorithms and Dimensionality Reduction

Erik Dovgan
Department of Intelligent Systems
Jožef Stefan Institute
Jamova cesta 39
SI-1000 Ljubljana, Slovenia
erik.dovgan@ijs.si

Mitja Luštrek
Department of Intelligent Systems
Jožef Stefan Institute
Jamova cesta 39
SI-1000 Ljubljana, Slovenia
mitja.lustrek@ijs.si

ABSTRACT
Health status and well-being of persons are significantly influenced by their physical fitness since, for example, low fitness is related to several health risks. Identification of unfit persons enables us to personalize advice or recommendations to improve their fitness. However, this identification is not straightforward or standardized. For this purpose, we propose a combination of dimensionality reduction methods and clustering algorithms on data from a test battery used in schools. Such an approach enables teachers, parents and policy makers to identify clusters of fit and unfit students, and better target actions for improving student fitness.

Keywords
Clustering algorithms, Dimensionality reduction, Student clustering

1. INTRODUCTION
Physical fitness has significant impact on health and well-being, since several health risk factors are related to low fitness (see examples in [4] and [2]). In order to reduce the risks, the physical fitness should be improved. This can be achieved, for example, in schools, where teachers, parents and policy makers can create and provide advice and recommendations. Although general advice and recommendations are possible, they are more efficient, when they are personalized and better-targeted. However, to achieve this, we firstly need to identify unfit students, which is not trivial.

Identification of fit and unfit students is not straightforward or standardized. There are some metrics, such as body mass index (BMI) [3] and the related Overweight and Obese Adolescents (OOA) categories [1], which enable identifying underweight, normal weight, overweight, and obese persons. However, these metrics are not directly related to the physical fitness and consequently cannot be effectively used to identify unfit students. To cluster students into fit and unfit, we propose to apply dimensionality reduction methods and clustering algorithms on data from widely used test battery. The identification of unfit students will enable decision makers to personalize actions for improving students' fitness.

The paper is further organized as follows. The procedure for identification of fit and unfit students is presented in Section 2. Section 3 describes the experiments in terms of the used dataset and the obtained results. Finally, Section 4 concludes the paper with ideas for future work.

2. IDENTIFICATION OF FIT AND UNFIT STUDENTS
The goal of the developed procedure is the identification of fit and unfit students. Since there are different risks between underweight, normal weight and overweight students, we focus only on one of these categories, namely overweight students. Note that this category also includes obese persons.

The developed procedure takes into account a set of physical fitness measurements, known as SLOfit test battery[1], which are performed yearly in Slovenian schools. The set of measurements is shown in Table 1. These attributes are given as raw data or as percentiles, where the attribute's percentile, i.e., the attribute's quantile, is the rank of the student based on this attribute within the set of students of the same sex and age. In addition to the measurement data, the procedure can also take into account a set of additional attributes that are shown in Table 2. Note that Fitness index (FI) is the quantile rank of the sum of the quantile ranks of the fitness measurements.

Table 1: Measurements of the test battery.

Measurement	Short name
Thickness triceps skinfold	TTSF
Reaction time during arm plate tapping	TAPT
Distance during standing broad jump	DSBJ
Time pass polygon backwards on all fours	TOCB
Number sit-ups in 60 s	NSU
Distance fingertips-toes, bending forward	DSR
Time bent arm position, hanging from bar	TBAH
Time run 60 m	T60m
Time run 600 m	T600m

The proposed procedure searches for fit and unfit students as follows. The input data consist of the measurement attributes and (a subset of) additional attributes. These data are clustered into groups of students. However, the true clusters are not given thus the quality of clustering cannot be easily assessed. As a solution, we apply dimensionality reduction to obtain two-dimensional representation of data, which is then visually assessed in terms of meaningfulness of the obtained clusters. The meaningfulness of the clusters is assessed based on two clusters' properties:

[1]http://en.slofit.org/measurements/test-battery

Table 2: Additional attributes.

Attribute	Short name
Sex	SEX
Height (raw or percentile)	H
Weight (raw or percentile)	W
Grade	GRD
Age	AGE
Fitness index	FI
OOA categories	OOA
Body Mass Index	BMI

1. We aim at obtaining at least two clusters that are separable in the reduced-dimensional space.

2. The obtained clusters should not be correlated to discrete attributes, i.e., SEX, GRD, AGE, and OOA. Such correlation is not wanted due to the fact that the easiest way to cluster or reduce dimensions is to focus on attributes that are already separable, e.g., discrete attributes. However, such a clustering/dimensionality reduction is meaningless for decision making, e.g., it makes no sense to find clusters of males and females since these clusters are already known.

In our procedure, we apply the following clustering algorithms: KMeans, Affinity Propagation, Mean Shift, and Birch. In addition, we use the following dimensionality reduction methods: Factor analysis (FA), Principal component analysis (PCA), Singular value decomposition (SVD), Independent component analysis (ICA), Isometric mapping (ISOMAP), and Uniform manifold approximation and projection (UMAP). The developed procedure works as follows. For each combination of clustering algorithms and dimensionality reduction methods we apply the following steps.

Step 1: A subset of data is randomly selected to make clustering and dimensionality reduction feasible (due to the fact that some methods are computationally intensive).
Step 2: This subset is clustered and the model for data clustering is obtained.
Step 3: Dimensionality reduction method is applied on the subset and the model for dimensionality reduction is obtained.
Step 4: The entire dataset is clustered with the clustering model.
Step 5: The model for dimensionality reduction is applied on the entire dataset.
Step 6: The entire dataset is presented in the reduced-dimensional space. Clusters are marked with different colors.
Step 7: This representation is visually assessed in terms of meaningfulness of the obtained clusters.

3. EXPERIMENTS AND RESULTS

The proposed procedure was evaluated on two relevant sets of students from the SLOfit dataset: a) High school students (ages 15–19), and b) Elementary school students (up to age of 11). In addition, only data from the most recent year was used, i.e., 2018. The attributes are shown in Tables 1–2. Note that the data of high school students included GRD, but not AGE, while data of elementary school students included AGE, but not GRD. The following sections present the analysis of the data, which was performed in two steps.

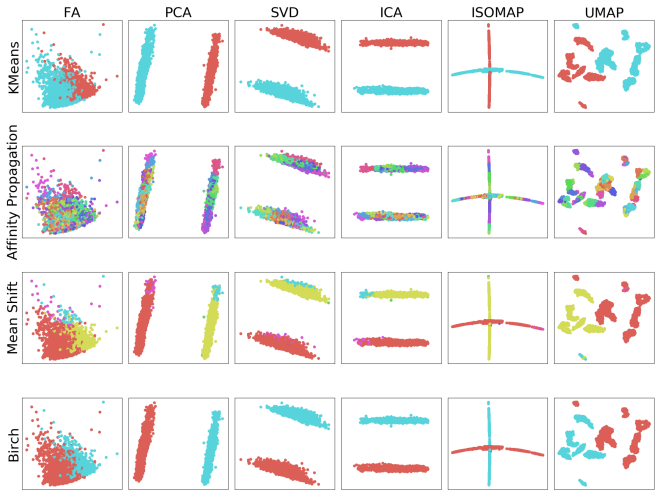

Figure 1: Dimensionality reduction and clustering on nonpercentile data from high school students. Reduced dimensions are presented with position in the new space, while clusters are shown with different colors.

3.1 Initial Analysis

The initial analysis was performed on all the attributes and with all the clustering algorithms and dimensionality reduction methods. The results are presented in Figure 1. This figure shows that several dimensionality reduction methods found two (separable) clusters, i.e., PCA, SVD, and ICA. In addition, KMeans and Birch identified those two clusters. However, the highest correlation with these clusters are obtained by SEX (see Figure 2). In addition, Figure 3 confirms that the two clusters obtained with reduced dimensions represent two sexes. Such clusters are obvious and thus not interesting for the decision makers.

Additional tests were performed on subsets of attributes and the results showed that only some subsets produced interesting clusters. Therefore, we decided to systematically assess various subsets of attributes. Since the measurements (see Table 1) are (probably) the most suitable for determining the physical fitness, we evaluated only the subsets of additional attributes (see Table 2), while measurements were always considered.

The results also indicated that some clustering algorithms and dimensionality reduction methods were redundant or uninformative. For example, Mean Shift and Affinity Propagation found more clusters than needed, while KMeans and Birch discovered the best (but the same) clusters (see Figure 1). Therefore, KMeans and Birch are redundant and we prefer KMeans among them due to its simplicity. In addition, dimensionality reduction methods can be divided into two groups: a) components/factor based (FA, PCA, SVD, and ICA), and b) projection based (ISOMAP and UMAP). The former, for example, aim at maximizing the variance

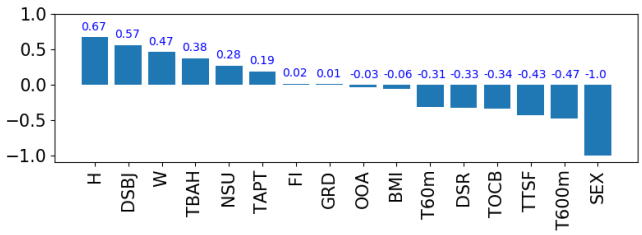

Figure 2: Correlation of clusters found by KMeans with the attributes on nonpercentile data from high school students.

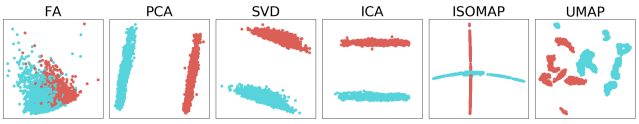

Figure 3: Relation between data in reduced dimensions and the SEX attribute on nonpercentile data from high school students, where SEX is shown with different colors.

in the dataset (PCA), while the latter try to maintain the distances between the data. Therefore, one representative of each group should be used, i.e., PCA as the most widely used from the first set, and UMAP (that is faster than ISOMAP) from the second set.

3.2 Systematic Analysis

Based on the results from the initial analysis, we decided to systematically evaluate all subsets of additional attributes (see Table 2) using only KMeans for clustering, and PCA and UMAP for reducing dimensions. Note that measurements' data (see Table 1) were always taken into account. In this way, the number of tested subsets of attributes was $2^7 = 128$ (due to 7 additional attributes). Each subset of attributes was evaluated four times: 1) percentile attributes of elementary school data, 2) nonpercentile attributes of elementary school data, 3) percentile attributes of high school data, and 4) nonpercentile attributes of high school data. This resulted in $4 \times 128 = 512$ tested subsets.

To additionally simplify clustering and find meaningful clusters, we grouped subsets of attributes with respect to their results and cluster the data with respect to clusters obtained with UMAP. More precisely, the procedure was as follows:

Step 1: Reducing dimensionality with UMAP.
Step 2: Clustering students for all subsets of attributes. Each subset represents one instance. Many instances cluster students similarly, therefore instances should be grouped in order to find only representative instances, i.e., one instance for each group of instances (Step 3).
Step 3: Grouping similar instances with respect to clusters of discrete attributes, i.e., OOA, SEX, and GRD/AGE, based on visual inspection. For example, grouping together instances with two clusters which represent two SEX-es, or instances with four clusters which represent four GRD-es.
Step 4: Clustering students for each group found in Step 3. The input for this clustering are the cluster ids of instances

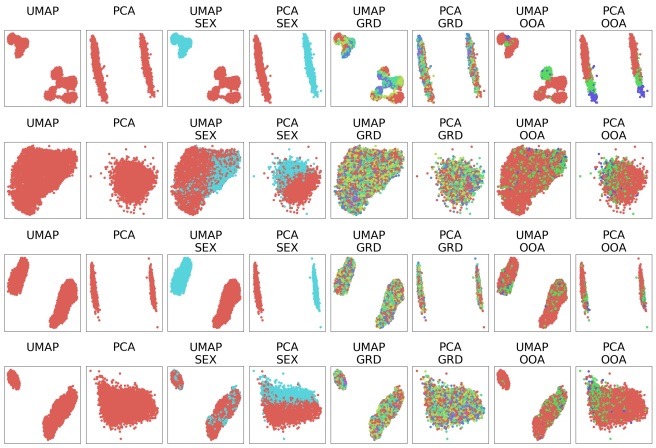

Figure 4: Examples of clusters found by PCA and UMAP on nonpercentile data from high school students. Columns 3–8 show the data with respect to discrete attributes, where attribute values are represented with different colors. Rows show the clusters with respect to four set of attributes that were used by dimensionality reduction methods. First row: all attributes, clusters represent SEX; Second row: measurements only, no clusters found; Third row: measurements and SEX, clusters represent SEX; Fourth row: measurements and FI, no connection between clusters and discrete attributes. According to definition of meaningful clusters given in Section 2, only the fourth row shows good clusters.

within the group, where these cluster ids were found in Step 2. The instances within the group might not produce similar clusters of students, therefore, such non-homogenous groups should be split (Step 5).
Step 5: Visual inspection of the obtained clusters for each group. If the found clusters are not consistent among the instances, the group is split and Step 3 is repeated. For example, if the group contains instances with two clusters and only half of these instances are correctly clustered into two clusters (i.e., consistent with clustering for the whole group), the correctly clustered instances represent a good subgroup and Step 6 should be applied. On the other hand, the incorrectly clustered instances are not a good subgroup and Step 3 has to be repeated.
Step 6: Finding representative instances of each group found in Step 5. For example, if attributes (e.g., SEX, GRD, W, H) of first instance are subset of the attributes of the second instance, the first instance (with a lower number of attributes) is more representative and should be used for further analysis.

This procedure enabled us to reduce the number of instances, i.e., subsets of attributes, that need to be analyzed, since it found the most representative subsets of attributes only.

The representative clusters are described in Table 3 in terms of additional attributes that were used (in addition to measurements), attribute type (raw or percentiles), the number of obtained clusters and the attribute with the highest cor-

Table 3: Properties of representative subsets of attributes obtained by combining dimensionality reduction methods and clustering algorithms.

Students	Additional attributes	Attribute type	Number of clusters	Attribute with highest correlation
high school	OOA	raw	2	OOA
high school	GRD, H, BMI	raw	2	GRD
high school	SEX	raw	2	SEX
high school	FI	raw	2	FI
high school	OOA, BMI	percentile	2	OOA
high school	SEX	percentile	2	SEX
high school	FI, GRD, W	percentile	2	FI
elementary school	AGE, H, W, FI, BMI, OOA	raw	2	OOA
elementary school	AGE, H, W, FI, BMI	raw	2	AGE
elementary school	SEX, AGE, W, FI, OOA	raw	3	SEX
elementary school	AGE, W, BMI, OOA	raw	3	OOA
elementary school	SEX, H, FI, BMI, OOA	raw	4	OOA
elementary school	AGE, W, FI	raw	5	AGE
elementary school	SEX, AGE, H, BMI, OOA	raw	6	SEX
elementary school	SEX, AGE, H, W, BMI, OOA	raw	7	SEX
elementary school	SEX, AGE, W, OOA	raw	8	SEX
elementary school	SEX, AGE, BMI, OOA	raw	10	SEX
elementary school	SEX, BMI, OOA	percentile	2	SEX

relation with clusters. Examples of obtained clusters are shown in Figure 4.

The results in Table 3 show that the majority of the obtained clusters is correlated with discrete attributes such as OOA, GRD/AGE, and SEX. The only exceptions are the two subsets of attributes whose clusters are correlated with FI. These subsets are also the most interesting ones, since the goal was to cluster the students based on their fitness into fit and unfit groups. These interesting clusters can be also seen in Figure 4 in fourth row. This figure confirms that the obtained clusters are not correlated with discrete attributes, i.e., SEX, GRD, and OOA.

Although we were able to find interesting clusters, i.e., those that are not correlated with discrete attributes, these clusters were obtained by using only specific subsets of attributes and only for high school data (see Table 3). In addition, no interesting clusters were obtained from elementary school data. It should be also noted that out of 512 possible attribute subsets, only two subsets were interesting, i.e., produced at least two clusters that were not correlated with discrete attributes.

4. CONCLUSION

This paper presented an approach for identification of fit and unfit students. This approach analyzes data from test battery used in schools by combining dimensionality reduction methods and clustering algorithms. By visually inspecting the results of data analysis, it enables us to find combinations of attributes that produce meaningful clusters of fit and unfit students. The identification of unfit students supports teachers, parents and policy makers in better targeting actions for improving fitness of those students.

In our future work, we will aim at developing a method for automatic assessment of obtained clusters, which will be used instead of visual assessment. To this end, an analytical approach for assessing the quality of clusters will be developed. This approach will also aim at determining to which extent the obtained clusters represent fit and unfit students. This task is especially challenging since we do not have the true values.

5. ACKNOWLEDGMENTS

This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 727560. We would also like to acknowledge the help of Bojan Leskošek, Gregor Jurak, Gregor Starc, and Maroje Sorić, from the Faculty of Sport, University of Ljubljana, Slovenia, who provided the SLOfit dataset and helped defining the problem.

6. REFERENCES

[1] F. Bacha, R. Saad, N. Gungor, J. Janosky, and S. A. Arslanian. Obesity, regional fat distribution, and syndrome X in obese black versus white adolescents: Race differential in diabetogenic and atherogenic risk factors. *The Journal of Clinical Endocrinology and Metabolism*, 88:2534–2540, 2003.

[2] S. W. Farrell, C. E. Finley, N. B. Radford, and W. L. Haskell. Cardiorespiratory fitness, body mass index, and heart failure mortality in men. *Circulation: Heart Failure*, 6(5):898–905, 2013.

[3] M. Kallioinen and S. I. Granheim. Overweight and obesity in the western pacific region. Technical report, World Health Organization, 2017.

[4] J. R. Ortlepp, J. Metrikat, M. Albrecht, P. Maya-Pelzer, H. Pongratz, and R. Hoffmann. Relation of body mass index, physical fitness, and the cardiovascular risk profile in 3127 young normal weight men with an apparently optimal lifestyle. *International Journal of Obesity*, 27:979–982, 2003.