Generic Approach for Interpretation of PCA Results - Use Case on Learner's Activity in Social Media Tools

Marian Cristian MIHĂESCU, Paul Ștefan POPESCU, Mihai Lucian MOCANU
University of Craiova, A.I. Cuza, 200-585 Craiova, Romania
mihaescu@software.ucv.ro

***Abstract*—Intensive usage of social media tools for educational purposes transformed many previously tackled issues from classical e-Learning systems. Among the most general challenging issues reside in building classification models having the performed activities set as independent variables and final grade as dependent variable. A critical step in data analysis process regards building interpretable models in terms of explaining feature values and ranges along with their influence on target class. We asked whether dimensionality reduction techniques may be effectively used such that high quality interpretable models are obtained. Principal component analysis (PCA) dimensionality reduction technique, scaling and several classical classification techniques were used to create a data analysis pipeline that produces classification models with similar accuracy of initial classification models built on raw available data. Experimental results show that features that characterize the activity performed on each social tool and on all tools are highly interpretable in our classification context. The proposed approach is flexible and can be adapted to similar practical use cases in which a large number of features is difficult to be interpreted and a digest is required as being more useful for bringing a better insight on data.**

***Index Terms*—data engineering, knowledge representation, machine learning, social network services, social computing.**

I. Introduction

On-line educational systems - classical e-Learning platforms or based on social tools - aim continuously to improve in terms of bringing a clearer insight on activities performed by learners. The usual approach relies on getting and analyzing data from available tools or sources that can bring better intuition about learner's engagement and activity. One method for accomplishing this task reduces to using machine learning algorithms on collected data. When there is a small number of features and a small number of target classes, the activity level can be obtained in a very intuitive way. For example, if we know how many hours a learner spent learning and how many tests he takes we may try to predict the final grade considering training data is available. The issue in this case is the small number of features that can bias our image regarding their activity although high interpretable classification models may be obtained. In our currently practical situation, machine learning algorithms can build models with better accuracy but taking into consideration the considerable number of features we face the problem of building highly interpretable classification models.

Learner modeling is accomplished in most of the cases using a set of features that describe the learners. Furthermore, after gathering all the data, data analyst can build models using machine learning algorithms. The models are strictly dependent on the number of features that describes an entity and also the number of instances used to build them. As the number of features and also the number of instances increases it becomes harder to build accurate models. The proposed approach from this paper analyses the possibility to use PCA dimensionality reduction technique that can produce a more interpretable model.

Learner's social activity is a key indicator regarding the engagement in the learning process and also offers a good intuition regarding how they perceive the learning material. This paper addresses the problem of creating a framework for interpretability for describing learner's social activity that can be estimated based on several features that can be extracted from three platforms: Twitter, blogspot and MediaWiki.

As dimensionality reduction method we employed PCA, a statistical procedure that aims to reduce the number of features without a significant impact on the dataset in terms of information loss and for classification accuracy. We used this approach in order to reduce the number of features that describe a learner and produce a more intuitive overview of their activity. Based on several preprocessing steps and a partial ordering approach, we can rank the learners and get an intuition regarding their activity performed.

Proposed approach has been designed for data provided by three social learning environments that were used in the learning process. Learners engaged in this study used these three tools continuously for one semester course and based on their activity we computed several features for each tool.

Since performed activities on available social tools are represented by a fairly large number of features we investigate the possibility of downsizing dimensionality while obtaining new reduced interpretable features that provide a very intuitive numerical interpretation.

Besides this goal, we aim to obtain a partial ordering of learners in terms of performed activities that validates the learning outcome and does not alter the quality and the interpretation of the initial data model that takes into account all available features.

The data analysis process is based on previously high quality obtained model [1] that accurately identified "*spam*"

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS-UEFISCDI, project number PN-II-RU-TE-2014-4-2604.

Digital Object Identifier 10.4316/AECE.2018.02004

and "*don't care*" users. The experimental setup uses PCA dimensionality reduction [2] technique for features available at tool level and scaling [3] for proving the partial ordering of learners by capturing the correlation between PCA values and distances from origin along with several classification models that used as target class the obtained discretized grades.

This paper presents in section two the related work in terms of other usages of PCA and other dimensionality reduction techniques. Section three presents the proposed approach that consists of a data analysis pipeline that uses PCA, scaling and classification models to obtain highly interpretable features. Section four presents experimental results that follow up a previously published paper which built high quality classification models considering a large number of features for each social media tool. Finally, conclusions and future work is presented by pointing out several key aspects that may further improve the results.

II. Related work

The use of PCA [4] in applied research [5] was referred in journals since 1964. Since then, many other usage scenarios [6] were reported, even mixtures of PCA [7]. Other approaches referred combination of PCA and other classification techniques [8], in this case the authors applied PCA for dimension reduction to honors learner dataset and then use the result as input for a RFF Neural Network that aims to build a predictive model. The results of using PCA show a faster convergence speed, a better algorithm accuracy and stronger generation ability.

Usage of PCA was made also in [9], where the authors analysed the use of social media in Higher education institution using Technology Acceptance Model [10]. The approach was a combination of PCA and Structural Equation Modeling (SEM) [11] used to analyse the relationships between determinants of these technologies. Another PCA related paper by Carlo Giovanella et.al. [12] refers analyzing the learner performance indicators such as learner's active involvement with web 2.0 tools and learners' learning styles. The authors used PCA to identify a subset of learner activities that are relevant to support a prediction of learner success. They also tested if there is a significant correlation between the learning styles and learner's performance and found that there is no correlation. Still in the area of social media and academic performance, S. C. Nsizwana et.al. [13] published a recent study conducted at the university of Zululand regarding the effects of social media use on academic performance of undergraduate learners. They had 68 participants and used a five-level Likert scale to determine if the use of social media have effects on the academic performance. They involved PCA to determine the extent of contribution of Likert scale items to the variables under study. The results indicated that the familiarity with social networks results in intensive usage of social networks and academic activities. The results also showed that learners that spent on social media predicts the academic results with a better academic pass rate.

Other related papers refer estimating the learner engagement in [14] and evaluate if there are patterns of learner engagement consistent with grade levels and id the class subject matter. The results show that the learner's activity and patterns are consistent across grade levels and the subject of the classroom influence directly the learners' engagement and because of that we considered only one course to analyse the learners because we needed a good control regarding the procedure. Another factor took into consideration that may influence the results presented in the paper is the learner's' background addressed by Páivi M. Tikka et. al [15]. The paper analyses if the learner's background influence the learner's activity along with other important factors like learners' attitudes and knowledge concerning the environments. The results showed that there are major variations among learners between genders or educational background; regarding our paper the learners had a similar academic and high school background and the study environment was consistent.

Regarding the learner modeling and machine learning [16], Chien-Sing Lee et. al. [17] published a paper that involves PCA and self-organizing map (SOM) clusters in order to produce better learner models. They applied two techniques and compared them in order to provide meaningful analysis and class labels for learner's clusters. The first technique used PCA and the second one involved a two-level clustering: a SOM clusters [18] at the first level and PCA at the second, so PCA was involved in both of the techniques. The experimental results revealed that the second approach, that combines SOM clusters with PCA improves the quality of cluster analysis.

Previous research in this area aimed to find "*spam*" and "*don't care"* users [1] was made on less instances (285) by using classification algorithms, and more exactly decision trees. In this paper we used a classification approach to find problematic users based on their activity level but without knowing what the exact value of activity is, the ranking of a learner among his colleagues or having a comparative scale. In the current paper, we extend the study on more learners and we propose an extended generic solution to estimate the activity level and also offer an intuitive overview regarding their activity.

We choose PCA because it was reported among the most used dimensionality reduction techniques [19] in the mean-square error sense [20, 21]. More recent research compares PCA, KPCA and ICA for dimensionality reduction [22] and the paper presents that SVM algorithm performs better with PCA then without it.

Isomap [23] is another machine learning algorithm used for dimensionality reduction that extends classical multidimensional scaling by using geodesic distance instead of Euclidean distance. Because the geodesic distance matrix can be interpreted as a kernel matrix, Isomap can be solved by a kernel eigenvalue problem.

Regarding the use of PCA and other dimensionality techniques, in 2009, Laurens van der Maaten et. al. [24] published a review in which referred both convex and non-convex dimensionality reduction techniques. The paper aimed to address the limitation of classical dimensionality reduction techniques like PCA and other newer techniques by describing them and testing on several datasets which are divided in two main categories: natural and artificial. The presented results showed that on synthetic datasets, PCA wins in four out of five cases but it may be outperformed by other algorithms in other cases. In the case of natural

datasets, the situation changes and PCA obtain the maximum score all five cases, so we can conclude that even though PCA may not be the best algorithm for our approach it is for sure close to the best result and is one of the most robust dimensionality reduction algorithms.

III. Proposed Approach

The proposed approach is based on a previous result [1] that created an accurate data model for identifying and deleting the "*spam*" and "*don't care*" users and obtaining the input dataset that is used for further processing. Thus, the input dataset consists of a set of instances (i.e., learners) defined by three sets of features, one set for each social tool (i.e., Twitter, blog and Wiki) and one target variable represented by the exam grade.

The data processing pipeline is presented in figure 1. The baseline model has been obtained as previous result in [1] and is further used for checking the impact on data and its ability of classify learners after computing and scaling the PC values. The PCs based classifier is compared with the baseline classifier in terms of their ability to correctly classify learners.

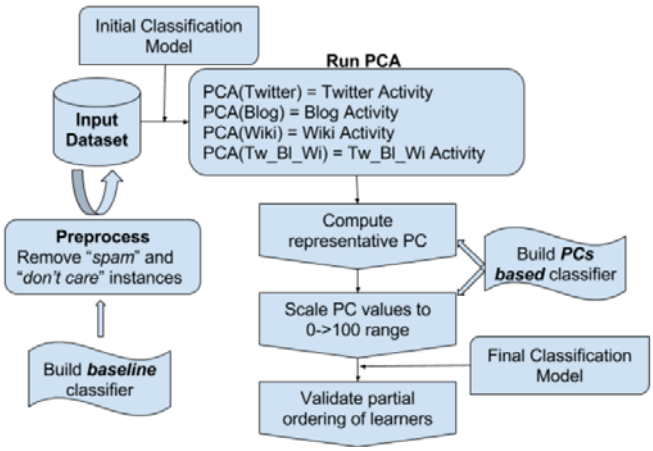

Figure 1. Data Analysis Pipeline

As PCA is used as dimensionality reduction technique, we need to make sure that along the data processing pipeline the newly computed features can also build a high-quality classifier. This approach is regarded as a "sanity check", such that in the situation in which the PCs based model has a high decrease in accuracy it would be a clear indication that processing of the initial input dataset has degraded the data quality in a large extent and thus it cannot be further used.

Computing the PCs is performed in two ways. One approach computes the PCs at tool level, reducing the number of features from the available ones at tool level to the most significant one. The obtained PCs, for a tool is further regarded as a digest of input features.

After obtaining the PCs from input features we choose the most representative one, such that it captures a high percentage of the variance in the input data. Figure 2 presents the general approach of PCA procedure that transforms the original raw *k* features from a tool (i.e., Twitter) to *k* orthogonal (i.e., independent) features. For example, if PC1 (the most representative component from *k* orthogonal computed components) component captures over 90% of variance from computed PCs it means that we may reduce the number of initial features from k down to one with a minimum loss of information.

From this perspective, we may be in the situation when summing over 90% needs more than one feature and that is why we need more PCs (i.e, two or more) to represent the input data. As a goal of our data analysis process we aim having only one PC as representative for each tool or even for all tools. This approach has the advantage of obtaining a single reduced value for each instance instead of k initial input features.

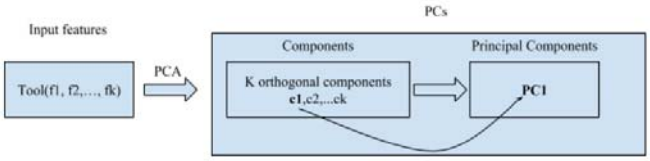

Figure 2. Dimensionality Reduction at Tool Level

As the percentage of the PC is higher we are more confident that the obtained value is better representing the initial input data. The main drawback of this approach is the poor interpretability of the computed PC value. In conclusion, this step has the advantage of computing a minimal number of features representing the data and the disadvantage of losing interpretability from initial features.

Another useful interpretation of results that may be performed at this step regards identifying the initial features that have the largest impact on obtained PC.

Coping with the lack of interpretability on obtained PC values has a large negative impact on further analysis. For example, given two instances with initial input feature values a domain expert may visually compare in an empirical way the activity values at feature level and interpret them in such a way that the most active learner can be assessed. From this perspective, obtained PC values for same two learners cannot be compared such that we cannot say which one had more activity. Another key issue is that we cannot interpret the PC values for obtaining an indication whether one learner or another had a small or large activity level on particular tool.

The proposed solution for tackling the problems of poor interpretability is to scale PC values to [0, 100] range in such a way that 0 *means "no activity"* and 100 means "*large activity*". The scaling is proposed by the following formula.

$$VALUE' = 100 - \frac{VALUE - \min V}{\max V - \min V} * 100 \qquad (1)$$

Where ***VALUE'*** represents the new value to be computed, ***VALUE*** represents the actual feature value, ***minV*** represents the minimum features' value from the dataset and ***maxV*** represents the maximum features' value from the dataset.

Usage of this approach orders learners in terms of their newly computed PC value and keep the values of original input features in relation with its initial PC value. We perform this transformation in two ways: visual and formal. The visual analytics picks several instances with low, average and high values in initial PC value and makes a

comparative analysis on original input features. Having in mind that a domain expert has a clear intuition on the meaning of the values of the original input features the partial ordering of learners in terms of their activity level can be performed visually.

We also prove by formal analysis of the entire dataset of learners' correlation between the actual performed activity values and corresponding scaled PC values.

For the original activity values, we consider a learner with zero activity on all features as the origin and compute the Euclidian distance from it to all learners. Intuitively, as the distance from origin to a learner is larger is an indication that more activity has been performed.

Definition: We define the "*origin learner*" to be the learner with all original feature values set to zero. This is a virtual learner that has not performed any action on any tool.

Definition: We define the "*activity distance*" to be the Euclidean distance from the "*origin learner*" to an actual learner.

$$ED = \sqrt{\text{twitter}^2 + \text{blogspot}^2 + \text{mediawiki}^2} \quad (2)$$

We compute Euclidean distanc (ED) e from origin to the learner by using as origin (0,0,0) coordinates all three computed activity values for twitter, blogspot and mediawiki.

D_i = Euclidean Distance ("*origin learner*", Learner_i)

This definition allows us to compare the actual performed activities of learners. Intuitively, if learner A is further from "*origin learner*" than a learner B it means that learner A performed more activity than learner B.

This approach allows us to obtain for each learner the value of the distance from "*origin learner*".We formally evaluate the correlation between distances from "origin learner" and scaled PC values.

TABLE I. INPUT DATA FOR ASSESSING THE CORRELATION BETWEEN D AND SPC

Learner	Original feature values	D_i	SPC
L_1	V_{11} V_{12} … V_{1k}	D_1	PC_1
L_2	V_{21} V_{22} … V_{2k}	D_2	PC_2
L_3	V_{N1} V_{N2} … V_{Nk}	D_3	PC_3

Where:
D_i = distance, from „origin learner
SPC = scaled PC value from original raw data

Table I presents the learners with their original feature values, their distance from "origin learner" and scaled PC values. We further evaluate the correlation between Di values and scaled PC values by plotting these values and fitting a regression line.

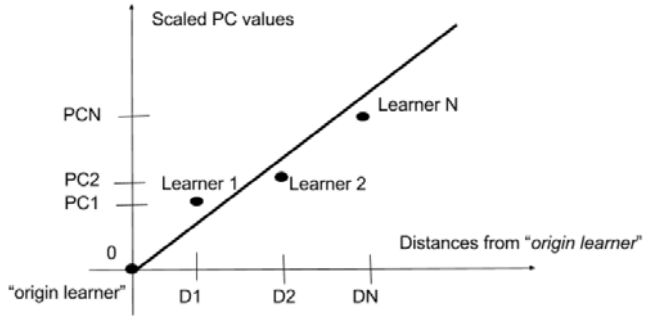

Figure 3. Correlation of distances from "*origin learner*" with scaled PC values

From this perspective, evaluating the correlation in terms of confidence interval validates the ranking obtained by the scaled PC values by assessing its quality in terms of partially ordering learners according with their performed activities. The worst-case scenario may occur when a large residual error is obtained, thus there is no correlation between initial raw data and obtained PC values which can be explained by a large information loss in the dimensionality reduction process. A positive correlation is a clear indication that original data may have been reduced with a minimal loss of information. A negative correlation would be also not satisfactory for the data transformation since this would mean that an increase in actual performed activity is regarded as a decrease in distance.

Finally, from classification perspective, we verify that the final obtained classification model has similar accuracy as the initial one. This validation makes sure that all performed data analysis steps have not significantly altered the initial model and thus can be further used with the same confidence. From data analytics perspective, the impact of running PCA and scaling is highly dependent on initial data quality. Running PCA should keep only one or maximum two PCs that highly represent the initial data, while scaling should make the values highly interpretable. The final two steps - correlation of distances with scaled PC values and rebuilding for evaluation the initial classification model - provide a sound evidence on the quality and interpretability of the obtained scaled PC values and associated classification model. At this step, the focus is not on obtaining a highly accurate model, but on the decrease in accuracy. Thus, a high decrease of accuracy - as well as a low correlation - is a clear indication that initial processing steps (i.e., PCA and scaling) have altered the data too much and the resulted features cannot be further used for classification in a reliable way. For improvements of the baseline classification model further processing may necessary regarding feature selection, outlier/noise reduction and proper fitting of the used classification algorithms.

The overall data analysis procedure makes sure that scaled computed PC values (at tool level or general) become interpretable and have a high degree of correlation with actual performed activities. Intuitively, a worst-case scenario may occur when scaled PC values represent in a poor way the input data and thus exhibit a low correlation with a small confidence interval. In this unfortunate situation, the scaled PC values cannot be used to represent as a digest the performed activity since the partial ordering is not valid and therefore other dimensionality reduction

techniques need to be used.

IV. Experimental Results

The experimental results were performed on 367 instances (i.e., learners) that used Twitter, blogspot and Wiki social media tools as learning environments at Web Application Design bachelor course. For the experiments we used the data gathered in 7 years, from 2010 to 2016 as Table II presents a sample of the features. The meaning of each feature along with its range values and sample dataset is presented in detail in [1] paper.

Table II. Sample Feature's list for Social Media Tools

Twitter features	Blogspot features	Wiki features
NO_TWEETS	NO_BLOG_POSTS	AVG_NO_ REV_PAGE
AVG_TWEET_ LENGTH	NO_BLOG_COM	NO_ACTIVE_ DAYS_WIKI
FRQ_TWEETS	FRQ_BLOG_POSTS	NO_ACTIVE_ DAYS_WIKI_REV
NO_ACTIVE_ DAYS_ TWITTER	AVG_RESPONSE_ TIME_ BLOG_COM	NO_ACTIVE_ DAYS_WIKI_FILES
NO_WIKI_REV	AVG_BLOG_POST_ LENGTH	NO_ACTIVE_ WEEKS_WIKI
NO_WIKI_FILES	AVG_BLOG_COM_ LENGTH	NO_ACTIVE_ WEEKS_WIKI_FILES
	NO_ACTIVE_DAYS_ BLOG	NO_ACTIVE_ WEEKS_WIKI_REV

The first processing step regards choosing the principal component for each subset of feature.

For the current study all the features are numeric and represent counts of performed activities. For example, NO_WIKI_REV represent the number of wiki page revisions and NO_WIKI_FILES represent the number of files uploaded on wiki.

Figure 4 presents the proportion of variance computed for each component. On the horizontal axis we have all the principal components computed for each tool (the number of components is equal with the number of dimensions) and on the vertical axis is presented the proportion of variance value. Thus, in Figure 4 (a), while in for other tools we have five and nine components, respectively. The maximum number of components is twelve and appears in the case of blogspot tool, the twitter has only five and in case of MediaWiki, we have nine. In Figure 4 a), the line represents the value of the proportion of variance for the Blogspot variable; most of the variance (0.9879) is represented by the first component and we don't need to consider other one. In the case of Twitter Figure 4 (b) the first component shows more than half of the variance, a cumulative proportion between first two components represents 0.9859 of the overall proportion of variance. In the last case, we have MediaWiki tool with 0.9454 of the variance is represented by the first component. Therefore, in one practical scenario, the principal component is represented only by the first component, and this is the one that is further used for obtaining an interpretation of the PCA reduction process. The goal of this analysis is to obtain the PCs that are further used in the data analysis process with a minimum information loss from original raw data.

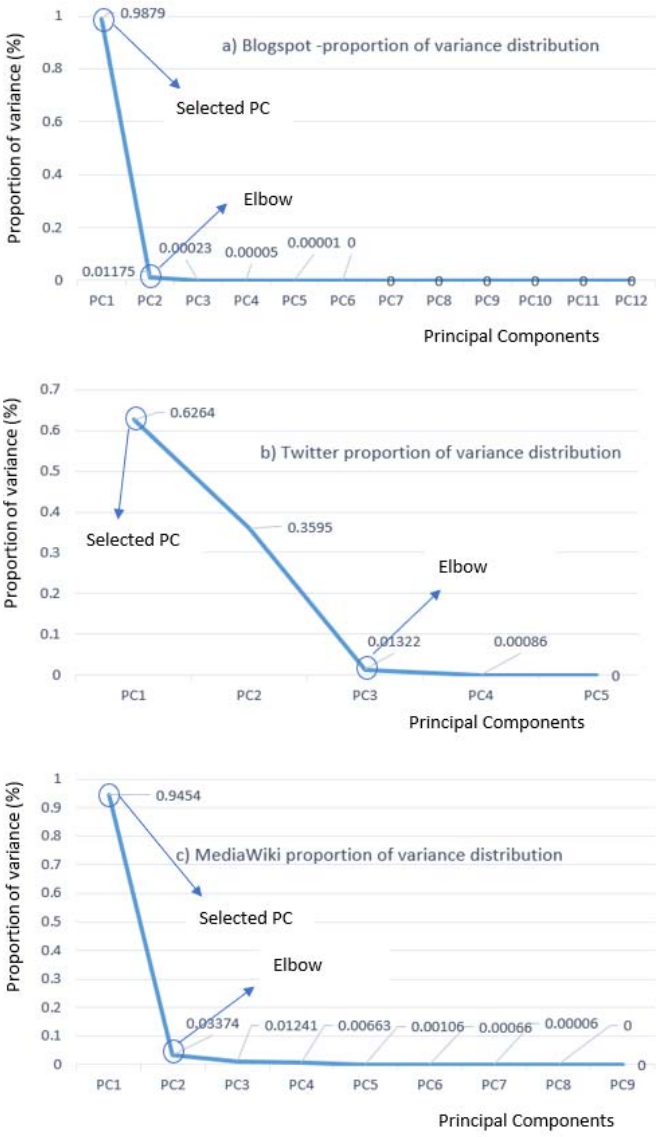

Figure 4. The proportion of variance distribution for each PC

Table III. Original and Scaled Values for Features and Corresponding Distances

Item id	Original PC values					Scaled PC values				
	twitter	blog	wiki	all	distance	twitter	blog	wiki	all	distance
209	0.90	1.81	2.86	3.39	3.50	10	3	0	5	10.44
207	2.32	0.32	1.74	2.25	2.92	0	13	9	11	15.81

206	-0.10	1.40	1.71	1.96	2.21	17	6	9	12	20.15
197	0.62	-0.33	1.92	1.20	2.04	12	17	7	16	21.95
195	0.96	-2.19	0.27	-0.74	2.41	10	29	20	26	36.62
196	-0.41	-0.17	-2.34	-1.73	2.38	20	16	39	31	46.66
200	-1.31	-3.63	-2.20	-4.36	4.44	26	39	38	44	60.34
208	-0.90	-0.93	-5.69	-4.56	5.83	23	21	64	45	71.18
199	-1.78	-3.13	-4.33	-5.48	5.63	29	36	54	50	71.08
204	-1.67	-4.62	-4.87	-6.75	6.92	28	45	58	56	78.57
198	-2.75	-3.24	-6.25	-7.04	7.56	36	36	68	**58**	**84.95**
203	-1.22	-6.29	-4.14	-7.13	7.63	25	56	52	**58**	**80.41**
201	-1.79	-4.52	-5.61	-7.24	7.43	29	45	63	59	82.67
202	-2.13	-5.17	-4.93	-7.36	7.46	32	49	58	59	82.40
205	-4.61	-11.29	-9.75	-15.48	15.61	49	89	94	100	138.41

Table III presents a snip from the reduced activity value. First column represents the learner's id from the dataset, the next three columns represent the originally computed PC activity values for each of those three tools, the fourth column presents the PC value obtained from all available features from all tools and the fifth column presents the Euclidean distance computed for them. The last five columns represent the same values but scaled at a scale from 0 to 100; 0 means no activity and 100 means the maximum activity achieved for that tool. These values present in an intuitive with a clear meaning the performed activities. The distance computed for the scaled values is also intuitive and has a range of values from 0 to 140 where 0 is the closest to the "*origin learner*" and 140 is the most further one.

As the learners are ordered in nondecreasing order of scaled all features column, we expect that the values in the last column (i.e., distance from "*origin learner*") to be also in nondecreasing order. Still, as we observe in the results, we may face the situation in which for the same - or even increasing - activity level we may have a decrease in distance. This situation may be observed clearly in figure 5 where, for the same - or slight increase - activity value we may obtain a slight decrease in in distance. Ideally, if such a situation would not occur it would infer a high quality fitted regression line with a close to zero error. The worst-case scenario occurs when the fitted line exhibits a large error with a very small confidence interval and thus many situations in which a significant increase in activity level is accompanied by a large decrease in distance from "*origin learner*".

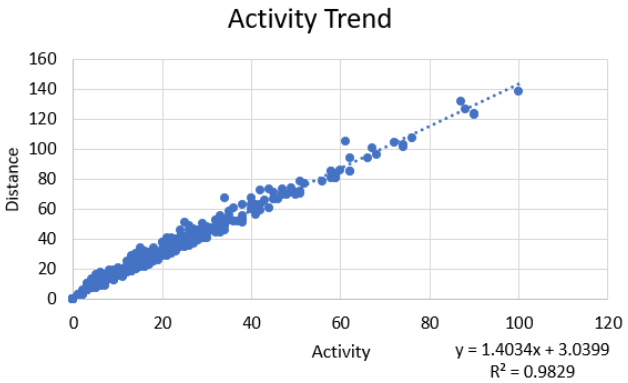

Figure 5. The activity vs. distance to "origin learner" trend

Figure 5 presents the computed activity trend based on the learners' activity computed for all features and the distance from origin. On the OX axis, we have the overall learner activity, and on the OY, we have the computed Euclidean distance from origin using the three values computed for each tool. Each blue dot is represented by an instance (i.e., learner) and we can see in the chart how close are to the trend line. Analyzing the chart, we can see that all the dots stay very close to the trend line and also follows the trend, most of them are close to the origin and as we go further they became rarer and a bit sparse because as they produce more activity, they are further from the origin. The value for intercept is 3.03, showing that the error for "origin learner" is small training into consideration the range values for distance which is from 0 to 140. The high positive correlation is also suggested by the large value of R^2, which is close to one.

This chart offers a visual analytics solution to validate our approach. A data exhibiting a higher variance (i.e., items that are not close to the trend line) would represent a sign for the lack of correlation. A large intercept value would be interpreted as predicting a large distance from "*origin learner*" for a learner that has not performed any activity, which is misleading in data analysis process. On the other

hand, a lower value in R^2 would be interpreted as a clear indication that the fitted line exhibits low correlation.

TABLE IV. ACCURACY OF CLASSIFICATION MODELS: INITIAL, WITH THREE PCS AND ITH ONE PC

Algorithm name	Initial features (no PCA)	With PCs (one for each tool)	With one PC from all features
Random forest	0,72	0,64	0,64
j48	0,78	0,67	0,67
c5.0	0,72	0,70	0,70
ctree	0,75	0,75	0,75
evtree	0,78	0,72	0,72
JRip	0,83	0,78	0,78
OneR	0,70	0,72	0,72
kk-NN	0,72	0,64	0,64

Table IV presents the results for a set of classification algorithms. The purpose of this analysis is to evaluate the impact of PCA reduction process in terms of classification accuracy. For each used algorithm there was computed the classification accuracy for a trained model on initial feature values (before PCA reduction), on PCs obtained for each tool, and on one PC obtained from all available features. The dependent variable is represented by the final grade, which is nominal and has three values: *low, average* and *high.* The results show that accuracy may remain the same (e.g., ctree algorithm) but in most cases it slightly decreases. In one situation (i.e., oneR algorithm) the accuracy slightly increases. Thus, we conclude that PCA reduction process has no significant impact in classification accuracy and thus, from this point of view the reduced data highly represents the original data. Although, the obtained classifiers may not be used in a practical context due to their low accuracy, the purpose of this analysis is aimed to obtain a clear indication regarding the decrease in data quality, not to train a high-quality classifier.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we have designed a data analysis mechanism that reduces the number of features with the goal of making them interpretable and reliable when building a classification model. The data analysis is based on PCA dimensionality reduction and scaling procedures. The validation of newly obtained data-set is performed by assessing the correlation between instances (i.e., learners' performed activity) and their Euclidean distance from "origin learner" and by comparing initial and final classification models (i.e., the one that uses the reduced and scaled features).

Further improvements in the data analysis process may take into consideration only the features that have larger impact in PCA dimensionality reduction process or employ other feature selection techniques. From this perspective, we could further take into consideration general categorical features along numerical ones, but this would need further factor analysis preprocessing. Other particular cases that may also be investigated regard the situation of binary variables or combinations of numeric and categorical variables. Depending on the used data types, it may be worth using other dimensionality reduction techniques in the same validation context and with the same goals regarding interpretability of feature values and high quality in obtained classification model.

Another key aspect of current work regards proving the correlation between newly obtained scaled and reduced feature values and Euclidean distance from "*origin learner*". Further studies may take into consideration other distance metrics (i.e., cosine, Minkowski, etc.) for computing activity distance or other correlation mechanisms (i.e., Pearson, Kendall, Spearman).

Thus, a well-planned, evaluated and validated usage of proposed data analysis process will lead to useful, interpretable and high-quality classification models in other application domains and for other machine learning tasks.

ACKNOWLEDGMENT

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS-UEFISCDI, project number PN-II-RU-TE-2014-4-2604.

REFERENCES

[1] M. C. Mihăescu, P.S. Popescu, E. Popescu, "Data analysis on social media traces for detection of "spam" and "don't care" learners", The Journal of Supercomputing, vol. 73, no. 10, pp. 1-22, 2017.

[2] A. Hervé, J. W. Lynne, "Principal component analysis", Wiley interdisciplinary reviews: computational statistics vol. 2, no. 4 pp. 433-459, 2010.

[3] W. K. Lim, K. Wang, C. Lefebvre, A. Califano, "Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks", Bioinformatics vol. 23, no. 13, pp. i282-i288, 2007.

[4] V. René, Y. Ma, S. S. Sastry. "Principal component analysis." Generalized Principal Component Analysis. Springer, New York, vol. 40, pp. 25-62, 2016.

[5] R. C. Radhakrishna. "The use and interpretation of principal component analysis in applied research", Sankhyā: The Indian Journal of Statistics, Series A, pp. 329-358, 1964.

[6] I. T. Jolliffe, J. Cadima. "Principal component analysis: a review and recent developments." Phil. Trans. R. Soc. A. 374, no. 2065, 2016: 20150202.

[7] T. J. Webster, "A principal component analysis of the US News & World Report tier rankings of colleges and universities", Economics of Education Review vol. 20 no. 3 pp. 235-244, 2001.

[8] M. E. Tipping and C. M. Bishop. "Mixtures of probabilistic principal component analyzers", Neural computation, vol. 11, no. 2, pp. 443-482, 1999.

[9] X. Moke, Y. Liang, W. Wu, "Predicting Honors Student Performance Using RBFNN and PCA Method", International Conference on Database Systems for Advanced Applications, Springer, pp. 364-375, 2017.

[10] D. Z. Dumpit and C. J. Fernandez, "Analysis of the use of social media in Higher Education Institutions (HEIs) using the Technology Acceptance Model", International Journal of Educational Technology in Higher Education vol. 14, no. 1, pp. 5, 2017.

[11] N. Marangunić, A. Granić. "Technology acceptance model: a literature review from 1986 to 2013." Universal Access in the Information Society vol. 14, no. 1, pp. 81-95, 2015.

[12] P. B. Lowry, J. Gaskin. "Partial least squares (PLS) structural equation modeling (SEM) for building and testing behavioral causal theory: When to choose it and how to use it." IEEE transactions on professional communication vol. 57, no. 2, pp. 123-146, 2014.

[13] C. Giovannella, F. Scaccia, E. Popescu, "A PCA study of student performance indicators in a Web 2.0-based learning environment", Advanced Learning Technologies (ICALT, 2013 IEEE 13th International Conference on Advanced Learning Technologies, pp. 33-35, 2013.

[14] S. C. Nsizwana, K. D. Ige, N. G. Tshabalala, "Social Media Use and Academic Performance of Undergraduate Students in South African Higher Institutions: The Case of the University of Zululand." Journal of Social Sciences vol. 50, no. 1-3, pp. 141-152, 2017.

[15] M. H. Marks, "Student engagement in instructional activity: Patterns in the elementary, middle, and high school years." American educational research journal vol. 37, no. 1 pp. 153-184, 2000.

[16] T. M. Páivi, T. K. Markku, M. T. Salla, "Effects of educational background on students' attitudes, activity levels, and knowledge concerning the environment." The journal of environmental education vol. 31, no. 3, pp. 12-19, 2000.

[17] A. Alireza, R. Lister, H. Haapala, A. Vihavainen. "Exploring machine learning methods to automatically identify students in need of assistance." Proceedings of the eleventh annual International Conference on International Computing Education Research. ACM, pp. 121-130, 2015.

[18] C-L. Lee, P.S. Yashwan, "Student modeling using principal component analysis of SOM clusters", In Advanced Learning Technologies, Proceedings. IEEE International Conference, pp. 480-484, 2004.

[19] P. Mangiameli, S. K. Shaw, D. West, "A comparison of SOM neural network and hierarchical clustering methods", European Journal of Operational Research vol. 93, no. 2 pp. 402-417, 1996.

[20] C. Girish, F. Sahin. "A survey on feature selection methods." Computers & Electrical Engineering vol. 40, No. 1. pp. 16-28, 2014.

[21] J. E. Jackson, “A user's guide to principal components”, John Wiley & Sons, vol. 587, 2005.

[22] I. T. Jolliffe, "Principal Component Analysis and Factor Analysis", Principal component analysis. Springer New York, pp. 115-128, 1986.

[23] L. J. Cao, K.S. Chua, W.K. Chon, H.P. Lee, Q.M. Gu, "A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine", Neurocomputing vol. 55, no. 1, pp. 321-336, 2003.

[24] H. Shi, B. Yin, Y. Kang, C. Shao, J. Gui, "Robust L-Isomap with a Novel Landmark Selection Method." Mathematical Problems in Engineering 2017. Vol. 2017, Article ID 3930957, pp. 12, 2017.

[25] L. van der Maaten, E. Postma, J. van den Herik. Dimensionality Reduction: A Comparative Review”, TiCC, Tilburg University, vol. 10, pp. 66-71, 2009.