Multivariate and Categorical Analysis of Gaming Statistics

Pin-Yu Chen*, Zhengling Qi†, Yanxin Pan†, Shin-Ming Cheng‡
*Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, USA
pinyu@umich.edu
†Department of Statistics, University of Michigan, Ann Arbor, USA
{qizl,yanxinp}@umich.edu
‡Department of Computer Science and Information Engineering,
National Taiwan University of Science and Technology, Taipei, Taiwan
smcheng@mail.ntust.edu.tw

***Abstract*—This paper provides exploratory analysis on gaming statistics via various multivariate and categorical data analysis approaches. The clustering results show that the principal components associated with the gaming data are related to player expertise and game camp characteristics. More importantly, the player level possesses only limited discriminant power for reflecting player expertise, and hence it indicates that gaming expertise classification is beyond player level. This paper therefore sheds new light on gaming statistic analysis, player expertise evaluation, and player level design.**

***Keywords*—exploratory data analysis, game statistics, multivariate and categorical data**

I. Introduction

Behavioral analysis on human activities is a fundamental research topic in social science [1]. Among various kinds of experimental studies, online video gaming turns out to be an appealing type of experiment that describes players' gaming strategies and real-time reactions [2]–[6]. In particular, gaming statistics are used for analyzing a player's gaming expertise and behaviors. Applications include but are not limited to gaming result prediction, player expertise inference, and data mining, to name a few.

The purpose of this paper is to apply commonly used multivariate and categorical analysis tools to the gaming statistics collected in [5]. Gaming behaviors were collected from some players of a game named *StarCraft 2* [5]. This dataset contains 3395 instances and 18 player attributes [7]. These attributes describe players' gaming behaviors and characteristics such as actions per minute (APM), action latency, age, play hours per week and so on. In addition, players are categorized into 7 expertise categories (gaming levels/rankings). Since the variables have different scale, each variable is standardized for multivariate analysis.

By investigating the gaming data, we are interested in player expertise category classification via dimension reduction and clustering techniques, especially in the the following aspects:

1) Can this dataset be well represented by a few principal components? More specifically, does there exists some principal components or attributes that can retain most variance of the original dataset? If so, how can we interpret these principal components?
2) Can we extract ordinal information by reducing data dimension via data approximation approaches?
3) Is there any polynomial relationship between variables? Should we include the quadratic term of variables before we performance dimension reduction?
4) Can clustering techniques reveal more informative clusters? How do the results compare to the original assigned gaming levels?
5) Is player level a good label for classification? If not, what are its shortcomings?

The results from the following data analysis tools are particularly addressed:

- Principal Component Snalysis (PCA) [8]
- Polynomial PCA [8]
- Non-metric Multi-Dimensional Scaling (non-metric MDS) [9]
- K-means Clustering [10]
- Random Forest [11]
- Support Vector Machine (SVM) [12]

II. Principal Component Analysis (PCA)

Using PCA, the eigenvalues of the sample covariance matrix are shown in Fig. 1. It is observed that there is a gap between the first two eigenvalues and the other eigenvalues, and the first two eigenvalues capture roughly 22% of total variance. By projecting the dataset onto the first two principal components in Fig. 2, we observe that the projection of first principal component (x axis) explains a player's gaming expertise in terms of his/her level. In other words, a player's level relates to the projection on the first axis. For instance, the players from level 1 and level 7 can be well separated by looking at first axis. Note that a player of low level does not necessarily imply that he/she has poor gaming expertise. It is likely that a player with outstanding gaming expertise belong to low levels simply because he/she does not have enough gaming records to be elevated to the next level. Therefore, it is relatively easy to distinguish a player from low levels (e.g.,

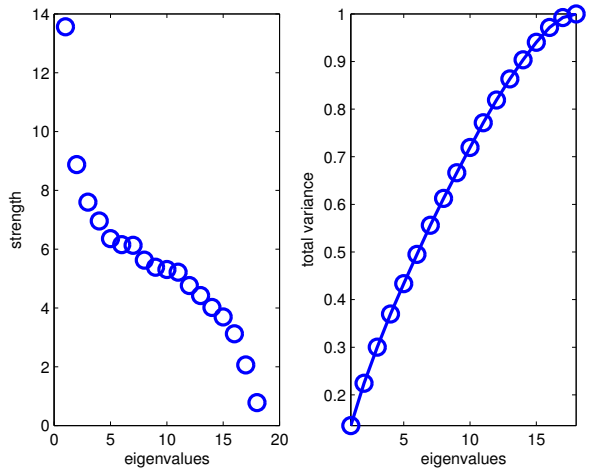

Fig. 1. Eigenvalue distribution and total variance using PCA. The first two principal components capture roughly 22% of total variance.

TABLE I. FIRST TWO PRINCIPAL COMPONENTS

Variables	PC1	PC2
Age	-0.1085	0.0855
HoursPerWeek	0.1301	-0.0927
TotalHours	0.0295	-0.0447
APM	0.3968	-0.2440
SelectByHotkeys	0.2768	-0.2405
AssignToHotkeys	0.2994	-0.0628
UniqueHotkeys	0.2224	0.0852
MinimapAttacks	0.1429	-0.0340
MinimapRightClicks	0.1692	-0.0720
NumberOfPACs	0.3590	0.1126
GapBetweenPACs	-0.3003	0.2056
ActionLatency	-0.3834	0.0487
ActionsInPAC	0.0987	-0.3645
TotalMapExplored	0.2205	0.4124
WorkersMade	0.1993	-0.0651
UniqueUnitsMade	0.1766	0.4628
ComplexUnitsMade	0.1702	0.3828
ComplexAbilitiesUsed	0.1548	0.3493

levels 1 and 2) and high level (e.g., levels 6 and 7), while it is difficult to distinguish a player from middle levels (e.g., levels 3, 4 and 5).

The high variance of the first axis in each level also indicates that the gaming expertise of players in each level may vary. However, the variance of the first axis decreases as the level increases, which means that the gaming expertise should be more consistent in high levels. By looking at the principal components in Table I, we have some interpretations for the first two principal components. The first principal component relates to gaming expertise based on player reactions. Some player reaction parameters such as Actions Per Minute (APM) and action latency have larger weights than others. These results can be explained by the facts that more actions and shorter action latency imply better gaming expertise, rendering the corresponding entry of the latter to be negative. In addition, age is shown to have negative impacts on gaming expertise, but the effects are minor since the corresponding weight is small.

It is interesting to see that the second axis does not play a role in distinguishing gaming expertise. The dominating

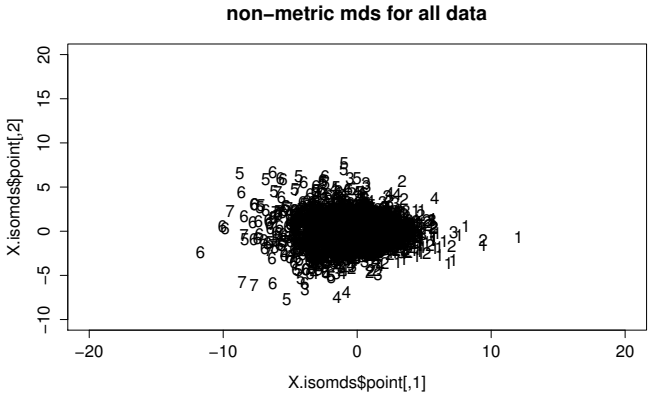

Fig. 3. Two dimensional data projection using non-metric MDS.

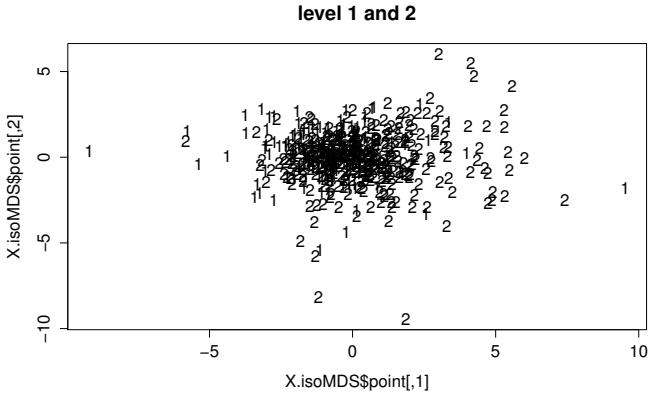

Fig. 4. Non-metric MDS on players from levels 1 and 2.

variables in the second principal components are the last three variables, which are associated with units production and gaming strategy (i.e., use of complex ability). The reason that the second principal component does not provide much information on gaming expertise is that for each game, a player is able to choose three camps. Each camp has different characteristics such as the unit production cost is low but lack of complex ability and short of attack damage, or vice versa. The indistinguishability of the second axis means that players in each level are composed of different camps, and the game setting is quite balanced in the sense that no camp has absolute advantage over other camps, otherwise a biased trend on the second axis should be observed.

III. NON-METRIC MULTIDIMENSIONAL SCALING (NON-METRIC MDS)

We first explain why metric MDS would not work in our case. When applying metric MDS, one cannot get a fair STREE index unless we choose large factor number. However, large factor number results in less interpretation of the dataset. Since the dataset is composed of players' gaming statistics in terms of their skills and reactions, non-metric MDS is adopted to capture the ordinal properties of the data [13]. In this report, three dimensional non-metric MDS is carried out, measuring dissimilarity by Euclidean distance on the standardized variables. The stress is 14.3%, which is fairly

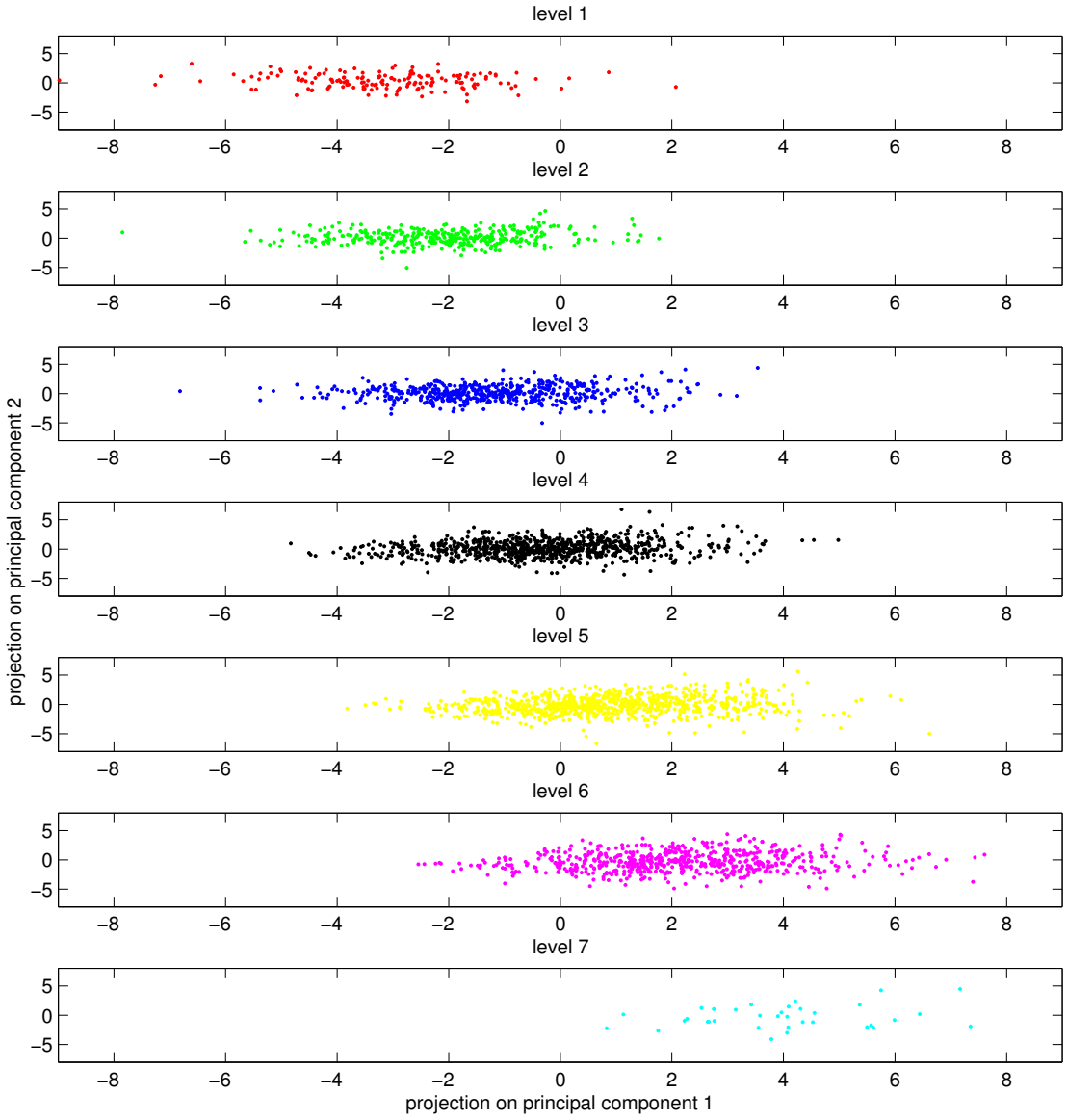

Fig. 2. Data projection on the first two principal components using PCA. The projection of first principal component (x axis) relates to gaming expertise based on player reactions, and the second axis (y axis) relates to gaming expertise based on units production and gaming strategy.

good. By plotting each players position in the configuration, we can easily interpret that the x axis represents the overall level of game players, where smaller score means higher level.

However, as mentioned in PCA analysis, it is relatively easy to distinguish a player from grouping some levels (e.g., levels 1 and 2, levels 3, 4 and 5, levels 6, 7). Therefore we apply non-metric MDS to each level group. The stress value of each one is below 15%, and the grouping of level 6 and 7 is less than 6%, which indicates that non-metric MDS gave good approximation of the data.

Looking at Fig. 3, the players position in each group with respect to their two dimensional scaling, we observe similar scatter pattern as in PCA. Figs 4, 5 and 6 display non-metric MDS on grouping players. For example, when grouping level 1 and 2, we can find that the first axis represents players overall level, where larger x value, means higher league index. For y axis, consistent with PCA, is explained by different camps characteristics. As for level 3, 4 and 5, as indicated before, it is hard to interpreted due to their highly overlapping nature, which is one of the hardest classification nowadays, and classification for highly overlapped clusters will be our future work for the second project. On the other hand, since the number of players in level 7 is small, we can find that

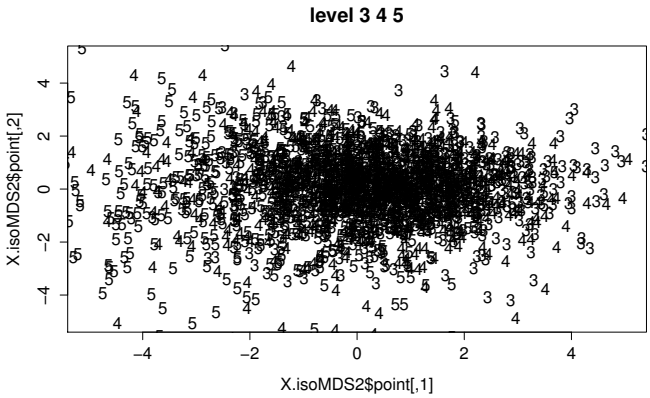

Fig. 5. Non-metric MDS on players from levels 3,4, and 5.

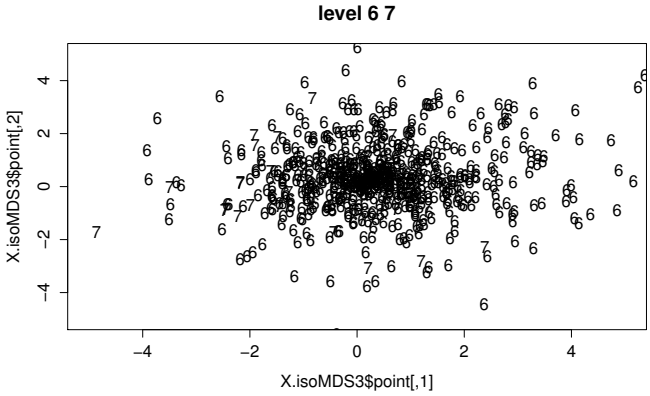

Fig. 6. Non-metric MDS on players from levels 6 and 7.

these players mainly concentrate at the righthand side of the figure, which means that these players are of higher overall level.

To sum up, non-metric MDS gives us a good low-dimensional data approximation by capturing the ordinal properties of the data. By grouping the players into three super-groups based on their levels, more structured and interpretable data patterns are revealed.

IV. Polynomial PCA

The intuition of using polynomial PCA lies in the fact that some variables are related to human physiological limit, such as Actions Per Minute (APM). Generally speaking, the differences between high level game players are less significant than the differences between low level game players. Consequently, variables in quadratic forms are expected to capture this relationship. By applying polynomial PCA to the dataset involving quadratic terms of variables, the first 10 eigenvalues are shown in Fig. 7. It is observed that there is a gap between the first two eigenvalues and the other eigenvalues, and the first two eigenvalues capture roughly 12% of total variance. By projecting the dataset onto the first two principal components in Fig. 8, we observe that the projection of first principal component (x axis) also explains a players gaming expertise in terms of his/her level. Similar to traditional PCA, The players

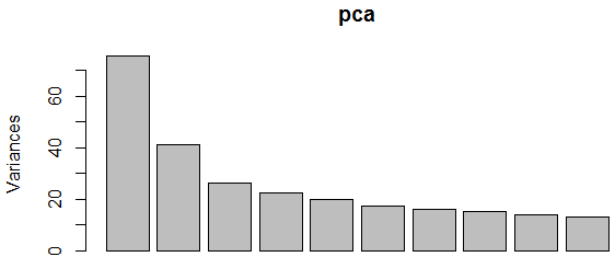

Fig. 7. First 10 eigenvalues using polynomial (quadratic) PCA.

from level 1 and level 7 can be well separated by looking at the first axis. However, it is still difficulty to distinguish players among middle levels (e.g., level 3, 4 and 5), since the variance captured by PC1 is not large enough.

Unlike traditional PCA, the second principal component in polynomial PCA also play a role in distinguishing gaming expertise. It can be seen from Fig. 9 that the variance of PC2 scores in higher level groups are much larger. The reason is that the PC2 describes the style of game player. PC2 can be interpreted as describing the style of a player. Those variables (including their quadratic forms) that characterize the skills of players have higher loadings, such as SelectByHotkeys, TotalMapExplored and ComplexAbilitiesUsed. However the TotalHours and HoursePerweek related variables has lower loadings. It can be seen from Fig. 9 that the variance of PC2 scores in higher level groups are much larger, which means that higher level game play are able to develop personal gaming style. For example some player may prefer to use hotkey when they do not have high Actions Per Minute (APM). Another reason is that when players get familiar with the game and having higher levels, they have better understanding about the advantages and disadvantages of different camps, especially in the aspects of unit production cost and complex ability. All in all, the advantage of polynomial (quadratic) PCA is that it can capture the nonlinear relationship between variables. Also quadratic PCA can use more information about data. The shortage of polynomial PCA is that this technique adding the number of original variables, which makes dimension reduction more challenging and less variance can be captured by the first few PCs.

V. K-means Clustering

We first apply K-means clustering technique to cluster the data points. The average within-cluster distance to the cluster centroids with respect to the number of clusters is shown in Fig. 10. A reasonable choice of the number of clusters is 20, where the average within-cluster distance enters the knee part (i.e., the point from which the average within-cluster distance has slow decreasing rate). However, 20 clusters are barely interpretable, and it also implies strong heterogeneity among data points of the same cluster. For interpretability, we cluster the players into two groups and analyze the Silhouette value in Fig. 11. It is observed that we have more strong agreement in group 2 where all Silhouette values are positive. On the

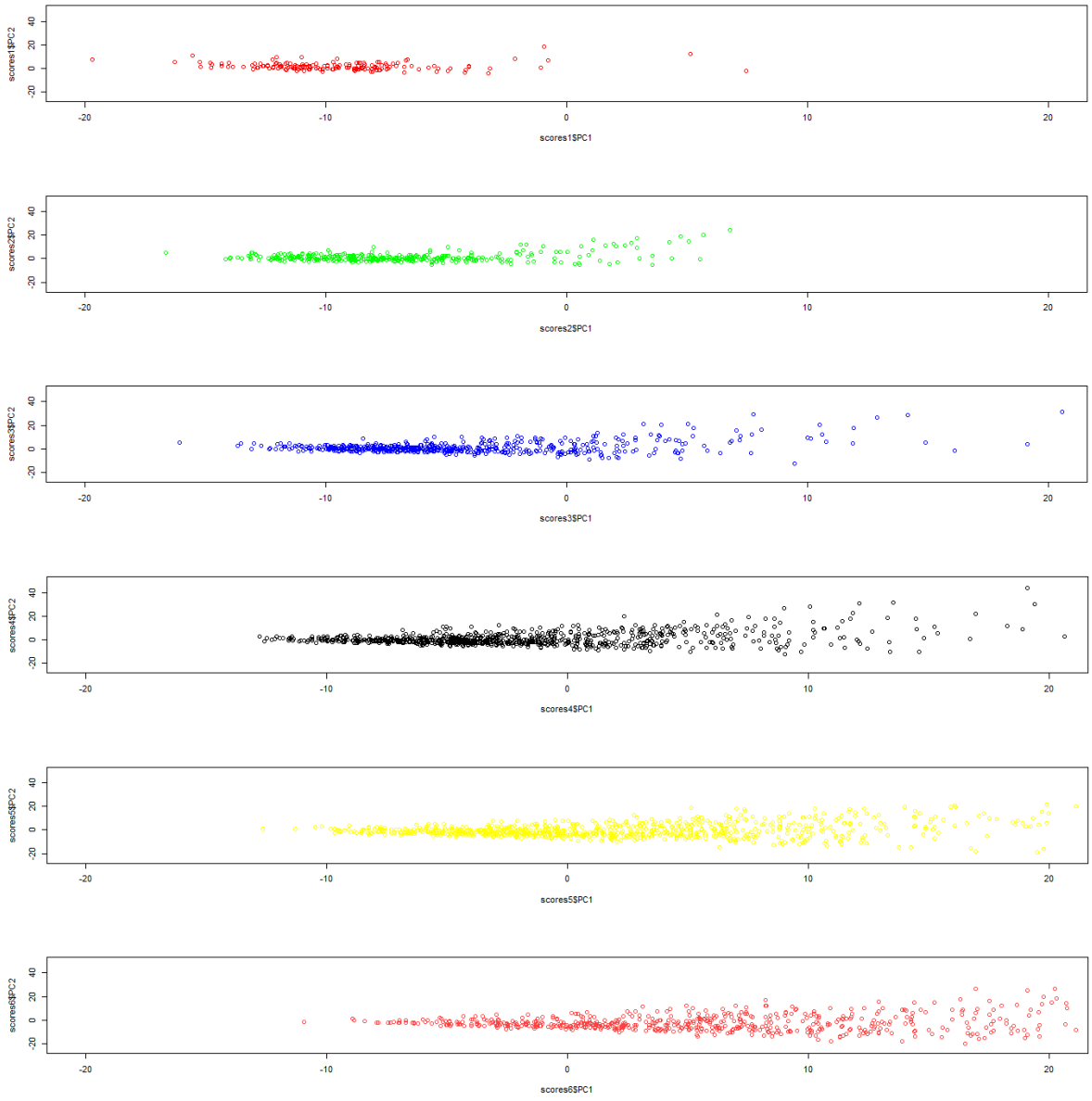

Fig. 8. Two dimensional data projection using polynomial PCA.

other hand, in group 1 almost half of the data points have negative Silhouette values, which shows apparent clustering inconsistency.

Fig. 2 displays two-dimensional PCA given gaming levels. It is clear that the data points of players of level 3,4, and 5 are overlapped since their gaming skills are nearly indistinguishable. Assigning the cluster labels using K-means clustering in Fig. 12, we also observe that the a great portion of data points in level 3,4, and 5 have opposite cluster labels. These results suggest that it is difficult to classify the gaming expertise for players in middle levels, whereas players in starter levels (levels 1 and 2) and expert levels (levels 6 and 7) can be better specified. Moreover, we are able to reach the conclusion that players in starter or expert levels tend to have more similar gaming statistics than players in middle levels.

VI. Classification using Random Forest

As we can see from PCA in the first report, the predictor variables are highly correlated. Since random forest is a useful technique that can deal with correlated predictor variables and easy to train and tune, we try to implement it to classify the players League Index according to the attributes of players.

The output is shown in Fig. 13. The out of bag error rate is more than 58%, which is not a good classifier. When observing the confusion matrix, we find that random forest classifies all

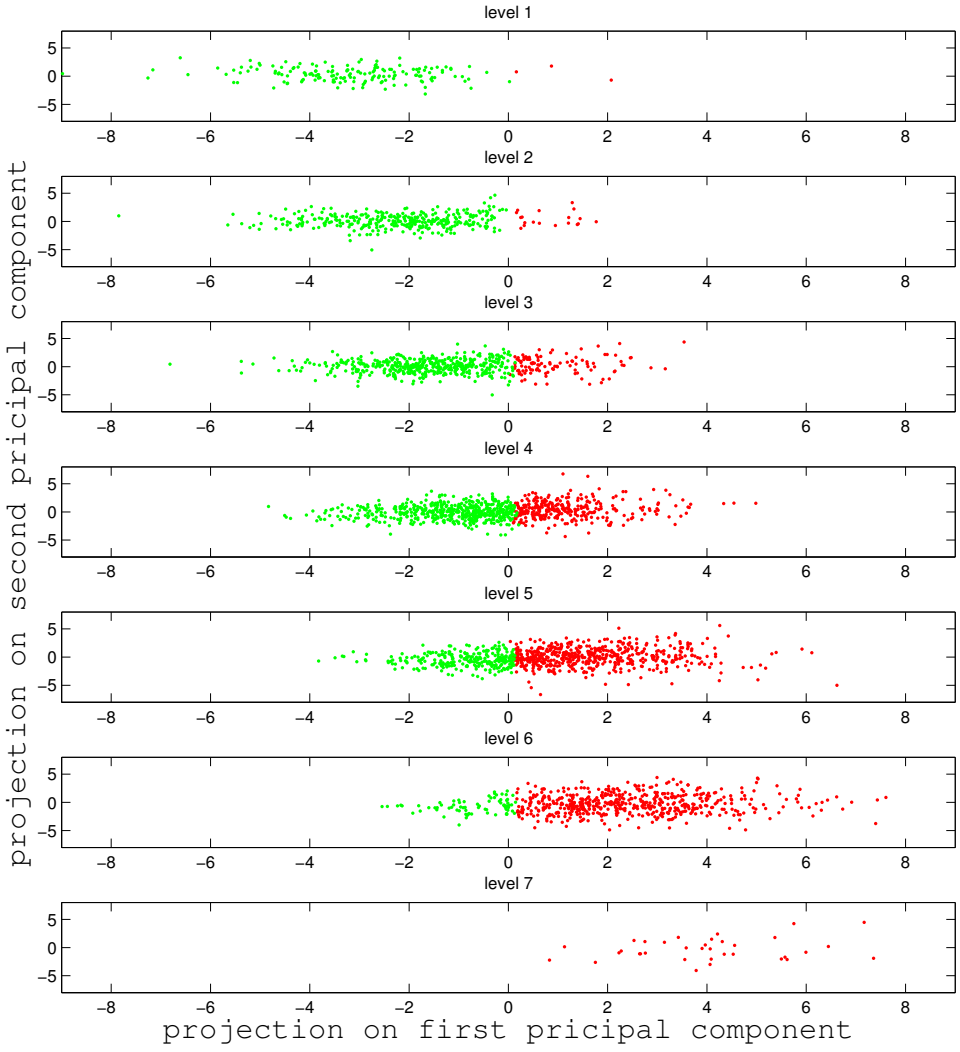

Fig. 12. K-menas clustering with 2 groups and data projection on the first two principal components using PCA. The projection of first principal component (x axis) relates to gaming expertise based on player reactions, and the second axis (y axis) relates to gaming expertise based on units production and gaming strategy (i.e., gaming camp options).

the cases in class 7 incorrectly because of small sample size. Class 2, 3, 4 , 5 and 6 are aggregated, which means that random forest cannot separate them. This can be further verified by Fig. 14 when we omit data points from labels 1 and 7.

From previous analysis, we find that it is likely that a player with outstanding gaming expertise belong to low levels simply because he/she does not have enough gaming records to be evaluated to the next level. Therefore, it may be more effective to classify players from low level (e.g., levels 1 and 2) and high level (e.g., levels 6 and 7), while the middle level is 3, 4, and 5. As shown in Fig. 15, the error rate is about 26%, which we can conclude that we can classify most of the cases correctly. When printing the importance for each variable and classes, we find some interesting result (shown in Fig. 16). The most important variable related to class 1 is the Total Hours Playing. The reason may be that players can increase their playing hours to get their League Index improved. Also, APM, ActionLatency, WorkersMade etc. are the most basic skills that player should improve themselves. On the other hand, for middle level, Total Hours Play becomes less important, but number of Perception action cycle becomes more important, which is a higher level technique. In high level, most of the technical skills including ActionLatency, MinimapAttacks etc. are relevant, which represents its highest level.

To sum up, we implement random forest to do classification on players level. Since there exist common phenomena that players do not have enough gaming records to be evaluated to the next level, we combine some level together to do further classification, which shows higher accuracy. By examining the importance of variables, we find some important variables

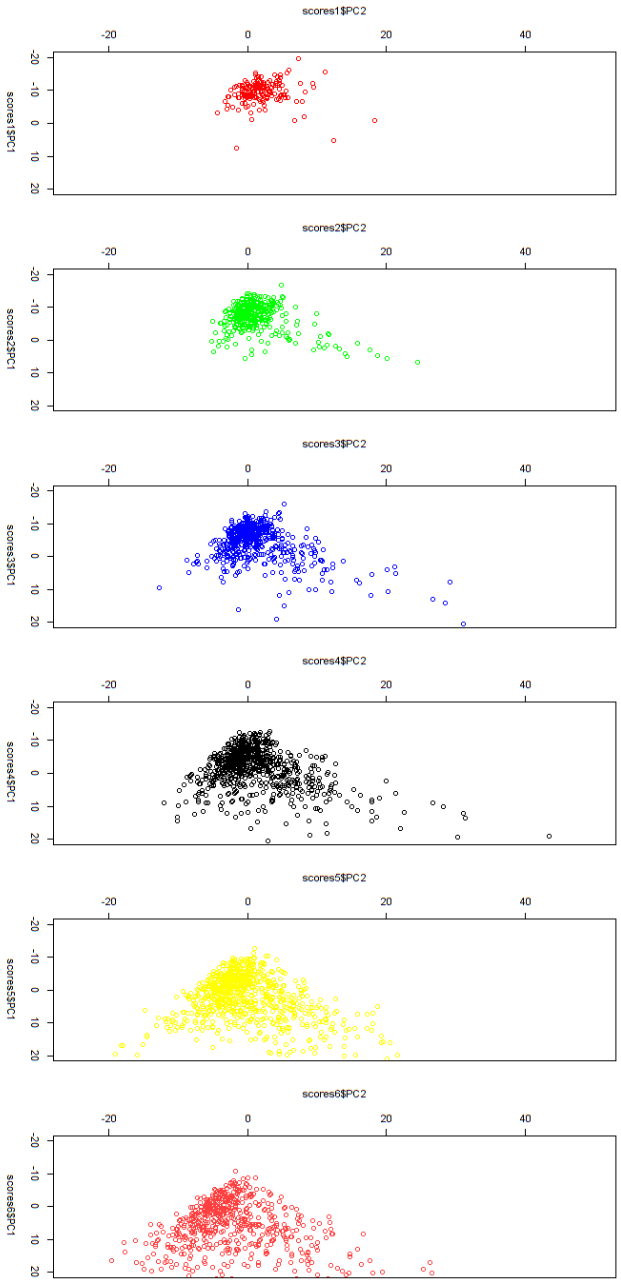

Fig. 9. Second principal component in polynomial PCA.

related to each class, which indicated what is the most importance skill to improve the level.

VII. Classification using SVM

The result of PCA and MDS show that the clusters of gaming data overlapping in low-dimensional projection. However, there is a more distinct difference between clusters in data projection by PCA. So it would be better to classify gaming data in the original data space or even higher data space. In this context, we apply classical and kernelized SVM to perform the classification task.

First of all, we classify the data by classical SVM and use cross-validation to get the accuracy of classification. Since

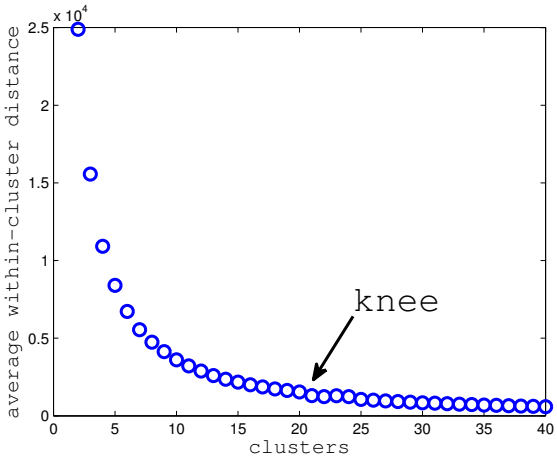

Fig. 10. Average within-cluster distance to the cluster centroids with respect to the number of clusters. A reasonable choice of the number of clusters is 20, where the average within-cluster distance enters the knee part (i.e., the point from which the average within-cluster distance has slow decreasing rate).

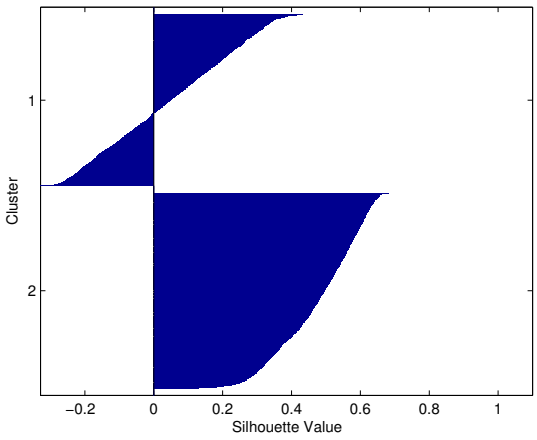

Fig. 11. Silhouette value of K-means clustering with 2 groups.

```
Call:
 randomForest(formula = LeagueIndex ~ ., data = input, ntree = 1000,      mtry = 4, importance = T)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 4

        OOB estimate of  error rate: 58.65%
Confusion matrix:
   1   2   3   4   5   6 7 class.error
1 50  63  37  16   0   1 0   0.7005988
2 32 102 114  89   9   1 0   0.7060519
3 19  69 166 229  63   7 0   0.6998192
4  6  40 137 377 204  47 0   0.5351418
5  1   8  39 235 340 180 0   0.5765878
6  0   0   6  49 220 345 1   0.4444444
7  0   0   0   0   2  33 0   1.0000000
```

Fig. 13. Confusion matrix for random forest.

```
Call:
 randomForest(formula = LeagueIndex ~ ., data = subdata, ntree = 1000,      mtry = 4, importance = T)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 4

        OOB estimate of  error rate: 55.31%
Confusion matrix:
   2   3   4   5   6 class.error
2 137 119  81   9   1   0.6051873
3  82 169 233  60   9   0.6943942
4  42 128 390 206  45   0.5191122
5   6  41 224 344 188   0.5716065
6   0   5  51 204 361   0.4186795
```

Fig. 14. Confusion matrix for random forest without labels 1 and 7.

the numbers of each class are different, we take it into

```
Call:
 randomForest(formula = LeagueIndex ~ ., data = new_data, ntree = 1000,      mtry = 4, importance = T)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 4

        OOB estimate of  error rate: 26.37%
Confusion matrix:
    1    2   3 class.error
1 227  287   0   0.5583658
2 103 1926 138   0.1112137
3   0  352 304   0.5365854
```

Fig. 15. Confusion matrix for random forest by aggregating gaming levels into 3 groups.

	1	2	3	MeanDecreaseAccuracy	MeanDecreaseGini
Age	-1.7985149	6.218988	6.2157170	7.411080	54.14380
HoursPerWeek	-5.5757141	4.678818	22.5183454	14.722317	63.17020
TotalHours	58.8748122	-3.356263	48.6927594	51.833569	142.36804
APM	38.0349124	20.151436	21.2597883	45.837361	174.62728
SelectByHotkeys	22.1171723	6.497315	27.5449160	34.198179	124.51382
AssignToHotkeys	29.4985732	-1.725799	25.2604715	28.255917	110.74291
UniqueHotkeys	0.6937983	-4.575996	23.5530140	10.624840	50.17590
MinimapAttacks	17.8046429	-5.916711	33.6229196	19.602588	93.39868
MinimapRightClicks	5.4209570	7.565492	3.4535942	10.441693	79.13885
NumberOfPACs	27.6309921	26.262350	31.7336182	51.935895	147.88485
GapBetweenPACs	26.0445873	-2.215446	37.4837615	36.565378	131.25055
ActionLatency	39.7603811	15.080935	41.9669325	54.582453	192.18544
ActionsInPAC	-2.4794030	24.855582	-1.4220234	22.320794	81.61790
TotalMapExplored	-8.1805844	13.784889	2.1556553	10.020715	63.94634
WorkersMade	21.1430599	2.845508	6.6885645	16.054375	89.09304
UniqueUnitsMade	-7.3163662	11.586986	4.2118510	8.890173	41.16932
ComplexUnitsMade	12.2106083	7.181147	-3.6215570	10.561953	29.52121
ComplexAbilitiesUsed	17.3029993	4.270216	0.3581305	12.293109	51.99496

Fig. 16. Importance for each variable and classes.

		PREDICTION						
		1	2	3	4	5	6	7
TRUE	1	15	21	9	8	1	0	0
LABLE	2	11	33	33	32	5	0	0
	3	11	18	37	83	26	5	0
	4	3	7	32	127	84	19	1
	5	0	2	9	86	118	64	0
	6	0	0	3	16	73	105	2
	7	0	0	0	0	2	9	3

Fig. 17. Confusion matrix using SVM.

consideration, and use the proportion of class as the class weights. The accuracy is only 40.53% for linear SVM and 41.52% for SVM with kernel sigmoid. Further observation of confusion matrix (shown in Fig. 17) shows that SVM usually misclassifies the individuals into similar class. For example 2 individuals whose League Index are 6 are misclassified into class 7, which is also a class of high level game players.

The inaccuracy can be explained by the fact that the gaming level might not be appropriately described by gaming capability. In other words, there are variables that are not indicators of gaming capability may influent the gaming level. For example, game player can buy fancy game equipment, which greatly affect the game level. With the help of fancy equipment, players are more likely to achieve higher game level.

In order to eliminate the ambiguity brought by factors rather than gaming capability, we merge the classes with similar gaming levels. Level 1 and level 2 are merged into new class, named low-level. Level 3, 4 and 5 are merged into middle-level class. The other levels are high-level class. Then the accuracy increased to 81.59% with radial kernel. The increase of classification accuracy also confirm the fact that the game level can only reflect the rough game capacity, which is more helpful when distinguish a player from low-level, middle-level and high level. To better reflect game players capacity, we need to construct new cluster structure as suggested in Sec. V, or assess cluster reliability for improved clustering performance [14], [15].

VIII. Conclusion

In this paper, we use various multivariate and categorical data analysis tools to analyze the relationships between gaming statistics and players' expertise level. It is observed that the first principal component can be interpreted as gaming expertise while the second principal component relates to camp characteristics. We also find out that the level indexes provided by the data results in highly overlapped clusters, which challenges the applicability of classification tools.

Furthermore, we verify that gaming statistics for players in middle levels have high variation and tend to overlap in low-dimensional projections using PCA, whereas players in starter and expert levels have much similar gaming patterns. We also show that the original 7 gaming levels are not fully representative, especially for middle level players. These results lead to high classification error rate for gaming expertise predictions, mainly caused by indistinguishability for middle level players. Nonetheless, it is relatively easy to cluster and classify a player as a starter or an expert by the gaming statistics.

IX. Acknowledgements

This work is supported in part by Minstry of Science and Technology, Taiwan, under contract MOST 103-2221-E-011-008-MY3.

References

[1] D. E. Hinkle, W. Wiersma, and S. G. Jurs, "Applied statistics for the behavioral sciences," 2003.

[2] J. Kirriemuir, "Video gaming, education and digital learning technologies," *D-lib Magazine*, vol. 8, no. 2, pp. 25–32, 2002.

[3] C. Kolo and T. Baur, "Living a virtual life: Social dynamics of online gaming," *Game studies*, vol. 4, no. 1, pp. 1–31, 2004.

[4] Y.-C. Chen, P. S. Chen, J.-J. Hwang, L. Korba, R. Song, and G. Yee, "An analysis of online gaming crime characteristics," *Internet Research*, vol. 15, no. 3, pp. 246–261, 2005.

[5] J. J. Thompson, M. R. Blair, L. Chen, and A. J. Henrey, "Video game telemetry as a critical tool in the study of complex skill learning," *PLoS ONE*, vol. 8, no. 9, p. e75129, 09 2013.

[6] E. Q. Yan, J. Huang, and G. K. Cheung, "Masters of control: Behavioral patterns of simultaneous unit group manipulation in starcraft 2," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 3711–3720.

[7] "Skillcraft1 master table dataset." [Online]. Available: http://archive.ics.uci.edu/ml/datasets/SkillCraft1+Master+Table+Dataset

[8] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[9] J. B. Kruskal and M. Wish, *Multidimensional scaling*. Sage, 1978, vol. 11.

[10] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Applied statistics*, pp. 100–108, 1979.

[11] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[12] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[13] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2005.

[14] P.-Y. Chen and A. O. Hero, "Phase transitions in spectral community detection," *arXiv:1409.3207*, 2014.

[15] ——, "Universal phase transition in community detectability under a stochastic block model," *Phys. Rev. E*, vol. 91, p. 032804, Mar 2015.