Analyzing for Patterns in the Cattell's 16 Personality Factors Dataset using Social Segmentation

EMINE YAMAN
Computer Science and Engineering
International University of Sarajevo
Hrasnička cesta 15, 71210 Sarajevo
BOSNIA AND HERZEGOVINA
eyaman@ius.edu.ba

ZAID ZERDO
Computer Science and Engineering
International University of Sarajevo
Hrasnička cesta 15, 71210 Sarajevo
BOSNIA AND HERZEGOVINA
zzerdo@student.ius.ba

Abstract:- Personality psychology has an essential prediction of behaviors. Cattell's personality theory is implemented to almost 49259 testers by asking 162 questions which occurred by 16 different question group types and testers answered these questions by using dataset for finding related pieces of information between these questions and forecast behaviors of the matter. These questions are established for answering the subject with appropriateness and some of these questions are based on the predicted questions for the same aim. Two social segments are completed by using the social segmentation with dividing the whole dataset with few mathematical equations and as a result, the general procedure is done. These two groups are compared for seeing the result in detail and which questions are mostly answered. The results show that different question groups play varied roles in how they correlate to other questiongroups.

Key-words: - Cattell's test; Psychological data; Preprocessing; Social segmentation

1 Introduction

The branch of personality psychology has always seen people as a construct that can be grouped into multiple personalities, then model and predict their behavior according to the type of their personality [6].

This paper attempt to predict somebody's behavior by the questions to which the subject answered with a high agreeableness factor. The whole model depends on the data that is supplied by an already completed test of 49159 testers.

Cattell's 16 Personality Factors test data is constructed on the theory that Cattell constructed in 1970 [4], while the questions for the actual test are also obtained from his handbook [5].

Hirschfeld et al. used the same data to select relevant items for a model of personality named Big Five, analysing the relevance of factor loadings that result in stabilization. [3] Other papers [1],[2],[10],[11] use associative rule analysis and high-dimensional data correlation to predict behaviour.

The dataset is a 49159 by 168 table, that has 162 questions divided into 16 groups, that test various universal psychological factors, such as extraversion, conscientiousness or feeling vs. thinking. Each question is answered by a number from 1 to 5, 1 being "I completely disagree", while 5 being "I completely agree".

Besides that, every single question group type contains ten or more questions each. The first few questions are positive to the attitude of the question type, while the remaining questions are negative to the attitude of the question, ex. if J question type is about the attitude about laws and rules, meaning that the first part

of the questions asks whether the subject has a positive attitude towards laws and rules while the second part is negative.

2 Research Method

2.1 Preprocessing

A lot of preprocessing has been done in order to only capture the essential data that can be used for prediction. From the starting 49159 rows the data was reduced to 11536 rows.

The first step was to homogenize the data and only select rows that are done by a tester from the US. This way the cultural impact does not play a role in prediction. US was chosen since most of the testers come from the US.

Since the dataset also contains the time it needed to finish with the test, all tests which are considered to be done too quickly or too slowly are eliminated from the data. In our case, a test is done too quickly if it was finished before ten minutes or done too slowly if the elapsed time is greater than 60 minutes.

Plenty of data was available at all times, so all missing data in the rows resulted in the removal of the entire row.

2.2 Prediction

Assume that a set Q_T contains a small number of questions to which the tester replied with a high level of agreeableness, i.e. the answer is either a 4 or a 5, predict set Q_P that contains answers to which the tester would also provide high levels of agreeableness.

The first thing that we do is construct two social segments, i.e. divide the whole data into two groups, S_0 and S_1. One of the groups, S_1 also has high agreeableness with the questions from Q_T, while the second group S_0 mostly disagreed with the Q_T set. The sets can be defined as following:

$$S_0 = \{t: \sum_{i=0}^{n} q_{ti} < 5\,\sigma\,|Q_T|\} \quad (1)$$

Where t is one the testers of the actual test. The value q_{ti} represents the answer to the question i by tester t, the test answer being $t \in [1,5]$. Coefficient σ represents the threshold for the summation of the answers to be considered for the S_0 set. The usual value for it is 0.75. Same applies for S_1 group:

$$S_1 = \{t: \sum_{i=0}^{n} q_{ti} \geq 5\,\sigma\,|Q_T|\} \quad (2)$$

With the only difference being in the range that is selected.

After segmenting the testers into two groups, the next step is to calculate their average coefficients for each question:

$$C_0 = \{q \to \frac{\sum S_0}{|S_0|} : q \in Q_T\} \quad (3)$$

$$C_1 = \{q \to \frac{\sum S_1}{|S_1|} : q \in Q_T\} \quad (4)$$

Where C_0 and C_1 represents the mapping of questions to coefficients. The next step is to calculate the difference map.

$$D_M = \{q \to C_0(q) - C_1(q) : q \in Q_T\} \quad (5)$$

The final answers that are chosen for the set Q_P is:

$$Q_P = \{q : D_M(q) > \varepsilon\} \quad (6)$$

Where threshold ε separates the values from predicted and non-predicted by the set Q_P.

A question is deemed positive or negative if:

$$Q_+ = \{q : q_{id} < threshold(q_{id})\} \quad (7)$$

$$Q_- = \{q : q_{id} \geq threshold(q_{id})\} \quad (8)$$

$$threshold = \{8,9,6,7,7,6,6,7,7,8,6,8,6,8,6,8\} \quad (9)$$

The threshold values are taken from the description of the questions. The value of $threshold(X) = Y$ represents the first question of type X that is contradicting the attitude of the type. If the value is 8, then the eighth question is the first one to contradict the attitude.

3 Results

Results show that the algorithm produces the prediction sets with credibility. Even though the algorithm results

are difficult to present, examples can show the credibility to some extent.

Given a question set that contains the following questions to which the subject answered with high agreeableness:

1. G10 – "I keep in the background"
2. F2 – "I try to follow the rules."

The algorithm produces the following table of predicted questions to which the subject should also show high agreeableness:

Table 1 - Coefficients for the given groups

Code	Questions	Coefficient
K3	I don't talk a lot	-0.9824
G9	I am quiet around strangers	-0.9567
K5	I keep my thoughts to myself	-0.7065
E8	I don't like crowded events	-0.6960
D10	I let myself be pushed around	-0.6113
L1	I am afraid that I will do the wrong thing	-0.5185

Although the table only shows a small portion of the predicted questions, it does show that the predicted set makes sense. The question G9 belongs to the same group as question G10, which is provided in the initial question set. The other three predicted answers belong to other groups, i.e. K and E, which are not present in the initial set. That indicates that the algorithm connects questions from multiple groups.

It is possible to test how each question group takes into consideration other question groups when predicting answers. The following graph shows how each question group predicts questions from its own group.

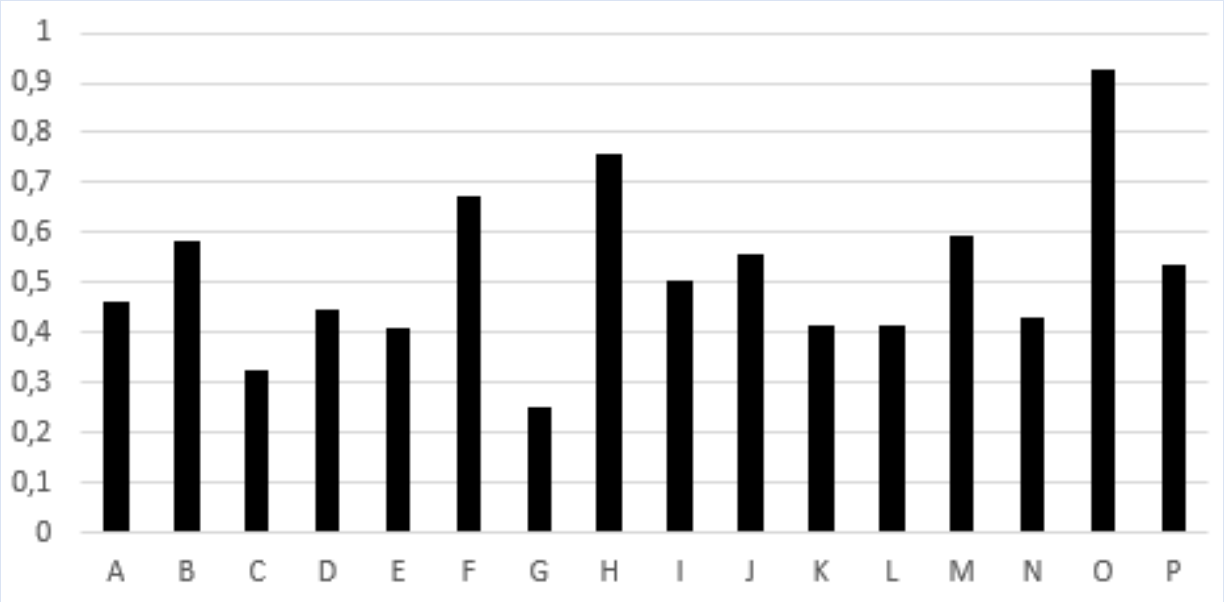

Figure 1 - Shows how each question group predicts answers from its own group

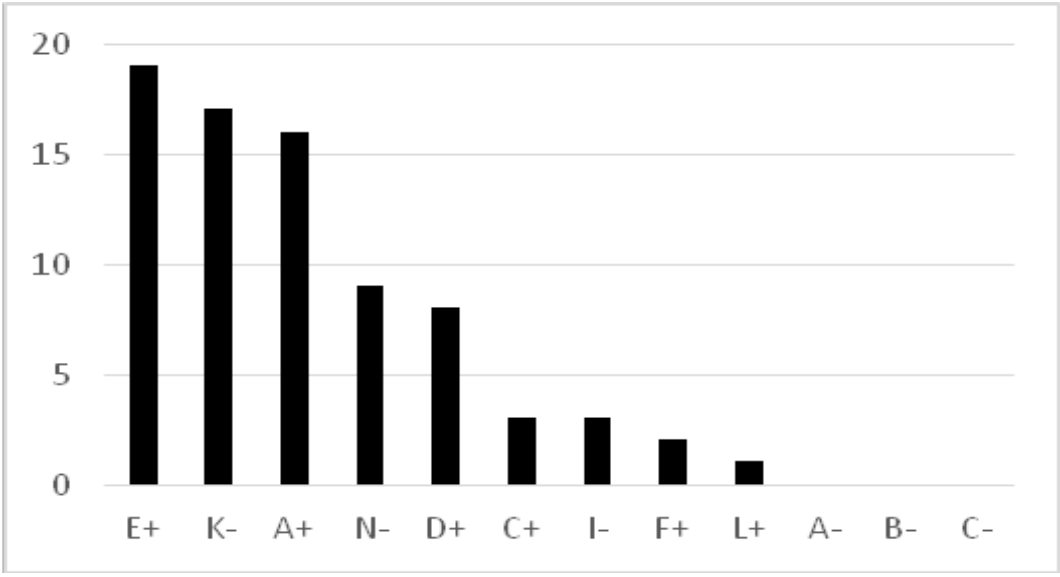

Figure 2 - Correlation of G+ questions with other question types

If we take into consideration two question groups like G and O, we can say that they heavily vary in the percentage of questions that they predict from their own group. Group O consists of question that ask the subject about order, conscientiousness and organization. It is interesting to note that almost all questions from this group only predict and correlate to the questions from the same group. Group G asks questions related to extraversion and social well-being and the graph predicts that whenever a question from this group is chosen, the predictions can be cast all around the place and include other question groups.

Since G question types have the most connections to other question types, I have decided to analyze that type further.

The previous figure shows the count of connections of G+ questions with other questions. Naturally, the most connections are with the same type, but in this graph, it is omitted.

Types of E+, K- and A+ show the most correlation with the G+ type. The G+ questions are connected to people who are open about their feelings, easily make friends and are quite extraverted.

Table 2 - Example questions for the given types

Type	**Question examples**
E+	“I am the life of the party”, “I joke around a lot”, “I act wild and crazy”, “I love large parties”
K-	"I disclose my intimate thoughts", "I show my feelings", "I am willing to talk about myself"
A+	"I know how to comfort others", "I enjoy bringing people together", "I cheer people up"

It does seem logical that the three groups are connected to the G+ group and we can see that the algorithm does produce results that make sense.

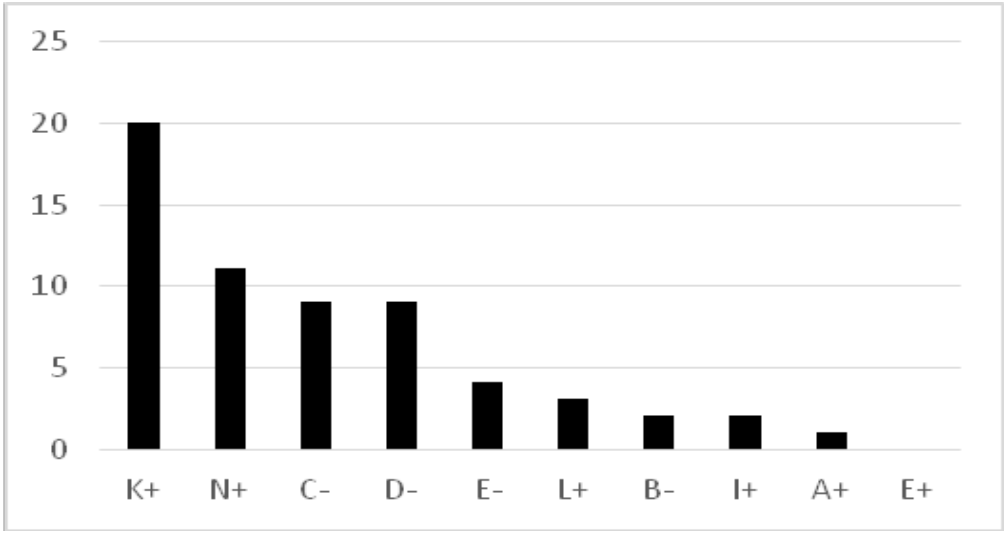

Figure 3 - Correlation of G- questions with other question types

The G- group shows large correlations with the K+ group, but also with N+, C- and D-.

Table 3 - Description of the given types

Type	**Question examples**
K+	"I don't talk a lot", "I keep my thoughts to myself", "I bottle up my feelings"
N+	"I want to be left alone", "I prefer to do things by myself", "I seek quiet"
C-	"I often feel blue", "I dislike myself", "I am easily discouraged"
D-	"I let myself be pushed around", "I wait for others to lead the way"

Table 4- A sample of the adjacency matrix for a subset of the questions of A type. Take note of the negative and positive values. The further the value goes into the negative, the more it is correlated to the question type, as explained previously in social segmentation. A8 question is the first question to negate the main theme of the question, as such, resulting in a positive value, rather than a negative.

	A1	*A2*	*A3*	*A4*	*A5*	*A6*	*A7*	*A8*
A1	**-1,96**	-0.73	-0,82	-0,66	-0,90	-0,82	-0,61	0,50
A2	-0,88	**-1,75**	-0,69	-0,78	-0,84	-0,71	-0,66	0,49
A3	-0,92	-0,65	**-1,97**	-0,77	-0,61	-0,58	-0,66	0,55
A4	-0,76	-0,76	-0,81	**-1,76**	-0,59	-0,55	-0,68	0,76
A5	-1,12	-0,87	-0,67	-0,63	**-1,69**	-0,93	-0,65	0,44
A6	-0,96	-0,70	-0,60	-0,55	-0,88	**-1,55**	-0,54	0,31
A7	-0,89	-0,79	-0,84	-0,84	-0,74	-0,66	**-1,82**	0,53
A8	0,48	0,37	0,45	0,59	0,33	0,24	0,35	**-1,95**

The given types do contradict the G question type and going well along the G- type.

If the social segmentation part of this paper is used to construct an adjacency matrix, then it is possible to apply various clustering algorithms on the data. The resulting coefficients can be used to map each question to every other, thus resulting in an adjacency matrix.

In order to symmetrize the matrix [7], [9], which is a needed step in further analysis, the following equation is used:

$$newcell[i,j] = \frac{oldcell[i,j] + oldcell[j,i]}{2} \quad (10)$$

The adjacency matrix results in a graph that can be used to analyze for clusters. One way to analyze is to use a community detection algorithm, such as Louvain [8].

A connection between two question types is considered to be as strong if $coeff < -0.55$.

Table 5 – A sample table that shows interconnections of A+, G+, E+, K- and B+ with E+, A+, G+ and K- question types. A higher number indicates that the question type of the first type has many interconnecting vertices with the other type. The resulting Harmony cluster can be deducted from this table.

X	*A+*	*G+*	*E+*	*K-*	*B+*
E+	9	27	25	1	0
A+	42	26	13	21	0
G+	16	20	19	17	0
K-	4	13	2	15	0

The given table is just a sample of a localized set of vertices that belong with each other, except vertex B+, which is given as a demonstration of a typical vertex that cannot possibly belong to the same cluster as the other types. It can be concluded that A+, G+, E+ and K- belong to the same cluster or community, since they contain many interconnections.

After further analysis, the highest connecting communities, along with the number of interconnections, are the following:

1. Depression, I = 412 (L+, C-, P+, I+); Cluster with four question types that indicate an individual with a need to disconnect from society. It is a combination of anti-social behavior, depression and low selfconfidence.
2. Harmony, I = 377 (A+, K-, G+, E+); A series of questions that indicate harmonious and extraverted behavior. Individuals that score high in this cluster are harmonic, social, understanding, fun and energetic.
3. Introversion, I = 292 (G-, N+, K+); An interconnection of question types that show a need to be alone. It is not as hardline as antisocial behavior, but it still does feel like social detachment.
4. Rebellious, I = 83 (J+, F-); A group of questions that, if scored high, indicate a personality that does not like rules and has no problems to break such rules. Authority is of no importance to these individuals.
5. Confidence, I = 74 (C+, L-); A small cluster that shows high self-esteem and self-confidence. Individuals that score highly in this group are generally quite bold and courageous.

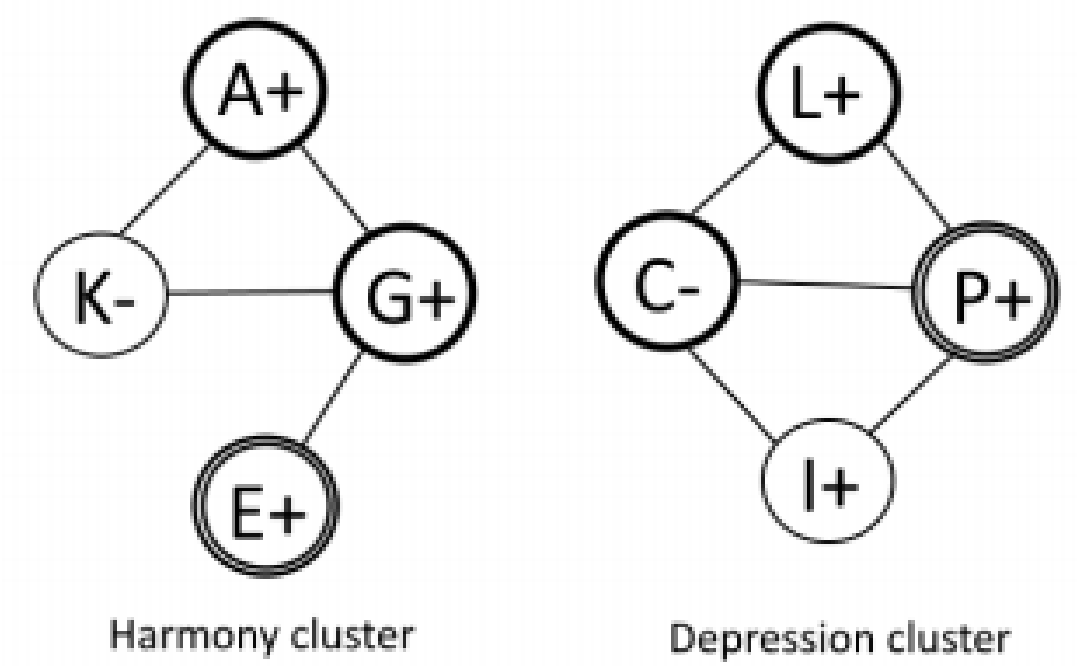

Figure 4 – Two of the most interconnecting communities and how each question type connects to another question. Bold lines indicate extremely interconnected question types, while the slim lines indicate types with low interconnections.

If we take into consideration the vertices that form the harmony cluster then we can see that they describe the following:

• A+ represents the ability to comfort others and understand their feelings.
• K- is the negative to introversion, indicating the ability to show understandable emotion to others.
• G+ indicates individuals that feel comfortable around others and don't mind being at the center of attention.
• E+ indicates the type of people that enjoy loud crowds and don't mind acting wild and crazy. It is interesting to note that this question type is loosely connected to the cluster, through G+.

The four-question type, when combined, result in a cluster we call harmony, since it describes people who are harmonious and comfortable with other people and their emotional states. Same analysis can be done with other clusters to see that the communities do make sense when combined with each other.

Besides the given five clusters, the other ones do not show enough interconnections to be considered worthy in analysis. All clusters with more than 50 interconnections are taken into consideration.

4 Discussion and Conclusion

Predicting behavior by social segmentation might be an interesting way to such predictions. By constructing two lists for the two social segments, the data is divided into two, thus making the differences between the two a lot more evident.

Even a relatively short list of high-agreeableness questions results in a large prediction set that provides questions that the tester might also see as something to which it is possible to agree.

Further work can be done to see whether other models, such as the Big Five, are compatible with this way of analysis. Factors such as Extraversion can be predicted by the given questions to see whether they all fit under the umbrella of that factor.

Same applies for the MBTI model. Cases of the 16 personality types can be used to predict more behavioral patterns according to the model. Given agreed questions such as "I let others' to lead the way", "I keep my thoughts to myself" and "I know how to comfort others" are a classic indication of the ISFJ personality in the

MBTI model. Predicting more behavioral patterns might be useful.

A new type of test could be constructed in the same manner. The tester might be provided with a small set of random questions to which the tester chooses only one. The algorithm then predicts other possible questions that the tester might agree to. The second iteration would contain a set of random questions as well as predicted questions predicted by the algorithm. This would also represent an extremely efficient way of testing whether the algorithm works.

Further analysis can be done on the relevance of the question type groups on other question group types. This has been done in different ways [1], [3], but not in this way. The algorithms perform efficiently when it comes to performance, indicating that it can go through a lot more data without wasting too much time. This would mean that the inclusion of more data from the preprocessing step can be left, ex. instead of removing all rows with missing values, just leave them there.

Clustering analysis can be done on the questions or the question group types to see what kind of behavioral patterns are connected to each other. The distance between the points can be estimated by a similar way that the Figure 1 graph shows.

The resulting clustering analysis, through community detection, can identify general and global behavioral patterns. The most interconnecting communities in this paper are the harmony, depression, introversion, rebellious and confidence clusters.

Acknowledgements

The author would like to thank Prof. Klimis Ntalianis for his helpful advice on various technical issues examined in this paper

REFERENCES

[1] G. Ver Steeg, A. Galstyan, "Discovering Structure in High-Dimensional Data Through Correlation Explanation." *Advances in Neural Information Processing Systems*, 2014.

[2] S. K. Perwez, H. M. Zubahir, M. R. Ghalib, K. Ahmed, M. Iftekar, "Association Rule Mining Technique for Psychometric Personality Testing and Behaviour Prediction." *International Journal of Engineering & Technology* 5.5 (2013): 4349-4361.

[3] G. Hirschfeld, R. Brachel, M. T. Thielsch, "Selecting Items for Big Five Questionnaires: At What Sample Size Do Factor Loadings Stabilize?." *Journal of Research in Personality* 53 (2014): 54-63.

[4] R. B. Cattell, H. W. Eber, M. M. Tatsuoka, "Handbook for the sixteen personality factor questionnaire (16 PF): In clinical, educational, industrial, and research psychology, for use with all forms of the test". Institute for Personality and Ability Testing, Champaign, 1970.

[5] "The Items in the 16 Preliminary IPIP Scales Measuring Constructs Similar to Those in Cattell's 16 Personality Factor Questionnaire (16PF)". http://ipip.ori.org/new16PFKey.htm

[6] S. Mark, K. Deaux, "Personality and Social Psychology." The Oxford Handbook of Personality and Social Psychology, ISBN: 9780199364121, 2012.

[7] I. S. Jutla, L. G. S. Jeub, P. J. Mucha, "A Generalized Louvain Method for Community Detection Implemented in MATLAB," http://netwiki.amath.unc.edu/GenLouvain (2011- 2016).

[8] P. J. Mucha, T. K. Richardson, K. M. A. Macon, M. A. Porter, J. P. Onnela, " Community Structure in Time-Dependent, Multiscale, and Multiplex Networks", Science, 328 (2010), pp. 876–878.

[9] B. Norman, "Algebraic Graph Theory", Vol. 2. Cambridge: Cambridge university press, ISBN: 0521458978, 1974.

[10] A. Pentland, A. Liu, "Modeling and Prediction of Human Behavior", Neural Computation 11, 229-242 (1999).

[11] J. Evermann, J. R. Rehse, P. Fettke, "Predicting Process Behavior using Deep Learning", Decision Support Systems, vol.100, 129-140(2017).