Datamining Techniques 2018 Advanced Project

Changxin Miao 11853018 Cmo440, Michael Mo 10770518 Mmo740, and Féliciën Veldema 10739335 Fva350

University of Amsterdam

Abstract. This paper describes the first advanced project of data mining techniques by modeling mood from data of the previous days.

Keywords: Data mining, mood, decision tree, ARIMA· SARIMAX

1 Introduction

In this report we investigate whether provided user activity data could be used to predict the user's mood. We achieve this objective by training a temporal and non-temporal model on the data. Afterwards we compare and conclude which model performs better. Furthermore, we also provide insights into the effect of different variables in predicting mood, which could lay the foundation for physiologists and physicians to conduct advanced research.

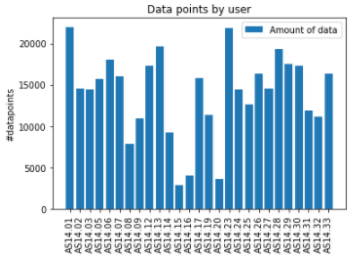
2 Data preprocessing

2.1 The Data

The provided dataset consists out of user data on their mental health. With a mobile application the users indicate their daily mood while their phone activity is monitored. There are 19 possible variables in the data set with their own range values. Mood is the important value we would like to predict by day and it's value ranges from 0 to 10. Circumplex arousal and circumplex valence both range from minus two to two and both model the James Russell emotional classification model [1]. Other inputs measure the activity, time spend on an app in seconds and the amount of times an action is performed, such as calling or texting.

2.2 Data transformation and analysis

The original data has columns indicating the user id, time stamp, variable and the value. The first processing step is to split the time stamp column into date and time. This is done to ensure that grouping on day becomes possible which will help to predict the mood of the following day.

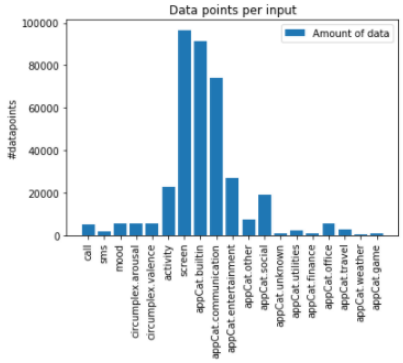



```
mood                                  circumplex.arousal                    circumplex.valence
Amount of data inputs:  5641          Amount of data inputs:  5643          Amount of data inputs:
             value                                 value                                 value
count  5641.000000                    count  5597.000000                    count  5487.000000
mean      6.992555                    mean     -0.098624                    mean      0.687808
std       1.032769                    std       1.051868                    std       0.671298  ...
min       1.000000                    min      -2.000000                    min      -2.000000
25%       7.000000                    25%      -1.000000                    25%       0.000000
50%       7.000000                    50%       0.000000                    50%       1.000000
75%       8.000000                    75%       1.000000                    75%       1.000000
max      10.000000                    max       2.000000                    max       2.000000
```

Fig. 1. (Left) Data points per patient. (Right) Descriptive statistics of the subset.

The statistics from the data shows that there are 27 unique users with varying amounts of data entries associated with them. Figure 1 on the left, shows the amount of data points for each user. Keeping in mind that there are 18 unique possible inputs per day, excluding mood, we see that user $AS14.01$, $AS14.23$ and $AS14.13$ deliver the most amount of data. Users $AS14.15$, $AS14.16$ and $AS14.20$ provide the least amount of data.

To further process the data, the amount of valid inputs are compared with the total amount of data points per variable. Within circumplex arousal and circumplex valence there are 46 and 156 Not Applicable (NA) entries respectively. This is seen in Figure 1 on the right.



	id	call p-val	activity p-val	circumplex.valence p-val
0	AS14.01	0.346188	0.061622	1.274928e-11
1	AS14.02	0.574067	0.114649	2.143439e-08
2	AS14.03	0.759739	0.989526	2.328591e-05
3	AS14.05	0.210856	0.136632	1.875132e-19
4	AS14.06	0.712663	0.932283	9.594017e-08
5	AS14.07	0.917852	0.035707	3.483144e-16
6	AS14.08	0.308922	0.265304	1.143010e-16
7	AS14.09	0.415889	0.000029	1.600325e-13
8	AS14.12	0.214285	0.681544	1.635620e-10
9	AS14.13	0.929014	0.077460	3.114415e-14
10	AS14.14	0.518193	0.902835	7.950326e-09

Fig. 2. (Left) Statistics of the frequency of the input variables. (Right) A subset of the correlation p-values.

Then new data frame is created where the input values are ordered by day and user id, resulting in one value per input per day. Mood and the circumplex values are averaged over the day per user. For the activity variable the maximum is taken per day while the other variable inputs are summed per day for all the users. This downscales the original 376710 row data to a dataframe of 15520 rows.

We were able to analyze the frequency of the inputs, displayed in the left of Figure 2. It shows that the inputs screen, appCat.builtin and appCat.communication are highly frequently used. Subsequently, we investigate correlations by calculating the correlation coefficient and p-values between the variables and the mood. The p-values yielded the result that only the circumplex valence variable is highly correlated to the mood of all users. Other variables are not significant correlated to the mood with the resulting values dependent on the user. Circumplex valence correlating with the mood makes sense as it is a scale of pleasure and displeasure [1]. A subset of the correlation p-values is displayed on the right of Figure 2. To summarize, circumplex valence is a potential feature to predict the mood of the user.

2.3 Data preprocessing for the non-temporal data

For the non-temporal model, we continue processing the data so that it becomes an instance-based dataset. Hereby each instance consists out of the target (mood at time t) and some corresponding features. The features are created by summarizing information of the table at a number of previous time steps. To maximize the number of data instances which can be created as well as having enough temporal history available, only information of three previous time stamps are considered. This also means that to create a single instance, data for four consecutive time steps are needed. If for any sequence of four time steps a missing value for mood is encountered, no data-instance will be created. Any missing values for circumplex arousal or valence are estimated as 0 (neutral mental state), and all other variables are set to a default value of 0 (count / time). The mood variable is expected to be correlated with that of previous days. So the first four features consist of: previous mood, average mood, and estimates for first and second derivative of mood in the past three days. Variables modeling mental state are also expected to have influence, so circumplex arousal and valence make up the next two features. A high number of total number of sms and calls could signal the occurrence of some event. These are the next two included. At last the maximum time spent on an mobile application of travel, finance, weather, office and the average of screen time and max activity scores are used since their deviation across the day varies a lot, and thus might cause a change in mood. Including those gives a total of 14 features. The resulting dataset consists of 1127 data instances covering all patients.

2.4 Data preprocessing for temporal data

Time-series models will be hand-crafted for each patient, but it is not feasible to create models for each patient. Therefore, we select two patients (AS14.01 and AS14.23) with highest amount of data records to construct customized ARIMA and SARIMAX models. Two extra data frames consisting of information for the two patients are created, with 46 and 44 observation respectively. **NAs** unavoidably occur in the resulting data frames, since different variables

are recorded for a specific day. In order to construct a temporal model which perceives the maximum amount of information, we deal with NA by interpolate it as the mean of the previous and next day. For the ARIMA model, there is only one endogenous variable-mood and its previous lags. For SARIMAX model, the relevance of variables will be identified by the granger causality test. We run the pairwise granger causality test for potential each exogenous and mood. Results indicate that only $circumplex.valence$ is granger caused by $Mood$. To identify more variables, the inversed relationship is also examined. In the end, $activity, appCat.entertainment, appCat.finance, circumplex.arousal, circumplex.valence$ All remained variables were treated as potential exogenous variables in the SARIMAX model. Final features will be chosen as by evaluation metrics of the SARIMAX model.

3 Model creation

Mood value of the previous day will be utilized as a benchmark model.

3.1 Non-temporal model

The features in the instance-based dataset for the non-temporal model still have interpretable meaning attached to them. As predictive model we thus opt to choose a Decision Tree (DT) regressor. The DT will predict by splitting on values of any of the available features. This thus naturally keeps the whole prediction process interpretable, where the exact reasons for prediction can be read from the DT itself.

Decision Tree regressor The implementation for the decision tree used is based on CART [2], which makes binary splits so that information gain is maximized. As metric for information gain, the highest reduction in sample variance after split is used. To train the decision tree, the instance-based dataset is first split in a training, validation and test set. The test and validation set contain all instances of 5 randomly selected patients each. The training set covers all instances from the remaining 17 patients, so that each set contains 175, 225 and 727 instances respectively. To prevent the DT from over-fitting on the training data, a grid search on the hyper parameter "min_samples_leaf" is performed. The MSE is used as metric to evaluate each resulting DT on the validation set. As outcome we get that setting the hyper parameter min_samples_leaf=70 gives the best performance on the validation set.

3.2 Temporal Model

As there exists a significant high degree of heterogeneity within individuals, people experience different mood cycles within a week. We decide to construct different models for different people [3]. After the data preprocessing, we obtained

the dataset for patient $AS14.01$. In the end, there remains 47 data points whose date starts from "2014-03-26" and ends on "2014-05-05". With a dataset of this scale, it is not feasible for us to exploit the neural networks to train models. Hence, ARIMA and SARIMAX models will be implemented for the prediction.

ARIMA An ARIMA model merely takes the time lags and errors of endogenous variables into account to construct the model. There is only one endogenous variable-mood in our model. Several parameters, the autoregressive terms (time lags) p, nonseasonal differencing factor d and the number of lagged forecast errors q for the model, need to be identified before the test.
At first, the plot of mood against time is presented in the following graph. As we could observe, the cyclic period could be 7 days. After we plot the mean of mood with a window size of 7 days, the underlying trend becomes significant. The seasonal decomposition graph further emphasizes the cyclic change of mood during one week.
Subsequently, the ACF plots and PACF graph provide more insights into selecting the p, d, q parameters. The possible selecting range for p will be (1,2,5,7) and the q parameter could be (1,2). Furthermore, result from Dickey-Fuller test indicates there is no significant evidence for supporting the existence of unit root in the mood variable. The variable is already stationary.
With the above evidence, we are safe to construct the ARIMA model. We set the 80% of the data points as training set and the rest 20% as the test set. In the case of patient $AS14.01$, there will be 10 test points. The graph below illustrates the expected and predicted values. It is obvious that there is no huge discrepancy and the error is always within one unit. Besides, we also calculate the MSE as a criteria for selecting the model.
Different parameters are engineers through the testing process, the optimal model is selected based on lower AIC and BIC values in addition to the MSE. The final parameters selected for this particular case will be (5, 1, 1).

SARIMAX The SARIMAX model takes "Mood" as endogenous variable and other relevant variables as exogenous variables. Since the p, d, q parameters are highly dependent on the trend and stationarity of predicted variable, we decide to set those parameters set the same as ARIMA model. The data set is also divided into 80% training set and 20% test set.
Subsequently, we engineered and test it with different parameters and variables. The exogenous variables are $appCat.entertainment, circumplex.arousal$, $circumplex.valence$ and the optimal parameter set will be (7,1,2) for the final model. The model is selected based on the lower AIC and BIC values.

4 Results

4.1 Decision Tree

We evaluate the optimized decision tree on the test set, and compare it with the benchmark model. The MSE for the test set of the decision tree is 0.404,

for the benchmark it is 0.470. Performing a t-test on the predictions the models made gives as conclusion the DT can not be ruled significant better than the benchmark.

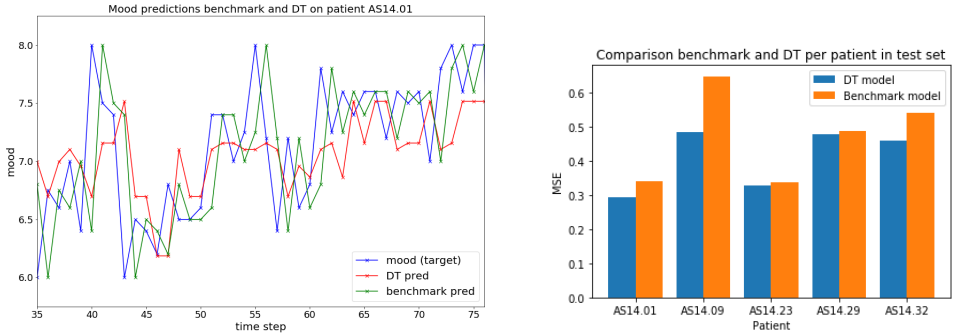
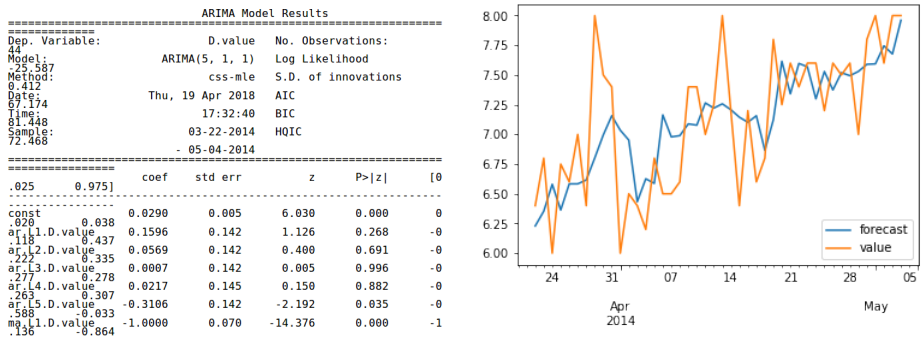


Fig. 3. Predictions for patient AS14.01 (left). Performance of DT on test set (right).

Looking at Figure 3, we see that the DT model does not seem to be able to predict any better than the benchmark. Looking at the DT itself, we see that most splits are made on prev_mood and avg_mood. Not many splits are made on other features and this suggests that any influence of those on the mood is more likely to be patient dependent.

4.2 ARIMA

The final model of ARIMA for patient one could be summarized in the picture below. The model concludes the coefficient and significance of lags, which enable us to draw insights into the cyclic movement of mood of patients.



ARIMA Model Results

Dep. Variable:	D.value	No. Observations:	44
Model:	ARIMA(5, 1, 1)	Log Likelihood	-25.587
Method:	css-mle	S.D. of innovations	0.412
Date:	Thu, 19 Apr 2018	AIC	67.174
Time:	17:32:40	BIC	81.448
Sample:	03-22-2014	HQIC	72.468
	- 05-04-2014		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0290	0.005	6.030	0.000	0.020	0.038
ar.L1.D.value	0.1596	0.142	1.126	0.268	-0.118	0.437
ar.L2.D.value	0.0569	0.142	0.400	0.691	-0.222	0.335
ar.L3.D.value	0.0007	0.142	0.005	0.996	-0.277	0.278
ar.L4.D.value	0.0217	0.145	0.150	0.882	-0.263	0.307
ar.L5.D.value	-0.3106	0.142	-2.192	0.035	-0.588	-0.033
ma.L1.D.value	-1.0000	0.070	-14.376	0.000	-1.136	-0.864

Fig. 4. ARIMA model overview.

Statistics indicates that lag 1 and lag 5 are significant, which implies that the mood of previous day and 5 days ago significantly influence the person's current day. The MSE value for the model is 0.09. The image on the right hand side presents the original (bench-mark model) and forecasting values for all 46 days. It moves with similar trend as the original values change.

4.3 SARIMAX

Similarly, results for SARIMAX and the performance are summarized in the figure below.

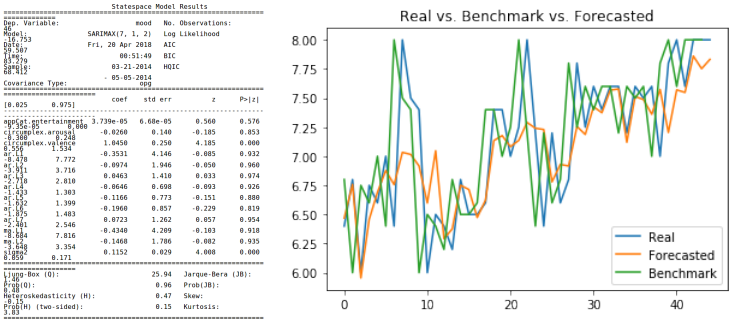


Fig. 5. SARIMAX model overview.

Among all three variables, similarly, only valence appears to be significant. MSE values of the final model approaches 0.104. The performance appears to be close to the real value.

5 Evaluation

During the experiment process, four models (Decision tree, ARIMA, SARIMAX and Benchmark model) were created. To compare model performance, we calculate the MSE between real values and predictions of last ten observations of the selected patient. Then t-tests were performed based on paired models. The table below summarizes test statistics and the corresponding p-value.

	Benchmark model		Decision tree model		ARIMA model		SARIMAX model	
	t-value	p-value	t-value	p-value	t-value	p-value	t-value	p-value
Benchmark model	-		1.01	0.33	2.83	0.01	0.63	0.53
Decision tree model	-		-		2.07	0.053	-2.01	0.059
ARIMA	-		-		-		-3.62	0.00
SARIMAX	-		-		-		-	

Table 1. t-test for paired models of patient 01 ($test - mean = row - column$).

	Benchmark model		Decision tree model		ARIMA model		SARIMAX model	
	t-value	p-value	t-value	p-value	t-value	p-value	t-value	p-value
Benchmark model	-		0.03	0.98	-0.08	0.93	-1.32	0.2
Decision tree model	-		-		-0.094	0.92	-1.05	0.2
ARIMA	-		-		-		-1.07	0.30
SARIMAX	-		-		-		-	

Table 2. t-test for paired models of patient 23 ($test - mean = row - column$).

Although most differences are not significant, we could still draw a general conclusion based on the performance graph and above tables. Model prediction power in sequence will be $SARIMAX > ARIMAX > Benchmark > Decision\,tree$.

6 Conclusion

Above analysis implements different models and concludes the prediction power for each model. In this section, we are going to analyze results and draw insights into models. The decision tree does not split a lot on most handcrafted features, suggesting that any big influence of those features (i.e. time spent on certain mobile applications) will be patient dependent.

Insights from time-series analysis will be that Circumplex arousal and Circumplex valence of previous will be most significant in predicting mood of the next day. Final optimal p, d, q parameters also indicate that the patient mood moves cyclic within a week and it is highly correlated with the mood of 1,5 and 7 days ago.
Model evaluation indicates time-series analysis (temporal-model) should be the most appropriate method in addressing the problem of predicting mood.

References

1. Russell, James A.: A circumplex model of affect. Journal of personality and social psychology 39.6 (1980)
2. Breiman, Leo. Classification and regression trees. Routledge, 2017.
3. Bauer, Michael, et al. "Temporal relation between sleep and mood in patients with bipolar disorder." Bipolar Disorders 8.2 (2006): 160-167.