Are We Living in a Happy Country: An Analysis of National Happiness from Machine Learning Perspective

Theresia Ratih Dewi Saputri
Department of Computer Engineering
Ajou University
Suwon, South Korea
trdsaputri@ajou.ac.kr

Seok-Won Lee
Department of Software Convergence Technology
Ajou University
Suwon, South Korea
leesw@ajou.ac.kr

***Abstract*— National happiness has been actively studied during last ten years. The factor of happiness could be different due to different human perspective. The factors used in this work include both physical needs and the mental needs of humanity such as educational factor. This work identified more than 90 features that can be used to predict the country happiness. Unfortunately, manually analyzing the features is difficult and needs a lot of resources. Due to numerous size of the features, it is unwise to rely on the prediction of national happiness by manual analysis. That process will result in the high cost of analysis. Therefore, this work used machine learning technique which is a Support Vector Machine to learn and predicts the country happiness. Dimensionality reduction is also done in this work. Using the information gain technique, the features can be reduced. This technique is chosen due to its ability to explore the interrelationships among a set of variables. The selected features are also evaluated using the SVM classifier. Using the data of 187 countries from the UN Development Project, this work is able to identify which factor needed to be improved by a certain country to increase the happiness of their citizens.**

Keywords-data mining; classification; feature selection; principal component analysis; support vector machine

I. INTRODUCTION

National happiness has been actively studied throughout the last ten years. The work in [8] argues that the government of a country is usually driven by the happiness of their citizens. Some factors that are controlled or authorized by the government positively correlate with the happiness level. That work shows that the key role to determine the citizen happiness is the improvement of public policy. Understanding happiness factors will help governments to make a better policy and legislation.

However, the factors that influence happiness could be different due to different human perspectives. We cannot just simply say that The United State is happier than Indonesia country because The United State has higher GDP. Peggy in [1] stated that happiness is correlated with national economic and cultural living conditions. The work in [2] determined happiness using three factors which are life expectancy, experienced well-being and Ecological Footprint. Other work in [9] shows a new measurement to improve the happiness of a country. Unlike the previous work, this work studies that happiness is not only related to physical but also mental needs. Therefore, they also consider mental health, which includes stress, depression, and emotional problems.

As a result of the increase of human social complexity, the factors proposed by [2] and [9] may not be reliable anymore. Additional factors such as health and human development index should be examined carefully. However, analyzing the factor to determine happiness of a particular country is not a trivial problem. A single factor can have a bigger impact than another. NEF organization in [2] proposes an equation to calculate the happiness index. However, this equation does not consider the economical aspects. Therefore, this work proposes an approach by extending the factors and adopting machine learning techniques to learn about those factors.

Due to numerous size of the features, it is unwise to rely on the prediction of a national happiness done by manual analysis. That process will result in high cost of analysis. Therefore, this work also proposes the use of machine learning to predict national happiness. Machine learning is a widely known technique to learn about patterns in data. There are several machine learning techniques which can be used to perform a prediction task [3]. One of the remarkable techniques is the support vector machine. This work uses support vector machine because its outstanding ability to perform a classification task.

II. RELATED WORK

This section briefly explains the related work in this project. Firstly, the national happiness analysis is described. It will discuss the importance of happiness analysis. Secondly, the used machine learning is introduced. Lastly, the proposed factor analysis is discussed.

DOI reference number: 10.18293/SEKE2015-224

A. National Happiness Analysis

The work in [4] mentioned that happiness could be a good indicator for how well a society is doing. This becomes important because Betham [5] said that the best society is the one where the citizens are happiest. Several researches have been conducted on positive aspects and the matters of happiness in policy making [6][7]. As mentioned in the previous section, happiness can be determined based on various factors. Unfortunately, these factors were analyzed manually [8]. The complexity of the factor leads to the expensive cost of analysis. Therefore, the automatic analysis is needed.

B. Support Vector Machine

Due to its capability to learn from the past experiences, machine learning has been used in various areas. Support vector machine is one of the powerful machine learning algorithm. Support Vector Machine (SVM) is a learning technique which is used for classifying unseen data correctly. It is a learning technique which usually used for classifying the unseen data correctly. This technique has been used in various research field due to its remarkable performance. In order to perform the classification task, support vector machine builds a hyperplane which separates the data into different categories [9].

One of the important advantages of support vector machine is its ability to handle the scarcity of the data. Moreover, support vector machine is able to learn about the complex decision boundaries in the high dimensional feature space efficiently. Due to the complex features used to predict the national happiness, it is important to apply the technique with ability to handle the complex features.

C. Factor Analysis

As mentioned in the previous section, there are a large number of features used to predict the national happiness. However, some of the features may have no significant contribution to the prediction. Therefore, it is unwise to use the entire features to analyze the happiness. This work uses factor analysis to analyze the related features. Factor analysis aims to determine the contribution of a certain feature. This technique does not focus on dimension reduction. Therefore, there will be no features removed. The works in [10] and [11] have introduced the advantages of factor analysis. The first advantages mentioned is the ability to identify latent dimensions or constructs that cannot be done using direct analysis. Moreover, this approach is easy to run and inexpensive in term of resources.

III. Proposed Approach

The aim of this project is to predict the national happiness of a particular country using machine learning techniques. The proposed approach contains four main steps in the data mining process which are data collection, data preprocessing, data analysis, and classification process as seen in Figure 1.

The first process in the process approach is collecting the data. The data used in this project are gathered from the UN Human Development Project. The data contains of the human development index, GDI, healthy index of each country in the world. However, these data are quite dirty. It cannot be used directly as the input data for the learning process. Therefore, the second process is data preprocessing. Data preprocessing is used to increase the data quality. By increasing the quality of data results to the increasing number of prediction accuracy and consistency. The processes included in this process are data cleaning and data integration. The routine processes that should be done are filling the missing values, reduce the noise and identify the outliers.

Figure 1. The proposed Approach

The third process in the proposed approach is data analysis. This explanatory data analysis is used for finding the relationship among the attributes of the features. This analysis is done by visualizing the data. Dimensional reduction also be done in this step. Using the information gain technique, the features can be reduced. Information gain technique is used because it can explore the interrelationships among a set of variables. The last part is this work is the classification process. In this classification process, SVM technique is used to predict the happiness of the data based on the important features. The validation process using k-fold cross validation technique is used to measure the performance of the data based on the accuracy, sensitivity and specificity values.

IV. Result and Analysis

This section discusses about the result of each step in the proposed approach. Moreover, this section also presents the analysis of significant factors to determine the happiness of a particular country.

A. Data Collection

As mentioned in the proposed approach section, the data used for this work are gathered from the UN Development Project. In total there are 187 countries listed in the data. Different types of factors are also mentioned in this data, such as human development index, education, environment, health care. The data consists of 105 types of features from 14 different factors.

However, we know that there is no perfect data. This data consists of various missing values, especially for the relatively small country such as Liechtenstein, as seen in Figure 2. This data does not only consist of missing value, but also some of outlier data. Therefore, this work needs a data preprocessing in order to improve the quality of data analysis.

17 Japan	94.8	5.2	..	9.2	3.9	0.0	68.(
18 Liechtenstein	..	..	..	..	..	..	43.'
19 Israel	96.7	4.8	99.7	9.3	3.9	0.3	7.1

Figure 2. Example of Missing value

B. Data Preprocessing

In order to increase the data quality, the data preprocessing is needed. The collected data are scattered in various tables. Therefore, creating an integrated data is needed. The single

integrated table consists of the entire features and sample that we are going to use in the next process. As seen in the previous sub-section, there are some missing values in the data. Those missing values may reduce the quality of the classification process. There are some ways to handle missing values such as pairwise deletion, listwise deletion, and mean substitution.

However, these approaches are not adequate to handle the missing values. The first and second approaches are not adequate because it needs to remove some of the data. Those approaches result in a massive decrease in the sample size for analysis and classification. It may not have a big impact when the number of missing values is small. However, there are a big number of missing values in the data used for this project. Therefore, choosing pairwise or listwise deletion is an unwise decision. One possible approach is to find the substitute of the missing values. Mean substitution can be one of possible solutions. However, adding data with mean will be useless. The overall mean, with or without replacing my missing data, will be the same. Moreover, using mean substitution makes only a trivial change in the correlation coefficient and no change in the regression coefficient.

Therefore, the approach used in this project is regression substitution. Instead of adding a trivial value for the missing value, regression substitution tried to predict the value based on other variables. This technique uses existing variables to make a prediction, and then substitute that predicted value as if it were an actual obtained value. In this case, seven variables are used to predict the missing value.

C. *Feature Selection*

Selecting the related features is important in order to improve the performance of the classifier. In order to perform this process, WEKA package for attribute selection is used [12]. The evaluator used in this work is information gain. This technique is chosen due to its ability to measure the amount of information in bits about the class prediction [13]. Therefore, it measures the expected reduction in entropy.

NO	Attributes Name	NO	Attributes Name
1	inequality in life expectancy	19	Pupil-teacher ratio
2	homeless person	20	Adolescent birth rate
3	Life expectancy at birth	21	Employment to population ratio
4	coefficient of human inequality	22	Gross National Income
5	HDI	23	female suicide rate
6	adult male mortality	24	male suicide rate
7	adult female mortality	25	Sex ratio at birth
8	Total fertility rate	26	Youth unemployment
9	maternal mortality ratio	27	youth literacy rate
10	adult with HIV	28	Electrification rate
11	Orphaned children	29	female secondary education
12	child with HIV	30	Homicide rate
13	under five mortality rate	31	Community
14	Dependency ratio (young age)	32	Internet users
15	Standard of living	33	Birth registration
16	infant mortality rate	34	expected years of schooling
17	Dependency ratio(old age)	35	Health care quality
18	Gender Inequality		

Figure 3. Selected features for classification

There are two main properties in the ranker evaluator. The first property is numToSelect property, which defines the number of attributes to keep, an Integer number that is -1 (all) by default. The next property is the threshold which defines the minimum value that an attribute has to get in the evaluator in order to be kept. In this case, the threshold is set to 0.

After running this process, the number of remaining attributes is 36 includes class attribute. Those attributes are listed in Figure 3. The selected attributes such as inequality in life expectancy (inequality in the distribution of the expected length of life based on data from life tables estimated using the Atkinson inequality index), the number of homeless people, and the mortality rate is used to classify the happiness of a certain country.

D. *Classification*

In order to evaluate the selected attribute, this work also runs the classification using the entire gathered attributes. The same parameter used to compare both scenarios. The kernel for SVM is chosen based on cross validation. Table 1 shows the result for comparing different types of kernel. We can see that the normalized poly kernel gave an outperform result. Therefore, this kernel is chosen for this work.

TABLE I. COMPARISON RESULT OF THE KERNEL

Kernel Type	**Accuracy Rate**
Normalize Poly Kernel	68.456 %
Poly Kernel	60.402 %
RBF Kernel	38.926 %
String Kernel	43.624 %

As mentioned before, this work also runs the classification using the entire attributes in order to validate the performance of selected attributes. Using the entire attributes we can see that the classification process result in 58 % of accuracy. This result shows the improvement of accuracy rate and true positive rate. It also shows that using the selected attribute is able to reduce the mean square error. It means that the selected attributes have strong correlation with the class attribute that can be used to predict the class which is the happiness of the country.

E. *Analysis*

Instead of classifying the data into happy and unhappy countries, this work classified the data into three categories which are happy, mid, and unhappy. Based on the classification results, it shows that most countries in the world are not in a happy state. As seen in Figure 4, 39 %, 38%, 23% of the countries are happy, mid-happy, and unhappy, respectively.

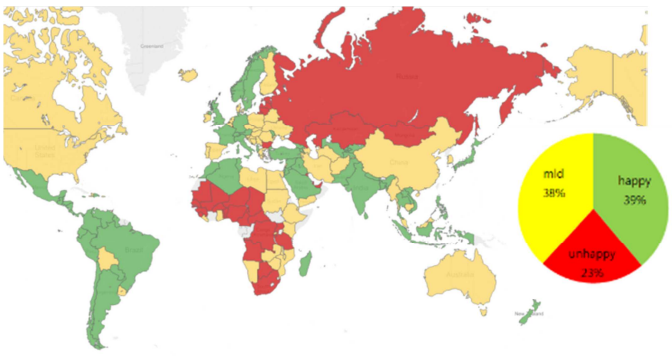

Figure 4. Distribution of Country Happiness

In order to analyze the happiness factor, this work also shows the distribution of the happiness based on the country. A country with red shade is a country in an unhappy state such as

Russia and Nigeria. Moreover, the country with yellow shade is mid-state country such as the United State and Australia. Lastly, the country with green shade has happy state such as Brazill and Indonesia. This figure showed one surprising fact that even though a country is developed, it does not mean that it has a happy state.

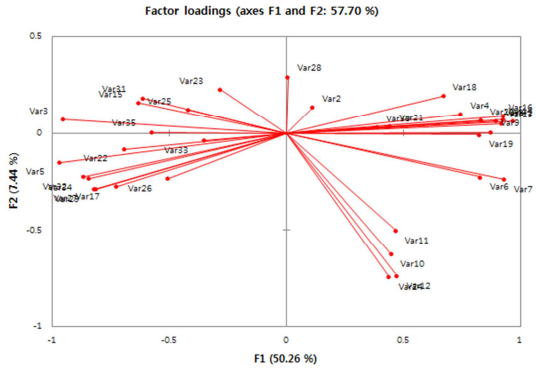

Figure 5. Factor Loadings based on PCA

After the classification process has been done, the next step is factor analysis. This process is done to determine the factor that drives the happiness of a country. There are various approaches in statistical learning to analysis the factor. This work uses factor analysis with principal component analysis (PCA) to evaluate the factors. Using this approach, the selected features are grouped into several factors as seen in Figure 5. The evaluation of the factors is done based on the classification results. In order to evaluate the factors, each of the sample was observed based on its class attributes. Using this evaluation, we can clearly see which features make a significant contribution to determine the happiness of a particular country.

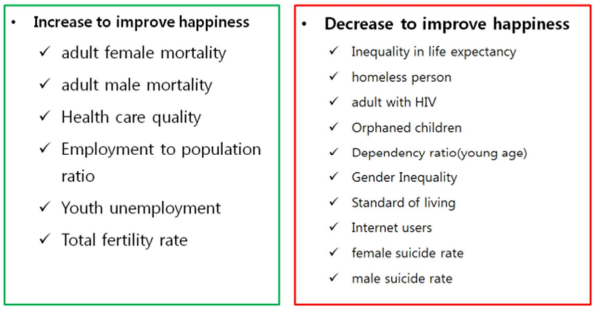

Figure 6. Summary of Significant Factors

In order to determine the significant factors, the factor loading values are used. Based on those values, some features should be increased to improve the happiness of a certain country such as health care quality and employment population ratio. We can also determine which features should be decreased in order to improve the happiness such as inequality in life expectancy, homeless person, adults with HIV and gender inequality. Figure 6 shows a summary of the features that need to be decreased or increased based on the results of the factors analysis process. In order to improve the state of happiness in for a country, the government can refer to the listed significant factors in the policy making process. Using this list, it will be easier for them to determine which factors they need to improve in order to increase the happiness of their citizens.

V. CONCLUSION

The happiness of a country cannot simply determine by its development index. This work showed that there are various factors can be used to determine the happiness. Due to the various factors to classify the happiness, prediction is hard to perform. Therefore, this work proposed the use of machine learning technique to learn about the factor to predict the national happiness combined with feature selection approach.

This work showed that the feature selection process using information gain is able to increase the performance of the classification process. The performance improvement is proved by the increase of accuracy rate when the selected features were used. This work also shows that using SVM classifier the happiness of each country can be determined effectively and efficiently.

Even though the accuracy rate increased due to the use of selected features, unfortunately, the accuracy of the result is still inferior. We argue that the inadequate of accuracy is caused by the large number of missing values. Therefore, the future work should implement a reliable approach for handling the missing values.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2013R1A1A2009801)

REFERENCES

[1] Schyns, Peggy. "Crossnational differences in happiness: Economic and cultural factors explored." Social Indicators Research 43.1-2, pp. 3-26, 1998.

[2] http://www.happyplanetindex.org/assets/happy-planet-index-report.pdf

[3] Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[4] Gudmundsdottir, Dora Gudrun. "The impact of economic crisis on happiness." Social indicators research 110.3. pp: 1083-1101, 2013.

[5] Bentham, J. (1789/1996).An Introduction of the principles of morals and legislation. Oxford: Clarendon Press. (Originally from 1789)

[6] Diener, E., Lucas, R. E., Schimmack, U., & Helliwell, J. Well-being for public policy. New York: Oxford University Press, 2009.

[7] Dolan, Paul, and Mathew P. White. "How can measures of subjective well-being be used to inform public policy?." Perspectives on Psychological Science 2, no. 1 pp.71-85.2007.

[8] Viinamäki, H., Kontula, O., Niskanen, L., & Koskela, K. "The association between economic and social factors and mental health in Finland." Acta Psychiatrica Scandinavica 92, no. 3, pp.208-213, 1995.

[9] Malhotra, R., & Jain, A. "Software Effort Prediction using Statistical and Machine Learning Methods." International Journal of Advanced Computer Science and Applications 2., pp. 1451-1521, 2011.

[10] Garson, G. David, "Factor Analysis," from Statnotes: Topics in Multivariate Analysis.

[11] Tucker, L. R., & MacCallum, R. C.. "Exploratory factor analysis." Unpublished manuscript, Ohio State University, Columbus, 1997.

[12] http://weka.sourceforge.net/doc.dev/weka/filters/supervised/attribute/AttributeSelection.html

[13] Roobaert, D., Karakoulas, G., & Chawla, N. V. "Information gain, correlation and support vector machines." In Feature Extraction, pp. 463-470. Springer Berlin Heidelberg, 2006.