International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 18 (2017) pp. 7790-7796

Optimal Risk Factor Selection Approach to Predict Heart Disease Accurately for the Immediate Health Care System

G. Purusothaman G [1] and Dr. A. Nithya [2]

[1]Ph.D, Research Scholar, Assistant Professor, School of Computer Studies (MCA), Bharathiar University, RVS College of Arts Science, Sulur, Coimbatore, Tamil Nadu, India.

[2]Assistant Professor, School of Computer Studies – Research and CS, Bharathiar University, RVS College of Arts Science, Sulur, Coimbatore, Tamil Nadu, India.

Abstract

Heart disease becomes the most found disease in the world over people with all age. There is various research methods has been introduced earlier for the accurate prediction of heart disease. In the existing research method heart disease prediction is done by using Particle Swarm Optimization (PSO) algorithm which selects risk factors selection and the disease prediction is done by using decision tree method. However this method degrades in its performance with reduced accuracy due to more computational complexity of the PSO and decision tree algorithms. This method reduced in its performance in case of presence of more noises and less information present in the dataset. These problems are resolved in the proposed research methodology by introducing novel research framework namely Noise aware Optimal Risk Factor Selection for Heart Disease Prediction System (NORFS-HDPS). In this work, initially Pre-processing is done by using K-means clustering Algorithm to make the dataset cleaner, consistent and noise free. Then Feature reduction of the dataset is done by using Fuzzy based rough set theory. The fuzzy rough sets extend the rough set concept through the use of fuzzy equivalence classes. From this reduced feature set, optimal feature selection is done by using Improved Bee Colony approach. Finally, classification of dataset is done by using Support vector machine. The overall evaluation of the proposed research methodology is done by using the Matlab simulation environment from which it is proved that the proposed work namely NORFS-HDPS can lead to optimal outcome than the existing research techniques.

Keywords: Risk factors, Feature reduction, Optimal feature selection, Noise, Classification, Heart disease prediction

INTRODUCTION

Heart disease is the most common found disease in the real world environment which might cause various effects on the human health level. Heart disease presence on humans might cause sudden death which requires to be identified as soon as possible for the prevention. More importantly older peoples frequently affected by heart disease and hospitalized which requires to be identified as soon as possible for the accurate and early identification of the heart disease. Finding the heart disease early is most difficult task which requires knowledge on heart disease parameters. It is required to focus on the varying heart disease parameter values which varies based on varying age people health levels.

However the optimal management and handling of heart disease factors gathered from multiple patients who are hospitalized in nature would be more difficult task. In this proposed research method, optimal handling of heart disease prediction is focused in order to deviate the heart disease prediction outcomes. The proposed research method attempts to focus on the varying heart disease characteristics and personal details about the patients to differentiate the effects of heart disease occurring on patients.

There is various research methods has been proposed earlier which attempts to focus on accurate and reliable prediction of heart disease. However those techniques cannot predict accurately due to presence of varying risk factors such as pollution effects, varying noises present in the environment and so on. And also available research method are more cost in nature which cannot response well on time, thus the earlier prediction of heart disease would be more difficult task.

In the proposed research methodology, accurate and early prediction of heart disease system is implemented which attempts to predict the occurrence of heart disease earlier based on risk factors that reflects the heart disease presents on humans. This study focuses on various samples that are gathered from the multiple patients for the accurate prediction of heart disease. From those samples, heart disease prediction is done based on the various risk factors such as pulse rate, blood pressure level and so on. This is ensured by introducing novel research framework namely Noise aware Optimal Risk

Factor Selection for Heart Disease Prediction System (NORFS-HDPS)

The overall organization of the proposed research methodology is given as follows: In section 2, detailed discussion about the varying related research methodologies has been given. In section 3, proposed research method is discussed in detailed along with suitable examples and explanations. In section 4, experimental evaluation of the proposed research method is given. Finally in section 5, final conclusion of the proposed research method based on experimental results has been given.

RELATED WORKS

In this section, different research methodologies have been discussed in detailed which attempts to predict the heart disease occurred on humans.

In [8], Ischemic Heart Disease Identification process is improvised by introducing the efficient feature selection techniques namely Multi Layer Perceptron Neural Network. This method adapts behaviour of Artificial Neural network to select the most optimal features from the set of attributes present in the heart disease database.

In [9], the performance cardiovascular disease prediction has been improved by focusing on the feature selection process. The main goal of this research method is, classification with interesting features instead of entire feature would increase the prediction rate.

In [10], authors introduced the automated method for the important feature selection process. The main goal of this research method is to select the features which are disease specific one. This method performs both filtering and wrapping to reduce the feature dimensionality.

In [11], authors attempted to predict the coronary heart disease by selecting the acoustic features from the heart disease database. These features are found based on overlapping frequency band values. This is done by introducing the quadratic discriminant function which is used to evaluate the discriminant behaviour between various frame values.

In [12] automatic introduced automatic prediction system for heart disease whose performance is improved considerably by integrating Probabilistic Principal Component Analysis based feature selection process. This method can guarantee the efficient selection of different kind of features which can lead to better classification rate.

In [16], authors introduced particle swarm optimization procedure for the optimal selection of risk factors that are associated with the proposed research method. This method has been evaluated on STULONG database from which it is confirmed that the proposed research method leads to provide the selection of more optimal risk factors feature set than the existing research methods.

In [17], authors introduced local lexicalized rules for the finding the risk factors that are more reasonable for the heart disease occurrence. This research method attempts to find the more unique factors by generating rules that tends to identify the variation present between the features set.

In [18], authors have introduced the similarity difference identification techniques based on which heart disease can be identified more accurately. This analysis has been carried out on japan heart disease database from which it is proved that the proposed research method can lead to increased performance result than the existing research methods.

Risk Factors Aware Heart Disease Prediction System

In the proposed research method, early heart disease prediction system is focused which attempts to retrieve the heart disease risk factors in the accurate way. This is done with the concern of risk factors of heart disease which reflects the heart disease presence in the accurate way. This is ensured in the proposed research method by introducing the Noise aware Optimal Risk Factor Selection for Heart Disease Prediction System (NORFS-HDPS). This research work can ensure the optimal prediction of heart disease presence earlier, thus the proper treatment can be carried out on right time. However the information gathered from online regarding heart disease would consist of various noises and irrelevant information which would cause the accurate detection rate of heart disease. This can be avoided by removing the noisy data and making them in the structured format, thus the accurate heart disease prediction can be ensured. The steps followed in the proposed research methodologies are given as follows:

- Preprocessing using K-means clustering Algorithm to make the dataset cleaner, consistent and noise free.
- Feature Selection of the dataset using Fuzzy based Rough set theory. The fuzzy rough sets extend the rough set concept through the use of fuzzy equivalence classes.
- Optimal feature selection using Improved Bee Colony approach
- Classification using Support vector machine is used for classification of the dataset

The overall flow of the proposed research method is given as follows:

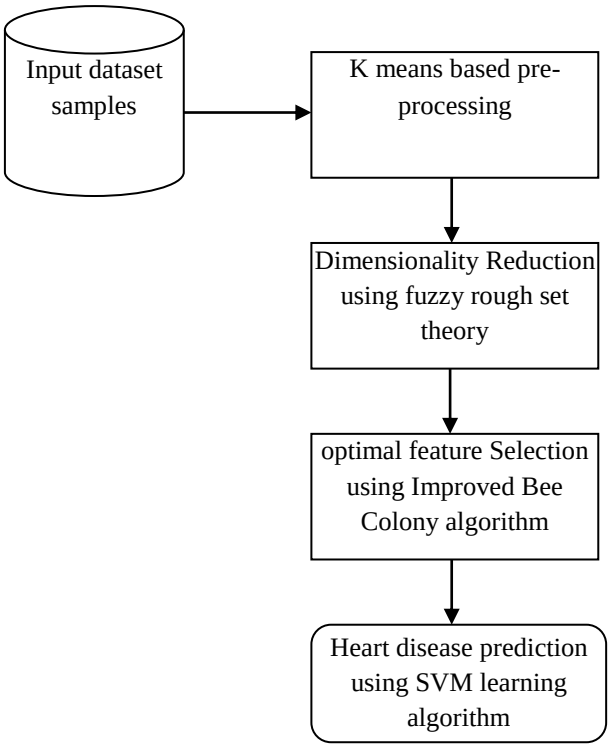

Figure 1: Overall proposed research work flow

Preprocessing Using K-Means Algorithm

Missing data values plays a more important role in the prediction process, where the incomplete information would lead to wrong prediction result of heart disease prediction. Incomplete information would lead to wrong prediction outcome which might cause the treatments, thus the dangerous effects patients might occur. In the proposed research method, missing value handling is done by using the K means algorithm.

K means algorithm is an most popular clustering techniques from the data mining field whose main goal is to group the similar kind of data that are having same characteristics together. Here similarity is identified by using the Euclidean distance which can find the similarity between the multiple objects. Thus the grouping can be done more effectively. This distance similarity metric can be utilized to replace the missing values present in the database, thus the complete information can be gained in the accurate way. To find the more accurate value there are more normalization techniques are available. In this research method min-max normalization technique is utilized for the accurate finding of missing values. It is used to find the boundary range for the missing value data. The algorithm is given as follows:

Algorithm 1: Missing Value Handling with K Means Algorithm

Input: Dataset with missing values

Output: Missing values filled clusters

1. Initialize clustering

2. Find the maximum and minimum range of each feature present in the dataset

3. Find the missing value using Max and Min range values using following equation

$$v' = \frac{v - \min(e)}{\max(e) - \min(e)}$$

4. Calculate the average score of each data point.

1) di= x1, x2,x3,x4…xn

2) di(avg)=(w1*x1+w2*x2+w3*x3+…..wm*xm)/m where x= attribute's value , m= no of attributes, w= weight to multiply to ensure fair distribution of cluster.

5. Calculate mean values

6. Replace missing values

The above algorithm can replace the missing values present in the dataset gathered based on which accurate prediction can be done. After missing value replacement, feature reduction is done which is explained in the following sub section.

Feature Reduction Using Fuzzy Based Rough Set Theory

Presence of the irrelevant information present in the database would lead to inaccurate prediction outcome and also would cause more computational complexity. For the accurate prediction outcome, it is required to remove the irrelevant information present in the dataset. In the proposed research method fuzzy rough set theory is utilized to select the more relevant data information which would remove the irrelevant information's. Thus the heart disease prediction outcome can be optimized.

The main goal of fuzzy rough set theory is to find the relevant data's by avoiding the irrelevant information present in the database and to present small volume of dataset. The fuzzy rough set theory is based on the equivalence relationship present between the multiple objects present in the database. In order to tackle the uncertainty problem this work attempts to derive the structural characteristics similarity present between multiple objects. Here fuzzy relationship is derived in the numerical representation format. These fuzzy relations are utilized then to derive the fuzzy rough set which would include only relevant features.

There are two types of fuzzy relations are considered in this work for generating the fuzzy based rough set. Those are multiplicative preference relations and fuzzy preference relations. For example consider U as nonempty universe and R as an fuzzy relation factor. The minimum and maximum bound values of fuzzy can be calculated as

(1) S-lower approximation operator:

$$\underline{R_S A(x)} = \inf_{u \leftarrow U} S(N(R(x, u)), A(u))$$

(2) T-upper approximation operator:

$$\overline{R_T} A(x) = \sup_{u \leftarrow U} T(R(x, u), A(u))$$

(3) θ-lower approximation operator:

$$\underline{R_\theta} A(x) = \inf_{u \leftarrow U} \theta(R(x, u), A(u))$$

(4) σ-upper approximation operator:

$$\overline{R_\sigma} A(x) = \sup_{u \leftarrow U} T(N(x, u), A(u))$$

Where S, T, θ and σ are mapping functions which is used to establish the relationship between the structural characteristics of attributes present in the database. The mail goal of this research method is to establish the relationship between attributes to remove the irrelevant data which is achieved in this research method.

Fuzzy preference relation R is defined as the fuzzy product set U×U which is an product relation between the nonempty universe data. This relationship can be represented as a membership function $\mu_R: U \times U \rightarrow [0, 1]$. Here it can assumed U as finite dataset for which relationship degree is measured as absolute product operation outcome.

Input: Decision table S = (CDUDT)

Output: One feature subset sub

Step 1*:* sub ← ∅, RS_1 ← sub, P_1 ← {RS_1} and US_1 ← U;// sub is the pool to conserve the selected attributes

Step 2: While EvalF (sub, DT) ≠ EvalF (CD, DT)

Do

{

Compute the positive region of forward approximation $POT_{P_i}^{U}(DT)$

US_{i+1} ←US- $POT_{P_i}^{U}(DT)$,//Remaining sets in the universal set

A ← c − sub;

Select a_0∈A which satisfies Sig (a_0, sub, DT, US_i)=max {Sig (a_0, sub, DT, US_i), a_k ∈ A},

If

{

Sig (a_0, sub, DT, US_i) > 0;

then sub← sub ∪ {a_0},

}

Else

{

RS_i ← RS_i ∪ {a_0};

PO_i ← {RS_1, RS_2, … , RS_i};

}

Step 3: Return sub and end

Optimal Feature Selection Using Improved Bee Colony Approach

Heart disease outcome purely depends on the types of features given to the training process. The risk factors of heart disease given to training process would increase the prediction accuracy and also it can be processed with reduced time complexity. Thus it is required to select the risk factors from the database, which can predict the heart disease in the accurate manner. In this research method Improved Bee Colony Approach is utilized for the optimal risk factor selection.

Bee colony algorithm is an swarm intelligence algorithm which is based on food foraging behavior of bee. There are three types of bee are present in the bee colony technique namely employee bee, onlooker bee and scout bees. Employee bees are used to search and retrieve the food to the nest. Onlooker bee is used to find the possible food location where employee bee can find food. And scout bee role is to predict the maximum food source if onlooker bee failed predict the food source. These behaviors are adapted in this research method to predict the risk factors which can optimally predict the heart disease occurrence.

Here initializing population terms to found as most important section which would mostly affect the outcome. Initially population would be initialized with random combination of bees from this most optimal solution would be identified. Here fitness value is calculated as

$$P_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n}$$

Where

Fit_i → fitness solution

i → food source

SN → number of food source

By using the above equation most solution optimal set of risk factors would be identified.

Optimal Heart Disease Prediction Using Support Vector Machine

After prediction optimal set of risk factors, those values would be learned in order to make more reliable and optimal prediction outcome. In this research method SVM classifier is utilized for the optimal prediction of heart disease. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

SVM is supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other.

Formula

To reduce the error minimization we can use given below formula

$$\Phi(w) = \frac{1}{2}||w||^2 \quad (2)$$

Estimating function

$$F(x) = \sum_{i=1}^{nsv}(x_i, y_i) \quad (3)$$

Algorithm

Given dataset X=(x1, y1),…….,(xn, yn), C // x and y –labeled sequence and C-class

Initialize vector v=0, b=0; class // v-vector and b-bias

Train an initial SVM

For each $x_i \in X$ do // xi is a vector containing features describing example i

Classify x_i using f (x_i)

If y_i f (x_i) < 1 // prediction class label

Find w', b' for known data // w', b' for new data

Add x_i to known data

If the prediction is wrong then retrain

Repeat

End

Error values are minimized using equation (1)

Classify results using equation (2)

EXPERIMENTAL RESULTS

The experiments are conducted in the matlab simulation environment where the performance evaluation between the proposed method namely Heart disease prediction using Fuzzy Rough Set Theory combined with Support Vector Machine (FRS - SVM) and the existing methods namely Coronary Heart Disease prediction using Particle Swarm Optimization Approach (CHD-PSO), Artificial Neural Network based Prediction (ANN) is done. The performance metrics that are considered in this research work for the comparison evaluation are, "Accuracy, Precision, Recall, F-Measure". The performance evaluation are shown and explained in the following sub sections.

Accuracy

Accuracy is defined as the proportion of true positives and true negatives among the total number of results obtained. Accuracy is evaluated as,

$$\text{Accuracy} = \frac{(\text{True positive} + \text{True negative})}{(\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative})}$$

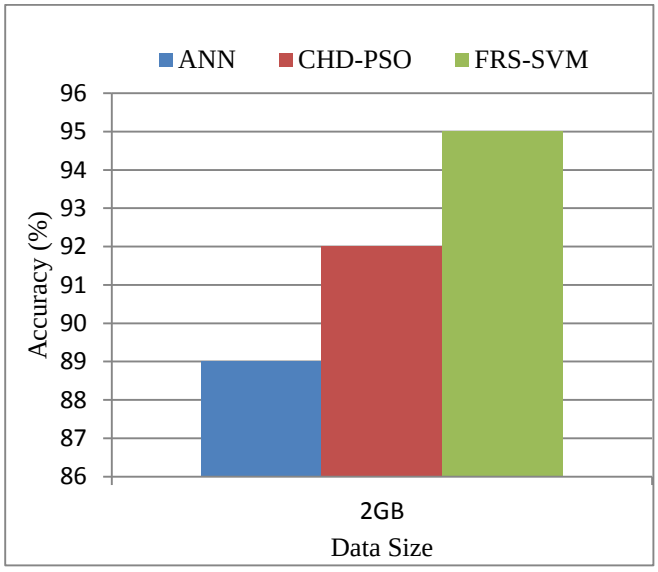

Figure 2: Accuracy

From the above graph mentioned in **[Figure-3],** it is proved that the proposed method namely FRS-SVM can provide better performance result than the existing research methodologies. From the analytical evaluation of the results, it can be proved that the proposed methodology FRS-SVM shows 6% better than CHD-PSO, 14% better performance than ANN method.

Precision

Precision value is evaluated according to the relevant information at true positive prediction, false positive.

$$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})}$$

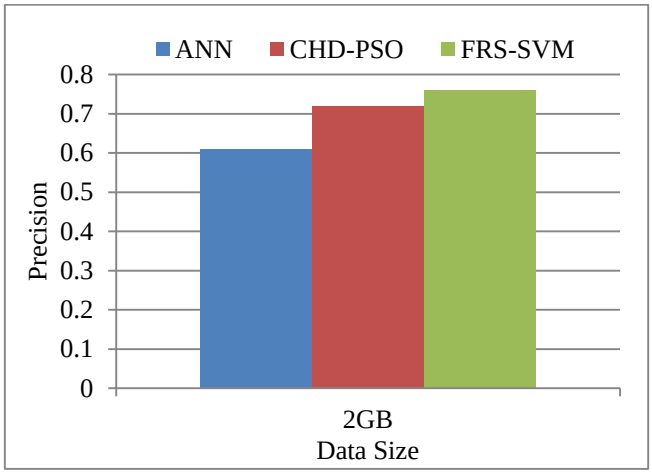

Figure 3: Precision

From the above graph it is proved that the proposed method namely FRS-SVM can provide better performance result than the existing research methodologies in terms of precision. From the analytical evaluation of the results, it can be proved that the proposed methodology FRS-SVM shows 13% better than ANN, 17% better performance than CHD-PSO method.

Recall

Recall value is evaluated according to the retrieval of information at true positive prediction, false negative.

$$\text{Recall} = \frac{\text{True Positive}}{(\text{True positive} + \text{False negative})}$$

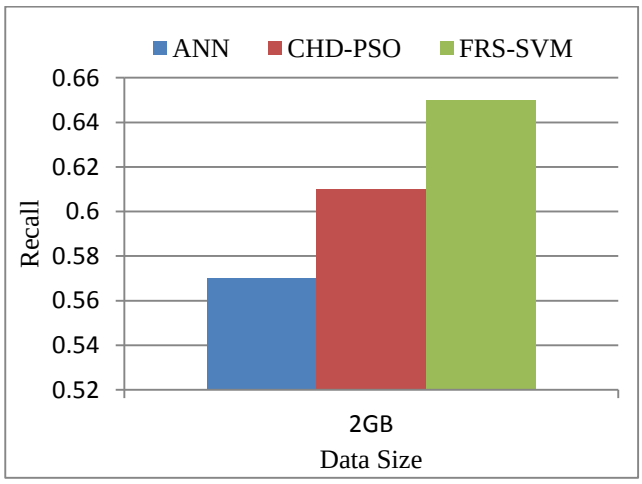

Figure 5: Recall

From the above graph it is proved that the proposed method namely FRS-SVM can provide better performance result than the existing research methodologies. From the analytical evaluation of the results, it can be proved that the proposed methodology FRS-SVM shows 17% better performance than CHD-PSO, 25% better performance than ANN method

F-Measure

The F-Measure computes some average of the information retrieval precision and recall metrics

$$\text{F} - \text{Measure} = \frac{2 * \text{precision}.\,\text{recall}}{\text{precision} + \text{recall}}$$

Figure 6: F-measure

From the above graph it is proved that the proposed method namely FRS-SVM can provide better performance result than the existing research methodologies. From the analytical evaluation of the results, it can be proved that the proposed methodology FRS-SVM shows 16% better than CHD-PSO, 23% better performance than ANN method.

CONCLUSION

Heart disease prediction is the most common problem found in the real world scenario which attempts to focus on the early prediction of heart disease, thus the patients can be treated well. This is done in this proposed work by introducing Noise aware Optimal Risk Factor Selection for Heart Disease Prediction System (NORFS-HDPS). In this research work heart disease prediction is done with the concern of risk factors that are reasonable for the sudden cardiac arrest. In this work, initially Pre-processing is done by using K-means clustering Algorithm to make the dataset cleaner, consistent and noise free. Then Feature reduction of the dataset is done

by using Fuzzy based rough set theory. The fuzzy rough sets extend the rough set concept through the use of fuzzy equivalence classes. From this reduced feature set, optimal feature selection is done by using Improved Bee Colony approach. Finally, classification of dataset is done by using Support vector machine. The overall evaluation of the proposed research methodology is done by using the Matlab simulation environment from which it is proved that the proposed work namely NORFS-HDPS can lead to optimal outcome than the existing research techniques.

REFERENCE

[1] K.Rajeswari, Dr.V.Vaithiyanathan, Dr. T.R. Neelakantan, "Feature Selection in Ischemic Heart Disease Identification using Feed Forward Neural Networks", International Symposium on Robotics and Intelligent Sensors 2012, Procedia Engineering 41 (2012) 1818 – 1823

[2] Swati Shilaskar, Ashok Ghatol, "Feature selection for medical diagnosis : Evaluation for cardiovascular diseases", Expert Systems with Applications 40 (2013) 4146–4153

[3] Zhang, Z., Dong, J., Luo, X., Choi, K. S., & Wu, X. (2014). Heartbeat classification using disease-specific feature selection. *Computers in biology and medicine*, *46*, 79-89.

[4] Schmidt, S. E., Holst-Hansen, C., Hansen, J., Toft, E., & Struijk, J. J. (2015). Acoustic features for the identification of coronary artery disease. *IEEE Transactions on Biomedical Engineering*, *62*(11), 2611-2619.

[5] Syed Muhammad Saqlain Shah, Safeera Batool, Imran Khan, Muhammad Usman Ashrafa, Syed Hussnain Abbas, Syed Adnan Hussain, "Feature Extraction through Parallel Probabilistic Principal Component Analysis for Heart Disease Diagnosis", Physica A (2017), http://dx.doi.org/10.1016/j.physa.2017.04.113

[6] V. Sree Hari Rao, and M. Naresh Kumar, "Novel Approaches for Predicting Risk Factors of Atherosclerosis", IEEE Journal Of Biomedical And Health Informatics, Vol. 17, No. 1, January 2013

[7] George Karystianis, Azad Dehghan, Aleksandar Kovacevic, John A. Keane, Goran Nenadic, "Using Local Lexicalized Rules to Identify Heart Disease Risk Factors in Clinical Notes", Journal of Biomedical Informatics (2015), doi: http://dx.doi.org/10.1016/j.jbi.2015.06.013

[8] Masaaki Matsunaga, Hiroshi Yatsuya, Hiroyasu Iso, Kentaro Yamashita, Yuanying Li, Kazumasa Yamagishi, Naohito Tanabe, Yasuhiko Wada, Chaochen Wang, Atsuhiko Ota, Koji Tamakoshi, Akiko Tamakoshi, "Similarities and differences between coronary heart disease andstroke in the associations with cardiovascular risk factors: The Japan Collaborative Cohort Study", Atherosclerosis xxx (2017) 1-7