Global Journal of Computer and Engineering Technology (GJCET)

Vol. 1, No. 1, January, 2021 ISSN: 2767-1933

A ROBUST DIVORCE PREDICTIVE MODEL USING DATA MINING WITH FEATURE SELECTION TECHNIQUE

A.K. Shrivas, Guru Ghasidas Vishwavidyalaya, India (akhiesh.mca29@gmail.com)
Vineet Kumar Awasthi, Dr. C. V. Raman University, India (vineet99kumar@gmail.com)
Devendra Singh Suman, Dr. C. V. Raman University, India (suman.bit12@gmail.com)

ABSTRACT

Nowadays, various problems are occurring during living in family where divorce case is one of them. The divorce case is very serious problem between husband and wife which destroy the family life and affects the children and other family members. This research work focuses on classification of divorce and non-divorce cases using classification techniques. Different classification techniques based on data mining have used for classification of divorce cases and compared the performance of classifiers in terms of accuracy. An Experiment results show that SVM has given better classification accuracy for classifying divorce and non-divorce cases. We have also applied Info gain and Gain ratio ranking based feature selection technique (FST) to make computationally efficient model and achieved the better accuracy for classifying the divorce and non-divorce cases with high accuracy. The suggested SVM-Gain Ratio model gives 98.23% of accuracy with reduced 5 feature subset.
Keywords: Data Mining, Feature Selection Technique (FST), Support Vector Machine (SVM), Classification, Classification and Regression Tree(CART).

1. INTRODUCTION

The divorce case is very serious matter in human life which creates problems among family members. A divorce problem occurs after deciding of husband and wife when they don't want to live together in any situation and don't want to solve the problems of each other. The divorce is very sensitive matter in human life, which destroys the family and relationship. A marriage is a relationship which came out from ancient time, it makes the pair of two people with family. The relations make the two families, cultures and societies. This marriage couples makes decision to live alone in lifestyle. Now days, divorce is shocking for marriage couples which is highly increasing rates in human life. These divorce problems are occurring in different countries and cultures day by day.Now days, identification and diagnosis of this problem is very challenging task due to increase number of misunderstanding between them. Data mining based classification techniques are useful to extract the knowledge from database. The main objective of this research work is to analysis, identification, and classification of divorce and non-divorce cases.

The figure 1 shows that the general architecture of proposed model where divorce dataset applied for classifying the data into divorce and non-divorce cases. The divorce prediction is an important problem for society and identification of reasons and classification is very challenging task for family and society.

Figure 1: Classification of divorce case

Various authors have worked on prediction of divorce due to increasing number of cases. They have used different data mining approach for identification and classification of divorce cases. Yontem et al.(2019) have suggested Artificial Neural Network (ANN) , Radial Basis Function (RBF) Neural Network and Random Forest (RF) for classification of divorce cases with correlation based FST and compared the accuracy of classifiers where ANN gives better classification accuracy with reduced feature subset. Kong et al. (2020) have suggested SVM, RF and Natural Gradient Boosting (NG Boost) techniques on a divorce prediction where proposed NG Boost gives better accuracy compared to others. Goel et al. (2019) have proposed Augur Justice technique for classification of divorce cases with Hindu, Muslim and Christian. They have also compared the performance of proposed classifier with Naive bayes, Decision Tree (DT) and RF where proposed model gives better accuracy compare to others. Nasser (2019) has suggested ANN technique to develop a model whether a couple is going to get divorced or not. Hafidz et al.(2020) have suggested SVM and ANN technique for divorce prediction and compared the performance in terms of accuracy and kappa value where SVM gives better accuracy compare than ANN. The above literature shows that divorce prediction is very challenging task and various researchers are using different data mining techniques for developing robust model for identification and classification of divorce cases.

2. PROPOSED ARCHITECTURE

This section explores the proposed architecture for predicting the divorce case or not. Figure 2 shows that the proposed architecture of research work. We have used divorce dataset collected from UCI repository. The data set consist 170 numbers of instances and 54 number of features. The nature of class is binary as divorce or non-divorce. Cross validation (Han et al., 2006) is technique for partition the data into training and testing. In k-fold cross validation, dataset is partition into k- fold where each fold used as testing and rest of folds used as training dataset. In case of classification, the accuracy estimate is the overall number of correct classifications from the k iterations, divided by the total number of tuples in the initial data. We have partitioned the dataset into training and testing using 10-fold cross validation.

Various data mining based techniques have used for classification of divorce cases. Data mining based classification technique works on training and testing stages in which training data is used to train the classifiers and test data is used to testing of trained classifiers. CART (Pujari,2001) is a DT algorithm and builds a binary decision tree by splitting the record at each node. CART uses the gini index for determining the best split. C4.5 (Pujari,2001) is DT classifier that handles the continuous attribute value ranges, pruning of decision trees and rule derivation. It produces tree with variable branches per node. RF (Parimala et al., 2011) is an ensemble classifier that consists of many decision trees. RF is often used for large dataset when number of input variables are large. Multilayer Perceptron (MLP) (Pujari, 2001) is also a classifier where extra hidden layers are added between input and output layer. The MLP method is used for both classification and regression problem.SVM (Hanet al., 2006) is a supervised learning method for classification of data into different categories. In classification, nonlinear kernel functions are often used to transform the input data to a high dimensional feature space in which the input data becomes more separable (i.e., linearly separable) compared to the original input space. Then, the maximum-margin hyper planes are constructed to optimally separate the classes in the training data. Naive bayes (Han et al., 2006) is a statistical classifier that can predict class membership probabilities. Naive bayes classifier exhibited high accuracy and speed when applied to large databases. We have supplied the dataset into various classifiers and compared the performance in terms of accuracy, sensitivity and specificity. We have selected the best classifier as SVM which predict the divorce cases with high accuracy.

Feature selection (Cios et al., 1998) is used to reduce the features form dataset and computationally increase the accuracy of model. This research work have used Information gain and Gain ratio FST to reduce the feature set. The Information gain or Info gain (Han et al., 2006) measures prefers to select attributes having a large number of values. The extension to info gain known as Gain ratio based on ranking, which attempts to overcome bias. We have also applied the Info gain and Gain ratio FST on divorce dataset to achieve better performance with best classifier SVM in terms of accuracy, sensitivity and specificity

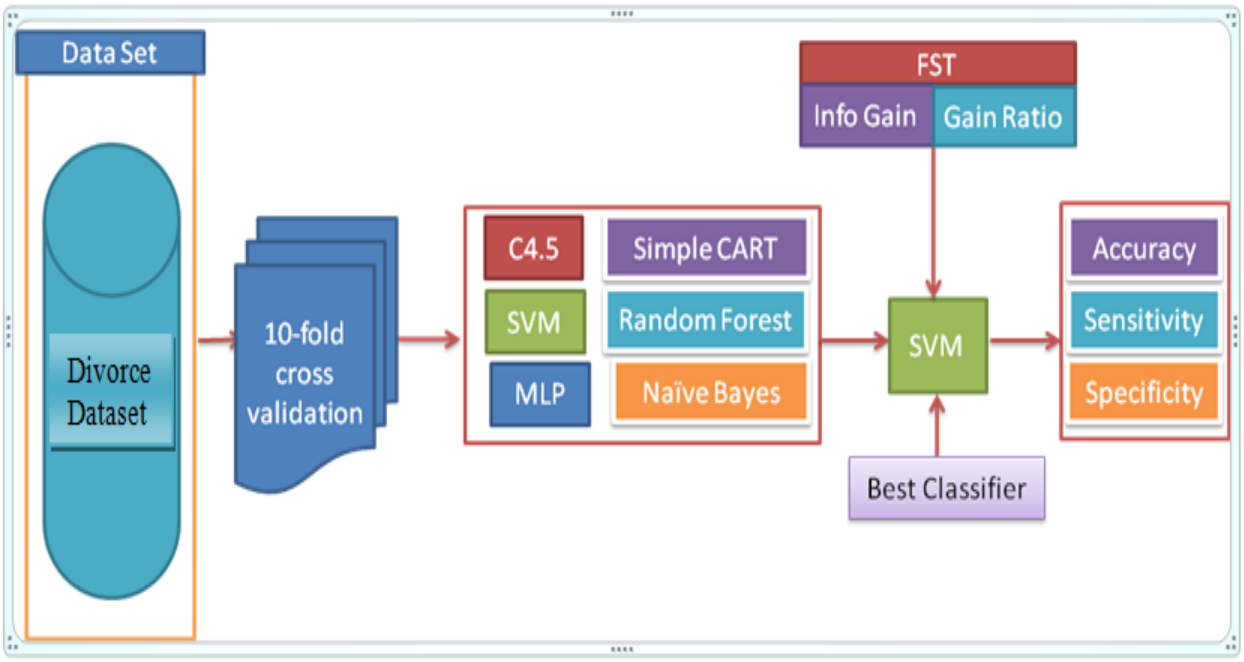

Figure 2: Proposed architecture for divorce prediction

3. EXPERIMENTAL RESULTS

This experiment was done in window 7 environment using WEKA data mining tool. We have partitioned data into training and testing using 10-fold cross validation technique. This experiment work divided into two sections: Firstly, analysis and compare of classifiers for classification of divorce and non-divorce cases. Secondly, applied the Info gain and Gain ratio FST on best classifier and proposed new computationally efficient classifier.

Section 1: This section have used C4.5, , RF, SVM (Poly Kernel Function), MLP (Single layer) , CART and Naïve Bayes for analysis and classification of divorce cases. We have compared the performance of classifiers in the term of accuracy, sensitivity and specificity with divorce dataset. Table 1 shows that performance measures of different classifiers with correctly classified instances and incorrectly classified instances with percentage. This table also shows the computational time of different classifiers.
We have achieved the best classification accuracy as 98.23% with SVM classifier. Figure 3 shows that correctly and incorrectly classified samples in percentage where SVM gives better performance and less computational time.

Table 1: Accuracy individual classifier with 10-fold cross validation

Classifier	Correctly classified	Incorrectly classified	Time taken to build model (In second)
C4.5	95.29%	4.70%	0.03
Random Forest	97.64%	2.35%	0.05
Simple Cart	95.88%	4.11%	0.08
Naïve Bayes	97.64%	2.35%	0.01
MLP	97.64%	2.35%	0.36
SVM (PolyKernel)	**98.23%**	**1.76%**	**0.01**

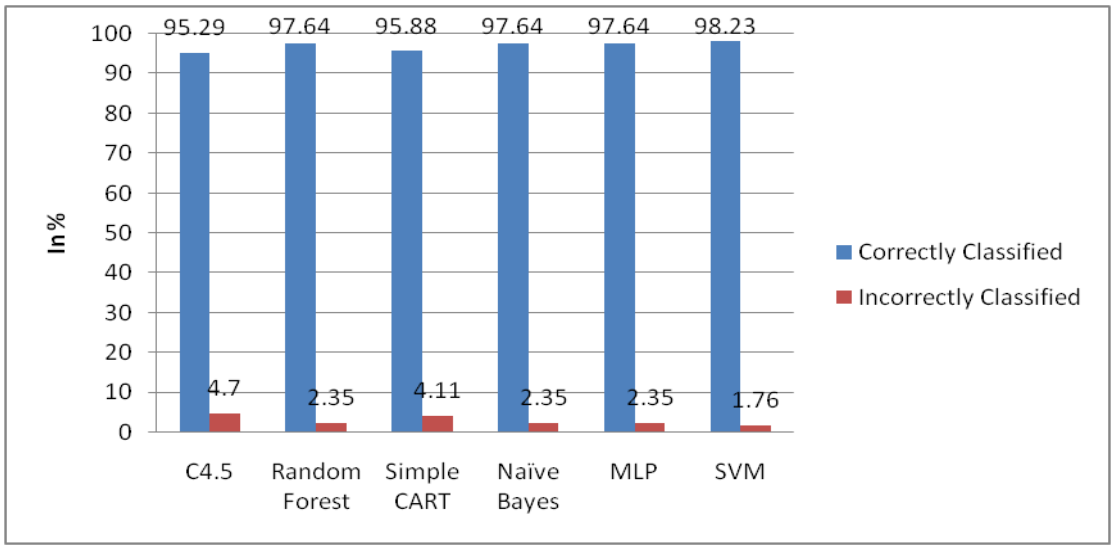

Figure 3: Correctly and incorrectly classified of samples in percentage

Table 2 shows that confusion matrix of best classifier as SVM with divorce data set where 3 samples of divorce are misclassified while no samples of non-divorce is misclassified. We have calculated the other measures like sensitivity, specificity and accuracy with the help of confusion matrix as shown in table 3. Figure 4 shows that the performance measures of best SVM classifier with divorce dataset.

Table 2 : Confusion matrix of best classifier SVM

Actual Vs. Predicted	**Divorce**	**Non_Divorce**
Divorce	81	3
Non-Divorce	0	86

Table 3: Performance measures of best classifier SVM

Performance Measures	**In %**
Accuracy	98.23%
Sensitivity	96.42%
Specificity	100%

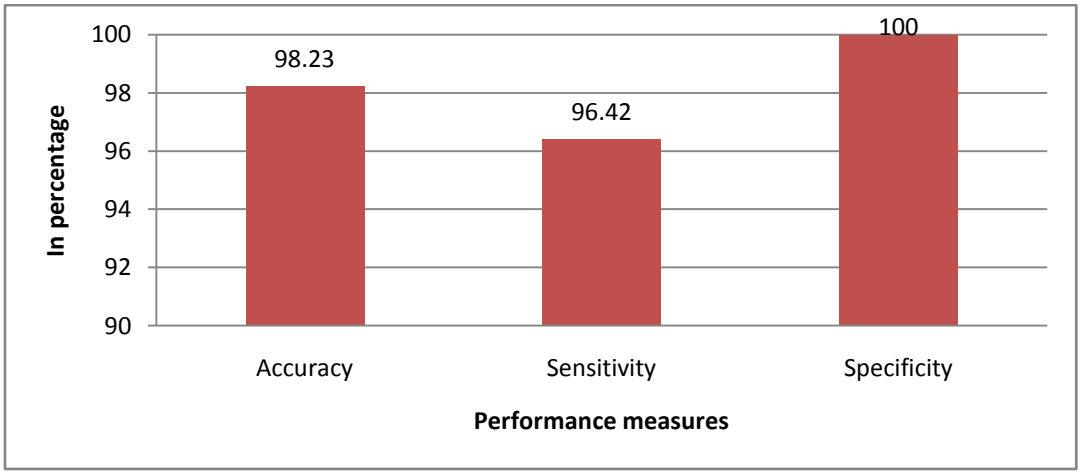

Figure 4: Performance measures of best classifier SVM

Section 2: In this section, we have used FST to select the relevant features from original feature space and improve the performance of classifiers. This research work has used Info gain and Gain ratio FST to rank the features of divorce data set; hence we can remove the less important feature from data sets. The ranking of features of divorce dataset was done by Info gain and gain ratio FST. The rank of features from important to less important as shown in table 4. We have applied the different feature subsets on best SVM classifier with divorce data set using Info Gain and Gain Ratio FST. The proposed model is known as SVM-InfoGain and SVM-GainRatio. The proposed SVM-InfoGain model is combination of SVM classifier and Info gain FST, where Info gain FST is used to reduce the feature from 54 to 50, then accuracy of SVM is improved from 98.23% to 98.82% and achieved same accuracy as 98.23% with 40 feature subset with reduced divorce dataset. The second proposed SVM-GainRatio model is combination of SVM and Gain ratio FST, where Gain ratio FST is used to reduce the feature from 54 to 50, then accuracy of SVM is improved from 98.23% to 98.82% and achieved better accuracy as 98.23% with 5 feature subset with SVM classifier. Finally, we can say that the suggested model SVM- GainRatio gives satisfactory accuracy with 5 feature subset. Table 5and figure5 show that accuracy of proposed model with reduce number of features.

Table 4: Ranking of features of dataset using raking based FST

Name of feature selection technique	Ranking of features(higher to lower)
InfoGain	20,18,40,11,19,17,9,15,29,16,26,12,39,30,5,36,27,25,41,14,21,28,4,22,10,13,8, 1,38,44,35,24,37,23,33,32,2,34,3,31,54,50,42,49,51,53,7,52,48,47,43,45,6,46
Gain ratio	18,11,17,19,9,16,26,25,14,28,21,40,8,20,38,37,2,36,31,29,39,54,41,30,12,4,27,22, 5,1,35,13,50,34,15,24,23,10,33,44,7,32,3,42,47,51,53,43,49,45,48,52,6,46

Table 5 : Accuracy of proposed model with reduced feature subset

Number of feature	SVM-Info Gain	SVM-Gain ratio
50	98.82%	98.82%
40	98.23%	98.23%
30	97.64%	97.64%
20	97.64%	97.64%
10	97.64%	98.23%
5	97.64%	98.23%

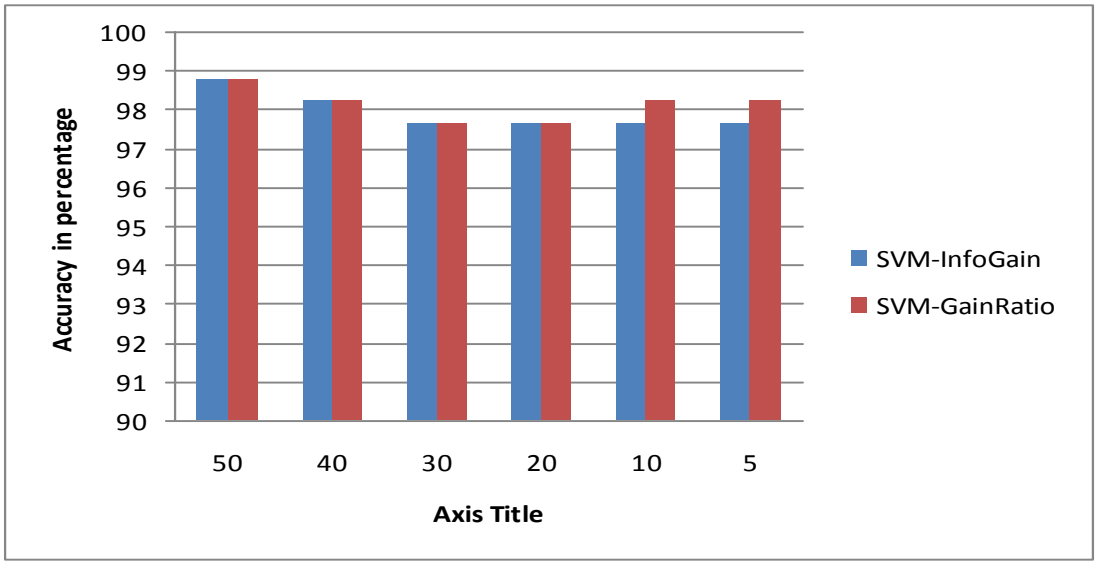

Figure 5: Graphical representation of accuracy of proposed model with FST

4. CONCLUSION

Divorce prediction is very challenging task due to increasing number of cases in society. We have proposed two models as SVM-InfoGain and SVM-GainRatio for classification of divorce and non-divorce cases. The proposed classifier based on SVM and ranking based FST to improve the performance of the classifier with high accuracy and less computational time. We have compared the performance of proposed and exiting classifiers where our proposed classifier gives better accuracy compared to others. In future, we will propose new hybrid model that will gives high accuracy compared to existing models. We will use FST like PCA, Genetic algorithm and other ranking based FST to improve accuracy with less number of features. We will also use various multiclass dataset for developing models with multiclass classification problem.

REFERENCES

Cios, K. J., Pedrycz ,W.,. W. and Swiniarski , R. W. (1998) . Data Mining Methods for Knowledge Discovery. *Kluwer Academic Publishers*, 3rd edition.

Goel S., Roshan S., Tyagi R. and Agarwal S. (2019). Augur Justice : A Supervised Machine Learning Technique To Predict Outcomes of Divorce Court Cases. *2019 Fifth International Conference on Image Information Processing (ICIIP)*, 280-285.

Hafidz N., Sfenrianto,Pribadi Y.,Fitri E. and Ratino (2020). ANN and SVM algorithm in Divorce Predictor. *International Journal of Engineering and Advanced Technology (IJEAT)*, 9(3) , 2523-2527.

Kong J. and Chai T. (2020). Is Your Marriage Reliable?: Divorce Analysis with Machine Learning Algorithms. *Proceeding of 2020 6th International Conference on Computing and Artificial Intelligence*, 1–4.

Nasser I. M. (2019). Predicting Whether a Couple is Going to Get Divorced or Not Using Artificial Neural Networks.*International Journal of Engineering and Information Systems (IJEAIS)*, 3(10), 49-55.

Divorce Dataset from UCI Repository: https://archive.ics.uci.edu/ml/datasets/Divorce+Predictors+data+set (Browing date: Feb 2020).

Yöntem, M. K., Adem, K., İlhan, T. veKılıçarslan, S. (2019). Divorce Prediction Using Correlation Based Feature Selection and Artificial Neural Networks.*NevşehirHacıBektaşVeliÜniversitesi SBE Dergisi*, 9(1), 259-273.

Parimala, R. and Nallaswamy, R. (2011). A Study of Spam e-mail Classification using Feature Selection Package. *Global Journal of Computer Science and Technology*. 11,1-11.

Pujari A. K. (2001). Data Mining Techniques. *Universities Press (India) Private Limited*, 4th edition,

Han J. and Micheline K. (2006). Data mining: Concepts and Techniques. *Morgan Kaufmann Publisher*.

WEKA Data Mining Tools: http://www.cs.waikato,ac.nz/~ml/weka/ (Browsing date: Oct. 2018).