Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 116 (2017) 113–120

Procedia Computer Science

www.elsevier.com/locate/procedia

2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017, 13-14 October 2017, Bali, Indonesia

A Comparative Study to Evaluate Filtering Methods for Crime Data Feature Selection

Masita @ Masila Abdul Jalil, Fatihah Mohd*, Noor Maizura Mohamad Noor

School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu, Malaysia.

Abstract

In this study, we present a comparative study on correlation and information gain algorithms to evaluate and produce the subset of crime features. The main objective of the study is to find a subset of attributes from a dataset described by a feature set and to classify the crimes into three different categories; low, medium and high. The experiment is carried out on the communities and crime dataset using WEKA, an open source data mining software. Based on attributes chosen by five features selection methods, the accuracy rates of several classification algorithms were obtained for analysis. The results from the experiment demonstrated that, the correlation method out performed information gain and human expert with a mean accuracy of 96.94% for entire classifier and FSs with 13 optimal features selection. This subset feature is important information for classification and can be effectively applied to crime dataset to predict crime category for different state and directly support decision making in crime prevention system.

Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.

Keywords: Correlation; crime prevention; feature selection; filter method;

* Corresponding author. Tel.: +609-668-3274; fax: +609-669-3326
E-mail address: mpfatihah@gmail.com

1877-0509
Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.
10.1016/j.procs.2017.10.018

1. Introduction

Crime prevention refers to the a series of programs that are involved with individuals, communities, businesses, non-government organizations and all levels of government in addressing the various social and environmental factors that contribute to the risk of community's crime, disorder and victimization [1 2 3]. There are many different approaches to crime prevention that focus in the intervention, the types of activities that are organized, and the mechanisms that are applied cover the environmental, social and economic, and criminal justice system features. All of the approaches aim to reduce the opportunities for crime to occur through community environments [4]. These features may seek to involve multiple types of race and ethnic categories (white Americans, black African and Asian.), different incomes class (low, middle and high income), various types of age categories, structure of family (single, married partners, unwed partners, parents with kids), education level (primary, secondary, university), population of town or locality in people live (housing price, types of house, home size), number of civil law enforcement assigned to a town, number of people working and the unemployment rate and others [5].

Because of the huge number of features included in the communities and crime data, many factors may affect the outcome of the crime prediction system. Thus, selecting the most relevant features and information is critical to improving the accuracy of prediction systems. Feature selection (FS) is a method of discovering the relevant features and removing the irrelevant features, often motivates to the performance of the learning algorithm. FS is also able to gain information about the process, reduce the data, storage and cost. There are two main models of feature selection: filter methods and wrapper methods. While filter models rely on the general characteristics of the training data to select features with independence of any predictor, wrapper models involve optimizing a predictor as part of the selection process [6].

In this paper, a number of filter methods are used over crime datasets with different number of relevant features. The results obtained for the filters studied; correlation attributes evaluator, correlation-based features subset evaluator and information gain are compared and discussed. This paper is organized as follows. Some related works are discussed in Section 2. Section 3 discusses the materials and methods used, containing communities and crime dataset, crime dataset preprocessing, and feature selection used in this study. Section 3, the experiments and results produced by features selections are presented and discussed. Finally, the conclusion of the study is concluded in Section 4.

2. Related Works

For many years, various studies have been done to communities and crime data. Buczak and Gifford [7] discovered a relationship between various crimes attributes by applying fuzzy association rule mining in crime pattern application. Halawa [8] explored multilayer perceptron into communities and crime dataset attribute for predicting the number of crimes (per capita violent crime). Iqbal et al. [5] also applied crime dataset from UC Irvine (UCI) machine learning repository for crime prediction. They used the manual method for attribute selection based on human expert. From 128 attributes, only twelve (12) attributes are chosen, namely country state, median family income, median household income, per capita income, number of people under the federal poverty level, percentage of people 25 and over with less than a 9th grade education, percentage of people 25 and over that are not high school graduates, percentage of people 25 and over with a bachelor's degree or higher education, percentage of people 16 and over in the labor force and unemployed, percentage of people 16 and over who are employed, population of community, total number of violent crimes per 100K population. In experimental works, they found that decision tree performed well than the naïve Bayesian for the crime dataset with twelve (12) selected features.

Anuar et al. [9] applied a particle swarm optimization (PSO) as a FS methods to communities and crime dataset. They proposed a hybrid crime classification model for crime prediction by combining artificial neural network (ANN), PSO and grey relation analysis (GRA). The study aimed to identify the significant features of the specific crimes and to classify the crimes into three (3) different categories. Another study developed crime location forecast method by means of calculating probability on socio economic and frequent closed item set lattice (FCIL) algorithm to locate the crime locations using the UCI data [10]. The FS methods are also explored as a hybrid algorithm in order to gain the optimum selected features such as combined information gain and sequential backward floating [11], hybrid generalized F-score (GF) with sequential forward search (SFS), sequential forward floating search (SFFS) and sequential backward floating search (SBFS) [12]. To further improve the study in crime dataset, we explore the crime features by filtering