Fuzzy based binary feature profiling for modus operandi analysis

Mahawaga Arachchige Pathum Chamikara, Akalanka Galappaththi, Roshan D Yapa, Ruwan D Nawarathna, Saluka Ranasinghe Kodituwakku, Jagath Gunatilake, Aththanapola Arachchilage Chathranee Anumitha Jayathilake, Liwan H Liyanage

It is a well-known fact that some criminals follow perpetual methods of operations, known as modi operandi (MO) which are commonly used to describe the habits in committing something especially in the context of criminal investigations. These modi operandi are then used in relating criminals to other crimes where the criminals have not yet been recognized. This paper presents a method which is focused on identifying the perpetual modi operandi of criminals by analyzing their previous convictions. The method involves in generating a feature matrix for a particular suspect based on the flow of events of his/her previous convictions. Then, based on the feature matrix, two representative modi operandi are generated: complete modus operandi and dynamic modus operandi. These two representative modi operandi will be compared with the flow of events of the crime in order to investigate and relate a particular criminal. This comparison uses several operations to generate two other outputs: completeness probability and deviation probability. These two outcomes are used as inputs to a fuzzy inference system to generate a score value which is used in providing a measurement for the similarity between the suspect and the crime at hand. The method was evaluated using actual crime data and ten other open data sets. Then ROC analysis was performed to justify the validity and the generalizability of the proposed method. In addition, comparison with nine other classification algorithms showed that the proposed method performs competitively with other related methods.

Fuzzy based binary feature profiling for modus operandi analysis

M. A. P. Chamikara[1, 2*], A. Galappaththi[1], Y.P.R.D. Yapa[1, 2], R.D. Nawarathna[1, 2], S. R. Kodituwakku[1, 2], J. Gunatilake[1, 3], A.A.C.A. Jayathilake[4] , L.H. Liyanage[5]

[1]Postgraduate Institute of Science, University of Peradeniya, Sri Lanka,

[2]Department of Statistics and Computer Science, University of Peradeniya, Sri Lanka,

[3]Department of Geology, University of Peradeniya, Sri Lanka,

[4]Department of Mathematics, University of Peradeniya, Sri Lanka,

[5]School of Computing, Engineering and Mathematics, University of Western Sydney, Australia.

Corresponding author:

M.A.P. Chamikara,
No.22, Nikaketiya, Menikhinna, 20000, Sri Lanka
Email address: pathumchamikara@gmail.com

Abstract

It is a well-known fact that some criminals follow perpetual methods of operations, known as modi operandi (MO) which are commonly used to describe the habits in committing something especially in the context of criminal investigations. These modi operandi are then used in relating criminals to other crimes where the criminals have not yet been recognized. This paper presents a method which is focused on identifying the perpetual modi operandi of criminals by analyzing their previous convictions. The method involves in generating a feature matrix for a particular suspect based on the flow of events of his/her previous convictions. Then, based on the feature matrix, two representative modi operandi are generated: complete modus operandi and dynamic modus operandi. These two representative modi operandi will be compared with the flow of events of the crime in order to investigate and relate a particular criminal. This comparison uses several operations to generate two other outputs: completeness probability and deviation probability. These two outcomes are used as inputs to a fuzzy inference system to generate a score value which is used in providing a measurement for the similarity between the suspect and the crime at hand. The method was evaluated using actual crime data and ten other open data sets. Then ROC analysis was performed to justify the validity and the generalizability of the proposed method. In addition, comparison with nine other classification algorithms showed that the proposed method performs competitively with other related methods.

Introduction

Scientists have long played a role in examining deviant behavior in society. "Deviance behaviour" is a term used by scientists to refer to some form of "rule-breaking" behaviour [1]. It can be the behaviour of violating a social norm or the law. Criminal behaviour is also a form of deviance, one that is defined as the breaking of legal rules. Nevertheless, there is a difference between deviance and crime. Deviance involves breaking a norm and evoking a negative reaction from others. Crime is a deviance that breaks a law, which is a norm stipulated and enforced by government bodies [1]. However, crimes affect the society negatively. Therefore, law enforcement authorities take necessary actions to mitigate crimes in an environment where high crime frequencies are observed each year. In this exercise the application of technology for crime analysis is being widened in the world. Locard's Exchange principle states that every contact of the perpetrators of a crime scene leaves a trace. The perpetrators will both bring something into the scene and leave with something from the scene [2]. However, the cognitive abilities of criminals will always make them minimize their risks of apprehension by conducting the perfect crime and maximizing their gain [3]. Modus operandi or method of operation such as preparation actions, crime methods and weapons are frequently used in criminal profiling because the past crime trends show that, after criminals get used to a certain method of operation, they try to use the same modus operandi in committing his/her next crime [4].

The criminals develop a set of actions during the performance of a series of crimes which we refer to as “modus operandi” (MO). MO is developed with the crimes he/she commits and the nature of trying to stick with the developed MO that has worked throughout the previous crimes [5]. In any criminal career, the MO happens to evolve, no matter what the circumstances. Also, it is a common behaviour that serial offenders tend to exhibit significant behaviour known as his/her signature. Therefore, MOs of criminals plays a major role in investigating crimes [5]. It is a known fact that features such as criminal signature and physical appearance are used in crime investigations in almost all the police departments around the world. Sri Lanka police also use MOs of criminals to identify the suspects who have conducted crimes. Currently Sri Lanka Police use a manual crime recording and investigation system. This manual system has many problems such as data redundancy, inefficiency, tediousness, inability to support crime investigation and many other problems which are associated with a conventional manual system. To overcome these problems, a web-based framework was proposed with geographical information support containing a centralized database for crime data storage and retrieval, named SL-CIDSS: Sri Lanka Crime Investigation and Decision Support System [6]. The proposed system accompanies a collection of data mining algorithms which effectively support the crime investigation process. Fuzzy based binary feature profiling (BFPM) for modus operandi analysis is one novel algorithm which is integrated with the system to provide an effective way to find the similarity between crimes and criminals.

According to the penal code of Sri Lanka first enacted in 1882 and amended subsequently several times in later years [7], Sri Lanka police classifies crimes into two categories: Grave crimes and Minor offences. Until 2014, grave crimes were classified under 21 crime categories and in 2015 another 5 new crime categories were introduced, making it 26 categories of grave crime types. Kidnapping, Fraud or mischief causing damage greater than 25000 rupees, Burglary, Grievous hurt, Hurt by sharp weapon, Homicide, Rape, Robbery, Cheating by trust, Theft are 10 of the most frequent crime types. To identify the patterns involved in crimes, a collection of subtypes were identified under these 26 crime types. These subtypes have been created mainly for the purpose of modus operandi analysis. Most frequent behaviors of criminals/crimes are considered as crime subtypes. When a crime is logged in the Grave Crime Record (GCR) book, it is classified under one of the 26 main categories. But, under the section of “nature of crime” in the GCR book, the police officers record the flow of the crime incident including the subtypes.

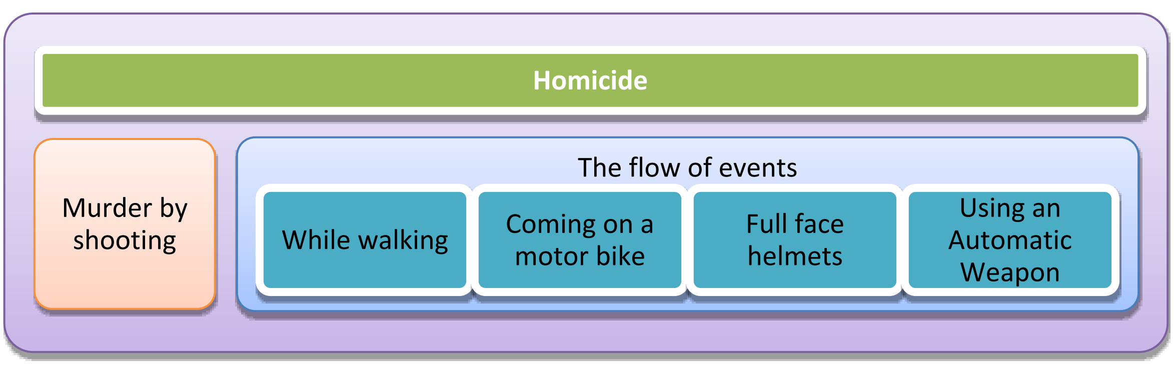

Figure 1. *Relationship between main crime type, subtypes and crime flows*

A subtype is a sub category of one of the main crime types. This flow of crime is written in the GCR book in the nature of the crime section as a description. For investigation, the nature of the crimes is broken into subtypes and flows according to their frequency of occurrence and uniqueness. These sub categorizations have been introduced mainly to minimize the broadness of main type and to improve clarity. Fig.1. depicts the relationship of the subtypes and flows where there can be a flow of events to a crime recorded as one of the 26 main crime types. For the simplicity and easy handling of data, the investigators have provided subtype codes and flow codes. The flow of events provides a modus operandi which is most of the time unique to an offender. Each subtype is provided with a code under the main type, to make the crime investigation process easier. For example, ROB/S001 denotes a subtype that is Highway robbery; here ROB denotes the main type under which the corresponding subtype appears. In this case, it is Robbery. Crime types are further subdivided into sub types to make the analysis and processing simpler. In this manner, crime subtypes and flows have been identified under all the 26 crime types. The space for adding more subtypes and flows under these crime types exists. A new subtype or a flow is introduced to a particular main crime, if the same subtype or the flow happens to persist for a prolonged time.

This paper proposes a novel method of criminal profiling using the modus operandi which can be used to identify associations between crimes and chronic criminals. The method is based on a new technique named, “binary feature vector profiling”. Key relationships between a criminal and the conducted crimes are analyzed using binary feature profiling and association rule mining techniques. Due to the impreciseness and vagueness of these extracted attributes, a fuzzy inference system is used in making the final decision. The newly proposed method was adapted into a classification algorithm in order to test its accuracy. An actual crime data set which was obtained from Sri Lanka Police was used in testing the performance of the newly proposed method and it was compared against nine well established classification algorithms. Comparisons were done using Friedman’s rank tests. The results confirmed that the proposed method produced competitive results compared to the other nine classification algorithms.

The rest of the paper is organized as follows. Related work section presents a summary of the work that has been conducted on modus operandi analysis as well as a brief discussion on crime investigation using link analysis and association mining in general. Materials and Methods section discusses the main steps of the newly proposed algorithm. Next, Results and Discussion section provides a validation and performance evaluation of the newly proposed method along with a performance comparison with nine other classification algorithms. Finally, some concluding remarks and future enhancements are outlined in the conclusion section.

Related work

Literature shows many methods which have been developed in the area of automated crime investigation. Our major concern has been laid upon the research carried out on crime investigation using association mining as our research considers on developing a model to find the associations between the criminals and the crimes depending on the modes operandi. C Bennell and DV Canter [8] have proposed a method to use statistical models to test directly the police practice of utilizing modus operandi to link crimes to a common offender. The results indicated that certain features such as the distance between burglary locations, lead to high levels of predictive accuracy. Craig Bennell, et. al. [9] have tried to determine if readily available information about commercial and residential serial burglaries, in the form of the offender's modus operandi, provides a statistically significant basis for accurately linking crimes committed by the same offenders. Benoit Leclerc, et al. [10] have reviewed the theoretical, empirical, and practical implications related to the modus operandi of sexual offenders against children. They have presented the rational choice perspective in criminology followed by descriptive studies aimed specifically at providing information on modus operandi of sexual offenders against children.

Clustering crimes, finding links between crimes, profiling offenders and criminal network detection are some of the common areas where data mining is applied in crime analysis [11], [12], [13] . Association analysis, classification and prediction, cluster analysis, and outlier analysis are some of the traditional data mining techniques which can be used to identify patterns in structured data. Offender profiling is a methodology which is used in profiling unknown criminals or offenders. The purpose of offender profiling is to identify the socio-demographic characteristics of an offender based on information available at the crime scene [14] [15]. Association rule mining discovers the items in databases which occur frequently and present them as rules. Since this method is often used in market basket analysis to find which products are bought with what other products, it can also be used to find associated crimes conducted with what other crimes. Here, the rules are mainly evaluated by the two probability measures, support and confidence [16], [17]. Association rule mining can also be used to identify the environmental factors that affect crimes using the geographical references [18]. Incident association mining and entity association mining are two applications of association rule mining. Incident association mining can be used to find the crimes committed by the same offender and then the unresolved crimes can be linked to find the offender who committed them. Therefore, this technique is normally used to solve serial crimes like serial sexual offenses and serial homicide [19].

Similarity-based association mining and outlier-based association mining are two approaches used in incident association mining. Similarity-based association mining is used mainly to compare the features of the crime with the criminal's behavioral patterns which are referred as modus operandi or behavioral signature. In outlier-based association mining, crime associations will be created on the fact that both the crime and the criminal have the possibility of having some distinctive feature or a deviant behavior [20]. Entity association mining/link analysis is the task of finding and charting associations between crime entities such as persons, weapons, and organizations. The purpose of this technique is to find out how crime entities that appear to be unrelated at the surface, are actually linked to each other [19]. Link analysis is also used as one the most applicable methods in social network analysis [21] in finding crime groups, gate keepers and leaders [22].

Attribution can be used to link crimes to offenders. If two offences in different places involve the same specific type, those may be readily attributed to the same offender [11].There are three types of link analysis approaches, namely Heuristic-based, Statistical-based and Template-based [19]. Sequential pattern mining is also a similar technique to association rule mining. This method discovers frequently occurring items from a set of transactions occurred at different times [23]. Deviation detection detects data that deviates significantly from the rest of the data which is analyzed. This is also called outlier detection, and is used in fraud detection [23] [24].

In classification, the data points will be assigned to a set of predefined classes of data by identifying a set of common properties among them. This technique is often used to predict crime trends. Classification needs a reasonably complete set of training and testing data since a high degree of missing data would limit the prediction accuracy [23]. Classification comes under supervised learning method [19], [25] which includes methods such as Bayesian models, decision trees, artificial neural networks [26] and support vector machines. String comparison techniques are used to detect the similarity between the records. Classification algorithms compare the database record pairs and determine the similarity among them. This concept can be used to avoid deceptive offender profiles. Information of offenders such as name, address, etc. might be deceptive and therefore the crime database might contain multiple records of the same offender. This makes the process of identification of their true identity difficult [23].

Systems and Methods

This section provides a description about the systems and methods used in developing the fuzzy based binary feature profiling for modus operandi analysis. First, an overview about how SL-CIDSS captures the logics of modus operandi is explained. Then a detailed description about the steps of the newly proposed algorithm is explained.

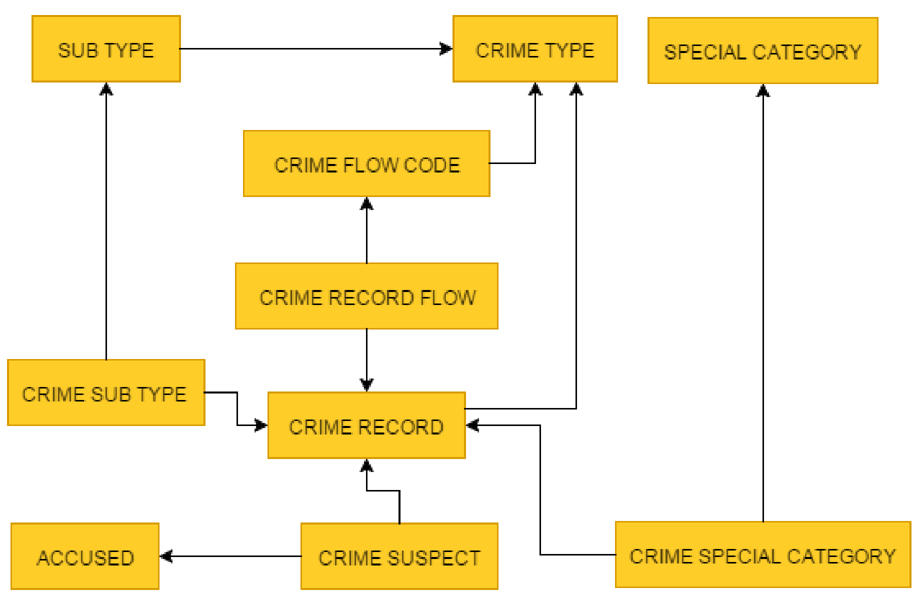

Figure 2. *Crime flow entity arrangement of SL-CIDSS*

Figure 2 shows how SL-CIDSS database captures the crime types and subtypes. A crime record has a crime record flow. Typically, a crime is committed by a criminal and a particular accused might commit one or more crimes. A CRIME RECORD can be of one the 26 crime types. A particular CRIME RECORD will be considered under one main CRIME TYPE with the highest precedence in the order of seriousness. For example, a crime incident that includes a murder and a robbery will be categorized as a murder though a robbery has also taken place. But in the nature of crime section, all crimes followed by the main type will be stated. Therefore, the CRIME RECORD FLOW captures all the steps of the crime as a sequence of steps recorded. The crime flows that have been previously registered are mapped under CRIME FLOW CODE. Also, a particular CRIME RECORD instance can contain multiple SUB TYPES which are recorded as CRIME SUB TYPE. The SPECIAL CATEGORY captures the crimes with special features such as crimes occurring at the same location or retail shop. A crime may involve several special categories which are saved in the CRIME SPECIAL CATEGORY. The ACCUSED entity records the information of suspects and accused and they are related to crime through the CRIME SUSPECT entity.

As the first step of the newly employed method, a feature matrix is generated, resulting in a binary matrix representing the crime flows. This binary feature matrix is composed of the binary patterns generated on previous convictions of a particular criminal/suspect. As the MOs are represented in binary all the analyses are conducted on this binary feature matrix. This binary form of the feature matrix provides a provision to direct application of computer algorithms with methods such as Apriori based association rule mining as the matrix is already in the binary format. The reduced complexity of the binary feature matrices provides an easy manipulation over the categorical and continuous valued features. Figure 3 shows the steps of the proposed MO analysis algorithm.

Step 1: Generate the feature matrix.

Step 2: Generate the dynamic MOs (DMO) of the criminals.

Step 3: Generate the complete MO profile (CMOP) of the criminals.

Step 4: Find deviation probability (DP) of CMOP from the crime MO under consideration (UMO).

Step 5: Find completeness probability of UMO against DMO.

Step 6: Use the two values obtained from step 4 and 5 as inputs of a fuzzy Inference system to obtain the final similarity value (out of 100).

Step 7: Classify the UMO under the class with highest similarity score for validation.

Figure 3. *Steps of the newly employed algorithm*

Generating the binary feature matrix

Table 1 shows how the feature matrix is generated and provides the way to generate modus operandi of criminals as binary sequences. According to the table, events of the crime scene are observed starting from its crime type. After a particular crime type is identified, the feature matrix is updated with ones for each subtype and flow code that is available in the crime or suspect modus operandi. The matrix will be filled by zeros in places which the modus operandi does not have any contact with. The column names to the feature matrix are generated in such a way that it covers the collection of main types, sub types, crime flows and special categories at hand. For example, if we consider the list of crime types, subtypes, crime flows and the special category in Table 1, it results in a feature matrix of a 21-bit vector shown in the last two columns.

Table 1. *An instance of feature selection for the feature matrix generation*

Main Semantic	Crime flow element code	Description	Suspect 1	Suspect 2
Crime types	HB	House Breaking	0	1
	HK	Hurt by Knife	0	0
	RB	Robbery	1	0
	TH	Theft	0	0
Sub types	ABD/S003	Abduction from the legal guardian	1	0
	ABD/S004	Abducting to marry	0	0
	ABD/S005	Abducting for sexual harassment	0	0
	BGL/S004	Use of stealth	0	1
	BGL/S011	Burglary in business places	0	0
	ROB/S001	Organized vehicle robbery	1	0
Crime Flows	BGL/F001	Entering from the window	0	1
	BGL/F002	Entering from the Fanlights	0	0
	BGL/F003	Removing grills	0	1
	BGL/F004	Breaking glasses	0	0
	ROB/F001	Showing identity cards	1	0

	ROB/F003	Wearing uniforms	1	0
	ROB/F004	Robbery using identity cards, uniforms and chains	0	1
	ROB/F009	Seizing inmates	0	0
	ROB/F010	Appearing as CID officers	0	0
Special Category	Retailer 1	Attacking/ robbing retailer 1's stores	0	0
	Retailer 2	Attacking/ robbing retailer 2's stores	0	0

In this manner we can produce binary MO patterns based on the crimes committed by different criminals as shown in the last two columns of Table 1. According to Table 1, Accused 1 has committed a robbery with the subtypes, ABD/S003 (an abduction of a child from the legal guardian), ROB/S001 (an organized vehicle robbery) and the flows, ROB/F001 (Identity cards have been shown),ROB/F003 (accused has been wearing uniforms). Accused 2 has committed a house breaking with the sub type BGL/S011 (use of stealth), and the flows, BGL/F001 (Entering from the window), BGL/F003 (Removing Grills).

In this manner, depending on the complete crime MO under consideration, it may generate modus operandi of different lengths. According to the full crime MO, criminal based MOs can be generated and taken into a full feature matrix of binary patterns. *ct, st, fl* and *sc* in Table 2 represent the abbreviations for crime type, sub type, flow and special category respectively.

Table 2. *Feature matrix generated using the selected modus operandi attributes in Table 1.*

	ct1	ct2	ct3	ct4	st1	st2	st3	st4	st5	st6	fl1	fl2	fl3	fl4	fl5	fl6	fl7	fl8	fl9	sc1	sc2
Accused 1	0	0	1	0	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0
Accused 2	1	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	0	0

Generating the dynamic MOs (DMOs) of the criminals

Dynamic MO is a binary feature vector which is generated on bit patterns of the feature matrix of a particular criminal. The main purpose of the DMO is to obtain a criminal specific crime flow which captures the crime patterns which are frequently followed by a particular criminal. It was named as the dynamic modus operandi as it is subject to change when the new crime flows are added to the feature matrix. Therefore, this addresses the changing nature of the patterns used by the criminals in committing crimes. First, a frequency threshold is generated using characteristic features of the sub matrix at hand which is the matrix of all crimes committed by the same criminal under consideration. If the corresponding accused is convicted for four crimes, the four bit patterns related to those four crimes will be available in the feature matrix.

The matrix shown in Table 3 is an example to a situation of a feature matrix generated on the previous convictions of a criminal. For simplicity let's consider a feature matrix of 10 columns.

Table 3. *Feature matrix generated on four previous convictions of a criminal*

A	B	C	D	E	F	G	H	I	J
1	0	0	0	1	1	1	0	1	0
1	0	0	1	0	1	1	0	1	0
1	0	0	0	1	1	0	1	1	0
1	0	0	1	0	1	1	1	1	0

If we consider A to J of Table 3 as crime flow features of the corresponding MOs, we can understand that in the first MO the criminal has followed a flow of A-E-F-G-I. The same criminal has followed a crime flow of A-D-F-G-I in his second crime. Likewise the other two flows are, A-E-F-H-I and A-D-F-G-H-I respectively.

The DMO of a particular criminal is generated using the Apriori method [27]. Apriori method is used to find the crime entities with the frequency threshold (frt) which is generated according to Equation 2. Apriori is a method to find frequent item sets in transactions in association rule mining [27]. A demonstration of the generation of D in Equation 1 on the properties of feature matrix is shown in Table 4.

$$D = \left\{ d \middle| d = \sum_{i=1}^{n} y_i \right\} \tag{1}$$

$$frt = {}^{M_D}\!/_{n} \tag{2}$$

Where,

y_i = cells in each column

M_D = Median of D,

$n = \sum f$ = number of values or total frequencies,

c = cumulative frequency of the median class

h = class interval size.

Table 4. *Column-wise addition of the feature matrix of the suspect under consideration*

A	B	C	D	E	F	G	H	I	J
1	0	0	0	1	1	1	0	1	0
1	0	0	1	0	1	1	0	1	0
1	0	0	0	1	1	0	1	1	0
1	0	0	1	0	1	1	1	1	0
4	0	0	2	2	4	3	2	4	0

Summation

The column-wise addition of the matrix shown in Table 4 gives 4, 0, 0, 2, 2, 4, 3, 2, 4 and 0. The distinct numbers are selected from the resulting vector which results in 4, 0, 2 . The vector with the distinct numbers is named as D. The median of D is then divided by the number of instances (rows) in the matrix as the frt, which is 2.5/4 = 0.625 for the above case. Therefore, frt will range from 0 to 1. This value provides an insight to a fair threshold value for the Apriori method to generate the dynamic modus operandi with the most frequent elements. frt is used as the frequency threshold in finding the lengthiest MO with a probability of 0.625 because this value suggests that there is a moderate possibility of one feature having 0.625 probability in each of MO. This results in a dynamic modus operandi (DMO) as shown in Equation 4, because the only transaction of crime attributes which provides a support of 0.625 is $\sigma\,(A,F,G,I)$ as shown in Equation 3.

$$s = \frac{\sigma(A,F,G,I)}{|T|} = \frac{3}{4} = 0.75 \tag{3}$$

$$\text{DMO} = [\,1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0\,] \tag{4}$$

Generating the complete MO profile (CMOP) of the criminals

The complete MO profile (CMOP) is obtained by the OR operation between the bits of each column of the feature matrix of the corresponding criminal. CMOP guarantees the provision of a composite crime flow by considering all of the previous crime flow entities of a particular criminal. For example, the complete profile for the feature matrix shown in Table 3 is obtained as shown in Table 5.

Table 5. *OR operation on the columns to obtain the complete MO profile*

A	B	C	D	E	F	G	H	I	J
1	0	0	0	1	1	1	0	1	0
1	0	0	1	0	1	1	0	1	0
1	0	0	0	1	1	0	1	1	0
1	0	0	1	0	1	1	1	1	0
1	0	0	1	1	1	1	1	1	0

OR Operation

Therefore, $CMOP = [1\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 0]$. CMOP contains 1s for each place for which a particular crime flow entity has taken place at least once.

Finding deviation probability (DP) of CMOP from the crime MO under consideration (UMO)

First, the deviation of each individual flow in the criminals feature matrix is obtained according to Equation 5. As the binary feature vectors are commonly used to represent patterns, many methods have been invented to find their similarity and distance [28]. Euclidean distance, Hamming distance, Manhattan distance, Jaccard, Sorensen and Tanimoto are few of the frequently used measures in that domain [28]. This probability value which is named as the deviation probability (DP), is used to obtain a measurement as to what extent of information is available in the UMO, extra to what is already available in the CMOP of a particular criminal. Let's assume that the bit pattern to be compared with the suspect's modus operandi profile under consideration is UMO = [1 0 0 0 1 1 1 0 0 1]. Therefore, DP provides the probability of 1s which are available in UMO but not in CMOP.

The deviation probability, DP can be given as,

$$DP = \frac{\sum_{i=1}^{n} x_i - y_i}{n} \text{ ,for } x_i = 1, y_i = 0 \text{ ; } i = 1,2,\ldots,n \tag{5}$$

Where,

$x_i = elements\ of\ the\ UMO$

$y_i = elements\ of\ the\ CMOP$

If we consider the feature matrix on Table 5,

$$Deviation = [1\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 0\ 1] - [1\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 0]$$

$$Deviation = [0\ 0\ 0 - 1\ 0\ 0\ 0\ \ - 1 - 11] \tag{6}$$

Define $AD = 1$, where AD is the number of positive 1s.

Therefore, $DP = 1/10 = 0.1$

As it appears in Expression 6, it produces positive 1s for the places with the features available in UMO but not in CMOP. The higher the DP, higher the amount of extra information available in UMO. Hence, a DP value close to 0 indicates the absence of extra features in UMO.

Finding completeness probability (CP) of UMO against DMO

For the same feature matrix which was considered in Table 3, the CP is obtained according to Equation 7. Here, the UMO is compared with DMO to obtain a probability to determine what extent of features in CP is available in UMO. Therefore, it is derived by the percentage of attributes which are present in both UMO and DMO.

Let $DMO = \{x_i\}_{i=1}^{n}$ and $UMO = \{y_j\}_{j=1}^{n}$ be two binary sequences.

Define, $z_k = \begin{cases} 1; x_i = y_j \\ 0; otherwise \end{cases}$. Then, $CP = \frac{\sum_{k=1}^{n} z_k}{n}$ is the completeness probability. (7)

For example, if we consider $DMO = [1\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 1\ 0]$, then for the $UMO = [1\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 0\ 1]$ a CP of $3/10 = 0.3$ is generated as in the 1st, 6th and 7th positions there are ones in both DMO and UMO. Higher the CP value, the more the UMO is composed of crime flow entities which are available in the DMO. Therefore, a CP value close to 1 indicates that the completeness of UMO compared to DMO is 100%.

Building a fuzzy inference system to obtain the final similarity score

The vagueness of the two measurements CP and DP generates a difficulty in calculating a similarity score using crisp logic. Therefore, the two parameters CP and DP were adapted into a fuzzy inference system which accepts two inputs and provides a score for the similarity between the DMO and UMO. Figure 4 shows a block diagram of the proposed fuzzy inference system. Mamdani fuzzy inference was proposed as an attempt to solve a control problem by a set of linguistic rules obtained from experienced human operators [29]. First, the rule base of the fuzzy controller was defined by observing the variations of CP and DP. The membership functions of the inputs and outputs were then adjusted in such a way that, the parameters which seem to be wrong can be fine-tuned, which is a common practice in defining fuzzy inference systems [30]. Literature shows many methods used in fine tuning the fuzzy parameters. Usage of adaptive networks [31] and Neuro-fuzzy systems [32] in fine tuning the fuzzy parameters have received more attention. The problem at our hand was to generate a fuzzy inference system which generates the highest similarity score when the DP value goes down and CP value goes up. We conducted a manual mapping procedure for the fuzzy membership functions. Therefore, the input and output space of the two inputs CP and DP and the output were partitioned into 3 subsets. Namely, LOW, MODERATE and HIGH. Center of gravity was used as the defuzzification strategy of the fuzzy controller. Mamdani fuzzy inference was especially selected for the similarity score generation procedure, for the highly intuitive knowledge base it offers due to the fact that both antecedents and the consequents of the rules are expressed as linguistic constraints [33]. First, we selected all of these membership functions with 50% overlap. Then the tuning procedure was conducted during which we adjusted either the left and/or right spread and/or overlapping to get the best possible similarity score for the given DP and CP. This procedure was conducted until the FIS generated satisfactory results.

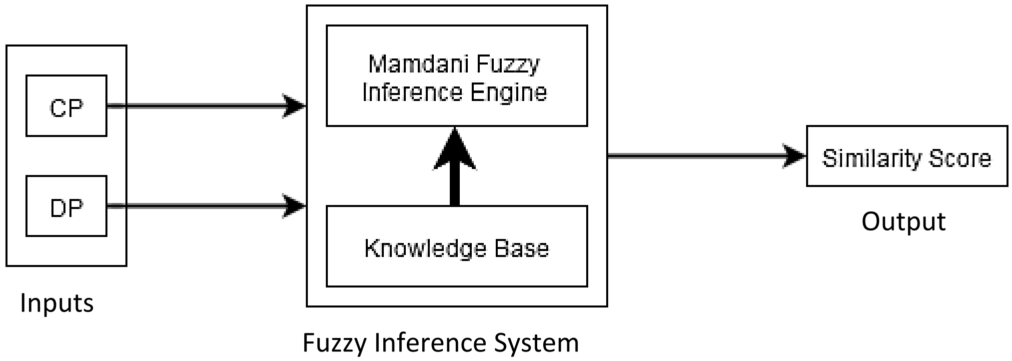

Figure 4. *Block diagram of the proposed fuzzy inference system.*

Figures 5 and 6 show the fuzzy inputs of the Fuzzy Inference System (FIS) which correspond to CP and DP values respectively. Figure 7 depicts the fuzzy output of the FIS. As the Figures 5, 6 and 7 depict, all the different levels of membership functions under each input and the output are selected to be triangular and trapezoidal functions as triangular or trapezoidal shapes are simple to implement and computationally efficient [34].

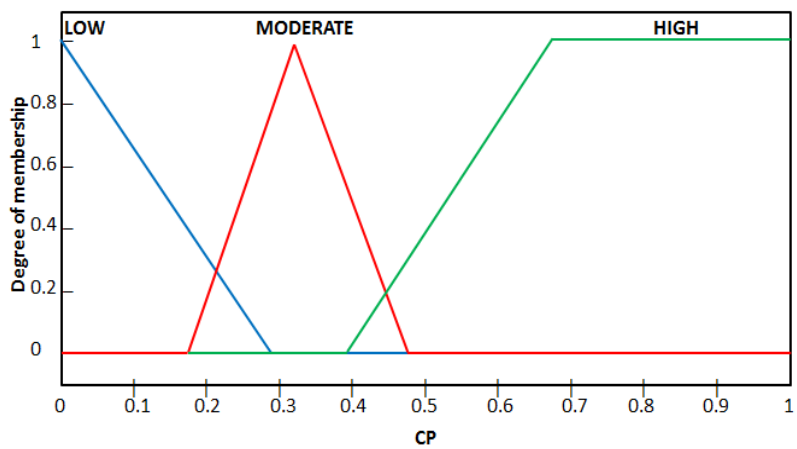

Figure 5. *Input fuzzy variable 1: CP*

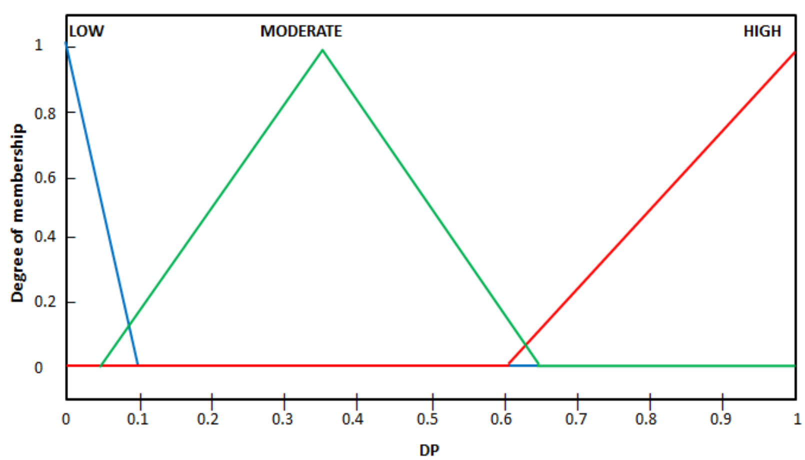

Figure 6. *Input fuzzy variable 2: DP*

.

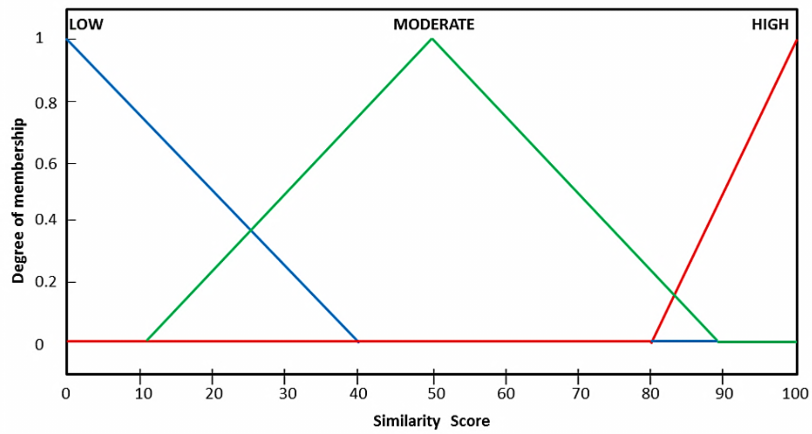

Figure 7. *Output fuzzy variable: similarity score.*

As shown in Figure 7, the universe of discourse of similarity score (fuzzy output) ranges from 0 to 100. The defuzzifier score which is generated from the FIS is considered as the measurement for how close the modus operandi under consideration is to a particular suspect's profile. A higher score value close to 100, which is generated from the FIS provides a good indication about a high similarity between the modus operandi of the crime and suspect under consideration.

The fuzzy rule derivation of the fuzzy controller is heuristic in nature. According to the calculations of the two inputs, higher values of CP, close to 1 and lower values of DP close to 0, positively affect the final similarity score. Therefore, the rule base of the fuzzy model is generated accordingly. Our rule base provides a non-sparse rule composition of 9 combinations as illustrated in Figure 8.

CP \ DP	***LOW***	***MODERATE***	***HIGH***
LOW	MODERATE	LOW	MODERATE
MODERATE	HIGH	MODERATE	LOW
HIGH	HIGH	MODERATE	LOW

Figure 8. Fuzzy rule set of the rule base of the inference system.

The rule surface of the fuzzy controller depicted in Figure 9, shows the variation of the score value with the changes of the two inputs CP and DP. According to the figure it's perfectly visible, for higher values of CP (close to 1) and for lower values of DP (close to 0), the fuzzy controller generates higher values for the score which are close to 100.

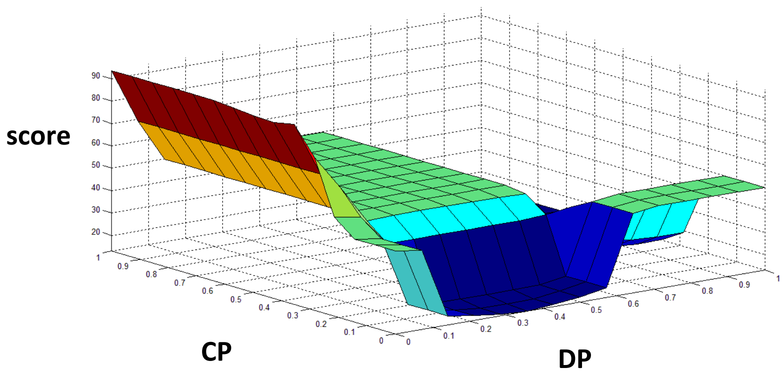

Figure 9. Rule surface of the fuzzy controller.

Classification of the UMO under the class with the highest similarity score

When the algorithm is used to find associations between modus operandi of criminals and modus operandi of crimes, the similarity score which is generated from the newly proposed method can be used directly. A similarity score which is close to 100 would suggest that the criminal has a very high tendency to have committed the crime which is under investigation. Therefore, the similarity scores generated can be used to classify a particular modus operandi to a most probable suspect with the highest similarity score.

The proposed method was developed by using MATLAB 7.12.0 (R2011a) [35]. All the necessary implementations were conducted using the MATLAB Script editor [36] apart from the FIS which was implemented using the MATLAB fuzzy toolbox [37]. The nine classification algorithms which were used for the performance comparison were classification algorithms which are already packaged with the WEKA 3.6.12 tool [38].

Results and Discussion

The method was tested with a crime data set obtained from Sri Lanka Police. Figure 10 shows the crime frequencies in Sri Lanka by the crime types from 2005 to 2011. It shows only 21 crime types because the 5 new crime types were introduced in 2015. 4^{th} column denoting House Breaking and Theft shows the highest number of occurrences. 14: Theft of property, 10: Robbery, 13: Cheating/ Misappropriation, 6: Hurt by Knife, 7: Homicide, 8: Rape/ Incest, 5: Grievous Hurt, 3: Mischief over Rs. 5000/=, 1: Abduction/Kidnapping comes next. For the validation of the algorithm, 7 crime types out of these 10 types were selected for the testing data set. They are, House Breaking and Theft, Theft of property, Robbery, Homicide, Rape/

Incest, Grievous Hurt, Abduction/Kidnapping. 31 crime flows were selected which are common to the seven selected crime types. The data set is also composed of 8 sub types and 2 special categories. Altogether the data set consisted of 67 instances in which each instance is composed of 48 attribute values. The data set is distributed over 20 classes (criminals).

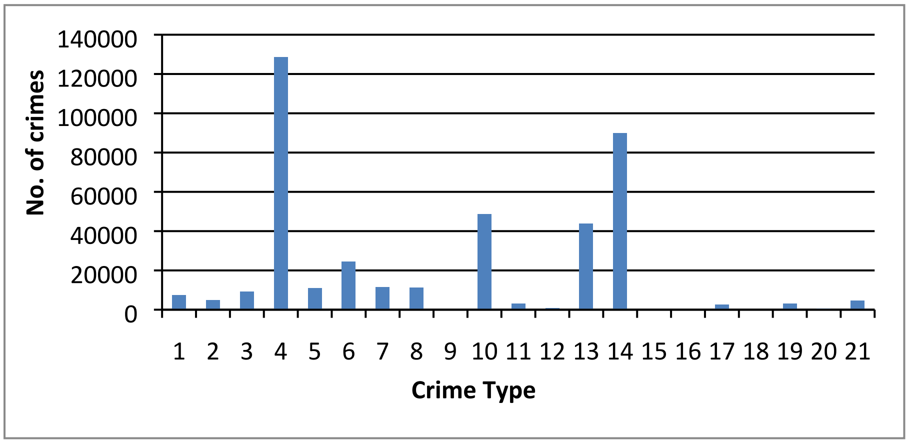

Figure 10. *Frequency of different crime types from year 2005-2011*

Figure 11. *Distribution of modus operandi instances over the classes of the dataset*

All the tests were performed in a Windows computer with Intel (R) Core (IM) i7-2670QM CPU of 2.20 GHz and a RAM of 8GB. The histogram of the instance distribution over the classes is shown in Figure 11.

10 fold cross validation [39] was used on the data set for a fair testing procedure. In 10-fold cross validation, the data set is divided into 10 subsets, and the holdout method is repeated 10

times. Each time, one of the 10 subsets is used as the test set and the other 9 subsets are put together to form a training set. Then the average error across all 10 trials is computed [39].

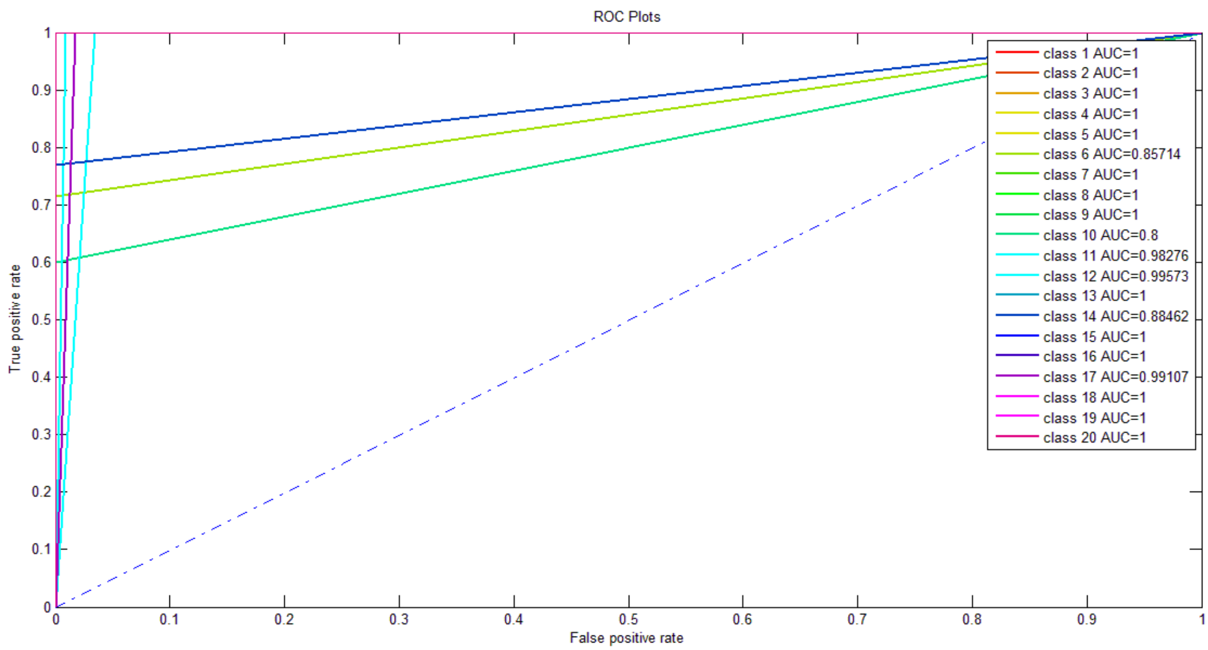

Figure 12. *ROC curves returned by the newly proposed method on the 20 classes of crime data set*

The test results of modus operandi classifications in Area Under Curve (AUC) [40], Root Mean Squared Error (RMSE) and time elapsed for the classification are shown in Table 6. A Receiver Operating Characteristic (ROC) curve is a two dimensional graphical illustration of the trade-off between the true positive rate (sensitivity) and false positive rate (1-specificity). Figure 12 depicts the ROC curve plotted on the classification results obtained by the newly proposed method on the crime data set. In the particular instance which is shown in Figure 12, all the ROC curves related to the crime data set are plotted well over the diagonal line and all of them have retuned AUC values which are either equal to 1 or very close to 1, providing a very good classification.

The sole intention of this research was to find out relationships among the modus operandi of criminals with the modus operandi found in the crime scenes to find the associations. To prepare the data set which was used under this research, a crime data set of around 3000 instances was analysed. Due to limitations of the real crime data set, it was quite a complex task to prepare a data set with a collection of sufficient modus operandi where each instance has a considerable flow of crime flows. Therefore, only a sample of 67 instances could be filtered from the population to generate a representative data set and it was verified by a domain expert before being used in the analysis. As the number of instances was around 67, it can be assumed to be an under-represented data sample. Another reason for the data set to become under- represented was the challenge in finding classes/criminals with more than one crime committed. The actual crime data set which is used for the testing purposes is imbalanced as it is apparent in Figure 11. For example the data set is composed of 4 single instance classes while there are seven classes with four instances each. With a classification data set like this, there is a very high tendency of getting biased results. To make the classification process unbiased, we used the concept of oversampling. Oversampling and under-sampling are two concepts which are used in overcoming class imbalance problems in input data sets. Oversampling and under-sampling are two different categories of resampling approaches, where in oversampling the small classes are incorporated with repeated instances to make them reach a size close to lager classes, whereas in under-sampling, the number of instances is deceased in such a way that the number of instances reach a size close to the smaller classes [41].

Table 6 shows the results returned by the fuzzy based binary feature profiling which was conducted on the actual crime data set. As shown in the table, there is an increase in the accuracy when the input data set undergoes oversampling. Since the maximum number of instances available under one suspect is equal to 5, under-sampling does not provide a good accuracy. The results prove that the new algorithm works well for a balanced data set as the new method showed an increase in performance when the data set is subjected to an oversampling greater than or equal to 5 which is the highest number of instances under a particular class.

Table 6. *Results returned by the fuzzy based binary feature profiling for the modus operandi analysis on actual data*

Data set (Number)	Oversampling or Under-sampling value	AUC	Root Mean Squared Error	Average time elapsed
1	2	0.5417	0.6986	0.0015
2	3	0.5562	0.4969	0.0011
3	4	0.5965	0.4303	0.0014
4	N/A	0.6937	0.7000	0.0010
5	5	0.6612	0.4126	0.0011
6	6	0.7063	0.2959	0.0011
7	10	0.8033	0.1396	0.0012
8	20	0.9339	0.0941	0.0013
9	30	0.9661	0.0853	0.0014
10	40	0.9637	0.0551	0.0015
11	50	0.9756	0.0578	0.0016
12	60	0.9626	0.0638	0.0018
13	70	0.9365	0.0792	0.0019
14	80	0.9391	0.0784	0.0023

15	90	0.9671	0.0752	0.0029

Figure 13 shows the change in AUC with the increase of sampling which starts from under-sampling of 2 and goes on to an over sampling of 90. According to the plot it can be observed that the ROC values are increased when the oversampling is increased.

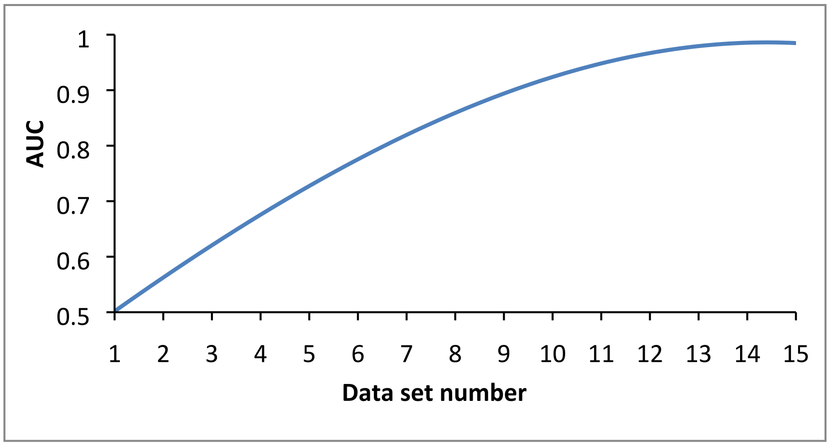

Figure 13. *Change of ROC values with oversampling*

Figure 14 depicts the change of RMSE with the increase of sampling. The plot clearly illustrates that the RMSE values are decreased with the oversampling.

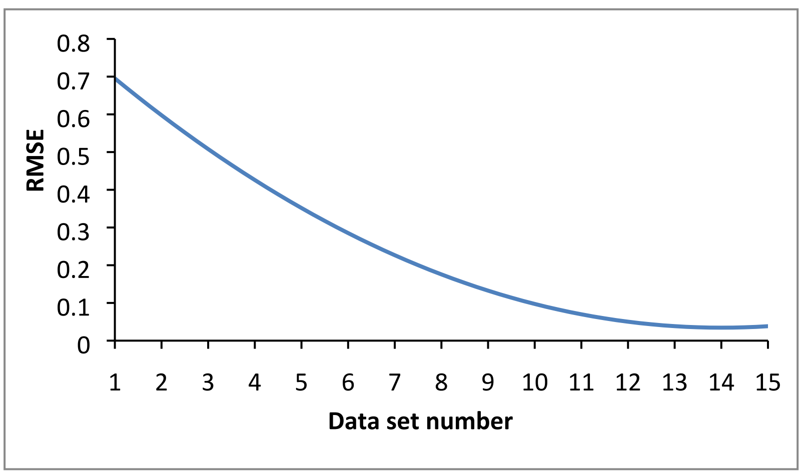

Figure 14. *Change of RMSE value with oversampling*

The execution time of the algorithm was 0.001s when there is no oversampling or under-sampling. The maximum execution time is 0.0031 when there is an oversampling of 90. According to the plot shown in Figure 15, it is clear that there is an increase of execution time as the oversampling size increases. But, the overall execution time is always remained under 3 milliseconds.

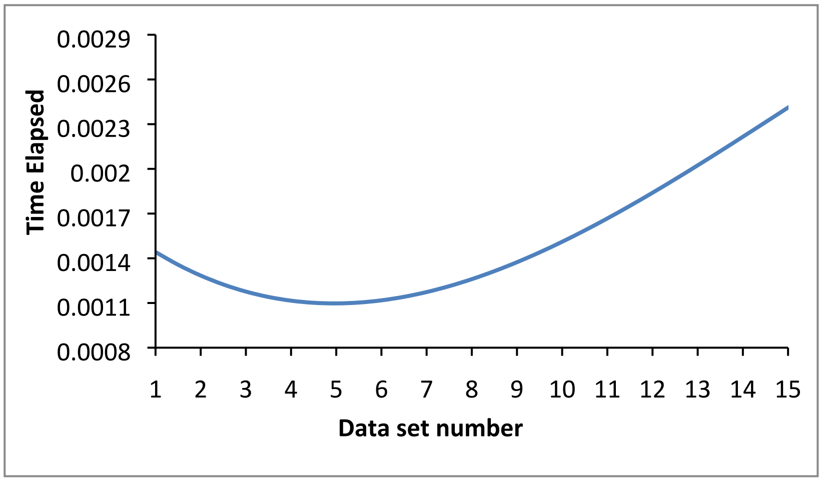

Figure 15. *Change of time elapsed for the 15 data sets.*

Overview of the classification algorithms used for the comparison

It is a known fact that there is no single algorithm which can be categorized as the best to solve any problem. Different classification algorithms may perform differently in different situations [42]. Therefore, the newly proposed method was then tested against ten other open classification data sets and the performance was evaluated against the results obtained with nine well-known classification techniques, thereby assessing the quality of the newly proposed method The nine classification techniques used for this purpose include, Logistic Regression, J48 Decision Tree, Radial Basis Function Network (RBFNetwork), Multi-Layer Perceptron, Naive Bayes Classifier, Sequential Minimal Optimization (SMO) algorithm, KStar instance based classifier, Best-first decision tree (BFTree) classifier, and LMT (Logistic Model Tree) classifier. These classifiers represent four classes of classification algorithms. Namely, function based classifiers, Tree based classifiers, Bayesian classifiers and Lazy classifiers.

Logistic Regression learns conditional probability distribution. Relating qualitative variables to other variables through a logistic cumulative distribution functional form is logistic regression [43]. J48 is an open source java implementation of the C4.5 decision tree algorithm [44]. A decision tree consists of internal nodes that specify tests on individual input variables or attributes that split the data into smaller subsets, and a series of leaf nodes assigning a class to each of the observations in the resulting segments. C4.5 algorithm constructs decision trees using the concept of information entropy [45]. Neural networks are flexible in being modeled virtually for any non-linear association between input variables and target variables [46]. Both Radial basis function networks and Multilayer perceptron (MLP) networks are neural networks

[47]. Bayesian classifiers assign the most likely class to a given example described by its feature vector [48]. SMO is an implementation of John Platt's sequential minimal optimization algorithm for training a support vector classifier. It globally replaces all missing values and transforms nominal attributes into binary one. It also normalizes all attributes by default [49] [50]. KStar (K*) is an instance-based classifier which uses an entropy –based distance function [51]. BFTree (Best- First Decision Tree) uses binary split for both nominal and numeric attributes [52]. LMT is a classifier for building 'logistic model trees', which are classification trees with logistic regression functions at the leaves [53], [54].

Table 7. *Description of the classification data sets for performance comparison*

Data set	Description	Number of Instances	No of Attributes
Dermatology Data Set [55]	This database has been created on a dermatology test carried out on skin samples which have been taken for the evaluation of 22 histopathological features. The values of the histopathological features have been determined by an analysis of the samples under a microscope. In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values.	336	33
Balance Scale Data Set [56]	This data set has been generated to model psychological experimental results. Each example is classified as having the balance scale tip to the right, tip to the left, or be balanced. The attributes are the left weight, the left distance, the right weight, and the right distance. The correct way to find the class is the greater of (left-distance * left-weight) and (right-distance * right-weight). If they are equal, it is balanced. There are 3 classes (L,B,R), five levels of Left-Weight (1,2,3,4,5), five levels of Left-Distance (1,2,3,4,5), five levels of Right-weight (1,2,3,4,5) and five levels of Right-Distance (1,2,3,4,5).	625	4
Balloons Data Set [57]	This data set has been generated using an experiment of stretching a collection of balloons carried out on a group of adults and children [58]. In the data set, Inflated is true if (color=yellow and size = small) or (age=adult and act=stretch). In the data set there are two main output classes, namely T if inflated and F if not inflated, two colors yellow and purple, two sizes, large and small, two act types,	20	4

	stretch and dip, and two age groups, adult and child.		
Car Evaluation Data Set [59]	Car Evaluation Database has been derived from a simple hierarchical decision model originally developed for the demonstration of DEX by M. Bohanec and V. Rajkovic [60]. The Car Evaluation Database contains examples with information that is directly related to CAR. They are buying, maint, doors, persons, lug_boot and safety. The attribute buying is the buying price which is considered to have four levels v-high, high, med, low. Maint is the price of the maintenance which contains the four levels, v-high, high, med, low. Doors have the four levels 2, 3, 4, 5-more. Person (capacity in terms of persons to carry), lug_boot (the size of luggage boot) and safety (estimated safety of the car) have 3 levels each.	1728	6
Soybean Data set [61] [62]	This is a small subset of the original soybean database. The data set is distributed over four classes, D1, D2, D3 and D4. The 35 categorical variables represent different levels of qualities of the soybean vegetable. These categorical variables include, plant-stand, precip, temp, hail, crop-hist, area-damaged, severity, seed-tmt, germination, lant-growth, leaves, leafspots-halo, leafspots-marg, leafspot-size,leaf-shread, leaf-malf, leaf-mild, stem, lodging, stem-cankers, canker-lesion, fruiting-bodies, external, mycelium, int-discolor, sclerotia, fruit-pods, fruit,seed, mold-growth, seed-discolor, seed-size, shriveling and roots. The number of levels represented by each variable varied from 2 to 3.	47	35
Lenses Data set [63]	Lenses data set is a small database about fitting contact lenses. The data set is composed of five attributes including the class variable. The data set has three classes. Age of the patient, spectacle prescription, astigmatic, tear production rate are the attributes of the data set. The attributes contain at least of two categories and at most of three categories.	24	4
Nursery Data set [64]	Nursery Database has been derived from a hierarchical decision model originally developed to rank applications for nursery schools. It has been used during several years in 1980's when there has been excessive enrollment to these schools in Ljubljana, Slovenia, and the rejected applications frequently needed an objective explanation. The final decision depended on three sub problems: occupation of parents and child's nursery, family structure and financial standing, and social and health picture of the family. The model has been developed within expert system shell for decision making [65].	12960	8
Tic-tac-toe	This database encodes the complete set of possible board	958	9

Data set [66]	configurations at the end of tic-tac-toe games, where "x" is assumed to have played first. The target concept is "win for x" (i.e., true when "x" has one of 8 possible ways to create a "three-in-a-row").		
SPECT Heart Data set [67]	The data set describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images. It has 267 instances that are described by 23 binary attributes including the class variable.	267	22
MONK's Problems Data set [68]	The MONK's problems have been the basis of a first international comparison of learning algorithms. The result of this comparison is summarized in "The MONK's Problems. There are three MONK's problems. The domains for all MONK's problems are the same. The data set is composed of 7 attributes and a binary class variable.	432	7

As the newly proposed method accepts only binary input variables, the data sets which are used for the analysis must be preprocessed into the acceptable format. For example, the "balance scale" data set is composed of 4 attributes. Table 8 shows the attributes and their information of the balance scale data set.

Table 8. *Attribute information of the balance data set*

Attribute	Number of Categories	Categories
Class Name	3	L, B, R
Left-Weight	5	1,2,3,4,5
Left-Distance	5	1,2,3,4,5
Right-Weight	5	1,2,3,4,5
Right-Distance	5	1,2,3,4,5

Therefore, the data set was adjusted as shown in figure 16, prior to using it with the proposed method. Each category of a particular attribute is represented by a dummy variable. For example, Left-Weight attribute results in 5 attributes in the preprocessed data set and each attribute is represented using 5 binary variables as LW1, LW2, LW3, LW4 and LW5 where the presence of the attribute denotes 1 and 0 otherwise. As depicted in Figure 16, if Left-Weight has a value of 2 in an instance it results in 1 for the corresponding derived attribute that is LW2. Therefore, if there is an instance where Left-Weight=2, Left-Distance=1, Right-Weight=3 and Right-Distance=4, Class Name=B, it is represented as LW1=0,LW2=1, LW3=0, LW4=0, LW5=0, LD1=1, LD2=0, LD3=0, LD4=0, LD5=0, RW1=0, RW2=0, RW3=1, RW4=0, RW5=0, RD1=0, RD2=0, RD3=0, RD4=1, RD5=0, Class Name=B.

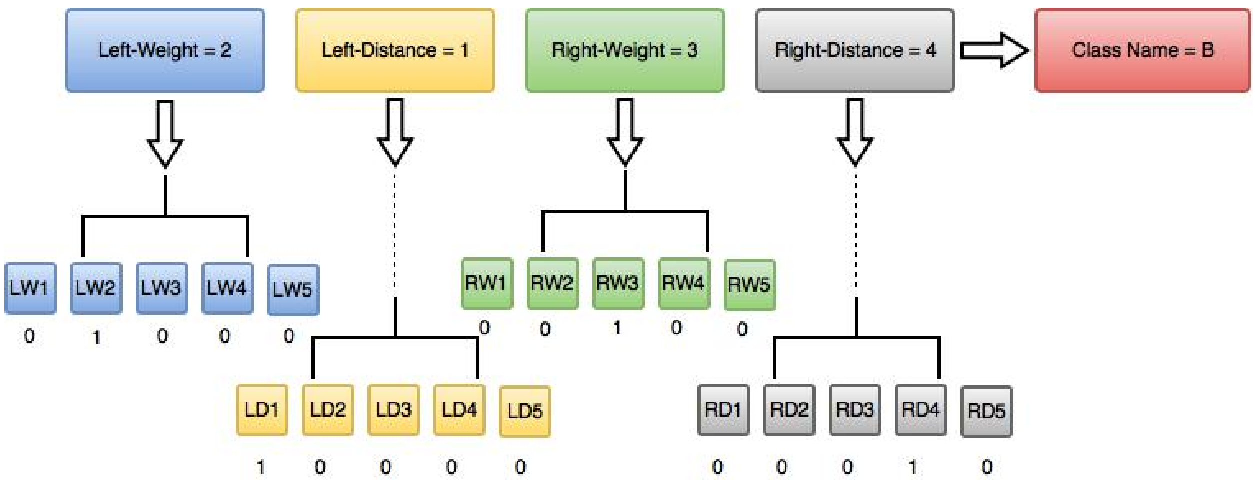

Figure 16. *Schematic diagram for pre-processing of the balance dataset in such a way that it matches the format of inputs of the newly proposed method*

The pre-processed data is then fed to the newly proposed algorithm and the nine other algorithms. Performances were compared based on AUC, root mean squared error (RMSE) and the processing time for model generation. 10 fold cross validation was used under each test for fair testing procedure. Then the receiver operating characteristic curves of the results were analysed. For simplicity the newly proposed modes operandi analysis algorithm was acronymed as BFPM (Binary feature profiling methodology).

As all the data sets which were used for the tests are composed of multi classes, weighted average AUC was used, where each target class is weighted according to its prevalence as given in Equation 8. Weighted average was used in order to prevent target classes with smaller instance counts from adversely affecting the results [69].

$$AUC_{weighted} = \sum_{\forall c_i \in C} AUC\,(c_i) \times p(c_i) \tag{8}$$

Table 9 shows the weighted average AUC values obtained for each data set under each classification algorithm.

Table 9. *Weighted average AUC values obtained by the algorithms on classifying the data sets*

	BFPM	Logistic Regression	J48	Radial Basis Function Network	Multi-Layer Perceptron	Naive Bayes Classifier	SMO	KStar	BFTree	LMT
Dermatology Data set	1	0.9990	0.9750	0.9860	0.9980	0.9980	0.9930	0.9970	0.9690	0.9960
Balance scale Data set	0.7945	0.9760	0.8110	0.9680	0.9770	0.9710	0.8830	0.9510	0.8130	0.9810
Balloons Data Set	1	1	1	1	1	1	1	1	1	1
Car evaluation Data set	0.8087	0.9900	0.9760	0.9740	1	0.9760	0.9550	0.9970	0.9940	0.9990
Soybean Data set	1	1	0.9860	1	1	0.9760	1	1	0.9740	1
Lenses Data set	0.9537	0.7470	0.8400	0.9170	0.8390	0.8700	0.7250	0.8870	0.8670	0.7980
Nursery Data set	0.9100	0.9880	0.9950	0.9870	1	0.9820	0.9640	0.9980	0.9990	0.9990
Tic-tac-toe Data set	0.9167	0.9960	0.8970	0.7340	0.9940	0.7440	0.9760	0.9990	0.9450	0.9920
SPECT Heart Data set	0.7857	0.8310	0.7560	0.8400	0.7860	0.8490	0.7070	0.7850	0.7230	0.8410

MONK's Problems Data set	0.833 3	0.7050	0.994 0	0.8130	0.9980	0.7120	0.746 0	0.997 0	0.955 0	0.988 0

Friedman's rank test is a nonparametric test analogous to a standard one-way repeated-measures analysis of variance [70]. The Friedman's rank test results returned on the AUC test data are shown in Table 10. According to the table, the highest mean rank is returned for MLP while the lowest mean rank is returned for SMO, proving that MLP provides the best performance while SMO provides the least performance for the 10 data sets tested. Therefore, it indicates that the new model provides a better performance than BFTree, J48 and SMO algorithms for the 10 data sets tested.

Table 10. *Friedman's mean rank values returned on the data available in Table 9*

Method	**Mean Rank**
MLP	7.70
LMT	7.10
KStar	6.95
LogisticRegression	5.95
RBFNetworks	5.05
NaiveBayesClassifier	5.05
BFPM	**4.95**
BFTree	4.50
J48	4.20
SMO	3.55

Table 11 shows the RMSE (Root Mean Squared Error) values obtained on the 10-fold cross validation results obtained by each algorithm. Friedman's rank test was conducted on the RMSE values obtained by the ten algorithms. The test returned the results shown in Table 12. According to the mean rank values, the new model provides a better accuracy than RBFNetworks, BFTree, KStar, NaiveBayesClassifier and SMO in the means of RMSE for the ten data sets tested.

Table 11. *RMSE values returned by each algorithm on the classification of the ten data sets*

	BFPM	Logistic Regression	J48	Radial Basis Function Network	Multi-Layer Perceptron	Naive Bayes Classifier	SMO	KStar	BFTree	LMT
Dermatology Data set	0.0881	0.0748	0.1163	0.1091	0.0764	0.0876	0.3110	0.0932	0.1327	0.0889
Balance scale Data set	0.3491	0.2092	0.3699	0.2464	0.2055	0.2793	0.3427	0.2741	0.3551	0.2227
Balloons Data Set	0	0	0	0.0010	0.0264	0.2385	0	0.2382	0	0.1618
Car evaluation Data set	0.1766	0.1520	0.1718	0.1983	0.0440	0.2162	0.3202	0.1990	0.1124	0.0920
Soybean Data set	0	0	0.1031	0	0.0329	0.1031	0.3118	0.0004	0.1061	0.1229
Lenses Data set	0.1737	0.4714	0.3249	0.3324	0.3812	0.3326	0.3967	0.3296	0.3337	0.3780
Nursery Data set	0.2981	0.1456	0.0951	0.1501	0.0174	0.1767	0.3204	0.1928	0.0387	0.0637
Tic-tac-toe Data set	0.4172	0.1289	0.3459	0.4391	0.1611	0.4319	0.1292	0.2788	0.2414	0.2055
SPECT Heart Data set	0.3928	0.3529	0.3866	0.3377	0.4069	0.4188	0.4327	0.3950	0.4022	0.3405

MONK's Problems Data set	0.3881	0.4202	0.1416	0.4067	0.0432	0.4179	0.5036	0.3026	0.2787	0.1955

Table 12. *Friedman's rank test values returned on the data available in Table 11*

Method	Mean Rank
MLP	3.70
LogisticRegression	4.00
LMT	4.60
J48	5.15
BFPM	**5.20**
RBFNetworks	5.40
BFTree	5.60
KStar	6.00
NaiveBayesClassifier	7.25
SMO	8.10

The average processing times for each algorithm on the classification of the ten data sets are given in Table 13. Friedman's rank test on the data of Table 13 returned the results shown in Table 14 in which the mean rank values prove better efficiency for the new method than J48, LogisticRegression, SMO, RBFNetworks, BFTree, MLP and LMT.

Table 13. *Average processing time for each algorithm on the classification of the ten data sets*

	BFPM	**Logistic Regression**	**J48**	**Radial Basis Function Network**	**Multi-Layer Perceptron**	**Naive Bayes Classifier**	**SMO**	**KStar**	**BFTree**	**LMT**
Dermatology Data set	0.0027	0.3900	0.0800	0.3800	2.7300	0.0500	0.1400	0.2800	0.1100	2.9000
Balance scale Data set	0.0030	0.0300	0.0800	0.2200	0.4800	0	0.0500	0	0.0600	0.8400
Balloons Data Set	0	0	0	0	0.0200	0	0.0200	0	0.0200	0.0500

Car evaluation Data set	0.0048	0.4200	0.0300	0.2700	13.4000	0.0200	0.1900	0	0.4800	13.3400
Soybean Data set	0.0042	0.0500	0.0700	0.1800	0.4300	0	0.0500	0	0.2300	0.9200
Lenses Data set	0.0009	0	0	0.0300	0.0800	0	0.0300	0	0.0100	0.0300
Nursery Data set	0.0013	8.8600	0.2500	16.4300	127.8600	0.0300	23.0300	0	7.9900	240.4600
Tic-tac-toe Data set	0.0091	0.1900	0.0100	0.1300	18.1800	0.0100	0.6000	0	0.6100	54.6700
SPECT Heart Data set	0.0073	0.0600	0.0100	0.0700	3.8800	0	0.0400	0	0.2300	2.1500
MONK's Problem Data set	0.0058	0.0800	0.0100	0.0600	5.5100	0	0.1300	0	0.2100	1.7200

Table 14. *Mean rank values returned by the Friedman's rank test on the time values available in Table 13*

Method	**Mean Rank**
KStar	2.10
NaiveBayesClassifier	2.35
BFPM	**2.75**
J48	4.15
LogisticRegression	5.35
SMO	6.25
RBFNetworks	6.35
BFTree	6.90

MLP	9.30
LMT	9.50

Friedman's rank test results for the three measurements, AUC, RMSE and time elapsed conclude that the newly proposed method provides acceptable results against the nine other well established classification algorithms.

Conclusion

The studies of modus operandi help crime investigation by letting the police officers to solve crimes by linking suspects to crimes. Though there are many descriptive studies available under modus operandi analysis, a very little amount of work is available under computer science. Many of these methods have been derived using the methods based on link analysis. But, the accuracy of these methods is always compromised due to the cognitive biases of the criminals.

A novel Fuzzy based Binary Feature Profiling method (BFPM) to find associations between crimes and criminals, using modus operandi is introduced. The newly proposed method subjects not only the properties of the present, but also the properties of his/her previous convictions. The concept of dynamic modus operandi which is available in the proposed method considers all the modi operandi of his/her previous convictions to provide a fair rectification to the errors which result due to the human cognition. Dynamic MO uses frequent item set mining to result in a generalized binary feature vector. Complete MO profile also encapsulates past modus operandi of a particular criminal by aggregating the modus operandi of all of his/her previous convictions to one binary feature vector. This feature also guarantees a usage of criminal's past crime record with more generalizability. Completeness probability measures how much information is available in the new crime which is not available in the complete MO profile. Therefore, this measurement provides the capability of measuring how much extra amount of information is carried by the MO of the new crime. The deviation probability provides a notion about how much the new MO deviates from the most frequent crime flows which are available in the dynamic MO of a particular criminal. The vagueness and the impreciseness prompted the fact that it is not possible to use crisp logic to generate the similarity score. Therefore a fuzzy inference system was modeled to generate the similarity score.

Due to the under-represented and imbalanced properties of the actual data set, the new method has returned a lower performance when it was proposed to the data set without any rectification on the data set. However, with the introduction of over sampling, the method returned a very good performance, allowing one to arrive at the conclusion that the method could provide acceptable results for a balanced data set.The method generated favorable results in providing a good similarity measurement to suggest the connections between crimes and criminals. Fuzzy controller of the new approach guarantees to resemble the human reasoning process by confirming the usage of human operator knowledge to deal with nonlinearity of the actual situation. The newly proposed method was then adapted into a classification algorithm for the validation and comparison with other classification algorithms. The comparison of the new method with the well-established classification algorithms confirmed the generalizability of the new method.

The method only provides the capability to process the categorical data sets. If there are any continuous variables in the data set, the values must be introduced with categories before further processing. The method can be further extended to directly accept the continuous attributes. As the center of gravity method is used for the defuzzification process, further optimizations can be done by simplifying the defuzzification procedure. Adapting the fuzzy inference engine to a Sugeno [71] type and converting the defuzzification method to a more computationally efficient method such as the weighted average [72] method would provide a less complex computation. This would result in even less processing time when the sophistication of the data set rises.

References

[1] Holdaway, S., *Issues in Sociology: Crime and Deviance*.: Nelson Thornes Ltd, 1993.

[2] Chisum, W.J. , Turvey, B., "Evidence dynamics: Locard's exchange principle & crime reconstruction.," *Journal of Behavioral Profiling*, vol. 1, no. 1, pp. 1-15, 2000.

[3] Paternoster, R., Bachman, R., *Explaining Criminals and Crime: Essays in Contemporary Criminological Theory*, Ronet Bachman Raymond Paternoster, Ed.: Roxbury Publishing Company, 2001.

[4] Palmiotto, M.J., "Crime Pattern Analysis: An Investigative Tool," *Critical Issues in Criminal Investigation*, vol. 2, pp. 59-69, 1988.

[5] Douglas, J. E., Munn, C., "Modus operandi and the signature aspects of violent crime," in *Crime Classificatio Manual*, 2nd ed.: John Wiley & Sons, 2006, pp. 19-30.

[6] Chamikara, M.A.P., Galappaththi, A., Yapa, Y.P.R.D., Nawarathna, R.D., Kodituwakku, S.R., Gunathilake, J., Liyanage, L.H., "A Crime Data Analysis Framework with Geographical Information Support for Intelligence Led Policing," *Manuscript submitted to PeerJ, PeerJ PrePrints 3:e1909 https://doi.org/10.7287/peerj.preprints.1529v1*, July 2015.

[7] The 'Lectric Law Library. The 'Lectric Law Library. [Online]. http://www.lectlaw.com/files/int20.htm

[8] Bennell, C., Canter, D. V., "Linking commercial burglaries by modus operandi: Tests using regression and ROC analysis," *Science & Justice*, vol. 42, no. 3, pp. 153-164, 2002.

[9] Bennell, C., Jones, N. J., "Between a ROC and a hard place: A method for linking serial burglaries by modus operandi," *Journal of Investigative Psychology and Offender Profiling*, vol. 2, no. 1, pp. 23-41, 2005.

[10] Leclerc, B., Proulx, J., Beauregard, E., "Examining the modus operandi of sexual offenders against children and its practical implications," *Aggression and violent behavior*, vol. 14, no. 1, pp. 5-12, 2009.

[11] Oatley,G., Ewwart, B., "Data mining and crime analysis," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 2, pp. 147-153, 2011.

[12] King, R.D., Sutton, G.M., "High times for hate crimes: Explaining the temporal clustering of hate-motivated offending," *Criminology*, vol. 51, no. 4, pp. 871-894, 2013.

[13] Borg, A., Boldt, M., Lavesson, N., Melander, U., Boeva, V., "Detecting serial residential burglaries using clustering," *Expert Systems with Applications*, vol. 41, no. 11, pp. 5252-5266, 2014.

[14] (2011) Offender Profiling. [Online]. http://www.liv.ac.uk/psychology/ccir/op.html

[15] Canter, D., Hammond, L., Youngs, D., Juszczak, P., "The efficacy of ideographic models for geograhical offender profiling," *Journal of Quantitative Criminology*, vol. 29, no. 3, pp. 423-446, 2013.

[16] Agrawal, R., Imielinkski, J., Swami, A., "Mining Association rule between sets of items in large databases," in *Proceedings of the ACM SIGMOD International Conference of Management of Data*, New York, 1993, pp. 207-216.

[17] Yi,X., Rao, F.Y., Bertino, E.,Bouguettaya, A., "Privacy preserving association rule mining in cloud computing," in *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*, 2015, pp. 439-450.

[18] Koperski, K., Han, J., "Discovery of spatial association rules in geographic information databases," in *Proceeding of the 4th International Symposium on Spatial Databases*, 1995, pp. 47-67.

[19] Chen, H., *Intelligence and security informatics for international security*, 1st ed.: Springer US, 2006, vol. 10.

[20] Lin, S., & Brown, D. E., "An outlier-based data association method for linking criminal incidents," *Decision Support Systems*, vol. 41, no. 3, pp. 604-615, 2006.

[21] Berry, M. J., Linoff, G., *Data Mining Techniques: For Marketing, Sales, and Customer Support*, 3rd ed.: Wiley, 2011.

[22] Chen, H., Zeng, D., Atabakhsh, H., Wyzga, W., & Schroeder, J., "COPLINK ： Managing Law Enforcement Data and Knowledge," *Communications of the ACM*, vol. 46, no. 1, 2003.

[23] Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., Chau, M., "Crime Data Mining: A general framework and some examples.," *Computer*, vol. 37, no. 0018-9162, pp. 50-56, April 2014.

[24] Capozzoli, A., Lauro, F., Khan, I., "Fault detection analysis using data mining techniques for a cluster of smart office buildings," *Expert Systems with Applications*, vol. 42, no. 9, pp. 4324-4338, 2015.

[25] Chikersal, P., Poria, S. and Cambria, E., "SeNTU: sentiment analysis of tweets by combining a rule-based classifier with supervised learning," in *Proceedings of the International Workshop on Semanti Evaluation, SemEval*, 2015, pp. 647-651.

[26] Chen, H., "Machine learning for information retrieval: neural," *Journal of the American Society for Information Science*, vol. 46, no. 3, pp. 194-216, 1995.

[27] Adamo, J. M., *Data mining for association rules and sequential patterns, Sequential and Parallel Algorithms*, 1st ed. New York: Springer Science & Business Media, 2001.

[28] S.S., Cha, S.H., Tappert, C.C., Choi, "A survey of binary similarity and distance measures," *Journal of Systemics, Cybernetics and Informatics*, vol. 8, no. 1, pp. 43-48, 2010.

[29] Mamdani, E.H., Assilina, S., "An experiment in linguistic synthesis with a fuzzy logic controller," *International Journal of Man-Machine Studies*, vol. 7, no. 1, pp. 1-13, 1975.

[30] Godjevac, J., *Neuro-fuzzy Controllers: Design and Application*.: PPUR presses polytechniques, 1997.

[31] Sun, C.T., "Rule-base structure identification in an adaptive-network-based fuzzy inference system," in *Fuzzy Systems, IEEE Transactions on*, vol. 2, 1994, pp. 64-73.

[32] Abraham, A., Nath, B. and Mahanti, P.K., "Hybrid intelligent systems for stock market analysis," *Computational science-ICCS*, pp. 337-345, 2001.

[33] Zadeh, L.A., "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets and Systems*, vol. 90, pp. 117-117, 1997.

[34] MathWorks, Inc. (1994-2015) MathWorks. [Online]. http://in.mathworks.com/help/fuzzy/foundations-of-fuzzy-logic.html

[35] MathWorks. (1994-2015) MathWorks. [Online]. https://in.mathworks.com/

[36] MathWorks. (1994-2015) MathWorks Documentation. [Online]. http://in.mathworks.com/help/matlab/ref/edit.html

[37] MathWorks. (1994-2015) MathWorks Fuzzy Logic Toolbox. [Online]. http://in.mathworks.com/help/matlab/ref/edit.html

[38] Machine Learning Group at the University of Waikato. WEKA. [Online]. http://www.cs.waikato.ac.nz/ml/weka/

[39] Refaeilzadeh, P., Tang, L., Liu, H., "Cross-validation," in *Encyclopedia of database systems*, M. Tamer

Özsu Ling Liu, Ed.: Springer US, 2009, pp. 532-538.

[40] Hanley, J. A., McNeil, B. J., "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29-36, 1982.

[41] Estabrooks, A., Jo, T., Japkowicz, N., "A multiple resampling method for learning from imbalanced data sets," *Computational Intelligence*, vol. 20, no. 1, pp. 18-36, 2004.

[42] Wolpert, D.H., "The Lack of a Priori Distinctions Between Learning Algorithms," *Neural Computation*, vol. 8, pp. 1341-1390, 1996.

[43] Chang, Y. C. I., Lin, S. C., "Synergy of Logistic Regression and Support Vector Machine in Multiple-class Classification," in *Intelligent Data Engineering and Automated Learning - IDEAL 2004*, vol. 5, Exeter, UK, 2004, pp. 132-141.

[44] Machine Learning Group at the University of Waikato. Class J48. [Online]. http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html

[45] J.R. Quinlan, "C4.5 programs for machine learning," *Machine Learning*, vol. 16, no. 3, pp. 235-240, 1993.

[46] Bishop, C.M., *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press, 1995.

[47] Jayawardena, A. W., Fernando, D. A. K., Zhou, M.C., "Comparison of Multilayer Peceptron and Radial Basis Function Networks as Tools for Flood Forecasting," in *Destructive Water: Water- Caused Natural Disasters, their Abatement and Control*, California, 1996, pp. 173-181.

[48] Rish, I., "An empirical study of the naive Bayes classifier," , vol. 3, New York, 2001, pp. 41-46.

[49] Platt, J., "Fast Training of Support Vector Machines using Sequential Minimal Optimization," in *Advances in Kernel Methods - Support Vector Learning* , B., Burges, C., Smola, A., Schoelkopf, Ed., 1998.

[50] Keerthi, S.S., Shevade, S.K.,Bhattacharyya, C., Murthy, K.R.K., "Improvements to Platt's SMO Algorithm for SVM Classifier Design," *Neural Computation*, vol. 13, no. 3, pp. 637-649, 2001.

[51] Cleary, J.G.,Leonard, E.T., "K*: An Instance-based Learner Using an Entropic Distance Measure," in *12th International Conference on Machine Learning*, 1995, pp. 108-114.

[52] Friedman, J., Hastie, T., Tibshirani, R., "Additive logistic regression: A statistical view of boosting," *Annals of statistics*, vol. 28, no. 2, pp. 337-407, 2000.

[53] Landwehr, N., Hall, M., Frank, E., "Logistic Model Trees," *Machine Learning*, vol. 95, no. 1-2, pp. 161-205, 2005.

[54] Sumner, M., Frank, E., Hall, M., "Speeding up Logisti Model Tree Induction," in *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2005, pp. 675-683.

[55] Ilter,N., Guvenir, H.A. (1998) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/Dermatology

[56] Hume, T. (1994) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/Balance+Scale

[57] Pazzani, M. (1991) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/Balloons

[58] M. Pazzani, "The influence of prior knowledge on concept acquisition: Experimental and computational results," *Journal of Experimental Psychology: Learning, Memory & Cognition*, vol. 17, no. 3, pp. 416-432, 1991.

[59] Bohanec, M., Zupan, B. (1997) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

[60] Bohanec,M., Rajkovic, V., "Expert system for decision making," *Sistemica*, vol. 1, no. 1, pp. 145-157, 1990.

[61] Fisher, D. (1987, Jan.) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/Soybean+%28Small%29

[62] Michalski, R.S., "Learning by being told and learning from examples: an experimental comparison of the two methodes of knowledge acquisition in the context of developing an expert system for soybean desease diagnoiss," *International Journal of Policy Analysis and Information Systems*, vol. 4, no. 2, pp. 125-161, 1980.

[63] Julien, B. (1990, Aug.) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/Lenses

[64] Bohanec, M., Zupan, B. (1997, June) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/Nursery

[65] Bohanec,M., Rajkovic, V., "Expert system for decision making," *Sistemica*, vol. 1, no. 1, pp. 145-157, 1990.

[66] Aha, D. W. (1991, Aug.) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame

[67] Kurgan,L.A., Cios, K.J. (2001, Oct.) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/SPECT+Heart

[68] Thrun, S. (1992, Oct.) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/MONK's+Problems

[69] Hempstalk, K., Frank, E., "Discriminating Against New Classes: One-class versus Multi-class Classification," in *AI 2008: Advances in Artificial Intelligence: 21st Australasian Joint Conference on Artificial Intelligence*, Auckland, 2008, pp. 325-336.

[70] Howell, D.C., *Fundamental Statistics For The Behavioral Sciences Focuses*, 8th ed. Belmont: Wadsworth, Cengage Learning, 2013.

[71] Takagi, T., Sugeno, M., "Fuzzy identification of systems and its applications to modeling and control," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 1, pp. 116-132, 1985.

[72] Wu, D., Mendel, J. M., "Aggregation using the linguistic weighted average and interval type-2 fuzzy sets," *Fuzzy Systems, IEEE Transactions on*, vol. 15, no. 6, pp. 1145-1161, 2007.