International Journal of Computer Applications Technology and Research
Volume 9–Issue 05, 172-176, 2020, ISSN:-2319–8656

Automated Crime Patterns Analysis Framework for Predictive Policing using Data Mining Techniques

Duncan Nyale
Directorate of Computing and e-Learning
The Cooperative University of Kenya
Nairobi, Kenya

Michael M Kangethe
e-Kraal
Cyber Security Innovation Hub
Nairobi, Kenya

Abstract: The aim of this research is to study and develop an automatable technological framework that can be used to identify contributing attributes, patterns and trends from reported cases using data mining techniques. A combination of classification and association rules based data mining approach has been proposed for this study due to its effectiveness in bringing out patterns and trends that are interlinked, related and near each other**.**

Keywords: ANN: - Artificial Neural Network, KNN: - K Nearest Neighbour, MAP: - Maximum A Posteriori, GUI: - Graphical User Interface

1. INTRODUCTION

Through the digitization of reported cases by several law enforcement and public oversight agencies, the need for faster and reliable methods of sifting through massive data and cases to identify attributes and patterns that could lead to a future occurrence arises. Currently when policing and conducting security operations in response to reported crimes most agencies either use the previous reports and will have to sift through a lot of records to find patterns or conduct blanket policing which both are inefficient and laborious. This research details a general framework developed from the use of a combination of several data mining techniques to map occurrences to their dominant attributes and combinations.

In this age of vast data generation, it is imperative that we should find new and novel ways of effectively and quickly analyzing data and give appropriate feedback for decision making in any sphere. The increased computing power coupled with artificial intelligence and machine learning can be used for data mining, or knowledge discovery in databases to bring out previously unknown and potentially useful information from data. Intelligent data analysis is to extract useful knowledge, a process which demands a combination of several things including extraction, analysis, conversion, classification, organization and reasoning. This is precisely what this research has managed to do by creating an intelligent analysis framework that can be universally applied to any data by combining two data mining techniques to leverage on both to create a reliable model applicable to law enforcement through intelligence based policing.

2. BACKGROUND CONCEPT

Data mining is a relatively new data analysis technique that has the ability to discover patterns stored within historical data and is now considered a catalyst for enhancing business processes by avoiding failure patterns and exploiting success patterns. Several data mining techniques have been developed over the last decade. Generally, the data mining techniques can be categorized in four categories, depending on their functionality: classification, clustering, numeric prediction, and association rules. The main difference between the different techniques is in the way they extract information (algorithms and methods used) and how results (knowledge discovery/rules) are expressed. (Khaled Nassar, March 2007)

Instance Based Learning**:** This is the approximation of the target function from the training examples, as the approximation process is repeated with each and every query. Each time a new instance is encountered, its relationship to the previously stored examples is examined to assign a target function value for the new instance. There are several algorithms which include the Locally Weighted Regression, Case Based Reasoning, and the one to observe, the K- Nearest Neighbor and the Radial Basis Functions.

Classification**:** problems are essentially predictive models used to analyze an existing database to determine categorical divisions or patterns in the data. Classification problems are focused on identifying the characteristics indicating the group or class to which each record in the database belongs. On the other hand, when there is no pre-identified class or group, the clustering technique is used to group items that seem to fall naturally together. Several algorithms are inherently designed and suited for this purpose which include the KNN, ANN, Radial Basis Functions

The data mining technique used in this research will be a combination of instance based association and classification machine learning algorithms. In association learning, the goal is to discover any interesting patterns in the data by discovering association rules. Association rules differ from classification rules in two ways: they can predict any attribute (not just the group or class), and they can predict more than one attribute's value at a time. A typical association rule is represented in the following way:

Cause_1, Cause_2 => Result (or consequence)

That is, if Cause_1 and Cause_2 hold then Result (the association rule) applies, for n% of cases with x% confidence.

Each rule extracted is usually provided with a confidence level and a support. The confidence is the statistical value presenting the probability of a certain rule and the support is the number of cases/projects in which the rule is found. A pattern is defined as several identical or similar rules indicating a trend. Most of the data mining techniques use statistical tests when constructing rules or patterns and also for correcting models that depend too strongly on particular records in producing the rules and patterns *(Feldens 2002).* Since the goal when analyzing the dataset collected here was to detect any potentially useful patterns within the target industry based on reported incidences and registered complaints, association learning was the data mining

technique selected to analyze the dataset collected in this research.

(Bruno Agard, Catherine Morency and Martin Trépanier 2007) Conducted a research on data mining methords for the transport industry user behavior using Smart Card Data with desirable results. The limitations to their research based on the proposed research were that theirs focused on commuters behavioral patterns for economic and financed based planning of their transport system. Their observations during their research showed that the public transport users of this study can rapidly be devided into four major behavioural groups, whatever type of tickets used.

(Vikas.Grover et al 2009) Examined the current techniques that are used to predict crime and criminality. They were able to narrow down their research of possible techniques to three main categories:

- ***Statistical Methods***, these mainly relate to the journey to crime, age of offending and offending behavior.
- ***Techniques using Geographical Information Systems*** that identify crime hot spots, repeat victimization, crime attractors.
- ***Crime Generators;*** a miscellaneous group which includes machine learning techniques to identify patterns in criminal behavior and studies involving reoffending. Although their research provided a great insight on methords of crime patterns analysis. Their approach was not focused on any particular industry and thus followed no formal government policy.

3. METHODOLOGY

The design of this prototype to reflect the system framework will be done in layers and components which will be designed individually and developed independently to reflect the independent framework sub processes. Each subcomponent will take a collection of different discreet valued, real valued, inputs and produce real valued outputs that will at times be the input values of other or the same components based on the instance of computation.

The main framework will be based on a hybrid model derived partially from the combination of the two case based reasoning techniques which is the Bayesian Belief Networks from the Bayesian Based learning class of algorithims, the main concept of the Bayesian Based Learning which in principle is as below.

Features of Bayesian learning methods:

- Each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct, unlike algorithms which completely eliminate a hypothesis if it is inconsistent with any single example
- Prior knowledge can be combined with observed data to determine the final probability of a hypothesis. Prior probability is provided by asserting a prior probability for each candidate hypothesis and also a probability distribution over observed data for each possible hypothesis
- Bayesian learning can accommodate hypotheses which make probabilistic predictions e.g. this patient has a 93 % chance of recovery
- New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities. The main principle borrowed from this methord is the Maximum A Posteriori (MAP) Hypothesis and Maximum Likelihood in which the main goal of the methord is To find the most probable hypothesis ***h*** from a set of candidate hypotheses ***H*** given the observed data ***D***.

  MAP Hypothesis, $h_{MAP} = \underset{h \in H}{argmax}\, P(h|D)$

  $= \underset{h \in H}{argmax}\, P(D|h)P(h)/P(D)$

  $= \underset{h \in H}{argmax}\, P(D|h)P(h)$

  Where If every hypothesis in *H* is equally probable a priori, we only need to consider the *likelihood of the data D* given *h*, ***P(D|h).*** Then, h_{MAP} becomes the ***Maximum Likelihood***, $h_{ML} = \underset{h \in H}{argmax}\, P(D|h)$

  in essence the formular is explained as below:

  - To determine the most probable hypothesis, given the data *D* plus any initial knowledge about the prior probabilities of the various hypotheses in *H*.
  - ***Prior probability of h, P(h):*** it reflects any background knowledge we have about the chance that *h* is a correct hypothesis (before having observed the data – independent of D). If ***P(h)*** is not known, we can assign same probability to each hypothesis
  - ***Probability of D, P(D):*** it reflects the probability that training data *D* will be observed given no knowledge about which hypothesis *h* holds.
  - ***Conditional Probability of observing D, P(D|h):*** it denotes the probability of observing data *D* given some world in which hypothesis *h* holds.

And Concept Learning which is Inferring a boolean-valued function from training examples of its input and output.

Concept learning can be formulated as a problem of searching through a predifined search space of potential hypothesis for the hypothesis that best fits the training examples. The hypothesis might be represented in the form below where:

- Hypothesis h is a conjunction of constraints on attributes
- Each constraint can be:
  - A specific value : e.g. *Gender=Male*
  - Range Value e.g. *Age=19-25*
  - A don't care value : Marital Status=?
  - No value allowed (null hypothesis): e.g. *Height*

Example: hypothesis *h*

Time	Gender	Marital Status	Location	Age	Crime
1800 - 1900Hrs	M	Single	Nairobi	19-25	Armed Robbery
0.9	0.99	0.3	0.6	0.8	

The main approach to this research will be in two main stages*;*

i. Theoretical model framework design and

ii. System prototype development and model testing

3.1 Theoretical Model Framework Design

This will be done in three sub-stages as structured below.

Framework research and design

This is the first stage of the overall research as it will involve the research of existing systems. Identification of successful designs in related areas of application. Identification of the relevant frameworks approach development of discreet overall steps and stages of the proposed framework.

The outcome of this stage will be a high level description of the system framework that will be used to solve the said problem. It will be a diagrammatic representation of the whole system framework as a flowchart with generalized descriptions of the individual components that make up the final system.

Input data structure and format design

This is the second sub-stage as it will involve the identification of all the possible outcomes (which in this case will be cases or reported incidences). The identification of all attributes that affect the outcomes. And also the structure of the attributes and their significance in how and by what magnitude they affect the effective outcome. From there the identified attributes will have to be converted into discreet or semi-discreet value or variable based data inputs so that the proposed model will be able to mathematically compute the outcomes based on the associated attributes.

The outcome of this stage will be the creation of a discreet and continuous value based attributes input data structure that will eventually dictate the final systems database design.

This stage will overlap with the third sub-stage which is the Mathematical model design as the model will dictate the data and type of data needed as model inputs.

Mathematical model design

This will be the final stage of the theoretical system framework research process. This stage will involve the actual mathematical computation model formulation. It will involve "mapping" the attributes to their occurrences and identification of their associations to the overall result outcomes. Each attribute will be analyzed separately and together with all others to identify the magnitude effect on the overall outcome they have. Once the attributes, their relationships and effects to the final outcome have been discretely identified, they will be combined with respect to their magnitude and effect to the overall outcome to provide the actual mathematical formula that will be mapped on to the framework to provide the complete theoretical blueprint of the system that can be implemented by a variety of different programming technologies and approaches that observe the algorithmic process. This process will involve several mathematical processes. The processes will focus on two main things.

1. Attribute incident association calculation.

Each associated attribute impacts the overall outcome differently as some have a greater impact on the outcome than others. This will be achieved by computing the magnitude effect of the attribute individually and by combination with other attributes that result to the same outcome.

The general mathematical rule to be observed will be as below:

$$w_{c_a} = \sum_{1}^{n} a_{c_i}$$

Equation 1: Proposed Attribute incident equation

Where:

- w_{c_a} Is the overall weight of the attribute in terms of how it affects the outcome.
- n Is the number of possible combinations of the attributes.
- a_{c_i} Is the combination instance of the attribute combination.

This initial model is subject to alteration and improvement during the research process. The result of this model will be the mathematical determination of the impact of specific attributes or conditions to an outcome.

2. Attribute Effect Calculation

This model will also be used to identify the most dominant attribute that leads to an occurrence or incident. The reverse association rule will be as below:

$$a_{Dominant} = a_{i_{MAX}}$$

Equation 2: Attribute Effect Magnitude Equation

This means the most dominant attribute that contributes to an occurrence is that attribute that has the highest number of correlation values in the said instance of computation.

Where:

- $a_{Dominant}$ is the dominant attribute
- $a_{i_{MAX}}$ is the particular attribute instance *i* of all the available attributes

This first main stage will involve the most research effort and time of implementation as the whole concept and purpose of the research is based on this stage.

3.2 Model Testing and Evaluation

It will involve the actual development of a "proof of concept" system prototype that will not only be independently testable but will demonstrate the practicability of the whole research as an artificially intelligent computer system. This stage will involve several structured independent and dependent sub-stages that will be followed to come up with an actual working model of the system. The stages involved will be as follows.

Database and input and output data structure design

This will be the actual database design process of the overall system. The proposed database technology to be used will be the MySQL database Technology. Reasons for following this approach are due to the fact that MySQL has the following but not limited to, advantages.

- It is FREE: one does not need a license to use this technology as no costs are needed to implement this technology.

- It is SCALABLE: the database can grow and accommodate large volumes of data at an exponentially increasing rate.
- It is FLEXIBLE: the ease of implementation of this database to different data structure rules and methods is almost seamless as the technology is not rigid by nature.
- Easy to use: the amount of technical skill needed to implement this database technology is minimal as there is enough literature on nearly every imaginable scenario.

Prototype Interface design

This stage will involve the actual system GUI design. It be done in two stages

1. Input control design

This sub-stage will involve the actual design of the main system user interface and the required controls needed to for the easy interaction and manipulation by the user in computing the output. The main focus of this stage is to design the controls and selection rules needed to compute the final outcome based on the selection criteria. This is where the user will interact with the system in performing the analysis.

2. Output results design

This sub-stage will involve the designing of the results presentation interface. Factors to be considered in this stage are the ease of interpretation of the results in such a way a layman can deduce the results of the analysis. The output will be presented as a collection of charts and histograms

System internal function design and implementation

This is the third sub-stage and it is the most vital process of this research stage this will be the actualization of the theoretical framework as working executable code. It will involve defining and implementing the relevant functions that will automatically compute the results using the artificial intelligence (Machine Learning) algorithms.

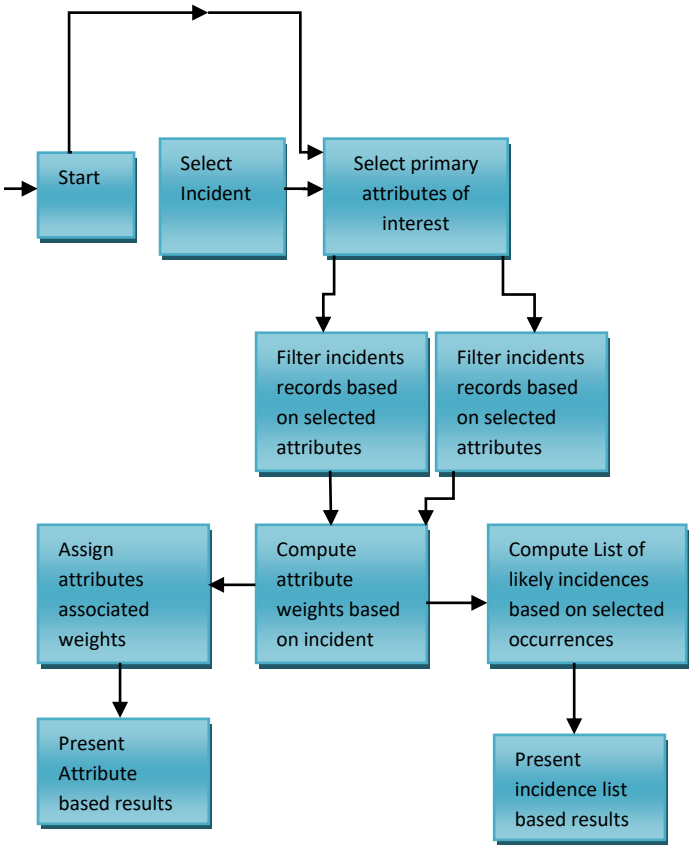

Figure 1 System Framework internal control flow

Data generation Prototype testing

This stage will involve:

1. The actual generation of the simulation data for the model accuracy testing purposes.
2. The system prototype testing and comparison to externally predefined cases and conditions.
3. Observation of the results and comparison to the test cases for accurate reflection of the models validity.

Results Documentation

This is the final stage of the research project. This will involve the detailed documentation of the projects simulation and testing observations results and conclusions. This will be the final conclusion of the project as it is and the summarization of the results and observations will be documented for external analysis and possible implementation of the proposed research product.

4. KEY OUTPUTS

If the model can be borrowed and implemented effectively on a data set, it should improve:

- Intelligence analysis of crime data sets
- Strategic policy formulation using results of analysis above
- Intelligence led policing of masses

5. KEY ACHIEVEMENTS

The research has resulted to some noted achievements which include:

- The development of a quantifiable mathematical Model that can be used to design and implement a predictive model
- An Automatable framework that can be applied to any data set to bring out certain patterns and trends
- Extending the body of knowledge by introducing a novel hybrid data mining concept

6. KEY CHALLENGES

This research like any other has not been without its own challenges, both technical and resource wise as they include.

- Unavailability of actual real world data to compute real world events.
- Limited Research time as this research has proved to be wide and has a lot of factors that still need both interpretation and analysis
- Framework testing was limited due to the inadequacy of testing data required; occasioned by the sensitivity of the sector in which the intelligence analysis framework is designed to function in.

7. ASSUMPTIONS AND LIMITATIONS OF SCOPE

The completion of this research and development of this framework will present some challenges which are stated below.

- The framework will work the same if applied to any data set and prototype
- The accuracy of the framework model cannot be guaranteed since it has not been tested with real data.
- When this framework is adopted and used all policies within the law will allow its usage

8. CONCLUSION AND RECOMMENDATIONS

The development of this framework has and will enable the immediate and automated analysis of data sets. This will provide a justifiable basis of the policies that will be made by stakeholders within the relevant industry as it uses the factors that are of interest to events and individuals. If implemented it would considerably increase the capacity to quickly, easily and reliably analyse data. This will reduce the needless effort to develop half-baked rules and policies while ensuring that every major contributing attributes and entity sets are catered for.

9. REFERENCES

[1] Bruno Agard, Catherine Morency, Martin Trépanier. 2007. Mining Public transport user behavior from smart card data.

[2] Vikas Grover, Richard Adderley, Max Bramer. 2009. Review of Current Crime Prediction Techniques.

[3] Nassar, Khaled 2007. Data-Mining of State Transportation Agencies Projects Databases. 12.

[4] Tom M. Mitchell 1997. Machine Learning.

[5] Van der Veer, H.T. Roos, A. van der Zanden 2009. Data mining for intelligence led policing.

[6] Dale Dzemydiene, Raimundas Vaitkevicius, Ignas Dzemyda 2010. Pattern recognition based on statistics and structural equation models in multi-dimensional data warehouses of social behavioral data pg 4 -10.

[7] Han J, Kamber 2006. Data mining: concepts and techniques.

[8] Lee BS, Snapp RR, Musick R, Critchlow T. 2002 Metadata models for ad hoc queries on terabyte-scale scientific simulations.