# Fast outlier detection using rough sets theory

F. Shaari, A. A. Bakar & A. R. Hamdan
*Centre for Artificial Intelligent Technologies (CAIT),*
*Faculty of Information Science and Technology,*
*The National University of Malaysia, Malaysia*

## Abstract

In many Knowledge Discovery applications, finding outliers is more interesting than finding inliers in a dataset. The perception of outliers is rare cases in dataset in which is being described as abnormal data in the information table. Outliers detections are applied in many important applications like fraud detection systems to uncover the suspicious objects which may have important knowledge hidden in the system. A new outlier detection technique based on Rough Sets Theory (RST) is hereby proposed. RSetOF is a new measure for the outlier factor based on RST. By employing this factor, a new formulation for detecting outlier is established. The outlyingness of outliers objects in a dataset using this measurement is identified. To detect outliers, two measurements which are the top *n* ratio and the coverage ratio are presented. Finding top *n* outliers from all objects allow searching of outliers from top ranked records based on the least outlier factor value. The capability in detecting outliers at top *n* number of outliers will indicate how fast the detection is. The efficiency of this technique by obtaining the coverage ratio value is then tested. The maximum percentage of coverage obtained shows the maximum number of outliers detected belonging to rare cases. A comparison is hence carried out to examine the performance of the RSetAlg with a selective outlier detection method, the Frequent Pattern method referred to as FindFPOF. Ten benchmark datasets for assessing the outlier detection technique are used for this purpose. The experimental result shows that the proposed technique is competitive and proven to be better in speed of detection than the other technique. The fast and efficient detection of outliers has proven its potential as a new outliers detection technique based on RST.

*Keywords: outlier detection, rare, deviate, exception, deviation, anomaly, infrequent, small, imbalance.*

WIT Transactions on Information and Communication Technologies, Vol 40, © 2008 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA080031

# 1 Introduction

An outlier as defined by Hawkins [1] is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Outlier mining focuses on the rare data whose behaviour is very exceptional when compared with the rest of the large amount of data. This exception identification can lead to the discovery of unexpected knowledge.

Outlier mining has been realized from several approaches or technologies in the field of statistics, machine learning, artificial intelligence, visualization and database management. Finding these outliers in large datasets has drawn increasing attention among researchers [2–10].

Although many techniques have been proven useful and effective in detecting outlier pattern, the following problems which occurred remain for further explorations among the data mining researchers. As data change its size and dimension, it is found that most algorithms developed faced the problems of handling the in-scalability of the dataset. The curse of dimensionality had caused the using of distances of points inappropriate to discover outliers in high dimensional space [3]. The concept of locality [4, 5] becomes difficult as data become sparse in high dimensional datasets.

The projection in lower density [6] fails to detect outliers in different projections. The clustered-based is found as a method which detect outliers as by product which does not able to interpret the abnormality of the outliers detected [7]. Although the problems of inefficiencies can be improved by hybriding [2] two techniques or more, yet this method is still in study and further research are indeed needed until today. The Frequent Pattern method [8] utilized the frequent patterns in different subspaces, in defining outliers in high dimensional space however the detection process is time consuming and computationally expensive. In Local Search Algorithm (LSA) [12], the detection of outliers proves able to be detected from feasible solution based on Optimization approach, however the process is reported by He et al. [11] as time consuming on very large datasets. In comparison, the Greedy Algorithm [11], is found faster in order of magnitude than the previous LSA method.

In this paper, it is assumed that objects which cluster in small group and away from the common objects are outliers. Objects which cluster in majority are grouped as a large class (common occurring cases). On the other hand, objects cluster in minority is grouped as a rare class (rare cases) where outliers are recited. A new method is proposed for detecting outlier by discovering the concept of *Non-Reduct* from the Rough Set Theory (RST) approach. In computing *Non-Reduct*, a new concept is hereby defined in calculating the Indiscernibility Matrix Modulo (iDMM) and Indiscernibility Function Modulo (iDFM). The foundation of these concepts can be found in [13]. The elements obtained from the proposed matrix iDMM are a different set of objects (*Non-Reduct*) where these objects contain non-interesting or redundant set of attributes in the IS. The different mechanism here can be related to the different behavior of rare objects or so called outliers. Thus, this set of *Non-Reduct* is expected to be able to uncover the outlier knowledge from the information system (IS). The

capability and outstanding knowledge found from the computation of *Non-Reduct* extend the approach and concept of RST.

The experimental results on the datasets show that: (1) The proposed method is a competitive method compared to the FindFPOF outlier detection algorithms on identifying outliers, (2) The algorithm has a good detection rate of outliers for large datasets and has the ability in predicting the presence of rare case where outliers are recited.

This paper is organized as follows. Section 2 discusses the concept of Rough Set Theory (RST) in brief. Section 3 introduces the detection of outlier using RST by defining the definition of Indiscernibility Matrix Modulo, Indiscernibility Function Modulo, *Non-Reduct* and the Rough Set Outlier Factor Value(RSetOF). The experimental design and experimental results explain detailed descriptions on ten datasets and the results are presented. The effectiveness of performance is measured by comparing both methods and the results are recorded and discussed in section 4. The conclusion of this paper is discussed in section 5.

## 2 Rough Sets Theory (RST)

RST is a mathematical tool introduced by Pawlak in 1980's [14]. It concerns with the analysis and modelling of classification and decision problems involving with vagueness, imprecise and uncertain or incomplete information. RST invokes the concept of approximation reasoning, hence has fundamental importance to Artificial Intelligent and Cognitive Sciences especially in areas of machine learning, knowledge acquisition, decision analysis, knowledge discovery from database, expert system, decision support systems, inductive reasoning and pattern recognition. Many important applications found to be developed effectively applying RST are like medicine, pharmacology, business analysis, banking, meteorology and security systems. Several important notations of RST which involve discern objects, indiscernibility relation, equivalence class, discernibility matrix and function, discernibility matrix modulo(DMM) and function modulo(DFM) and computation of Reduct can be found in [13–15].

## 3 Outlier detection based on rough set theory

*Reduct* is determined from the set of prime implicants of the discernibility function. *Reduct* does not contain redundant attributes. The set of attributes is usually interesting attributes and is used in attribute selection process. The computation of *Reduct*, can be used to represent Information System(IS) or Decision System(DS).

In this work, the set of attributes in which is referred as superfluous and redundant is of prime interest. Although this set of attributes is usually considered as not interesting, it is presumed that these attributes are important and able to detect outliers in datasets.

In the following sub-sections, the new concept of *Non-Reduct* is introduced by presenting new definitions and computation of *Non-Reduct*. An example of a

DS will be used to demonstrate the process and the results obtained are recorded. Here, a new outlier detection technique is proposed based on the concept of *Non-Reduct* in RST.

### 3.1 The concept of *Non-Reduct*

The concept of *Non-Reduct* introduced in this section is originated from the concept of *Reduct*. In RST, a DS is similar to an IS, but a distinction is made between the condition and the decision attributes. In an IS, the information is not interpreted but in a DS, each object of the domain is assigned with a value of an expert classification attribute. A simple DS with distribution of equivalence classes is as shown in Table 1 [17].

Table 1: Example of an Equivalence of a DS.

| Class | a | b | c | Dec | Num of Objects |
|---|---|---|---|---|---|
| E1 | 1 | 2 | 3 | 1 | 50 |
| E2 | 1 | 2 | 1 | 2 | 5 |
| E3 | 2 | 2 | 3 | 2 | 30 |
| E4 | 2 | 3 | 3 | 2 | 10 |
| E5,1 | 3 | 5 | 1 | 3 | 4 |
| E5,2 | 3 | 5 | 1 | 4 | 1 |

Table 2: Indiscernibility Matrix Modulo (iDMM) Class from Decision System, $\mathcal{A}$.

| | **E1** | **E2** | **E3** | **E 4** | **E5** | $f'$ |
|---|---|---|---|---|---|---|
| E1 | x | x | x | x | x | - |
| E2 | x | x | b | x | x | { b } |
| E3 | x | b | x | a, c | x | {a,b} {a,c} |
| E4 | x | x | a,c | x | x | {a,c} |
| E5 | x | x | x | x | x | - |

As mentioned in section 2, the discernibility function $f$ which determined *Reduct* is computed from the process of DMM and DFM. Similarly, *Non-Reduct* can be computed using a new formulation of iDMM and iDFM. The following subsection 3.1.1 explains the new creation of the Indiscernibility matrix modulo decision (iDMM), and the Indiscernibility function matrix modulo(iDFM), while the subsection 3.1.2 describes on *Non-Reduct*.

### 3.2 Indiscernibility Matrix Modulo Decision (iDMM) and Indiscernibility function $f'$ (iDFM)

The concept of iDMM is to find a set of attributes from every pair of equivalence classes which are indiscern in attribute values from the matrix and represent these attributes in the form of Conjunctive Normal Form(CNF). In obtaining the set of attributes as mentioned above, iDMM is hereby defined as below :

Definition 1 (*Indiscernibility Matrix Modulo D (iDMM)*)**.** Given a decision system(DS), $A = (U,(C,D))$ where $U$ is a nonempty finite set called the universe, while $C$ and $D$ are nonempty finite set of attributes. These attributes are separated into disjoint sets of condition attributes $C$ and decision attributes $D$ (where $C \cap D = \phi$). An indiscernibility relation by attribute *C, IND(C)* allow objects to be classified into set of equivalence classes, where $n = |U/IND(C)|$.

Therefore, the *Indiscernibility Matrix Modulo D* of *A*, $M'_{[C,D]}$ is defined as in equation(1) below:

$$M'_{[C,D]} = \{m'_{[C,D]}(E_i, E_j) : 1 \le i \le n, 1 \le j \le n\},$$
$$where, m'_{[C,D]}(E_i, E_j) = \begin{cases} m'_{[C,D]}(E_i, E_j) = \{a \in C : a(E_i) = a(E_j)\}, if\ \delta_C(E_i) = \delta_C(E_j) \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad , otherwise \end{cases} \quad (1)$$

Where the entry $m'_{[C,D]}(i, j)$ in the iDMM is the set of attributes from *C* that indiscerns between object classes $E_i, E_j \in U/IND(C)$ and if the decision attributes $\delta$ are also indiscerns between the classes where $\delta_c(E_i) = \delta_c(E_j)$.

Table 2 above illustrates the iDMM from a decision system *A*. The simplification of the disjunction and conjunction of the matrix gives the Indiscernibility function modulo D(iDFM), $f'$ as shown in the rightmost column in the table. The function $f'$ can be expressed as in equation (2) below.

$$f'_{[C,D]} = \bigwedge_{i,j \in \{1,\ldots,n\}} \vee\, \tilde{m}_{[C,D]}(E_i, E_j), \text{ where } n = |U / IND(C)| \quad (2)$$

The generated function $f'$ indicates the computation of *Non-Reduct*. In the following section the notion of *Non-Reduct* is described.

### 3.3 Definition of *Non-Reduct*

*Reduct* is used, which is defined as follows: Given *A = (U,A), let B* $\subseteq$ *A*, a *Reduct* of *B* is a set of attributes *B'*$\subseteq$ *B* such that all attributes *a* $\in$ *B-B'* are dispensable, and *IND(B') = IND(B)*. The set of *Reduct* of *B* is denoted *Red(B)* [13,15,16]. The very intuitive definition of *Non-Reduct* can be reflected as follows:

Definition 2 (*Non-Reduct*)**.** Given A = *(U,A), let B* $\subseteq$ *A*, a *Non-Reduct* of *B* is a set of attributes *B** $\subseteq$ *B,* such that all attributes *a* $\in$ *B-B** are indispensable, and *IND(B*) = IND(B)*. The set of *Non-Reduct* of *B* is denoted *Non-Reduct(B*).

Table 3 depicts the computation of *Reducts* based on a DS. The first column in the table lists the five equivalence classes from E1 to E5, each of which contains a number of objects from universe that are indiscernible by attributes *a* through *c*. A set of *Reducts* is as shown in the rightmost column in the table. Each of the *Reduct* is the prime implicants (*f*) of the CNF as shown in the third column in the table.

Correspondingly, *Non-Reducts* can be translated into computing the prime implicants of a boolean function as shown in the rightmost column depicted from table 4.

Table 3: *Reducts* of a Decision System (DS).

| Equiv. Class | CNF of Boolean Function | Prime Implicants (*f*) | *Reducts* |
|---|---|---|---|
| E1 | $c \wedge a \wedge (a \vee b) \wedge (a \vee b \vee c)$ | $c \wedge a$ | {a,c} |
| E2 | $c \wedge ( a \vee b )$ | $c ( a \vee b)$ | {a,c} {b,c} |
| E3 | $a \wedge (a \vee b \vee c )$ | a | {a} |
| E4 | $(a \vee b) \wedge ( a \vee b \vee c)$ | $a \vee b$ | {a} {b} |
| E5 | $(a \vee b \vee c) \wedge (a \vee b) \wedge (a \vee b \vee c) \wedge (a \vee b \vee c)$ | $a \vee b$ | {a} {b} |

Table 4: *Non-Reducts* in a Decision System (DS).

| Equiv. Class | CNF of Boolean function | Prime Implicants $(f')$ | *Non-Reducts* |
|---|---|---|---|
| E1 | - | - | - |
| E2 | b | b | {b} |
| E3 | $b \wedge ( a \vee c )$ | $b ( a \vee c )$ | {a,b} {b ,c} |
| E4 | $a \vee c$ | $a \vee c$ | {a} {c} |
| E5 | - | - | - |

It is obvious here that *Reducts* and *Non-Reducts* depicted in Table 3 and Table 4 are distinct set of attributes obtained by representing the DS. As previously defined and explained, *Reduct* is a set of interesting attributes that is capable of representing the knowledge in a DS, therefore *Non-Reduct* can be defined as a non-interesting set of attributes which is presumed to contain undiscovered important knowledge. The set of *Non-Reducts* is potentially contributed to the concept of outliers as explained in the previous section.

## 4 Experimental design and results

The experimental design is constructed based on the explanation on the process of the *Non-Reduct* computation and description of the datasets chosen for the experiment. In the following subsection 4.1, all datasets are stated. The datasets are then prepared for testing in the proposed method as discussed in subsection 4.2. In subsection 4.3, the experimental results are presented.

### 4.1 Data description and preparation

Ten datasets from Machine Learning Repository[18] are chosen. The ten datasets are Lymphography(LYM), Breast Cancer(BRE), Cleaveland(CLV), Heartdisease(HDE) and Echoli(ECO) Iris Plant (IRP), Zoo(ZOO), Glass(GLS), COIL2000(COL) and Australian Credit Card(ACC).

First, all the discretized datasets are formed into equivalence classes based on RST. In the following step, these datasets are being processed in creating imbalance class distributions in them. This is done by removing certain amount of data from certain classes from the DS. The creation of imbalance class

distributions produced two cases which are common and rare cases describing the characteristic of each dataset. This step reflects the process of preparing dataset for outlier detection by William et al. [10] & He et al. [8]. The percentage for rare cases in each dataset which ranged from 1% to 10% is based on the definition given from the literature [6,10] whereas, the percentage of common classes are within the range of 90% to 99%. According to Lazarevic et al. [19] the frequency of rare cases are defined with smaller percentages from 0.1% to less than 10%. Generally, outliers of this range are those datasets from intrusion domain application [9,20,21].

### 4.2 Design and evaluation method

The process of computing the *Non-Reduct* and the detection of outliers as explained in the above section 3 and subsection 4.1 can be illustrated by a model referred to as Outlier Detection Modeling Process as in Figure 1.

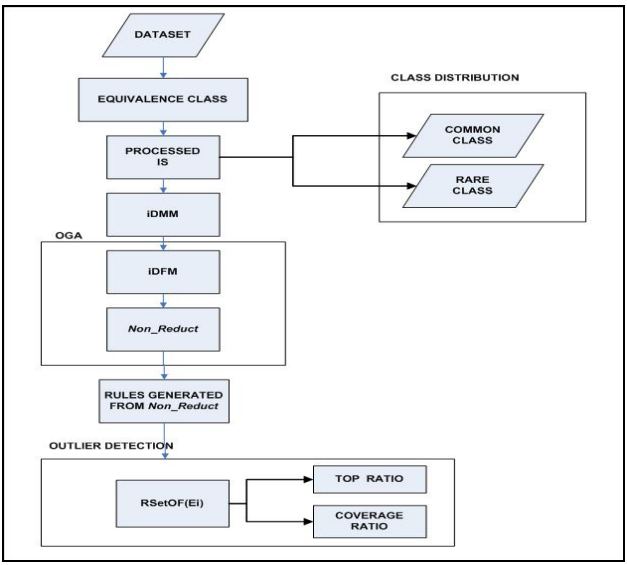


Figure 1: Outlier detection modeling process.

The model can be a base model in outlier mining using Rough Set Theory. In implementing the model based on Rough Sets Theory, a new method referred as RSetAlg is proposed. The RSetAlg method is tested upon ten datasets and the performance is evaluated. The evaluation of performance is conducted by comparing the RSetAlg method with a method referred as FindFPOF chosen from the literature review [8]. As mentioned in subsection 3.3, three measurements used in the proposed method in identifying and detecting outliers are RSetOF value, *top ratio* and *coverage ratio*(CR). The detection of outliers are recorded from search based on *top-n* outliers to the highest percentage of all records in a dataset. Outliers detected belonging to rare classes are measured by

*coverage ratio* during the search. In this work, the notion detection rate is the preference used to explain the detection of outliers from the two measurements *top ratio* and *coverage ratio*.

During the comparison, the measurement *top ratio* used a number of records instead of equivalence class. This will allow equivalence comparison among the two methods. The following subsection 4.3 explains the experimental results.

### 4.3 Experimental results

In exploring the reliability and robustness of the newly developed RSetAlg method, several experiments were conducted. Table 5 depicts the experimental results tested upon ten datasets for both RSetAlg and FindFPOF methods. The results performances are analysed by comparing the detection rates and F-measure of both methods. The first column showed a list of ten datasets while the second, third , fourth and fifth columns described the detection rate and F-measure of both methods respectively. Each dataset in the first column, represented 100% CR where all outliers detected belong to rare cases. The results showed that RSetAlg has better detection rate compared to FindFPOF when tested upon BRE, LYM, CLV, ZOO, HDE, ACC and COL datasets. However, the result tested upon IRP, GLS and ECO datasets showed that FindFPOF method has better detection rate than RSetAlg in comparison.

Table 5: Comparison of Detection Rate and F-measure between two outlier detection methods (RSetAlg and FindFPOF) tested upon ten datasets.

| Outlier at 100% CR Dataset | **Detection Rate** | | **F- Measure** | |
|---|---|---|---|---|
| | RSetAlg | FindFPOF | RSetAlg | FindFPOF |
| | | | | |
| BRE | 8.07% | 14.0% | 1 | 0.84786 |
| IRP | 12.62% | 11.65% | 1 | 0.666666 |
| LYM | 4.05% | 8.11% | 1 | 0.16651 |
| CLV | 4.30% | 53.64% | 1 | 0.22222 |
| ZOO | 23.76% | 35.64% | 1 | 0.608696 |
| HDE | 6.25% | 11.88% | 0.84210 | 0.8 |
| ACC | 16.82% | 18.35% | 0.97435 | 0.5 |
| GLS | 66.20% | 56.34% | 0.875 | 0.66666 |
| ECO | 66.57% | 4.48% | 0.07862 | 0.1395 |
| COL | 5.03% | 10.06% | 0.11787 | 0.706410 |

The second metric of the performance which is F-measure as in [22,23] are adopted as to give prominent results for rare cases problem. In this work, True Positive rates are defined as number of rare classes correctly classified among all positive samples during the test and False Positive rates are common classes incorrectly identified as rare classes in the test among all negative samples. The results show that the performance of RSetAlg is encouraging for eight(8) datasets BRE, IRP, LYM, CLV, ZOO, HDE, ACC AND GLS where F-measure indicate the values between 0.8 to one(1). It is learned that the range values from 0.6 to 1 for F-measure suggests promising performance in predicting the

presence of rare cases thus showing RSetAlg as a distinctive method in mining rare cases based on detection of outliers. In turn, the results performance tested upon COL and ECO show poor F-measure.

## 5 Conclusion

In this paper a new outlier detection method is proposed based on the computation of *Non-Reducts* in Rough sets Theory. Ten datasets were tested using the proposed RSetAlg method and the effectiveness is evaluated by comparing with FindFPOF method. Two metrics of performance used in the comparison are the Detection Rate and F-measure. It is observed that RSetAlg method achieves better detection rate than FindFPOF in detecting outliers. The proposed method also show better prediction of outliers in rare cases based on F-measure values compared to FindFPOF. In conclusion RSetAlg is an effective and fast outlier detection method compared to FindFPOF method.

## References

[1] Hawkins, D. M., Identifications of Outliers, Monograph on Applied Probability and Statistic, Reading, London Chapman and Hall, 1980.
[2] Hodge, V. J. & Austin, J., A Survey of Outlier Detection Methodologies. Artificial Intelligence Review (22: 2004) 85–126, 2004.
[3] Knorr, E. M. & Ng, R. T., Algorithms for Mining Distanced-Based Outliers in Large Datasets. 24th VLDB Conference, New York, pp. 392–403, 1998.
[4] Breunig, M. M., Kriegel, H.-P., Ng, R. T. & Sander, J. OPTICS OF: Identifying Local Outliers, SIGMOD US ,93-104, 2000.
[5] Chiu, A. L.-M. & Fu, A. W.-C, Enhancements on Local Outlier Detection. Seventh International Database Engineering and Applications Symposium (IDEAS'03), 2003.
[6] Aggarwal, C. C. & Yu, P. S., Outlier Detection for High Dimensional Data. SIGMOD'01, Santa Barbara, pp. 37–46, 2001.
[7] He, Z., Xu, X. & Deng, S., Discovering Cluster Based Local Outliers. Pattern Recognition Letters., 2003.
[8] He, Z., Huang, J. Z., Xu, X. & Shengchun, D. A Frequent Pattern Discovery Method for Outlier Detection. (Vol. 3129/2004): Springer Berlin/ Heidelberg 2004.
[9] Hawkins, S., He, H., Williams, G. & Baxter, R. 2002. Outlier Detection using Replicator Neural Networks. DaWak 2002.
[10] Williams, G., Baxter, R., He, H., Hawkins, S. & Gu, L. A Comparative Study of RNN for Outlier Detection in DataMining, 2nd IEEE Inf. ICDM'02, Japan pp. 709–712, 2002.
[11] He,Z., Deng, S., Xu, X. & Huang, J., A Fast Greedy Algorithm for Outlier Mining, PAKDD pp. 567–576, 2006.
[12] He, Z., Deng, S. & Xu, X. An Optimization Model for Outlier Detection in Categorical Data. International Conference on Intelligent Computing, pp. 400–495, 2005.

[13] Mollestad, T., A Rough Set Approach to Data Mining: Extracting a Logic of Default Rules from Data, The Norwegian University of Science and Technology, 1997.
[14] Pawlak, Z., Rough Sets. Int. Journal of Computer and Information Sciences 11:5, 341–356 1982.
[15] Mollestad, T. & Komorowski, J., A Rough Set Framework of Prepositional for Mining Default Rules Springer-Verlag Singapore, 1998.
[16] Pawlak, Z., Grzymala-Busse, J., Slowinski, R. & Ziarko, W., Rough Sets, Communication of the ACM [Electronic Version], 38, 1995.
[17] Bakar, A. A., Propositional Satisfiability Method in Rough Classification Modelling for Data Mining. PhD Thesis, University of Putra, Malaysia, Kuala Lumpur, 2002.
[18] Murphy, M. P., UCI Machine Learning Repository (online). Retrieved, from http://www.ics.uci.edu/~mlearn/MLRepository.html (1 Mac 2005), 1995.
[19] Lazarevic, A., Srivastava, J. & Kumar, V., Data Mining for Analysis of Rare Events: A Case Study in Security, Financial and Medical Applications, PAKDD, 2004.
[20] Yamanishi, K., Takuechi, J.-I. & Williams, G., On-line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms, KDD 2000, Boston, MA USA, 2000.
[21] Yamanishi, K. & Takeuchi, J. I., Discovering Outlier Filtering Rules from Unlabeled Data. KDD 2001, San Francisco, CA USA, 2001.
[22] Otey, M. E., Parthasarathy, S. & Ghoting, A., An empirical Comparison of Outlier Detection Algorithms, ACM SIGKDD Workshop on Data Mining Methods for Anomaly Detection, pp: 45–52, 2005.
[23] Lin, S. & Brown, D. E., An Outlier-based data association method for linking criminal incidents, Decision Support System, p. 12, 2004. www.sciencedirect.com.