Technical Report

An Outlier-based Data Association Method For Linking Criminal Incidents

Song Lin & Donald E. Brown

Department of Systems Engineering
University of Virginia

SIE 020010

An Outlier-based Data Association Method For Linking Criminal Incidents

Song Lin Donald E. Brown
sl7h@virginia.edu brown@virginia.edu
Department of Systems and Information Engineering
University of Virginia, Charlottesville, VA 22903, USA

Abstract

Data association is an important data-mining task and it has various applications. In crime analysis, data association means to link criminal incidents committed by the same person. It helps to discover crime patterns and catch the criminal. In this paper, we present an outlier-based data association method. An outlier score function is defined to measure the extremeness of an observation, and the data association method is developed based upon the outlier score function. We apply this method to the robbery data from Richmond, Virginia, and compare the result with a similarity-based association method. Result shows that the outlier-based data association method is promising.

Keywords

Outlier, data association, crime analysis

1 Introduction

Associating records in the database according to a certain mechanism is a major data-mining task. Different applications, including target tracking and document retrieval can be treated as data association problems.

An important activity in law enforcement is to link criminal incidents committed by the same person or a group of people (gang). It is also referred to as tactical analysis in criminology. This analysis will help to understand the behavior of the criminals and the patterns of crimes, make predictions for future crimes, and even arrest the criminals. A number of systems have been developed to tackle this association problem. Integrated Criminal Apprehension Program (ICAP) [10] was introduced by R.O.Heck, and it allowed police officers to perform a matching between suspects and arrested criminals using the Modus Operandi (MO) features; Armed Robbery Eidetic Suspect Typing (AREST) [3] employed an expert system approach; a similarity-based data association method was proposed by Brown and Hagen [6]. In practice, analysts normally build SQL (Structured Query Language) strings to retrieve all records in the database matching the searching criterion.

One common feature in conventional crime association methods is that they concentrate on "association". When all the information about the crime incidents and suspects is perfectly observed and recorded, above methods perform well. For example, if the fingerprint of the offender is known, only an exact matching is needed. However, that seldom occurs in practice. The information available usually is insufficient to separate two different criminals. We may have a number of crimes with the same suspect description "white male", but it is not credible to say that these "white males" are the same person. Therefore, the association method should consider

not only linking records together, but also being able to distinguish a small group of observations from other incidents. In other words, this method should be capable to perform both "association" and "separation".

In this paper, we present an outlier-based association method. Outliers are observations highly different from other records. When a group of records have some common characteristics, and these common characteristics are "outliers", we associate these records. Hence our method takes both "association" and "separation" into account.

The rest parts of the paper are organized as follows: in the next section, we briefly review traditional and recent outlier detection approaches; the outlier-based association method is given in section 3; in section 4, we apply this outlier-based association method to the robbery dataset of Richmond city, 1998, and we compare this method with a similarity-based association method; and section 5 concludes the paper.

2 Related work

An outlier is "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" [9]. Finding outliers is important in a variety of applications, including credit fraud detection and network intrusion detection. Most studies on outlier detection lie in the field of statistics, and a number of discordancy tests have been developed. [4][9] In practice, a 3σ rule is widely used. The mean μ and the standard deviation σ are calculated, and if one observation lies outside the $(\mu-3\sigma, \mu+3\sigma)$ range, it is treated as an outlier. These methods are developed to detect a single outlier. When multiple outliers exist, the median and the MAD scale are suggested instead of the mean and the standard deviation. [2]

These studies focus on finding outliers in a univariate dataset, and generally an assumption is made that the data points follow some standard distribution. (Mostly a Gaussian distribution is assumed) However, in real-world problem, datasets are usually multivariate and it is hard to make assumptions on the underlying distributions.

Several approaches that detect outliers in multivariate data without the a-priori assumption of the distribution have been proposed by some researchers recently. Knorr and Ng introduced the notion of distance-based outliers [11][12]. They call an object *DB(p,D)* outlier if at least a fraction of *p* of the objects in the dataset having a distance greater than distance *D*. They also prove that their notion of outlier unified outlier definitions in some standard statistical distributions [11]. Then they give several algorithms for detecting distance-based outliers [12]. Ramaswamy et al. [15] point out that the *DB(p,D)* outlier is sensitive to the parameter *p* and D, and they are difficult to be set. They presented a k-nearest neighbor outlier. The distance from each data points to its *k*-th nearest neighbor is calculated. Then all data points are ranked according to the distance, and the top *n* data points are selected as outliers. Breunig et al. proposed the concept of "local" outlier [5]. They argued that outliers should be considered locally not globally. They define the concept of outlier factor, and for each object in the dataset they assign a "score" to measure the extremeness instead of giving binary results ("yes" or "no"). Aggarwal et al. claimed that for high-dimension dataset, the notion of outliers exists in sub-space projections. They use an evolutionary algorithm to find these outliers in sub-spaces [1].

These works concentrate on detecting individual outliers, and they are generally for continuous numerical variables. Association among outliers has not been studied. In this paper, we present an outlier-based association method. This method is also capable of detecting outliers. The method is

developed for categorical variables, since we hope to apply this method to crime association problem and many variables, especially MO features, are categorical in crime analysis.

3 Outlier-based association method

3.1 Basic Idea

The basic idea of this method comes from a "Japanese sword" claim, first introduced by Brown and Hagen [6]. Consider the weapon used in a series of robberies. When the weapon is a "gun", we can hardly asserting that these incidents result from the same person because "gun" is very common and everybody uses "guns". However, if we have two incidents with an unusual weapon like a "Japanese sword", we are more confident to link these two incidents.

We generalize the Japanese sword claim as follows: When a group of records have some common characteristics and these characteristics are outliers, these records are more likely to result from the same cause (criminal in crime analysis).

3.2 Definitions

In this section we formally define the concepts and notations used in the reminder of this paper.

$A_1, A_2, ..., A_m$ are m attributes that we consider relevant to our study, and $D_1, D_2, ..., D_m$ are their domains respectively. Currently, these attributes are confined to be categorical. Let $z^{(i)}$ be the i-th incident, and $z^{(i)}.A_j$ is the value on the j-th attribute of incident i. $z^{(i)}$ can be represented as $z^{(i)} = (z_1^{(i)}, z_2^{(i)}, ..., z_m^{(i)})$, where $z_k^{(i)} = z^{(i)}.A_k \in D_k$, $k \in \{1,...,m\}$. $\boldsymbol{Z}$ is the set of all incidents.

Definition 1. Cell

Cell c is a vector of the values of attributes with *dimension t*, where $t \le m$. So a cell is a subset of the Cartesian product of $D_1 \times D_2 \times \cdots \times D_m$. A cell can be represented as $c = (c_{i_1}, c_{i_2}, ..., c_{i_t})$, where $i_1, ..., i_t \in \{1,...,m\}$, and $c_{i_s} \in D_{i_s}$, $s = 1,...,t$. The concept of cell originates from the area of On-Line Analytical Processing (OLAP) [7]. For example, if the attributes are quantity, time, product and geography, then (Sales, January 1994, Candy Bars and the United States) will be a cell, which describes the value of candy bar sales in the United States for the month of January 1994 [14]. In order to standardize the definition of a cell, for each D_i, we add a "wildcard" element "*". Now we allow $D'_i = D_i \cup \{*\}$. For cell $c = (c_{i_1}, c_{i_2}, ..., c_{i_t})$, we can represent it as $c = (c_1, c_2, ..., c_m)$, where $c_j \in D'_j$, and $c_j = *$ if and only if $j \notin \{i_1, i_2, ..., i_t\}$. $c_j = *$ means that we do not care about the value on the *j*-th attribute. $\boldsymbol{C}$ denotes the set of all cells. Since each incident can also be treated as a cell, we define a function *Cell*: $\boldsymbol{Z} \rightarrow \boldsymbol{C}$. If $z = (z_1, z_2, ..., z_m)$, $Cell(z) = (z_1, z_2, ..., z_m)$.

Definition 2. Dimension of a cell

We call a cell *c* a *t*-dimensional cell or a cell of dimension *t* if cell *c* take non-* values on *t* attributes.

Definition 3. Contain

We say that cell $c=(c_{i_1},c_{i_2},...,c_{i_t})$ *contains* incident z if and only if $z.A_j=c_j$, $j\in\{i_1,...,i_t\}$. With the "wildcard" element *, we can also say that cell $c=(c_1,c_2,...,c_m)$ *contains* incident z if and only if $z.A_j=c_j$ or $c_j=*$, $j=1,2,...,m$. Then we generalize the concept *contain* to cells. We say that cell $c'=(c_1',c_2',...,c_m')$ contains cell $c=(c_1,c_2,...,c_m)$ if and only if $c_j'=c_j$ or $c_j'=*$, $j=1,2,...,m$

Definition 4. Content of a cell

We define function *content* where *content(c)*: $\boldsymbol{C}\rightarrow 2^{\boldsymbol{Z}}$, which returns all the incidents that cell *c* contains. $content(c)=\{z\mid cell\,c\,contains\,z\}$.

Definition 5. Count of a cell

Function *count* is defined in a natural way over the non-negative integers. *count(c)* is the number of incidents that cell *c* contains. *count(c)*=|*content(c)*|.

Definition 6. Parent cell

Cell $c'=(c'_1,c'_2,...,c'_m)$ is the *parent cell* of cell *c* on the *k*-th attribute when: $c'_k=*$ and $c'_j=c_j$, for $j\neq k$. Function *parent(c,k)* returns *parent cell* of cell *c* on the *k*-th attribute.

Obviously, *parent(c,k)* contains cell *c*.

Definition 7. Neighborhood

P is called the *neighborhood* of cell *c* on the *k*-th attribute when ***P*** is a set of cells that takes the same values as cell c in all attributes but *k*, and does not take the wildcard value * on the *k*-th attribute, i.e., $\boldsymbol{P}=\{c^{(1)},c^{(2)},...,c^{(|P|)}\}$ where $c_l^{(i)}=c_l^{(j)}$ for all $l\neq k$, and $c_k^{(i)}\neq *$ for all $i=1,2,...,|P|$. Function $neighbor(c,k)$ returns the neighborhood of cell *c* on attribute *k*. Neighborhood can also be defined in another way: the neighborhood of cell *c* on attribute k is a set of all cells whose parent on the k-th attribute are same as cell *c*.

Definition 8. Relative frequency

We call $freq(c,k)=\dfrac{count(c)}{\sum_{c'\in neighbor(c,k)}count(c')}$ *relative frequency* of cell *c* with respect to attribute k.

Relative frequency can also be defined as: $freq(c,k)=\dfrac{count(c)}{count(parent(c,k))}$

Definition 9. Uncertainty function

We use function U to measure the uncertainty of a neighborhood. This uncertainty measure is defined on the relative frequencies. If we use $P = \{c^{(1)}, c^{(2)}, ..., c^{(|P|)}\}$ to denote the neighborhood of cell c on attribute k, then

$$U : R^{|P|} \to R^{+},$$

where $U(c,k) = U(freq(c^{(1)},k), freq(c^{(2)},k), ..., freq(c^{(|P|)},k))$ Obviously, U should be symmetric for all $c^{(1)}, c^{(2)}, ..., c^{|P|}$. U takes a smaller value if the "uncertainty" in the neighborhood is low.

An uncertainty function that satisfies the above properties is entropy: $H(X) = -\sum p_i \log(p_i)$. Then, $U(c,k) = H(c,k) = -\sum_{c' \in proj(c,k)} freq(c',k) \log(freq(c',k))$ for the above case. This is also the formula for the entropy conditional on the neighborhood projection. For the $freq = 0$, we define $0 \cdot \log(0) = 0$, as is common in information theory.

3.3 Outlier score function (OSF)

A function $f : C \to R^{+}$ is used to measure the extremeness of a cell. We call it an outlier score function. The more extreme a cell is, the higher outlier score it gets. The outlier score function provides us the capability to separate the records contained in the cell from all other records. By introducing the outlier score function, we are able to do "association" and "separation" at the same time.

We recursively define function *f* as:

$$f(c) = \begin{cases} \max\limits_{k \text{ takes all non-* dimension of } c} (f(parent(c,k)) + \dfrac{-\log(freq(c,k))}{H(c,k)}) \\ 0 \quad c = (*,*,...,*) \end{cases} \tag{1}$$

When *H(c,k)* = 0, we say $\dfrac{-\log(freq(c,k))}{H(c,k)} = 0.$

It is simple to verify that this function satisfies the following properties.

I. If $c^{(1)}$ and $c^{(2)}$ are two one-dimension cells, and both of them take non-* value on the same attribute, then $f(c^{(1)}) \geq f(c^{(2)})$ holds if and only if $count(c^{(1)}) \leq count(c^{(2)})$.

II. Assume that $c^{(1)}$ and $c^{(2)}$ are two one-dimension cells, and they take non-* values on two different attributes, say i and j respectively. If $freq(c^{(1)},i) = freq(c^{(2)},j)$, then $f(c^{(1)}) \geq f(c^{(2)})$ holds if and only if $U(c^{(1)},i) \leq U(c^{(2)},j)$, where $c^{(1)}$ takes non-* value on *i*-th, and $c^{(2)}$ takes non-* value on *j*-th attribute respectively. If we define the uncertainty function in an entropy format: $U(c,k) = H(c,k)$, then property II can be rewritten as: $f(c^{(1)}) \geq f(c^{(2)})$ if and only if $H(c^{(1)},i) \leq H(c^{(2)},j)$.

III. $f(c^{(1)}) \geq f(c^{(2)})$ always holds if $\exists k$, $c^{(2)} = parent(c^{(1)}, k)$.

These three properties can be understood as follows. Assume we have a number of robbery incidents with their MO features. The first property means incidents with unusual MOs are more meaningful than common MOs. If we have 100 robberies, 95 happen with the weapon "gun" and the rest 5 are associated with "Japanese sword". Those 5 "Japanese sword" incidents are more likely to be committed by the same criminal. The first property can also be called "Japanese sword" property.

The second property means extremeness level gets reinforced when the uncertainty level is low. Also for the 100 robberies, this time we consider two MO features: weapon used and method of escape. For weapon MO, we have 95 with a value "gun" and 5 with value "Japanese sword"; for method of escape MO, we have 20 values, "by foot", "by car", etc., and each of them have 5 incidents. Although both "Japanese sword" and "by car" cover 5% of all the incidents, they should not be treated equally. "Japanese sword" deserves a higher outlier score because the uncertainty level on "weapon" is lower than "method of escape". We call the second property "augmented Japanese sword" property.

We call the third property "more evidence" property. It says that incidents are more likely to be done by the same person if we have "more evidences", i.e. these incidents take the same value on more attributes. Again for the robbery example, consider two MOs: weapon used and method of escape. We have 5 robberies with "Japanese sword", and 3 out of them also have a same method of escape -- "bicycle", the outlier score for the combination of "Japanese sword" and "bicycle" should be greater than "Japanese sword" only. In our outlier score function, a maximum is used to guarantee this.

3.4 Data association method based on OSF

Based on the outlier score function, we propose a new data association method.

Data association rule

The following rule is used to associated data: for two incidents $z^{(1)}$ and $z^{(2)}$, we say $z^{(1)}$ and $z^{(2)}$ are associated with each other if and only if there exist a cell *c*, *c* contains both $z^{(1)}$ and $z^{(2)}$, and $f(c)$ exceeds some threshold value τ.

Definition 10. Union

$c^{(1)}$ and $c^{(2)}$ are two cells. We call cell *c* the *union* of $c^{(1)}$ and $c^{(2)}$ when both $c^{(1)}$ and $c^{(2)}$ are contained in *c*, and for any c' containing $c^{(1)}$ and $c^{(2)}$, c' contains *c*. It is easy to prove that the *union* exist for any two cells $c^{(1)}$ and $c^{(2)}$. Assume $c^{(1)} = (c_1^{(1)}, c_2^{(1)}, ..., c_m^{(1)})$, $c^{(2)} = (c_1^{(2)}, c_2^{(2)}, ..., c_m^{(2)})$, then $c = (c_1, c_2, ..., c_m)$ will be the *union* of $c^{(1)}$ and $c^{(2)}$, where

$$c_i = \begin{cases} c_i^{(1)}, & if\ c_i^{(1)} = c_i^{(2)} \\ *, & otherwise \end{cases}, i \in \{1,2,...,m\}.$$

Function $Union(c^{(1)}, c^{(2)})$ returns the union of $c^{(1)}$ and $c^{(2)}$. Since each incident can also be treated as a cell, we generalize the *Union* concept to combinations of incident-incident and incident-cell, where $Union(z^{(1)}, z^{(2)}) = Union(Cell(z^{(1)}), Cell(z^{(2)}))$.

From definition of the *union* cell, if any cell c contains both $z^{(1)}$ and $z^{(2)}$, it contains $Union(z^{(1)}, z^{(2)})$. From the *more evidence* property, $f(Union(z^{(1)}, z^{(2)}) \geq f(c)$. Therefore, we can write the following equivalent data association rule:
associate $z^{(1)}$ and $z^{(2)}$, iff $f(Union(z^{(1)}, z^{(2)}) \geq \tau$.

3.5 Discussion about the computational complexity

From the formula (1), we can see that in order to calculate the outlier score for an incident, we need to compute the outlier scores for all its parent cells; and all parents of those parent cells need to be calculated accordingly. There are totally 2^m cells that contain the incident. Therefore, for the worst case, we need to calculate $n \times 2^m$ outlier scores (or $n \times (2^m - 1)$ cells, since the outlier score of all-* cell (*,*,…*) is defined as 0). The computational complexity is linear with respect to the number of the records, and it is exponential with respect to the number of attributes for the worst case. (The worst case does not always happen; actual computational complexity depends on the distribution of the incidents, i.e. how sparse the data is.) Some data dimensionality reduction methods can be applied to lessen the computational burden. We will give this procedure in detail in section 4.

4 Application

4.1 Dataset description

We apply our outlier-based association method to a robbery dataset. The dataset contains information of robbery crimes that occurred in Richmond city, Virginia in 1998. Robbery data is selected for two reasons. First, compared with some "violent" crime types such as murder of sexual attack, multiple robberies are more frequent. Second, there is a sufficient portion of robbery incidents that are solved (with criminals arrested) or partially solved (with an identified suspect). These two points make it preferable to verify our algorithm.

The original crime database is maintained by the police department of Richmond city. Both incident and suspect information are included in the database. For each incident, time, location (both street address and latitude/longitude), and MO features are recorded. Some incidents may have one or more associated suspects; the name, height, weight, and other information about the suspects are stored in the database. There are totally 1198 incidents, and 170 out of them have identified suspect(s), i.e. the name(s) of the suspect(s) are known. Some incidents have more than one suspect, and there are 207 unique (suspect, incident) pairs.

4.2 Selection of attributes and dimensionality reduction

We pick 6 MO features in our analysis, since MOs are typically used in tactical crime analysis. These MO features are listed in table 1(a), and they are categorical. We also incorporate the census data, including demographic and consumer expenditure data. These data are used because some census statistics are helpful to discover the criminals' preference. For example, some

criminals may like to attack some "high-income-level" areas. There are 83 census features. Finally, we use some "distance" features in our analysis. These "distance" features may represent the criminals' spatial preference. For example, we have one "distance" feature called "distance to highway". Criminals may like to initiate the attack at a certain distance range from major highways so that nobody can watch them during the attack and escape as fast as possible after the attack. The MO features and distance features are listed in table 1.The detail description of the census features is given in appendix I. Both census attributes and distance attributes are numerical, and they have also been used in a previous study on prediction of break and entering crime events by Liu. [13]

Table 1. Attributes used in analysis

(a) MO attributes

Name	Description
Rsus_Acts	Actions taken by the suspects
R_Threats	Method used by the suspects to threat the victim
R_Force	Actions that suspects force the victim to do
RVic_Loc	Location type of the victim when robbery was committed
Method_Esc	Method of escape the scene
Premise	Premise to commit the crime

(b) Distance attributes

Name	Description
D_Church	Distance to the nearest church
D_Hospital	Distance to the nearest hospital
D_Highway	Distance to the nearest highway
D_Park	Distance to the nearest park
D_School	Distance to the nearest school

A data dimensionality reduction procedure is performed before applying our outlier association algorithm to the data. Redundant features are unfavorable to data-mining algorithms in terms of both efficiency and accuracy. In our dataset, redundant features exist heavily, since we are using census features. They come from two major sources: one feature may be highly dependent on another feature, e.g. PCARE_PH (expense on personal care per household) and PCARE_PC (expense on personal care per capita); also, several features could be highly linearly dependent, e.g. POP_DST (population density), FEM_DST (female density), and MALE_DST (male density). We use the principle component analysis (PCA) to reduce the dimensions.

PCA [8] is widely applied in many applications. PCA replace the old features with a series of "new" features. Each "new" feature, called a component, is a linear combination of old features, and all these "new" features are orthogonal. The first few components explain most of the variance in the data. Therefore, we can transform the data from original coordinates to these components without losing much information. (Actually, the *k*-th component is the eigenvector with respect to the *k*-th largest eigenvalue of the covariance or correlation matrix of the original dataset, and the eigenvalue represents the proportion of variance explained by the *k*-th component.) We apply PCA to census features and distances, since PCA work on numerical attributes only. The result is given in Table 2.

Table 2. Eigenvalue and variance coverage proportion for first 15 components

	Eigenvalue	Proportion of explained variance	Cumulative proportion of explained variance
1	29.462	0.3348	0.3348
2	16.603	0.1887	0.5235
3	7.431	0.0844	0.6079
4	4.253	0.0483	0.6562
5	3.571	0.0406	0.6968
6	3.121	0.0355	0.7323
7	2.299	0.0261	0.7584
8	2.148	0.0244	0.7828
9	1.837	0.0209	0.8037
10	1.478	0.0168	0.8205
11	1.274	0.0145	0.835
12	1.187	0.0135	0.8485
13	1.112	0.0126	0.8612
14	1.013	0.0115	0.8727
15	0.9540	0.0108	0.8835

The first 4 components are selected because the first 4 components cover almost 2/3 of the variance of the data. Therefore, there are totally 10 attributes — 6 categorical MO attributes and 4 component attributes are chosen in our final analysis.

Our algorithm is designed for categorical attributes. We convert the 4 numerical attributes into categorical ones by dividing them into 11 equal non-overlapping intervals or bins. Obviously, this procedure is same to generating a histogram for density estimation [17]. The number of bins 11 is determined by applying Sturge's number of bins rule [18].

4.3 Evaluation criteria

We want to evaluate whether associated incidents detected by our method corresponds the true result. We use the incidents with one or more identified suspects (whose names are known) for evaluation. All "identified" incident pairs are generated. When two incidents have the same suspect, we say that this pair of incidents is a "true association"; otherwise we call it a "non-association". There are totally 33 "true associations".

Two measures are used to assess the method. The first measure is *number of detected true associations*. We hope the algorithm can discover the "true associations" as many as possible. The second measure is a little more complicated than the first one. We call it *average number of relevant records*. Given one incident, the algorithm will return a list of "relevant records". (We call two incidents are relevant to each other when the incident pair is determined to be a "true association" by the algorithm). The algorithm is more accurate when the length of the list is short. Also, when the result is presented to an end user or crime analyst, the analyst would prefer a short list because that means less effort that they need to put for further investigation. Therefore, we use the average number of "relevant" records as our second criterion. In information retrieval area [16], two most important measures for evaluate the effectiveness of a retrieval system are *recall*

and *precision*. The former is the ability of the system to present all relevant items, and the latter is the ability to present only the relevant items. Our first evaluation criterion can be treated as a *recall* measure and the second one is a *precision* measurement. In addition, our second criterion is a measurement of user effort, which is also an important evaluation criterion used in information retrieval.

One point that we need to mention is that the above measures can be used as evaluation criteria not only for our algorithm, but for any association method as well. Therefore, they can be employed in comparing different methods.

4.3 Result and comparison

We set different threshold levels to test our method. Obviously, when we set the decision threshold τ to 0 all incidents will be determined as relevant by the algorithm, and the corresponding number of detected true associations is 33; on the contrary, if we set the decision threshold to infinity we will get no relevant incident, and the corresponding number of detected true associations is 0. This rule holds for all data association algorithms. As the threshold increases, we expect a decrease in both number of discovered true associations and average number of related records. For different threshold values, the result is given in table 3.

Table 3. Result for outlier-based association method

Threshold	Detected true associations	Avg. number of related records
0	33	169.00
1	33	122.80
2	25	63.51
3	23	29.92
4	17	15.14
5	13	8.05
6	7	4.74
7	4	2.29
∞	0	0.00

We compare our method with a similarity-based association approach. This method was previously proposed by Brown and Hagen, [6]. The idea of similarity-based approach is to calculate similarity scores between incident pairs and perform the data association based upon the similarity score. Using the same two evaluation measures, the result of the similarity-based method is given in table 4.

Table 4. Result for similarity-based association method

Threshold	Detected true associations	Avg. number of related records
0	33	169.00
0.5	33	152.46
0.6	27	84.72
0.7	16	47.41
0.8	8	19.78
0.9	1	3.86
∞	0	0.00

The comparison between outlier-based and similarity-based association method is given in Fig. 1.

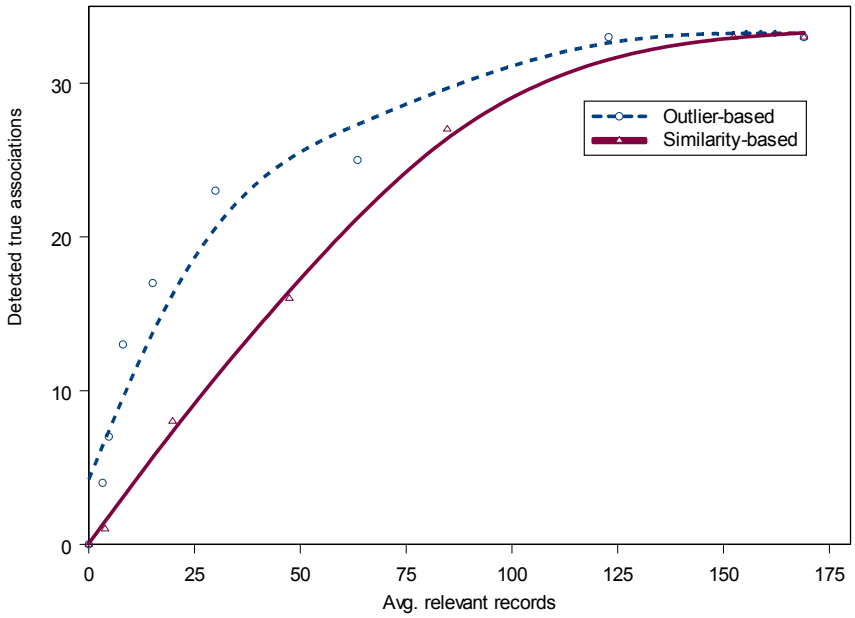

Fig.1 Comparison between outlier-based and similarity-based method

From Fig.1, we can see that the curve for outlier-based method lies above the similarity-based method. That implies given a same "accuracy" level outlier-based method return less number of relevant records, and keeping the same number of relevant record level, outlier-based method is more accurate. The outlier-based data association method outperforms the similarity-based method. This result will help police officer for further investigation.

5 Conclusion

In this paper, we present an outlier-based data association method. When a group of observations have some common characteristics and these characteristics are very different from others (given by the outlier score function), we associate these observations. An outlier score function is built to measure the extremeness level, and based on the outlier score function the data association method is developed. This method is applied to the robbery incident dataset of Richmond, Virginia of 1998, and is compared to a similarity-based association method. Result shows that the outlier-based association method promising.

Acknowledgement

The work reported in this paper was partially supported by the grant from National Institute of Justice.

Appendix I. Census attributes and the description (1997)

Attribute name	Description
General	
POP_DST	Population density (density means that the statistic is divided by the area)
HH_DST	Household density
FAM_DST	Family density
MALE_DST	Male population density
FEM_DST	Female population density
Race	
RACE1_DST	White population density
RACE2_DST	Black population density
RACE3_DST	American Indian population density
RACE4_DST	Asian population density
RACE5_DST	Other population density
HISP_DST	Hispanic origin population density
Population Age	
POP1_DST	Population density (0-5 years)
POP2_DST	Population density (6-11 years)
POP3_DST	Population density (12-17 years)
POP4_DST	Population density (18-24 years)
POP5_DST	Population density (25-34 years)
POP6_DST	Population density (35-44 years)
POP7_DST	Population density (45-54 years)
POP8_DST	Population density (55-64 years)
POP9_DST	Population density (65-74 years)
POP10_DST	Population density (over 75 years)
Householder Age	
AGEH1_DST	Density: age of householder under 25 years
AGEH2_DST	Density: age of householder under 25-34 years
AGEH3_DST	Density: age of householder under 35-44 years
AGEH4_DST	Density: age of householder under 45-54 years
AGEH5_DST	Density: age of householder under 55-64 years
AGEH6_DST	Density: age of householder over 65 years
Household Size	
PPH1_DST	Density: 1 person households
PPH2_DST	Density: 2 person households
PPH3_DST	Density: 3-5 person households
PPH6_DST	Density: 6 or more person households
Housing, misc.	
HUNT_DST	Housing units density
OCCHU_DST	Occupied housing units density
VACHU_DST	Vacant housing units density

Attribute name	Description
MORT1_DST	Density: owner occupied housing unit with mortgage
MORT2_DST	Density: owner occupied housing unit without mortgage
COND1_DST	Density: owner occupied condominiums
OWN_DST	Density: housing unit occupied by owner
RENT_DST	Density: housing unit occupied by renter
Housing Structure	
HSTR1_DST	Density: occupied structure with 1 unit detached
HSTR2_DST	Density: occupied structure with 1 unit attached
HSTR3_DST	Density: occupied structure with 2 unit
HSTR4_DST	Density: occupied structure with 3-9 unit
HSTR6_DST	Density: occupied structure with 10+ unit
HSTR9_DST	Density: occupied structure trailer
HSTR10_DST	Density: occupied structure other
Income	
PCINC_97	Per capita income
MHINC_97	Median household income
AHINC_97	Average household income
School Enrollment	
ENRL1_DST	School enrollment density: public preprimary
ENRL2_DST	School enrollment density: private preprimary
ENRL3_DST	School enrollment density: public school
ENRL4_DST	School enrollment density: private school
ENRL5_DST	School enrollment density: public college
ENRL6_DST	School enrollment density: private college
ENRL7_DST	School enrollment density: not enrolled in school
Work Force	
CLS1_DST	Density: private for profit wage and salary worker
CLS2_DST	Density: private for non-profit wage and salary worker
CLS3_DST	Density: local government workers
CLS4_DST	Density: state government workers
CLS5_DST	Density: federal government workers
CLS6_DST	Density: self-employed workers
CLS7_DST	Density: unpaid family workers
Consumer Expenditures	
ALC_TOB_PH	Expenses on alcohol and tobacco: per household
APPAREL_PH	Expenses on apparel: per household
EDU_PH	Expenses on education: per household
ET_PH	Expenses on entertainment: per household
FOOD_PH	Expenses on food: per household
MED_PH	Expenses on medicine and health: per household
HOUSING_PH	Expenses on housing: per household
PCARE_PH	Expenses on personal care: per household
REA_PH	Expenses on reading: per household

Attribute name	Description
TRANS_PH	Expenses on transportation: per household
ALC_TOB_PC	Expenses on alcohol and tobacco: per capita
APPAREL_PC	Expenses on apparel: per capita
EDU_PC	Expenses on education: per capita
ET_PC	Expenses on entertainment: per capita
FOOD_PC	Expenses on food: per capita
MED_PC	Expenses on medicine and health: per capita
HOUSING_PC	Expenses on housing: per capita
PCARE_PC	Expenses on personal care: per capita
REA_PC	Expenses on reading: per capita
TRANS_PC	Expenses on transportation: per capita

Reference

[1] Aggarwal, C., Yu, P., *Outlier Detection for High Dimensional Data*, SIGMOD Conference Proceedings, 2001

[2] Andrews, D. Bickel, P., Hampel, F., Huber, P., Rogers, W., and Tukey, J. *Robust Estimate of Location*, Princeton University Press, 1972

[3] Badiru, A.B., Karasz, J.M. and Holloway, B.T., *AREST: Armed Robbery Eidetic Suspect Typing Expert System*, Journal of Police Science and Administration, 16, 210-216, 1988

[4] Barnett, V. and Lewis, T., *Outliers in Statistical Data*, John Wiley, 1994

[5] Breunig M. M., Kriegel H.P., Ng R., Sander J.: *LOF: Identifying Density-Based Local Outliers*, Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2000), 93-104, 2000

[6] Brown, D.E and Hagen, S., *Data Association Methods with Applications to Law Enforcement*, Decision Support Systems, 2002, to be appeared

[7] Gray, J., Chaudhuri, S., Bosworth, A. Layman, A., Reichart, D., Venkatrao, M., Pellow, F., and Pirahesh, H., *Data Cube: A Reletinal Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals*, Data Mining and Knowledge Discovery, 1, 29-53, 1997

[8] Hastie, T. Tibshirani, R., Friedman, J., *The Element of Statistical Learning: Data Mining, Inference, and Prediction,* Springer-Verlag, New York, 2001

[9] Hawkins, D., *Identifications of Outliers*, Chapman and Hall, London, 1980

[10] Heck, R. O., *Career Criminal Apprehension Program: Annual Report* (Sacramento, CA: Office of Criminal Justice Planning), 1991

[11] Knorr, E. and Ng R., *A Unified Notion of Outliers: Properties and Computation*, In Proc. of the Int. Conf. on Knowledge Discovery and Data Mining, 219-222, 1997

[12] Knorr, E., Ng, R., *Algorithms for Mining Distance-based Outliers in Large Datasets,* VLDB Conference Proceedings, September 1998

[13] Liu, Hua, *Space-Time Point Process Modeling: Feature Selection and Transition Density Estimation*," Dissertation for Systems Engineering University of Virginia, 1999

[14] Olap council. *www.olapcouncil.org*.

[15] Ramaswamy, S., Rastongi, R., and Shim, K., *Efficient Algorithms for Mining Outliers from Large Data Sets,* Proc. of the ACM SIGMOD Conference, 427-438, 2000

[16] Salton, G. and McGill, M. *Introduction to Modern Information Retrieval,* New York: McGraw-Hill Book Company, 1983

[17] Scott, D. *Multivariate Density Estimation: Theory, Practice and Visualization*, New York, NY: Wiley, 1992

[18] Sturges, H.A., *The Choice of a Class Interval*, Journal of American Statistician Association, 21, 65-66, 1926