Technical Report

Outlier-based Data Association: Combining OLAP and Data Mining

Song Lin & Donald E. Brown

Department of Systems Engineering

University of Virginia

SIE 020011

December. 2002

Outlier-based Data Association:
Combining OLAP and Data Mining

Song Lin Donald E. Brown
sl7h@virginia.edu brown@virginia.edu
Department of Systems and Information Engineering
University of Virginia, Charlottesville, VA 22904

Abstract

Both data mining and OLAP are powerful decision support tools. However, people use them separately for years: OLAP systems concentrate on the efficiency of building OLAP cubes, and no statistical / data mining algorithms have been applied; on the other hand, statistical analysis are traditionally developed for two-way relational databases, and have not been generalized to the multi-dimensional OLAP data structure. Combining both OLAP and data mining may provide excellent solutions, and in this paper, we present such an example – an OLAP-outlier-based data association method. This method integrates both outlier detection concept in data mining and ideas from OLAP field. An outlier score function is defined over OLAP cells to measure the extremeness level of the cell, and when the outlier score is significant enough, we say the records contained in the cell are associated to each other. We apply our method to a real-world problem: linking criminal incidents, and compare our method with a similarity-based association algorithm. Result shows that this combination of OLAP and data mining provides a novel solution to the problem.

Keyword: OLAP, data mining, data association, outlier

1. Introduction

The concept of data warehousing was introduced in 1990's. It is a collection of technologies that assist managers of an organization to make better decisions. Online analytical processing (OLAP) is a key feature supported by most data warehousing systems. (Chaudhuri and Dayal, 1997; Codd, et al., 1993; Shoshani, 1997; Welbrock, 1998)

OLAP is quite different from its ancestor, online transaction processing (OLTP) systems. OLTP focuses on automation of data collecting procedure. Keeping detailed, consistent, and up-to-date data is the most critical requirement for an OLTP application. Although as the fundamental building blocks these transactional records are important to an organization, a decision maker is more interested in the summary data than investigating a particular record. Traditional relational database management system (DBMS) is not efficient enough to satisfy the requirement of OLAP since to acquire summary information need a number of aggregation SQL queries with group-by clauses.

The OLAP concept was introduced to satisfy the requirement of efficiency. Summary or aggregation data, such as sum, average, max, and min, is pre-calculated and stored in a *data cube*. Compared with two-way relational tables normally used in OLTP, a data cube is *multidimensional*. Each *dimension* consists of one or more categorical attributes, and hierarchical structures generally exist in the dimensions. The architecture of a typical OLAP application is showed in Fig. 1.

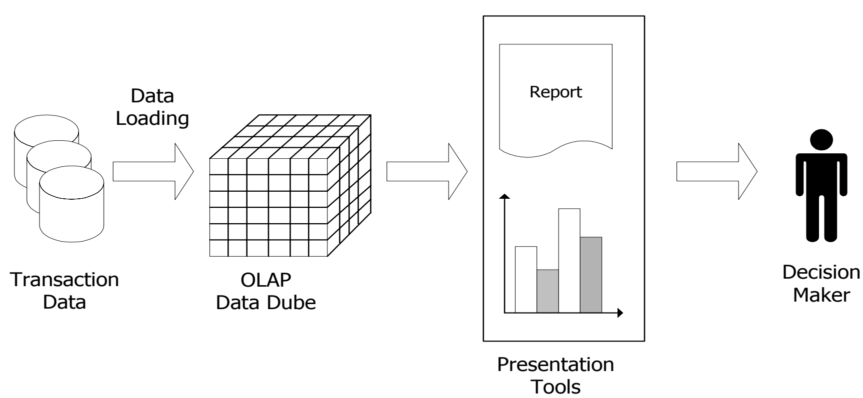

Fig. 1. OLAP Architecture

Although OLAP is capable to provide summary information efficiently, how to make the final decision is still an art of applying the domain knowledge, sometimes common sense, of the decision-maker. Few quantitative data mining methods, like regression or classification, have been introduced into the OLAP arena. On the other hand, traditional data mining algorithms are mostly designed for two-dimension dataset, and OLAP is not involved in developing the data mining algorithm. Since both OLAP and data mining are powerful tools for the decision making process, the ideal situation is to combine both of them to solve the real-world problem, as illustrated in Fig. 2.

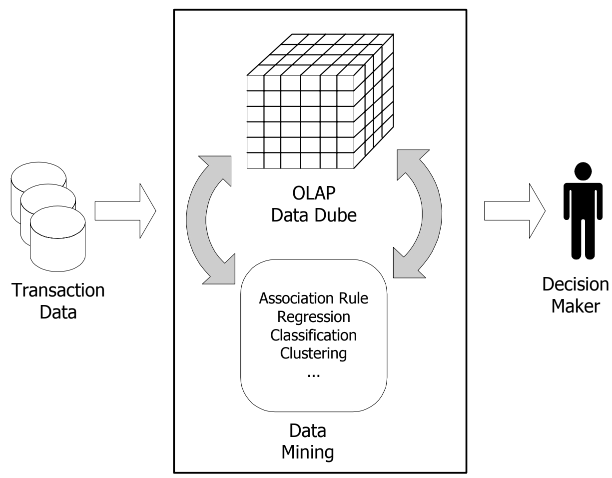

Fig. 2. Ideal situation: combining both OLAP and data mining

In this paper, we present an example of combining both the data mining and OLAP to solve the data association problem. Data association involves linking or grouping records in the database according to similarity or other mechanisms, and many applications can be treated as data association problems. For example, in multiple-sensor or multiple-target tracking (Avitzour, 1992; Jeong, 2000; Power 2002), we want to associate different tracks of the same target; in document retrieval system (Salton, 1983), we want to associate documents with the given searching string; in crime analysis (Brown and Hagen, 1999), we want to associate crime incidents together did by the same criminal. Different approaches have been proposed to solve the data association problem. In this paper, we present a new data association method – an OLAP outlier based data association method. This method integrates both the concept of outlier detection from data mining field and OLAP techniques seamlessly. The data is first modeled into an OLAP data cube, and an outlier score function is built over OLAP cells. The outlier score function measures the extremeness level of the OLAP cell, and we associate the records in the cell when the outlier score is exceptionally large. We apply our method to a real-world problem: linking criminal incidents. Result shows that it potentially provides an excellent solution to incorporate both OLAP and data mining. Also, we will show that some data mining techniques, including feature extraction and selection, can be employed in building the OLAP cubes.

The rest parts of this paper are organized as follows: in section 2, we briefly look through the background and the initial motivation of the study, as well as previous works on outlier detection and OLAP data mining; our OLAP outlier association method is given in section 3; in section 4, we apply this method to the robbery dataset of Richmond city, and compare our method with a similarity-based method; section 5 concludes the paper.

2. Background and related work

2.1 Background and motivation

Data mining is a collection of "techniques that can be used to find underlying structure and relationship in a large amount of data" (Kennedy et al., 1998) and it has been applied to many real-world applications, such as manufacturing process control, medical diagnosis, and credit card fraud detection (Kennedy et al., 1998; Ripley 1996; Hastie et al., 2001) Various data mining techniques have also been introduced to law enforcement field. A number of models and systems have been developed by researchers. (Corman and Mocan, 2000; Gunderson and Brown, 2001; Brown et al., 2001; Osgood, 2000), and the Regional Crime Analysis Program (ReCAP) at the University of Virginia represents one example.

ReCAP (Brown, 1998) is a shared information and decision support system that assists the police departments to analyze and prevent crimes. The system is an integration of three parts: a database management system, a Geographic Information System (GIS), and a statistical analysis toolkit. Our study in this paper was initially motivated by adding a new tactical crime analysis component to the system. Tactical analysis, a term used in criminology, means to associate crime incidents committed by the same person. This association is important in crime analysis, and it will help to discover the crime patterns, make predictions for future crimes, and even catch the criminal.

Different methods have been proposed and some software programs have been developed to solve this crime association problem in the past two decades. R.O.Heck introduced the Integrated Criminal Apprehension Program (ICAP) (Heck, 1991). ICAP enables police officer to perform a matching between the suspects and the arrested criminals using Modus Operandi (MO) features. Similar to ICAP, the Armed Robbery Eidetic Suspect Typing (AREST) program (Badiru et al., 1988) is also capable to make a suspect matching. AREST employed an expert system approach: a set of key characteristics and rules were set by crime analysts, and the system classify a potential offender into three categories: probable suspect, possible suspects, and non-suspects. Different to these two suspect matching systems, FBI developed the Violent Criminal Apprehension Program (ViCAP) (Icove, D.J., 1986), which focus on associating crime incidents. An expert approach is used in ViCAP. In the COPLINK project (Hauk et al., 2002), a concept space model was adopted. Brown and Hagen (Brown and Hagen, 1999) proposed a similarity-based crime association method. A similarity score is evaluated for each field, and a total similarity score is calculated as the weighted average of similarity scores across all attributes. Their method can be applied to both suspect and incident matching. Besides those theoretical methods, crime analysts normally use the Structured Query Language (SQL) in practice. They build the SQL string and make the system return all records that match their searching criteria.

When every piece of the information about the crime incident is observed and recorded, above methods perform well. For example, if the fingerprint or DNA sequence of the criminal is known, we only need to make a precise match. However, that happens rarely in practice. Some descriptions about the suspect may look like “white male with blonde hair and blue eyes”. If we set the matching criteria as “Gender = ‘male’ and Race = ‘white’ and hair_color = ‘blonde’ and eye_color = ‘blue’”, we will expect a long list of names. Definitely we cannot conclude all those people are the same criminal. The reason is that the combination of "white male, blond hair, and blue eyes" is quite *common*, and these *common* features make the record not distinguishable. We hope to develop a new method that is capable to identify this distinctiveness, and that leads to our outlier-based data association approach.

2.2 Existing work on outlier detection

Outliers exist extensively in real world, and they are generated from different sources: a heavily tailed distribution or errors in inputting the data. Mostly, they are "so different from other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins, 1980). Finding outliers is important in different applications, such as credit fraud detection and network intrusion detection. Traditional studies on detecting outliers lie in the field of statistics, and a number of statistical tests, called discordancy tests, are developed (Barnett and Lewis, 1994; Hawkins, 1980). In some practices like monitoring a manufacturing process, a 3σ rule is generally adopted. The 3σ rule is: calculating the mean μ and the standard deviation σ, and if one observation lies outside the $(\mu-3\sigma, \mu+3\sigma)$ range, we say it an outlier. All these methods are developed to detect a single outlier, and they may fail when multiple outliers exist. Some researchers suggest using the median and the MAD scale instead of the mean and the standard deviation for detecting multiple outliers (Andrews, et al., 1972).

These approaches are designed to find outliers in a univariate dataset, and the data points in the dataset are assumed to follow some standard distribution, such as a normal or Poisson distribution. However, most real-world data are multivariate and it is difficult to make assumptions of the underlying distribution.

Some researchers proposed different methods for detecting outliers in multivariate data without the a-priori assumption of the distribution. Knorr and Ng gave their definition of distance-based outliers (Knorr and Ng, 1997, 1998), A points is called *DB(p,D)* outlier if at least a portion of *p* of the points in the dataset keeping a distance from greater than *D*.

They also proved that their notion of the distance-based outlier unified the outlier definitions in some standard distribution. Then they gave several algorithms, including an OLAP version, to find all distance-based outliers. Ramaswamy et al. (Ramaswamy et al., 2000) argued that the *DB(p,D)* outliers are too sensitive to the parameter *p* and *D*. They defined a k-nearest neighbor outlier. They calculate the *k*-th nearest distances for all data points and rank the points according to these distances, and then pick the top *n* as outliers. Breunig et al. (Breunig et al., 2000) proposed another notion of "local" outliers. They think that a data point is an "outlier" only when we consider a "local" neighborhood of the points. They assign each object with an outlier degree, which they call *local outlier factor*. Thus, they use a continuous score to measure the outlier instead of give the binary result yes or no. Aggarwal et al. (Aggarwal et al., 2001) claim that both the *distance-based* and *local* outliers do not work well for high dimensional dataset since the data are sparse, and outliers should be defined in sub-space projections. They proposed an evolutionary algorithm to find the outliers.

These methods are developed to detect individual outliers, and the association of outliers has not bee studied. In this paper, we present an outlier-based data association method. Instead of defining outlier for individual record, we consider to build the outlier measure for a group of data points. These data points are "similar" on some attributes and are "different" on other attributes. If these common characteristics are quite "unusual", or in other words, they are "outliers", these data points are well separated from other points. The "weird" characteristics strongly suggest that these data points are generated by a particular mechanism, and we should associate these points.

2.3 Studies on combining OLAP and data mining

Some researchers began to generalize some data mining concepts on OLAP cubes in recent years. These works include the cubegrade problem (Imielinski et al., 2000), the constrained gradient analysis (Dong et al., 2001), and data-driven OLAP cube exploration (Sarawagi, et al. 1998). We will review these studies briefly in this section.

The cubegrade problem was posed by Imielinski et al (Imielinski et al., 2000). It is a generalized version of association rule (Agrawal et al. 1993). Two important concepts in association rule are support and confidence. Let us take the market basket example. Support is the fraction of transactions that contains a certain item (bread and butter), and confidence is that the proportion of transactions that contains another item B given that these transactions contain A. Imielinski et al. declare that the association rule can also be viewed as the change of the *count* aggregates when imposing another constraint, or in OLAP terminology, making a drill-down operation on an existing cube cell. They think other aggregates like *sum, average, max*, and *min* can be studied in addition to the *count*. Also, other OLAP operations, like roll-up and one-dimension mutation can be incorporated. They argued that the cubegrade could support the "what if" analysis better, and they introduced two query languages to retrieve the cubegrade sets.

Constrained gradient analysis (Dong et al., 2001) is similar to the cubegrade problem. It focuses on extracting pairs of OLAP cube cells that are quite different in aggregates and similar in dimensions (usually one cell is the ascendant, descendent, or sibling of the other cell). Instead of dealing the whole OLAP cube, some constraints (significance constraints and probe constraints) are added to limit the search range, and more than one measures (aggregates) can be considered simultaneously.

The discovery-driven explorations were proposed by Sarawagi et al. (Sarawagi, et al. 1998), and it aims at finding exceptions in the cube cells. They define a cell as an exception as the measure (aggregate) of the cell differs significantly from its anticipated value. The anticipated value is calculated by some formula and they suggest an additive or multiplicative form. They also give the formula to estimate the standard deviation. When the difference between the cell value and its anticipated value is greater than 2.5 standard deviation, the cell is an exception. Their method can be treated as an OLAP version of the 3σ rule.

Similar to above works, we also focus on OLAP cube cells in our analysis. We define a function on OLAP cube cells to measure the extremeness of the OLAP cell. When the cell is an "outlier", we say the data points contained in this cell are associated. Hence this method combines both outlier detection in data mining and concepts from OLAP. We hope to apply this technique to resolve a real-world problem: associating crime incidents.

3. The OLAP-outlier-based data association

3.1 Basic idea

The basic idea of this method originates from the "Japanese sword" claim, first proposed by Brown and Hagen (Brown and Hagen, forthcoming). Assume we have a number of robbery incidents. If the weapon used in some incidents is a "gun", we cannot associate these incidents because "gun" is too common. However, if we have a couple of incidents with some special weapon, say "Japanese sword", we can confidently assess that these two incidents are done by the same person.

We generalize this claim and restate it in OLAP terms as follows: if we have a group of records contained in a cell, and this cell is very "different" from other cells (or this cell is an outlier), then these records are probably generated from a same causal mechanism and hence they are associated with each other.

3.2 Definitions

In this section, we give the mathematical definitions of the concepts and notations that will be used in the reminder of the paper. People familiar with OLAP concepts can see that our notations are same or very similar from the terms used in OLAP field.

$A_1, A_2, \ldots, A_m$ are m attributes that we consider relevant to our study, and $D_1, D_2, \ldots, D_m$ are their domains respectively. Currently, these attributes are confined to be categorical. (these attributes are "dimensions" of the OLAP cube). 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 $c_{i_s} \in D_{i_s}$. 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.

The term dimension may bring confusion there is another "dimension" in OLAP. We still use the term dimension here because this paper is for both people from OLAP and people from other field. In rest of the paper, we will use the term *OLAP dimension* explicitly.

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., ***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.

One candidate 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 above case. This is also the formula for the entropy conditional on the neighborhood. When the $freq = 0$, we define $0 \cdot \log(0) = 0$, as is common in information theory.

Definition (1) to (7) comes directly from the OLAP area, (some words may not be exactly the same as used in OLAP. For example, we use the term "neighbor" instead of "sibling"), and we rewrite it in a more mathematical and formal manner so that people from fields other than OLAP can understand them as well.

3.3 Outlier score function (OSF)

A function $f: C \rightarrow R^{+}$ 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.

We recursively define function *f* as:

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

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 $\exists k$, $c^{(2)} = parent(c^{(1)},k)$.

The following example gives some explanation that why these properties need to be satisfied. Assume we have 100 robbery incidents with their MO features:

The first property says that unusual attribute values provide more information. If out of these 100 incidents, 95 have a “gun” involved, and 5 are “Japanese swords”. Then the 5 “Japanese sword” incidents are more likely to be done by the same person, and the “Japanese sword” cell deserves a higher outlier score than the “cell”. The first property is also call “Japanese sword” property for simplification.

The second property means that when we define the concept of “outlier”, we need to consider not only this cell, but other cells as well. The extremeness level gets reinforced when the uncertainty level is low. Now we consider two MO features: weapon and method of escape. For weapons, we have 95 “guns” and 5 “Japanese swords”; for method of escape, we have 20 values, “by foot”, “by car”, etc., and each of them cover 5 incidents. Obviously, we should value “Japanese sword” and “by foot” differently. The cell “Japanese sword” is more unusual because the uncertainty level on this attribute is lower. We call this property “augmented Japanese sword” property.

Property I and II leads to a natural formula of the outlier score function for one-dimension case: $f = f_1 / f_2$, where f_1 increase as the count (frequency) of the cell decrease, and f_2 increase as the uncertainty level increase. We use *–log(frequency)* as f_1 and entropy as f_2. Both of them come from information theory.

Then we generalize the formula to high dimension. An intuitive thought is to take the summation over all one-dimension as the overall outlier score. However, that brings a problem when the attributes are not independent. For example, we have one attribute “eye color” and another attribute “hair color”. For the former we have a value “black” and for the latter we have “blonde”. Both people with black eyes and people with blonde hair are not rare, but their combination, “blond people with black eyes” are quite unusual. That is the reason why we bring the concept “neighbor” and “relative frequency” in our formula. When all attributes are independent, it is obvious that our definition is that same as take summation over all attributes.

The third property is easy to understand, and we call it "more evidence" property. For the 5 "Japanese swords", there are 3 incidents that also have a same method of escape "by bicycle". The outlier score for the cell of the combination of "Japanese sword" and "bicycle" should be greater than cell "Japanese sword" only, because we have "more evidence" to show that these 3 incidents result from the same person. In our outlier score function definition, a maximum is used to guarantee this.

In our method, there is no numerical measure and "count" is selected as the aggregate function. Apparently, this method can be generalized to numerical measures typically used in most OLAP applications like sales and other aggregates functions like sum or average with some slight modification. One modification is to discretize the numerical aggregation into bins or use some probability density estimation techniques (Scott, 1992). Additionally, this method can be generalized when there is a hierarchical structure in some dimensions.

3.4 Data association method

We associate incidents in a cell when the cell is an outlier, and the rule is as follows: for 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.

This rule requires checking all cells, which is computationally inefficient. From the "more evidence" property, we know that we only need to verify the "smallest" cell that contains $z^{(1)}$ and $z^{(2)}$.

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 simple to prove that the union cell always exists.

Therefore we have another association rule as follows: a*ssociate* $z^{(1)}$ *and* $z^{(2)}$, *iff* $f(Union(z^{(1)}, z^{(2)})) \geq \tau$.

4. Application

4.1 Dataset description

We apply this OLAP-outlier-based data association method to the robbery dataset of Richmond city, Virginia in 1998. The original database is maintained by the police department of Richmond city, and information of different types of crimes is stored in the database. We choose robbery mainly for two reasons: first, compared with some "violent" crime types such as murders or sexual attacks, multiple robberies are more "frequent"; second, there is a sufficient portion of robbery incidents that are solve (with the criminals arrested) or partially solved (with one or more identified suspects). These two features make it preferable to verify our algorithm.

Both incident and suspect information are included in the database. For incidents, the time, location (both street address and latitude/longitude), and MO features are recorded; for suspects, height, weight, and other information are recorded (height and weight are sometimes estimated by the victim/witness). Most robberies are associated with one or more suspects. Some of the suspects are identified, and if a suspect is identified, his or her name is stored in the dataset. There are totally 1198 robberies, and 170 have identified suspects. Some incidents have more than one suspect, and there are totally 207 unique (suspect, incident) pairs.

4.2 Select attributes for analysis

Three types of attributes are selected for our analysis: MO features, census features, and distance features. MOs are important in tactical analysis and we pick 6 MO features, as listed in table 1 (a). All MO features are categorical. Census data is acquired from "Census CD + maps" (1998) and it contains two part: demographic and consumer expenditure data. Census statistics are helpful to reveal the criminals' preference. For example, some criminals may prefer to attack some "high-income-level" areas. There are totally 83 census attributes, and the detail description is given in appendix I. Finally, we incorporate distance attributes. Distance attributes are distances from the crime incidents to some spatial landmarks, such as a highway, and they represent criminals' spatial behavior preferences. Criminals may like to initiate the attack at a certain distance range from major highway so that nobody can watch them during the attack and then escape as fast as possible after the attack is completed. Distance attributes are calculated using the latitude/longitude and the ArcView GIS software. Distance attributes are listed in table 1 (b). Both the census and distance attributes are employed in a previous study on prediction of break and entering crimes by Liu (Liu, 1999).

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

4.3 Building the OLAP dimensions / feature extraction and selection

One important issue in building OLAP data cube is to select the OLAP dimensions. In our dataset, census and distance features cannot be used as OLAP dimensions for two reasons: first, many features are redundant and these redundant features are unfavorable in terms of both accuracy and efficiency; second, they are numerical and an OLAP data cube dimension has to be categorical. We apply some data mining approaches, including feature extraction, feature selection, and density estimation, to resolve this problem.

4.3.1 Redundant features

Redundant features exist heavily in our dataset because we are using census features. The redundancy normally represents as linear dependency among attributes. For example, both PCARE_PH (expense on personal care per household) and PCARE_PC (expense on personal care per capita) are statistics of personal care expenses, and they are similar to each other. So the key idea here is to remove the linear dependency. We consider using two methods: principal component analysis (PCA) and feature selection.

Using PCA to build the OLAP dimensions

PCA is widely applied in many applications (Hastie, 2001). PCA replaces 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.) For census and distance features, we apply PCA, and result is given in table 2.

We pick the first 4 components since they cover almost 2/3 of the variance in the data. Therefore, we have totally 10 dimensions (6 MOs and 4 PCA components).

Table 2. PCA for census and distance features (first 10 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

Using feature selection to build the OLAP dimensions

Another candidate approach to build the OLAP dimensions is feature selection. Compared with PCA, feature selection select a subset of the features from the original feature set instead of generating a series of linear combinations. One advantage about feature selection is that the selected features are easy to interpret for most cases.

Various feature selection methods have been proposed for both supervised and unsupervised learning (Liu and Motoda, 1998). We adopt a "selecting features by clustering" approach. The idea is to partition the original feature set into some clusters, and each cluster consists of a number of features that are similar to each other. One representative feature is selected for each cluster and we take all representative features as our result. The similar idea has been applied in previous study by Mitra et al. (Mitra, 2002).

In our analysis, we use the correlation coefficient ρ to measure how similar two features are, and the k-medoid partitioning method (for detail about different clustering methods, see (Everitt, 1993)) is selected as the clustering algorithm. We choose the k-medoid method for three reasons: first, it works on both similarity/distance matrix and coordinate data while some other methods work only on coordinate data; second, it tends to group similar records together; third, it returns medoids, based on which we can give representative features. By checking the silhouette plot (Kaufman and Rousseeuw, 1990), we finally get three clusters, as given in Fig. 3.

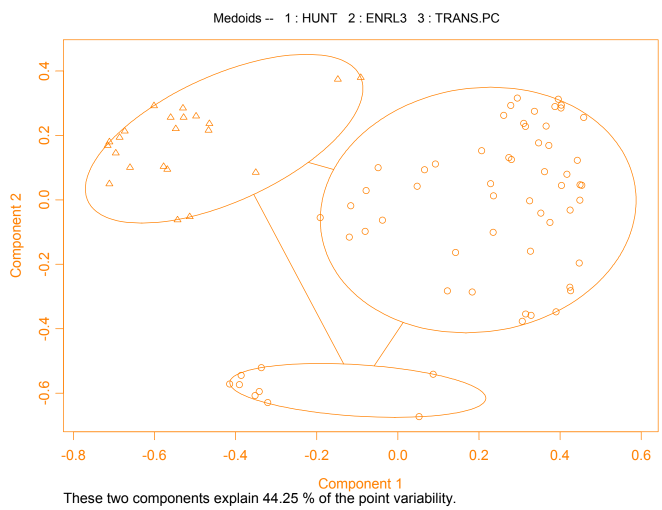

Fig. 3. Bivariate plot of the medoid clustering

The three medoids given by the clustering algorithms are HUNT_DST (housing unit density), ENRL3_DST (public school enrollment density), and TRAN_PC (expenses on transportation: per capita). Some adjustment is made based on this result. For ENRL3, we replace it with another feature POP3_DST (population density: age 12-17) (they are in the same cluster). The reason is that age 12-17 is more meaningful in criminology. Crime analysts consider people in this age range are likely to be both attackers and victims. Another reason is that the correlation coefficient of ENRL3_DST and POP3_DST are 94%, which means they are very similar. For the same reason, we change the TRAN_PC to MHINC (median household income): the latter is easier to interpret.

4.3.2 Discretization

The selected features (either through PCA or feature selection) are numerical, and we transform them into categorical ones by dividing them into 11 equal-length intervals or bins. This procedure is same to generating the histogram for density estimations (Scott, 1992). The number of intervals is derived by applying the Sturges' number of bins rule (Sturges, 1926).

4.4 OLAP implementation

OLAP is typically implemented through three strategies: relational OLAP (ROLAP), multidimensional OLAP (MOLAP), and hybrid OLAP (HOLAP). Our method takes a ROLAP strategy. Cells are stored in tables in a relational database. The algorithm is implemented in Visual Basic and Microsoft Access.

4.4 Evaluation criteria

We want to evaluate whether associations given by our method correspond the true result, and 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" or they are "relevant" to each other; 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 (apparently it cannot exceed the total number of true associations). The second measure is *average number of relevant records*, and it is slightly complicated than the first one. It is explained as follows:

For any association algorithm, given one incident, the algorithm will return a list of "relevant" incidents. Since the number of true associations is fixed, 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 shorter list because that means less effort that they need to put for further investigation. (Think about the search engines like *google*. It is better that the user can find the document in 5 pages than in 10 pages.) Therefore, we take the average of all the "relevant records lists" as the second measure.

In information retrieval studies (Salton, 1983), two most important measures for evaluating 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.

4.5 Result and comparison

Different threshold values are set 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.

We compare our method with a similarity-based association approach. This method was previously proposed by Brown and Hagen, (Brown and Hagen, 1999). The idea of similarity-based approach is to calculate similarity scores between incident pairs and a total similarity score is calculated as the weighted average of the similarity scores. Whether a pair of incidents is associated is determined by the total similarity score.

The PCA and feature selection is used for both our method and the similarity-based method, and we compare the results, as given in table 3 and 4 respectively.

Table 3. Comparison: features generated by PCA

(a) Outlier-based 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

(b) Similarity based 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

Table 4. Comparison: features selected by clustering

(a) Outlier-based method

Threshold	Detected true associations	Avg. number of related records
0	33	169.00
1	32	121.04
2	30	62.54
3	23	28.38
4	18	13.96
5	16	7.51
6	8	4.25
7	2	2.29
∞	0	0.00

(b) Similarity-based method

Threshold	Detected true associations	Avg. number of related records
0	33	169.00
0.5	33	112.98
0.6	25	80.05
0.7	15	45.52
0.8	7	19.38
0.9	0	3.97
∞	0	0.00

If we set the average number of relevant records as the X-axis and set the detected true associations as the Y-axis, the comparisons can be illustrated as in Fig. 4 and 5. Obviously, both the similarity-based and the outlier-based method have the same starting and ending points.

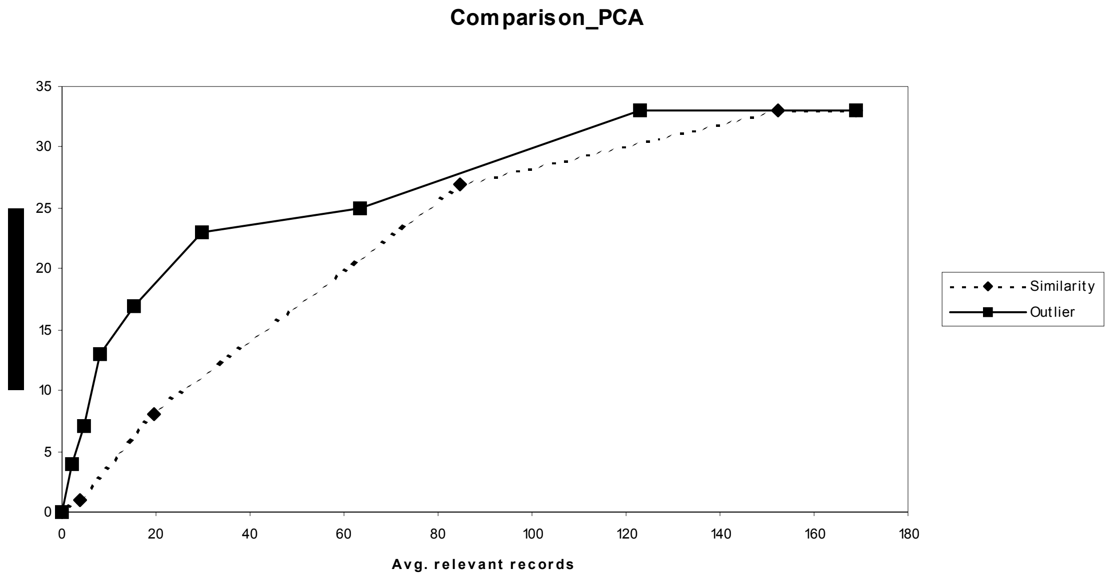

Fig. 4. Comparison: dimension generated by PCA

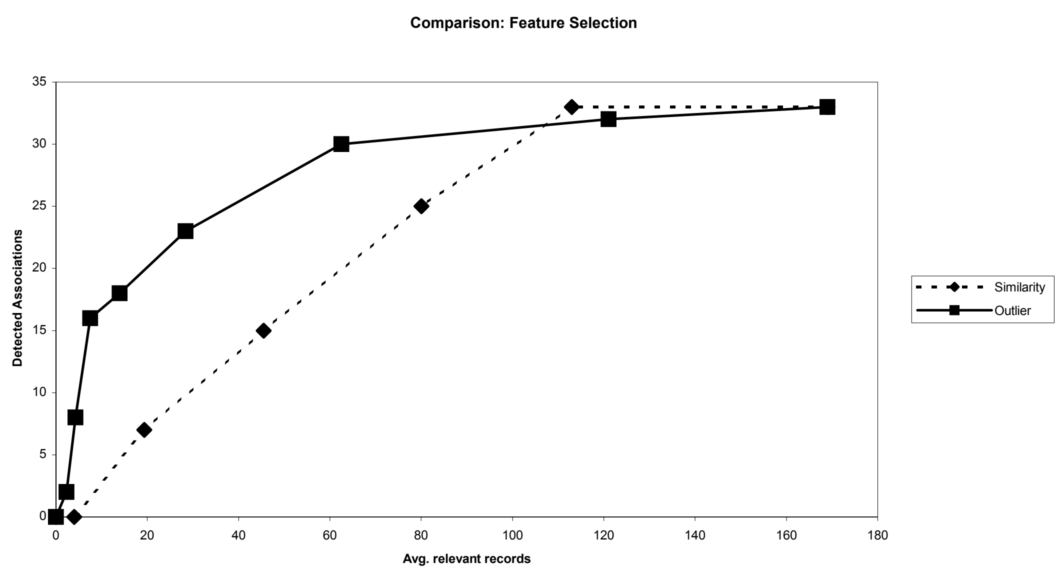

Fig. 5. Comparison: dimension generated by feature selection

4.6 Discussion

From Fig. 4, we can see that the curve of outlier-based method lies above the curve of similarity-based method. That implies that given the same "accuracy" level outlier-based method return less number of relevant records, and keeping the same number of relevant records level, outlier-based method is more accurate. Hence the outlier-based association method outperforms the similarity-based method.

For Fig. 5, we can see that for most cases, the outlier-based method lies above the curve of similarity-based method. Similarity-based method sits slightly higher than the outlier-based method when the average number of relevant records level is set above 100, which means the algorithm is expected to determine 100 crime incidents "relevant" to the given incidents. Given there totally 170 identified suspects, 100 is not a number that the crime analysts hope to investigate. So, we still consider outlier-based method a better approach.

For OLAP dimensions generated by PCA and feature selection, the OLAP-outlier-based association appears to be a promising method, and these results will help police officers for further investigation.

5 Conclusion

In this paper, we present a new data association method. This method combines both OLAP concepts and outlier detection idea from data mining. An outlier score function is defined on OLAP cube cells and it measures the extremeness of the cell. We associated the records in the cell when the cell is "unusual". This method is applied to the robbery dataset of Richmond city, and is compared with a similarity-based association method. Result shows that:

- The outlier-based method outperforms the similarity-based method.
- Combining OLAP and data mining potentially provides a powerful tool to solve real-world problem.

Also, in our analysis, we show that some data-mining techniques, such as feature extraction (PCA) and selection, can be applied in building the OLAP more efficiently.

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

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

Agrawal, R., Imielinski, R., and Swami, A., "Mining association rules between sets of items in large databases", *Proc. of ACM SIGMOD Conference on Management of Data*, 207-216, Washington D.C., May 1993.

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

Avitzour, D., "A Maximum Likelihood Approach to Data Association", IEEE Trans. Aerospace and Electronic Systems, Vol. 28, 560-565, 1992

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

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

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

Brown, D.E., "The Regional Crime Analysis Program (ReCAP): a framework for mining data to catch criminals", *Proc. of IEEE International Conference on Systems, Man, and Cybernetics*, 2848 –2853, 1998

Brown, D.E.; Hagen, S.C., "Correlation analysis for decision support with applications to law enforcement", *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, 1074 –1078, 1999

Brown D.E. and Hagen S.C., “Data Association Methods with Applications to Law Enforcement”, *Decision Support Systems*, forthcoming

Brown, D.E., Liu, H. and Xue, Y., “Mining Preference from Spatial-temporal Data”, *Proceedings of the First SIAM International Conference of Data Mining*, 2001

Chaudhuri, S. and Dayal, U., “An Overview of Data Warehousing and OLAP Technology”, *ACM SIGMOD Record* 26(1), March 1997.

Codd, E.F., Codd, S.B., and Salley, C.T., “Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate”. *Technical report*, 1993

Corman, H. and Mocan, H.N., “A Time-Series Analysis of Crime, Deterrence and Drug Abuse in New York City”, *The American Economic Review*, 90, 584-604, 2000

Dong, G., Han, J., Lam, J. Pei, J., and Wang, K., “Mining Multi-Dimensional Constrained Gradients in Data Cubes”, *Proc. of the 27th VLDB Conference*, Roma, Italy, 2001.

Everitt, B. *Cluster Analysis*, John Wiley & Sons, Inc., 1993

Gunderson. L.F. and Brown, D.E. “Using Cluster Specific Salience Weighting to Determine the Preferences of Agent for Multi-agent Simulations”, *Proceedings of IEEE International Conference on Systems, Man, and Cybrenetics*, 1465-1470, 2001

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

Hauck, R., Atabakhsh, H., Onguasith, P., Gupta, H., and Chen, H., “Using Coplink to Analyse Criminal-Justice Data”, *IEEE Computer*, vol. 35, 30-37, 2002

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

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

Icove, D. J., “Automated Crime Profiling”, *Law Enforcement Bulletin*, 55, 27-30, 1986

Imielinski, T., Khachiyan, L., and Abdul-ghani, A., “Cubegrades: Generalizing association rules”. *Technical report*, Dept. Computer Science, Rutgers Univ., Aug. 2000.

Jeong, H. and Park, J., A Multiple-Target Tracking Filter Using Data Association Based on a MAP Approach, IEICE Trans. Fundamentals, Vol. E83-A, 1203-1210, 2000

Kaufman, L. and Rousseeuw, P. *Finding Groups in Data*, Wiley, 1990

Kennedy, R., Lee, Y., Roy, B., Reed, C., and Lippman, R., *Solving Data Mining Problems through Pattern Recognition*, Prentice Hall, 1998

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

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

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

Liu, H. and Motoda, H., *Feature selection for knowledge discovery and data mining*, Kluwer Academic Publishers, Bonston, 1998

Mitra, P., Murthy, C.A., and Pal, S.K., "Unsupervised Feature Selection Using Feature Similarity", *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 24, 301-312, 2002

Osgood, D. Wayne, "Poisson-Based Regression Analysis of Aggregate Crime Rates", *Journal of Quantitative Criminology*, 16, 21-43, 2000

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

Ripley, B., *Pattern Recognition and Neural Networks*, Cambridge University Press, 1996

Power, C.M. and Brown D.E., "Context-based Methods for Track Association", Proceedings of the 5th International Conference on Information Fusion, 1134-1140, 2002

Sarawagi, S., Agrawal, R., and Megiddo. N., "Discovery-driven exploration of OLAP data cubes", *Proc. of the Sixth Int'l Conference on Extending Database Technology (EDBT)*, Valencia, Spain, March 1998.

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

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

Shoshani, A., "OLAP and Statistical Databases: Similarities and Differences", *Proc. ACM PODS '97*, 185-196

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

Welbrock, P.R., Strategic Data Warehousing Principles Using SAS Software, Cary, NC: SAS Institute inc., 1998.