Discovery of Regularized Areas with Maximal Confidence from Location Data

Hiroya Inakoshi
Artificial Intelligence Laboratory, Fujitsu Laboratories Ltd.
inakoshi.hiroya@jp.fujitsu.com

Tatsuya Asai
(affiliation as previous author)
asai.tatsuya@jp.fujitsu.com

Takuya Kida
Graduate School of Information Science and Technology, Hokkaido University.
kida@ist.hokudai.ac.jp

Hiroki Arimura
(affiliation as previous author)
arim@ist.hokudai.ac.jp

keywords: Optimized association rule mining, Location data, GPS

Summary

We propose a new approach to discovering regions optimizing the expected responses from data with a strong spatial bias. The methods available thus far do not work well on data of that nature because they assume that coordinates and responses are uniform and isotropic. To relax this assumption, we employ a hypothesis that cells in an irregularly sized mesh are connected transitively. However, it requires considerable computation and possibly overfits data because there are exponentially many transitive closures. Our contributions to overcome these problems are twofold: we prove the maximal property that shows how irrelevant cells are removed without enumerating candidates in the hypothesis space, and we propose a description length of transitive closure based on which the remaining regions are regularized. We show via experiments that our algorithms do not decrease the precision with unknown data, even when such data are neither uniform nor isotropic. In addition, we show that the regularized region improves the precision by more than 20% compared to the unregularized one.

1. Introduction

Cars, smartphones, and drones are equipped with GPS these days, which helps record detailed traces of persons and things, such as time and coordinates as well as a few measured responses. There are strong needs and demands to take full advantage of GPS data [Derekenaris 01, Mamei 10, Zheng 09, Zheng 10], to make the society more efficient by extracting novel information from it.

Various studies pertaining to knowledge discovery from relatively low-dimensional data have been conducted. In geospatial information and epidemiological literature, the cluster detection test (CDT) has been formalized [Kulldorff 97]. This test is performed to extract connected administrative districts maximizing their log-likelihood, when the baseline and measure count of each district are given [Besag 91, Cressie 93]. In data mining and machine learning literature, optimized association rule mining [Fukuda 96, Rastogi 02], box or patient rule induction [Friedman 99, Hastie 09], and bichromatic discrepancy [Dobkin 96] have been formalized.

These methods are used to extract regions from a hypothesis set, which is a rectangular or x-monotone class of regions, for example, to maximize the expected response, confidence, support, or gain, given a properly fine grid mesh. The left and center images in Figure 1, respectively, show rectangular and x-monotone regions.

Solving an optimization problem in which the object function is estimated from data is difficult because it is hard to estimate said function. Instead, those methods try to directly discover the optimized regions from the data, but some assumptions are required to accomplish this. However, the said methods will not work on GPS data for the following reasons; they assume that the optimized regions are isotropic and axis-aligned rectangular or follow a x-monotone hypothesis; they assume that the coordinates are uniformly distributed in the two-dimensional space. These assumptions are unrealistic for application to GPS data, in which external factors such as geographical or weather conditions dominate, or in scenarios where the mutual effects between people or objects dominate.

In this paper, we propose a method to discover regions optimizing the expected response from data with strong
spatial biases. Our method employs a hypothesis set consisting of transitively connected cells in an irregularly shaped mesh. The picture on the right in Figure 1 shows examples of the regions discovered by the proposed method. By building the mesh adaptively for a given set of data such that its cells have approximately the same numbers of coordinates, we can prevent areas containing a small number of coordinates from being assigned high priority, even if the coordinates are ununiformly distributed. Moreover, because this hypothesis considers a variety of shapes, the proposed method allows us to discover hotspots with non-isotropic responses.

By contrast, relaxing the assumption likely requires much computation, and the discovered regions tend to overfit the data because the number of transitively connected cells can be large. To overcome those problems, our contributions are as follows:

- We prove the maximal property that holds for cells outside of the optimized region and propose a peeling procedure that iteratively removes irrelevant cells without enumerating candidate regions.
- We introduce a description length of connected cells and propose a gradient descent based pasting procedure that prevents the remaining regions from overfitting by employing this length as a regularization term.

The remainder of this paper is organized as follows. After providing a brief review of related works in Chapter 2, we propose the framework of our algorithm and describe its key features along with a few preliminaries in Chapter 3. Next, we explain the algorithms in Chapter 4 and present our experimental results in Chapter 5 in addition to describing application of the algorithm to taxi fleet control. Finally, we present our concluding remarks in Chapter 6.

### 2. Related Works

#### 2.1 Cluster detection test

Cluster detection test (CDT) was developed in the fields of epidemiology. A CDT method usually consists of two phases; first, the region with highest likelihood ratio between the inside and the outside of that region is discovered by means of standard mortality rate represented using binomial or Poisson distributions. Then, a statistical test is performed to confirm whether the data inside and outside of the discovered region follow different probability distributions. Kulldorff, et al. [Kulldorff 97] proposed this framework and employed the circular hypothesis space. Their method has been utilized frequently in epidemiological studies up to the present day. In some cases, however, diseases spread along winds or rivers, and for this reason, a more flexible hypothesis space than the circular one is required. Tango, et al. [Tango 07, Tango 12] proposed a hypothesis set consisting of transitively connected segments in administrative districts. Both Kulldorff and Tango’s methods naively enumerated connected regions, while other studies employed simulated annealing [Duczmal 04, Duczmal 06], upper-level set [Patil 04], or echelon structure [Ishoka 13] for enhancing computational efficiency. These methods, however, consider the empirical average response and consequently execute a statistical test to validate the significance of the discovered region.

#### 2.2 Bump hunting and bichromatic discrepancy

When we are interested in the regions with high response rates, it is easier to obtain them directly rather than estimating the joint probability of variables followed by discovering the regions using the estimated probability. Friedman, et al. [Friedman 99] studied the maximization of averaged response in an axis-parallel rectangular hypothesis from point data. Patient rule induction method
Friedman 95, Hastie 09) iteratively peels either lower or higher end of the variables’ domain one at a time until the averaged response of the remaining rectangular region stops decreasing and then performs pasting to ensure that the rectangular region does not overfit the concrete training data. This procedure is a type of heuristic and does not find the exactly optimized solution.

Dobkin, et al. [Dobkin 90] studied bichromatic discrepancy of the axis-parallel rectangular hypothesis in the context of machine learning. Bichromatic discrepancy is the number of differences between positive and negative examples. They proposed an algorithm to discover the region with the maximum bichromatic discrepancy by employing its monotonous property. Notably, this is equivalent to the minimization of disagreement between true and predictive class by definition, as well as the optimized gain association rule mining. Agrawal, et al. [Agrawal 95] proposed an algorithm to maximize the likelihood ratio between the inside and the outside of a domain by using a technique similar to that proposed by Dobkin.

2.3 Optimized association rule mining

Optimized association rule mining was first introduced by Fukuda, et al. They formulated the discovery of association rules for a single numeric attribute [Fukuda 96] and two-dimensional numeric attributes [Fukuda 01]. The goal is to obtain optimized instantiations in terms of support, confidence, or gain of the rule, given an uninstantiated association rule. A concrete method first aggregates the measurements and baselines for every cell of an equal-sized grid and then discovers the regions with the highest score among a certain hypothesis set such as a rectangle. Their algorithms include one for a $x$-monotone region that maximizes the gain achieved by using dynamic programming with fast matrix search, one for a single rectangular region that maximizes gain, and approximation algorithms for a rectangular region that maximizes support or confidence.

Rastogi and Shim generalized the optimized association rule mining [Rastogi 02] such that it has the disjunctive normal form over either categorical or numeric attributes, and they showed that the problem is NP-hard. They proposed a branch-and-bound algorithm to prune the search space as well as enumerating instantiations of numeric attributes before starting processing so that the identified instantiations do not overlap each other.

Neill, et al. [Neill 04] proposed the branch-and-bound algorithm over an overlap-wrap $k$-dimensional tree to discover exact solutions of extended densities of regions in a two-dimensional plane.

3. Proposed Method

3.1 Preliminaries

Consider a point dataset $S = \{ (x_i, l_i) \mid i = 1, 2, \ldots, n \}$ where $x_i \in \mathbb{R}^2$, $l_i \in \{0, 1\}$, each of which is drawn iid from a probability distribution $p(X, L)$ of random variables $X \in \mathbb{R}^2$ and $L \in \{0, 1\}$. Let $U \subseteq \mathbb{R}^2$ be a two-dimensional region containing all the data. Let $D \subseteq U$ be a region and $\overline{D} = U \setminus D$ be the complement region of $D$.

We need a mesh whose cells have approximately the same number of coordinates even when they scatter uniformly on the region $U$. Let us consider a mesh and its cells as $M = \{ c_j \mid j = 1, 2, \ldots, m \}$ where $U = \bigcup_{j=1}^{m} c_j$ and $c_i \cap c_j = \emptyset$ if $i \neq j$. Note that this definition only requires the cells to be exhaustive and exclusive, and that either grid, polygonal, or quad-tree mesh satisfies the definition.

Obviously, two cells may share their boundary and we call the shared part in the boundary a border. In other words, by stitching borders so that they enclose a single cell, we have a boundary of that cell. We call the border and the point at which two consecutive borders meet an edge and a corer of that cell, respectively.

We can enclose an arbitrary region $D$ on the mesh $M$ by similarly stitching borders. Such closed regions comprise the hypothesis set that we are interested in. By using a relation $R : M^2 \mapsto \{\text{True}, \text{False}\}$ where $R(c_j, c_k) = \text{True}$ iff those two cells share a border, we formally define the hypothesis $H$ such that $D \in H$ iff $R^+(c_j, c_k) = \text{True}$ for all $c_j, c_k \in D$, where $R^+$ indicates the transitive closure of $R$. We denote this hypothesis $H$ as $R^+$, as it is related to the transitive closure.

$D$ is denoted as a column vector $x_D$:

$$x_D = (x_D)_{j=1}^{m}, \text{ where } x_D = \begin{cases} 1 & \text{if } c_j \in D, \\ 0 & \text{otherwise} \end{cases}$$

We also define $x_D \in \mathbb{R}^m$ to introduce the continuous version of $x_D = \sigma(z_D) \in (0, 1)^m$ where $\sigma(\cdot)$ is a sigmoid function. The continuous version is used when we execute the pasting in Section 3.2 to regularize the discovered regions.

Note that either cell $c_j$ or region $D$ includes some points in $S$. The number of points in $c_j$ is called support of the cell and denoted as $s_j = \sup(c_j)$. The number of points in $c_j$ such that $l_j = 1$ is called hit of the cell and is denoted as $h_j = \hit(c_j)$. Similarly, $\sup(D)$ and $\hit(D)$ denote the support and the hit of region $D$. $\sup(c_j)$ and $\hit(c_j)$ are represented by the column vectors, $s = (s_j)_{j=1}^{m}$ and $h = (h_j)_{j=1}^{m}$, respectively.

We briefly introduce Rastogi and Shim [Rastogi 02] and
A rule is of the form \( \text{if } A \rightarrow B \). We will show this as follows. An association rule in the form \( A \rightarrow B \) has \( \text{conf}(A \rightarrow B) \) and \( \text{sup}(A \rightarrow B) \). Our methods discover closed regions each of which is comprised of transitivity connected cells in \( \mathbb{R}^+ \), so that they optimize an object function with a regularization term as introduced in the following section.

### 3.2 Problem description

Using the preliminaries introduced in the previous section, the optimization problem is formalized as follows:

**Problem 1** (Maximization problem of spatial cluster)

Given a point dataset \( S \), a mesh set \( M \), an object function \( f \), and a hypothesis set \( H \), which is a transitive closure of cells, find \( D \in H \) to maximize \( f(D) \).

We employ the expected average response as the object function \( f \), as given below.

\[
E[L \mid X \in D] = \int p(l \mid x \in D) \, dl
\]
\[
= \int p(x \in D, l) \, dl \quad \text{with } p(x \in D) \quad \text{ and } \quad p(x) \text{, from the examples.}
\]
\[
\text{At the fourth equality, we estimate either the joint or marginal probabilities, respectively, } p(x, y) \text{ and } p(x), \text{ from the examples.}
\]

The average response is equivalent to the confidence of an association rule. We will show this as follows. An association rule is of the form \( C_1 \rightarrow C_2 \) if \( C_1(1, 2) \) are logical conditions on \( X \) and \( L \). Their Boolean values are fixed by instantiating \( X \) and \( L \) with some coordinates and a response, although we omit either \( X \) or \( L \) for simplicity. By letting the support of condition, denoted as \( \text{sup}(C_i) \), be the number of records in \( D \) satisfying \( C_i \), the support and confidence of the rule are defined as follows:

\[
\text{sup}(C_1 \rightarrow C_2) = \text{sup}(C_1),
\]
\[
\text{conf}(C_1 \rightarrow C_2) = \frac{\text{sup}(C_1 \cap C_2)}{\text{sup}(C_1)}.
\]

A rule is **confident** if its confidence is not less than a given confidence threshold \( \theta \). A rule is **ample** if its support is not less than a given support threshold \( Z \).

Let \( C_1 \) be \( \text{True} \) for all \( x_i \in D \) and \( C_2 \) be \( \text{True} \) if \( l_i = 1 \). Then, we obtain \( \text{sup}(C_1) = \sum_{x_i \in D} 1 \) and \( \text{sup}(C_1 \cap C_2) = \sum_{x_i \in D} l_i \). Obviously, they are equivalent to the denominator and numerator of (5), respectively. Hence, we interchangeably use the expected average response and confidence.

Let us go revert to the object function. As mentioned earlier, we employ transitive closures, \( \mathbb{R}^+ \), as the hypothesis to discover hotspots from neither uniform nor isotropic data. This requires us to avoid overfitting because the hypothesis is highly complex. Therefore, the object functions should include a regularization term as follows:

\[
f(D) = E[L \mid X \in D] + \lambda ||D||
\]

where \( ||D|| \) is the description length which will be defined in Section 3.3, and \( \lambda \) is a hyper parameter.

Obviously, it is difficult to obtain the optimized solution explicitly. Instead, our algorithms first obtain an initial solution and then perform a gradient descent. In Section 3.3, we introduce the key theorem to find the initial solution quickly, and then, we define the description length concretely in Section 3.4. Thereafter, we explain the gradient descent based algorithms.

### 3.3 Maximal property of regions

In this section, we introduce the maximal property and prove the inequality that holds in the cases of cells satisfying the maximal property.

**[Definition 1]** For given a hypothesis \( H \), a threshold \( Z \), and \( D \in H \), we define \( \widetilde{D} \) as the maximized region with respect to confidence satisfying

\[
\text{sup}(\widetilde{D}) \geq Z \text{ and conf}(D) < \text{conf}(\widetilde{D}) \text{ for any } D \supset \widetilde{D}.
\]

Informally, this means that \( \widetilde{D} \) is maximal if there is no area \( D \) containing \( \widetilde{D} \) with confidence greater than or equal to \( \widetilde{D} \). The following lemma trivially holds:
[Lemma 1] Given a mesh \( M = \{c_i\}_{i=1}^m \) and an association rule \( C_1 \to C_2 \), let \( s_i \) and \( h_i \) be the support and hit of cell \( c_i \), respectively. For any \( E \in H \),

\[
\text{conf}(E) = \frac{\sum_{c_i \in E} h_i}{\sum_{c_i \in E} s_i} \leq \frac{1}{|E|} \sum_{c_i \in E} \frac{h_i}{\min_{c_j \in E} s_j},
\]

where \(|E|\) is the number of cells in \( E \).

The following theorem presents the condition that each cell should satisfy when it is not included in the maximum region. The region remaining after peeling the irrelevant cells by using this inequality obviously never misses the maximum region.

[Theorem 1] For \( \tilde{D}, D \in H \) such that \( \tilde{D} \subset D \), let \( E = D \setminus \tilde{D} \). Then, \( \text{conf}(D) < \text{conf}(\tilde{D}) \) holds if

\[
\frac{h_i}{\min_{c_j \in E} s_j} < \text{conf}(D) \text{ for } \forall c_i \in E.
\]

[Proof] To keep the formulas simple, we introduce some invariants: \( S_D = \sum_{c_i \in D} s_i, H_D = \sum_{c_i \in D} h_i, S_E = \sum_{c_i \in E} s_i \) and \( H_E = \sum_{c_i \in E} h_i \). Since \( D = \tilde{D} \cup E \) and \( \tilde{D} \cap E = \phi \),

\[
S_{\tilde{D}} = S_D - S_E, \\
H_{\tilde{D}} = H_D - H_E.
\]

Now we evaluate \( \text{conf}(\tilde{D}) = \text{conf}(D) \):

\[
\text{conf}(\tilde{D}) = \frac{H_{\tilde{D}}}{S_{\tilde{D}}} = \frac{H_D - H_E}{S_D - S_E} = \frac{S_D (H_D - H_E) - (S_D - S_E) H_D}{S_D S_E} = \frac{S_E S_D H_D - S_D H_E}{S_D} = \frac{S_E}{S_D} \left( \frac{H_D}{S_D} - \frac{H_E}{S_E} \right) = \frac{S_E}{S_D} (\text{conf}(D) - \text{conf}(E)).
\]

By summing (11) over \( c_i \in E \) and combining this with (10), we get

\[
\text{conf}(E) \leq \frac{1}{|E|} \sum_{c_i \in E} \frac{h_i}{\min_{c_j \in E} s_j} < \text{conf}(D).
\]

From (12) and (15), \( \text{conf}(\tilde{D}) > \text{conf}(D) \) is proved. 

This section discusses the regularization term of object function (8) that penalizes the region by its complexity on shape. It is well understood that the VC-dimension of polygons is related to the number of their edges, or equivalently, number of their corners [Mohri 12]. We employ this measure as the regularization term denoted by \( \|D\| \) and explain how to evaluate it.

[Lemma 2] The number of corners of transitive closure \( D \in \mathbb{R}^+ \) is given as follows:

\[
(1 - x_D)^T W x_D \text{ where } W = 2U - V
\]

where \( 1 \) is an \( m \)-dimensional column vector with all elements equaling 1, and \( U = (u_{jk})_{j=1,k=1}^{m,m} \) and \( V = (v_{jk})_{j=1,k=1}^{m,m} \) represent \( m \times m \) matrices, such that

\[
u_{jk} = \begin{cases} 
1 & \text{if } c_j \text{ and } c_k \text{ share their borders}, \\
0 & \text{otherwise}.
\end{cases}
\]

\[
v_{jk} = \begin{cases} 
2 & \text{two corners of } c_j \text{ are on a edge of } c_k, \\
1 & \text{one corner of } c_j \text{ is on a edge of } c_k, \\
0 & \text{otherwise}.
\end{cases}
\]

[Proof] Equation (16) is decomposed as

\[
(1 - x_D)^T W x_D = \sum_{c_j,c_k \in D} (1 - x_{Dj}) w_{jk} x_{Dk}
\]

Moreover,

\[
+ \sum_{c_j,c_k \in D} (1 - x_{Dj}) w_{jk} x_{Dk} + \sum_{c_j,c_k \in D} (1 - x_{Dj}) w_{jk} x_{Dk}.
\]
Equation (15) represents the case in which both k and c_j are in D. In this case, this term is always 0 because 1 - x_{Dj} = 0. Similarly, (20) is always 0 because x_{Dk} = 0. Therefore, we only need to investigate (21), where either c_k or c_j is in D, whereas the others are not. Without loss of generality, we have c_k \in D and c_j \in \overline{D}.

Owing to space constraints, we provide a sketch of the proof. Twice the number of shared edges between c_k and \overline{D} equals the number of corners on the border between c_k and the cells in \overline{D} if no corner of c_k is a corner of D (left column in Figure 2). The former number differs from the later by the number of c_k’s corners that are also those of D as well (right column in Figure 2). With this in mind, note that the j-th element of UXD is the number of the edges that c_j shares with the region D and that accordingly, (1 - x_D)^TWXD gives the number of edges shared between D and \overline{D}. Similarly, note that the k-th element of (1 - x_D)^TVXD is the number of corners on the border between c_k and the region \overline{D} and that accordingly, (1 - x_D)^T VXD gives the number of corners on the border between D and \overline{D}. Consequently, (16) was proved to give the number of corners in D.

The matrices U and V were obtained by employing sweep-line algorithms in a manner similar to that used for evaluating transitive closure from a collection of cells. For detail, refer to [Inakoshi 14].

4. Algorithms

This chapter explains the procedures used to discover the optimized hypothesis (Algorithm 1). The proposed method involves building a quad-tree comprising irregularly sized cells (line 2), peeling the cells irrelevant to the maximal regions (line 3), and pasting cells by running the gradient descent method to regularize the regions (line 4). The notations are the same as those used in Chapter 3. These three processes are detailed in the following sections.

4.1 Data adaptive mesh forming

Lines 5–14 describe the procedure to build a quad-tree. Because the cells are split iteratively into four along the horizontal and vertical borders so long as they have more points than the threshold \alpha (line 12), the cells in the final quad-tree are expected to have roughly the same number of points as that in the fixed mesh of size m. In the case where the points are distributed uniformly in the two-dimensional space, the quad-tree converges to a grid mesh as more points become available. It is easy to obtain h and s by counting the points in cells during quad-tree forma-

| Algorithm 1 Proposed method |
|-------------------------------|
| 1: \textbf{procedure} MAIN(S, U, m, Z) |
| 2: \hspace{1em} M, h, s \leftarrow BUILDQUADTREE(S, U, m) |
| 3: \hspace{1em} G \leftarrow PEELING(M, h, s, Z) |
| 4: \hspace{1em} return \textsc{pasting}(G, h, s) as $\tilde{D}$ |
| 5: \textbf{procedure} BUILDQUADTREE(S, U, m) |
| 6: \hspace{1em} \alpha \leftarrow |S|/m |
| 7: \hspace{1em} initialize queue Q by \{U\}, M by empty set |
| 8: \hspace{1em} while Q is not empty do |
| 9: \hspace{2em} pop cell from Q as U |
| 10: \hspace{2em} if sup(U) > \alpha then |
| 11: \hspace{3em} split U into four cells and push them to Q |
| 12: \hspace{2em} else |
| 13: \hspace{3em} append U to M |
| 14: \hspace{1em} return M as well as its h and s |
| 15: \textbf{procedure} PEELING(M, h, s, Z) |
| 16: \hspace{1em} initialize maximal regions, G, by \{M\} |
| 17: \hspace{1em} repeat |
| 18: \hspace{2em} initialize M by empty set |
| 19: \hspace{2em} for all D \in G do |
| 20: \hspace{3em} peel irrelevant cells from D by using (11) |
| 21: \hspace{3em} add cells in D to M |
| 22: \hspace{2em} initialize G by closures discovered in M |
| 23: \hspace{1em} until sup(M) < Z or no cell peeled at line 20 |
| 24: \hspace{1em} return G |
| 25: \textbf{procedure} PASTING(M, G, h, s) |
| 26: \hspace{1em} initialize W from M |
| 27: \hspace{1em} initialize z_D in accordance with G |
| 28: \hspace{1em} repeat |
| 29: \hspace{2em} z_D \leftarrow z_D - \eta \nabla z_D \cdot f(x_D) |
| 30: \hspace{2em} until z_D converges |
| 31: \hspace{1em} return \{c_j | z_{Dj} > 0\} |

tion (line 14).

4.2 Peeling

Lines 15–24 describe the peeling procedure. All maximal regions G are reported by iteratively removing the irrelevant cells in accordance with Theorem 1 until no peeling occurs at line 20. The process for obtaining new transitive closures (line 23) is implemented efficiently by using the sweep-line algorithm [Hertz 08, Inakoshi 14]. h and s are needed at lines 20 and 23.

4.3 Pasting

Lines 25–31 describe the pasting procedure. First, we obtain the matrices W = 2U - V needed to evaluate the regularization term \|D\| based on their definitions in Section 3.2. Owning to space constraints, the process of ef-
Table 1 Parameters of synthetic data.

| Name  | Values | Description             |
|-------|--------|-------------------------|
| n     | 4096   | #data                   |
| K     | { 1, 2, \cdots, 9 } | #Gaussian components |
| m     | 64 × 64 | #segments               |
| F     | { 0.2, 0.4, 0.6, 0.8, 1.0 } | flatness of components |
| q     | 0.65   | probability to split cells |

ficient implementation by using the sweep-line algorithm is omitted from this paper. Second, we initialize \( x_{Dj} \) a positive number if \( c_j \) is in one of the regions in \( G \), or a negative number otherwise (line 27). Finally, we run the gradient descent method (line 28-40). To make the object function differentiable, we replace \( x_{Dj} \) with its continuous version as explained in Section 3.1. With this trick, the gradient is given as follows:

\[
\nabla_{x_D} f(x_D) = \left( \frac{\partial f}{\partial x_{Dj}} \right)_{j=1}^m (x_{Dj} - \frac{1}{m}) \left( \frac{\partial f}{\partial z_{Dj}} \right)_{j=1}^m
\]

\[
\Delta \left( x_{Dj} + \frac{1}{m} \right) \left( \frac{\partial f}{\partial x_{Dj}} \right)_{j=1}^m
\]

5. Experimental Results

This chapter presents the experimental results showing that the proposed method captures the probability \( p(X,L) \) well and achieves a higher average response by using synthetic data. We compared the method with and without regularization, denoted as MTCR and MTC, respectively. MTC stands for "Maximized Transitive Closure". In addition, we present the results obtained by RASTOGI and FUKUDA. Note that RASTOGI, FUKUDA, and MTC can represent CDT introduced in Section 2.1 because they are supposed to output the similar regions as CDT in the following experimental settings.

5.1 Generation of synthetic data

The synthetic data is used to emulate spatially isotropic or non-isotropic cases by making the coordinate distribution \( p(X) \) uniform or ununiform, as well as by making the response probability \( p(L | X) \) round or squashed, respectively. Table 1 shows the parameters needed to achieve this.

In the uniform case, \( p(X) \) follows an uniform distribution. In the ununiform case, on the contrary, a quad-tree is employed to determine the distribution of coordinates. It is built by letting cells split with the probability \( q^d \), where \( d \) is the depth of quad-tree. \( p(X | c_j) \) and \( p(c_j) \) follow individual uniform distributions. This implies that the cells have the same number of coordinates, either large or small, which makes the coordinates sparse or dense from place to place. Note that the parameter \( q \) was determined such that the average number of cells was \( m \) for ensuring fair comparison between the ununiform and uniform cases.

The response probability \( p(L | X) \) follows mixed Gaussian probability with \( K \) components whose mixture coefficients are equal. Their covariance matrices are elliptic, and the flatness of these components is \( F \), but the angles of these components are determined randomly and differ from each other.

Once the joint probability \( p(X,L) = p(L | X)p(X) \) is fixed, we generate \( n \) coordinates with their responses such that they follow the joint probability using a random process.

5.2 Experimental configurations

Simply evaluating the average responses for the data from which the regions are discovered is not sufficient for demonstrating the aim of the present study. Instead, we must evaluate the expected average responses. To this end, we designed the following experimental configurations: first, we generated 11 independent datasets by using the procedure explained in Section 2.1. Then, by using one of these datasets and the algorithms, we discover the maximal regions with the parameter \( m \), which determines the fineness of the grid or quad-tree mesh these algorithms employ. Finally, we evaluate the precision of discovered regions for the remaining 10 datasets. This test process is repeated on 20 differently parameterized probabilities and the average of these trials is calculated.

5.3 Results

§ 1 Uniform or ununiform distributions

We first compared each algorithm between the uniform and the ununiform cases. The response was fixed to be round (\( F=1.0 \)) to eliminate factors other than uniformity. We investigated both unimodal (\( K=1 \)) and multimodal (\( K=5 \)) cases (upper table in Table 2). With this experiment, we expect that quad-tree meshes benefit in the ununiform cases by taking dense and sparse areas fairly. The precision of MTC and MTCR improved significantly by around 10% in the unimodal cases, and by 1.5%–3.0% in the multimodal cases. Naturally, the improvement was lower in the multimodal cases because the number of coordinates per component was lower than that in the unimodal cases. Although RASTOGI enjoyed its highest precision in the unimodal and uniform cases, MTCR was superior in the ununiform cases and comparable to FUKUDA in the uniform cases.
Table 2  Precision of four algorithms which compare the ununiform with uniform cases (upper) and squashed to round cases (lower). MTC and MTCR are the proposed methods, which are without and with regularization, respectively.

|                | unimodal (K=1) | multimodal (K=5) |
|----------------|----------------|------------------|
|                | MTC | MTCR | FUKUDA | RASTOGI | MTC | MTCR | FUKUDA | RASTOGI |
| uniform        | 25.80 | 27.43 | 28.76 | 34.18 | 7.37 | 8.14 | 8.47 | 6.68 |
| ununiform      | 36.56 | 38.22 | 30.53 | 36.22 | 8.89 | 11.16 | 8.97 | 7.14 |
| lift           | 10.76 | 10.79 | 1.77 | 2.04 | 1.52 | 3.02 | 0.50 | 0.46 |

|                | uniform | MTC | MTCR | FUKUDA | RASTOGI | MTC | MTCR | FUKUDA | RASTOGI |
|----------------|---------|-----|------|--------|--------|-----|------|--------|--------|
| round (F=1.0)  | 25.80   | 27.43 | 28.76 | 34.18 | 36.58 | 38.22 | 30.53 | 36.22 |
| squashed (F=0.2)| 29.16  | 29.01 | 30.41 | 26.97 | 37.70 | 37.64 | 28.64 | 26.44 |
| lift           | 3.36    | 1.58  | 1.65 | -7.21 | 1.12  | -0.58 | -1.89 | -9.78 |

Fig. 3  Precision of MTC and MTCR with K=5 and variable flatness.

§ 2  Round or squashed responses

Next, we compare each algorithm with the squashed and round responses. To eliminate factors other than the flatness, the responses were fixed as unimodal (K=1). We investigated both uniform and ununiform cases (lower table in Table 2). With this experiment, we expect that the hypothesis of transitive closure captures regions with high expected average response even when the response probability is squashed. The precision did not change significantly between the squashed and the round cases with both MTC and MTCR, while it decreased by 7.21%–9.78% when using RASTOGI. Notably, FUKUDA did not degrade because the x-monotone hypothesis can also represent squashed regions.

§ 3  Advantage of regularization

Thus far, our methods, MTC and MTCR, advantageous in the case with non-isotropic distribution.

Finally, we show the improvement resulting from regularization by comparing the algorithms in the case of a non-isotropic distribution with variable flatness (Figure 3). The improvement rate made by MTCR from MTC were 11.6%–23.0%.

5.4  Application using realistic data

Based on the results shown in Figure 4, one feasible application is taxi fleet control based on expected demand. Herein, the taxis regularly report their location every minute or so. The points indicate the locations of the taxis. A point is enabled when a taxi picks up a passenger, otherwise, it is disabled. We believe our method is more effective than the alternatives because the points are distributed ununiformly, as can be seen in the figure. Knowing the maximal regions would increase the probability of taxis picking up passengers in the discovered regions because there should be greater demand for taxis and fewer taxis available. Our method helps model the spatial patterns of demands on weekly or specified period bases.

6. Conclusion

We proposed a new approach to maximize the expected average response. In this method, we introduced a regularization term that penalizes the complex regions and relaxes the assumptions related to coordinate distribution. These enhancements made our method adaptive to observed data while improving its predictive performance. The experiments showed that our algorithms did not decrease the precision in the case of unseen data, even when it was neither uniform nor isotropic. In addition, they showed that the regularization improved precision by more than 20%.

In the future, this works can be continued in two directions: one involves extending it to higher dimensional
data. The other is theoretical analysis. Both the quad-tree and the regularization term are currently apriori. They need justifications based on the theory of statistical machine learning.

◇ References ◇

[Agrawal 06] Agrawal, D., Phillips, J. M., and Venkatasubramanian, S.: The hunting of the bump: on maximizing statistical discrepancy, in SODA '06 Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithm, pp. 1137–1146, Miami, Florida, Association for Computing Machinery (2006)

[Berg 08] Berg, de M., Chong, O., Kreveld, van M., and Overmars, M.: Computational Geometry - Algorithms and Applications, pp. 147–171, 3rd edition, Springer (2008)

[Besag 91] Besag, J. and Newell, J.: The detection of clusters in rare diseases, Journal of the Royal Statistical Society, Series A (Statistics in Society), Vol. 154, No. 1, p. 143 (1991)

[Cressie 93] Cressie, N. A. C.: Statistics for Spatial Data, Wiley Series in Probability and Statistics, pp. 381–452, John Wiley & Sons, Inc., Hoboken, NJ, USA (1993)

[Derekenaris 01] Derekenaris, G., Garofalakis, J., Makris, C., Prentzas, J., Siotas, S., and Tsakalidis, A.: Integrating GIS, GPS and GSM technologies for the effective management of ambulances, Computers, Environment and Urban Systems, Vol. 25, No. 3, pp. 267–278 (2001)

[Dobkin 96] Dobkin, D. P., Gunopulos, D., and Maass, W.: Computing the maximum bichromatic discrepancy, with applications to computer graphics and machine learning, Journal of Computer and System Sciences, Vol. 52, No. 3, pp. 453–470 (1996)

[Duczmal 04] Duczmal, L. and Assunção, R.: A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters, Computational Statistics & Data Analysis, Vol. 45, No. 2, pp. 269–286 (2004)

[Duczmal 06] Duczmal, L., Kulldorff, M., and Huang, L.: Evaluation of spatial scan statistics for irregularly shaped clusters, Journal of Computational and Graphical Statistics, Vol. 15, No. 2, pp. 428–442 (2006)

[Friedman 99] Friedman, J. H. and Fisher, N. I.: Bump hunting in computational statistics with a restricted likelihood ratio for detecting disease clusters, Statistics in Medicine, Vol. 31, No. 30, pp. 4207–4218 (2012)

[Zheng 09] Zheng, Y., Zhang, L., Xie, X., and Ma, W.-Y.: Mining interesting locations and travel sequences from GPS trajectories, in Proceedings of the 19th International Conference on World Wide Web - WWW '09, p. 791, New York, New York, USA, ACM Press (2009)

[Zheng 10] Zheng, V. W., Zheng, Y., Xie, X., and Yang, Q.: Collaborative location and activity recommendations with GPS history data, in Proceedings of the 19th International Conference on World Wide Web - WWW ’10, p. 1029, New York, New York, USA, ACM Press (2010)

(担当委員：馬場 雪乃)

Received May 22, 2018.
Author’s Profile

Hiroya Inakoshi
received the B.S. and M.E. degrees in Engineering from Kyoto University in 1993 and from Graduate School of Engineering in Kyoto University 1995, respectively. He is currently engaged in the research on data mining and machine learning with Fujitsu Laboratories Ltd. He is a member of IEICE and IPSJ.

Tatsuya Asai (Member)
received the M.E. degree from Nagoya University in 2001 and received Dr. (Information Science) degree from Kyushu University in 2004. He was engaged in research on XML database systems with Fujitsu Laboratories Ltd. from 2004. His current research interests include discovery science. He is a member of IPSJ.

Takuya Kida
received the B.S. degree in Physics, the M.S. and Dr. (Information Science) degree all from Kyushu University in 1997, 1999, and 2001, respectively. He was a full-time lecturer of Kyushu University Library from October 2001 to March 2004. He is currently an associate professor of the division of computer science, Graduate School of Information Science and Technology, Hokkaido University, since April 2004. His research interests include text algorithms and data structures, information retrieval, and data compression. He is a member of IEICE, IPSJ, and DBSJ.

Hiroki Arimura (Member)
received the B.S degree in 1988 in Physics, the M.S. and the Dr.Sci. degrees in 1990 and 1994 in Information Systems from Kyushu University. From 1990 to 1996, he was at the Department of Artificial Intelligence in Kyushu Institute of Technology, and from 1996 to 2004, he was at the Department of Informatics in Kyushu University. Since 2006, he has been a professor of the Graduate School of Information Science and Technology, Hokkaido University. His research interests include data mining, computational learning theory, information retrieval, and algorithm design. He is a member of IPSJ and ACM.