高次元ピラミッド構築問題とデータマイニングへの応用

Danny Z. Chen[1], 全 眞嬉 [2], 加藤直樹 [3], 徳山 豪 [2]
[1] University of Notre Dame, Dept. of Computer Science and Engineering **chen@cse.nd.edu**
[2] 東北大学大学院情報科学研究科システム情報科学専攻 **(jinhee,tokuyama)@dais.is.tohoku.ac.jp**
[3] 京都大学大学院工学研究科建築工学専攻 **naoki@archi.kyoto-u.ac.jp**

概要 最適領域ルールは福田らによって提案されている。[9, 10] は数値属性データで構成されるデータベースに対する有効なデータマイニングツールである。一方、上記の方法では2つの欠点がある：（1）各ルールは高々2変数の数値属性に対応でき（2）与えられたデータの正確な位置ではなく、単純に領域 R の中にあるか、外にあるかだけに基づき判断される。本論文では、これらの欠点を取り除くための新しい方法の提案をする。グラフアルゴリズムの用いて、2つ以上の属性を持つ最適数値属性結合ルールと階層構造の数値属性結合ルールを与える。又、本論文のメソッドは異常なデータの除去とデータクラスタリングに適切である。

Higher-Dimensional Pyramid Construction Problem and Application to Data Mining

Danny Z. Chen[1], Jinhee Chun[2], Naoki Katoh[3], Takeshi Tokuyama[2]

[1] Dept. of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA.
chen@cse.nd.edu
[2] GSIS, Tohoku University, Sendai, Japan. **(jinhee,tokuyama)@dais.is.tohoku.ac.jp**
[3] Graduate School of Engineering, Kyoto University, Kyoto, Japan. **naoki@archi.kyoto-u.ac.jp**

Abstract. Optimized region rules developed by Fukuda et al. [9, 10] are effective tools for data mining in databases with numeric data. However, there are two drawbacks in the previous methods: (1) each rule can contain at most two numeric conditional attributes, and (2) the decision is made based only on whether a given data is inside or outside a region R, but not on the exact position of the data. In this paper, we propose a new method for removing these drawbacks. Indeed, by applying graph algorithms, we give optimized numeric association rules with more than two attributes, and give layered-structure numeric association rules. Our method is also applicable to removal of exceptional data and data clustering.

1 Introduction

Association rules are useful for determining correlations between attributes of a relation. Association rules, introduced in Agrawal-Imielinski-Swami[1], provide a useful mechanism for discovering correlations among the underlying data and have applications in business marketing. The form of a general association rule is $C_1 \to C_2$. This association rule can be viewed as being defined over attributes of a relation, where C_1 and C_2 are conjunctions of conditions.

Efficient generation of effective association rules is a major opic in data mining. In this paper, we are interested in association rules on numeric data. For treating numeric attributes, one possible method is to discretize each numeric attribute and transform it to a Boolean attribute or a categorical attribute without numeric information; however, the transformation often causes loss of information of the original data.

We consider a d-tuple data $\mathbf{x}$ ($\mathbf{x}$ consists of d numeric attributes) as a point in a d-dimensional space and construct a distribution in a voxel grid showing confidence/support information as seen in the top-left picture in Fig. 1. Then, we consider a rule $(\mathbf{x} \in R) \to (C = yes)$ for a suitable region R in the d-dimensional space. We call such a rule a *region rule*. Here, the target attribute C can be numeric (usually with certain monotonicity condition), for which the right-hand side is $X(C) = f$, where $X(C)$ indicates a random variable associated with C and f is a probabilistic distribution [14].

2 Problem formulation

Consider a d-tuple with numeric attributes $A_1, A_2, \ldots, A_d$. For each $k = 1, 2, \ldots, d$, the domain of A_k is divided into m_k buckets $\boldsymbol{b}_1^k, \boldsymbol{b}_2^k, \ldots, \boldsymbol{b}_{m_k}^k$, such that the A_k-values of data are distributed into these buckets. We indeed apply the equi-depth-bucketing algorithm [9] to create buckets such that the number of data whose A_k-values fall in $\boldsymbol{b}_j^k$ is almost N/m_k for each $1 \le j \le m_k$, where N is the total number of data d-tuples. We consider the Cartesian product of bucketings, and create a voxel grid Γ containing $n = m_1 \times m_2 \times \cdots \times m_d$ cells, in which the cell c_I indexed by $I = (i_1, i_2, \ldots, i_d)$ corresponds to the Cartesian product $\boldsymbol{b}_{i_1}^1 \times \boldsymbol{b}_{i_2}^2 \times \cdots \times \boldsymbol{b}_{i_d}^d$.

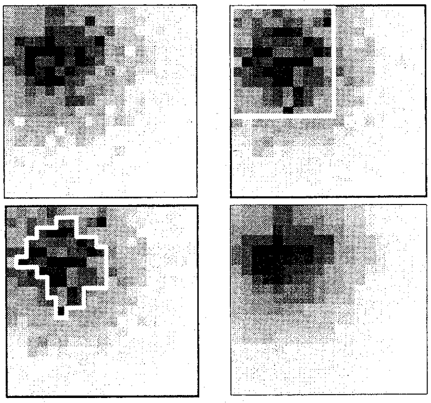


Figure 1: Region rules and a layered structure

For simplicity of presentation, we assume from now on that $m_k = m$ for $k = 1, 2, \ldots, d$ and $n = m^d$. This restriction can be easily removed.

We fix a target attribute C, which we assume to be Boolean for simplicity. We can also consider the case where C is a numeric attribute analogously to [14]. For each cell $c \in \Gamma$, let $\mu(c)$ be the number of data x satisfying that $A_k(x) \in \boldsymbol{b}^k_{i_k}$ for $k = 1, 2, \ldots, d$ if c is indexed by $I = (i_1, i_2, \ldots, i_d)$. Let $\rho(c)$ be the number of data x satisfying that $C(x) = yes$ and $A_k(x) \in \boldsymbol{b}^k_{i_k}$ for $k = 1, 2, \ldots, d$.

The generation of such ρ and μ has been discussed by Fukuda et al. in [10]. Hence, we assume that we have such ρ and μ on the voxel grid Γ.

For a region R consisting of cells of Γ, let $\mu(R) = \sum_{c \in R} \mu(c)$ and $\rho(R) = \sum_{c \in R} \rho(c)$. We call $\rho(R)$ and $conf(R) = \rho(R)/\mu(R)$ the *support* and *confidence* of the rule $(\mathbf{x} \in R) \rightarrow (C = yes)$, respectively. We also call $\mu(R)$ the *antecedent support.*[1]

Introducing a nonnegative parameter t,

$$g_t(R, \rho, \mu) = \rho(R) - t \cdot \mu(R)$$

is called the *parametric gain* of ρ against μ within the region R.

We consider a family $\mathcal{F}$ of regions in Γ, and let $R^{opt}(t)$ be a region in $\mathcal{F}$ maximizing $g_t(R, \rho, \mu)$. Naturally, the support $\rho(R^{opt}(t))$ is a nonincreasing function in t, while the confidence $conf(R^{opt}(t))$ is is a nondecreasing function in t.

In [10], for the family of x-monotone regions in a two-dimensional grid, an algorithm for computing t such that the confidence $conf(R^{opt}(t))$ is maximized under the condition that the support $\rho(R^{opt}(t))$ is above a given threshold value θ is given. Also, we can compute t together with $R^{opt}(t)$ maximizing the support under the condition that the confidence is above a given threshold. The corresponding rules are called *maximum confidence rule* and *maximum support rule*, respectively. The same problem on the family of rectilinear convex regions in a two-dimensional grid has been considered in [18]. Moreover, it can be shown that the optimal subdivision maximizing a convex objective function (e.g., entropy or GINI index) is obtained as $R^{opt}(t)$ for a suitable t, and such a t can be efficiently computed for the above mentioned families of regions. Morimoto et al. gave a construction of accurate decision trees by using the region rules maximizing GINI index [15].

A layered structure of the data (we often call it *pyramid structure* or *pyramid*) with respect to a region family $\mathcal{F}$ is a series of regions $\mathcal{P} = P_0 \supset P_1 \supset \cdots \supset P_h \supset P_{h+1} = \emptyset$ together with an increasing series of nonnegative real numbers $t_0 < t_1 < \cdots < t_h$ called *heights* satisfying that $conf(P_i \setminus P_{i+1}) = t_i$ for $i = 0, 1, \ldots, h$. The regions P_i $(i = 0, 1, \ldots, h)$ are called *flats* of the pyramid. The optimal pyramid is the pyramid structure approximating the data distribution such that the standard squared L_2 measurement of the error (defined later) is minimized.

Since $\rho(R^{opt}(t))$ is nonincreasing in t, one might expect that $R^{opt}(t) \subseteq R^{opt}(t')$ for $t > t'$. If this were true, we would have proved (in Section 2.1) that the series $R^{opt}(t_0), R^{opt}(t_1), \ldots, R^{opt}(t_h)$ and heights $t_0 < t_1 < \cdots < t_h$ for the sorted set of all transition values t_i of $R^{opt}(t)$ give the optimal pyramid structure.

[1] In [9, 10], this is called "support", and $\rho(R)$ is called "hit".

Unfortunately, this is not always true for each of the families of x-monotone regions and rectilinear convex regions.

In this paper, we investigate the conditions under which the optimal pyramid can be efficiently constructed. Especially, we define the following region families: (1) stabbed-union of orthogonal regions, (2) generalized base-monotone regions, and (3) digitized star-shaped regions. Surprisingly, for each of these region families, the optimal pyramid can be computed in polynomial time for any fixed d dimensions, by reducing the problem to the minimum s-t cut problem in a directed graph. Moreover, we can flexibly control the size of the family of regions by using a graph associated with the grid.

3 The optimal pyramid problem

For the sake of convenience, we consider an abstract situation that μ and ρ are arbitrary nonnegative distribution functions on the d-dimensional voxel grid Γ of $n = m^d$ cells such that $\mu(c)$ and $\rho(c)$ are integers satisfying $\mu(c) \geq \rho(c)$ for every cell $c \in \Gamma$ and $\mu(\Gamma) \leq N$. We call $\rho(R)/\mu(R)$ the confidence of R abusing the convention when ρ and μ correspond to the support and antecedent support in a database.

We fix a family $\mathcal{F}$ of regions in Γ. Without loss of generality, we assume that $\emptyset \in \mathcal{F}$ and $\Gamma \in \mathcal{F}$.

Definition 1 *Consider a series $\mathcal{P}$ of regions $P(t_i)$ $(i = 0, 1, 2, \dots, h)$ in $\mathcal{F}$ associated with a series of increasing nonnegative numbers (called heights) $t_0 < t_1 < t_2 < \cdots < t_h$ satisfying that $P(t_0) = \Gamma$ and $P(t) \subseteq P(t')$ for $t > t'$. $\mathcal{P}$ is said to be a* pyramid *(or* pyramid structure*) approximating ρ with respect to μ if $conf(P(t_i) - P(t_{i+1})) = t_i$, with $P(t_{h+1}) = \emptyset$.*

The approximation error of $\mathcal{P}$ is defined by

$$\sum_{i=0}^{h} \sum_{c \in P(t_i) - P(t_{i+1})} (conf(c) - t_i)^2 \mu(c).$$

This is the squared L_2 distance between the confidence ρ/μ and the surface function f_P of the pyramid defined by $f_P(x) = t_i$ if $x \in P(t_i) - P(t_{i+1})$, considering μ as the density function. A pyramid $\mathcal{P}$ is *optimal* if it has the minimum approximation error among all pyramids.

The problem is very intuitive if μ is a constant function. In this case, the optimal pyramid can be considered as a unimodal reformation of ρ/μ minimizing the loss of positional potential. Although this is a basic problem in computational geometry and geography (especially for $d = 2$), this specialized problem has not been theoretically investigated before. In general, the pyramid can be considered as a unimodal approximation of ρ relative to μ. In Figure 2, we give an example of reforming a function ρ to a pyramid (where $\mu \equiv 1$) for $d = 1$.

Constructing an optimal pyramid (in two or higher dimensions) is a natural extension of the problem of region segmentation, and will be useful in several applications besides data mining (e.g., statistics, geomorphology, and computer vision [4]).

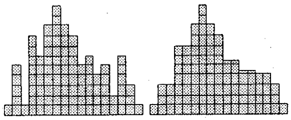


Figure 2: A reformation of ρ (the left) to a pyramid (the right) for $d = 1$

3.1 Optimal pyramid and parametric gain

Lemma 3.1 (See [6]) *Consider a function $P(t)$ from $(0, \infty)$ to $\mathcal{F}$ satisfying that $P(t) \subseteq P(t')$ for $t > t'$ and maximizing the objective function*

$$J(P) = \int_{t=0}^{\infty} \rho(P(t)) - t \cdot \mu(P(t)) dt.$$

Then, it has at most $n + 1$ transition values $0 = t_{-1} < t_0 < t_1 < \cdots < t_h$ satisfying that $P(t) = P(t_i)$ for $t \in (t_{i-1}, t_i]$ and $P(t) = \emptyset$ for $t > t_h$. Moreover, $\mathcal{P}$ consisting of $P(t_0), P(t_1), \dots, P(t_h)$ is the optimal pyramid.

Thus, intuitively, the optimal pyramid $\mathcal{P}$ is obtained by piling up horizontal sections $P(t)$ with as large parametric gains as possible. Consider the region $R^{opt}(t)$ in $\mathcal{F}$ maximizing the parametric gain $g_t(R, \rho, \mu)$. Intuitively, if t increases, $R^{opt}(t)$ is shrunk. If $\{R^{opt}(t)\}$ (precisely speaking, the transitions of the function $R^{opt}(t)$) forms a pyramid, it is obviously the optimal pyramid. Unfortunately, if we pile up the maximum gain regions $R^{opt}(t)$, they do not always form a pyramid, since $R^{opt}(t) \subset R^{opt}(t')$ does not always hold for $t > t'$. This makes it very difficult to compute the optimal pyramid.

3.2 Closed family

We now consider region families for which the relation $R^{opt}(t) \subset R^{opt}(t')$ holds for $t > t'$, and utilize them for designing layered region rules.

A (discrete) family $\mathcal{F}$ of regions in $\mathcal{R}^d$ is called a *closed family* if it is closed under the intersection and union operations, that is, $R \cap R' \in \mathcal{F}$ and $R \cup R' \in \mathcal{F}$ for any two regions R, R' in $\mathcal{F}$.

Proposition 3.2 (See [6]) *Given a closed family $\mathcal{F}$, let $R^{opt}(t)$ be the region in $\mathcal{F}$ maximizing $g_t(R, \rho, \mu)$. If there are multiple regions in $\mathcal{F}$ maximizing $g_t(R, \rho, \mu)$, we take any one which is minimal under inclusion. Then, the series of transitions of $R^{opt}(t)$ gives the optimal pyramid $\mathcal{P}$ for $\mathcal{F}$.*

Thus, if a family $\mathcal{F}$ is constructed by using closed families as its building blocks, we can hope to design an efficient algorithm for computing the optimal pyramid for $\mathcal{F}$.

4 Higher-dimensional Problem

In a higher dimensional case ($d \geq 2$), the time complexity largely depends on the specific family $\mathcal{F}$ of regions. Indeed, it is not difficult to see that the problem of computing just a single flat of the optimal pyramid is NP-hard for some families even for $d = 2$ [3].

4.1 Family with a small number of regions

If $\mathcal{F}$ has M different regions, the optimal pyramid can always be computed in polynomial time in M and n for the d-dimensional case. We construct a directed acyclic graph $H(\mathcal{F}) = (\mathcal{F}, E)$ whose vertex set is $\mathcal{F}$. For each pair R and R' of $\mathcal{F}$, we give a directed edge $e = (R, R')$ if and only if $R \supset R'$. We compute $t(e)$ such that $\rho(R \setminus R') = t(e) \cdot \mu(R \setminus R')$. The value $t(e)$ is called the *height label* of e, and $r(e) = t^2(e)\rho(R \setminus R')/2$ is called the *profit* of e. A directed path $p = e_0, e_1, \dots, e_q$ is called *admissible* if $t(e_{i-1}) < t(e_i)$ for $i = 1, 2, \dots, q$. The profit of an admissible directed path is the sum of profit values of the edges on it.

Lemma 4.1 *The optimal pyramid is associated with the admissible path with the maximum profit in $H(\mathcal{F})$, such that $R \setminus R'$ is a flat of the pyramid if and only if (R, R') is an edge on that path.*

Thus, we can reduce the optimal pyramid problem to a maximum-weight-path problem in the directed acyclic graph $H(\mathcal{F})$. Note that each directed path in $H(\mathcal{F})$ has at most n edges. By using a dynamic programming algorithm, we obtain the following result:

Theorem 4.2 *The optimal pyramid for $\mathcal{F}$ of M different regions can be computed in $O(M^2n)$ time.*

Unfortunately, the above algorithm is seldom practical. For example, the family of rectangular regions has $O(n^2)$ regions, and hence the above time complexity is $O(n^5)$. Moreover, for computing accurate layered region rules, we want to consider families in which M is exponentially large in n.

Thus, we seek more efficient algorithms for some special families of regions.

4.2 Stabbed unions of orthogonal regions

We consider a typical closed family of regions in Γ. For a fixed cell c of Γ, a region R in Γ is called a *stabbed union* of orthogonal regions at c (in short, a stabbed union region) if R can be represented as the union of orthogonal regions each of which contains c. The cell c is called the center cell of R. Fig. 3 gives an example of the two-dimensional case.

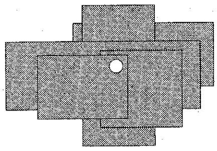


Figure 3: A stabbed union of rectangles

It is clear that the family of all stabbed unions at a cell c is a closed family; in fact, it is the closure of the family of all rectangles containing c. The pyramid given in Fig. 1 is based on the family of stabbed unions at a point. Naturally, the center cell (or point) gives the peak of the pyramid. To design an algorithm for computing the optimal pyramid, we need an efficient algorithm for computing the maximum parametric gain region in the family of stabbed unions. For this purpose, we generalize the problem and apply graph-theoretic methods.

4.3 Domination-closures in a grid graph

In the voxel grid Γ, we fix a center cell $c = (c_1, c_2, \ldots, c_d)$, and define a directed graph $G(c)$, whose vertex set consists of all voxels in Γ. For voxels $p = (p_1, p_2, \ldots, p_d)$ and $q = (q_1, q_2, \ldots, q_d)$, the L_1 distance between p and q is $dist(p,q) = \sum_{i=1}^{d} |p_i - q_i|$. The neighboring cells of p are the cells whose distance from p is one. For a cell p and its neighbor q, a directed edge is defined: its direction is (p,q) (i.e., from p to q) if $dist(p,c) = dist(q,c) + 1$ and otherwise (q,p). The graph $G(c)$ is a weakly-connected directed graph with $d(m-1) \times m^{d-1} = O(n)$ edges (we assume d is a constant), and c is its unique sink vertex (i.e., a vertex without outgoing edges).

A subgraph $H = (V,E)$ of $G(c)$ is called a *rooted subgraph* if there exists a directed path in H from each vertex v of V to c.

Given a rooted subgraph $H = (V,E)$ of $G(c)$, we say that a vertex u is *H-dominated* by another vertex v if there exists a directed path from v to u in H. An *H-domination closure* W is a subset of V satisfying the condition that every Clearly, each H-domination closure defines a connected region containing the cell c in Γ, and we like to identify such regions. Given a rooted subgraph H of $G(c)$, we consider the family $\mathcal{F}_H$ that is the set of all H-domination closures. Since the domination closure property is closed under union and intersection, we have the following proposition:

Proposition 4.3 *For a rooted subgraph H of $G(c)$, $\mathcal{F}_H$ is a closed family of regions.*

The following lemmas are straightforward from the definitions of $G(c)$ and $\mathcal{F}_H$.

Lemma 4.4 *A region R in Γ is a stabbed union at c if and only if it is a $G(c)$-domination closure.*

Lemma 4.5 *Let H and H' be two spanning rooted subgraphs of $G(c)$. Then, if H is a subgraph of H',*

4.3.1 Algorithms for computing the optimal pyramid with respect to $\mathcal{F}_H$

Let us fix a rooted subgraph H of $G(c)$. We consider a parameter value t defining the height of a flat of the optimal pyramid with respect to $\mathcal{F}_H$, and we give a weight $\rho(p) - t \cdot \mu(p)$ to each voxel p (and its corresponding vertex in H). Due to the following lemma, we can assume that t is a rational number with small denominator and numerator.

Lemma 4.6 *If t is a height defining a flat of the optimal pyramid for a certain region family, then t is a rational number represented by a quotient of two integers less than or equal to N.*

Proof: Suppose that t defines a flat P_i. Then, $\rho(P_i \setminus P_{i+1}) = t \cdot \mu(P_i \setminus P_{i+1})$. Since ρ and μ take integer values at most N, we have the lemma. □

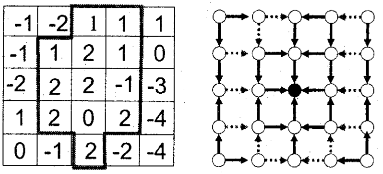
By definition, the maximum (parametric) gain region $P^{opt}(t) \in \mathcal{F}_H$ is the H-domination closure maximizing the sum of weights of voxels in the region. In graph-theoretic terminology, this is the *maximum domination closure* of the weighted directed graph H, and is obtained as the connected component of H containing c by removing a set of edges (the *cut set*). See Fig. 4 for a region $P^{opt}(t)$ in $\mathcal{F}_H$ (here, $H = G(c)$ and $d = 2$) and the corresponding cut set; the number in each pixel p of the left picture is the weight $\rho(p) - t \cdot \mu(p)$.

The following theorem is due to Hochbaum [12], and it was applied by Wu and Chen [17] to a geometric segmentation problem in a setting different from this paper.

Theorem 4.7 (Hochbaum [12], See [6]) *Given a directed graph G with n vertices having real-valued vertex weights and m edges, its maximum domination closure can be computed in $O(T(n,m))$ time, where $T(n,m)$ is the time for computing a minimum s-t cut in an n-vertex, m-edge directed graph with nonnegative edge weights.*

Theorem 4.8 (See [6]) *The optimal pyramid for the family $\mathcal{F}_H$ can be computed in $O(n^{1.5} \log n \log^2 N)$ time.*

Corollary 4.9 *The optimal pyramid for the family of stabbed unions of orthogonal regions at a given cell c can be computed in $O(n^{1.5} \log n \log^2 N)$ time.*



-1	-2	1	1	1
-1	1	2	1	0
-2	2	2	-1	-3
1	2	0	2	-4
0	-1	2	-2	-4

Figure 4: An optimal region and its corresponding cut set (dashed arrows) in $G(c)$

We note that we have a more efficient $O(n \log n)$ time algorithm for the family of stabbed unions of rectangles at a grid point p for the two-dimensional case [6]; we omit it in this paper, since it does not work for a higher-dimensional case. We have a more efficient algorithm if H is a tree.

Theorem 4.10 (See [6]) *If H is a tree, the optimal pyramid for the family $\mathcal{F}_H$ can be computed in $O(n \log N)$ time.*

4.3.2 Further remarks

In the association rule generation, m is usually small (up to 100 and typically about 20 in demonstrations of the SONAR system [8]) and d is seldom larger than 4, since very fine mesh is not necessary for learning a rule from a database. Thus, $n = m^d$ is not very large (typically, in the range of 400 – 160,000), while the database size N may be huge. Fortunately, the database size only affects the time complexity by some $\log N$ factors. Moreover, each $\log N$ factor can be replaced by $\log \gamma + \log n$ if we discretize the values μ and ρ into integers in $[0, \gamma]$ for a suitable $\gamma \ll N$ to compute an approximation solution in a computation time independent of the database size N (assuming ρ and μ are given).

The algorithm given in Theorem 4.8 assumes that we know the peak c of the optimal pyramid. If we do not know the peak, a naive method is to examine all possible n cells of Γ as the candidates of the peak of the optimal pyramid, and report the one with the maximum objective function value. This method takes $O(n^{2.5} \log n \log^2 N)$ time. However, we can practically guess the peak by using some method to reduce the number of candidates of the peak to $n^\delta \ll n$. This improves the running time to $O(n^{1.5+\delta} \log n \log^2 N)$. We give a typical trick in [5].

4.4 Based monotone regions

In the two dimensional case, the family of based monotone regions is one of the region families adopted in the SONAR data mining system [10, 15] in order to represent two-dimensional association rules.

A based monotone region is a connected region R in Γ such that the intersection of R with each column of Γ is either empty or a continuous subcolumn whose lowest edge lies on the lower boundary of Γ; i.e., R is a connected region such that for each fixed i, $\{(i,j) \in R\} = \{(i,j) : j \leq f(i)\}$ for some $f(i) \in [0, n]$, where i is the column index and j is the row index (counting from bottom to the top). In other words, the connected region is defined by a function $y \leq f(x)$ in the digital (x, y)-plane.

Consider the family of based monotone regions containing a fixed cell $c = (c_1, 1)$ on the bottom row. It is easy to see that this family is a closed family.

We generalize this family as follows. We fix a cell $c \in \Gamma$ in the d-dimensional voxel grid, and consider the graph $G(c)$ (defined in Section 4.3). For each vertex $p = (p_1, p_2, \ldots, p_d) \neq c$, we find the largest index i such that $p_i \neq c_i$, and select the outgoing edge of p corresponding to its i-th coordinate. Thus, we obtain a spanning tree $T_0(c)$ of $G(c)$, which we call *lexicographic minimum bend* spanning tree. In the voxel grid, the path from p to c in $T_0(c)$ consists of at most d segments.

It is easy to see that a region R is a based monotone region containing $c = (c_1, 1)$ in a two-dimensional pixel grid if and only if it is a domination closure in $T_0(c)$. This gives a d-dimensional analogue of the family of based monotone regions. The following theorem is obvious from Theorem 4.10.

Theorem 4.11 *The optimal pyramid with respect to the family of domination closures of $T_0(c)$ can be computed in $O(n \log N)$ time.*

If the peak c is not given, we can apply the strategy for guessing the peak discussed in Section 4.3.2.

It is possible to consider other ways of extending the family of based monotone regions to higher dimensions. For example, suppose that we add all adjacent edges of $G(c)$ for each vertex p satisfying that $p_d = c_d$ to $T_0(c)$ to obtain a new graph H. Then every region R corresponding to a domination closure of H is monotone with respect to each coordinate axis except x_d. In other words, the projection of R to the plane Z defined by $x_d = c_d$ is a stabbed union in the $d-1$ dimensional grid, and the intersection of R with a vertical column defined by $x_i = p_i$ $(i = 1, 2, \ldots, d-1)$ is either empty or a segment penetrating Z.

4.5 Digitized star-shaped regions

In Section 4.3, we presented an algorithm for the optimal pyramid in the region family $\mathcal{F}_H$, and, especially, showed that the graph $G(c)$ itself defines the family of stabbed unions of orthogonal regions.

In Euclidean geometry spaces (i.e., not the voxel grid settings), a popular closed family of regions is the family of star-shaped regions centered at a given point q. Recall that a region R is said to be *star-shaped* centered at q if for any point $v \in R$, the line segment $\overline{qv}$ is entirely contained in R [13]. For example, a convex region containing q is a star-shaped region centered at q; more precisely, the family of star-shaped regions is the closure (with respect to the union and intersection operations) of the family of convex regions containing q. Thus the family of star-shaped regions is a quite rich family.

Hence, we would like to define a digital analogue of the star-shaped regions. However, this is not a well-defined problem in general, since it depends on the definition of the digital line segment between two voxels. Unfortunately, it does not form a closed family if we adopt one of the popular definitions of digital line segments. A typical definition is to let the set of all voxels intersecting a (true) line segment form the corresponding digital line segment; however, the intersection of two such digital line segments through a fixed cell c (c contains the center point q) is not always connected, and thus cannot be represented as the union of suitable digital line segments through c.

Here, we give a probabilistic construction of the digital star-shaped regions by using a modified spanning tree of $G(c)$. Let c be the center of the digital star-shaped regions. The first idea is to construct a spanning tree T of $G(c)$ rooted at c such that for each vertex v (i.e., a voxel of Γ) of T, the unique directed path $path_T(v, c)$ in T from v to c simulates a (true) line segment $\overline{vc}$ in the digital space.

Each vertex of $G(c)$ has at most d outgoing edges. We obtain a spanning tree of $G(c)$ by selecting exactly one outgoing edge for each vertex (except c). Consider a vertex corresponding to the voxel $p = (p_1, p_2, \ldots, p_d)$. Let $dist(p) = \sum_{i=1}^{d} |p_i - c_i|$ be the L_1 distance from p to c. Note that $dist(p)$ is also the distance from p to c in T. Our strategy for constructing T is to select the edge corresponding to the i-th coordinate with the probability $|p_i - c_i|/dist(p)$.

Let $x_i(k, p)$ be the number of edges corresponding to the i-th coordinate encountered during walking on the path $path_T(p, c)$ from p by k steps. Since T is constructed in a randomized fashion, $x_i(k, p)$ is a random variable.

Lemma 4.12 *The expectation $E = E(x_i(k,p))$ is $k|p_i - c_i|/dist(p)$. Moreover, the probability that $|x_i(k,p) - E| \geq \sqrt{2ak}$ is smaller than $2e^{-a}$ for any positive number a, where e is the natural logarithm base.*

Proof: Since the process is Martingale, the expectation can be obtained by an induction on k. The deviation bound is obtained from Azuma's inequality [16]. □

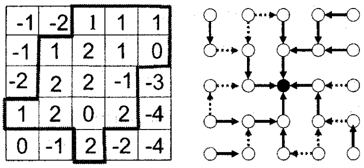
This implies that $path_T(p, c)$ approximates the line segment $\overline{pc}$ for each vertex p. Since T is a tree, the optimal pyramid for the family of T-domination closures can be computed in $O(n \log N)$ time by Theorem 4.10. However, the set of T-domination closures thus defined has some serious defects in simulating the star-shaped regions, even in the two dimensional case. Note that, geometrically, a T-domination closure is not always a simply connected region in Γ. This is because there are some leaf vertices in T that do not correspond to any boundary cells (i.e., cells touching the grid boundary) of Γ.

For this purpose, for any leaf vertex v of T corresponding to an internal cell of Γ, we need to add an incoming edge to v in order to obtain a new graph $\tilde{T}$. Again, we choose one of the incoming edges for v in a probabilistic fashion. A simple calculation shows that the expected number of additional edges needed is $n/6 + o(n)$ if $d = 2$ and bounded by $n/3$ for $d \geq 3$. We then have a graph $\tilde{T}$ as constructed above, and treat the set of domination closures in the graph $\tilde{T}$ as the family of *randomized digitized star-shaped regions* with the center c. See Fig. 5 for an example, and compare it with Fig. 4.

Corollary 4.13 *The optimal pyramid for the family of a randomized digitized star-shaped regions with the center c can be computed in $O(n^{1.5} \log n \log^2 N)$ time.*

5 Concluding remarks

The pyramid structure is useful in several ways. First, it gives a nice visualization of the tendency of the distribution of confidence. Second, by selecting the part of the pyramid above a threshold height t (we call this operation *clipping*), we can generate a region giving an association rule together with information



-1	-2	1	1	1
-1	1	2	1	0
-2	2	2	-1	-3
1	2	0	2	-4
0	-1	2	-2	-4

Figure 5: An optimal region and its corresponding cut set in $\tilde{T}$ (only a part near c is displayed)

on the strength of influence of the rule on each data depending on its actual geometric position inside the region. The height t is determined by using a support threshold or some optimization criteria such as the GINI maximization [15]. Third, once we obtain a region rule, we can easily remove the influence of this rule by subtracting the pyramid from the original data distribution and refilling the same antecedent support with an average confidence. Then, we can further search for a weaker rule, or can consider pyramids with different peaks simultaneously to extract more than one layered rule from the data. This automatically gives a clustering that covers a majority of data by clipping a high part from each of such pyramids. Fourth, it potentially can be used to clean data since a data item which is an outlier in the pyramid approximation for every possible peak can be considered as an "exceptional data" (possibly some input error) or a "confused data".

The selection of a suitable region family is a very important problem. The digitized star-shaped regions seem to be a reasonable family for $d = 2$. For $d \geq 3$, a family corresponding to a graph with more edges may be useful. We will do experiments in the future research to identify good parameters for the numbers of outgoing and incoming edges of each vertex for $d \geq 3$ (especially, $d = 3$) that help define a nice family for our data mining applications.

References

[1] R. Agrawal, T. Imielinski, A. Swami, Mining Association Rules between Sets of Items in Large Databases, *Proc. SIGMOD* (1993) 207–216.

[2] A. Amir, R. Kashi, N. S. Netanyalm, Analyzing Quantitative Databases: Image Is Everything, *27th Proc. VLDB Conference* (2001).

[3] T. Asano, D. Chen, N. Katoh, and T. Tokuyama, Efficient Algorithms for Optimization-Based Image Segmentation, *Int'l J. of Computational Geometry and Applications* **11**(2001) 145-166.

[4] I. Bloch, Spatial Relationship between Objects and Fuzzy Objects using Mathematical Morphology, in *Geometry, Morphology and Computational Imaging*, 11th Dagsthul Workshop on Theoretical Foundations of Computer Vision, April 2002.

[5] D. Chen,J. Chun, N. Katoh, and T. Tokuyama, Layered Data Segmentation for Numeric Data Minig, *Presented at Submitted.*

[6] J. Chun, N. Katoh, and T. Tokuyama, How to Reform a Terrain into a Pyramid, *Presented at DIMACS Workshop on Geometric Graph Theory* (2002).

[7] S. Dasgupta, Learning Mixtures of Gaussians *Proc. 40th IEEE FOCS* (1999), pp. 634–644.

[8] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama, SONAR: System for Optimized Numeric Association Rules, *Proc. SIGMOD 1996* (1996) p.553.

[9] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama, Mining Optimized Association Rules for Numeric Attributes, *Journal of Computer and System Sciences* **58** (1999) 1-12.

[10] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama, Data Mining with Optimized Two-Dimensional Association Rules, *ACM Transaction of Database Systems* **26** (2001) 179-213.

[11] A. Goldberg and S. Rao, Beyond the Flow Decomposition Barrier, *Proc. 38th IEEE FOCS* (1997) 2–11.

[12] D. S. Hochbaum, A New-old Algorithm for Minimum Cuts in Closure Graphs, *Networks* **37** (2001) 171-193.

[13] D. T. Lee and F. P. Preparata, An Optimal Algorithm for Finding the Kernel of a Polygon, *Journal of the ACM* **26** (1979) 415-421.

[14] Y. Morimoto, H. Ishii and S. Morishita, Construction of Regression Trees with Range and Region Splitting, *The 23rd VLDB Conference* (1997) 166-175.

[15] Y. Morimoto, T. Fukuda, S. Morishita, and T. Tokuyama, Implementation and Evaluation of Decision Trees with Range and Region Splitting, *Constraints* (1997) 402-427 (a preliminary version in VLDB'96).

[16] R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, 1995.

[17] X. Wu and D. Z. Chen, Optimal Net Surface Problems with Applications, *Proc. 29th International Colloquium on Automata, Languages and Programming* (2002) 1029-1042.

[18] K. Yoda, T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama, Computing Optimized Rectilinear Regions for Association Rules, *Proc. KDD97* (1997) 96-103.