Многомерный поиск в обучающих системах

Б. Д. Аширматов

В данной работе исследуется задача многомерной близости в базе данных на специфической модели данных. Вводится модель данных, исследуются ее свойства. Модель обобщается на непрерывный случай, исследуются также и ее свойства. На основе свойств моделей предложен и реализован алгоритм поиска в базе данных.

Ключевые слова: многомерный поиск, обучающие системы, модель данных, алгоритм поиска, задача ближайшего соседа, база данных, многомерная близость, свойства, моделирование, разбиение пространства.

1. Введение

В работе исследуется задача многомерного поиска. Дано множество точек из многомерного дискретного пространства $Z_{m+1}^n = \{0, 1, \ldots, m\}^n$. Имеется точка-запрос, в соответствии с которой необходимо выдать все точки, отстающие от данной не более чем на фиксированное расстояние в манхэттенской метрике, либо выдать, что таких точек нет. Так звучит задача в ее классической постановке.

Помимо классической постановки существуют другие варианты постановки задачи:

– выдать ближайшую точку

– выдать несколько ближайших точек

– выдать хотя бы одну точку из ближайших точек

В данной работе рассматривается задача многомерного поиска для специфичной модели данных — модели «объект — свойство».

Предполагается, что каждый объект характеризуется определенными свойствами. С течением времени возрастает конкретное свойство объекта. Предполагается, что у всех объектов одинаковые начальные свойства и свойства объектов не уменьшаются. База данных объектов хранит копию каждого объекта на каждом конкретном этапе. К базе данных объектов подается объект-запрос, в соответствии с которым нужно выдать либо один из ближайших объектов, либо сообщить, что ближайших объектов нет.

В работе используется подход, предложенный П. А. Алисейчиком: для исследуемой модели данных изучить распределение записей в многомерном кубе и для этого распределения построить алгоритм, являющийся оптимальным в среднем.

Автор выражает глубокую благодарность научному руководителю к.ф.-м.н., старшему научному сотруднику кафедры МаТИС Алисейчику П. А. за постановку задачи и помощь в работе.

2. Модель данных

Предположим, что каждый объект имеет k свойств. Далее предположим, что каждое свойство характеризуется определенным целым числом из интервала $[0, l]$. Будем считать, что свойства объекта возрастают поочередно, причем на определенную величину.

Представим свойства объекта при помощи набора $x = (x_1, x_2, \ldots, x_k)$ с k координатами с целыми значениями от 0 до l, то есть поставим в соответствие точку из пространства $Z_{l+1}^k = \{0, 1, \ldots, l\}^k$.

Введем меру похожести двух объектов — манхэттенское расстояние между соответствующими наборами свойств из пространства Z_{l+1}^k.

Определение 1. *Кривой* изменения свойств объекта будем называть возрастающую последовательность точек $z = (z_0, z_1, \ldots, z_p)$ такую, что

$$z_{i-1} < z_i, |z_{i-1} - z_i| = 1, z_{i-1} \in Z_{l+1}^k, z_i \in Z_{l+1}^k, \quad i = 1, 2, \ldots, p.$$

Будем считать, что начальные свойства у всех объектов совпадают. Представим начальные свойства точкой $O = (0, 0, \ldots, 0)$, а максимальные свойства точкой $L = (l, l, \ldots, l)$.

Введем понятия зависимых и независимых свойств, а также понятие сжатия двух свойств.

Определение 2. Два свойства будем называть *независимыми*, если значения по этим двум свойствам объекта могут изменяться произвольным образом.

Определение 3. Два свойства будем называть *зависимыми*, если значения по этим двум свойствам объекта могут изменяться единственным образом.

Определение 4. Будем говорить, что два свойства со значениями от 0 до l *могут быть сжаты в одно*, если существует функционал $T : \{0, 1, \ldots, l\}^2 \to R$, такой что $\forall x_1, x_2 \in S \subseteq \{0, 1, \ldots, l\}^2$

$$|x_1 - x_2| = |T(x_1) - T(x_2)|,$$

где S — множество всевозможных пар значений по этим двум свойствам.

Утверждение 1. *Никакие два независимых свойства не могут быть сжаты в одно.*

Доказательство. Пусть два свойства независимы и значения по ним могут быть от 0 до l. Предположим, что они могут быть сжаты в одно. Тогда по определению независимых свойств существует как минимум две кривые $z^1 = (z^1_0, \ldots, z^1_{2l})$ и $z^2 = (z^2_0, \ldots, z^2_{2l})$ из O в L в пространстве Z^2_{l+1}.

Поскольку $z^1_0 = z^2_0 = O$ и $z^1_{2l} = z^2_{2l} = L$, найдем такие точки z^1_k и z^2_k, в которых кривые различаются на k-ом шаге. Такие точки существуют, поскольку тогда бы кривые совпадали.

Пусть тогда $T(O) = x$. Тогда $T(L) = x + 2l$ или $T(L) = x - 2l$. Рассмотрим случай $T(L) = x + 2l$ (случай $T(L) = x - 2l$ аналогичен). Очевидно, что $T(z^1_k)$ и $T(z^2_k)$ находятся в отрезке $[T(0), T(L)]$. Тогда поскольку $|T(O) - T(z^1_k)| = |T(O) - T(z^2_k)|$, получаем, что $T(z^1_k) = T(z^2_k)$, а мы предположили обратное, противоречие. Утверждение доказано.

Замечание 1. Очевидно, что поскольку никакие два независимых свойства не могут быть сжаты в одно, то и никакие k свойств не могут быть сжаты в одно, если какие-то два из них независимы.

Замечание 2. Очевидно, что любые два и более зависимых свойств могут быть сжаты в одно: поскольку существует единственная кривая изменения свойств, будем ставить каждой точке z ее номер в кривой, а именно сумму ее координат. Так мы переведем пространство $Z_{l+1}^k = \{0, 1, \ldots, l\}^k$ в пространство $Z_{kl+1}^1 = \{0, 1, \ldots, kl\}$.

Таким образом, после сжатия n блоков по p зависимых свойств в n свойств вместо пространства $Z_{l+1}^k = \{0, 1, \ldots, l\}^k$ получим пространство $Z_{m+1}^n = \{0, 1, \ldots, m\}^n$, где $m = pl$.

3. Особенности модели данных

Мы будем моделировать наши данные в пространстве Z_{m+1}^n при помощи случайной величины. Поскольку невозможно заранее определить, какие кривые будут наиболее вероятны, а какие нет, мы будем считать, что все кривые равновероятны.

Определение 5. *Весом* точки $x = (x_1, x_2, \ldots, x_n)$ будем называть сумму её координат. Обозначать будем $||x||$.

$$||x|| = \sum_{i=1}^{n} x_i.$$

Подсчитаем число возможных кривых из точки $O = (0, 0, \ldots, 0)$ в точку $x = (x_1, x_2, \ldots, x_n)$. Будем обозначать это число $C(O, x)$.

$$C(O, x) = \binom{||x||}{x_1}\binom{||x|| - x_1}{x_2}\ldots\binom{x_n}{x_n} = \binom{||x||}{x_1\ \ldots\ x_n}.$$

Тогда общее число возможных кривых из точки O в точку $M = (m, m, \ldots, m)$ равно $C(O, M)$. Поскольку все кривые равновероятны, тогда вероятность появления каждой кривой равна $1/C(O, M)$.

Подсчитаем число возможных кривых из точки O в точку M, проходящих через точку $x = (x_1, x_2, \ldots, x_n)$. Будем обозначать это число $C_x(O, M)$.

$$C_x(O, M) = C(O, x)C(x, M) = C(O, x)C(O, M - x).$$

Определение 6. *Характеристикой* точки $x = (x_1, x_2, \ldots, x_n)$ будем называть вероятность прохождения кривой через точку x. Обозначать характеристику точки x будем $P(x)$.

$$P(x) = \frac{C_x(O, M)}{C(O, M)}.$$

Определение 7. *Слоем* будем называть множество точек, сумма координат которых совпадает. Будем обозначать слой точек с суммой координат s G_s.

$$G_s = \{x = (x_1, x_2, \ldots, x_n) : ||x|| = s\}.$$

Очевидно, что слои G_s, $s = 0, 1, \ldots, mn$ задают полное разбиение точек пространства $Z^n_{m+1} = \{0, 1, \ldots, m\}^n$ и не пересекаются друг с другом.

Определение 8. *Центром* слоя G_s будем называть точку g_s

$$g_s = (\left\lfloor \frac{s}{n} \right\rfloor, \left\lfloor \frac{s}{n} \right\rfloor, \ldots, \left\lfloor \frac{s}{n} \right\rfloor).$$

Очевидно, что центр слоя g_s принадлежит слою G_s, когда s делится на n.

Заметим, что соседние точки для точки $x = (x_1, x_2, \ldots, x_n)$, $x \in G_s$ (соседними будем называть те, которые находятся на расстоянии 2) и принадлежащие G_s, могут находиться как на том же расстоянии от центра слоя G_s, так и на большем или меньшем расстоянии.

Определение 9. Будем говорить, что точка $y = (y_1, y_2, \ldots, y_n) \in G_s$ получена из точки $x = (x_1, x_2, \ldots, x_n) \in G_s$ *путем удаления относительно центра слоя*, если $y_a = x_a + 1$, $y_b = x_b - 1$, $y_i = x_i$, $i = 1, 2, \ldots, n$, $i \neq a$, $i \neq b$, где $x_a = \left\lfloor \frac{s}{n} \right\rfloor + k$, $x_b = \left\lfloor \frac{s}{n} \right\rfloor - l$, $k \geqslant 0$,

$l \geqslant 0$. Множество точек, полученных из x путем удаления относительно центра слоя будем обозначать $U_s(x)$. Фактически мы определили множество соседних точек для x на слое, находящихся на большем расстоянии от центра слоя, чем x.

Утверждение 2. $\forall G_s,\ s = 0, 1, \ldots, mn,\ \forall x \in G_s\ u\ \forall y \in U_s(x)$ $(U_s(x) \neq \varnothing)$

$$P(x) > P(y) \tag{1}$$

Доказательство. Докажем для произвольных точек x и y, $x \in G_s$, $y \in G_s$, $y \in U_s(x)$. Пусть без ограничения общности $x = (x_1, x_2, x_3, \ldots, x_n)$, $x_1 = \left\lfloor \frac{s}{n} \right\rfloor + k$, $x_2 = \left\lfloor \frac{s}{n} \right\rfloor - l$, $k \geqslant 0$, $l \geqslant 0$, $y = (x_1 + 1, x_2 - 1, x_3, \ldots, x_n)$. Тогда $y = (\left\lfloor \frac{s}{n} \right\rfloor + k + 1, \left\lfloor \frac{s}{n} \right\rfloor - l - 1, x_3, \ldots, x_n)$.

Тогда

$$\frac{P(x)}{P(y)} = \frac{\dfrac{s!(mn-s)!}{x_1!x_2!\ldots x_n!(m-x_1)!(m-x_2)!\ldots(m-x_n)!}}{\dfrac{s!(mn-s)!}{(x_1+1)!(x_2-1)!\ldots x_n!(m-x_1-1)!(m-x_2+1)!\ldots(m-x_n)!}} =$$

$$= \frac{(x_1+1)(m-x_2+1)}{x_2(m-x_1)} = \left(1 + \frac{k+l+1}{\left\lfloor \frac{s}{n} \right\rfloor - l}\right)\left(1 + \frac{k+l+1}{m - \left\lfloor \frac{s}{n} \right\rfloor - k}\right) > 1 \tag{2}$$

Умножив обе части неравенства (2) на $P(y)$, получим (1). Утверждение доказано.

Определение 10. *Окружностью радиуса* p на слое G_s назовем множество точек из G_s, находящихся от центра слоя на расстоянии p. Будем обозначать $O_{s,p}$.

$$O_{s,p} = \{x = (x_1, \ldots, x_n) : \sum_{i=1}^{n} |x_i - g_{s_i}| = p, x \in G_s\}.$$

Утверждение 3. *При* $p_1 < p_2$, $O_{s,p_1} \neq \varnothing$, $O_{s,p_2} \neq \varnothing$

$$\min_{x \in O_{s,p_1}} P(x) > \min_{x \in O_{s,p_2}} P(x),$$

$$\max_{x \in O_{s,p_1}} P(x) > \max_{x \in O_{s,p_2}} P(x).$$

Доказательство. Найдем для произвольного p $\min P(x)$ и $\max P(x)$, $x \in O_{s,p}$.

Рассмотрим следующие две точки x' и x'', $x' \in O_{s,p}$, $x'' \in O_{s,p}$, $0 \leqslant j < k < i$

1)
$$x' = (\left\lfloor \frac{s}{n} \right\rfloor + j, \left\lfloor \frac{s}{n} \right\rfloor + i, x_3, \ldots, x_n),$$
$$x'' = (\left\lfloor \frac{s}{n} \right\rfloor + j + 1, \left\lfloor \frac{s}{n} \right\rfloor + i - 1, x_3, \ldots, x_n).$$

Тогда $\dfrac{P(x')}{P(x'')} =$

$$= \frac{\dfrac{s!(mn-s)!}{(\lfloor \frac{s}{n} \rfloor + j)!(\lfloor \frac{s}{n} \rfloor + i)! \ldots (m - \lfloor \frac{s}{n} \rfloor - j)!(m - \lfloor \frac{s}{n} \rfloor - i)! \ldots}}{\dfrac{s!(mn-s)!}{(\lfloor \frac{s}{n} \rfloor + j + 1)!(\lfloor \frac{s}{n} \rfloor + i - 1)! \ldots (m - \lfloor \frac{s}{n} \rfloor - j - 1)!(m - \lfloor \frac{s}{n} \rfloor - i + 1)! \ldots}} =$$
$$= \frac{(\lfloor \frac{s}{n} \rfloor + j + 1)(m - \lfloor \frac{s}{n} \rfloor - i + 1)}{(\lfloor \frac{s}{n} \rfloor + i)(m - \lfloor \frac{s}{n} \rfloor - j)} =$$
$$= \left(1 + \frac{j + 1 - i}{\lfloor \frac{s}{n} \rfloor + i}\right)\left(1 + \frac{j + 1 - i}{m - \lfloor \frac{s}{n} \rfloor - j}\right) < 1. \quad (3)$$

2)
$$x' = (\left\lfloor \frac{s}{n} \right\rfloor - j, \left\lfloor \frac{s}{n} \right\rfloor - i, x_3, \ldots, x_n),$$
$$x'' = (\left\lfloor \frac{s}{n} \right\rfloor - j - 1, \left\lfloor \frac{s}{n} \right\rfloor - i + 1, x_3, \ldots, x_n).$$

Тогда $\dfrac{P(x')}{P(x'')} =$

$$= \frac{\dfrac{s!(mn-s)!}{(\lfloor \frac{s}{n} \rfloor - j)!(\lfloor \frac{s}{n} \rfloor - i)! \ldots (m - \lfloor \frac{s}{n} \rfloor + j)!(m - \lfloor \frac{s}{n} \rfloor + i)! \ldots}}{\dfrac{s!(mn-s)!}{(\lfloor \frac{s}{n} \rfloor - j - 1)!(\lfloor \frac{s}{n} \rfloor - i + 1)! \ldots (m - \lfloor \frac{s}{n} \rfloor + j + 1)!(m - \lfloor \frac{s}{n} \rfloor + i - 1)! \ldots}} =$$
$$= \frac{(\lfloor \frac{s}{n} \rfloor - j - 1)(m - \lfloor \frac{s}{n} \rfloor - i + 1)}{(\lfloor \frac{s}{n} \rfloor + i)(m - \lfloor \frac{s}{n} \rfloor - j)} =$$

$$= \left(1 + \frac{j+1-i}{\left\lfloor \frac{s}{n} \right\rfloor - j}\right)\left(1 + \frac{j+1-i}{m - \left\lfloor \frac{s}{n} \right\rfloor + i}\right) < 1. \qquad (4)$$

Таким образом, из (3) и (4) мы получили $\min P(x)$ и $\max P(x)$, $x \in O_{s,p}$.

Заметим, что если точки x' и x'' таковы, что

$$P(x') = \min_{x \in O_{s,p_1}} P(x), P(x'') = \min_{x \in O_{s,p_2}} P(x),$$

то $x'' \in U_s(x')$ и в силу утверждения 3 $P(x') > P(x'')$.

Аналогично получаем, что если точки x' и x'' таковы, что

$$P(x') = \max_{x \in O_{s,p_1}} P(x), P(x'') = \max_{x \in O_{s,p_2}} P(x),$$

то либо $x'' \in U_s(x')$, либо $x'' \in U_s(y)$, $P(x') > P(y)$, и в силу утверждения 3 $P(x') > P(x'')$. Утверждение доказано.

Определение 11. *Кривой* изменения свойств *с определенным количеством петель* t будем называть неубывающую последовательность точек $z = (z_0, z_1, \ldots, z_k)$

$$z_{i-1} <= z_i, |z_{i-1} - z_i| \leqslant 1, z_{i-1} \in Z^n_{m+1}, z_i \in Z^n_{m+1}, \;\; i = 1, 2, \ldots, k,$$

причем $z_{i-1} = z_i$ ровно в t случаях.

Определение 12. *Кривой* изменения свойств *с ограниченным количеством петель* t будем называть неубывающую последовательность точек $z = (z_0, z_1, \ldots, z_k)$

$$z_{i-1} <= z_i, |z_{i-1} - z_i| \leqslant 1, z_{i-1} \in Z^n_{m+1}, z_i \in Z^n_{m+1}, \;\; i = 1, 2, \ldots, k,$$

причем $z_{i-1} = z_i$ не более чем в t случаях.

Таким образом, дав возможность объекту не изменять свои свойства не более (ровно) t раз, мы делаем нашу модель «объект — свойство» более приближенной к реальности.

Утверждение 4. *Характеристики точек совпадают для определения кривой, кривой с определенным количеством петель и кривой с ограниченным числом петель.*

Доказательство. Покажем, что характеристики точек для определения кривой и кривой с определенным количеством петель совпадают.

Подсчитаем число возможных кривых с определенным количеством петель t из точки $O = (0, 0, \ldots, 0)$ в точку $x = (x_1, x_2, \ldots, x_n)$. Будем обозначать это число $C'(O, x, t)$.

$$C'(O, x, t) = C(O, x)\binom{||x|| + t}{t}.$$

Тогда общее число возможных кривых с определенным количеством петель t из точки O в точку M равно $C'(O, M, t) = C(O, M)\binom{||M|| + t}{t}$. Поскольку все кривые опять же равновероятны, тогда вероятность появления каждой кривой равна $1/C'(O, M, t)$.

Подсчитаем число возможных кривых с определенным количеством петель t из точки O в точку M, проходящих через точку $x = (x_1, x_2, \ldots, x_n)$. Будем обозначать это число $C'_x(O, M, t)$.

$$C'_x(O, M, t) =$$
$$= C(O, x)C(O, M - x)\sum_{i=0}^{t}\binom{||x|| + i}{i}\binom{||M|| - ||x|| + t - i}{t - i} =$$
$$= C(O, x)C(O, M - x)\binom{||M|| + t}{t}.$$

Тогда характеристика точки для кривых с определенным количеством t петель совпадает с характеристикой точки для обычных кривых

$$P(x) = \frac{C'_x(O, M, t)}{C'(O, M, t)} = \frac{C_x(O, M)}{C(O, M)}.$$

Теперь покажем, что характеристика точки для для определения кривой и кривой с ограниченным количеством t петель совпадают.

Подсчитаем число возможных кривых с ограниченным количеством петель t из точки $O = (0, 0, \ldots, 0)$ в точку $x = (x_1, x_2, \ldots, x_n)$. Будем обозначать это число $C''(O, x)$.

$$C''(O, x, t) = C(O, x)\sum_{i=0}^{t}\binom{||x|| + i}{i}.$$

Тогда общее число возможных кривых с ограниченным количеством петель t из точки O в точку M равно $C''(O,M,t) = C(O,M)\sum_{i=0}^{t}\binom{||M||+i}{i}$. Поскольку все кривые снова равновероятны, тогда вероятность появления каждой кривой равна $1/C''(O,M,t)$.

Подсчитаем число возможных кривых с ограниченным количеством петель t из точки O в точку M, проходящих через точку $x = (x_1, x_2, \ldots, x_n)$. Будем обозначать это число $C''_x(O,M,t)$.

$$C''_x(O,M,t) =$$

$$= C(O,x)C(O,M-x)\sum_{i=0}^{t}\binom{||x||+i}{i}\sum_{j=0}^{t-i}\binom{||M||-||x||+j}{j} =$$

$$= C(O,x)C(O,M-x)\sum_{i=0}^{t}\sum_{j=0}^{i}\binom{||x||+j}{j}\binom{||M||-||x||+i-j}{i-j} =$$

$$= C(O,x)C(O,M-x)\sum_{i=0}^{t}\binom{||M||+i}{i}.$$

Тогда характеристика точки для кривых с ограниченным количеством петель t совпадает с характеристикой точки для обычных кривых

$$P(x) = \frac{C''_x(O,M,t)}{C''(O,M,t)} = \frac{C_x(O,M)}{C(O,M)}.$$

Утверждение доказано.

4. Обобщение модели данных

Обобщим модель, описанную в главе 2. Предположим, что каждый объект имеет k свойств, причем каждое свойство характеризуется определенным вещественным из интервала $[0,1]$.

Тогда свойствам объекта поставим в соответствие точку $x = (x_1, x_2, \ldots, x_k)$ из пространства $Z^k_{[0,1]} = [0,1]^k$.

Аналогично вводится понятие кривой изменения свойств (без ограничения на расстояние между соседними точками), независимых и зависимых, понятие сжатия свойств в одно.

Для обобщенной модели также справедливы утверждение 1, замечания 1 и 2.

Определение 13. *Характеристикой* точки $x = (x_1, x_2, \ldots, x_n)$ будем называть сумму следующих двух вероятностей:

а) вероятность события, что $y = (y_1, y_2, \ldots, y_n) \in [0,1]^n$, $y <= x$,

б) вероятность события, что $y = (y_1, y_2, \ldots, y_n) \in [0,1]^n$, $y >= x$,

$$P(x) = \prod_{i=1}^{n} x_i + \prod_{i=1}^{n} (1 - x_i).$$

Аналогично вводится понятие слоя.

Определение 14. *Центром* слоя G_s будем называть точку g_s

$$g_s = (\frac{s}{n}, \frac{s}{n}, \ldots, \frac{s}{n}).$$

Определение 15. Будем говорить, что точка $y = (y_1, y_2, \ldots, y_n) \in G_s$ получена из точки $x = (x_1, x_2, \ldots, x_n) \in G_s$ *путем удаления относительно центра слоя*, если $y_a = x_a + \varepsilon$, $y_b = x_b - \varepsilon$, $y_i = x_i$, $i = 1, 2, \ldots, n$, $i \neq a$, $i \neq b$, где $x_a = \frac{s}{n} + \delta_1$, $x_b = \frac{s}{n} - \delta_2$, $\delta_1 \geqslant 0$, $\delta_2 \geqslant 0$, $\varepsilon > 0$.

Утверждения 5 и 6 являются аналогами для обобщенной модели утверждений 2 и 3.

Утверждение 5. $\forall G_s$, $s = 0, 1, \ldots, mn$, $\forall x \in G_s$ *и* $\forall y \in U_s(x)$ $(U_s(x) \neq \varnothing)$

$$P(x) > P(y).$$

Доказательство. Докажем для произвольных точек x и y, $x \in G_s$, $y \in G_s$, $y \in U_s(x)$. Пусть без ограничения общности $x = (x_1, x_2, x_3, \ldots, x_n)$, $x_1 = \frac{s}{n} + \delta_1$, $x_2 = \frac{s}{n} - \delta_2$, $\delta_2 \geqslant 0$, $\delta_2 \geqslant 0$, $y = (x_1 + \varepsilon, x_2 - \varepsilon, x_3, \ldots, x_n)$, $\varepsilon > 0$. Тогда $y = (\frac{s}{n} + \delta_1 + \varepsilon, \frac{s}{n} - \delta_2 - \varepsilon, x_3, \ldots, x_n)$.

Тогда $P(y) =$

$$= (\frac{s}{n}+\delta_1+\varepsilon)(\frac{s}{n}-\delta_2-\varepsilon)\prod_{i=3}^{n} x_i+(1-\frac{s}{n}-\delta_1-\varepsilon)(1-\frac{s}{n}+\delta_2+\varepsilon)\prod_{i=3}^{n}(1-x_i) =$$

$$= P(x) + (\varepsilon(\frac{s}{n} - \delta_2) - \varepsilon(\frac{s}{n} + \delta_1) - \varepsilon^2)\prod_{i=3}^{n} x_i +$$

$$+ (-\varepsilon(1 - \frac{s}{n} + \delta_2) + \varepsilon(1 - \frac{s}{n} - \delta_1) - \varepsilon^2)\prod_{i=3}^{n}(1 - x_i) =$$

$$= P(x) - \varepsilon(\delta_1 + \delta_2 + \varepsilon)\prod_{i=3}^{n} x_i - \varepsilon(\delta_1 + \delta_2 + \varepsilon)\prod_{i=3}^{n}(1 - x_i) < P(x).$$

Утверждение доказано.

Утверждение 6. *При* $p_1 < p_2$, $O_{s,p_1} \neq \varnothing$, $O_{s,p_2} \neq \varnothing$

$$\min_{x\in O_{s,p_1}} P(x) > \min_{x\in O_{s,p_2}} P(x),$$

$$\max_{x\in O_{s,p_1}} P(x) > \max_{x\in O_{s,p_2}} P(x).$$

Доказательство. Найдем для произвольного p $\min P(x)$ и $\max P(x)$, $x \in O_{s,p}$.

Рассмотрим следующие две точки x' и x'', $x' \in O_{s,p}$, $x'' \in O_{s,p}$, $0 \leqslant \delta_1 < \delta_2$, $\varepsilon \leqslant \frac{\delta_2-\delta_1}{2}$

1) $x' = (\frac{s}{n}+\delta_1, \frac{s}{n}+\delta_2, x_3, \ldots, x_n)$, $x'' = (\frac{s}{n}+\delta_1+\varepsilon, \frac{s}{n}+\delta_2-\varepsilon, x_3, \ldots, x_n)$.

Тогда $P(x'') =$

$$= (\frac{s}{n}+\delta_1+\varepsilon)(\frac{s}{n}+\delta_2-\varepsilon)\prod_{i=3}^{n} x_i+(1-\frac{s}{n}-\delta_1-\varepsilon)(1-\frac{s}{n}-\delta_2+\varepsilon)\prod_{i=3}^{n}(1-x_i) =$$

$$= P(x') + (\varepsilon(\frac{s}{n} + \delta_2) - \varepsilon(\frac{s}{n} + \delta_1) - \varepsilon^2)\prod_{i=3}^{n} x_i +$$

$$+ (\varepsilon(1 - \frac{s}{n} - \delta_1) - \varepsilon(1 - \frac{s}{n} - \delta_2) - \varepsilon^2) \prod_{i=3}^{n}(1 - x_i) =$$

$$= P(x') + \varepsilon(\delta_2 - \delta_1 - \varepsilon) \prod_{i=3}^{n} x_i + \varepsilon(\delta_2 - \delta_1 - \varepsilon) \prod_{i=3}^{n}(1 - x_i) > P(x'). \quad (5)$$

2) $x' = (\frac{s}{n} - \delta_1, \frac{s}{n} - \delta_2, x_3, \ldots, x_n)$, $x'' = (\frac{s}{n} - \delta_1 - \varepsilon, \frac{s}{n} - \delta_2 + \varepsilon, x_3, \ldots, x_n)$.

Тогда $P(x'') =$

$$= (\frac{s}{n} - \delta_1 - \varepsilon)(\frac{s}{n} - \delta_2 + \varepsilon) \prod_{i=3}^{n} x_i + (1 - \frac{s}{n} + \delta_1 + \varepsilon)(1 - \frac{s}{n} + \delta_2 - \varepsilon) \prod_{i=3}^{n}(1 - x_i) =$$

$$= P(x') + (-\varepsilon(\frac{s}{n} - \delta_2) + \varepsilon(\frac{s}{n} - \delta_1) - \varepsilon^2) \prod_{i=3}^{n} x_i +$$

$$+ (-\varepsilon(1 - \frac{s}{n} + \delta_1) + \varepsilon(1 - \frac{s}{n} + \delta_2) - \varepsilon^2) \prod_{i=3}^{n}(1 - x_i) =$$

$$= P(x') + \varepsilon(\delta_2 - \delta_1 - \varepsilon) \prod_{i=3}^{n} x_i + \varepsilon(\delta_2 - \delta_1 - \varepsilon) \prod_{i=3}^{n}(1 - x_i) > P(x'). \quad (6)$$

Таким образом, из (5) и (6) мы получили $\min P(x)$ и $\max P(x)$, $x \in O_{s,p}$.

Заметим, что если точки x' и x'' таковы, что

$$P(x') = \min_{x \in O_{s,p_1}} P(x), \quad P(x'') = \min_{x \in O_{s,p_2}} P(x),$$

то $x'' \in U_s(x')$ и в силу утверждения 5 $P(x') > P(x'')$.

Аналогично получаем, что если точки x' и x'' таковы, что

$$P(x') = \max_{x \in O_{s,p_1}} P(x), \quad P(x'') = \max_{x \in O_{s,p_2}} P(x),$$

то либо $x'' \in U_s(x')$, либо $x'' \in U_s(y)$, $P(x') > P(y)$, и в силу утверждения 5 $P(x') > P(x'')$. Утверждение доказано.

5. Описание алгоритма поиска

Основным результатом данной работы является алгоритм разбиения объектов из пространства $Z_m^n = \{0, 1, \ldots, m\}^n$ на такие фигуры, что при попадании объекта-запроса в одну из фигур, перебор объектов внутри данной фигуры был примерно одинаков.

Фактически нужно задать такое отображение точек $F : \{0, 1, \ldots, m\}^n \to K \subseteq Z$, $K = \{0, \ldots, k\}$, чтобы в соответствии с характеристиками точек суммарная характеристика фигуры была примерно одинаковой для всех фигур.

Каждой точке x пространства Z_n^n поставим в соответствие номер s слоя, которой принадлежит точка.

Каждой точке x пространства Z_m^n поставим в соответствие расстояние r до точки $(\frac{s}{n}, \ldots, \frac{s}{n})$ на слое.

Разобьем слой s на n частей: каждую часть будем образовывать при помощи $n-1$ угловых точек слоя $(0, \ldots, s, \ldots, 0)$ и точки $(\frac{s}{n}, \ldots, \frac{s}{n})$. Заметим, что данные точки не обязательно принадлежат Z_m^n. Занумеруем эти части числами от 0 до $n-1$. Таким образом, каждой точке x пространства Z_m^n поставим номер части k.

Теперь каждой точке x соответствует тройка (s, r, k).

Будем высчитывать номер фигуры, которой принадлежит точка $x \in Z_m^n$ в зависимости от параметров (s, r, k). Для начала определим параметр $step$ следующим образом: для точек с $r = 0$ $step$ равен 0, для точек с $0 < r \leqslant 1$ $step$ равен 1, для точек с $1 < r \leqslant 3$ $step$ равен 2. Таким образом, склеим окружности, увеличивая каждый размер блока на 1. Теперь осуществим склейку слоев: поставим каждой точке параметр g равный $\lfloor \frac{s}{step+1} \rfloor$. Теперь вычислим по параметрам $(g, step, k)$ номер num. Это и будет номером фигуры.

Сложность вычисления номера фигуры по вышеописанному алгоритму будет $O((m+n)n)$.

В случае, если фигура с номером num не содержит объектов, необходимо найти соседние фигуры, в которых нужно будет продолжить поиск.

Для начала преобразуем num в тройку $(g, step, k)$. Пусть множество $next$ будет содержать соседние фигуры для num. Под добавление в $next$ тройки $(g, step, k)$ будем понимать добавление номера фигуры

num для этой тройки. Рассмотрим соседние фигуры по k. Добавим в $next$ тройки $(g, step, i), 0 \leqslant i < n, i \neq k$. Также проверим фигуры, соседние по g. Если $g > 0$, добавим в $next$ тройку $(g-1, step, k)$. Если $(g+1)(step+1) \leqslant mn$, добавим в $next$ тройку $(g+1, step, k)$. Осталось рассмотреть фигуры, соседние по $step$. Если $step > 0$, то добавим в $next$ те тройки $(g', step-1, k)$, которые имеют хотя бы один общий слой с тройкой $(g, step, k)$. Вычислим r — номер максимальной окружности, которая входит в блок $step$. Если $r < mn$, то добавим в $next$ те тройки $(g', step+1, k)$, которые имеют хотя бы один общий слой с тройкой $(g, step, k)$.

Сложность вычисления номеров соседних фигур по вышеописанному алгоритму будет $O(mn)$.

Таким образом, используя алгоритм вычисления номера фигуры, алгоритм вычисления номеров соседних фигур и алгоритм обхода в ширину графа, мы можем устроить обход нашего разбиения по фигурам.

Данный алгоритм поиска может быть немного модифицирован и перенесен на случай обобщенной модели данных.

6. Результаты экспериментов

Алгоритм поиска был реализован на языке C++ для экспериментов на искусственных данных. Искусственные данные представляли собой кривые изменения свойств t объектов с n начальными $(0, 0, \ldots, 0)$ и конечными свойствами $(m, m, \ldots, m)$. Каждая кривая представляла собой кривую с неограниченным количеством петель.

Изначально база объектов заполнялась t кривыми, удовлетворяющими вышеуказанным свойствам. Каждый элемент кривой вносился в базу объектов как отдельный объект. Согласно алгоритму вычисления номера фигуры каждому объекту ставился в соответствие номер фигуры. После этого к базе поступало r запросов. По запросу высчитывался номер фигуры, которой принадлежит объект-запрос и поиск велся внутри этой фигуры. Эксперимент происходил в 2 этапа: а) запрос был случайной точкой из пространства Z_m^n и б) запрос был слу-

чайной точкой на случайной кривой (учитывались только объекты, которые не принадлежали кривой объекта-запроса).

В ходе эксперимента высчитывались среднее avd, минимальное $mind$ и максимальное $maxd$ расстояние от точки-запроса до точки-ответа, средний диаметр фигуры $diam$, среднее avc, минимальное $minc$ и максимальное $maxc$ количество точек в фигуре, а также частота удачных запросов $reqc$ в случае а) и частота удачных запросов $reqr$ в случае б).

Таблица результатов эксперимента для $n = 10$ и $m = 15$:

t	r	avd	$mind$	$maxd$	avc	$minc$	$maxc$	$reqc$	$reqr$	$diam$
5000	5000	5	0	29	1685	0	7673	0.99	0.79	15.99
10000	10000	5	0	25	3343	0	15348	0.99	0.82	16.01
15000	15000	4	0	28	5011	0	23088	0.99	0.84	15.94

Таблица результатов эксперимента для $n = 15$ и $m = 10$:

t	r	avd	$mind$	$maxd$	avc	$minc$	$maxc$	$reqc$	$reqr$	$diam$
5000	5000	10	0	36	1803	0	13489	0.99	0.87	18.81
10000	10000	9	0	36	3597	0	27129	0.99	0.89	18.67
15000	15000	8	0	37	5476	0	40854	0.99	0.91	18.48

Основываясь на результатах эксперимента, можно заметить, что в записи в среднем распределены примерно равномерно по фигурам и средний диаметр фигуры примерно одинаков. Кроме того, при достаточно большом количестве объектов в базе поиск ближайшего объекта ограничивается лишь фигурой объекта-запроса, и с увеличением количества записей расстояние от объекта-запроса до объекта-ответа уменьшается.

Список литературы

[1] Алисейчик П. А., Вашик К., Кнап Ж., Кудрявцев В. Б., Шеховцов С. Г., Строгалов А. С. Моделирование процесса обучения // Интеллектуальные системы. Т. 10, вып. 1–4. 2006. С. 189–270.

[2] Риордан Дж. Комбинаторные тождества. М.: Наука, 1982.