Федеральное государственное автономное
образовательное учреждение высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт математики и фундаментальной информатики
Базовая кафедра вычислительной и информационной технологии

УТВЕРЖДАЮ
Заведующий кафедрой
________ /В.В.Шайдуров

«17» июня 2016 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

СЕТ-РЕГРЕССИОННЫЙ АНАЛИЗ ЗАВИСИМОСТЕЙ СЛУЧАЙНЫХ СОБЫТИЙ

Направление 02.04.01 Математика и компьютерные науки

Магистерская программа 02.04.01.01 Математическое и компьютерное моделирование

Научный руководитель
кандидат физико-математических наук,
доцент ________ /Д. В. Семенова
17.06.2016

Выпускник ________ / А. И. Иванова
17.06.2016

Красноярск 2016

АННОТАЦИЯ

Объём диссертационной работы 50 страниц, на которых размещены 13 рисунков и 20 таблиц. При написании работы использовалось 25 источников.

Ключевые слова: случайное множество, вероятностное распределение, совместное распределение случайных множеств, сет-регрессия, сет-средние, суперпозиция, ассоциативные функции, рекуррентный метод.

Целью работы является разработка и исследование сет-регрессионных моделей для анализа зависимостей случайных событий в сложных статистических системах. Объектом исследования в представленной диссертации является моделирование сложных статистических систем. Предмет исследования – построение сет-регрессионных отображений конечных множеств событий.

В магистерскую диссертацию входит введение, три главы, заключение, список использованных источников. Во введении раскрывается актуальность исследования по выбранному направлению, ставится проблема, цель и задачи исследования, определяются объект, предмет научных поисков, указывается методологическая база исследования, его теоретическая, практическая значимости. В первой главе изложены элементы теории конечных случайных множеств, даны определения, которые используются в дальнейшем исследовании. Разобраны числовые примеры. Во второй главе рассмотрен алгоритм решения задачи сет-регрессии в классической постановке. Поставлена и решена задача суперпозиции регрессионных отображений. В третьей главе для задачи сет-регрессии используется рекуррентный метод построения вероятностных распределений случайных множеств событий с помощью аппарата ассоциативных функций. Доказаны теоремы о виде условного распределения случайного множества. Рассмотрен пример сет-регрессионного анализа медицинских данных. Заключение посвящено основным выводам, приведены сведения об апробации работы.

ANNOTATION

Volume of the thesis 50 pages to 13 figures and 20 tables are available. When writing the diploma has been used 25 sources.

Keywords: random set of events, probability distribution, joint distribution, set-regression-set average, superposition, associative function, recursive method.

The main objective is to develop and study set-regression models to analyze the dependencies of random events in a sophisticated statistical systems. The object of research when writing the work served as statistical modeling of complex systems. The subject of the research work was the task set-regression.

In his master's thesis includes an introduction, three chapters, conclusion, list of references.

The introduction reveals the relevance of studies in the chosen direction, raise the problem, the purpose and objectives of the study are determined by the object, the subject of scientific research, indicates the methodological base of research, its theoretical, practical significance. In the first chapter, proposed elements of finite random sets, provides definitions that are used in the further study and dismantled at the training examples. The second chapter includes a statement of net-regression and the superposition of tasks set for the regression problem, given algorithms for solving the tasks implemented in practice. The third chapter describes the recurrent method for constructing probability distributions of random sets of events with the help of the device associative functions, we developed an algorithm for constructing the method applied in the academic and practical example (medical data), set out and prove two auxiliary theorems. The conclusion is devoted to the basic conclusions are lists of abstracts and publications.

СОДЕРЖАНИЕ

Введение

Актуальность работы. В настоящее время практически любое исследование подразумевает при обработке результатов наблюдений, экспериментов и измерений использование технологий анализа данных. Их применение позволяет получить новое знание об объекте исследования, выявить скрытые закономерности. Данные нечисловой природы часто встречаются в мониторинге, прогнозировании и управлении в экологических, финансовых, медицинских исследованиях. И для их обработки следует привлекать корректные методы, основанные на соответствующих вероятностных моделях, что зачастую в реальных исследованиях, как в России, так и за рубежом игнорируется. Нередко при обработке данных возникает ситуация, когда исходная информация об объекте исследования представлена неколичественными признаками. В этом случае, адекватной математической моделью данных являются случайные множества, относящиеся к одному из объектов статистики нечисловой природы. Для выборочных данных, описываемых неколичественными признаками, приходится решать те же самые задачи, что и в классических разделах математической статистики: классификация объектов без указания учителя, распознавание образов объектов, оценивание регрессионной зависимости и другие.

Обзор литературы. Случайное множество событий — это случайный элемент со значениями из множества всех подмножеств конечного множества избранных событий $\mathfrak{X}$. Основная идея современной теории случайных множеств состоит в том, что структура статистических взаимозависимостей подмножества конечного множества полностью определяется распределением случайного множества, заданного на множестве всех его подмножеств. Распределение случайного множества — это удобный математический аппарат для описания всех способов взаимодействия элементов между собой. А зависимость среднего значения от какой-либо величины, от некоторой другой величины или от нескольких величин дает нам регрессия. Регрессия [2, 3] — это зависимость среднего значения какой-либо величины от некоторой другой величины или от нескольких величин. Для исследования функциональной зависимости между случайными множествами событий в [2, 3, 18] было предложено искать решение задачи построения регрессионного отображения (функции сет-регрессии) в виде условных сет-средних случайных множеств событий. В [7, 8, 9, 14, 15] сет-регрессия использовалась для моделирования потребительских предпочтений и сегментации рынка. Входными данными для задачи является совместное распределение случайных множеств событий, которое было получено из имеющейся статистики. Однако, чтобы статистически оценить распределение случайного множества N событий на основе наблюдений, приходится сталкиваться с проблемой хронической недостаточности имеющегося количества наблюдений для надлежащей оценки всех 2^N вероятностей $p(X)$, образующее данное вероятностное распределение. Оптимизация данных, как элемент предобработки (препроцессинга), включает снижение размерности входных данных [22]. В [16, 13] предложен рекуррентный метод построения вероятностных распределений случайных множеств событий с помощью аппарата ассоциативных функций. Основная идея рекуррентного метода [16, 13] — выразить вероятности пересечений множества событий функционально через вероятности самих событий, что приводит к уменьшению числа параметров, необходимых для построения вероятностных распределений случайных множеств событий.

Объект и предмет исследования. Объектом исследования в представленной диссертации является моделирование сложных статистических систем. Предмет исследования — построение сет-регрессионных отображений конечных множеств событий.

Цели и задачи. Основной целью работы является разработка и исследование сет-регрессионных моделей для анализа зависимостей случайных событий в сложных статистических системах.

Данная цель достигается путем решения следующих задач:

1. разработать алгоритм для решения задачи сет-регрессии одного случайного множества на другое в классической постановке;
2. разработать алгоритм построения суперпозиции сет-регрессий;
3. разработать метод предобработки данных для задачи сет-регрессии;
4. построить примеры использования сет-регрессионных моделей для анализа данных в экономических и медицинских системах.

Методы исследования основаны на теории вероятностей, математической статистике, методах анализа данных и теории случайных множеств.

Теоретическая значимость работы заключается в том, что получен аналитический вид функции сет-регрессии для вероятностных распределений случайных множеств, рекуррентным методом.

Практическая ценность. Результаты, полученные в работе могут быть применены для модели разделения рынка на потребительские предпочтения, медицинского анализа данных.

Личное участие автора в получении результатов, изложенных в диссертации. Постановка изложенных в диссертации задач была сделана научным руководителем магистранта, к.ф.-м.н., доцентом Семеновой Д.В.. Доказательство и обоснование полученных в диссертации результатов, математические выкладки, численные расчеты выполнены лично автором.

Общая характеристика работы. Во введении раскрывается актуальность исследования по выбранному направлению, ставится проблема, цель и задачи исследования, определяются объект, предмет научных поисков, указывается методологическая база исследования, его теоретическая, практическая значимости. В первой главе изложены элементы теории конечных случайных множеств, даны определения, которые используются в дальнейшем исследовании. Разобраны числовые примеры, иллюстрирующие используемые инструменты теории случайных конечных множеств. Во второй главе рассмотрен алгоритм решения задачи сет-регрессии в классической постановке. Поставлена и решена задача суперпозиции регрессионных отображений. Построена сет-регрессионная модель распределения потребительских предпочтений между фирмами. В третьей главе для предобработки данных для задачи сет-регрессии предлагается использовать рекуррентный метод построения вероятностных распределений случайных множеств событий с помощью аппарата ассоциативных функций. Доказаны две теоремы о виде условного распределения случайного множества, полученного рекуррентным методом с помощью аппарата ассоциативных функций. Разработан алгоритм рекуррентного построения вероятностных распределений для случайных множеств событий. Рассмотрен пример сет-регрессионного анализа медицинских данных. Заключение посвящено основным выводам, приведены сведения об апробации работы.

Структура диссертации. Объём диссертационной работы 51 страниц, на которых размещены 13 рисунков и 20 таблиц. При написании диплома использовалось 25 источников.

1 Элементы теории конечных случайных множеств

1.1 Определение случайного множества событий и его характеристики

Пусть $(\Omega, \mathcal{F}, \mathbf{P})$ — вероятностное пространство, $(\mathbf{U}, \mathcal{A})$ — измеримое пространство, где $\mathbf{U}$ — произвольное множество, а $\mathcal{A}$ — некоторая σ-алгебра его подмножеств. Будем говорить, что функция $K = K(\omega)$, определенная на Ω и принимающая значения в $\mathbf{U}$, есть $\mathcal{F}/\mathcal{A}$-измеримая функция, или случайный элемент [4, 23, 24] (со значениями в множестве $\mathbf{U}$), если для любого $A \in \mathcal{A}$ верно $\{\omega : K(\omega) \in A\} \in \mathcal{F}$.

Случайное множество K на конечном множестве событий $\mathfrak{X} \subset \mathcal{F}$ определяется как отображение $K : \Omega \to 2^{\mathfrak{X}}$, измеримое относительно пары алгебр $\left(\mathcal{F}, 2^{2^{\mathfrak{X}}}\right)$, такое что $\mathbf{P}\left(X\right) = \mathbf{P}\left(K^{-1}(X)\right)$.

Каждое событие $x \in \mathfrak{X}$, разбивает пространство элементарных исходов на два непересекающихся множества: $\Omega = x + x^c$ — событие x и его дополнение — событие $x^c = \Omega - x$. Всё множество событий $\mathfrak{X}$ разбивает пространство элементарных событий на непересекающиеся **события-терраски** следующего вида [3]:

$$\mathrm{ter}\left(X\right) = \left(\left(\bigcap_{x \in X} x\right) \bigcap \left(\bigcap_{x \in X^c} x^c\right)\right), \tag{1.1}$$

$X \subseteq \mathfrak{X}$, где $X^c = \mathfrak{X} - X$ — дополнение подмножества событий X до $\mathfrak{X}$. Таким образом, $\Omega = \sum\limits_{X \subseteq \mathfrak{X}} \mathrm{ter} X$, где

$$\mathrm{ter}(X) \cap \mathrm{ter}(Y) = \oslash \Longleftrightarrow X \neq Y.$$

Распределения случайного множества — это класс специальных вероятностных распределений, характеризующих множества событий. В распределении множества событий содержится всеобъемлющая информация о всех видах вероятностных зависимостей событий из данного множества.

Распределение случайного множества K, определенного на множестве избранных событий $\mathfrak{X}$ можно представить несколькими эквивалентными распределениями вероятностей, порожденным этим множеством событий [3]:

– **распределение I-го рода** случайного множества K на $\mathfrak{X}$ — это набор $\{p(X) = \mathbf{P}(K = X), X \subseteq \mathfrak{X}\}$ из $2^{|\mathfrak{X}|}$ вероятностей вида

$$p(X) = \mathbf{P}\left(\left(\bigcap_{x \in X} x\right) \bigcap \left(\bigcap_{x \in X^c} x^c\right)\right); \tag{1.2}$$

– **распределение II-го рода** случайного множества K на $\mathfrak{X}$ — это набор $\{p_X = \mathbf{P}(X \subseteq K), X \subseteq \mathfrak{X}\}$ из $2^{|\mathfrak{X}|}$ вероятностей вида

$$p_X = \mathbf{P}\left(K \supseteq X\right) = \mathbf{P}\left(\bigcap_{x \in X} X\right), \tag{1.3}$$

которые удовлетворяют системе из $2^{|\mathfrak{X}|}$ неравенств Фреше

$$0 \leq \max\left\{0, 1 - \sum_{x \in X}\left(1 - \mathbf{P}\left(x\right)\right)\right\} \leq p_x \leq \min_{x \in X} \mathbf{P}\left(x\right).$$

Распределения I-го и II-го рода связаны формулами обращения Мёбиуса [3]

$$p_X = \sum_{X \subseteq Y} p(Y), \tag{1.4}$$

$$p(X) = \sum_{X \subseteq Y} (-1)^{|Y|-|X|} p_Y. \tag{1.5}$$

Пример 1. Определим на произвольном триплете событий $\mathfrak{X} = \{x, y, z\}$ распределение I-го рода случайного множества K следующим образом

$$p(\oslash) = \frac{1}{24},$$

$$p(\{x\}) = \frac{3}{24},\ p(\{y\}) = \frac{5}{24},\ p(\{z\}) = \frac{1}{24},$$

$$p(\{x, y\}) = \frac{1}{24},\ p(\{x, z\}) = \frac{2}{24},\ p(\{y, z\}) = \frac{4}{24},$$

$$p(\{x, y, z\}) = \frac{7}{24}.$$

Заметим, что $\sum\limits_{X\subseteq\mathfrak{X}} p(X) = 1$.

Для сокращения записи примем обозначения: $p(\{x\}) = p(x)$, $p(\{x, y\}) = p(xy)$, $p(\{x, y, z\}) = p(xyz)$ и т.д.

Распределение II-го рода для триплета

$$\{p_{\oslash},\ p_x,\ p_y,\ p_z,\ p_{xy},\ p_{xz},\ p_{yz},\ p_{xyz}\}\,,$$

можно получить из распределения I-го рода по формуле обращения Мёбиуса (1.4). Распишем формулу (1.4) для данного триплета:

$$p_{\oslash} = p_x+p_y+p_z+p_{xy}+p_{xz}+p_{yz}+p_{xyz} = \frac{1}{23}+\frac{3}{23}+\frac{5}{23}+\frac{1}{23}+\frac{1}{23}+\frac{2}{23}+\frac{4}{23}+\frac{7}{23} = 1,$$

$$p_x = p(x) + p(xy) + p(xz) + p(xyz) = \frac{3}{24} + \frac{1}{24} + \frac{2}{24} + \frac{7}{24} = \frac{13}{24},$$

$$p_y = p(y) + p(xy) + p(yz) + p(xyz) = \frac{5}{24} + \frac{1}{24} + \frac{4}{24} + \frac{7}{24} = \frac{17}{24},$$

$$p_z = p(z) + p(xz) + p(yz) + p(xyz) = \frac{1}{24} + \frac{2}{24} + \frac{4}{24} + \frac{7}{24} = \frac{14}{24},$$

$$p_{xy} = p(xy) + p(xyz) = \frac{1}{24} + \frac{7}{24} = \frac{8}{24},$$

$$p_{xz} = p(xz) + p(xyz) = \frac{2}{24} + \frac{7}{24} = \frac{9}{24},$$

$$p_{yz} = p(yz) + p(xyz) = \frac{4}{24} + \frac{7}{24} = \frac{11}{24},$$

$$p_{xyz} = p(xyz) = \frac{7}{24}.$$

Обратное преобразование (1.5) распределения II-го рода приводит к исходному распределению I-го рода. Действительно,

$$p(\oslash) = 1 - p_x - p_y - p_z + p_{xy} + p_{xz} + p_{yz} - p_{xyz}$$

$$= 1 - \frac{13}{24} - \frac{17}{24} - \frac{14}{24} - \frac{8}{24} + \frac{9}{24} + \frac{11}{24} - \frac{7}{24} = \frac{1}{24},$$

$$p(x) = p_x - p_{xy} - p_{xz} + p_{xyz} = \frac{13}{24} - \frac{8}{24} - \frac{9}{24} + \frac{7}{24} = \frac{3}{24},$$

$$p(y) = p_y - p_{xy} - p_{yz} + p_{xyz} = \frac{17}{24} - \frac{8}{24} - \frac{11}{24} + \frac{7}{24} = \frac{5}{24},$$

$$p(z) = p_z - p_{xz} - p_{yz} + p_{xyz} = \frac{14}{24} - \frac{9}{24} - \frac{11}{24} + \frac{7}{24} = \frac{1}{24},$$
$$p(xy) = p_{xy} - p_{xyz} = \frac{8}{24} - \frac{7}{24} = \frac{1}{24},$$
$$p(xz) = p_{xz} - p_{xyz} = \frac{9}{24} - \frac{7}{24} = \frac{2}{24},$$
$$p(yz) = p_{yz} - p_{xyz} = \frac{11}{24} - \frac{7}{24} = \frac{4}{24},$$
$$p(xyz) = p_{xyz} = \frac{7}{24}.$$

$\diamond$

1.2 Средние характеристики случайного множества событий

В [2, 3] были введены средне-множественные характеристики случайных конечных множеств, которые обладают экстремальными свойствами, то есть минимизируют некоторые показатели близости детерминированного множества к случайному.

Сет-ожиданием случайного множества K на $\mathfrak{X}$ называется множество

$$\mathcal{E}K = \left\{x \in \mathfrak{X} : p_x \geq h\right\}, \tag{1.6}$$

где h выбирается таким образом, чтобы мощность сет – ожидания была ближайшим целым числом к математическому ожиданию мощности случайного множества K, то есть $h = \min\limits_{\alpha \in [0,1]} \big| |\mathcal{E}K| - \mathbf{E}|K| \big|$.

Сет-квантилем порядка $\alpha \in [0, 1]$ случайного множества K на $\mathfrak{X}$ называется множество

$$\mathbf{Q}_\alpha K = \left\{x \in \mathfrak{X} : p_x \geq \alpha\right\}. \tag{1.7}$$

Сет-квантиль включает те события из множества $\mathfrak{X}$, вероятность которых больше чем α.

Сет-модой случайного множества K на $\mathfrak{X}$ называется множество $\mathrm{Mod}K$, на котором вероятность $\mathbf{P}(K = X)$ принимает максимальное значение, то есть

$$\mathrm{Mod}K = \max_{X \subseteq \mathfrak{X}} \mathbf{P}(K = X). \tag{1.8}$$

Заметим, что для произвольного распределения сет – мода, вообще говоря, не единственна.

Сет-медианой случайного множества событий K на $\mathfrak{X}$ называется множество

$$\mathrm{Med}K = \left\{ x : p_x \geq \frac{1}{2} \right\}. \tag{1.9}$$

Сет-медиана является сет-квантилем порядка $\alpha = \frac{1}{2}$.

Пример 2. Найдем средние характеристики для случайного множества, заданного на триплете событий из примера 1.

1. Сет-квантиль уровня $\alpha = \dfrac{15}{24}$
$$\mathbf{Q}_{\frac{15}{24}} K = \left\{ x \in \mathfrak{X} : p_x \geq \frac{15}{24} \right\} = \{y\}.$$

2. Сет-медиана $\mathrm{Med}K = \left\{ x : p_x \geq \dfrac{1}{2} \right\} = \{x, y, z\}$

3. Сет-мода $\mathrm{Mod}K = \max\limits_{X \subseteq \mathfrak{X}} \mathbf{P}(K = X) = \{y\}$.

4. Для вычисления сет-ожидания по (1.6) введем случайную величину ξ, которая принимает значения равные мощности случайного множества событий на $\mathfrak{X}$, т.е. $\xi = |K|$. Для нашего примера $\xi \in \{0, 1, 2, 3\}$. Найдем вероятностное распределение ξ.
$$\mathbf{P}(\xi = 0) = \frac{1}{24};$$
$$\mathbf{P}(\xi = 1) = p(x) + p(y) + p(z) = \frac{3}{24} + \frac{5}{24} + \frac{1}{24} = \frac{9}{24};$$
$$\mathbf{P}(\xi = 2) = p(xy) + p(xz) + p(yz) = \frac{1}{24} + \frac{4}{24} + \frac{2}{24} = \frac{7}{24};$$
$$\mathbf{P}(\xi = 3) = p(xyz) = \frac{7}{24}.$$
Математическое ожидание ξ равно
$$\mathbf{E}\xi = \mathbf{E}|K| = 0 \cdot \frac{1}{24} + 1 \cdot \frac{9}{24} + 2 \cdot \frac{7}{24} + 3 \cdot \frac{7}{24} = \frac{44}{24}.$$

Уровень h в (1.6) выбирается таким образом, чтобы мощность сет-ожидания была ближайшим целым числом к математическому ожиданию мощности случайного множества K. Если все вероятности событий имеют различные значения, то условие на уровень h можно записать следующим образом [2]:

$$h \ : \ \Big| |\mathcal{E}K| - \mathbf{E}|K| \Big| \leq \frac{1}{2}. \tag{1.10}$$

В качестве значений h будем выбирать вероятность событий. Тогда

– Пусть $h = p_x = \dfrac{13}{24}$. Тогда $\mathcal{E}K = \{x, y, z\}$ и условие (1.10) примет вид $\left|3 - \frac{44}{24}\right| \geq \frac{1}{2}$.

– Пусть $h = p_y = \dfrac{17}{24}$. Тогда $\mathcal{E}K = \{y\}$ и условие (1.10) примет вид $\left|1 - \frac{44}{24}\right| \geq \frac{1}{2}$.

– Пусть $h = p_z = \dfrac{14}{24}$. Тогда $\mathcal{E}K = \{y, z\}$ и условие (1.10) примет вид $\left|2 - \frac{44}{24}\right| \leq \frac{1}{2}$.

Отсюда следует, что необходимо выбрать $h = p_z = \dfrac{14}{24}$ и сет-ожидание будет равно $\mathcal{E}K = \{y, z\}$.

$\diamond$

Следующие теоремы об экстремальных свойствах сет-средних доказаны в работах [2, 3].

Теорема 1.1 (Экстремальность сет–квантиля). *Сет-квантиль порядка α доставляет минимум среднему расстоянию до K*

$$\mathbf{E}\rho_\alpha\left(K, \mathbf{Q}_\alpha K\right) = \min_{X \subseteq \mathfrak{X}} \mathbf{E}\rho_\alpha\left(K, X\right),$$

где ρ_α — асимметричное расстояние на пространстве множеств событий

$$\rho_\alpha\left(X, Y\right) = |X \Delta Y|_\alpha = \alpha|X \setminus Y| + (1 - \alpha)|Y \setminus X|, \qquad \alpha \in (0,\ 1).$$

Теорема 1.2 (Экстремальность сет—моды). *Сет-мода доставляет минимум среднему расстоянию до K*

$$\mathbf{E}\rho\left(K, \mathrm{Mod}K\right) = \min_{X \subseteq \mathfrak{X}} \mathbf{E}\rho\left(K, X\right),$$

где ρ — дискретное расстояние на пространстве множеств событий

$$\rho\left(X, Y\right) = \begin{cases} 1, & X = Y \\ 0, & X \neq Y. \end{cases}$$

Теорема 1.3 (Экстремальность сет—медианы). *Сет-медиана доставляет минимум среднему расстоянию до K*

$$\mathbf{E}\rho\left(K, \mathrm{Med}K\right) = \min_{X \subseteq \mathfrak{X}} \mathbf{E}\rho\left(K, X\right),$$

где ρ — расстояние на пространстве множеств событий

$$\rho\left(X, Y\right) = |X \Delta Y| = |X \setminus Y| + |Y \setminus X|.$$

Из теорем следует, что сет-средние обладают экстремальным свойством минимизировать определенные характеристики близости случайного множества и некоторого детерминированного множества.

1.3 Совместное распределение случайных множеств событий

Введем обозначение $X /\!\!/ \mathfrak{X}$, показывающее, что рассматривается некоторое подмножество X из множества избранных событий $\mathfrak{X}$.

Пусть задано $(\Omega, \mathcal{F}, \mathbf{P})$ — вероятностное пространство. Рассмотрим непересекающиеся множества $\mathfrak{X}, \mathfrak{Y} \subset \mathcal{F}$, $\mathfrak{X} \cap \mathfrak{Y} = \emptyset$. Совместное распределение двух случайных множеств K и L, значения которых содержаться в конечных множествах событий $\mathfrak{X}$ и $\mathfrak{Y}$ соответственно, есть набор из $2^{|\mathfrak{X}|+|\mathfrak{Y}|}$вероятностей $\{p(X, Y), X \subseteq \mathfrak{X}, Y \subseteq \mathfrak{Y}\}$, где

$$p(X, Y) = \mathbf{P}(K = X, L = Y) = \mathbf{P}(\mathrm{ter}(X + Y /\!\!/ \mathfrak{X} + \mathfrak{Y})).$$

Таблица 1 - Совместное распределение $\mathbf{P}(K = X, L = Y)$ двух случайных множеств событий K и L, где $X \subseteq \mathfrak{X}$ на $Y \subseteq \mathfrak{Y}$

	$\mathrm{ter}(\oslash /\!\!/ \mathfrak{Y})$	$\ldots$	$\mathrm{ter}(Y /\!\!/ \mathfrak{Y})$	$\ldots$	$\mathrm{ter}(\mathfrak{Y} /\!\!/ \mathfrak{Y})$	$\mathbf{P}(K = X)$
$\mathrm{ter}(\oslash /\!\!/ \mathfrak{X})$	$p(\mathrm{ter}(\oslash /\!\!/ \mathfrak{X} + \mathfrak{Y}))$	$\ldots$	$p(\mathrm{ter}(Y /\!\!/ \mathfrak{X} + \mathfrak{Y}))$	$\ldots$	$p(\mathrm{ter}(\mathfrak{Y} /\!\!/ \mathfrak{X} + \mathfrak{Y}))$	$p(\mathrm{ter}(\oslash /\!\!/ \mathfrak{X}))$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\mathrm{ter}(X /\!\!/ \mathfrak{X})$	$p(\mathrm{ter}(X /\!\!/ \mathfrak{X} + \mathfrak{Y}))$	$\ldots$	$p(\mathrm{ter}(X + Y /\!\!/ \mathfrak{X} + \mathfrak{Y}))$	$\ldots$	$p(\mathrm{ter}(X + \mathfrak{Y} /\!\!/ \mathfrak{X} + \mathfrak{Y}))$	$p(\mathrm{ter}(X /\!\!/ \mathfrak{X}))$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\mathrm{ter}(\mathfrak{X} /\!\!/ \mathfrak{X})$	$p(\mathrm{ter}(\mathfrak{X} /\!\!/ \mathfrak{X} + \mathfrak{Y}))$	$\ldots$	$p(\mathrm{ter}(\mathfrak{X} + Y /\!\!/ \mathfrak{X} + \mathfrak{Y}))$	$\ldots$	$p(\mathrm{ter}(\mathfrak{X} + \mathfrak{Y} /\!\!/ \mathfrak{X} + \mathfrak{Y}))$	$p(\mathrm{ter}(\mathfrak{X} /\!\!/ \mathfrak{X}))$
$\mathbf{P}(L = Y)$	$p(\mathrm{ter}(\oslash /\!\!/ \mathfrak{Y}))$	$\ldots$	$p(\mathrm{ter}(Y /\!\!/ \mathfrak{Y}))$	$\ldots$	$p(\mathrm{ter}(\mathfrak{Y} /\!\!/ \mathfrak{Y}))$	

В [21] показано, что совместное распределение двух случайных множеств событий удобно представлять в матричном виде (таблица 1).

Условную вероятность множества событий $\mathrm{ter}(Y /\!\!/ \mathfrak{Y})$ при условии, что наступило множество событий $\mathrm{ter}(X /\!\!/ \mathfrak{X})$ будем находить исходя из определения условной вероятности

$$\mathbf{P}\left(\mathrm{ter}(Y)|_{\mathrm{ter}(X)}\right) = \mathbf{P}\left(\{L = Y\}|_{\{K=X\}}\right) = \\ = \frac{\mathbf{P}(\mathrm{ter}(X + Y /\!\!/ \mathfrak{X} + \mathfrak{Y}))}{\mathbf{P}(\mathrm{ter}(X /\!\!/ \mathfrak{X}))} \tag{1.11}$$

Условное распределение случайного множества L при условии, что наступило событие $\mathrm{ter}(X /\!\!/ \mathfrak{X})$,можно представить в виде таблицы 2.

Таблица 2 - Условное распределение случайного множества K при условии, что наступило событие $\mathrm{ter}(Y /\!\!/ \mathfrak{Y})$

$\{L = Y\}$	$\mathrm{ter}(\oslash /\!\!/ \mathfrak{Y})$	$\cdots$	$\mathrm{ter}(Y /\!\!/ \mathfrak{Y})$	$\cdots$	$\mathrm{ter}(\mathfrak{Y} /\!\!/ \mathfrak{Y})$
$\mathbf{P}\left(\{L = Y\}\vert_{\{K=X\}}\right)$	$\frac{\mathbf{P}(\mathrm{ter}(\oslash + X /\!\!/ \mathfrak{X}+\mathfrak{Y}))}{\mathbf{P}(\mathrm{ter}(X /\!\!/ \mathfrak{X}))}$	$\cdots$	$\frac{\mathbf{P}(\mathrm{ter}(X+Y /\!\!/ \mathfrak{X}+\mathfrak{Y}))}{\mathbf{P}(\mathrm{ter}(X /\!\!/ \mathfrak{X}))}$	$\cdots$	$\frac{\mathbf{P}(\mathrm{ter}(\mathfrak{X}+\mathfrak{Y} /\!\!/ \mathfrak{X}+\mathfrak{Y}))}{\mathbf{P}(\mathrm{ter}(X /\!\!/ \mathfrak{X}))}$

Аналогичным образом можно определить условное распределение случайного множества K при условии, что наступило событие $\mathrm{ter}(Y /\!\!/ \mathfrak{Y})$

$$\mathbf{P}\left(\mathrm{ter}(X)\vert_{\mathrm{ter}(Y)}\right) = \mathbf{P}\left(\{K = X\}\vert_{\{L=Y\}}\right) = \\ = \frac{\mathbf{P}(\mathrm{ter}(X + Y /\!\!/ \mathfrak{X} + \mathfrak{Y}))}{\mathbf{P}(\mathrm{ter}(Y /\!\!/ \mathfrak{Y}))} \tag{1.12}$$

Резюме. В данной главе приведен краткий обзор основных определений теории случайных множеств, которые понадобятся в дальнейшей в работе.

2 Сет-регрессионный анализ

Если установлено, что случайные множества зависимы, то возникает естественная задача исследования характера этой зависимости. В работах [2, 18] исследовались задачи построения регрессионных отображений конечных множеств. Известно, что уравнение регрессии обычно выбирается из условного минимума некоторого функционала [10, 11]. В работе [2] рассматривается два подхода к построению регрессионных отображений. Первый состоит в том, что уравнение регрессии находится в виде условных сет – средних на основе абсолютных экстремальных свойств средних характеристик случайных конечных множеств. При втором подходе оптимальные регрессионные отображения выбираются из некоторых параметрических семейств отображений. В работе анализируется и исследуется первый подход.

2.1 Задача сет-регрессии

Если дано совместное распределение двух случайных множеств K и L, значения которых содержаться в конечных множествах $\mathfrak{X}$ и $\mathfrak{Y}$ соответственно, то регрессией L на K называется любой оператор Ψ, приближенно представляющий статистическую зависимость L от K [2, 3, 18]. Для определения зависимости между двумя случайными множествами событий в [2, 3, 18] было предложено использовать случайно-множественную регрессию (сет-регрессию), которая устанавливает вид средней функциональной зависимости между этими двумя случайными множествами событий.

В работе используется постановка задачи нахождения сет-регрессии случайных множеств, предложенная в [2].

Задача сет-регрессии. Пусть задано совместное распределение двух случайных множеств событий K и L, значения которых содержаться в конечных множествах $\mathfrak{X}$ и $\mathfrak{Y}$ соответственно. Пусть $\rho : 2^{\mathfrak{Y}} \to 2^{\mathfrak{Y}}$ — некоторая метрика в $\mathfrak{Y}$. Необходимо найти сет-функцию $\Psi : 2^{\mathfrak{X}} \to 2^{\mathfrak{Y}}$, которая доставляет минимум функционалу $\mathbf{E}\rho(L, \Psi(K))$, где $\mathbf{E}$ оператор математического ожидания.

По аналогии с многомерным статистическим подходом, регрессию одного случайного множества на другое можно искать в виде условных сет-средних [2]:

– условного сет-квантиля порядка α

$$\mathbf{Q}_\alpha \left(L \left|_{\{K=X\}}\right.\right) = \left\{y : \mathbf{P}\left(y \left|_{\mathrm{ter}(X)}\right.\right) \geq \alpha\right\}, \tag{2.13}$$

если $\rho_\alpha(X, Y) = \alpha|X \setminus Y| + (1-\alpha)|Y \setminus X|$;

– условной сет-медианы

$$\mathrm{Med}\left(L \left|_{\{K=X\}}\right.\right) = \left\{y : \mathbf{P}\left(y \left|_{\mathrm{ter}(X)}\right.\right) \geq \frac{1}{2}\right\}, \tag{2.14}$$

если $\rho(X, Y) = |X \Delta Y|$;

– условной сет-моды

$$\mathrm{Mod}\left(L \left|_{\{K=X\}}\right.\right) = \max_{Y \subseteq \mathfrak{Y}} \mathbf{P}\left(Y \left|_{\{K=X\}}\right.\right), \tag{2.15}$$

если $\rho(X, Y) = \begin{cases} 1, & X = Y \\ 0, & X \neq Y \end{cases}$.

В качестве теоретической основы для этого подхода выступают приведенные в первой главе экстремальные свойства сет-средних (теоремы 1.1 – 1.3). Следует заметить, что выбор того или иного условного сет-среднего в качестве решения задачи зависит от метрики ρ. Таким образом, регрессии одного случайного множества на другое в виде условной сет-моды, сет-медианы, сет-ожидания обладают экстремальным свойством минимизировать определенные характеристики близости случайного множества и его регрессии на другое случайное множество.

Найденное решение может быть представлено визуально, что является удобным, а в некоторых случаях и необходимым, способом представления зависимости событий, обеспечивая для исследователя интуитивно понятное и удобное восприятие, а также легкость сопоставления и сравнения.

В работе предложен общий алгоритм решения задачи построения сет-регрессии через условные сет-средние (алгоритм 1).

Алгоритм 1 Алгоритм решения задачи сет-регрессии

Вход: совместное распределение $\{p(X,Y), X \subseteq \mathfrak{X}, Y \subseteq \mathfrak{Y}\}$ случайных множеств K и L, значения которых содержаться в непересекающихся множествах $\mathfrak{X}, \mathfrak{Y} \subset \mathcal{F}$, соответственно.

Выход: Функция сет-регрессии $\Psi : 2^{\mathfrak{X}} \to 2^{\mathfrak{Y}} : \Psi(X) = \min\limits_{\Psi} \mathbf{E}\rho(L, \Psi(X))$.

1: Из совместного распределения распределения двух случайных множеств событий найти маргинальное распределение случайного множества K:

$$p(X /\!\!/ \mathfrak{X}) = \sum_{Y \subseteq \mathfrak{Y}} p(X,Y).$$

2: Найти условные распределения случайных множеств $L|_X$ при условии, что наступило подмножество событий $X \subseteq \mathfrak{X}$:

$$p\left(L|_X\right) = \mathbf{P}(\{L = Y\}|_{\{K=X\}}) = \frac{\mathbf{P}(\mathrm{ter}(X + Y /\!\!/ \mathfrak{X} + \mathfrak{Y}))}{\mathbf{P}(\mathrm{ter}(X /\!\!/ \mathfrak{X}))}.$$

3: Для всех $X \subseteq \mathfrak{X}$ вычислить условные сет-средние характеристики случайных множеств с распределениями $\{p\left(L|_X\right), Y \in \mathfrak{Y}\}$.

4: Выбрать условное сет-среднее и построить с его помощью функцию регрессии $\Psi(X)$, $X \subseteq \mathfrak{X}$.

5: **Возвратить** $\Psi(X)$.

Рассмотрим алгоритм решения задачи сет-регрессии на примере.

Пример 3. Пусть заданы

– множество событий $\mathfrak{X} = \{x_1, x_2, x_3\}$, $\mathfrak{X} \subset \mathcal{F}$;

– множество $\mathfrak{Y} = \{y_1, y_2, y_3\}$, $\mathfrak{Y} \subset \mathcal{F}$;

– совместное распределение

$$\{p(X,Y) = \mathbf{P}(K = X, L = Y), X \subseteq \mathfrak{X}, Y \subseteq \mathfrak{Y}\}$$

двух случайных множеств событий K и L, значения которых содержаться в конечных множествах $\mathfrak{X}$ и $\mathfrak{Y}$ соответственно (таблица 3).

Найти сет-функцию $\Psi : 2^{\mathfrak{Y}} \to 2^{\mathfrak{X}}$, которая доставляет минимум функционалу $\mathbf{E}\rho(K, \Psi(Y))$, где $\mathbf{E}$ оператор математического ожидания, а $\rho :\ 2^{\mathfrak{X}} \to 2^{\mathfrak{X}}$ — некоторая метрика в $\mathfrak{X}$.

Таблица 3 - Матричное представление совместного распределения двух случайных множеств событий K и L для примера 3.

$\mathfrak{Y}$ \ $\mathfrak{X}$	$\mathrm{ter}(\oslash)$	$\mathrm{ter}(x_1)$	$\mathrm{ter}(x_2)$	$\mathrm{ter}(x_3)$	$\mathrm{ter}(x_1x_2)$	$\mathrm{ter}(x_1x_3)$	$\mathrm{ter}(x_2x_3)$	$\mathrm{ter}(x_1x_2x_3)$
$\mathrm{ter}(\oslash)$	0.01	0.02	0.02	0	0.03	0.03	0.03	0.02
$\mathrm{ter}(y_1)$	0.02	0.02	0.03	0	0.01	0.03	0	0
$\mathrm{ter}(y_2)$	0.01	0.03	0.02	0.02	0	0.02	0.01	0.02
$\mathrm{ter}(y_3)$	0	0.03	0.01	0.02	0.01	0.02	0.02	0
$\mathrm{ter}(y_1y_2)$	0.02	0.02	0.01	0.03	0.02	0.01	0.03	0.02
$\mathrm{ter}(y_1y_3)$	0.01	0.02	0	0.03	0.03	0.03	0	0.01
$\mathrm{ter}(y_2y_3)$	0	0.01	0	0.03	0.01	0.02	0.01	0.02
$\mathrm{ter}(y_1y_2y_3)$	0.01	0.02	0	0	0.03	0.01	0.01	0

Будем искать решение задачи по алгоритму 1 в виде сет-средних по формулам (2.13) — (2.15).

На первом этапе решения из таблицы совместного распределения двух случайных множеств событий K и L (таблица 3) найдем маргинальные распределения случайных множеств K (таблица 4) и L (таблица 5), суммируя соответствующие столбцы или строки по формуле

$$p(X /\!\!/ \, \mathfrak{X}) = \sum_{Y \subseteq \mathfrak{Y}} p(X,Y).$$

Таблица 4 - Маргинальное распределение случайного множества событий K на $\mathfrak{X}$ для примера 3.

$\mathrm{ter}(X)$, $X \subseteq \mathfrak{X}$	$p(X) = \mathbf{P}(\{K = X\})$
$\mathrm{ter}(\oslash) = x_1^c \cap x_2^c \cap x_3^c$	0.151
$\mathrm{ter}(x_1) = x_1 \cap x_2^c \cap x_3^c$	0.115
$\mathrm{ter}(x_2) = x_1^c \cap x_2 \cap x_3^c$	0.138
$\mathrm{ter}(x_3) = x_1^c \cap x_2^c \cap x_3$	0.109
$\mathrm{ter}(x_1x_2) = x_1 \cap x_2 \cap x_3^c$	0.166
$\mathrm{ter}(x_1x_3) = x_1 \cap x_2^c \cap x_3$	0.143
$\mathrm{ter}(x_2x_3) = x_1^c \cap x_2 \cap x_3$	0.108
$\mathrm{ter}(x_1x_2x_3) = x_1 \cap x_2 \cap x_3$	0.068

Таблица 5 - Маргинальное распределение случайного множества событий L на $\mathfrak{Y}$ для примера 3.

$\mathrm{ter}(Y)$, $Y \subseteq \mathfrak{Y}$	$p(Y) = \mathbf{P}(\{L = Y\})$
$\mathrm{ter}(\oslash) = y_1^c \cap y_2^c \cap y_3^c$	0.093
$\mathrm{ter}(y_1) = y_1 \cap y_2^c \cap y_3^c$	0.168
$\mathrm{ter}(y_2) = y_1^c \cap y_2 \cap y_3^c$	0.104
$\mathrm{ter}(y_3) = y_1^c \cap y_2^c \cap y_3$	0.135
$\mathrm{ter}(y_1y_2) = y_1 \cap y_2 \cap y_3^c$	0.135
$\mathrm{ter}(y_1y_3) = y_1 \cap y_2^c \cap y_3$	0.162
$\mathrm{ter}(y_2y_3)) = y_1^c \cap y_2 \cap y_3$	0.111
$\mathrm{ter}(y_1y_2y_3)) = y_1 \cap y_2 \cap y_3$	0.092

На втором этапе найдем все условные распределения случайного множества K при условии, что наступило подмножество событий $Y \subseteq \mathfrak{Y}$, пользуясь формулой (1.12). Например, при фиксированных событиях $\{K = \mathrm{ter}(x_1)\}$ и

$\{L = \mathrm{ter}(y_1)\}$ условная вероятность согласно (1.12) равна

$$\mathbf{P}\left(\{K = \mathrm{ter}(x_1)\}\big|_{\{L=\mathrm{ter}(y_1)\}}\right) = 0.02/0.115 = 0.171.$$

Результаты расчетов сведены в таблице 6.

По формуле (1.4) для каждого из условных распределений таблицы 6 найдем соответствующие вероятности событий из $\mathfrak{X}$ (вероятности II-го рода). Результаты расчетов приведены в таблице 7.

Таблица 6 - Условные распределения случайного множества K при условии, что наступило множество событий $Y \subseteq \mathfrak{Y}$ для примера 3.

$\mathfrak{Y}$ \ $\mathfrak{X}$	$\mathrm{ter}(\varnothing)$	$\mathrm{ter}(x_1)$	$\mathrm{ter}(x_2)$	$\mathrm{ter}(x_3)$	$\mathrm{ter}(x_1x_2)$	$\mathrm{ter}(x_1x_3)$	$\mathrm{ter}(x_2x_3)$	$\mathrm{ter}(x_1x_2x_3)$
$\mathbf{P}(K\|_{L=\mathrm{ter}(\varnothing)})$	0.044	0.117	0.12	0.03	0.172	0.185	0.18	0.15
$\mathbf{P}(K\|_{L=\mathrm{ter}(y_1)}))$	0.174	0.171	0.251	0.027	0.094	0.231	0.039	0.013
$\mathbf{P}(K\|_{L=\mathrm{ter}(y_2)}))$	0.096	0.212	0.148	0.132	0.006	0.161	0.104	0.141
$\mathbf{P}(K\|_{L=\mathrm{ter}(y_3)}))$	0.03	0.248	0.13	0.189	0.079	0.156	0.156	0.012
$\mathbf{P}(K\|_{L=\mathrm{ter}(y_1y_2)}))$	0.138	0.133	0.072	0.138	0.123	0.08	0.178	0.093
$\mathbf{P}(K\|_{L=\mathrm{ter}(y_1y_3)}))$	0.103	0.167	0.018	0.209	0.206	0.179	0.016	0.103
$\mathbf{P}(K\|_{L=\mathrm{ter}(y_2y_3)}))$	0.042	0.091	0.034	0.247	0.125	0.222	0.086	0.152
$\mathbf{P}(K\|_{L=\mathrm{ter}(y_1y_2y_3)}))$	0.118	0.269	0.055	0.01	0.374	0.074	0.088	0.012

Найдем решение задачи в виде сет-квантиля уровня $\alpha = 1/4$. Формула (1.7) примет вид

$$\mathbf{Q}_{1/4}\left[K\left|_{L=\mathrm{ter}(Y)}\right.\right] = \left\{x : \mathbf{P}\left(x\left|_{\mathrm{ter}(Y)}\right.\right) \geq 1/4\right\}. \quad (2.16)$$

Введём функцию $I_\alpha : \mathfrak{X} \times 2^{\mathfrak{Y}} \to \{0, 1\}$, $\alpha \in [0, 1]$ значения которой определяются по формуле

$$I_\alpha(x, Y) = \begin{cases} 1, & \mathbf{P}\left(x|_{\mathrm{ter}(Y)}\right) \geq \alpha \\ 0, & \text{иначе.} \end{cases} \quad (2.17)$$

Функция (2.17) фактически является индикатором множества, построенного по формуле (1.12), то есть индикатором условного сет-квантиля порядка α.

Таблица 7 - Вероятности событий из $\mathfrak{X}$ для условных распределений случайного множества K при условии, что наступило множество событий $Y \subseteq \mathfrak{Y}$ для примера 3.

$\mathfrak{Y}$	$\mathbf{P}\left(x_1\vert_{\mathrm{ter}(Y)}\right)$	$\mathbf{P}\left(x_2\vert_{\mathrm{ter}(Y)}\right)$	$\mathbf{P}\left(x_3\vert_{\mathrm{ter}(Y)}\right)$
$\mathrm{ter}(\emptyset)$	0.625	0.623	0.546
$\mathrm{ter}(y_1)$	0.509	0.397	0.31
$\mathrm{ter}(y_2)$	0.52	0.399	0.538
$\mathrm{ter}(y_3)$	0.495	0.377	0.513
$\mathrm{ter}(y_1 y_2)$	0.428	0.466	0.534
$\mathrm{ter}(y_1 y_3)$	0.655	0.342	0.507
$\mathrm{ter}(y_2 y_3)$	0.59	0.379	0.708
$\mathrm{ter}(y_1 y_2 y_3)$	0.729	0.53	0.184

Решение задачи в виде условного сет-квантиля представим в виде таблицы 8, элементы которой находятся по формуле (2.17). Каждая строка таблицы является индикаторной функцией соответствующего условного сет-квантиля. По таблице 8 запишем решение задачи в виде сет-функции:

$$\Psi\left(Y\right) = \{x \in \mathfrak{X} : I_{1/4}\left(x, Y\right) = 1\},$$

для всех $Y \subseteq \mathfrak{Y}$.

На рисунке 1 представлено решение задачи сет-регрессии в виде условного сет-квантиля порядка $\alpha = 1/4$.

Используя таблицу 7 и (2.17) по аналогии запишем решение задачи в виде условной сет-медианы

$$\Psi\left(Y\right) = \{x \in \mathfrak{X} : I_{1/2}\left(x, Y\right) = 1\},$$

для всех $Y \subseteq \mathfrak{Y}$. Матричное представление дано в таблице 9, график решения представлен на рисунке 2.

Для нахождения сет-регрессии в виде сет-моды воспользуемся таблицей

Таблица 8 - Матричное представление решения задачи в виде условного сет-квантиля порядка $\alpha = 1/4$. Каждая строка таблицы является индикаторной функцией соответствующего условного сет-квантиля.

	x_1	x_2	x_3
$\mathrm{ter}(\varnothing)$	1	1	1
$\mathrm{ter}(y_1)$	1	1	1
$\mathrm{ter}(y_2)$	1	1	1
$\mathrm{ter}(y_3)$	1	1	1
$\mathrm{ter}(y_1 y_2)$	1	1	1
$\mathrm{ter}(y_1 y_3)$	1	1	1
$\mathrm{ter}(y_2 y_3)$	1	1	1
$\mathrm{ter}(y_1 y_2 y_3)$	1	1	0

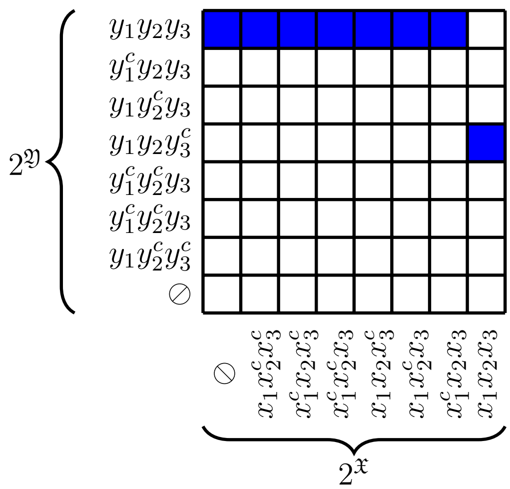

Рисунок 1 - Решение задачи сет-регрессии множества $\mathfrak{X}$ на $\mathfrak{Y}$ в виде условного сет-квантиля порядка $\alpha = 1/4$.

6 и формулой (2.15). Введём функцию $J : 2^{\mathfrak{X}} \times 2^{\mathfrak{Y}} \to \{0, 1\}$, $\alpha \in [0, 1]$ значения которой определяются по формуле

$$J(X, Y) = \begin{cases} 1, & \mathbf{P}(X) = \max\limits_{Z \subseteq \mathfrak{X}} \mathbf{P}\left(Z\,|_Y\right), \\ 0, & \text{иначе}. \end{cases} \tag{2.18}$$

Функция (2.18) фактически является индикатором множества, постро-

Таблица 9 - Матричное представление решения задачи в виде условной сет-медианы. Каждая строка таблицы является индикаторной функцией соответствующего условной сет-медианы.

	x_1	x_2	x_3
$\mathrm{ter}(\oslash)$	1	1	1
$\mathrm{ter}(y_1)$	1	0	0
$\mathrm{ter}(y_2)$	1	0	1
$\mathrm{ter}(y_3)$	0	0	1
$\mathrm{ter}(y_1y_2)$	0	0	1
$\mathrm{ter}(y_1y_3)$	1	0	1
$\mathrm{ter}(y_2y_3)$	1	0	1
$\mathrm{ter}(y_1y_2y_3)$	1	1	0

енного по формуле (2.15), то есть индикатором условной сет-моды. Решение задачи в виде условной сет-моды представим в виде таблицы 9, элементы которой находятся по формуле (2.18). График решения представлен на рисунок 3.

Таблица 10 - Матричное представление решения задачи в виде условной сет-моды.

$\mathrm{ter}(x_1x_2x_3)$	0	0	0	0	0	0	0	0
$\mathrm{ter}(x_2x_3)$	0	0	0	0	0	0	0	0
$\mathrm{ter}(x_1x_3)$	1	0	0	0	0	0	0	0
$\mathrm{ter}(x_1x_2)$	0	0	0	0	0	0	0	1
$\mathrm{ter}(x_3)$	0	0	0	0	1	1	1	0
$\mathrm{ter}(x_2)$	0	1	0	0	0	0	0	0
$\mathrm{ter}(x_1)$	0	0	1	1	0	0	0	0
$\mathrm{ter}(\oslash)$	0	0	0	0	0	0	0	0
	$\mathrm{ter}(\oslash)$	$\mathrm{ter}(y_1)$	$\mathrm{ter}(y_2)$	$\mathrm{ter}(y_3)$	$\mathrm{ter}(y_1y_2)$	$\mathrm{ter}(y_1y_3)$	$\mathrm{ter}(y_2y_3)$	$\mathrm{ter}(y_1y_2y_3)$

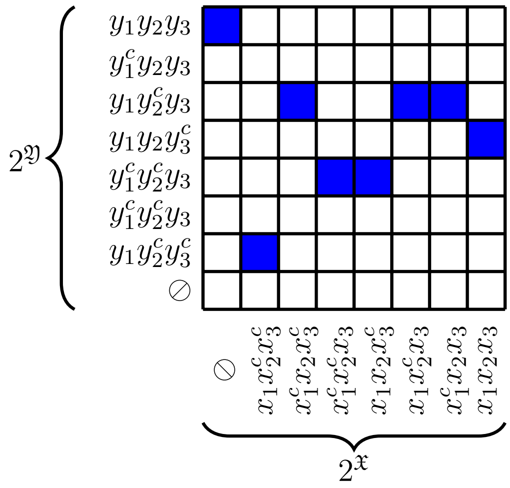

Рисунок 2 - Решение задачи сет-регрессии множества $\mathfrak{X}$ на $\mathfrak{Y}$ в виде условной сет-медианы.

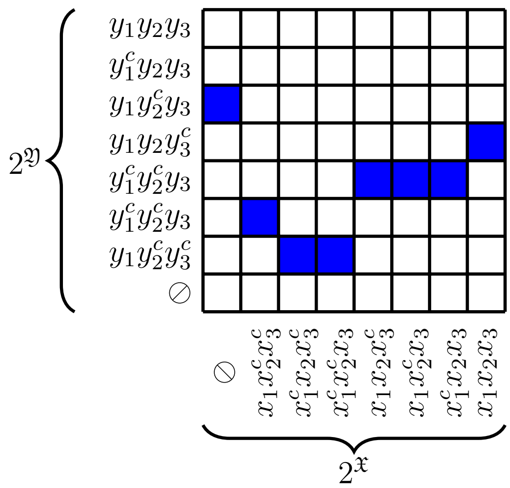

Рисунок 3 - Решение задачи сет-регрессии множества $\mathfrak{X}$ на $\mathfrak{Y}$ в виде условной сет-моды.

$\diamond$

2.2 Суперпозиция сет-регрессий

Задача суперпозиции сет-регрессий является модификацией для задачи сет-регрессии. Модификация заключается в следующим: имея сет-регрессионное отображение случайного множества K на случайное множество L и сет-регрессионное отображение случайного множества L на случайное множество M для одних и тех же типов сет-средних, можно получит сет-регрессионное отображение случайного множества K на случайное множество M.

Задача построения суперпозиции сет-регрессий. Пусть заданы

– случайное множество событий K на конечном множестве $\mathfrak{X} \subset \mathcal{F}$,

– случайное множество событий L на конечном множестве $\mathfrak{Y} \subset \mathcal{F}$

– и случайное множество событий M на конечном множестве $\mathfrak{Z} \subset \mathcal{F}$,

$\mathfrak{X} \cap \mathfrak{Y} \cap \mathfrak{Z} = \oslash$.

И пусть известны

– совместное распределение $\{p(X,Y), X \subseteq \mathfrak{X}, Y \subseteq \mathfrak{Y}\}$ случайных множеств событий K и L, где $p(X,Y) = \mathbf{P}(K = X, L = Y)$

– и совместное распределение $\{p(Y,Z), Y \subseteq \mathfrak{Y}, Z \subseteq \mathfrak{Z}\}$ случайных множеств событий L и M, где $p(Y,Z) = \mathbf{P}(L = Y, M = Z)$.

Найти: сет-функцию $\Psi : 2^{\mathfrak{X}} \to 2^{\mathfrak{Z}}$, которая устанавливает зависимость между случайными множествами K и M.

В работе предложен алгоритм, который позволяет строить суперпозицию сет-регрессий (Алгоритм 2). Алгоритм рассмотрен на примере распределение потребительских предпочтений между фирмами.

2.3 Сет-регрессионная модель распределения потребительских предпочтений между фирмами

Сегментация рынка, фундамент любого маркетингового исследования, заключается в разделении потребителей на группы так, чтобы покупатели внутри группы обладали примерно одинаковыми потребностями (потребительскими предпочтениями) и характеристиками. На основе полученного раз-

Алгоритм 2

Вход: – случайное множество событий K на конечном множестве $\mathfrak{X} \subset \mathcal{F}$,

– случайное множество событий L на конечном множестве $\mathfrak{Y} \subset \mathcal{F}$

– и случайное множество событий M на конечном множестве $\mathfrak{Z} \subset \mathcal{F}$, $\mathfrak{X} \cap \mathfrak{Y} \cap \mathfrak{Z} = \oslash$.

– совместное распределение $\{p(X, Y), X \subseteq \mathfrak{X}, Y \subseteq \mathfrak{Y}\}$ случайных множеств событий K и L, где $p(X, Y) = \mathbf{P}(K = X, L = Y)$

– и совместное распределение $\{p(Y, Z), Y \subseteq \mathfrak{Y}, Z \subseteq \mathfrak{Z}\}$ случайных множеств событий L и M, где $p(Y, Z) = \mathbf{P}(L = Y, M = Z)$.

Выход: сет-функция $\Psi : 2^{\mathfrak{X}} \to 2^{\mathfrak{Z}}$, которая устанавливает зависимость между случайными множествами K и M.

Найти решение задачи сет-регрессии для множеств K и L в виде некоторого условного сет-среднего: $\theta\left(\{L = Y\}|_{\{K=X\}}\right), X \subseteq \mathfrak{X}, Y \subseteq \mathfrak{Y}$.

2: Найти условные распределения случайных множеств $L|_X$ при условии, что наступило подмножество событий $X \subseteq \mathfrak{X}$:

$$p(L|_X) = \mathbf{P}(\{L = Y\}|_{\{K=X\}}) = \frac{\mathbf{P}(\mathrm{ter}(X + Y /\!\!/ \mathfrak{X} + \mathfrak{Y}))}{\mathbf{P}(\mathrm{ter}(X /\!\!/ \mathfrak{X}))}.$$

Найти решение задачи сет-регрессии для множеств L и M в виде некоторого условного сет-среднего: $\delta\left(\{M = Z\}|_{\{L=Y\}}\right), Y \subseteq \mathfrak{Y}, M \subseteq \mathfrak{Z}$.

4: Найти суперпозицию $\Psi(X) = \delta\left(\{M = Z\}|_{\theta\left(\{L=Y\}|_{\{K=X\}}\right)}\right)$.

биения рынка на сегменты разрабатываются маркетинговые стратегии, позволяющие фирме более полно использовать свои конкурентные преимущества, максимально эффективно использовать имеющиеся ресурсы.

Для сегментации рынка необходимы исходные данные. Отличительными особенностями таких данных являются многомерность, т.е. каждый исследуемый объект характеризуется некоторым набором признаков, разнотипность и неточность данных. Многомерность, разнотипность и неполнота данных значительно усложняют обработку и анализ, затрудняют процесс разработки модели сегментации рынка. Поэтому особенно актуальной становится разработка модели исходных данных, позволяющая учесть особенности данных. Цена, качество и выбор товара, репутация фирмы, уровень обслуживания и трудности пути — все это важнейшие характеристики для восприятия фирмы покупателем. Более того при принятии решения о поездки за покупками у индивидуумов варьируется приоритеты этих характеристик. Таким образом все эти характеристики нужно объединить в рамках общей модели. В работе предложен сет-регрессионный подход построения модели распределения потребительских предпочтений между фирмами, выпускающими косметическую продукцию.

Пусть дано

– множество покупателей

$$\mathfrak{X} = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}\},$$

где $x_i, i = \overline{1, 12}$. В таблице 11 представлены все признаки, которые используются в анкете. После анализа анкетных данных множество покупателей определено следующими признаками потребителей:

$$\begin{array}{lll}
-\,x_1 = (a_1, b_3, c_1); & -\,x_5 = (a_2, b_1, c_2); & -\,x_9 = (a_2, b_3, c_4);\\
-\,x_2 = (a_1, b_1, c_3); & -\,x_6 = (a_2, b_2, c_2); & -\,x_{10} = (a_3, b_1, c_1);\\
-\,x_3 = (a_2, b_1, c_1); & -\,x_7 = (a_2, b_3, c_2); & -\,x_{11} = (a_3, b_2, c_3);\\
-\,x_4 = (a_2, b_3, c_1); & -\,x_8 = (a_2, b_1, c_4); & -\,x_{12} = (a_3, b_2, c_4);
\end{array}$$

– множество «Признаков товара»

$$\mathfrak{Y} = \{a, b, c, d\} =$$

$$= \{\text{«Доступность»}, \text{«Высокое качество»},$$

$$\text{«Низкая цена»}, \ \text{«Известность производителя»}\};$$

Таблица 11 - Признаки потребителей

покупка туши		возраст покупателя		доход	
1 раз в месяц	a_1	от 18 до 25 лет	b_1	Высокий- св. 25000	c_1
раз в 3 месяца	a_2	от 26 до 38 лет	b_2	Выше среднего-15000-25000	c_2
раз в полгода	a_3	от 39 до 45 лет	b_3	Низкий- 6000-10000	c_3
				Очень низкий—2000-6000	c_4

– Множество «Фирм» $\mathfrak{Z} = \{u, v, w, t\} = \{$«Loreal», «Max Factor», «Bourjois», «Maybelline»$\}$

И пусть известны совместное распределение

$$\{p(X, Y) = \mathbf{P}(K = X, L = Y), X \subseteq \mathfrak{X}, Y \subseteq \mathfrak{Y}\}$$

двух случайных множеств событий K и L, значения которых содержаться в конечных множествах $\mathfrak{X}$ и $\mathfrak{Y}$ соответственно; совместное распределение

$$\{p(Y, Z) = \mathbf{P}(L = Y, M = Z), Y \subseteq \mathfrak{Y}, Z \subseteq \mathfrak{Z}\}$$

двух случайных множеств событий L и M, значения которых содержаться в конечных множествах $\mathfrak{Y}$ и $\mathfrak{Z}$ соответственно.

Требуется найти некоторую функцию множеств $\Psi : 2^{\mathfrak{X}} \longrightarrow 2^{\mathfrak{Z}}$, описывающую распределение потребительских предпочтений между фирмами, выпускающими косметическую продукцию.

Решение задачи. Решение состоит из трех этапов:

1. На первом этапе необходимо найти решение задачи сет-регрессии для множеств $\mathfrak{X}$ и $\mathfrak{Y}$ в виде некоторого сет-среднего $\Theta\left(\{L = Y\}\,\middle|_{\{K=X\}}\right)$, $X \subseteq \mathfrak{X}$, $Y \subseteq \mathfrak{Y}$.

На рисунке 4 представлено решение задачи сет-регрессии множества $\mathfrak{Y}$ на $\mathfrak{X}$ в виде сет-квантиля с уровнем $\alpha = 2/5$.

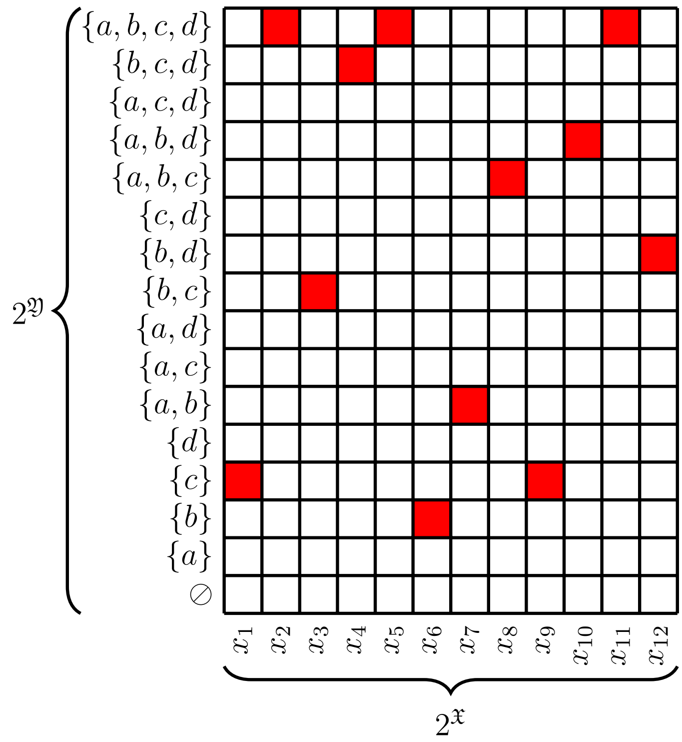

Рисунок 4 - Решение задачи сет-регрессии множества $\mathfrak{Y}$ на $\mathfrak{X}$ в виде сет-квантиля с уровнем $\alpha = 2/5$.

2. На втором этапе необходимо найти решение задачи сет-регрессии для множеств $\mathfrak{Y}$ и $\mathfrak{Z}$ в виде некоторого сет-среднего $\Phi\left(\{M = Z\}\big|_{\{L=Y\}}\right)$, $Y \subseteq \mathfrak{Y}$, $Z \subseteq \mathfrak{Z}$.

   На рисунке 5 представлено решение задачи сет-регрессии множества $\mathfrak{Z}$ на $\mathfrak{Y}$ в виде сет-квантиля с уровнем $\alpha = 2/5$.

3. Решение задачи определяется как суперпозиция

$$\Psi(X) = \Phi\left(\{M = Z\}\Big|_{\Theta\left(\{L=Y\}\big|_{\{K=X\}}\right)}\right).$$

   На рисунке 6 дано графическое представление суперпозиции сет-функций. На рисунке 7 представлено решение задачи — сет-функция $\Psi(X)$, $X \subseteq \mathfrak{X}$.

Сет-функция $\Psi(X)$ показывает какую фирмы предпочитают покупатели. Например $x_8 = (a_2, b_1, c_4)$— множестно покупателей в возрасте от 18 до 25 лет с доходом 2000-6000 и преобретакю тушь раз в три месяца предпочитают следующие фирмы туши v, w, t, где u—«Loreal», v— «Max Factor», w «Bourjois», t—«Maybelline».

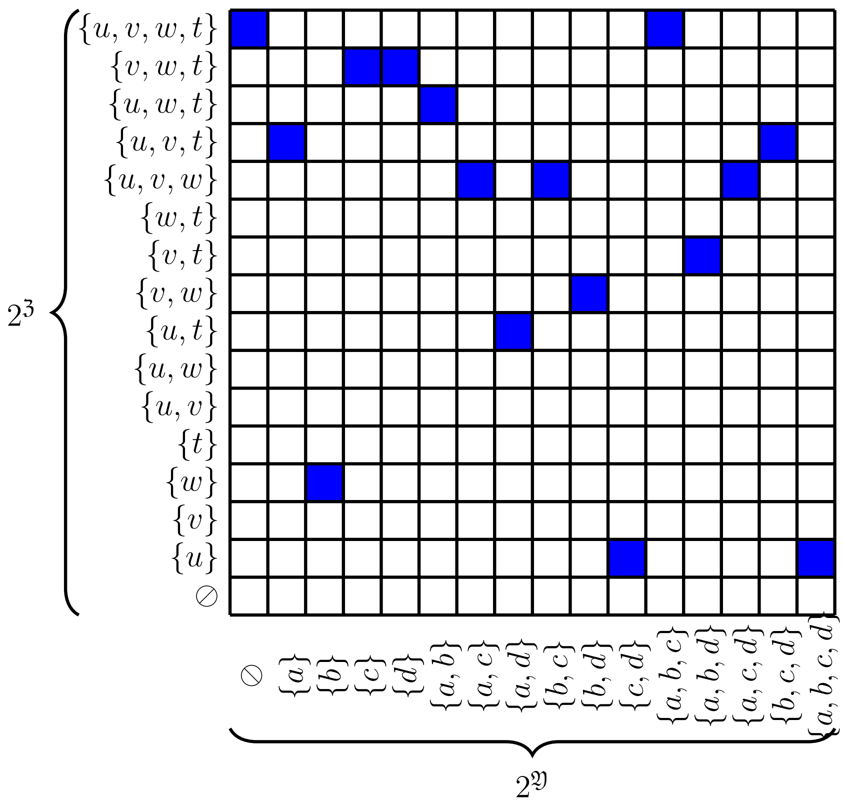

Рисунок 5 - Решение задачи сет-регрессии множества $\mathfrak{Z}$ на $\mathfrak{Y}$ в виде сет-квантиля с уровнем $\alpha = 2/5$.

Аналогично можно построить регрессионные отображения случайных множеств событий на основе других сет-средних [2, 17]. Анализ и сравнение предложенной сет-регрессионной модели с нечеткой моделью Леунга дан в работах [6, 14, 15].

Резюме. В главе рассмотрен алгоритм решения задачи сет-регрессии в классической постановке. Поставлена и решена задача суперпозиции регрессионных отображений. Построена сет-регрессионная модель распределения

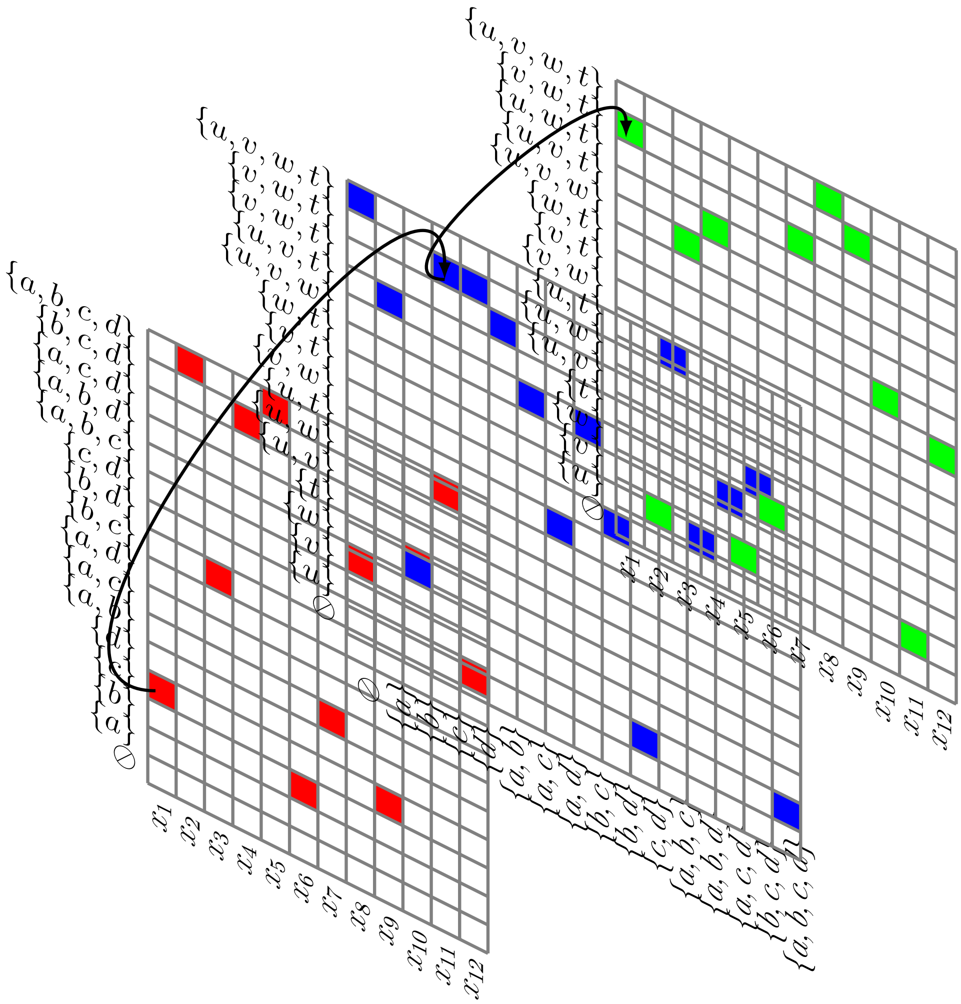

Рисунок 6 - Суперпозиция сет-функций.

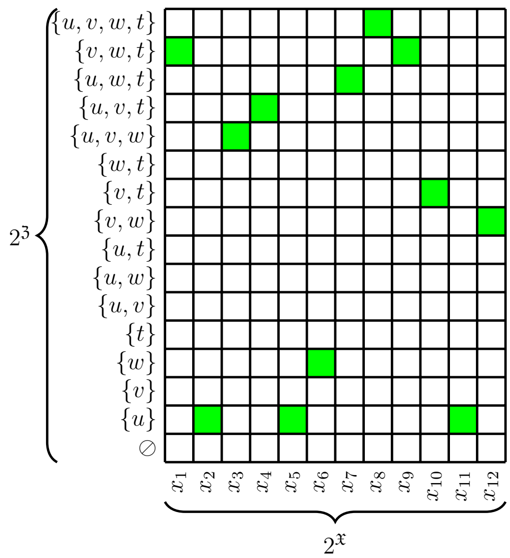

Рисунок 7 - График сет-фукции $\Psi(X)$, $X \subseteq \mathfrak{X}$.

потребительских предпочтений между фирмами, выпускающими косметическую продукцию.

3 Предобработка данных для задачи сет-регрессии

Подход к построению сет-регрессии в виде условных сет-средних на практике в большинстве случаев неприемлем. Это связано с тем, что этого необходимо знать либо совместное распределение двух случайных множеств, либо его статистическую оценку. Однако чтобы статистически оценить распределение случайного множества N событий на основе наблюдений приходится сталкиваться с проблемой хронической недостаточности имеющегося количества наблюдений для надлежащей оценки всех 2^N вероятностей. Оптимизация данных, как элемент предобработки, включает снижение размерности входных данных. В диссертации для предобработки данных для задачи сет-регрессии предлагается использовать рекуррентный метод построения вероятностных распределений случайных множеств событий с помощью аппарата ассоциативных функций, предложенный в [13, 16].

3.1 Рекуррентный метод построения распределений случайных множеств событий с помощью аппарата ассоциативных функций

Одной из главных проблем исследования случайных множеств событий является задача построения их вероятностных распределений, описывающих все способы взаимодействия элементов между собой в моделируемом множестве. Однако на пути решения этой задачи стоит известная преграда «проклятия размерности», которая заключается в экспоненциальном росте размерности вероятностного распределения случайного множества событий при возрастании числа событий, образующих это множество.

В [13, 16] было предложено организовать процесс построения вероятностных распределений случайных множеств событий на основе рекуррентного соотношения, полученного с помощью аппарата ассоциативных функций [20]. Основная идея построения — выразить вероятности пересечений множества событий функционально через вероятности самих событий, что приводит к уменьшению числа параметров, необходимых для построения вероятностных распределений случайных множеств событий.

Исходя из известных вероятностей событий $p_x = \mathbf{P}(x), x \in \mathfrak{X}$, формирование вероятностей пересечения множеств событий p_x осуществлять последовательно согласно рекуррентной формуле [13, 16]:

$$\mathbf{P}\left(\bigcap_{x\in X} x\right) = AF\left(p_x, \mathbf{P}\left\{\bigcap_{y\in X/\{x\}} y\right\}\right), \forall X \subseteq \mathfrak{X},$$

где AF — ассоциативная функция.

В [13, 16] исследованы проблемы рекуррентного построения класса вероятностных распределений случайного множества на конечном множестве из N событий с помощью следующих ассоциативных функций

1. $AF(a,b) = a \cdot b$;
2. $AF(a,b) = \min\{a,b\}$;
3. $AF(a,b) = \max\{a+b-1, 0\}$;
4. $AF(a,b) = \ln\left(1 + \frac{\left(e^{-\lambda a}-1\right)\left(e^{-\lambda b}-1\right)}{e^{-\lambda}-1}\right)^{-\frac{1}{\lambda}}, \lambda \neq 0.$

В [13, 16] приводятся теоремы, устанавливающие условия построения, а также вид и условия существования полученных вероятностных распределений случайных множеств событий.

Приведем алгоритм рекуррентного метода построения вероятностных распределений случайных множеств событий с помощью аппарата ассоциативных функций (Алгоритм 3) и рассмотрим пример применения данного метода к задаче сет-регрессии.

Пример 4. Пусть задано совместное распределение двух случайных множеств событий K и L, значения которых содержаться в конечных мно-

Алгоритм 3 Алгоритм рекуррентного построения распределений случайных множеств ассоциативными функциями

Вход: N вероятностей событий p_x, $x \in \mathfrak{X}$ и вид ассоциативной функции $AF(a,b)$.

Выход: Сет-функция $\{f(X),\ X \subseteq \mathfrak{X}\}$ на конечном множестве событий $\mathfrak{X}$.

Положить $f(\emptyset) = 1$.
Для всех $x \in \mathfrak{X}$ **выполнять**
    положить $f(x) = p_x = \mathbf{P}(x)$.
Конец цикла
Для всех $X \in 2^{\mathfrak{X}}$, $|X| > 1$ **выполнять**
    положить $f(X) = \mathrm{AF}\left(p_x, x \in X\right) = AF\left(p_x, P\left\{\bigcap\limits_{y \in X/\{x\}} y\right\}\right)$.
Конец цикла

жествах $\mathfrak{X}$ и $\mathcal{Y}$ соответственно, где $\mathfrak{X} = \{a, b, c\}$ и $\mathcal{Y} = \{x, y\}$. Задается исходное совместное распределение двух случайных множеств событий K и L показанных в таблице 1.

Таблица 12 - Совместное распределение двух случайных множеств событий K и L, значения которых содержаться в конечных множествах $\mathfrak{X}$ и $\mathfrak{Y}$ соответственно.

$\mathfrak{Y}$ \ $\mathfrak{X}$	$\{\oslash\}$	$\{a\}$	$\{b\}$	$\{c\}$	$\{a,b\}$	$\{a,c\}$	$\{b,c\}$	$\{a,b,c\}$
$\{\oslash\}$	0,014	0,040	0,061	0,005	0,005	0,006	0,040	0,066
$\{x\}$	0,024	0,017	0,027	0,053	0,028	0,044	0,018	0,064
$\{y\}$	0,040	0,068	0,027	0,005	0,033	0,023	0,071	0,056
$\{x,y\}$	0,017	0,022	0,008	0,020	0,002	0,061	0,020	0,013

На рисунке 8 показано решение задачи сет-регрессии для совместного распределения двух случайных множеств событий K и L, представленного в таблице 12.

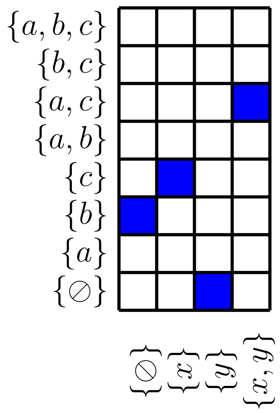

Рисунок 8 - Сет-регрессия в виде сет-квантиля с уровнем $\alpha = 3/5$ для совместного распределения двух случайных множеств событий K и L, представленного в таблице 12.

Из исходного совместного распределения найдем N вероятностей событий $p_x = 0.440, p_y = 0.489$, $x, y \in \mathcal{Y}$, $p_a = 0.550, p_b = 0.540, p_c = 0.566$, $a, b, c \in \mathfrak{X}$. Применим рекуррентный метод с указанными выше ассоциативными функциями. Получим новые вероятностные распределения II-го рода таблицы 13 – 16.

Таблица 13 - Распределение II-рода с ассоциативной функцией $AF\left(a, b\right) = a{\cdot}b$.

$\mathfrak{Y}$ \ $\mathfrak{X}$	$\{\emptyset\}$	$\{a\}$	$\{b\}$	$\{c\}$	$\{a,b\}$	$\{a,c\}$	$\{b,c\}$	$\{a,b,c\}$
$\{\emptyset\}$	1,000	0,550	0,540	0,566	0,297	0,311	0,305	0,168
$\{x\}$	0,440	0,242	0,237	0,249	0,131	0,137	0,134	0,074
$\{y\}$	0,489	0,269	0,264	0,276	0,145	0,152	0,149	0,082
$\{x,y\}$	0,215	0,118	0,116	0,122	0,064	0,067	0,066	0,036

Применим формулу обращения Мёбиуса (1.5) для всех $X \in 2^{\mathfrak{X}}$, перейдем от распределения II-рода к распределению I-рода, тем самым получим:

– совместное распределение двух случайных множеств событий K и L с

Таблица 14 - Распределение II-рода с ассоциативной функцией $AF\left(a,b\right) = \min\left(a,b\right)$.

$\mathfrak{Y} \backslash \mathfrak{X}$	$\{\oslash\}$	$\{a\}$	$\{b\}$	$\{c\}$	$\{a,b\}$	$\{a,c\}$	$\{b,c\}$	$\{a,b,c\}$
$\{\oslash\}$	1,000	0,550	0,540	0,566	0,540	0,550	0,540	0,540
$\{x\}$	0,440	0,440	0,440	0,440	0,440	0,440	0,440	0,440
$\{y\}$	0,489	0,489	0,489	0,489	0,489	0,489	0,489	0,489
$\{x,y\}$	0,440	0,440	0,440	0,440	0,440	0,440	0,440	0,440

Таблица 15 - Распределение II-рода с ассоциативной функцией $AF\left(a,b\right) = \max\left(a,b\right)$.

$\mathfrak{Y} \backslash \mathfrak{X}$	$\{\oslash\}$	$\{a\}$	$\{b\}$	$\{c\}$	$\{a,b\}$	$\{a,c\}$	$\{b,c\}$	$\{a,b,c\}$
$\{\oslash\}$	1,000	0,550	0,540	0,566	0,090	0,116	0,106	0,000
$\{x\}$	0,440	0,000	0,000	0,006	0,000	0,000	0,000	0,000
$\{y\}$	0,489	0,039	0,028	0,054	0,000	0,000	0,000	0,000
$\{x,y\}$	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Таблица 16 - Распределение II-рода с ассоциативной функцией $AF\left(a,b\right) = \ln\left(1+\frac{\left(e^{-\lambda a}-1\right)\left(e^{-\lambda b}-1\right)}{e^{-\lambda}-1}\right)^{-\frac{1}{\lambda}}$.

$\mathfrak{Y} \backslash \mathfrak{X}$	$\{\oslash\}$	$\{a\}$	$\{b\}$	$\{c\}$	$\{a,b\}$	$\{a,c\}$	$\{b,c\}$	$\{a,b,c\}$
$\{\oslash\}$	1,000	0,550	0,540	0,566	0,287	0,301	0,295	0,154
$\{x\}$	0,440	0,232	0,227	0,239	0,118	0,124	0,121	0,062
$\{y\}$	0,489	0,258	0,253	0,266	0,132	0,138	0,136	0,070
$\{x,y\}$	0,205	0,106	0,104	0,109	0,053	0,056	0,055	0,028

ассоциативной функцией $AF\left(a,b\right) = a \cdot b$ (таблица 17);

– совместное распределение двух случайных множеств событий K и L с ассоциативной функцией $AF(a,b) = \max\{a,b\}$ (таблица 18);

– совместное распределение двух случайных множеств событий K и L — с ассоциативной функцией $AF(a,b) = \ln\left(1 + \frac{(e^{-\lambda a}-1)(e^{-\lambda b}-1)}{e^{-\lambda}-1}\right)^{-\frac{1}{\lambda}}$) (таблица 19).

Таблица 17 - Совместное распределение задано с помощью рекуррентного метода с ассоциативной функцией $AF(a,b) = a \cdot b$.

$\mathfrak{Y} \backslash \mathfrak{X}$	$\{\oslash\}$	$\{a\}$	$\{b\}$	$\{c\}$	$\{a,b\}$	$\{a,c\}$	$\{b,c\}$	$\{a,b,c\}$
$\{\oslash\}$	0,026	0,031	0,030	0,034	0,037	0,041	0,039	0,048
$\{x\}$	0,020	0,025	0,024	0,026	0,029	0,032	0,031	0,038
$\{y\}$	0,025	0,030	0,029	0,032	0,035	0,039	0,038	0,046
$\{x,y\}$	0,019	0,024	0,023	0,025	0,028	0,031	0,030	0,036

Таблица 18 - Совместное распределение задано с помощью рекуррентного метода с ассоциативной функцией $AF(a,b) = \min\{a,b\}$.

$\mathfrak{Y} \backslash \mathfrak{X}$	$\{\oslash\}$	$\{a\}$	$\{b\}$	$\{c\}$	$\{a,b\}$	$\{a,c\}$	$\{b,c\}$	$\{a,b,c\}$
$\{\oslash\}$	0,434	0,000	0,000	0,016	0,000	0,010	0,000	0,051
$\{x\}$	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
$\{y\}$	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,049
$\{x,y\}$	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,440

На рисунках 9, 10 представлены решения задачи сет-регрессии для совместного распределения заданного с помощью рекуррентного метода с различными ассоциативными функциями.

Таблица 19 - Совместное распределение задано с помощью рекуррентного метода с ассоциативной функцией $AF(a,b) = \ln\left(1 + \frac{\left(e^{-\lambda a}-1\right)\left(e^{-\lambda b}-1\right)}{e^{-\lambda}-1}\right)^{-\frac{1}{\lambda}}$.

$\mathfrak{Y}$ \ $\mathfrak{X}$	$\{\varnothing\}$	$\{a\}$	$\{b\}$	$\{c\}$	$\{a,b\}$	$\{a,c\}$	$\{b,c\}$	$\{a,b,c\}$
$\{\varnothing\}$	0,007	0,030	0,029	0,032	0,041	0,045	0,043	0,050
$\{x\}$	0,021	0,028	0,027	0,030	0,030	0,034	0,032	0,034
$\{y\}$	0,024	0,034	0,032	0,036	0,037	0,041	0,039	0,042
$\{x,y\}$	0,022	0,025	0,024	0,026	0,025	0,028	0,027	0,028

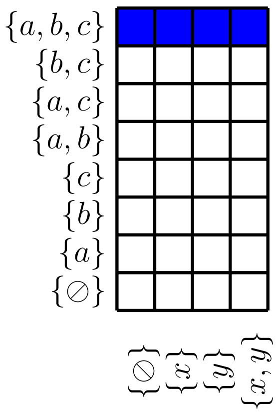

Рисунок 9 - Сет-регрессия в виде сет-квантиля с уровнем $\alpha = 1/4$, совместное распределение задано с помощью рекуррентного метода с ассоциативной функцией $AF(a,b) = a \cdot b$

В диссертационной работе используются ассоциативные функции 1 и 4, потому что в работе [16] показано, что полученные рекуррентным методом распределения с применением ассоциативных функций 2 и 3 являются не полными вероятностными распределениями. Следовательно при решении задачи сет-регрессии, а точнее при определение маргинального распределения будем получать нули. А это в свое время не дает определить условное распределение, так как происходит деление на ноль.

Зная вероятностное распределение II-го рода и применяя рекуррентный метод построения вероятностных распределений случайных множеств собы-

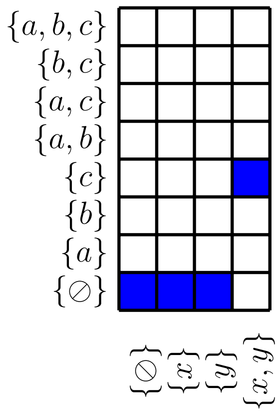

Рисунок 10 - Сет-регрессия в виде сет-квантиля с уровнем $\alpha = 1/4$, совместное распределение задано с помощью рекуррентного метода с ассоциативной функцией $AF\left(a, b\right) = \ln\left(1 + \frac{\left(e^{-\lambda a}-1\right)\left(e^{-\lambda b}-1\right)}{e^{-\lambda}-1}\right)^{-\frac{1}{\lambda}}$

тий с помощью аппарата ассоциативных функций для решения задачи сет-регрессии. В построенном для нее алгоритме можем опустить первый и второй шаг. Для этого приведем и докажем теорему о виде условного распределения случайного множества событий с использованием ассоциативных функций.

Теорема 3.1. *Пусть совместное распределение II-го рода случайных множеств K и L, заданных на непересекающихся множествах событий $\mathfrak{X}$ и $\mathfrak{Y}$ соответственно, построено рекуррентным методом*

$$\{p_{Z/\!\!/\mathcal{X}+\mathcal{Y}}, Z \subseteq \mathfrak{X} + \mathcal{Y}\},$$

где $p_{Z/\!\!/\mathfrak{X}+\mathcal{Y}} = f\left(Z\right) = AF\left(p_x, x \in Z\right)$.

Тогда условное распределение случайного множества событий L при условии наступлении события $\mathrm{ter}\left(X\right)$ *— это набор $\left\{p\left(Y|_{ter(X)}\right), Y \subseteq \mathfrak{Y}\right\}$, где*

$$p\left(Y|_{\mathrm{ter}(X)}\right) = \frac{\sum\limits_{X\subseteq W}\sum\limits_{Y\subseteq U}\left(-1\right)^{|V|-|Z|} f\left(f\left(W\right), f\left(U\right)\right)}{\sum\limits_{Y\subseteq\mathfrak{Y}}\sum\limits_{X\subseteq W}\sum\limits_{Y\subseteq U}\left(-1\right)^{|V|-|Z|} f\left(f\left(W\right), f\left(U\right)\right)}, \tag{3.19}$$

где $Z = X + Y, X \in \mathfrak{X}, Y \in \mathfrak{Y}, V = W + U, X \subseteq W, Y \subseteq U$.

Доказательство: Доказательство изъято по решению правообладателя.

Теорема 3.2. *Пусть совместное распределение случайных множеств K и L, определенных на непересекающихся множествах событий $\mathfrak{X}$ и $\mathfrak{Y}$ соответственно, задано рекуррентным методом с ассоциативной функцией Франка.Тогда условное распределение случайного множества событий $L|_X$ при условии наступлении события* $\mathrm{ter}\left(X\right)$ *— это набор* $\left\{p\left(Y|_{\mathrm{ter}(X)}\right), Y \subseteq \mathfrak{Y}\right\}$, *где*

$$p\left(Y|_{ter(X)}\right) = \frac{\ln \prod\limits_{Z \in 2^{\mathfrak{X}+\mathfrak{Y}}:(X+Y) \subseteq Z} \left(1 + \frac{\prod\limits_{x \in Z}\left(e^{-\lambda p_x}-1\right)}{e^{-\lambda}-1}\right)^{-1/\lambda}}{\sum\limits_{V \subseteq \mathfrak{Y}} \ln \prod\limits_{Z \in 2^{\mathfrak{X}+\mathfrak{Y}}:(X+V) \subseteq Z} \left(1 + \frac{\prod\limits_{x \in Z}\left(e^{-\lambda p_x}-1\right)}{e^{-\lambda}-1}\right)^{-1/\lambda}}. \qquad (3.20)$$

Доказательство: Доказательство изъято по решению правообладателя.

3.2 Пример сет-регрессионного анализа медицинских данных

Нередко при обработке медицинских данных возникает ситуация, когда исходная информация об объекте исследования представлена бинарными признаками. В этом случае, адекватной математической моделью данных являются случайные множества, относящиеся к одному из объектов статистики нечисловой природы.

Сет-регрессионный анализ в медицинских исследованиях может быть применен для прогнозирования: прогноз степени тяжести заболевания по множеству наблюдаемых у пациента симптомов; прогноз множества осложнений, возможных у пациента по множеству диагностических симптомов; прогнозирование процесса течения некоторой болезни, когда состояние пациента характеризуется набором дихотомических признаков; прогноз некоторого количественного показателя по множеству симптомов болезни.

Рассмотрим применение рекуррентного метода в задаче сет-регрессии на примере анализа медицинских данных, предоставленных кафедрой медицинской кибернетики Красноярского государственного медицинского университета.

Пример 5.

Дано: Пусть $\{a, b, c, x, y\}$ = {хронические заболевания, сахарный диабет, заболевания сердца, повышение веса, пассивный образ жизни}.

- Обозначим события $\mathfrak{X} = \{a, b, c\}$, $\mathfrak{X} \subset \mathcal{F}$ как множество заболеваний.
- Обозначим события $\mathfrak{Y} = \{x, y\}$, $\mathfrak{Y} \subset \mathcal{F}$ как образ жизни.
- Получена статистическая оценка вероятностей событий $p_a = 0.4734, p_b = 0.1739, p_c = 0.2077, p_x = 0.1787, p_y = 0.5797$;
- Для рекуррентного метода используется ассоциативная функция Франка

$$AF(a, b) = \ln\left(1 + \frac{\left(e^{-\lambda a} - 1\right)\left(e^{-\lambda b} - 1\right)}{e^{-\lambda} - 1}\right)^{-\frac{1}{\lambda}}.$$

Найти: функцию сет-регрессии $\Psi : 2^{\mathfrak{Y}} \to 2^{\mathfrak{X}}$.

В таблице 20 представлен набор условных распределений, полученных с помощью теоремы 3.2. На рисунке 11—13 представлен результат решения задачи сет-регрессии в виде условного сет-квантиля.

Таблица 20 - Условные распределения случайного множества K при условии, что наступило множество событий $Y \subseteq \mathfrak{Y}$ для примера 5.

$\mathfrak{Y}$ \ $\mathfrak{X}$	$\{\oslash\}$	$\{a\}$	$\{b\}$	$\{c\}$	$\{a,b\}$	$\{a,c\}$	$\{b,c\}$	$\{a,b,c\}$
$\{\oslash\}$	0,1339	0,0940	0,0843	0,1849	0,0215	0,0269	0,0062	0,0061
$\{x\}$	0,2852	0,0223	0,0052	0,0064	0,0051	0,0063	0,0015	0,0015
$\{y\}$	0,2198	0,1452	0,0331	0,0414	0,0324	0,0404	0,0094	0,0093
$\{x,y\}$	0,0343	0,0335	0,0078	0,0097	0,0077	0,0096	0,0023	0,0022

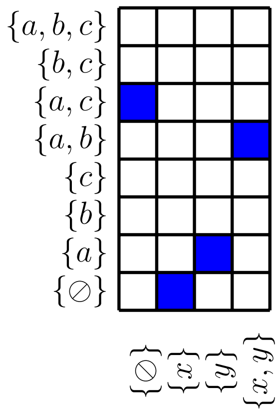

Рисунок 11 - Сет-регрессия в виде сет-квантиля с уровнем $\alpha = 1/4$, совместное распределение задано с помощью рекуррентного метода с ассоциативной функцией $AF\left(a,b\right) = \ln\left(1 + \frac{\left(e^{-\lambda a}-1\right)\left(e^{-\lambda b}-1\right)}{e^{-\lambda}-1}\right)^{-\frac{1}{\lambda}}$

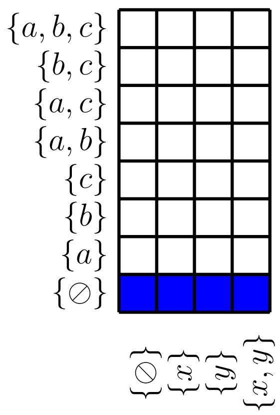

Рисунок 12 - Сет-регрессия в виде сет-медианы, совместное распределение задано с помощью рекуррентного метода с ассоциативной функцией $AF\left(a,b\right) = \ln\left(1 + \frac{\left(e^{-\lambda a}-1\right)\left(e^{-\lambda b}-1\right)}{e^{-\lambda}-1}\right)^{-\frac{1}{\lambda}}$

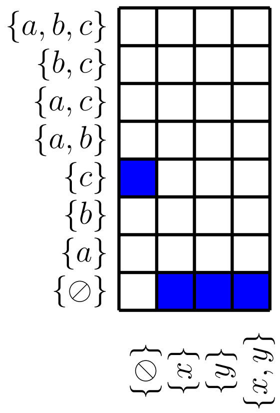

Рисунок 13 - Сет-регрессия в виде сет-моды, совместное распределение задано с помощью рекуррентного метода с ассоциативной функцией $AF\left(a,b\right)=\ln\left(1+\frac{\left(e^{-\lambda a}-1\right)\left(e^{-\lambda b}-1\right)}{e^{-\lambda}-1}\right)^{-\frac{1}{\lambda}}$

Резюме. В данной главе для предобработки данных для задачи сет-регрессии предлагается использовать рекуррентный метод построения вероятностных распределений случайных множеств событий с помощью аппарата ассоциативных функций. Доказаны две теоремы о виде условного распределения случайного множества, полученного рекуррентным методом с помощью аппарата ассоциативных функций. Разработан алгоритм рекуррентного построения вероятностных распределений для случайных множеств событий. Рассмотрен пример сет-регрессионного анализа медицинских данных.

Заключение

Основные результаты. В результате проделанной работы были получены следующие результаты.

1. Разработан алгоритм решения задачи сет-регрессии случайных множеств в классической постановке.

2. Поставлена и решена задача построения суперпозиции сет-регрессий.

3. Для предобработки входных данных предложено использовать рекуррентный метод построения распределений случайных множеств событий ассоциативными функциями. Доказана теорема о виде условного распределения случайного множества событий с использованием ассоциативных функций. Исследовано решение задачи сет-регрессии с использованием ассоциативной функции Франка.

4. Приведены примеры использования сет-регрессионных моделей для анализа зависимостей в экономических и медицинских исследованиях.

Апробация работы. Основные результаты и отдельные положения диссертационных исследований докладывались и обсуждались на научных и научно-практических конференциях международного и всероссийского уровней.

Публикации. По результатам научных исследований имеются следующие публикации в трудах конференций [5], [6], [7], [8], [9], [14], [15].

Список использованных источников

1. Белов, К. А. Методы и алгоритмы случайно-множественного анализа медицинских данных: диссертация кандидата технических наук : 05.13.01 / Константин Андреевич Белов. — Воронеж, 2005. — 121 с.

2. Воробьёв, О.Ю. Сет-регрессионный анализ зависимостей событий в статистических системах: учеб. пособие / О.Ю. Воробьев, А.Ю. Фомин — Красноярск: ИВМ СО РАНб КрасГУ, 2004. — 116 с.

3. Воробьёв, О.Ю. Эвентология:учеб. пособие / О. Ю. Воробьев — Красноярск: Сиб. фед. ун-т., 2007. — 435 с.

4. Прохоров, Ю.В. Вероятность и математическая статистика: Энциклопедия / Гл. ред. Ю.В.Прохоров. — Москва: Большая Российская энциклопедия, 1999. — 914 с.

5. Иванова, А. И. Препроцессинг данных для задачи сет-регрессии /А. И. Иванова // Проспект Свободный-2016: материалы науч. Конф., [Электронный ресурс] — Красноярск: Сиб. Федер. Ун-т, том «Математика, информатика: теория вероятностей, математическая статистика и финансово-актуарная математика», 2016. — С. 21-23.

6. Иванова, А. И. Суперпозиция сет-регрессий / А. И. Иванова // Проспект Свободный-2015: материалы науч. Конф., посвященной 70-летию Великой Победы [Электронный ресурс] — Красноярск: Сиб. Федер. Ун-т, том «Математика, информатика: теория вероятностей, математическая статистика и финансово-актуарная математика», 2015. — С. 21-23.

7. Иванова, А.И. О задаче сет-регрессии / А.И. Иванова, М.А. Атабеков // Труды XIV конференции по финансово-актуарной математике и эвентологии многомерной статистики. — Красноярск, 2015 — С. 285-292.

8. Иванова, А.И. Эвентологическая регрессия случайных множеств событий в виде сет-средних / А.И. Иванова // ”Молодежь и наука”: сборник материалов X Юбилейной Всероссийской научно-технической конференции студентов, аспирантов и молодых ученых с международным участием, посвященной 80-летию образования Красноярского края, [Электронный ресурс], № заказа 1644/отв. ред. О. А. Краев — Красноярск: Сиб. федер. ун-т., 2014

9. Иванова, А.И. Эвентологические регрессии случайных множеств событий в виде условного сет-квантиля / А.И. Иванова // Перспективы развития фундаментальных наук [Электронный ресурс]: труды XII Международной конференции студентов и молодых учёных. Россия, Томск, 21–24 апреля 2015 г. / под ред. Е.А. Вайтулевич. — Национальный Исследовательский Томский политехнический университет, 2015. — С. 597-599.

10. Кендалл, М. Дж. Теория распределений./ М. Дж. Кендалл, А.Стьюарт. — Москва: Наука, Физматлит, 1966. — 588 с.

11. Кобзарь, А.И. Прикладная математическая статистика. Для инженерных и научных специальносте./ А.И. Кобзарь. — Москва: Физматлит, 2006.—816 с.

12. Лукьянова, Н.А. Визуализация средневероятного события через матричное представление террасных событий /Лукьянова Н.А., Я.В.Нартов, Д.В.Семенова. // Труды XVI международной конференции по эвентологической математики и смежным вопросам— Красноярск: СФУ (под ред. Олега Воробьева), 2012. — С. 145-158.

13. Лукьянова, Н. А. Ассоциативные функции Франка в построении семейств дискретных вероятностных распределений случайных множеств событий / Н.А. Лукьянова, Д. В. Семенова // Прикладная дискретная математика, № 2(32), 2016. — С. 5-19.

14. Семенова, Д.В. Сравнение нечеткого и событийного подхода в задаче сегментации рынка /Д.В. Семенова, А.И. Иванова // Новые информационные технологии в исследовании сложных структур: Материалы Десятой российской конференции с международным участием. — Томск: Издательский Дом Томского государственного университета, 2014. — С. 128.

15. Семенова, Д.В. Нечеткая и сет-регрессионная модели распределения потребительских предпочтений между фирмами / Д.В. Семенова // Труды XIII ФАМЭМС'2014 конференции (под ред. О.Ю. Воробьёва). — Красноярск: СФУ, 2014. — С. 222-227.

16. Семенова, Д.В., Рекуррентное построение дискретных вероятностных распределений случайных множеств событий / Д. В. Семенова, Н.А. Лукьянова // Прикладная дискретная математика. № 4(26), 2014. — С. 47-58.

17. Орлов, А.И. Нечисловая статистика/ А.И. Орлов — Москва:МЗ-Пресс, 2014. — 513 с.

18. Тарасова, О. Ю. Сеточные и регрессионные алгоритмы аппроксимации сложных систем событий: автореф. дис. на соиск. учен. степ. канд. физ.-мат. н. по специальности 05.13.01 / Ольга Юрьевна Тарасова —- Красноярск, 2007. — 24 с.

19. Ширяев, А.Н. Вероятность: в 2-ч кн.— 4-е изд. / А.Н. Ширяев. — М.: МЦНМО, 2004. — 968 с.

20. Alsina, S. Associative functions: Triangular Norms and Copulas. / S. Alsina , M. Frank , B. Scueizer — Singapore: World Scientific Pubishing Co.Pte. Ltd., 2006. — 237 p.

21. Lukyanova, N.A. Eventological subdistributions: matrix representation of the events-terraces for a set of events / N.A. Lukyanova // Труды XV Междуна-

родной ЭМ’2011 конференция. 10–11 декабря 2011 — Красноярск: КГТЭИ, СФУ, 2011. — С. 20–26.

22. Han, J. Data Mining: Concepts and Techniques / J. Han , M. Kamber. — Elsevier Inc., 2006. —

23. Molchanov, I. The Theory of Random Sets Springer / I. Molchanov. — New York, 2011.— 488 p.

24. Nguyen, H. T. An Introduction to Random/ H. T. Nguyen. — Sets Taylor Francis Group, LLC, 2006. — 240 p.

25. Frank, M.J. On the simultaneous associative of $F(x, y)$ и $x+y-F(x, y)$/ M.J. Frank — Aequatines Math.№19, 1979 — 194-226 p.