Информационные процессы, Том 15, № 4, 2015, стр. 402–413

МАТЕМАТИЧЕСКИЕ МОДЕЛИ, ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ

Описание пространства дизайна с помощью экстремальных эллипсоидов в задачах представления данных[1]

А.А. Бедринцев, В.В. Чепыжов

Институт проблем передачи информации, Российская академия наук, Москва, Россия
e-mail: chep@iitp.ru

Поступила в редколлегию 10.11.2015

Аннотация—Рассмотрены задачи описания множества данных с помощью построения оптимального эллипсоида. Оптимизационные задачи сформулированы в виде задач выпуклого программирования с использованием линейных матричных неравенств. Предложенные методы сравнены с похожими ранее разработанными методами по двум критериям: объему эллипсоида и количеству точек в обучающей выборке, которые лежат вне эллипсоида.

КЛЮЧЕВЫЕ СЛОВА: экстремальные эллипсоиды, пространство дизайна, линейные матричные неравенства, выпуклое программирование.

1. ВВЕДЕНИЕ

В последнее время одними из главных средств при разработке сложных технических объектов являются математическое моделирование и анализ данных. Объект параметризуется многомерным вектором $x \in \mathbf{R}^d$. Компоненты вектора x включают геометрические и физические характеристики, параметры внешней среды и параметры его функционирования. Рассматривается некоторая характеристика $Y = F(x)$ объекта, и объект считается оптимальным, если его характеристика $Y(x)$ принимает наибольшее или наименьшее значение среди всех других допустимых объектов x. Другими словами, задача выбора наилучшего технического решения может быть сформулирована как оптимизация некоторой функции $F(x)$ на множестве объектов x.

Чтобы сформулировать оптимизационную задачу, необходимо задать ограничения на вектор x. Очевидно, что не все наборы из d вещественных чисел соответствуют корректному и физически осмысленному объекту. Для значений отдельных параметров инженеры могут указать интервалы допустимых значений, известные из предметной области. Заданная таким образом область – параллелипипед $\Pi = \{x \in \mathbf{R}^d | l^i \le x^i \le u^i, i = 1 \dots d\}$ – может содержать точки, далекие от известных векторов x, описывающих физически корректные объекты. Весьма вероятно, что эти точки не будут соответствовать корректным объектам.

Достаточно часто встречается ситуация, когда отдельные координаты вектора x не имеют самостоятельного значения и смысла, и, поэтому, сложно задать для них индивидуальные информативные ограничения. Вектор x может включать детальное описание поверхности объекта, которое состоит из координат точек на сетке, наложенной на поверхность объекта. Такие описания широко применяются в CAD-системах, программах для расчета аэродинамических характеристик и при визуализации.

В процессе разработки инженеры, выполняя эксперименты, собирают базу данных цифровых описаний объектов и параметров экспериментов $X = \{x_i \in \mathbf{R}^d, i = 1 \dots N\}$. Векторы

[1] Работа выполнена при поддержке Российского научного фонда (проект № 14-50-00150).

из множества X описывают реальные объекты. Желательно строить описание области данных (пространства дизайна) с корректными векторами, основываясь на известных векторах из множества X. Такие области должны обладать следующими свойствами. Во-первых, они должны содержать как можно больше векторов из X. Иначе будет потеряна важная информация о большом количестве корректных объектов. Во-вторых, пространство дизайна должно иметь небольшой объем, чтобы не содержать точки, удаленные от точек из X и которые с большой вероятностью не соответствуют физически осмысленному объекту.

Построенная таким образом область может быть использована в качестве ограничений на значения x в оптимизационных задачах. Это вызывает несколько дополнительных требований на область. Выпуклая оптимизация высоко развита, и задачи выпуклого программирования имеют важные свойства, такие как существование и единственность глобального минимума, локальный минимум является глобальным [1]. Для решения выпуклых задач разработаны эффективные алгоритмы и доступны их качественные реализации на многих языках программирования. Чтобы задача была выпуклой, необходимо, чтобы ограничения определяли бы выпуклое пространство дизайна. Поэтому в данной статье будет строится выпуклое описание множества данных X. Также важна простота описания области и легкость генерации случайных точек в ней. В работе предлагается искать описание пространства дизайна в виде эллипсоида [2, 3].

Другим применением описания области данных является задача определение выбросов. Выбросы – это точки, которые существенно отличаются от обычных точек. Строится описание множества X ординарных, т.е. нормальных точек с помощью эллипсоида небольшого объема. Если новая точка принадлежит ему, она классифицируется как ординарная. В обратном случае, делается вывод, что это выброс.

Обычно аномалии крайне редки, и обучающие выборки для задач детектирования выбросов содержат небольшое количество отрицательных примеров. В похожей задаче определения новизны обучающие выборки состоят исключительно из ординарных точек [4]. Таким образом, необходимо построить геометрическое тело, содержащее большинство точек из обучающей выборки, чтобы минимизировать вероятность ложных срабатываний. Но это тело должно иметь небольшой объем, так как иначе выбросы часто будут приниматься за ординарные точки. Задачи детектирования аномалий и обнаружения новизны широко применяются в статистике, задачах построения моделей для кредитного скоринга, автоматизированном детектировании мошеннических действий [5, 6].

Статья [7] дает еще одно применение эллипсоидов минимального объема. Доказано, что центр эллипсоида минимального объема, который содержит более половины точек множества X, является устойчивой аффинно-инвариантой оценкой местоположения (*location estimation*). Статьи [8] и [9] рассматривают другие свойства этой оценки: скорость сходимости, состоятельность и непрерывность по отношению к распределению, с помощью которого сгенерировано множество X. Главный недостаток этой оценки это высокая вычислительная сложность точного вычисления. В статье [10] описывается приближенный метод аппроксимации оценки местоположения с помощью минимального эллипсоида, которые имеет разумную сложность в пространствах небольшой размерности.

Настоящая статья организована следующим образом. Раздел 2 содержит формальную постановку задачи и доказательство существования точного решения с указанием наивного, но ресурсоемкого алгоритма. В разделе 3 приведен обзор двух известных методов для получения приближенного решения задачи. В разделе 4 мы предлагаем два обобщения известных методов. Будет показано, что некоторые свойства ранее известных методов сохраняются при обобщении. Это позволяет эффективно применять новые методы при практическом моделировании. Раздел 5 содержит результаты численных экспериментов на искусственных наборах

данных с разными статистическими характеристиками. Последний раздел 6 подводит итоги статьи.

2. ПОСТАНОВКА ЗАДАЧИ

Пусть дана выборка $X = \{x_i \in \mathbf{R}^d, i = 1 \dots N\}$ векторов из d-мерного пространства. Пусть E – эллипсоид, который предстоит построить для описания множества X. Эллипсоид может быть задан с помощью его центра $a \in \mathbf{R}^d$, симметричной положительно определенной матрицы $P = P^T \succ 0$ размера $d \times d$ и квадрата эффективного радиуса $R \geq 0$ в виде следующего неравенства:

$$E = \{x \in \mathbf{R}^d | \, (x-a)^T P^{-1} (x-a) \leq R\} \ . \tag{1}$$

Обозначим через $Vol\,(E)$ объем эллипсоида E, который вычисляется по формуле:

$$Vol\,(E) = w_d \sqrt{\det P} R^{\frac{d}{2}} \ , \tag{2}$$

где $w_d = \dfrac{\pi^{d/2}}{\Gamma(d/2+1)}$ – объем единичного шара в пространстве $\mathbf{R}^d$.

Введем дискретную функцию количества точек из множества X, которые не принадлежат эллипсоиду E:

$$K\,(E) = \#\{x \in X | x \notin E\} \ . \tag{3}$$

Поставим следующую двухкритериальную задачу оптимизации поиска эллипсоида небольшого объема, содержащего много точек из заданного множества X:

$$\min_E \left(Vol\,(E), K\,(E)\right) \ . \tag{4}$$

Невозможно уменьшить до нуля объем эллипсоида, не оставив за его пределами практически все точки из обучающего множества X. Поэтому решением задачи будет набор оптимальных по Парето эллипсоидов. Это множество эллипсоидов, которые не доминируют друг друга. Говорят, что эллипсоид E_1 доминирует эллипсоид E_2 тогда и только тогда, когда $K\,(E_1) \leq K\,(E_2)$ и $Vol\,(E_1) \leq Vol\,(E_2)$, и хотя бы одно из этих неравенств выполняется строго. Фронт Парето задачи образован парами объема (2) и количества (3) выбросов среди точек обучающего множества данных, посчитанных для Парето-оптимальных эллипсоидов. Выбор конкретного эллипсоида делается экспертом из множества оптимальных по Парето в зависимости от потребностей конкретной задачи анализа данных.

Следующая теорема устанавливает принципиальную разрешимость задачи (4).

Теорема 1. *Парето-фронт и набор оптимальных по Парето эллипсоидов для задачи* (4) *существуют и могут быть найдены.*

Доказательство. Перебирая все подмножества $G \subset X$, решая задачи

$$\begin{aligned} &\min_E \left(Vol\,(E)\right) \\ &s.t.\ G \subset E \end{aligned} \tag{5}$$

и выбирая среди решений (5) набор недоминируемых эллипсоидов, очевидно, можно решить задачу (4). Заметим, что ограничение $G \subset E$ и целевая функция $Vol\,(E)$ могут быть выражены в выпуклой форме [1]. Следовательно, каждая задача (5) имеет единственное решение если выпуклая оболочка множества G имеет положительный объем. Чтобы найти фронт Парето, нужно выбрать эллипсоид с минимальным объемом, который содержит ровно j точек из X

для каждого значения $j = 1 \dots N$, т.е. минимум ограниченного снизу множества (т.к. объем не отрицателен).

Возможно, что для некоторой точки (V, K) на фронте Парето есть два (или более) различных оптимальных эллипсоида $E_1 \neq E_2$, для которых $Vol\,(E_1) = Vol\,(E_2)$ и $K\,(E_1) = K\,(E_2)$. Так как они не доминируют друг друга, то оба включаются в набор Парето-оптимальных эллипсоидов задачи (4). □

3. ИЗВЕСТНЫЕ ПРИБЛИЖЕННЫЕ МЕТОДЫ

Использованный в доказательстве теоремы 1 метод нереализуем на практике. Он требует решения приблительно 2^N (количество подмножеств $G \subseteq X$) задач вида (5). В настоящей статье описываются методы приближенного решения задачи (4).

Мы можем заменить объем эллипсоида любой монотонно возрастающей функцией $\phi\,(Vol\,(E))$. Следующая задача оптимизации эквивалентна задаче (4):

$$\min_E \left(\phi\,(Vol\,(E)), K\,(E)\right) \quad .$$

В статьях, которые будут указаны далее в этом разделе, вместо решения задачи (4) авторы заменяют дискретную целевую функцию $K\,(E)$ другой непрерывной выпуклой функцией. Например, суммой величин ξ_i, где неотрицательная величина $\xi_i \geq 0$ есть мера удаленности точки x_i от эллипсоида. Эта мера удаленности может выбираться разными способами в зависимости от условий задачи. Если $x_i \in E$, то $\xi_i = 0$. Обозначая через $\xi = \{\xi_i\}_{i=1}^N$ вектор всех ξ_i, вместо задачи (4) получаем следующую формулировку:

$$\min_{E,\xi} \left(\phi\,(Vol\,(E)), \sum_{i=1}^N \xi_i\right) \quad . \tag{6}$$

Метод скаляризации – общепринятый способ решения многокритериальных задач оптимизации [1]. Для любой неотрицательной константы $C \geq 0$ рассмотрим следующую задачу с одной целевой функцией:

$$\min_{E,\xi} \phi\,(Vol\,(E)) + C \sum_{i=1}^N \xi_i \quad .$$

Для выпуклых задач (если все целевые функции являются выпуклыми и ограничения задают выпуклое множество в пространстве ξ_i и параметров эллипсоида P, a, R), изменяя параметр скаляризации C в диапазоне $[0; \infty)$, можно получить все оптимальные по Парето эллипсоиды [1]. Для невыпуклых задач скаляризация даст подмножество фронта Парето задачи (6).

При фиксированной матрице P эллипсоида или при фиксированном значении определителя $\det P$ объем есть функция эффективного радиуса эллипсоида, поэтому возьмем $\phi\,(Vol\,(E)) = R$.

В статье [11] описание области строится в форме шара радиуса $\sqrt{R}$. Точки из обучающего множества x_i лежат в шаре радиуса $\sqrt{R + \xi_i}$ для некоторого $\xi_i \geq 0$. Таким образом, матрица эллипсоида P является единичной, и скаляризованная форма задачи описания множества данных имеет следующий вид:

$$\begin{aligned}&\min_{R,a,\xi} R + C\sum_{i=1}^{N}\xi_i\\&s.t.\,(x_i-a)^T(x_i-a)\le R+\xi_i, i=1\ldots N,\\&R\ge 0,\xi_i\ge 0, i=1\ldots N\ .\end{aligned}\tag{7}$$

Авторы статьи [12] предлагают метод описания множества X, в котором множество данных описывается с помощью эллипсоида, чья матрица равна ковариационной матрице выборки X. Такой выбор матрицы связан с целью учесть корреляции между различными характеристиками моделируемых объектов. В работе [12] центр эллипсоида фиксирован в среднем арифметическом точек X: $\mu = \frac{1}{N}\sum\limits_{i=1}^{N} x_i$. Матрица P вычисляется по формуле:

$$P = Cov\,(X,X) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T\ .$$

В итоге, статья [12] предлагает решать следующую оптимизационную задачу:

$$\begin{aligned}&\min_{R,\xi} R + C\sum_{i=1}^{N}\xi_i\\&s.t.\,(x_i-\mu)^T Cov^{-1}(X,X)(x_i-\mu)\le R+\xi_i, i=1\ldots N,\\&R\ge 0,\xi_i\ge 0, i=1\ldots N\ .\end{aligned}\tag{8}$$

Обе задачи (7) и (8) могут быть записаны в выпуклой форме (см. раздел 4.2) и, следовательно, используя метод скаляризации, возможно найти их фронт Парето.

4. ПРЕДЛАГАЕМЫЙ НОВЫЙ ПОДХОД

Эллипсоид, построенный посредством задачи (8), имеет фиксированный центр. Если позволить ему меняться, как в задаче (7), мы, очевидно, получим эллипсоиды не хуже после решения следующей задачи:

$$\begin{aligned}&\min_{R,a,\xi} R + C\sum_{i=1}^{N}\xi_i\\&s.t.\,(x_i-a)^T Cov^{-1}(X,X)(x_i-a)\le R+\xi_i, i=1\ldots N,\\&R\ge 0,\xi_i\ge 0, i=1\ldots N\ .\end{aligned}\tag{9}$$

Описанные выше задачи (7), (8) и (9) имеют общий недостаток: матрица эллипсоида фиксирована. Мы предлагаем включить матрицу эллипсоида в список переменных оптимизации. Сделав это, получим следующую оптимизационную задачу:

$$\begin{aligned}&\min_{P,a,R,\xi} R + C\sum_{i=1}^{N}\xi_i\\&s.t.\,(x_i-a)^T P^{-1}(x_i-a)\le R+\xi_i, i=1\ldots N,\\&P=P^T\succ 0, R\ge 0,\xi_i\ge 0, i=1\ldots N\ .\end{aligned}$$

Параметры эллипсоида (1) включают и матрицу P, и квадрат эффективного радиуса R. Умножая элементы P на некоторое число $z>1$, можно без нарушения ограничения уменьшить

R и ξ_i, деля их на z, и уменьшить значение $R + C\sum_{i=1}^{N}\xi_i$ сколь угодно близко к нулю. Чтобы избежать этого, потребуем, чтобы эллипсоид с единичным эффективным радиусом был бы не слишком большим. Добавим одно неравенство на определитель матрицы эллипсоида в список ограничений задачи. Дополнительно, заменим C на другой «безразмерный» параметр $C = \frac{1}{N\nu}$, смысл которого будет объяснен позже. Таким образом, мы получаем задачу:

$$\begin{aligned}&\min_{P,a,R,\xi} R + \frac{1}{N\nu}\sum_{i=1}^{N}\xi_i \\ &s.t.\,(x_i-a)^T P^{-1}(x_i-a) \le R+\xi_i, i=1\dots N, \\ &P = P^T \succ 0, \det P \le 1, R \ge 0, \xi_i \ge 0, i=1\dots N\ .\end{aligned} \tag{10}$$

4.1. Интерпретация параметра скаляризации

Следующая теорема разъясняет смысл замены $C = \frac{1}{N\nu}$.

Теорема 2. *Доля точек множества X, которые не принадлежат эллипсоиду* (1)*, построенному с помощью решения задачи* (10) *не превосходит ν.*

Доказательство. Лагранжиан для задачи (10) имеет следующий вид:

$$\begin{aligned}L(P,a,R,\xi,\alpha,\beta,\gamma,\delta) =& R + \frac{1}{N\nu}\sum_{i=1}^{N}\xi_i \\ &+ \sum_{i=1}^{N}\alpha_i\left((x_i-a)^T P^{-1}(x_i-a) - R\xi_i\right) \\ &- \sum_{i=1}^{N}\beta_i\xi_i - \gamma R + q(P,\delta)\ ,\end{aligned}$$

где $\alpha = \{\alpha_i \ge 0\}_{i=1}^{N}$, $\beta = \{\beta_i \ge 0\}_{i=1}^{N}$, $\gamma \ge 0$ – множители Лагранжа, $q(P,\delta)$ – функция матрицы эллипсоида и множителей Лагранжа, соответствующая оставшимся ограничениям. В оптимальной точке $\nabla L = 0$. Получаем следующую систему уравнений:

$$\frac{\partial L}{\partial R} = 1 - \sum_{i=1}^{N}\alpha_i - \gamma = 0\ , \tag{11}$$

$$\frac{\partial L}{\partial \xi_i} = \frac{1}{N\nu} - \alpha_i - \beta_i = 0\ . \tag{12}$$

Также из условий Каруша-Куна-Такера имеем уравнения дополняющей нежёсткости:

$$\beta_i\xi_i = 0\ . \tag{13}$$

Точка $x_i \notin E$ тогда и только тогда, когда $\xi_i > 0$, что вместе с (13) означает, что $\beta_i = 0$. Из (12) получаем $\alpha_i = \frac{1}{N\nu}$ для выбросов (точек вне эллипсоида). Так как $\gamma \ge 0$, то, учитывая уравнение (11), заключаем, что $\sum_{i=1}^{N}\alpha_i \le 1$. Используя неотрицательность всех α_i мы утверждаем, что количество тех α_i, которые равны $\frac{1}{N\nu}$, не превосходит $N\nu$. □

Можно доказать похожие утверждения для задач (7), (8) и (9).

Три следующих замечания показывают, как теорема 2 может быть применена на практике.

Замечание 1. Из теоремы 2 следует, что количество точек множества X, которые лежат вне эллипсоида, не превосходит $N\nu$, что позволяет указать диапазон возможных значений параметра ν. На практике слишком малые или слишком большие значения ν вне отрезка $\left[\frac{1}{N}, 1\right]$ не используются. Для малых значений ν все точки принадлежат эллипсоиду, и изменение параметра скаляризации ниже значения $\frac{1}{N}$ не меняет количество (3) выбросов среди обучающего множества, оно остается $K = 0$.

Для значений $\nu > 1$ все точки могут лежать вне эллипсодида. Для больших значений ν коэффициент $C = \frac{1}{N\nu}$ для меры удаленности точки от эллипсоида становится малым и оптимальный эллипсоид вырождается в точку с $R = 0$. Действительно, уравнение (12) доказывает $0 \le \alpha_i \le \frac{1}{N\nu}$. Если $\nu > 1$, то $\sum\limits_{i=1}^{N} \alpha_i < 1$ и, следовательно (см. (11)), $\gamma > 0$. Условие ККТ $\gamma R = 0$ доказывает, что $R = 0$, т.е. эллипсоид действительно вырождается в точку.

Замечание 2. Решая задачу (10), можно аппроксимировать часть фронта Парето задачи (4). Теорема 2 позволяет выбрать значение скаляризационного параметра, если инженер хочет получить эллипсоид с заданным количеством выбросов (3).

Замечание 3. Возможно диагностировать изменения типичного поведения моделируемой системы. В задаче определения аномалий аналитики работают с имеющейся выборкой данных и обучают классификатор. Когда система внедрена, нормальное поведение системы может измениться, и параметры, которые ранее были аномальными, в будущем могут стать нормальными, и наоборот. Это может означать исчерпанный ресурс механизма. Или, альтернативно, эволюция системы может быть типична (например, изменение климата), и необходимо заново построить эллипсоид по свежим данным. Теорема 2 позволяет автоматически детектировать необходимость повторного обучения модели. Система может отслеживать интенсивность детектирования выбросов. Эллипсоид следует считать устаревшим, если доля выбросов превосходит значение ν, использованное при построении эллипсоида.

4.2. Выпуклая формулировка задачи оптимизации

Задача (10) может быть переформулирована в выпуклом виде. Пусть $Q = P^{-\frac{1}{2}}$, $b = Qa$. Тогда задача (10) эквивалента следующей:

$$\begin{aligned}
&\min_{Q,b,R,\xi} R + \frac{1}{N\nu}\sum_{i=1}^{N}\xi_i \\
&s.t.\,(Qx_i - b)^T I\,(Qx_i - b) \le R + \xi_i, i = 1\ldots N, \\
&Q = Q^T \succ 0, \det Q \ge 1, R \ge 0, \xi_i \ge 0, i = 1\ldots N\ .
\end{aligned}$$

Известно, что функция $f(Q) = -\ln\det Q$ выпукла на множестве положительно определенных симметричных матриц [1]. Также с помощью леммы Шура (*Schur's lemma*) квадратичное ограничение может быть переписано в виде линейного матричного неравенства (*linear matrix inequality*) (ЛМН). Множество решений ЛМН выпукло [13]. Следовательно, задача (10) может

быть сведена к задаче выпуклого программирования:

$$\begin{aligned}
&\min_{Q,b,R,\xi} R + \frac{1}{N\nu}\sum_{i=1}^{N}\xi_i \\
&s.t. \begin{pmatrix} R+\xi_i & (Qx_i-b)^T \\ (Qx_i-b) & I \end{pmatrix} \succ 0, i = 1\dots N, \\
&Q = Q^T \succ 0, -\ln\det Q \le 0, R \ge 0, \xi_i \ge 0, i = 1\dots N \ .
\end{aligned} \tag{14}$$

Аналогично в выпуклом виде могут быть сформулированы задачи (7), (8) и (9).

4.3. Уточнение решения

Задачи (9) и (14) являются аппроксимациями задачи (4). Эллипсоид E, на котором достигается соответствующее минимальное значение, содержит некоторое подмножество $U(E) \subseteq X$ векторов из обучающей выборки. Точки из $U(E)$ составляют множество тех точек, при котором эллипсоид E является оптимальным по Парето для задачи (6).

Известна классическая задача построения эллипсоида минимального объема, содержащего заданные точки (эллипсоид Левнера). Зафиксируем в уравнении эллипсоида (1) квадрат эффективного радиуса $R = 1$ и сделаем замену переменных $Q = P^{-\frac{1}{2}}, b = P^{-\frac{1}{2}}a$. Тогда параметры эллипсоида минимального объема, содержащего точки из $U(E)$ могут быть найдены путем решения следующей задачи [13]:

$$\begin{aligned}
&\min_{Q,b} -\ln\det Q \\
&s.t. \begin{pmatrix} 1 & (Qx_i-b)^T \\ (Qx_i-b) & I \end{pmatrix} \succeq 0, \ x_i \in U(E) \ .
\end{aligned} \tag{15}$$

Пусть имеется эллипсоид E, оптимальный для задачи (6). Найдем те точки из X, которые принадлежат эллипсоиду E: $U(E) = \{x \in X | x \in E\}$. С помощью процедуры (15) построим эллипсоид E' минимального объема, содержащий точки из $U(E)$. Тогда $Vol(E') \le Vol(E)$ и $K(E') \le K(E)$. На практике первое неравенство обычно выполняется строго. Эллипсоид E' в общем случае не принадлежит эллипсоиду E и, в дополнение к точкам из $U(E)$ может содержать дополнительные точки из X. Таким образом, применение процедуры (15) позволяет получить эллипсоид E', доминирующий эллипсоид E в смысле основной задачи (4).

5. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Методы, основанные на решении оптимизационных задач (9) и (14) после применения процедуры (15), сравнены с известными ранее методами (7) and (8). В дальнейшем мы будем называть методы (7) *«Шаром»* (*Ball*), (8) – *«Эллипсоидом главных компонент»* (*Principal Component Ellipsoid* или, сокращенно, *PCE*), и назовем наш метод (9) вместе с последующим применением процедуры (15) *«Эллипсоидом главных компонент с оптимальным центром»* (*PCE with optimal center*) и (14), (15) – *«Оптимальным эллипсоидом»* (*Optimal ellipsoid*).

Задачи выпуклого программирования решались с помощью пакета CVX для MATLAB [14].

Сравнение методов было проведено на следующих искусственных наборах данных:

1. Множество **box_10_100** состоит из 100 точек, равномерно распределенных в кубе $[0,1]^{10}$.
2. Чтобы сгенерировать множество **normal_COV_6_100** из 100 точек, было использовано многомерное нормальное распределение в пространстве $\mathbf{R}^6$ с фиксированной неединичной ковариационной матрицей $\Sigma \neq I$.

3. Двумерный набор данных **banana** состоит из 100 точек, взятых из невыпуклой области. Он показан на рисунке 1.

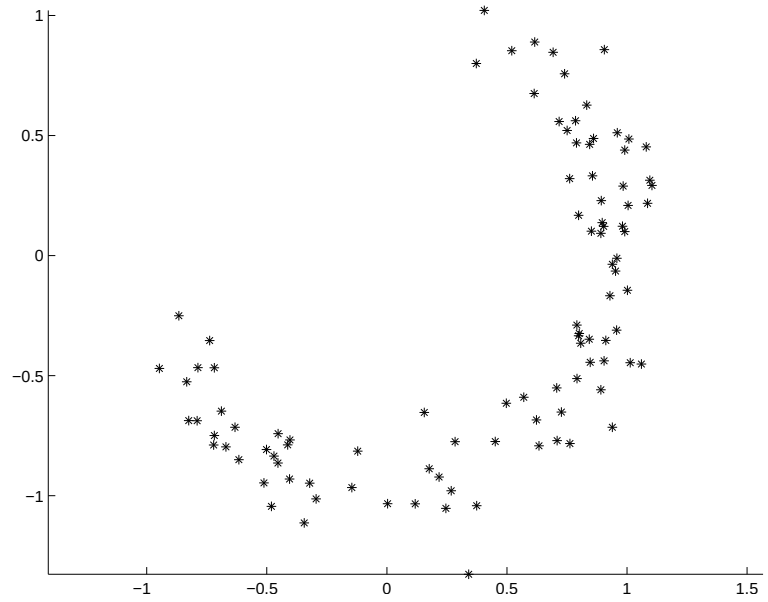

Рис. 1. Множество из 100 точек **banana** в $\mathbf{R}^2$, которые образуют невыпуклую область.

Графики 2, 3 и 4 показывают аппроксимации фронта Парето для задачи (4), полученные посредством рассмотренных методов для наборов данных **box_10_100**, **normal_COV_6_100** и **banana** соответственно. По оси ординат отложена доля $K(E)/N$ точек, которые не принадлежат эллипсоиду. По оси абсцисс – d-мерный объем $Vol(E)$ эллипсоида. Для каждого метода и набора данных было использовано 40 различных значений ν в диапазоне $(0, 0.5]$.

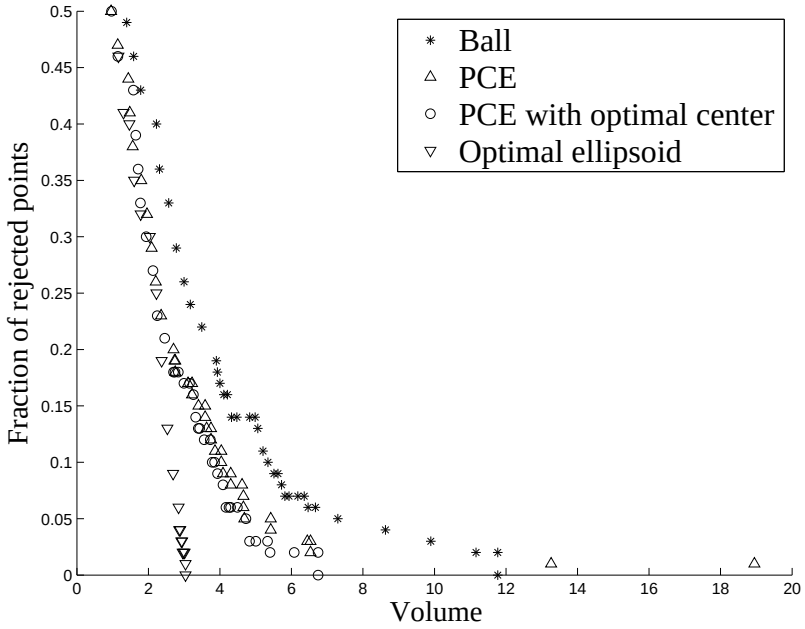

Рис. 2. Приближенный Парето-фронт для набора данных **box_10_100**.

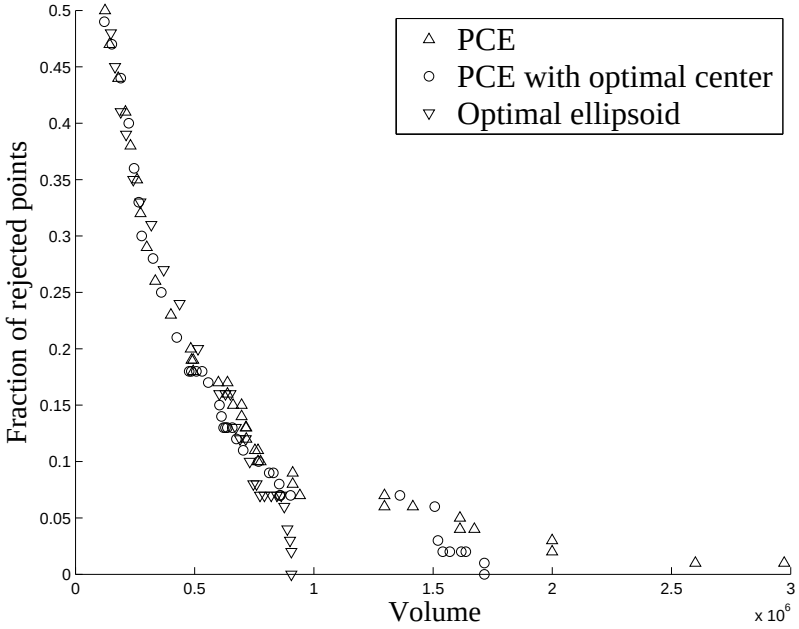

Рис. 3. Приближенный Парето-фронт для набора данных **normal_COV_6_100**.

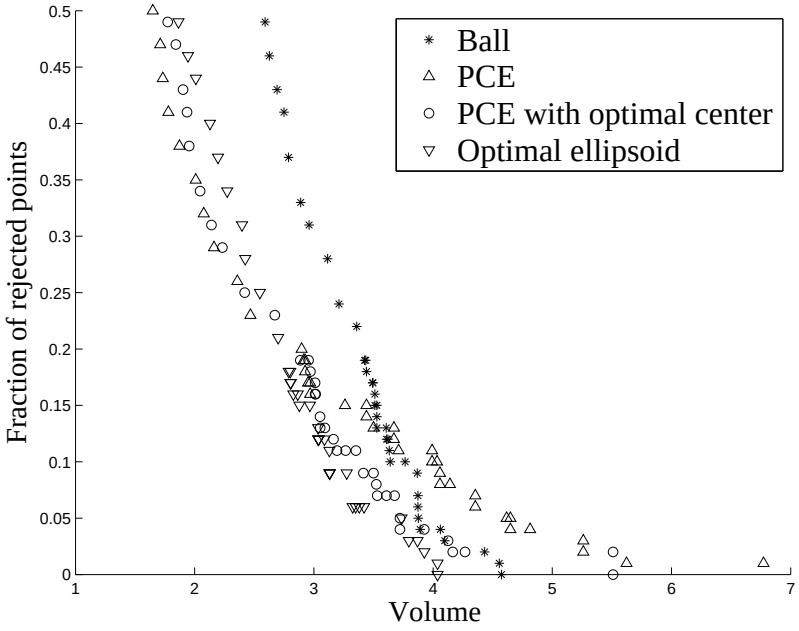

Рис. 4. Приближенный Парето-фронт для набора данных **banana**.

Анализируя результаты экспериментов, мы приходим к следующим выводам:

1. В пространствах низкой размерности d, когда размер выборки N достаточно большой, все методы дают близкие аппроксимации фронта Парето.
2. Метод *«Шар»* дает гораздо более далекие от фронта Парето точки, чем другие методы, если дисперсия данных вдоль одного направления существенно отличается от дисперсии вдоль других направлений.

3. Иногда наш метод *«Оптимальный эллипсоид»* дает точки, которые доминируются эллипсоидами, построенными другими методами. Это возможно, поскольку формулировка (6) – приближение задачи (4) и дает приближенное решение. Эллипсоиды, построенные с помощью одной аппроксимации, могут доминировать другие приближенные решения.
4. Если ν увеличивается, то есть количество выбросов в обучающем множестве велико, то все методы (кроме *«Шара»*) дают близкие фронты Парето в большинстве экспериментов.
5. *«Оптимальный эллипсоид»* дает существенно лучшие по сравнению с другими методами результаты в высокоразмерных пространствах, и когда множество X состоит из не очень большого количества точек. Такие ситуации возникают на практике. Компьютерный эксперимент может быть очень длительным, и в начале исследования база данных экспериментов содержит небольшое количество многомерных векторов.
6. Наш подход *«Оптимальный эллипсоид»* также существенно выигрывает у ранее известных методов, когда доля точек вне эллипсоида ν достаточно мала. Эта часть фронта Парето наиболее интересна на практике.

6. ЗАКЛЮЧЕНИЕ

Мы рассмотрели представление данных с использованием экстремальных эллипсоидов. Основная задача (4) с непрерывной и дискретной целевыми функциями решена приближенно путем замены исходных критериев на другие выпуклые функции. Для решения многокритериальной задачи использован метод скаляризации.

В перспективе стоит изучить альтернативные аппроксимации задачи (4). Минимизация сумм расстояний в метрике Махаланобиса от границы эллипсоида до точки вместо минимизации их квадратов может увеличить устойчивость оценки местоположения и дать более точные результаты для количества выбросов $K(E)$, потому что большие расстояния от точки до эллипсоида штрафуются слабее. Возможны также более точные, но невыпуклые аппроксимации данной задачи. Однако для них не существует универсальных быстрых методов решения.

СПИСОК ЛИТЕРАТУРЫ

1. Boyd S., Vandenberghe L. *Convex Optimization*. Cambridge: UniversityPress, 2004.
2. Бедринцев А.А., Чепыжов В.В., Чернова С.С. Экстремальные эллипсоиды как аппроксиматоры пространства дизайна в задачах предсказательного метамоделирования. *Искусственный интеллект и принятие решений*, 2015, №2, стр. 35–44.
3. Бедринцев А.А., Чепыжов В.В.. Двухкритериальная задача построения оптимальных эллипсоидов для представления данных. *Информационные технологии и системы – 2014. Сборник трудов.* Нижний Новгород, 2014.
4. David T.J. Tax, Robert P.W. Duin. Support Vector Data Description. *Machine Learning*, 2004, vol. 54, pp. 45–66.
5. Karanjit Singh, Dr. Shuchita Upadhyaya. Outlier Detection: Applications And Techniques. *International Journal of Computer Science Issues*, 2012, vol. 9, issue 1, no.3, pp. 307–323.
6. Poonam Rana, Deepika Pahuja, Ritu Gautam. A Critical Review on Outlier Detection Techniques. *International Journal of Science and Research*, 2014, vol. 3, issue 12, pp. 2394–2403.
7. Peter Rousseeuw. Multivariate Estimation with High Breadown Point. In: *Mathematical Statistics and Applications, Vol. B.* Eds. W. Grossmann, G. Pflug, I.Vincze. W.Wertz. Dordrecht: Reidel Publishing Company, 1985, pp. 283-297.
8. Laurie Davies. The Asymptotics of Rousseeuw's Minimum Volume Ellipsoid Estimators. *The Annals of Statistics*, 1992, vol. 20, no.4, pp. 1828–1843.

9. Xuming He, Gang Want. Cross-Checking Using the Minimum Volume Ellipsoid Estimator. *Statistica Sinica*, 1996, vol. 6, pp. 367-375.

10. Stefan Val Aelst, Peter Rousseeuw. Minimum volume ellipsoid. *WIREs Computational Statistics*, 2009, vol. 1, pp. 71–82.

11. Wei-Cheng Chang, Ching-Pei Lee and Chih-Jen Lin. *A Revisit to Support Vector Data Description.* Technical report. National Taiwan University of Science and Technology, 2013.

12. Yang Zhang, Nirvana Meratnia, Paul J.M. Havinga. Distributed online outlier detection in wireless sensor networks using ellipsoidal support vector machine. *Ad Hoc Networks*, 2012, vol. 11, issue 3, pp. 1063–1074.

13. Stephen Boyd, Laurent El Ghaoui, Eric Feron, Venkataramanan Balakrishnan. *Linear Matrix Inequalities in System and Control Theory.* SIAM studies in applied mathematics, 1994, vol. 15.

14. CVX: Matlab Software for Disciplined Convex Programming. http://cvxr.com/cvx/

Design Space Description by Extremal Ellipsoids in Data Representation Problems

Bedrintsev A., Chepyzhov V.

Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow, Russia

Problems of the data set description and outliers detection are solved by constructing the optimal ellipsoid. Optimization problems are formulated as convex programming problems using linear matrix inequalities. The proposed method is compared with the similar previously known methods in terms of two criteria: the volume of ellipsoid and the number of points in train data set that lie beyond the ellipsoid.

KEYWORDS: extremal ellipsoids, design space, linear matrix inequalities, convex programming.