УДК 519.6

А.П.Гожий

НЕПАРАМЕТРИЧЕСКОЕ ОЦЕНИВАНИЕ ЭКСПЕРТНЫХ ДАННЫХ В ЗАДАЧАХ СЦЕНАРНОГО АНАЛИЗА

В процессе построения и анализа сценариев при обработке экспертных данных могут возникать ситуации, когда отсутствует информация о функциях распределения вероятностных событий на прогнозных графах, а задаются только различия между ними.

В данных случаях целесообразно использовать аппарат непараметрической статистики. На данный момент непараметрические методы применяются в таких отраслях, как: биологические и медицинские исследования, социология, экономика, машиностроение, прогнозирование, сценарное планирование и т.д.

Рассмотрим основные типы непараметрических задач, которые могут решаться при выполнении анализа экспертной информации.

Прежде всего необходимо выделить сугубо непараметрическую задачу оценивания неизвестных распределений [6,8]. Ее не следует смешивать с задачей аппроксимации неизвестного распределения известными функциями, которая рассматривается в обычной (параметрической) статистике и которая в конечном счете сводится к оценке параметров этих функций. В непараметрической постановке эта задача формулируется следующим образом: задается достаточно широкий непараметрический класс распределений (например, класс всех непрерывных функций распределений или класс всех распределений, обладающих плотностью); требуется предложить процедуру, результатом которой является оценка функции распределения или плотности. Другими словами, задача требует определить различия между распределениями внутри заданного класса, причем эти различия вообще не конкретизируются.

Если же эти различия конкретизировать, то мы приходим к другому классу задач, аналогичных по своей форме классическим задачам оценки параметров. Аналогия здесь опять же чисто внешняя, так как о параметрах распределения в обычном смысле говорить нельзя, поскольку класс распределения непараметричен, т.е. фактически оценивается не параметр распределения, а параметр

 ISSN 1562-9945

различия между распределениями внутри заданного непараметрического класса [6].

Параметр различия может быть определен через известный функционал от неизвестного распределения. К настоящему времени известны следующие основные оценки таких функционалов:

- оцениваемый параметр известным образом связан с квантилями;
- оцениваемый параметр выражается как математическое ожидание известной функции по неизвестному распределению;
- параметр известным образом входит в неизвестное распределение;
- параметр есть нелинейный интегральный функционал (типа энтропии).

Третья категория непараметрических задач - проверка непараметрических гипотез.

Любая задача проверки непараметрических гипотез выглядит следующим образом. Из двух конкурирующих гипотез альтернатива всегда непараметрична, а нулевая гипотеза может быть либо простой, либо непараметрической. Поскольку, по крайней мере, одна гипотеза есть класс неизвестных распределений, различие между гипотезами задается в некотором общем виде, не связанном с конкретным видом функции распределения.

Требуется предложить процедуру (тест), результатом которой явилось бы решение об истинности одной из гипотез на основании предъявленной выборки (или нескольких выборок - при многовыборочных задачах).

Последний класс непараметрических задач в настоящее время наиболее развит и широко используется на практике. Поэтому перечислим основные непараметрические задачи проверки гипотез, приведенные в работе [3], чтобы продемонстрировать, как именно задаются гипотезы и различия между ними .

1. Задача согласия. Пусть задано известное непрерывное распределение F(x). Из неизвестного распределения G{x), принадлежащего классу всех остальных распределений, берется выборка $x_1., x_2,..., x_n$.

Конкурирующие гипотезы:

нулевая гипотеза $H_0 : F = G$ - простая гипотеза;

альтернатива
- а) $H^+_1 : F<G$
- б) $H^-_1 : F<G$

— односторонние гипотезы;

в) $H_1 : F \neq G$ - двусторонняя гипотеза

Как видим, нулевая гипотеза проста, альтернатива в любом из вариантов непараметрична, различие между ними задается односторонним или простым неравенством между F и G.

2. Задача сдвига (расположения). Иногда известно, что интересующий нас фактор приводит к сдвигу распределения в ту или иную сторону. Направление сдвига может быть известным или неизвестным. В таких обстоятельствах возникает задача обнаружения или локализации. В простейшей постановке задача расположения формулируется в том случае, когда известно, что альтернатива сводится только к сдвигу, т.е. $F(x|\theta)=F_0(x-\theta)$, где θ - параметр сдвига.

Нулевая гипотеза Но : $\theta = 0$.

Альтернатива:

а) $H'_1 : \theta>0$ }

б) $H''_1 : \theta<0$ } - односторонние гипотезы;

в) $H'''_1 : \theta \neq 0$ - двусторонняя гипотеза

В других случаях может быть не известно, проявляется ли влияние исследуемого фактора только в сдвиге, но известно, что сдвиг может иметь место.

Поскольку распределения неизвестны, среди них иногда могут встретиться и не имеющие моментов (по которым можно было бы судить о сдвиге); поэтому естественной мерой сдвига являются квантили того или иного уровня p. Возможны следующие варианты задачи сдвига.

Нулевая гипотеза $H^*_0 : F^{-1}(p) = x_p$.

Альтернатива:

а) $H'_1 : F^{-1}>x_p$ }

б) $H''_1 : F^{-1}<x_p$ } - односторонние гипотезы;

в) $H'''_1 : F^{-1} \neq x_p$ - двусторонняя гипотеза

В односторонних задачах иногда рассматривается несколько общая нулевая гипотеза

$H_0 : F^{-1}(p) \leq x_p$; $H_1 : F^{-1}(p)>x_p$

В некоторых случаях известна симметричность распределения относительно медианы. Тогда задача сдвига может быть сформулирована с учетом этой информации:

 ISSN 1562-9945

H_0 :$F^{-1}(0,5) = x_0$ и F(x) - симметрична относительно,

а) H'_1 : $F^{-1}(0,5) > x_0$, F(x) - симметрична,

б) H''_1 : $F^{-1}(0,5) \neq x_0$, F(x) - симметрична.

3. Задача расположения и симметрии. В отличие от задачи расположения в данной задаче альтернатива расширяется так, чтобы охватить как все сдвинутые, так и все несимметричные распределения:

H_0 :$F^{-1}(0,5) = x_0$ и F(x) - симметрична,

H_1 :$F^{-1}(0,5) = x_0$ и F(x) - несимметрична.

По существу, это задача проверки симметричности распределения F(x) относительно точки x_0.

4. Задача масштаба. В ряде случаев заранее известно, что исследуемый фактор приводит к изменению масштаба распределения.

Если изменяется только масштаб, имеем альтернативу вида $F(x/\theta) = F_0(\theta,x)$ и задачу следующего типа.

Нулевая гипотеза H_0 : $\theta = 1$.

Альтернатива:

а) H'_1 : $\theta>1$ }

б) H''_1 : $\theta<1$ } - односторонние гипотезы;

в) H'''_1 :$\theta \neq 1$ - двусторонняя гипотеза

Если кроме изменения масштаба могут происходить какие-либо другие изменения распределения, а нас интересует только сам масштаб, необходимо ввести меру масштаба. В качестве меры масштаба, как и при сдвиге, разумно использовать квантильную меру или меру типа размаха выборки. Обозначим выбранную меру через р. Тогда имеем следующую задачу.

Нулевая гипотеза H_0 : $\rho = \rho_0$.

Альтернатива:

а) H'_1 : $\rho>\rho_0$ }

б) H''_1 : $\rho<\rho_0$ } - односторонние гипотезы;

в) H'''_1 :$\rho \neq \rho_0$ - двусторонняя гипотеза

Для решения перечисленных задач используется ряд непараметрических методов и критериев проверки гипотез, основными из которых являются:

критерий сумм рангов Уилкоксона;

критерий знаков Фишера;

критерий Гупты;

критерий Ансари-Бредли;

критерий Краскела-Уолиса;

критерий Фридмана;

критерий Колмогорова-Смирнова.

Эти непараметрические методы являются наиболее распространенными, и эффективно применяются в разнообразных ситуациях [5,7,8]. Краткое описание критериев приведено в таб. 1.

Таблица 1

Непараметрические критерии проверки гипотез

Название критерия	Задача	Данные	Короткое описание
Критерий суммы рангов Уилкоксона	О положении (сдвиге) распределения.	Две выборки X_1 ., Xm и Y_1 ., Yn	Проверка гипотезы о том, что выборки X и Y выбраны из одной совокупности, то есть $\Delta = 0$ (разность медиан X и Y)
Критерий Фишера	О положении (сдвиге) распределения.	Парные повторные наблюдения (X_1, Y_1) , (Xn, Yn)	Проверка гипотезы об эффекте обработки θ, где θ – неизвестная медиана X - Y
Критерий Гупты	О положении (сдвиге) распределения.	Одна выборка Z_1 ., Zn	Проверка гипотезы о симметричности совокупности Z относительно медианы θ
Критерий Ансари-Бредли	О рассеянии (масштабе) распределения	Две выборки X_1 ., Xm и Y_1 ., Yn	Проверка гипотезы о том, что выборки X и Y выбраны из одной совокупности.
Критерий Краскела-Уолиса	Однофакторный дисперсионный анализ (k выборок)	N наблюдений Xij n = 1 ., nj, j = 1..., k .	Проверка гипотезы о том, что k выборок выбранны из одной совокупности ($\tau_1 = \tau_2 = \ldots = \tau_k$). τ_j –эффект обработки j выборки.
Критерий Фридмана	Двухфакторний дисперсионный анализ	n * к наблюдений Xij., n = 1 , j = 1..., к.	Проверка гипотезы о том, что k выборок выбранны из одной совокупности ($\tau_1 = \tau_2 = \ldots = \tau_k$). τ_j –эффект обработки j выборки.
Критерий Колмогорова-Смирнова	Критерии для альтернатив	Две выборки X_1 ., Xm и Y_1 ., Yn	Проверка однородности двух выборок против альтернативы, о том, что альтернативы отличны.

Изложенный в данной статье материал является частью системной методологии применения статистических методов в анализе сценариев, которая представлена в [1,2] .

 ISSN 1562-9945

Для реализации непараметрического оценивания в задачах сценарного анализа, представленные методы были реализованы в виде программного комплекса структура которого приведена на рис.1.

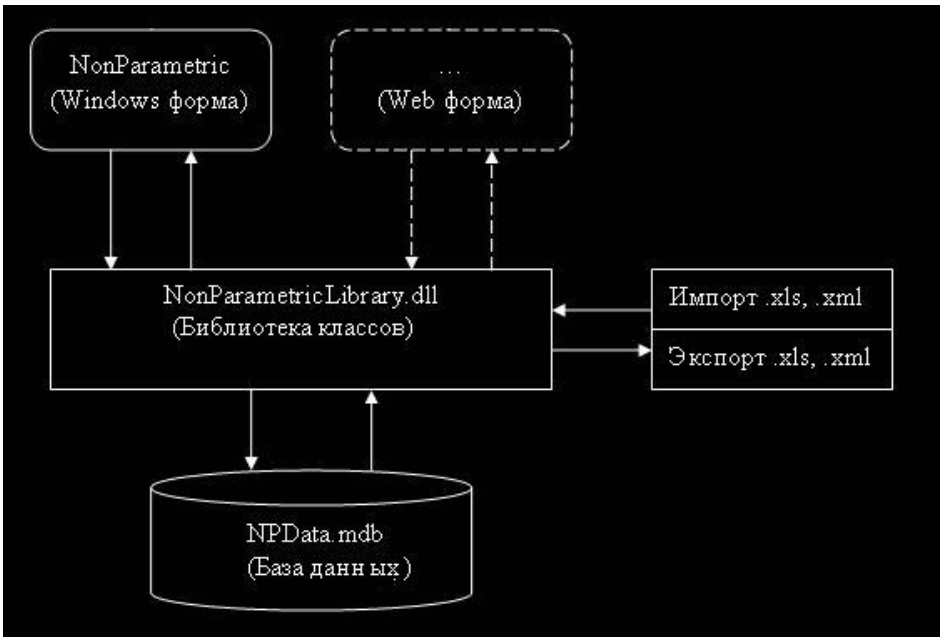

Рисунок 1 - Структура программного комплекса непараметрического оценивания

Главной частью программного комплекса является библиотека классов NonParametricLibrary.dll. В ней сохраняется иерархия классов, которая осуществляет анализ данных с помощью непараметрических методов. Каждому методу соответствует определенный класс. Кроме того реализованы классы которые инкапсулируют ввод и вывод данных. Эти классы могут использоваться любым приложением Windows (Windows-application): Windows форма, web форма, служба Windows или web служба. Необходимая для непараметрических критериев информация сохраняется в локальной базе данних NPData.mdb. Именно отсюда выбираются табличные значения для представленных выше методов. Кроме того, в базе даних сохраняется описание каждого реализованного метода и описание основной и альтернативной гипотез. Метод класса с .dll файла обращается к базе данних з запросом и получает набор данных (dataset) в ответ. Возможен ввод и вывод данных с помощью электронных таблиц MS Excel (.xls).

На рисунке 2 приведен внешний вид интерфейса программного комплекса.

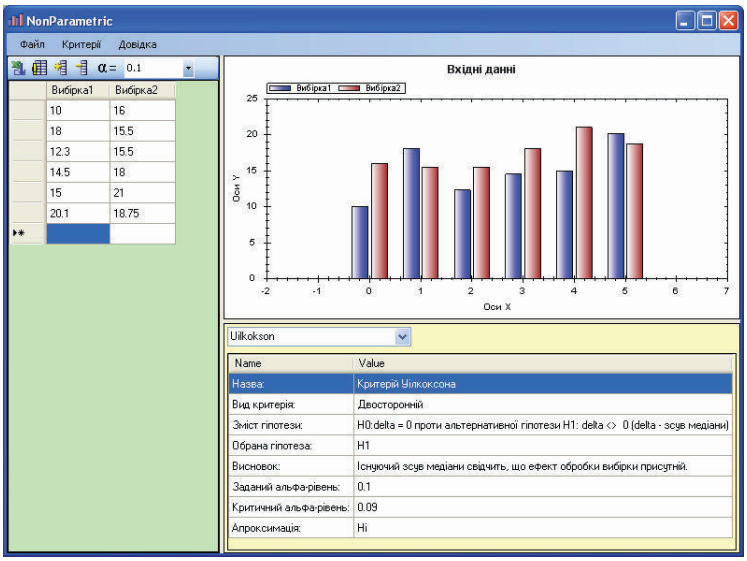

Рисунок 2 - Интерфейс программного комплекса непараметрического оценивания

Данный программный комплекс может использоваться автономно, а может применяться в качестве подсистемы непараметрического оценивания в системах поддержки принятия решений в задачах сценарного планирования.

ЛИТЕРАТУРА

1. Гожий О.П., Коваленко И.И. Системный подход к выбору статистических методов при генерации и анализе сценариев // Системные технологии. Региональный межвузовский сборник научных трудов.- Выпуск №5 (46). Днепропетровск, 2006- с.38-49.
2. Коваленко И.И., Гожий А.П. Системные технологии генерации и анализа сценариев: Монография. – Николаев: Изд-во НГГУ им. Петра Могилы, 2006. – 160с.
3. Леман Э. Проверка статистических гипотез. М. .: Финансы и статистика, 1964.
4. Орлов А.И. Теория принятия решений. Учебное пособие. - М.: Издательство "Март", 2004
5. Рунион Ф. Справочник по непараметрической статистике. Пер с англ. – М.:Финансы и статистика, 1982. – 198с.
6. Тарасенко Ф.П., Непараметрическая статистика. – Томск,: ТГУ, 1976.
7. Тюрин Ю.Н., Непараметрические методы статистики. – М.: «Знание» 1978.
8. Холлендер М., Вулф Д.А. Непараметрические методы статистики - М.: Финансы и статистика 1983 г.

Получено 18.03.2008 г.

 ISSN 1562-9945