ОЦЕНИВАНИЕ РИСКА РЕГРЕССИОННОЙ МОДЕЛИ В СЛУЧАЕ НЕИЗВЕСТНОГО РАСПРЕДЕЛЕНИЯ[1]

Татьяна Ступина, Виктор Неделько

***Аннотация**: В данной работе поднимается достаточно актуальная проблема оценивания качества решения в условиях отсутствия информации о распределениях. Для задачи регрессионного анализа рассматривается альтернативная функция риска, построенная ранговым методом. Отражены положительные и отрицательные стороны такого подхода. Статистическим моделированием получены точечные оценки эмпирической функции риска, отражающие обоснованность применения рангового метода в условия «полной неопределённости».*

***Ключевые слова**: функция риска, эмпирическая функция риска, ранговая регрессия, класс линейных решающих функций.*

***ACM Classification Keywords**: G3 Вероятность и Статистика – Корреляционный и Регрессионный анализ.*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

Введение

Подход к обработке экспериментальных данных зависит от специфики конкретной области и конечной цели, которая ставится в задаче. В различных областях знаний, целью которых является обнаружение причинно-следственных связей, могут быть использованы одинаковые методы не всегда приводящие к удовлетворительному решению. Чаще всего причина кроется в недостатке априорной информации об изучаемом объекте (явлении) или в некорректной применимости того или иного метода (алгоритма) к обрабатываемым данным. Уточнение же модели, как правило, происходит уже в процессе обработки данных экспертами или в случаях наличия достаточной априорной информации, что не всегда бывает возможным в случае автоматизированной обработки информации и необходимости быстрого принятия решения.

Таким образом, на первом этапе эффективней было бы предложить эксперту модель, полученную наиболее универсальным методом, для её последующего уточнения или вообще принятия решения об её концептуальном изменении. Неотъемлемым этапом в построении модели является её оценка – оценка качества модели. Хорошо известным и широко применяемым способом оценивания качества модели является функция риска [В.Н. Вапник, 1984]. Несмотря на достаточно широкое применение регрессионного анализа во многих прикладных областях знаний задача оценивания риска регрессионной модели и до настоящего времени остаётся актуальной. Это связано с отсутствием универсального метода оценивания качества модели, построенной по выборкам ограниченного объёма в условиях полной неопределённости (отсутствие какой-либо информации о распределениях) [Дж. Себер 1980]. Для задачи распознавания образов предложен подход к эмпирическому оцениванию риска методом численного моделирования, который даёт практически приемлемые оценки [В.М. Неделько, 2008].

[1] Работа выполнена при финансовой поддержке гранта РФФИ 08-01-00944-а

На практике для оценивания риска обычно используют оценки скользящего контроля, как точечные оценки без указания доверительной вероятности. При этом скользящий контроль во многих случаях полагается наилучшим способом оценивания риска, хотя к настоящему времени неизвестны имеющие практически приемлемую точность интервальные оценки риска, основанные на скользящем контроле. В работе [В.М. Неделько, 2008] для задачи распознавания двух образов было показано, что в некоторых случаях на основе эмпирического риска могут быть получены более точные интервальные оценки риска, чем на основе скользящего экзамена. Более того, метод построения эмпирических доверительных интервалов потенциально позволяет использовать не только рассмотренные эмпирические функционалы качества, но и другие характеристики выборки и метода обучения.

В представленной работе получены эмпирические оценки ранговой регрессионной модели из класса линейных функций. Построение решений в данном классе функций не предполагает выполнение классических требований как при восстановлении линейных регрессионных функций. И ещё одним положительным моментом является возможность построения решения в разнотипном пространстве переменных в классе логических решающих функций [Т.А. Ступина, 2006]. Результаты представлены графически и таблично. Проведена сравнительная характеристика эмпирического риска с риском, построенным по контрольной выборке.

Основные понятия

Пусть D_X – пространство значений переменных, используемых для прогноза, а D_Y – пространство значений прогнозируемых переменных, и пусть *C* – множество всех вероятностных мер на заданной σ-алгебре подмножеств множества $D = D_X \times D_Y$.

При каждом $c \in C$ имеем вероятностное пространство: $\langle D, B, P_c \rangle$, где *B* – σ-алгебра, $P_c[D]$ – вероятностная мера (в квадратных скобках мы указываем не аргумент функции, а множество, на котором задана σ-алгебра). Параметр c будем называть *стратегией природы*. Решающей функцией называется соответствие $f : D_X \to D_Y$ из некоторого класса решающих функций Φ.

Качество принятого решения оценивается заданной функцией потерь $L : Y^2 \to [0, \infty)$. Функция потерь задает цену ошибки как меру несоответствия принятого решения $f(x)$ и истинного значения *y*.

Под риском будем понимать средние потери:

$$R(c, f) = \int_D L(y, f(x)) dP_c[D].$$

Заметим, что значение риска зависит от стратегии природы *c* — распределения, которое в общем случае является неизвестным.

Пусть $v = \{(x^i, y^i) \in D \mid i = \overline{1, N}\}$ — случайная независимая выборка из распределения $P_c[D]$. Эмпирический риск определим как средние потери на выборке:

$$\tilde{R}(v, f) = \frac{1}{N} \sum_{i=1}^{N} L(y^i, f(x^i)).$$

Оценка риска на контрольной выборке определяется как

$$R^*(v^*, f) = \frac{1}{N^*} \sum_{i=1}^{N^*} L(y_i^*, f(x_i^*)),$$

где $v^* = \left\{\left(x_i^*, y_i^*\right) \in D \;\middle|\; i = \overline{1, N^*}\right\}$ – «новая» случайная независимая выборка из распределения $P_c[D]$.

Пусть $Q : \{v\} \to \Phi$ – алгоритм (метод) построения решающих функций, а $f_{Q,v} \in \Phi$ – функция из класса решающих функций Φ, построенная по выборке *v* алгоритмом *Q*.

Функционал скользящего экзамена определяется как

$$\breve{R}(v, Q) = \frac{1}{N} \sum_{i=1}^{N} L\left(y^i, f_{Q,v_i'}\left(x^i\right)\right),$$

где $v_i' = v \setminus \left\{\left(x^i, y^i\right)\right\}$ – выборка, получаемая из v удалением *i*-го наблюдения.

Задача построения решающей функции (модели) заключается в выборе подходящего алгоритма *Q* и в оценивании риска принятого решения.

Доверительный интервал для *R* будем задавать в виде $[0, \hat{R}(v)]$.

Здесь мы ограничиваемся односторонними оценками, поскольку на практике для риска важны именно оценки сверху. Таким образом, в данном случае построение доверительного интервала эквивалентно выбору функции $\hat{R}(v)$, которую будем называть оценочной функцией или просто оценкой (риска).

При этом должно выполняться условие:

$$\forall c, \quad P\left(R \le \hat{R}(v)\right) \ge \eta,$$

где η – заданная доверительная вероятность.

Известные на данный момент оценки риска строятся не как функции непосредственно выборки, а через композицию $\hat{R}(v) = R_e\left(\overline{R}(v)\right)$, то есть как функции значений некоторого эмпирического функционала $\overline{R}(v)$, в качестве которого обычно выступает эмпирический риск или скользящий экзамен [В.Н. Вапник, 1984].

Эмпирический функционал здесь выступает в роли точечной оценки риска, на основе которой строится интервальная оценка.

Функция риска построения ранговой регрессии

Пусть $y = f(x)$ — решающая функция, являющаяся некоторой аппроксимацией целевой зависимости, $f \in \Phi$.

Определим риск следующим образом

$$R(c, f) = \max_{A \in \Psi_X} \left| P\left(x \in A, y > f(x)\right) - P\left(x \in A, y < f(x)\right) \right|,$$

где $\Psi_X \subseteq \Lambda_X$ — некоторое подмножество Λ_X — σ-алгебры подмножеств из D_X.

Если $\Psi_X = \Lambda_X$, то

$$R(c, f) = \int_{D_X} \left| \beta^+(x) - \beta^-(x) \right| dP(x),$$

где $\beta^+(x) = P\left(y > f(x) \middle| x\right), \; \beta^-(x) = P\left(y < f(x) \middle| x\right)$.

Чтобы риск можно было оценить по выборке, нужно ограничить Ψ_X, например, множеством интервалов.

Как вариант, в качестве риска можно использовать расстояние Монжа между $\beta^+(x)$ и $\beta^-(x)$.Так же можно попробовать определить расстояние Монжа без использования дополнительной метрики в D_X. Очевидно, что всегда существует $f^*(x)$, для которой риск равен нулю. Это условная медиана, являющаяся оптимальной решающей функцией относительно заданного риска.

Учитывая, что $\beta^+(x) = 1 - \beta^-(x)$ функцию риска представим в следующем виде:

$$R(c, f) = \int_{D_X} |2\beta(x) - 1| dP(x),$$

где $\beta(x)$ - порядок квантили $f(x)$.

Без ограничения общности будем рассматривать $f \in \Phi$ - класс линейных функций. Приоритетной стороной рассматриваемого рангового риска является то, что решения, полученные относительно него являются робастными, т.е. устойчивыми к большим случайным выбросам. Отметим также, что при выполнении классических требований к восстановлению линейных регрессионных зависимостей (ошибки независимы и нормально распределены, регрессоры не случайны) оптимальная решающая функция, представленная условным математическим ожиданием, также является оптимальной относительно рангового критерия.

Выборочный функционал риска. Алгоритм построения решения

Алгоритмом Q по выборке ν объёма N строим эмпирическую функцию f из класса линейных функций Φ. Качество построенной функции будем оценивать по эмпирическому риску:

$$\tilde{R}_f = \sum_{i=1}^{M} \sum_{x \in D_X^i} |2\tilde{\beta}(x) - 1| \cdot \tilde{p}(x),$$

где $\tilde{p}(x) = \frac{N_i}{N}$, $N_i = |D_X^i|$, $\tilde{\beta}(x) = \frac{N_i^1}{N_i}$, $N_i^1 = |D_i^1|$, $D_i^1 = \{(x, y) \in V_N \mid y < f(x), x \in D_X^i, y \in D_Y\}$.

Тогда оптимальной решающей функцией в заданном классе относительно рангового критерия будет функция $\tilde{f}(x) = \arg\min_{f \in \Phi} \tilde{R}_f$.

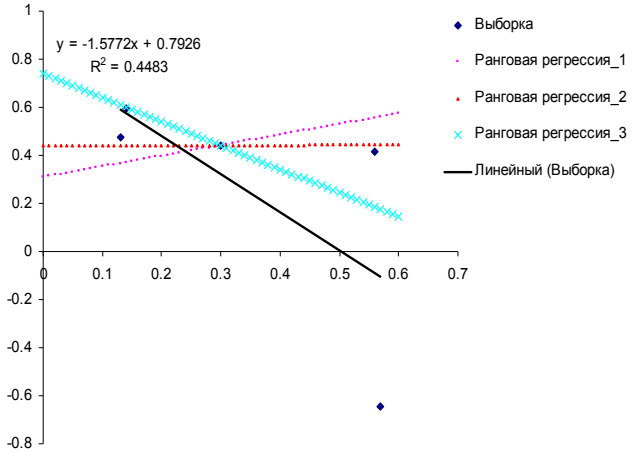

Рис. 1 Линейные регрессии, построенные ранговым методом и по МНК-методу

В целях изучения свойств эмпирического рангового риска будем рассматривать произвольный алгоритм построения линейной зависимости, процедуру и способ разбиения исходного признакового пространства $D_X = \bigcup D_X^i$. Тем самым мы практически охватываем всевозможные способы (алгоритмы) восстановления линейных зависимостей. Оценки эмпирического риска, полученные таким способом, будут являться практически оптимальными. Следовательно, появляется возможность исследования качества решения, построенного некоторым направленным алгоритмом относительно рангового критерия.

Для построения оценки эмпирического риска будем рассматривать оценку риска по контрольной выборке R^* как несмещённой оценки риска [В.Н. Вапник, 1984], представленной в первом параграфе. Риск по контрольной выборке задается аналогично эмпирическому риску, но для элементов контрольной выборки $v^* = \left\{ \left(x_i^*, y_i^* \right) \in D \;\middle|\; i = \overline{1, N^*} \right\}$.

На рисунке 1 мы приведём небольшой показательный пример, демонстрирующий приоритетное свойство линейной регрессии, построенной по ранговому методу в условиях малого объёма выборки, $N = 5$, равномерно распределенной случайной составляющей со среднеквадратическим отклонением равным 0,1 и с 20% выбросами. Истинная линейная функциональная зависимость в примере представляется простым уравнением $f(x) = 0.5$. Несмотря на неоднозначность решения, в примере ранговые регрессионные функции, очевидно, менее отличаются от истинной. По крайней мере, восстановленная по выборке функция достаточно близкая, в метрике L^2 или в метрике C, к истинной является элементом множества решений, имеющих одинаковые значения эмпирической функции риска. В принципе, при введении дополнительных условий, на основании некоторой априорной информации, можно построить алгоритм, определяющий единственное решение из данного множества. Этот вопрос в данной работе мы пока не рассматриваем.

Построение эмпирической оценки риска

Под эмпирической оценкой понимается величина, полученная оцениванием минимальной доверительной вероятности по некоторому эвристически выбранному множеству распределений. Если это множество выбрано достаточно «широким», то естественно ожидать, что полученная оценка будет близка к истинной. Возможность доверия таким оценкам может быть аргументирована следующим соображением. Если целенаправленным эвристическим поиском не удалось построить распределения, при котором доверительная вероятность была бы меньше заданной величины, то можно ожидать, что и в реальной задаче распределение окажется таким, что оценка останется справедливой.

$E\tilde{R}$	ER^*	σ^2
0.12	0.36	0.1
0.16	0.29	0.2
0.17	0.25	0.3
0.21	0.27	0.4

Таб. 1 Оценка эмпирического риска в зависимости от уровня шума

В таблице 1 приведены значения точечных оценок эмпирической функции, построенные статистическим моделированием. Результаты подчёркивают достаточно интересный факт. При плохих распределениях оценка «рангового» риска практически равна значению риска, полученного на контроле, как на распределении. Этот результат даёт нам основание применять эмпирическую оценку риска как

достаточно хорошую при построении ранговой регрессии в случае неизвестного распределения. Проведя дополнительное объёмное моделирование по всевозможным распределениям, можно построить эмпирические доверительные интервалы для функции риска, аналогично тому, как в это было сделано для задачи распознавания двух образов [В.М. Неделько 2008].

Заключение

Несмотря на достаточно хорошо изученные и широко применяемые методы регрессионного анализа, в данной работе поднимается достаточно актуальная проблема оценивания качества решения в условиях отсутствия информации о распределениях. Была рассмотрена и исследована альтернативная функция риска, построенная ранговым методом для задачи восстановления регрессионной зависимости. Отражены положительные и отрицательные стороны такого подхода. Статистическим моделированием получены точечные оценки эмпирической функции риска, отражающие обоснованность применения данного метода в условия «полной неопределённости». Нетривиальной и интересной задачей остаётся создание направленного алгоритма построения эмпирической ранговой регрессии относительно исследуемого риска. Некоторые идеи лежат прямо на поверхности и достаточно скоро будут реализованы авторами работы.

Благодарности

Работа выполнена при финансовой поддержке гранта РФФИ 08-01-00944-а.

Библиография

[Дж. Себер 1980] Дж. Себер. Линейный регрессионный анализ. Изд-во М: Мир, 450с.

[В.Н. Вапник, 1984] В.Н. Вапник. Алгоритмы и программы восстановления зависимостей. Изд-во, М: Наука, 805с.

[В.М. Неделько 2008] В.М. Неделько. Об интервальном оценивании риска для решающей функции. Таврический вестник информатики и математики, Изд-во НАН Украины, 2008, с. 97-103.

[T.A. Ступина 2006] T.A. Stupina. Recognition of the Heterogeneous Multivariate Variable. Proceeding of the international conference, 2006 (KDS'2006), Varna (Bulgaria), Vol 1 – pp. 199-202.

Информация об авторах

Татьяна Ступина *– Институт Нефтегазовой Геологии и Геофизики СО РАН, проспект Коптюга 3, Новосибирск, 630090, Россия, e-mail: stupinata@ipgg.nsc.ru*

Виктор Неделько *– Институт Математики СО РАН, проспект Коптюга 4, Новосибирск, 630090, Россия, e-mail: nedelko@math.nsc.ru*