МАТЕМАТИЧЕСКИЕ МЕТОДЫ В ГЕОФИЗИКЕ — ММГ-2003
Новосибирск, 2003 С. 81–86

Регрессионное моделирование данных сейсмологических наблюдений

К.В. Симонов, С.А. Перетокин,
А.Л. Щемель*, С.И. Шерман†

Аннотация. В работе представлены метод построения нелинейной многомерной аппроксимации и эффективный алгоритм для анализа данных каталога землетрясений, позволяющие учитывать неточности во входных значениях, соответствующих аргументам регрессионной функции. Приводится описание нейросетевого метода и эффективного алгоритма анализа данных сейсмологических наблюдений, основанных на построении нейросетевого гладкого гомоморфного преобразования из равномерного распределения в заданное, которое позволяет "размножать" сейсмические события в исследуемом районе. Обсуждаются результаты вычислительных экспериментов, проведенных с помощью указанного алгоритмического обеспечения, по моделированию данных сейсмического мониторинга сейсмоактивных зон Красноярского края.

1. Введение

Задача обработки данных сейсмического мониторинга формально ставится как преобразование из генеральной совокупности входных данных с заданными доверительными интервалами в выходные, более точные (с меньшими доверительными интервалами). Существует множество способов решения этой задачи; гибкие и мощные средства предлагаются в программном пакете Matlab. В инженерных и экспериментальных расчетах часто используются программы MathCad и Microsoft Excel. В последнем программном продукте можно создавать произвольные функционалы оценки и оптимизировать их.

Хотя используемые методы оптимизации довольно эффективны (метод Ньютона и сопряженных градиентов), но быстродействие указанных программ неудовлетворительное, возможно, из-за неиспользования принципа двойственности для нахождения градиентов, что ограничивает возможность оптимизировать большое число параметров. В работе предлагается краткое описание разработанного алгоритма на основе нейросетевых и других вычислительных технологий (метод упругих карт), способного эффективно обрабатывать неточно заданные и значительные по объему массивы данных сейсмического мониторинга.

2. Постановка задачи обработки данных

Как уже отмечалось, основной круг математических задач, встречающихся при обработке данных сейсмического мониторинга, связан с проблемой, кото-

*Институт вычислительного моделирования СО РАН, Красноярск.
†Институт земной коры СО РАН, Иркутск.

рая формулируется как заполнение пропусков в таблицах (каталоги землетрясений).

При традиционном подходе существует довольно значительное количество методик выполнения прямой регрессии. Чаще всего применяется метод наименьших квадратов: вводится функционал, характеризующий адекватность численной модели имеющимся данным, и он минимизируется путем изменения параметров модели. Имея модель, хорошо отражающую зависимость входов от выходов, содержащуюся в экспериментальных данных, мы по известным входам легко получаем выходы.

Очевидны ограничения для решения поставленной проблемы минимизацией с помощью взвешенного метода наименьших квадратов: не решается обратная задача, не учитываются неопределенности входов, значительный спектр поставленных задач будет плохо обусловлен, т. е. основная проблема состоит в том, что указанный функционал описывает довольно частный случай. Для удовлетворительного решения задачи заполнения пробелов в таблицах обычный метод регрессии, эмпирически аппроксимирующий искомую зависимость, неприемлем. В этом случае данные должны рассматриваться не как набор "задач-ответов", а как портрет явления – каждая строка таблицы является точкой, а в общем случае – эллипсоидом, в пространстве признаков.

Нейросетевая модель (или другое устройство) вырабатывает внутреннее представление о том, какие состояния корректны для заданного явления, а какие – нет, это может быть и функциональная зависимость одних обобщенных координат от других, и модель конечного автомата, и набор образов. На этапе решения обобщенной задачи устройство получает на вход эллипсоид в пространстве входов-выходов (строка с неопределенностями в некоторых компонентах) и ищет ближайшую к нему корректную точку. Естественно, результат поиска может и не быть геометрически самым близким, он зависит от реализации устройства, например, при использовании градиентного спуска вдоль градиента и будет получен первый попавшийся экстремум.

Конечно, решением обратной задачи, особенно многокритериальной, может быть и континиум, тогда может быть важным находить не только одну из подходящих точек, а различные статистические характеристики решения – распределение, меру и т. п. Итак, целью данной работы является разработка математических методов для возможно более независимой от пользователя (автоматичной) компьютерной системы обработки неточно определенных данных.

3. Принципы построения алгоритма обработки данных

В большинстве случаев в нейросетевом подходе используются оптимизационные принципы. Строится функционал, оценивающий качество нейросетевой модели, и оптимизируется градиентными, либо другими методами. Методы оптимизации довольно хорошо разработаны, но и не тривиальны. Основные проблемы состоят в быстродействии работы нейросети, сложном рельефе оценочной функции. Применение нейросетевого подхода и направлено на решение проблемы быстродействия.

Для оптимизации функции градиентными методами при большом числе параметров требуется находить множество частных производных, и с помощью метода множителей Лагранжа удается довести время нахождения градиента до CT, где T – время вычисления функции, традиционно, константа C порядка трех, в разработанном здесь алгоритме она приближается к единице, т. е. время одной итерации обучения примерно равно одной итерации функционирования.

Ландшафт оценочной функции обычно очень сложен, поэтому при недостаточной избыточности числа подстроечных параметров невозможно, используя градиентные методы, достигнуть глобального минимума. Для решения этой проблемы разработаны методы глобальной оптимизации, такие как метод динамических ядер, генетические алгоритмы, методы Монте–Карло. Однако существующая проблема переобучения показывает, что глобальный минимум обучающего функционала далеко не всегда обеспечивает оптимальное решение, регуляризационные методы и методы ранней остановки иллюстрируют это.

Для оценки экстраполяционных способностей обученных нейросетей типично разделение обучающей выборки на собственно обучающую и тестовую. С целью исключения эксперта из процесса обучения используется большое количество вариантов разделений выборки, при которых находятся в среднем оптимальные параметры регрессионной модели. Далее найденные параметры используются для построения окончательного варианта модели, причем в обучающей выборке участвуют все примеры задачника.

Например, в методе перекрестной проверки каждая задача поочередно перемещается из обучающей выборки в тестовую, поэтому вычисление обучающего функционала состоит из суммирования многих незначительно отличающихся значений. Можно, используя теорию возмущений, приблизительно вычислять обучающий функционал и его градиенты, что значительно увеличивает скорость обучения.

4. Обобщенный метод наименьших квадратов

В этой связи предлагается оптимизировать следующий функционал:

$$H = \sum_i^n \sum_j^m \left(\frac{x_{ij} - \tilde{x}_{ij}}{\delta \tilde{x}_{ij}} \right)^r + \left(\frac{f(\tilde{x}_i, \mathbf{p}) - \tilde{y}_i}{\delta \tilde{y}_i} \right)^r,$$

где i – номер эксперимента; y – экспериментальное значение; x – вектор переменных, для которых находится зависимость; $f(x, \mathbf{p})$ – аппроксимирующая функция; $\mathbf{p}$ – вектор настраиваемых параметров функции; δy – доверительный интервал по y; δx – доверительный интервал по x; n – количество задач; $m-$ размерность; r – четная степень ($r = 2$).

При оптимизации находятся не только уточненные объекты (y), но и их признаки (x). За это, конечно, приходится расплачиваться временными ресурсами, однако возможна реализация алгоритма, при котором дополнительные члены функционала приводят к набору независимых задач малой размерности.

В качестве предобработки над экспериментальными значениями и доверительными интервалами проводятся следующие действия: центрирование и нормирование x_i, где учитывается вес этой величины, обратно пропорциональный

доверительному интервалу. Одновременно этими же действиями оценивается линейная часть регрессии. В качестве базисной функции используется

$$a_{it} = b_i + c_i \sum_j \sin\Big(\varphi_{ij} + \sum_k w_{jk} X_{kt}\Big),$$

где X – входы, b, c, w, φ – подстраиваемые параметры (b и c, определяются при предобработке); a_{it} i-й выход задачи t; j меняется от единицы до числа нейронов.

Такой подход был предпринят из-за легкости вычисления производных, оценки интегральных характеристик (значимости входов, средней гладкости) и близости к интегральному преобразованию Фурье, для которого справедливы универсальность и т. п. Гладкость функции, а точнее *негладкость*, оценивалась как

$$U = \sum_j \sum_{k=1}^{n} w_{jk}^2.$$

На эту величину пользователем накладывается ограничение сверху, и при предобработке, если необходимо, матрица связей нормируется для удовлетворения заданного условия. Во время обучения подстраиваемые параметры w изменялись таким образом, чтобы не выходить за указанные пределы.

Для оптимизации использовался метод сопряженных градиентов, т. е. направление шага выбиралось линейной комбинацией антиградиента оценочной функции и шага на предыдущей итерации. На каждой итерации вдоль выбранного направления выбирался оптимальный шаг с помощью квадратичной аппроксимации оценочной функции. Из-за неточности такой аппроксимации, методом дихотомии выбирался не ухудшающий шаг в случае ошибочного определения оптимального. В процессе оптимизации выбираемое направление постепенно удаляется от наилучшего, поэтому на каждом шаге, номер которого равнялся числу подстроечных параметров, делался переход на направление антиградиента, т. е. использовался вариант алгоритма сопряженных градиентов Полака–Рибиера. Также использовалась ортогонализация сопрягаемых направлений, по возможности расчетные величины не пересчитывались, а к ним вычислялись добавки для оптимизации времени вычислений. Оценка преобразовывалась к виду, удобному для быстрого дифференцирования. На первом шаге выполняется линейная регрессия данных, поэтому оценка находится в интервале от нуля до единицы.

Как упоминалось выше, для наилучшего использования экстраполяционных возможностей регрессионных методов необходимо применение какого-либо из методов проверки получаемых моделей на данных, не включенных в обучающую выборку. С этой целью был разработан и проверен метод быстрого тестирования, основывающийся на методе перекрестной проверки. Оценкой правильности выбора глобальных параметров модели, таких как вид и сила ограничений, число нейронов и т. п., может служить ошибка модели на тестовой выборке, на тех примерах, где она не обучалась.

Довольно сложно оптимизировать глобальные параметры, так как модель приходится постоянно обучать заново. Существуют методы хороших разбиений задачника на подвыборки, однако оказалось возможным значительно ускорить

метод перекрестной проверки, используя теорию возмущений. Модификация других методов для автоматической разбивки задачника при сегодняшнем развитии вычислительной техники и способов программирования представляется нецелесообразным, поскольку значительное увеличение способов разбиения влечет за собой либо усложнение алгоритмов, либо снижение эффективности работы программы.

Можно заметить, что чем менее градиенты задач коррелируют между собой, тем лучше модель выдержит перекрестную проверку. На этом свойстве возможно дальнейшее ускорение за счет создания наиболее репрезентативного задачника (выбрасывая неважные задачи).

Данные результаты реализованы в пакете программ модели и опробованы в ряде численных экспериментах, которые подтвердили правильность высказанных предположений. Применение тестовых методов позволяет решить проблему "самодостаточности" в случае смешения входов и выходов, построения нейросетевых или иных функций, где "все является функцией от всего". Появляется возможность относительно полной автоматизации обработки данных, когда пользователю не надо задумываться даже о наличии и количестве причинно-следственных связей в данных, а нейросетевая модель сможет разобраться, что от чего зависит.

5. Визуализация данных сейсмического мониторинга методом упругих карт

В работе представлен подход для анализа сейсмичности региона средствами вычислительного эксперимента с целью оценки сейсмического риска и уточнения сейсмической опасности. Для сейсмоактивных зон территории Красноярского края решались задачи пространственно-временного анализа сейсмичности и моделирование сейсмогеологических данных на основе регрессионного подхода, а также эффективная визуализация полученных результатов методом упругих карт. Исходными данными для решения поставленных задач являлись каталоги сейсмических событий в регионе.

На первом этапе вычислительного эксперимента выполнялся пространственно-временной анализ сейсмичности на основе статистического анализа каталога землетрясений региона. Здесь осуществлялся поиск параметров пространственно-временного распределения сейсмических событий и долгосрочная оценка сейсмической опасности. Изучалось распределение сейсмических событий с учетом направления, а также распределение сейсмических событий с учетом времени.

Второй этап вычислительного эксперимента – регрессионное моделирование каталога землетрясений. Решением этой задачи являлось создание специализированного датчика случайных чисел, генерирующего параметры сейсмических событий, которое осуществлялось на основе регрессионного подхода. Датчик случайных чисел строился как регрессионное отображение из вспомогательного многообразия в многообразие реальных данных. При этом сначала определялось редкое распределение точек многообразия с общим количеством точек, соответствующим количеству зарегистрированных событий. Использовано пространство скрытых параметров в виде подмножества квадратной решетки, ограниченного кругом.

При моделировании использовался комплекс программ Модели. Как показано выше, программный комплекс Модели предназначен для оперативного синтеза по большим массивам экспериментальных данных аналитических моделей с регулируемым уровнем сглаживания этих данных. Программа работает в среде операционных систем MS Windows 95/98/2000/XP и Windows NT 4, а также в виде набора инструментальных функций вычислительной среды MatLab в UNIX и MacOS.

В математическом отношении программа осуществляет нелинейную многомерную регрессию. В качестве интерполирующего применяется один из вариантов многомерных представлений в виде интегралов Фурье с заменой интегралов конечными суммами. При оптимизации используется метод быстрого вычисления многомерных градиентов или метод множителей Лагранжа и вариант метода сопряженных градиентов.

Для расчета использовался вариант комплекса программ в виде надстройки к MS Excel. В ходе проведенных расчетов достигнутая точность аппроксимации данных каталога (координат эпицентра) составила 2,5–3 км. В качестве обучающих выборок был выбран каталог ИФЗ (В. И. Уломов и др.) и каталог КНИИГиМС.

Следующий шаг – визуализация связей геолого-геофизических факторов с сейсмичностью, где реализация алгоритма визуализации данных осуществлялась на основе метода упругих карт. Основным элементом модели упругой карты является упругая сетка, которая должна обладать: свойством близости к точкам данных; свойством упругости по отношению к растяжению; свойством упругости по отношению к изгибу.

Таким образом, решением поставленной задачи будет оптимизация следующего функционала:

$$D = \frac{D_1}{|x|} + \lambda\frac{D_2}{pq} + \mu\frac{D_3}{pq} \to \min,$$

где $|x|$ – число точек в X; λ, μ – коэффициенты упругости, отвечающие за растяжение и изогнутость сетки соответственно; D_1, D_2, D_3 – слагаемые, отвечающие за свойства сетки: D_1 – мера близости расположения узлов сетки к данным; D_2 – мера растянутости сетки; D_3 – мера изогнутости (кривизны) сетки.

В работе анализируется и сопоставляется результат визуализации суммарной выделившейся сейсмической энергии (с использованием метода упругих карт) и данные о геолого-геофизических факторах (и их визуализация) изучаемой области. Моделирование указанных данных с помощью регрессионного подхода позволило построить прогностическую модель сейсмической активности в регионе на основе регрессионной модели связи геолого-геофизических факторов и данных исследуемого каталога землетрясений.