

О замене дисперсионного анализа методом анализа углов наклона градиента аппроксимирующей функции.

Милехин С.А.
milekhinsa@gmail.com

Аннотация.

Для определения значимости влияния переменных на функцию вместо дисперсионного анализа предлагается метод, основанный на анализе угла наклона градиента функции к осям переменных и плоскостям, определяемым переменными.

Метод решает все задачи дисперсионного анализа при меньшем количестве экспериментальных данных.

Метод делает возможной новую концепцию планирования эксперимента, основанную на разделении факторного пространства на кластеры по степени влияния факторов на функцию.

Метод пригоден для извлечения экспертной информации (знаний) из анализа экспериментальных данных.

Ключевые слова:

Дисперсионный анализ, значимость факторов, ранжирование факторов по значимости, взаимодействия факторов, значимость взаимодействий факторов, параллельные опыты, план эксперимента, полный факторный эксперимент, дробный факторный эксперимент, неплановый эксперимент.

1. Введение

В 90-х годах автору пришлось заниматься созданием новых композиционных материалов (КМ), а именно через изменение состава материала получать КМ с нужными свойствами (задача состав — свойства).

Работа была главным образом связана со статистической обработкой результатов эксперимента и написанием программного обеспечения.

Опирались пришлось на отчёты по научной работе института и текущие результаты разработчиков КМ.

Эта информация была мало подходящей для классического статистического анализа, который предполагает план эксперимента, оценку значимости факторов, составление на этой основе уравнения регрессии и т. д.

Разработчик КМ, проверяя пришедшую ему мысль, для быстрого результата порою ставил эксперимент без плана и без параллельных опытов.

Хотелось использовать и эту информацию, хотя она и «неправильная».

Конечно разработчик ставил экспериментальные точки из своих соображений но для статистической обработки это просто случайные точки.

Можно ли задействовать всю имеющуюся экспериментальную информацию в том числе и эти случайные точки для оценки значимости факторов как будто они были поставлены по плану и с параллельными опытами?

Появилась идея оценки значимости факторов через градиент аппроксимирующей поверхности.

Мы можем аппроксимировать нужную функцию по всем имеющимся экспериментальным точкам.

Что есть градиент функции нескольких переменных?

Это вектор, указывающий направление наиболее быстрого возрастания функции, модуль которого характеризует скорость этого возрастания.

Если функция аппроксимирована в пространстве факторов, то по определению градиента, градиент в этом случае есть есть вектор с координатами частных производных функции по факторам [1, с. 394].

Угол наклона градиента к оси фактора определяет степень влияния этого фактора на функцию, а угол наклона градиента к плоскости, определяемой сочетанием факторов определяет степень влияния этого сочетания факторов на функцию.

Если это так, то по поведению аппроксимирующей поверхности можно ранжировать факторы и их взаимодействия (сочетания факторов) по значимости их влияния на функцию отклика.

Для определения степени влияния факторов и их взаимодействий на функцию существует дисперсионный анализ.

По определению Г. Фишера дисперсионный анализ это отделение дисперсии, приписываемая одной группе причин от дисперсии, приписываемой другим группам [2, с. 75].

У Г. Шеффе [3, с. 13] дисперсионный определяется как статистический метод анализа результатов наблюдений, зависящих от различных одновременно действующих факторов, выбор наиболее важных факторов и оценка их влияния.

Т.е. дисперсионный анализ это статистический метод, основанный на анализе дисперсий, имеющий свой математический аппарат и требующий определённое количество экспериментальной информации.

Предлагаемый метод решает все вопросы, которые решает дисперсионный анализ, а именно даёт полное решение задачи ранжирования значимости влияния факторов и их взаимодействий (сочетаний факторов) на функцию отклика.

Причём такое решение можно получить при меньшем количестве экспериментальной информации, что в случае создания КМ имело принципиальное значение из-за времени и дороговизны эксперимента.

2. Постановка задачи.

Нужно из имеющихся случайных экспериментальных точек извлечь то же количество информации, которую может дать ПФЭ с параллельными опытами.

Сформулируем задачу в наиболее общем виде.

Есть некая наблюдаемая функция, зависящая от N факторов x_i :

$$y = f(x_1, x_2, \dots, x_N).$$

Необходимо выяснить, какие факторы больше влияют на изменение функции в точке $x_1^j, x_2^j, \dots, x_N^j$, какие меньше, а также выяснить, какие взаимодействия факторов — двойные $x_i x_j$, тройные $x_i x_j x_k$ и, вообще говоря, какие возможные сочетания факторов $C_m^N = N! / (N - m)! m!$,

где $1 < m < N$ существенно влияют на функцию в рассматриваемой точке, а какие нет.

Полным решением задачи для точки $x_1^j, x_2^j, \dots, x_N^j$ будем считать составление упорядоченного по степени влияния на функцию в данной точке ряда, составленного из всех факторов и всех их сочетаний.

Назовём этот ряд рядом значимостей в точке.

Отметим, что ряд значимостей в точке сам есть функция точки $x_1^j, x_2^j, \dots, x_N^j$ и для различных областей поверхностей может меняться в зависимости от вида функции $y = f(x_1, x_2, \dots, x_N)$.

Решением задачи для всей области определения функции $y = f(x_1, x_2, \dots, x_N)$ будет совокупность всех рядов значимостей для каждой точки области определения функции.

Сегменты в области определения функции, которые имеют одинаковые ряды значимости, то есть сегменты с одинаковой степенью влиянием факторов на функцию назовём кластерами.

Определять точность решения будет точность аппроксимации функции, которая зависит от количества и расположения экспериментальных точек. В нашем случае расположение точек случайно а количество — сколько есть.

3. Решение задачи.

Для изучения значимости влияния факторов и их взаимодействий (сочетаний) на функцию $y = f(x_1, x_2, \dots, x_N)$ в статистике используется дисперсионный анализ. С его помощью можно решить задачу в самой широкой постановке, т. е. отделить значимые факторы а так же все их значимые взаимодействия (сочетания факторов) от незначимых, а также упорядочить их по степени значимости, т.е. в нашей терминологии составить ряд значимостей, что и является решением задачи для области, в которой проводится дисперсионный анализ.

Отметим что построение ряда значимости в дисперсионном анализе возможно не для точки, как в предлагаемом методе, а для области факторного

пространства, где планируется эксперимент.

Но для полного решения задачи методом дисперсионного анализа предъявляются жёсткие требования к данным.

Во-первых, это обязательная постановка эксперимента по плану, использование существующей разрозненной неплановой информации исключается.

Во вторых, для полного решения задачи необходима постановка полного факторного эксперимента (ПФЭ). Анализ результатов такого эксперимента, т. е. разделение и сравнение дисперсий от разных факторов и их сочетаний и даёт ответ на поставленный вопрос.

Количество экспериментов для ПФЭ будет равно I^N , где I – количество уровней эксперимента, N - количество факторов, причём строго говоря в каждой точке необходимы параллельные эксперименты...

Для проектирования композиционных материалов такое количество эксперимента было нереально из-за дороговизны эксперимента и сроков его проведения. Наверное та же картина будет и для большинства других практических задач.

Сокращение же количества экспериментов с применением дробного факторного эксперимента даёт возможность оценки значимости только для факторов, но исключает оценку значимости взаимодействий факторов, что делает решение задачи неполным.

Таким образом применение дисперсионного анализа на практике проблематично.

Предлагаемый метод решает все эти проблемы

Если мы имеем ряд экспериментальных точек $y^j = f(x_1^j, x_2^j, \dots, x_N^j)$, то имеем возможность аппроксимации функции методом наименьших квадратов и получаем некоторую поверхность. Насколько корректна будет эта аппроксимация в зависимости от количества точек и их взаимного расположения — отдельный вопрос.

Но, имея уравнение поверхности, мы можем для каждой её точки определить углы наклона градиента к осям — это будут характеристики влияния факторов на скорость изменения функции, т. е. есть возможность определить степень влияния или значимость каждого фактора на функцию по величине угла наклона.

Аналогично мы можем определить угол наклона градиента к двумерным плоскостям $x_i x_j$, что будет определять значимость двойного взаимодействия, к трёхмерным плоскостям $x_i x_j x_k$, что будет определять значимость тройного взаимодействия, и к $C_m^N = N! / (N - m)! m!$ - мерным плоскостям, что будет определять значимость влияния любого сочетания факторов на функцию.

Таким образом по величине угла наклона мы имеем возможность упорядочения

факторов и любых их сочетаний по степени влияния на функцию, т.е. без постановки ПФЭ, а лишь с помощью аппроксимации для каждой точки построить полный ряд значимостей, т.е. получить более полное решение задачи, чем позволяет дисперсионный анализ.

То есть если говорить о взаимодействиях, то у Г. Шеффе [3] и у С. Ахназаровой [2] рассматриваются максимум тройные взаимодействия факторов.

Предлагаемый метод позволяет рассматривать все взаимодействия.

Учёт множественных взаимодействий может менять картину влияния факторов.

Конечно, точность метода и правильность выводов будет целиком зависеть от того, насколько точна будет аппроксимация изучаемой функции.

В случае обработки существующей информации без возможности постановки дополнительного эксперимента придётся довольствоваться тем, что есть.

Если же такая возможность есть или предполагается с нуля планировать эксперимент с использованием предлагаемого метода, можно предложить следующую схему.

Количество экспериментальных точек определить из сметы эксперимента и расположить их равномерно.

Полученные экспериментальные данные аппроксимировать полиномом второй степени, как это рекомендуют в [2], что даст информацию о поведении функции и наличии особенностей.

Определить границы кластеров, если кластеры есть.

Идеальной будет возможность проведения повторного эксперимента для уточнения поведения функции, локализации особенностей, если они есть, улучшения аппроксимирующей функции и уточнения границ кластеров, если они есть, т. е. реализовать последовательное улучшение аппроксимирующей поверхности.

Точность аппроксимации предлагается определить по аналогии со сходимостью численного решения дифференциального уравнения — если при уменьшении шага решение практически не меняется значит задача решена.

В случае определения достаточной точности аппроксимации прибавление одной или нескольких экспериментальных точек не должно менять характер влияния факторов на функцию, т.е. ряды значимостей и границы сегментов не должны существенно изменяться.

Метод был опробован при разработке опытного КМ и показал свою состоятельность.

Что-то из вышесказанного было реализовано практически, что-то осталось как идея для будущего.

Было создано программное обеспечение, состоящее из базы данных по составам и свойствам композиционных материалов и программы обработки информации, в которой для произвольного количества факторов методом наименьших квадратов производился расчёт коэффициентов выбранной аппроксимирующей

функции.

В базе помечались нужные данные и отправлялись на обработку.

Для выбранных точек определялся градиент и ряд значимости. На основании этих данных выбирались точки для дальнейшего поиска.

Было доказано, что при использовании аппроксимации в переменных ряда значимости остаточная дисперсия была наименьшей из всех других вариантов.

Эффект уменьшения остаточной дисперсии при исключении малозначимых факторов говорит о улучшении интерполяционных свойств полинома [2, с.149].

Ещё одно приложение метода это возможность получения экспертной информации или генерация правил для базы знаний из анализа рядов значимости.

Например правило « Увеличение доли коротких волокон в композиции ведёт к увеличению прочности на изгиб» получается при анализе ряда значимости функции «прочность на изгиб», где фактор «доля коротких волокон», будет самым значимым.

Была мысль использовать использовать предложенную методику для построения экспертной системы по проектированию КМ путём обработки базы данных по составам и свойствам КМ, был создан прототип генератора правил, который анализируя ряд значимостей заполнял правилами базу знаний, написанную на языке ПРОЛОГ.

Простой прототип работал, но сама задача была большой и сложной для практического применения, запроса на экспертную систему не было, а сам институт в середине 90-х перестал существовать.

Но сама идея анализа аппроксимирующей поверхности для получения экспертной информации может быть интересной для Data Mining при малом количестве информации.

Выводы.

1. Предложенный метод делает возможной новую концепцию планирования эксперимента как построения достаточно точной аппроксимирующей поверхности.
2. Предложенный метод полностью решает задачу дисперсионного анализа при меньшем количестве информации. И информации любой, а не только плановой. Поэтому он представляется предпочтительным дисперсионному анализу как более универсальный и эффективный.
3. Предложенный метод даёт более точные вычисления для функции отклика в координатах ряда значимости.
4. Предложенный метод пригоден для извлечения экспертной информации из

анализа экспериментальных данных.

Список литературы:

1. Фихтенгольц Г.М. Курс дифференциального и интегрального исчисления- М.: Наука, 1969, т.1
2. Ахназарова С.Л., Кафаров В.В. Методы оптимизации эксперимента в химической технологии. — М.: Высшая школа, 1985
3. Шеффе Г. Дисперсионный анализ.-М., Наука, 1980