СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ И «ИНТЕРНЕТ ВЕЩЕЙ» — ВОСПОМИНАНИЕ О БУДУЩЕМ

ВАЛЕРИЙ МИЛЫХ
СЕРГЕЙ СТЕПАНОВ
v.milykh@quarta.ru

В последнее время концепция «Интернет вещей» (Internet of things, IoT) является одной из самых популярных тем для обсуждения. Однако, как это часто бывает у новых понятий, «звучное» определение может несколько опережать реальное состояние дел и иметь слегка искаженную трактовку.

Безусловно, развитие систем сбора и обработки информации происходит стремительно, непрерывно растет количество приборов, имеющих разнообразные датчики. Они обмениваются информацией с обрабатывающими центрами и друг с другом. Еще недавно «безмолвные» приборы и устройства получают системы датчиков и контроллеров, производящих колоссальные объемы информации и выбрасывающие их в сети передачи данных. Сами сети передачи данных стали всеобъемлющими, затрагивают практически все сферы человеческой деятельности. Централизованные и децентрализованные системы управления, в том числе и в режиме реального времени, обрабатывают эти потоки информации и формируют управляющие воздействия. Это реальный, стремительный прогресс: системы становятся все более управляемыми, контролируемыми и взаимосвязанными.

И все-таки есть ощущение, что чего-то не хватает. Увеличение скорости обработки данных, увеличение производительности обрабатывающих центров, даже рост «управляемости» различных технологических систем — все это не может быть глобальной конечной целью.

Позволим себе высказать предположение об отсутствии одного важного элемента — целостного видения ситуации, видения многообразия процессов взаимодействия различных систем и составляющих их элементов. Представляется, что реальная цель иная — создание удобной, надежной среды окружения, обладающей прогнозируемым поведением, самостоятельно решающей целый ряд задач и рационально использующей имеющиеся ресурсы во всем их сочетании.

Что дает основание сделать вывод о сугубо «технологическом» (не целостном) подходе в современном восприятии «Интернета вещей»? Понимание того, что складывающаяся в настоящее время ситуация, как это ни парадоксально, не столь нова. Нечто подобное, конечно с оговорками, но уже случалось в различных областях экономики в 80-е годы и связано это было с результатами научно-технической революции и внедрением вычислительной техники.

ИСТОРИЧЕСКИЙ ЭКСКУРС

Чтобы лучше понять суть происходящих сегодня процессов, сделаем краткий экскурс в прошлое: рассмотрим некоторые аспекты развития весьма сложной и важной отрасли — энергетики.

Итак, уже в 80-е годы XX в. стало понятно: для обеспечения надежной работы энергосистемы необходимо обеспечить возможность сбора и обработки информации о состоянии энергосистемы, создать механизмы управления топологией системы, сформировать эшелонированные системы управления. В то время построение такого набора систем, в частности информационно-управляющих, было весьма серьезной инженерной задачей. Однако получаемый эффект того стоил. Нельзя не вспомнить и тот факт, что Единая энергетическая система того времени была крупнейшей в мире, централизованной и требовала для обеспечения своей работы не менее серьезных усилий.

Безусловно, объем информации и степень вовлеченности различных элементов системы был количественно иной, однако, с точки зрения качественной оценки, ситуация имеет много схожего. Расчетные модели того времени уверенно показали, что наличие информации о потоках мощности, возможность управления перетоками энергии за счет изменения топологии энергетической сети обеспечивали существенную экономию, а наличие информационно-вычислительных управляющих систем обеспечивало прогнозируемую надежность функционирования. Все это направило вектор движения в сторону управляемых, обладающих способностью к оптимизации электрических сетей. Аналогичные работы проводились и в смежных областях энергетики: управление и оптимизация режимов работы генерирующих мощностей, оптимизация управления транспортировкой топлива для электрических станций и формирование его запасов и т. п.

Однако практика расчетов показала, что невозможно выполнять растущие требования по дальнейшему росту эффективности и надежности работы энергосистемы только путем «арифметического» роста количества анализируемых показателей. Более того, стало понятно, что существуют ситуации, когда большое количество данных, в частности, с разной степенью достоверности (некий «информационный» шум) может приводить к неверным выводам и последующим неверным управляющим воздействиям. Это привело к необходимости поиска методов анализа достоверности данных, к поиску оптимального набора показателей (главных показателей), которые оказались бы адекватны текущему состоянию системы и набора которых было бы «достаточно» для полноценного ее описания.

Не менее остро стояла и задача унификации описания параметров различных элементов системы для их последующего накопления и обработки, и необходимость визуализации результатов расчетов — отображение топологии и визуализация полученных результатов. Отдельным направлением стало создание методов оперативной коррекции моделей описания объектов. Очень часто объект не есть некая статическая совокупность, это развивающаяся система. Могут добавляться новые компоненты и подсистемы, взаимное влияние которых не всегда возможно учесть в инженерных расчетах (скрытые/неявные зависимости). Важное место занял и подход, связанный с формированием ресурсной модели объекта наблюдения (в данном случае энергосистемы) как набора ресурсов, имеющих свои граничные значения и диапазоны изменения, именно совокупность которых может быть использована для достижения общей целевой функции: оптимизации поведения объекта наблюдения при заданных/имеющихся ограничениях.

Итак, сделаем вывод из этого небольшого исторического экскурса. Многие из описанных задач были успешно решены с помощью использования статистических методов — методов построения статистических моделей объектов наблюдения.

СОВРЕМЕННАЯ СИТУАЦИЯ

Возвращаясь в наше время, можно уверенно констатировать схожесть стоящих сегодня задач с описанными выше:

- унификация данных;
- обработка больших объемов данных;
- выявление недостоверных данных и главных показателей;
- построение скоординированных моделей описания объектов;
- построение гибких моделей описания объектов.

Это позволяет нам применить ранее накопленный опыт использования статистических моделей и, в первую очередь, моделей фазовых состояний.

Как было показано ранее, отдельные инженерные подсистемы, датчики и блоки исполнительной автоматики могут иметь существенно отличающиеся структуры предоставляемых данных. Цель же статистической системы (в части наблюдения и регистрации фазовых состояний системы в целом) заключается в едином, целостном подходе к организации данных о наблюдаемых объектах для последующего статистического моделирования и управления.

Для выполнения данного условия существуют серьезные предпосылки: наличие отраслевых стандартов и унифицированных протоколов, специализированных серверных решений (например, OPC-серверы, выступающие в качестве шлюзов для сбора и первичной обработки данных, поступающих от разнородных инженерных систем). Мы же в описываемом подходе имеем в виду высшую степень унификации данных — на уровне унифицированного функционального описания инженерных систем.

Следующим важнейшим условием описываемого подхода является оценка существенности (главные показатели) наблюдаемых значений факторов, их взаимосвязи в виде статистической модели, в привязке к выбранным критериям, к значениям которых и необходимо привести наблюдаемую систему при помощи управляющих воздействий на исполнительные механизмы инженерных подсистем.

Главная же цель — на основе интегрированных данных наблюдений построить адекватную статистическую модель, включающую в себя описание всех инженерных подсистем, с оценкой существенности факторов и возможностью регламентированной автоматической корректировки модели по мере накопления данных по фазовым состояниям наблюдаемой системы.

Не менее важен и выбор критериев оптимальности поведения всей системы в целом. В применении к идеологии «Интернета вещей» решается задача обеспечения достижения или максимального приближения целевой функции к заданным критериям статистической модели, вне зависимости от конкретных используемых инженерных решений и систем. Накопление и последующий статистический анализ накопленной информации позволяют качественно иначе анализировать взаимосвязь инженерных систем, выявить скрытые зависимости (через физические факторы среды) и использовать эти зависимости, обеспечить понимание тенденций развития процессов во времени.

Применяемый метод статистического моделирования может основываться как на линейных, так и на нелинейных статистических моделях с редукцией количества существенных факторов (к примеру, не более шести, так как шесть факторов, без ограничения общности, покроют до 98% статистической достоверности и точности). Применимыми можно признать методы главных компонент, линейной и нелинейной регрессии, регрессии Кокса и проч. Отбор наиболее приемлемого метода можно осуществить как на этапе проектирования вручную, так и автоматически, на этапе построения и/или корректировки статистической модели по мере увеличения базы данных накопленных фазовых состояний в процессе эксплуатации. Выбор способа зависит только от ресурсных возможностей обеспечения базы данных фазовых состояний среднего уровня.

Основой и ядром функциональности всей системы является модуль построения и коррекции статистической модели фазовых состояний относительно критерия, выбранного в качестве цели оптимизации на основе значений факторов, оцениваемых в процессе построения статистической модели на степень существенности их влияния на выбранный критерий (критерии).

Однако для некоторых гипотез по применимости тех или иных методов статистического моделирования необходимо доказательство о значимости принадлежности наблюдаемых значений факторов к классу нормальных распределений. Эти доказательства, по необходимости, будут основываться на критериях нормальности по значениям тестов Колмогорова–Смирнова, Лиллиефорса и W-критерия Шапиро–Уилка.

В действительности для инженерных подсистем со сложным поведением гипотеза о нормальности распределения значений факторов оказывается очень часто недостоверной. В силу ряда причин, таких как инерционность процессов, переходные состояния и другие особенности, инженерные подсистемы демонстрируют в своем поведении значения факторов, распределенных зачастую по логнормальному закону, а в пределе иногда и по экспоненциальному. Подбор формы распределения на этапе проектирования и опытной эксплуатации покажет, насколько близки в исследуемой ситуации функции распределения к логнормальной форме или распределению Вейбулла в качестве обобщения этого класса распределений, что важно при выборе применяемых критериев и способов проверки гипотез, а также наличие цензурированных данных, то есть данных, в составе которых заведомо исключены недопустимые значения факторов (сенсорных датчиков), а это именно так, поскольку данные фазовых состояний берутся из реальных ситуаций функционирования инженерных подсистем. Значимая аппроксимация значений критерия распределением Вейбулла позволит утверждать, что в отношении исследуемых выборок фазовых состояний и оценки гипотез о рисках неадекватности правомерно применение теста Уилкоксона–Гехана.

Оценить степень влияния факторов на выбранный критерий можно, построив регрессионную модель и уравнение регрессии Кокса, коэффициенты которого при переменных, назначенных факторам, покажут направление влияния и его силу, а уровни статистической значимости (p<0,05) проверки нулевой гипотезы (о равенстве коэффициента нулю, т. е. отсутствии влияния этого фактора) покажут, насколько можно доверять суждению о степени и градиенте влияния. Коэффициенты больше единицы указывают на ухудшающее влияние соответствующих факторов относительно достижения оптимального значения критерия, а значения, меньшие единицы, — на улучшающее влияние. «Удаленность» показателя от единицы в ту или другую сторону показывает степень влияния фактора.

Уровень значимости всего уравнения p может оказаться меньше задаваемого уровня статистической значимости 0,05. Это будет означать,

что подобранные коэффициенты в уравнении регрессии Кокса достаточно точно описывают распределение значений критерия. Уровни значимости по отдельным факторам позволяют доверять (считать статистически значимыми) показателям коэффициентов влияния только для факторов со значениями p, меньшими заданного уровня значимости. Уровни значимости остальных факторов, значительно превосходящие заданный уровень 0,05, указывают на то, что влияние этих факторов на значения выбранного критерия на фоне других факторов значимо не определяется на имеющемся материале наблюдения фазовых состояний. Но это не значит, что мы можем утверждать, что такого влияния нет в действительности.

Предлагаемый подход в виде статистического моделирования фазовых состояний и отбора оптимальных по выбранному критерию состояний для формирования управляющих воздействий на исполнительные механизмы неоднородных инженерных подсистем является, по сути, методом апостериорного оценивания.

Априорные методы выбора форм и видов зависимостей с критерием неоднородных инженерных подсистем чреваты двумя существенными недостатками.

Во-первых, нет достоверной уверенности в том, что выбранная заранее формула взаимосвязи поведения подсистемы полно, адекватно и однозначно отображает ее влияние на выбранный критерий оптимальности системы в целом.

Во-вторых, любые изменения параметров и/или состава инженерных подсистем потребуют полного перепроектирования всей системы под новые условия и структуру с новыми сомнениями в адекватности по первому пункту.

Апостериорный подход статистического оценивания такими недостатками не обладает.

Поскольку статистическая модель критерия оптимальности с оцениванием существенности факторов строится динамически, по мере накопления базы данных фазовых состояний системы, полнота, адекватность и достоверность зависят только от качества применяемого алгоритма построения статистической модели.

Масштабируемость статистической модели не имеет ограничений общности в смысле появления новых инженерных подсистем или изменения их характеристик и основывается на регламентированном единообразии представления данных и управляемых элементов инженерных подсистем.

* * *

Таким образом, описываемый подход позволяет использовать статистические модели описания объектов для построения скоординированного описания модели, автоматическую модификацию модели при изменении критериев оптимальности, включение новых инженерных систем в цикл наблюдения и управления простым включением в список схожих, производить анализ развития систем во времени и прогнозировать развитие ситуаций, и, в конечном итоге, мы придем к «умной» технологической среде окружения человечества. Впрочем, и это — только очередной «шажок» человечества, аналогичный тому, что был сделан ранее при переходе от пара к электричеству. ⬢