ТЕХНОЛОГИЯ НЕЧЕТКОГО ПРОГНОЗИРОВАНИЯ ХАРАКТЕРИСТИК СЛОЖНЫХ ОБЪЕКТОВ И СИСТЕМ

Виталий Снитюк, Сергей Говорухин

***Аннотация**: В статье выполнена общая постановка задачи прогнозирования выходной характеристики сложных объектов по измеренным значениям входных факторов. Осуществлен анализ технологии решения поставленной задачи с использованием множественной линейной регрессии и методов кластеризации. Предложена альтернативная технология, использующая аппарат нечеткой логики, и выполнен сравнительный анализ результатов.*

***Ключевые слова**: прогнозирование, регрессия, кластеризация, нечеткая логика.*

***ACM Classification Keywords** I.2 Artificial Intelligence*

Введение

В современном машиностроении важную роль играют задачи измерения различных характеристик, играющие важную роль при диагностике состояния колесных пар, контроле износа стенок трубопроводов, прогнозировании жесткости металлических пластин и т. д.

Разработка технологии повышения эффективности и качества измерений остается актуальной научной проблемой. На практике повышение качества измерений достигается использованием большого количества измерений. Сложность задачи точного измерения характеристик пластин обусловлена следующими причинами: большая площадь измерений, сравнительно небольшая площадь контакта сенсора и объекта на измеряемом участке, ошибки измерений, вносимые субъектами, которые их выполняют, ошибки, обусловленные случайными факторами.

Анализ релевантных источников

Для решения задач такого типа традиционно используются методы кластеризации и регрессии. Поскольку данные о сложных объектах и процессах имеют многоплановый и разнотипный характер, то это значительно усложняет, а часто делает невозможным выявление и построение их адекватных моделей. С целью преодоления таких трудностей на первом этапе необходимо выполнить идентификацию топологической структуры таких объектов. Для этого применяют методы кластеризации, основными недостатками которых являются [Gordon, 1999]: директивное априорное предположение о количестве кластеров, сложность выбора оптимальной метрики для определения близости объектов и кластеров, необходимость большого количества парных сравнений объектов, недостаточное статистическое обоснование результатов кластеризации и соответственно значительное количество возможных вариантов разбиений на кластеры.

На втором этапе для решения задач идентификации зависимостей традиционно используют методы линейной множественной регрессии (ЛМР). Использование данного метода базируется на ряде допущений и ограничений, главным из которых является предположение о наличии линейной зависимости в многомерной структуре данных между выходной характеристикой и входными факторами. Однако такое предположение не всегда является правомерным.

Уменьшить влияние указанных недостатков традиционных методов при разработке технологии прогнозирования характеристик сложных объектов предлагается использовать методы нечеткой логики.

Постановка задачи

Рассмотрим постановку задачи в общем виде.

Пусть Y – результирующая характеристика, $X_1,...,X_n$ – входные факторы. С помощью оригинального сенсора проведены измерения n параметров m объектов. Для каждого объекта осуществлено k измерений. Таким образом, получим обучающую выборку данных, представленную матрицей M, которая имеет $n+1$ столбцов и $m \cdot k$ строк (табл. 1).

Таблица 1

№ объекта	№ измерения	Значения входных факторов			Выходная характеристика
		X_1	…	X_n	Y
1	1	x_{11}^1	…	x_{1n}^1	y_1^1
	…	…	…	…	…
	k	x_{k1}^1	…	x_{kn}^1	y_k^1
…	…	…	…	…	…
m	1	x_{11}^m	…	x_{1n}^m	y_1^m
	…	…	…	…	…
	k	x_{k1}^m	…	x_{kn}^m	y_k^m

Предположим, что информация, представленная в табл. 1, утеряна. Причиной этого могла бы стать также поломка или утеря сенсора. Тогда возникает необходимость использования другого сенсора, что приводит к осуществлению значительного количества измерений, причем значения результирующей характеристики оказываются смещенными.

Сделаем упрощающие предположения. Пусть при использовании других сенсоров получены малые выборки данных при небольшом количестве измерений для объектов. Априорно считаем, что значения Y утрачены, за исключением:

1. l объектов, $l << m$, причем для каждого объекта выполнено по одному измерению.
2. r объектов, $r <<< m$, для которых выполнено по несколько измерений p_r.

Тогда $l \approx p_r \cdot r$. Причем и в первом, и во втором случае значения Y известны.

Необходимо по результатам измерений восстановить значения Y для всех новых объектов.

Технология решения задачи с использованием МЛР

Рассмотрим следующий алгоритм построения прогнозирующей модели МЛР при известных допущениях.

Шаг 1. Рассчитать матрицу парных корреляций входных факторов: $M = cor(X_i, X_j)$.

Шаг 2. Выполнить процедуру удаления зависимых факторов по одному из алгоритмов:

1. Исключить те факторы, для которых $|cor\ (X)| > K > 0$, где K - пороговое значение.
2. Исключить те факторы, для которых $\sum_i |cor(X_i, X_j)| = \max_j \sum_i |cor(X_i, X_j)|$, $i \neq j$, а также те, для которых $\sum_i |cor(X_i, X_j)| > L$, где L – некоторая константа.

Шаг 3. Вычислить вектор корреляций выходной характеристики с каждым из оставшихся входных факторов: $cor(Y, X_i)$.

Шаг 4. Выбрать p факторов, для которых $|cor(Y, X_i)| > l_y$, где l_y – положительная константа, причем $p = 7 \pm 2$.

Шаг 5. Выполнить проверку оставшихся факторов на мультиколлинеарность.

Шаг 6. По данным обучающей выборки для выбранных p факторов построить уравнение множественной линейной регрессии: $Y_1 = F(X_1, X_2, ..., X_p)$.

Далее используем данные, полученные при использовании другого сенсора.

Шаг 7. Разыграть l равномерно распределенных случайных чисел.

Шаг 8. Выбрать соответствующие строки из таблицы.

Шаг 9. По МНК построить второе уравнение: $Y_2 = F_2(X_1, X_2, ..., X_p)$.

Шаг 10. Построить графики функций Y_1 и Y_2, проходящие через l выбранных точек (пример приведен на рис. 1).

Шаг 11. Далее определить значения ошибок регрессии Y_2 одним из двух способов:

1. Рассчитать значение ошибки как среднее ошибок во всех точках: $d = \frac{1}{l}\sum_{i=1}^{l} d_i$, в этом случае значение ошибки будет одним и тем же во всех точках.
2. Запомнить вектор значений ошибок во всех точках: $D = (d_1, d_2, ..., d_l)$, в этом случае каждой точке соответствует единственное значение ошибки.

В обоих случаях $d_i = |Y_1(O_i) - Y_2(O_i)|$, $i = \overline{1,l}$.

Шаг 12. Для нового объекта выполнить измерения входных факторов (получаем новое измерение O') и подставить их значения в Y_2. Тогда в зависимости от выбранного способа на шаге 11 соответственно рассчитываем значения Y_1:

1. $Y_1 = Y_2 + d$;
2. Во второй выборке найти то измерение, которое наиболее «близко» к данному в пространстве выбранных признаков. Для этого использовать один из известных методов кластерного анализа [Мандель, 1988], [Айвазян, 1989], [Jain, 1988]:
   а) выбрать метод расчета расстояния между измерениями – меру близости объектов: $q(O_i, O_j)$;

b) в малой обучающей выборке определить измерение O_i, расстояние до которого от данного измерения O' минимально: $q(O',O_i)=\min_j q(O',O_j)$, $i=\overline{1,l}$, $j=\overline{1,l}$.

c) тогда $Y_1=Y_2+d_i$.

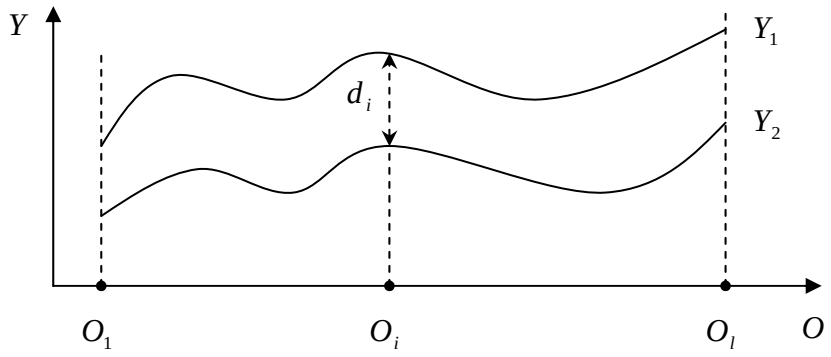

Рис.1. Графики ЛМР, построенные по данным первой выборки (Y_1) и по фрагменту другой выборки (Y_2)

Нечеткая идентификация

Используем для решения поставленной задачи методы теории нечетких множеств.

Пусть A_{ij}, $i=\overline{1,m}$, $j=\overline{1,n}$, - нечеткие термы (лингвистические переменные (ЛП)), которые строятся на множестве значений фактора X_j объекта i, то есть на множестве значений $x_{1j}^i,...,x_{lj}^i$, $l=\overline{1,k}$, которое является универсальным множеством данной ЛП. Пусть $T_{ij}=(T_1,T_2,...,T_s)$ - терм-множество ЛП A_{ij}, каждому элементу которого соответствует ФП $\mu_{T_i}(X_j)$, $i=\overline{1,s}$ (рис. 2).

В качестве функций принадлежности для элементов терм-множества предлагается использовать такие: $\mu_{T_1}(X)$ и $\mu_{T_s}(X)$ - линейные, $\mu_{T_i}(X)$, $i=\overline{2,s-1}$ - треугольного вида с параметрами (a,c).

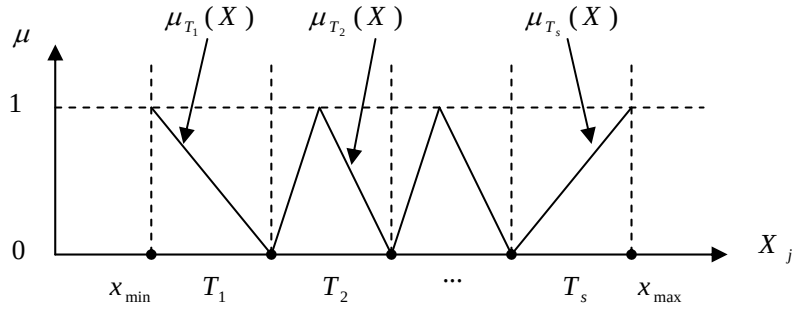

Рис. 2. ФП терм множества ЛП

Формируем нечеткую базу знаний (НБЗ). Пусть модель прогнозирования выходной характеристики задается набором правил R_i, $i=\overline{1,m}$:

$$Если\ X_1 \in A_{i1},\ X_2 \in A_{i2},\ X_n \in A_{in}\ то\ Y_i = C_i,$$

где A_{ij}, $i = \overline{1,m}$, $j = \overline{1,n}$ – ЛП, $C_i = const$.

Так как в данной НБЗ антецеденты заданы нечеткими множествами, а консеквенты заданы константами, то нечеткий логический вывод (НЛВ) производится по синглтонной базе знаний, которая является частным случаем НЛВ Сугено [Штовба, 2007].

В соответствии с правилами НЛВ Сугено получаем следующий алгоритм прогнозирования выходной характеристики Y при заданном входном объекте $X' = (X'_1, X'_2, ..., X'_n)$:

Шаг 1. Рассчитать степени выполнения посылки i-го правила для входного вектора X':

$$\mu_i(X') = \min\{\mu_1(X'_1), \mu_2(X'_2), ..., \mu_n(X'_n)\},\ i = \overline{1,m}.$$

В данном случае использована операция минимума (t-норма).

Шаг 2. В результате выполнения предыдущего шага для всей НБЗ получаем нечеткое множество, которое соответствует входному вектору X':

$$\tilde{y} = \left(\frac{\mu_1(X')}{C_1}, \frac{\mu_2(X')}{C_2}, ..., \frac{\mu_m(X')}{C_m} \right).$$

Шаг 3. Выполнить дефаззификацию $\tilde{y} \to y$ одним из способов:

1. Найти взвешенное среднее: $y = \dfrac{\sum_{i=1}^{m} \mu_i(X') \cdot C_i}{\sum_{i=1}^{m} \mu_i(X')}$;

2. Найти взвешенную сумму: $y = \sum_{i=1}^{m} \mu_i(X') \cdot C_i$.

Далее необходимо решить задачу построения модели для прогнозирования характеристик объектов при малой обучающей выборке, полученной с помощью нового сенсора. Один из подходов заключается в использовании НБЗ, в которой заключения заданы в виде линейных целевых функций. Основной недостаток данного подхода заключается в допущении о линейной зависимости выходных характеристик.

Другой подход базируется на идее использования методологии кластерного анализа. В [Аверкин, 1986] предлагается метод варьирования прототипов, который относится к косвенным методам построения ФП для одного эксперта. Пусть P - прототип, который характеризуется параметрами $x_1, ..., x_n$. Введем меру расстояния между объектом и прототипом: $Q(P, X) = \|P - X\|$. Для оптимизации вычисления расстояния от объекта до разных прототипов вводится штрафная функция $d(P)$. Пусть имеем m прототипов $P_1 ... P_m$. Тогда для каждого объекта вычисляется функция расстояния до ближайшего прототипа: $sim(O_i) = \min_j (Q(O_i, P_j) + d(P_j))$, $j = \overline{1,m}$.

Значения ФП объекта O прототипу P вычисляются по формуле:

$$\mu_P(O_i) = 1 - \frac{sim(O_i)}{\max_j sim(O_j)},\ j = \overline{1,m}.$$

Таким образом, для каждого прототипа получим ФП объектов прототипу. Для вычисления выходной характеристики для каждого нового объекта $O' = (x_1, x_2, ..., x_n)$ находим взвешенное среднее:

$$Y' = \frac{\sum_{i=1}^{m} \mu_{P_i}(O') \cdot Y_i}{\sum_{i=1}^{m} \mu_{P_i}(O')}.$$

Заключение

Предложенная технология решения поставленной задачи, которая базируется на использовании статистических методов, множественной линейной регрессии и кластерного анализа, является классической для решения задач такого типа. Идея использования аппарата нечеткой логики является перспективной [Асаи, 1993] и позволит повысить точность идентификации и прогнозирования жесткости пластин. В этом направлении планируется проводить дальнейшие теоретические и практические исследования.

Один из основных недостатков практического применения полученной нечеткой модели заключается в необходимости построения большой НБЗ. Например, в задаче прогнозирования жесткости металлических пластин $n = 73$, то есть одно правило состоит из 73 предусловий, а всего обучающая НБЗ имеет 98 правил. Решение задачи при таких условиях требует больших вычислительных затрат. Поэтому актуальной становится задача снижения размерности входных факторов при построении нечеткой модели, одним из возможных решений которой является использование методов нахождения главных компонент.

Перспективными для решения поставленной задачи являются также использование методов искусственного интеллекта, таких как нейронные сети и генетические алгоритмы.

Библиография

[Gordon, 1999] A.D. Gordon. Classification. – Boca Raton, 2nd ed., CRC Press LLC, 1999.

[Наконечный, 1998] С.И. Наконечный, Т.О. Терещенко, Т.П. Романюк. Econometrия – К.: КНЕУ, 1998.

[Мандель, 1988] И.Д. Мандель. Кластерный анализ. – М.: Финансы и статистика, 1988.

[Айвазян, 1989] Прикладная статистика: Классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1989.

[Штовба, 2007] С.Д. Штовба. Проектирование нечетких систем средствами MATLAB. – М.: Телеком, 2007.

[Аверкин, 1986] А.Н. Аверкин и др. Нечеткие множества в моделях управления и искусственного интеллекта / Под ред. Д.А. Поспелова. - М.: Наука, 1986.

[Асаи, 1993] Прикладные нечеткие системы / Под ред. Т. Тэрано, К. Асаи, М. Сугэно. – М.: Мир, 1993.

Информация об авторах

Снитюк Виталий *– Черкасский государственный технологический университет, зав. кафедрой информационных технологий проектирования; бул. Шевченко, 460, Черкассы, Украина;*
e-mail: snytyuk@gmail.com

Говорухин Сергей – *Черкасский государственный технологический университет, аспирант факультета информационных технологий и систем; бул. Шевченко, 460, Черкассы, Украина;*
e-mail: govorukhin@gmail.com