ПРИМЕНЕНИЕ НЕЧЕТКОГО МОДЕЛИРОВАНИЯ В ЗАДАЧЕ ИМПУТИРОВАНИЯ ДАННЫХ

Лучкова С.В., Перемитина Т.О., Ященко И.Г.

Федеральное государственное бюджетное учреждение науки Институт химии нефти Сибирского отделения Российской академии наук (ИХН СО РАН), г. Томск

sric@ipc.tsc.ru

Применение нечеткого моделирования в задаче импутирования является актуальной задачей. Использование нечеткой системы вместо статистических моделей позволяет тщательней относиться к обрабатываемым данным и повысить точность выполнения решаемой задачи. В настоящей статье предложен алгоритм импутирования данных с применением нечеткой системы на основе метода эволюционной стратегии и исследовано влияние параметров нечеткой системы на процесс импутирования.

1. Введение

В настоящее время технология нечеткого моделирования является одной из развивающихся областей обработки данных. Применение нечетких систем в задаче импутирования вызывает ряд вопросов касательно параметров системы: как правильно выбрать параметры построения базы правил, функции принадлежности, метода эволюционной стратегии, оптимизирующей параметры нечеткой системы, и другие. Данные вопросы очень важны, и являются объектами рассмотрения в данной работе.

Статья организована следующим образом. В разделе 2 вводятся базовые понятия, и рассматривается параметрическая идентификация нечеткой системы. В разделе 3 описывается задача импутирования данных. В разделе 4 представлены вычислительные эксперименты. Раздел 5 содержит заключение и выводы по работе.

2. Параметрическая идентификация нечеткой системы

Нечеткое моделирование осуществляется посредством системы нечеткого вывода, которая выполняет следующие действия [1, 2]:

1) система преобразует числовую информацию в лингвистические переменные (формирует базу правил);

2) система обрабатывает лингвистическую информацию, выполняя логические операции нечеткой конъюнкции, импликации и агрегации правил;

3) система формирует численные результаты.

2.1. Нечеткая система. В работе используется нечеткая система типа синглтон, в которой n входных переменных (антецеденты), m нечетких правил, каждое из которых имеет следующий вид:

ЕСЛИ $x_1=A_{1j}$ ***И*** $x_2=A_{2j}$ ***И*** … ***И*** $x_n=A_{nj}$ ***ТО*** r_j,

где r_j — действительное число, $r_j \in \Re$,

Aij - лингвистический терм, описываемый функцией принадлежности.

Нечеткая система осуществляет отображение $F:\Re^n \to \Re$, заменяя оператор нечеткой конъюнкции произведением, а оператор агрегации нечетких правил — сложением, получаем выходное значение $F(\mathbf{x})$:

$$F(\mathbf{x})=\frac{\sum_{j=1}^{m} r_j \cdot \prod_{i=1}^{n} \mu_{Aij}(x_i)}{\sum_{j=1}^{m} \prod_{i=1}^{n} \mu_{Aij}(x_i)}, \quad (1)$$

где $\mathbf{x} = [x_1,\ldots,x_n]^T \in \Re^n$ - значение i-го входа,

$\mu_{Aij}(x_j)$ - функция принадлежности лингвистического терма A_{ij},

r_j – значение выходного значения (консеквента) в *j*-м правиле.

2.2. Функция принадлежности. Существует свыше десятка типовых форм кривых для задания функций принадлежности. Наибольшее распространение получили: треугольная, трапецеидальная и гауссова функции принадлежности. В работе для идентификации используется треугольная функция, которая определяется тройкой чисел *(a, b, c)* и ее значение в точке *x* вычисляется согласно выражению:

$$\mu_{A_{ij}}(x_j) = \begin{cases} 1 - \dfrac{b-x}{b-a}, a \le x \le b \\ 1 - \dfrac{x-b}{c-b}, b \le x \le c \\ 0, в\ остальных случаях \end{cases} \quad (2)$$

При *(b-a)=(c-b)* имеем случай симметричной треугольной функции принадлежности, которая может быть однозначно задана двумя параметрами из тройки *(a, b, c)* [1].

2.3. Параметрическое представление. Нечеткая система может быть представлена как

$$y = f(\mathbf{x}, \boldsymbol{\theta}),$$

где $\theta = ||\theta_1,\ldots, \theta_N||$ – вектор параметров,

$N = k$ (число параметров, описывающих одну функцию принадлежности) * t (число термов, описывающих одну входную лингвистическую переменную, – является задаваемым параметром нечеткой системы);

y – скалярный выход системы.

Например, если взять n входных переменных, определённых на t термах, и треугольную функцию принадлежности, то для модели типа синглтон, вектор параметров будет выглядеть следующим образом:

$$\boldsymbol{\theta}_n = [a_{11}b_{11}c_{11}\ldots a_{1t}b_{1t}c_{1t}a_{21}b_{21}c_{21}\ldots a_{2t}b_{2t}c_{2t}\ldots a_{n1}b_{n1}c_{n1}\ldots a_{nt}b_{nt}c_{nt}],$$

где a_{ij}, c_{ij}, b_{ij} – параметры треугольной функции принадлежности формулы (2), *i*-й лингвистической переменной *j*-го терма. Параметры, входящие в данный вектор влияют на адекватность модели.

Параметрическая идентификация рассматривается как процесс оптимизации нечеткой модели, который сводится к нахождению таких параметров нечеткой системы, чтобы ошибка вывода была минимальной. При этом оценивается качество нечеткого вывода по значениям ошибки вывода, разницы между значениями выходной переменной из таблицы наблюдений $f(\mathbf{x})$ и значениями $F(\mathbf{x})$, полученными нечеткой системой. Ошибку вывода необходимо минимизировать[1], для этого используются:

1) среднеквадратичная ошибка (СКО): $\dfrac{\sqrt{\sum_i^N (f(\mathbf{x}_i) - F(\mathbf{x}_i, \boldsymbol{\theta}))^2}}{N}$.

2) средняя абсолютная ошибка (САО): $\dfrac{\sum_i^N |f(\mathbf{x}_i) - F(\mathbf{x}_i, \theta)|}{N}$

Задача параметрической идентификации - это определить параметры нечетких правил путем оптимизации работы нечеткой системы. Методы идентификации делятся на два типа: методы, которые используют производные от параметров нечеткой системы и

метаэвристические методы. К методам, использующие производные, относятся: градиентный метод, метод наименьших квадратов и другие числовые методы [1]. Ко второму же типу относят: генетический алгоритм [3], эволюционные стратегии [3, 4, 5], алгоритм муравьиной колонии [6]. В данной работе используется один из метаэвристических методов, а именно метод эволюционной стратегии.

2.4. Эволюционная стратегия. Эволюционная стратегия это эвристический метод оптимизации в разделе эволюционных алгоритмов, основанный на адаптации и эволюции. Стратегия основана на механизмах естественного отбора и наследования. В ней используется принцип выживания наиболее приспособленных особей. Преимущества алгоритма перед другими методами оптимизации заключаются в параллельной обработке множества альтернативных решений [4, 5].

Алгоритм работает с популяцией особей (хромосом), каждая из которых представляет собой упорядоченный набор параметров задачи, подлежащих оптимизации. Основной характеристикой каждой особи является ее мера приспособленности. Общий алгоритм для эволюционной стратегии выглядит так:

Общий алгоритм эволюционной стратегии.

Шаг1. Генерируется начальная популяция $P(0)$, устанавливается $i=0$.

Шаг2. Повторяется до тех пор пока не выполнится условие остановки:

(а) Рассчитывается приспособленности каждой хромосомы из популяции $P(i)$.

(б) Применяются генетические операторы, такие как скрещивание, мутация к родительской популяции и производятся потомки.

(в) Отбираются хромосомы для следующего поколения $P(i+1)$ из полученных потомков и возможно родителей.

В настоящее время имеются разные варианты метода эволюционной стратегии, чаще всего используют следующие:

1) $(\mu+\lambda)$ - эволюционная стратегия (набор родителей и потомков);

2) (μ,λ)- эволюционная стратегия (набор только потомков).

Рассмотрим свойства и сформируем алгоритмы стратегии. В $(\mu+\lambda)$ - эволюционной стратегии, μ-родителей может участвовать в воспроизводстве λ-потомков. Тогда $(\mu+\lambda)$-поколение будет уменьшен до μ-потомков следующего поколения селекцией. Основным преимуществом данного подхода является непринужденное использование адаптивных стратегических параметров. Однако есть и недостатки, а именно - частое «застревание» в локальном оптимуме. Например, в алгоритме это выражается так - поколение-родитель не дает поколение-потомок лучшее себя.

Для предотвращения этого недостатка был предложен метод (μ,λ)- эволюционной стратегии, где селекция подчинена условию $\lambda > \mu$. Предыдущие μ-родители будут полностью заменены, и не будут использоваться в следующем поколении. Недостатком этого алгоритма является то, что «лучшие» из μ-родителей могут быть заменены более «худшими» сгенерированными λ-потомками и будут потеряны, а это в итоге может дать не самый корректный результат. При выборе разновидностей метода эволюционной стратегии необходимо учитывать это свойство следующего поколения.

2.5. Алгоритм параметрической идентификации нечеткой системы. Алгоритм для настройки параметров нечеткой системы модели типа синглтон методом эволюционной стратегии выглядит следующим образом:

Вход: Таблица наблюдений.

Шаг 1. Загружаем таблицу наблюдения.

Шаг 2. Задать параметры нечеткой системы:

(а) количество термов;

(б) количество генерируемых хромосом для начальной популяции μ;

(в) параметр функции приспособленности, для построения базы правил.

Шаг 3. Строим нечеткую систему:

(а) базу правил для обучения, на основе нечетких термов, равномерно распределенных по каждому входному параметру из таблицы наблюдения;

(б) формируем начальную популяцию хромосом $\boldsymbol{\theta}_n$;

(в) находим консеквенты для каждого правила методом "ближайшего" из таблицы наблюдение для базы правил;

(г) рассчитываем адекватность полученной нечеткой системы.

Шаг 4. Задаем параметры метода эволюционной стратегии:

(а) критерий остановки для идентификации параметров нечеткой системы (количество итераций);

(б) количество генерируемых хромосом (потомков) λ из начальной популяции;

(в) параметры оператора скрещивания (количество точек скрещивание; алгоритм скрещивание);

(г) вероятность мутации;

(д) алгоритма селекции;

(е) свойства следующего поколения.

Шаг 5. Запускаем идентификацию методом эволюционной стратегии.

Шаг 6. Рассчитываем адекватность нечеткой системы. Если достигнуто условие выхода *Шаг 7*, иначе *Шаг 5*.

Шаг 7. Вывод решения – «наилучшей» хромосомы.

Выход: Оптимизированная база правил. Значение ошибки нечеткого вывода для "наилучшей" хромосомы.

3. Задача импутирования

Важной задачей в обработке данных является импутирование или восстановление пропущенных значений. Чаще всего она возникает при идентификации зависимостей с неполной априорной информацией о значении параметров. Объективными причинами этого являются поломки оборудования при измерении значений технических характеристик процессов, потеря ретроспективной информации, экстремальный характер функционирования, ограниченный доступ и другие. При попытке "отбросить" свойства с пропущенными значениями, "заменить" средним значением по свойству или воспользоваться статистическими методами часто приводит к потере информации, к искажению результатов или недостаточно точным значениям результатов [7].

Рассмотрим особенности структуры таблиц наблюдения для использования предложенного метода в задаче импутирования (табл.1).

Пусть $X=(X_1, X_2, \ldots X_n)$- вектор входных параметров, m – количество записей в табл.1, $A=(a_{ij})_{i=1,j=1}^{m\ n}$ - матрица исходной информации.

Исходная информация имеет пропуски, обозначенные звездочками в табл. 1. Допускается по одному пропуску на запись, так как пропущенное значение будет являться выходным значением для данной записи.

Таблица 1 - Структура входной информации

	X1	X2	X3	...	Xn-1	Xn
1	a11	a12	a13	.	*	a1n
2	a21	a22	a23	.	a2n-1	a2n
...	.	.	.	.	.	.
m	am1	am2	am3	.	amn-1	amn

Таким образом, задача восстановления пропусков в данных заключается в определении выходного значения для каждой записи с пропуском, на основе всех записей, кроме той, в которой идет восстановление.

Алгоритм импутирования будет выглядеть так:

Вход: таблица наблюдений с пропусками в записях.

Шаг 1. Загружаем входные данные (таблицу наблюдения).

Шаг 2. Задаем параметры нечеткой системы.

Шаг 3. Выбираем параметры метода эволюционной стратегии.

Шаг 4. Применяем эволюционную стратегию.

*Шаг 5.*Отбираем лучшую хромосому.

Шаг 6. Восстанавливаем пропуск на основе сформированной базы правил и лучшей хромосомы.

Шаг 7. Проверяем условие остановки. Если достигнуто условие выхода *Шаг 8*, иначе *Шаг 4*.

Шаг 8. Выводим решения.

Выход: таблица наблюдений с восстановленными значениями.

4. Вычислительные эксперименты

Для исследования влияния параметров системы были взяты следующие данные: выборка из общей базы данных «О свойствах нефтей» Института химии нефти СО РАН, включающей описание более 21000 образцов [8, 9]. В ней выделен массив вязких парафинистых нефтей, содержащий 141 запись по 5 характеристикам. В "полную" таблицу были специально введены пропуски. Такой подход позволяет рассчитывать точность восстановления, так как мы можем сравнить полученный результат с искусственно "пропущенными" данными.

На процесс импутирование данных в первую очередь влияет параметр нечеткой системы, то есть количество лингвистических термов (или просто термов), от которых зависит база правил для обучения системы. Проведя ряд тестов, были получены результаты, представленные в табл. 2.

Таблица 2 - Исследование параметров нечеткой системы

Количество термов на параметр	Ошибка вычисления	
	Ср. квадратичная (СКО)	Ср. абсолютная (САО)
3	0,080946	0,550576
4	0,072553	0,620107
5	0,062454	0,468006
6	0,057758	0,415331
7	0,059252	0,489001

Как видно из табл. 2, наилучшее начальное решение приходится на 6 термов, а 7 термов уже не приводит к улучшению результата, такое поведение нечеткой системы объясняется тем, что для каждого интервала данных существует "разумное" количество термов, на которое следует производить разбиение входных интервалов переменных функцией принадлежности.

Для наглядности приведем пример: пусть у нас есть интервал данных от 0 до 1, тогда при равномерном разбиении на 5 и 7 термов получим результаты, представленные на рис. 1а и рис. 2б соответственно. При 7 термах (рис. 1б) видно, что "частокол" из треугольной функции принадлежности сильно частый, что в данном случае не оптимально. Это приведет к увеличению времени расчетов, так как возрастет количество правил, например, при 2 переменных по 5 термов база правил состоит из 25 правил, а при 5 переменных по 5 термах – из 3125 правил и т.д.

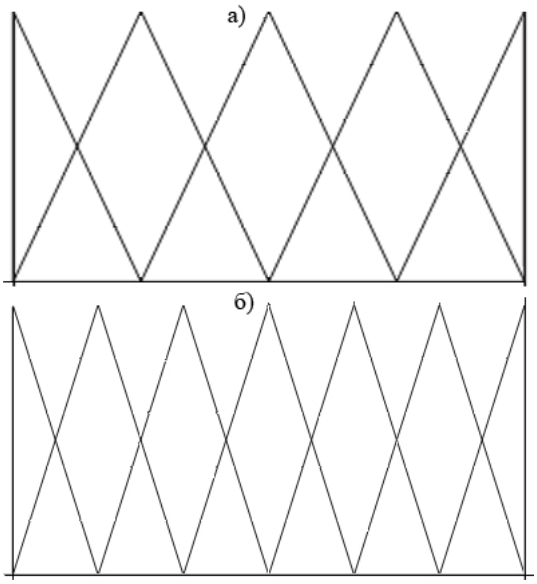

Рисунок 1 - Разбиение на термы

На построение базы правил влияет также свойство функции принадлежности - разрешение или запрет на выход за начальные границы интервалов входных данных. Это свойство больше влияет на настройку нечеткой системы, а не на выбор начальных параметров, так как первая хромосома в любом наборе строится по принципу равномерного разбиения внутри границ, поэтому это свойство рассмотрим в самом конце работы.

Далее, на процесс импутирования влияют параметры метода эволюционной стратегии, так как именно этим методом происходит настройка нечеткой системы или ее обучение. У эволюционной стратегии необходимо выбрать: алгоритм скрещивания (одноточечный, многоточечный или унифицированный), вероятность мутации, алгоритм селекции (турнирный отбор, случайный, рулеточный или элитарный), соотношение начального количества хромосом μ и генерируемых λ, и свойство следующего поколения ($(\mu+\lambda)$ или (μ,λ)). Результаты представлены в табл. 3-7.

Примечание: для экспериментов использовались следующие условия: 6 термов, разрешено выходить за начальные границы. Количество итераций для расчетов равно 1000, если не указано другое.

Таблица 3 - Исследование параметров эволюционной стратегии

Алгоритм скрещивания	**Усредненная ошибка вычисления**	
	СКО	**САО**
Алгоритм селекции: случайный отбор, вероятность мутации 0,07, свойство следующего поколения $(\mu+\lambda)$, μ = 20, λ = 40		
Одноточечный	2,55265	4,95727
Многоточечный (4 точки)	2,46616	4,85605
Унифицированный	1,99767	4,37111
Алгоритм селекции: турнирный отбор, вероятность мутации 0,07, свойство следующего поколения $(\mu+\lambda)$, μ = 20, λ = 40		
Унифицированный	2,14579	4,20841
Алгоритм селекции: рулеточный отбор, вероятность мутации 0,07, свойство следующего поколения $(\mu+\lambda)$, μ = 20, λ = 40		
Унифицированный	2,1356	4,57524
Алгоритм селекции: элитарный отбор, вероятность мутации 0,07, свойство следующего поколения		

$(\mu+\lambda)$, μ = 20, λ = 40		
Унифицированный	1,68311	3,3928

Как видно из табл. 3 унифицированный алгоритм скрещивания дал лучший результат при случайном отборе, поэтому для рассмотрения остальных алгоритмов отбора использовался только унифицированный алгоритм. Далее, наиболее точный результат в совокупности предоставил элитарный отбор, рассмотрим его более подробно. В табл. 4 представлены результаты влияния количества итераций на время вычисления.

Таблица 4 - Результаты влияния количества итераций на время вычисления

Количество итераций	Время вычисления (чч:мм:сс)	Усредненная ошибка вычисления	
		СКО	САО
0	-	2,52901	5,29846
50	00:04:43	2,30549	4,67326
100	00:09:27	2,26192	4,59192
500	00:47:51	2,21651	4,43482
1000	01:34:38	2,10361	4,09103
1500	02:23:46	2,09505	4,04241
2000	03:11:59	2,06241	3,85821

Из табл. 4 видно, что увеличение количества итераций работы алгоритма уменьшает ошибку, но увеличивает время работы системы. На время работы системы также влияет количество термов, так, например, если при 6 термах время вычисления 500 итераций составляет около 48 минут, то при 7 термах время уже около 90 минут.

Установлено, что вероятность мутации на время вычисления не влияет, однако влияет на точность работы. В табл. 5 представлен выбор вероятности мутации, для рассматриваемого примера наилучший результат дает вероятность мутации 0,15.

Таблица 5 - Результаты влияния вероятности мутации

Вероятность мутации	Усредненная ошибка вычисления	
	СКО	САО
0,03	2,21353	4,5846
0,07	2,12239	4,3686
0,09	2,15042	4,3765
0,10	2,14764	4,3898
0,15	2,05346	4,6121
0,20	2,20968	4,4008

В табл. 6 представлены данные анализа влияния соотношения начального количества хромосом μ к генерируемым λ и влияние свойства следующего поколения ($(\mu+\lambda)$ или (μ,λ)).

Таблица 6 - Влияние количества хромосом и свойства следующего поколения

Кол-во хромосом, и свойство следующего поколения	Усредненная ошибка вычисления	
	СКО	САО
(10+20)	2,18887	4,61617
(10+30)	2,17641	4,42053
(20+40)	2,18851	4,76471

(20+50)	2,06061	4,35609
(10, 20)	2,44669	5,06788
(10, 30)	1,95399	3,99733
(20, 40)	2,43172	5,18523
(20, 50)	2,22017	4,54274

Согласно исследованию, наименьшие ошибки расчетов получаются при комбинации (20+50) - эволюционной стратегии и (10, 30). Это подтверждает плюсы и минусы алгоритмов стратегий, описанных выше, и дает нам итоговые комбинации настроек. Следующий шаг – проверить, как же влияют свойства функции принадлежности (табл. 7).

Таблица 7 - Влияние свойства функции принадлежности

Свойство функции принадлежности	**Усредненная ошибка вычисления**	
	СКО	**САО**
Разрешено выходить за пределы начальных границ (Время вычисления 1000 итераций в среднем 95 минут)		
(20+50)	2,06061	4,35609
(10, 30)	1,95399	3,99733
Запрещено выходить за пределы начальных границ (Время вычисления 1000 итераций в среднем 155 минут)		
(20+50)	2,81386	4,14932
(10, 30)	3,27282	4,44925

Влияние свойств функции принадлежности значительно, при этом возросли и ошибки расчетов, и время вычисления. Это обусловлено тем, что комбинаций треугольников на ограниченном интервале меньше, а проверок их построения больше. Проверки построения имеют условия: разбиение термов должно покрывать весь заданный интервал, термы не должны перекрываться полностью и прочее. Следовательно, необходимо использовать алгоритм с разрешением выхода за начальные интервалы.

Итак, мы определили что, так как в алгоритме присутствует случайный компонент, как в построении нечеткой системы, так и в ее настройке, поэтому имеем две итоговые комбинации, каждая из которых может быть выигрышной в конечном итоге (табл. 8).

Таблица 8 - Итоговые комбинации параметров

Количество лингвистических термов нечеткой системы	6
Количество итераций	1000
Алгоритм скрещивания	Унифицированный
Алгоритм селекции	Элитарный
Вероятность мутации	0,15
Свойство следующего поколения	(20+50) и (10,30)
Свойство функции принадлежности	Разрешено выходить за начальные пределы

Заключение

В работе предложены алгоритм идентификации нечеткой системы на основе метода эволюционной стратегии для настройки параметров и алгоритм импутирования. На примере использования массива о свойствах вязких парафинистых нефтей были поставлены вычислительные эксперименты с искусственно «пропущенными» данными. Эксперименты показали, что для различных данных оптимальные параметры системы будут разными, и в каждом случае их необходимо подбирать. Поэтому было показано, как следует выбирать параметры нечеткой системы и эволюционной стратегии. Дальнейшие исследования

направлены на более детальный подход к обрабатываемым данным, например, возможность задавать количество термов для каждой переменной, что, возможно, сократит количество правил и снизит вычислительное время, а, следовательно, позволит обрабатывать одновременно выборки большего объема данных.

Литература

1. *Ходашинский И.А., Гнездилова В.Ю., Дудин П.А., Лавыгина А.В*. Основанные на производных и метаэвристические методы идентификации параметров нечетких моделей // Труды VIII международной конференции "Идентификация систем и задачи управления" SICPRO '08. Москва, 2009 г. С.501-528.
2. *S. Hoche, and S. Wrobel.* A Comparative Evaluation of Feature Set Evolution Strategies for Multirelational Boosting // Proc. 13th Int. Conf. on ILP, 2003 г.
3. *Рутковская Д., Пилиньский М., Рутковский Л.* Нейронные сети, генетические алгоритмы и нечеткие системы. М.: Горячая линия – Телеком, 2006.
4. *H.-P. Schwefel.* Numerical Optimization of Computer Models. John Wiley & Sons, 1981
5. *H.-P. Schwefel.* Evolution and Optimum Seeking. New York: John Wiley & Sons, 1995
6. *Дудин П.А*. Применение алгоритма муравьиной колонии для идентификации нечетких моделей // Материалы XLV Международной научной конференции «Студент и научно-технический прогресс» Информационные технологии. НГУ. Новосибирск, 2007. С. 188-189.
7. *Лучкова С.О.* Алгоритмы нечетких систем в задачах импутирования // Материалы III Всероссийской молодежной научной конференции «Современные проблемы математики и механики», г. Томск 2012. С.329-334.
8. *Козин Е.С., Полищук Ю.М., Ященко И.Г.* База данных по физико-химическим свойствам нефтей. Нефть. Газ. Новации. 2011. № 3. С. 13-16
9. *Ященко И.Г., Сваровская Л.И., Перемитина Т.О., Лучкова С.В.* Методы статистического анализа и нечетких систем в исследованиях влияния химического состава и условий залегания нефтей на численность и активность пластовой микрофлоры в задачах повышения нефтеотдачи // Химия нефти и газа: Материалы VIII Международной конференции, 24-28 сентября 2012 г., г. Томск. – Томск: Изд-во Томского государственного университета, 2012.С. 299-302