УДК 004.89

Н.М. Кораблев[1], Г.С. Иващенко[2]
ХНУРЭ, г. Харьков, Украина, [1]korablev.nm@gmail.com, [2]igs2005@rambler.ru

КРАТКОСРОЧНОЕ ПРОГНОЗИРОВАНИЕ ВРЕМЕННЫХ РЯДОВ, СОДЕРЖАЩИХ АНОМАЛЬНЫЕ ЗНАЧЕНИЯ, ПРИ ПОМОЩИ МОДЕЛЕЙ ИСКУССТВЕННЫХ ИММУННЫХ СИСТЕМ

В статье предложен метод прогнозирования временных рядов, содержащих аномальные значения, при помощи искусственных иммунных систем. Рассмотрена гибридная модель выявления аномальных значений и прогнозирования искаженных временных рядов на основе модели клонального отбора и метода вывода по прецедентам (CBR). Оценка эффективности модели выполнена путем сравнительного анализа, представлены результаты экспериментальных исследований, демонстрирующие особенности предлагаемого подхода.

ПРОГНОЗИРОВАНИЕ, ВРЕМЕННОЙ РЯД, АНОМАЛЬНОЕ ЗНАЧЕНИЕ, ИСКУССТВЕННАЯ ИММУННАЯ СИСТЕМА, ВЫВОД ПО ПРЕЦЕДЕНТАМ, МОДЕЛЬ КЛОНАЛЬНОГО ОТБОРА, АНТИТЕЛО, АНТИГЕН, АФФИННОСТЬ

Введение

Краткосрочное прогнозирование позволяет решить задачу определения будущего состояния различных систем на основе анализа уже имеющихся ретроспективных данных. Точность прогноза во многом определяется надежностью исходных данных, поэтому наличие аномальных выбросов в обрабатываемых ретроспективных данных является одной из проблем задачи прогнозирования, для решения которой необходимо использовать подход, позволяющий обрабатывать искаженные данные таким образом, чтобы они оказывали минимальное воздействие на результат работы прогнозирующей модели [1]. Анализ временного ряда после исключения аномальных значений может привести к получению некорректного прогноза, и как следствие, к принятию неверных решений [2].

Перспективным подходом является применение метода вывода на основе прецедентов (case based reasoning – CBR), в котором при рассмотрении новой задачи выполняется поиск подобного прецедента в предыстории в качестве аналога [4]. Если представить временной ряд как совокупность выборок, то прогнозирование при помощи CBR заключается в определении выборки, максимально соответствующей последним известным значениям временного ряда и последующей оценки его будущих значений. Выявление аномалий при помощи этого метода заключается в поиске в базе прецедентов выборки, подобной анализируемой. Если обнаруженная подобная выборка ранее была определена как содержащая аномальные значения, то делается вывод о присутствии аномальных выбросов и в анализируемой выборке. Недостатком данного подхода является требование базы прецедентов значительного объема, в то время как в реальных условиях набор исходных данных ограничен и часто не содержит примеров аномальных значений временного ряда.

В настоящее время активно развиваются подходы на основе методов вычислительного интеллекта, таких как искусственные нейронные сети и искусственные иммунные системы (ИИС), которые могут интегрироваться с другими подходами и отличаются быстродействием и адаптационными возможностями.

В [4] предложен подход, основанный на применении модели клонального отбора, использующей разнородные антитела, построенные на основе вывода по прецедентам и простейших методов прогнозирования, который использует сегментацию временного ряда и подбор для каждого участка наиболее эффективного метода прогнозирования. Однако данный подход не позволяет учитывать внешние факторы, влияющие на прогнозируемую величину, и получать прогноз для временных рядов, содержащих аномальные выбросы. В представленной работе получил дальнейшее развитие подход, предложенный в [4].

1. Постановка задачи

Набор данных, описывающий протекание какого-либо длительного процесса, может быть представлен в виде временного ряда $Z(t) = z_1, z_2, ..., z_N$ длины N, и S внешних факторов, представленных в виде рядов $X_1(t), X_2(t), ..., X_S(t)$, значения которых получены в моменты времени $t_1, t_2, .., t_N$. Набор последовательных значений $Z_t^L = z_t, z_{t+1}, ..., z_{t+L-1}$, лежащих внутри временного ряда, назовем выборкой из этого ряда, имеющей длину L, с моментом начала отсчета t, $L \in \overline{[1, N-1]}$, $t \in \overline{[1, N-L]}$. Аномальный выброс определяется как значение временного ряда, существенно отличающееся от других элементов выборки. Необходимо построить модель, которая позволит определять, содержит ли проверяемая выборка аномальные значения, и выполнять прогнозирование временного ряда в условиях его искажения аномальными значениями.

2. Выявление аномалий при помощи моделей ИИС

Существуют различные модели, основанные на принципах работы иммунной системы: модель клонального отбора, модель отрицательного отбора, модель иммунной сети и другие, которые можно использовать для решения задачи выявления аномалий и прогнозирования временных рядов.

Применение подхода на основе модели положительного отбора [5], заключается в генерации детекторов, соответствующих выборкам значений временного ряда, не содержащих аномальных выбросов. Метод включает в себя этапы обучения и распознавания. В ходе обучения случайным образом создаются детекторы, которые сопоставляются с выборками значений, составляющих анализируемый временной ряд. Если значение аффинности меньше заданного порогового значения, новый детектор удаляется из популяции.

Создание случайных детекторов и их последующий отбор проводится до формирования набора детекторов, описывающих анализируемый ряд. В процессе распознавания выборка принимается содержащей аномальные выбросы в том случае, если ее аффинность детекторам меньше порогового значения. Недостатком данного подхода на основе модели положительного отбора является необходимость обучающей выборки, включающей в себя информацию об аномальных выбросах [5].

В случае применения сочетания модели клонального отбора и CBR для решения задачи выявления аномалий [6], прецедент содержит выборку значений временного ряда и ее характеристику (содержит данная выборка аномальные выбросы, или нет). Модель клонального отбора основывается на поиске антител (вариантов решения), наиболее соответствующих антигену (поставленной задаче), основываясь на значении функции аффинности (мере близости между антителом и антигеном).

Представим популяцию антигенов в виде множества $Ag = \{Ag_1, Ag_2, ..., Ag_M\}$, где M – размер популяции антигенов, соответствующий количеству примеров в обучающей выборке.

Каждый элемент множества Ag – пример из обучающей выборки, проверяемый участок временного ряда, представленный в виде вектора фиксированной длины $Ag_j = \langle ag_1, ag_2, ag_3, ..., ag_L \rangle$, где j – индекс в популяции антигенов, L – число значений в составе антигена – длина выборки проверяемых значений ряда.

Популяция антител представлена в виде множества $Ab = \{Ab_1, Ab_2, ..., Ab_N\}$, где N – размер популяции антител. Антитело – основной элемент ИИС, в терминах подхода CBR исполняет роль прецедента, состоит из двух частей: $Ab_j = \langle ab_1, ab_2, ab_3, ..., ab_L, D_{Ab} \rangle$. Первая часть $ab_1, ab_2, ..., ab_L$ аналогична по структуре антигену, определяет меру близости антитела антигену, при помощи определения значения функции аффинности. Вторая часть антитела содержит его идентификатор (получаемый при создании), значение меры отклонения D_{Ab}, показывающей близость антитела к аномальному, $D_{Ab} \in (0,1]$, и идентификатор того антитела, на основании сравнения с которым определяется значение D_{Ab}. Чем ближе значение D_{Ab} к пороговому, тем вероятнее присутствие в составе антитела (той выборки, что формировала его первую часть) аномального значения. Данная часть не участвует в определении аффинности.

Аффинность – мера близости между антителом и антигеном, основной критерий отбора антител – определяется как

$$Aff = \frac{\sum_{k=1}^{L} \left(1 + |ab_k - ag_k|\right)^{-1}}{L} \in (0,1], \qquad (1)$$

Алгоритм выявления аномальных значений следующий:

1. Создание начальной популяции антител на основе значений ряда. При этом вторая часть созданных антител на данном этапе содержит только их собственный идентификатор. Предполагается, что созданные антитела не содержат аномальных значений. Неиспользуемые значения исходного ряда служат обучающей и контрольной выборками. При наличии тренда в исходном временном ряду рекомендуется использовать ряд приращений.

2. Формирование обучающей выборки. Возможна настройка модели как путем обучения с учителем (когда известно, какие антитела из обучающей выборки соответствуют аномальным участкам ряда), так и при помощи обучения без учителя. В том случае, когда изначально неизвестно, какие данные являются аномальными, ИИС самостоятельно формирует обучающую выборку путем вычисления матрицы аффинностей между антителами. Те из них, средняя аффинность которых наименьшая (меньше определенного порога, что говорит о том, что данные антитела существенно отличаются от остальных в популяции), принимаются за аномальные и составляют обучающую выборку.

3. Добавление в популяцию антител, содержащих аномальные значения. При этом происходит проверка аффинностей между ними и присутствующими в популяции антителами, значение D_{Ab} которых пока не задано или меньше единицы. В случае, если $D_{Ab'} * Aff < D_{Ab}$, где $D_{Ab'}$ – значение меры отклонения добавляемого в популяцию антитела, то $D_{Ab} = D_{Ab'}$.

4. Проверка антигена, о котором неизвестно, содержит аномальные значения, или нет. Выполняется отбор антител, имеющих наибольшую аффинность к антигену. Отбор так же проходят антитела, на которые ссылаются отобранные. Если значение D_{Ab} для антитела, имеющего наивысшую аффинность, выше порогового, принимается ре-

шение о присутствии в антигене аномальных значений. Вывод делается не об одном аномальном значении ряда, а об аномальной последовательности, поскольку проверяется весь антиген, а не его отдельная часть.

5. При наличии информации о том, является антиген аномальным или нет, следует коррекция модели, в зависимости от принятого решения:

6а. Антиген правильно определен как содержащий или не содержащий аномальные выбросы. Следует клонирование отобранных антител, применение ненаправленной обратно пропорциональной мутации, значение D_{Ab} клонов рассчитывается на основании аффинности с проверяемым антигеном или имеющимися в популяции антителами с наибольшими значениями меры отклонения.

6б. Антиген ошибочно определен как содержащий аномалии. Отобранные антитела клонируются и после этого устраняются из популяции. Значения D_{Ab} клонов рассчитывается на основании аффинности с присутствующими в популяции антителами, содержащими аномальные значения.

6г. Антиген ошибочно определен как не содержащий аномалий. В популяцию добавляется новое антитело, имеющее значение $D_{Ab}=1$, и происходит перерасчет меры отклонения для всех антител популяции.

7. При наличии нового проверяемого антигена происходит возврат к пункту 4.

Данный подход позволяет выявлять аномалии во временных рядах в условиях отсутствия примеров аномальных значений и недостаточной информации о критерии, по которому возможно различать нормальные и аномальные выборки значений временного ряда.

3. Гибридная модель прогнозирования временных рядов на основе модели клонального отбора

Прогнозирование на основе модели клонального отбора основано на поиске антител (вариантов решения), наиболее соответствующих антигену (поставленной задаче), основываясь на значении функции аффинности (мере близости между антителом и антигеном) [4-6].

Согласно гипотезе, сформулированной в [7], если мера подобия между выборками Z_t^L и Z_{t-k}^L имеет значение, близкое к единице, то мера подобия между выборками длины P, следующими за ними, Z_{t+L}^P и Z_{t-k+L}^P, также близка к единице. Тогда путем определения выборки, максимально соответствующей последним известным значениям временного ряда, возможна оценка его будущих значений.

Антиген представляет собой совокупность выборок значений прогнозируемого и сопутствующих рядов [8]. Выборка значений прогнозируемого ряда может включать в себя неизвестное число аномальных значений.

Антитело, используемое для решения задачи прогнозирования, состоит из двух частей. Первая часть $ab_1,ab_2,...,ab_L,ab'_1,ab'_2,...,ab'_L$ по структуре аналогична антигену (но включает в себя выборку значений только одного внешнего фактора $ab'_1,ab'_2,...,ab'_L$), представляет собой набор параметров, описывающих поставленную задачу (в нашем случае это выборки известных значений ряда, включая пропущенные значения) и используется при определении аффинности. Вторая часть $ab_{L+1},...,ab_{L+f}$, длина которой равна горизонту прогнозирования, не влияет на вычисляемое значение аффинности и описывает предлагаемый антителом прогноз для той выборки значений временного ряда, что составляет его первую часть.

Общее число антител в популяции при отсутствии пропущенных значений во временном ряду определяется следующим образом:

$$n=\sum_{i=L_{\min}}^{L_{\max}}\left(\left(N-\left(L_i+k\right)+1\right)N_m F\right), \qquad (2)$$

где N – число известных значений прогнозируемого временного ряда; L – длина части антитела, участвующей в определении аффинности, $L\in\left[L_{\min},L_{\max}\right]$; k – величина горизонта прогнозирования (длина выборки прогнозируемых значений ряда); N_m – число методов прогнозирования, используемых в модели [4]; F – число поддерживаемых моделью внешних факторов, потенциально влияющих на прогнозируемую величину.

Аффинность определяется с учетом значения меры отклонения D_{Ab}, и величины весовых коэффициентов для выборок, представляющих различные внешние факторы:

$$Aff\left(Ab\right)=D_{Ab}*\eta*\left(\eta_{Ab}*Aff_{Ab}+\eta_{Ab'}*Aff_{Ab'}\right), \qquad (3)$$

где η – значение коэффициента отбора; η_{Ab} и $\eta_{Ab'}$ – коэффициенты, определяющие влияние выборок исходного и сопутствующего рядов на аффинность антитела, при этом $\eta_{Ab}+\eta_{Ab}{}'=1$. В антигене могут быть не представлены некоторые внешние факторы, а в отдельном антителе представлена только одна сопутствующая выборка. Коэффициент отбора η предназначен для определения приоритета антител различных типов, т.к. антитела, созданные на основе одной и той же выборки, будут иметь одинаковое значение аффинности. Меры подобия выборок прогнозируемого временного ряда Aff_{Ab} и выборок рядов значений внешних факторов $Aff_{Ab'}$ определяется согласно (1).

В качестве результата (предлагаемого варианта прогноза) в данном поколении принимаются значения, входящие в состав антител $ab_{L+1},ab_{L+2},...,ab_{L+f}$, аффинность которых $Aff\left(Ab\right)\to 1$.

4. Получение прогноза и процесс обучения ИИС

На начальном этапе получения прогноза требуется определить аномальные значения в исследуемом временном ряду, согласно подходу, описанному в разделе 2. После определения аномальных

значений формируется антиген на основе выборки значений ряда, предшествующих прогнозируемым, и происходит создание популяции антител, обладающих аффинностью, выше пороговой. Результатом является прогноз того антитела среди них, которое имеет наибольшую аффинность к заданному антигену.

После получения реальных значений прогнозируемой величины происходит коррекция коэффициентов, влияющих на значения аффинностей антител. Приращение коэффициента отбора η получают антитела того типа, представитель которого среди популяции отобранных показал наименьшую ошибку прогноза, что дает преимущество при последующих отборах антител при определении аффинности, т.е. предпочтение будет отдано тем антителам, которые использовали для получения своего варианта прогноза метод, успешно показавший себя на предыдущих итерациях [4].

При $S > 0$ прогноз строится с учетом внешних факторов, представленных в виде сопутствующих временных рядов и происходит коррекция значений весовых коэффициентов, определяющих аффинность выборок, составляющих антитело. Коррекция весового коэффициента в ходе обучения ИИС позволяет снижать влияние того или иного внешнего фактора на предлагаемый вариант прогноза, путем вытеснения из популяции антител, которые включают в себя выборку значений наименее значимого внешнего фактора [8].

При определении аффинности приоритет отдается антителам, построенным на основе выборок с меньшим числом аномальных значений, и в процессе обучения шаблоны, имеющие аномалии, будут замещены близкими к ним, но с меньшим значением D_{Ab}.

Для антител, использующих CBR, применяется направленная прямо пропорциональная мутация, которой подвергается только та часть антитела, которая определяет его прогноз и не участвует в определении аффинности. Для антител, вычисляющих свой вариант прогноза самостоятельно, ненаправленной мутации подвергается только первая часть (и соответственно изменяется предлагаемый вариант прогноза), что частично решает проблему недостатка прецедентов в базе.

Обучение ИИС повторяется для каждого антигена из обучающей выборки заданное число раз, или до достижения некоторого заданного значения средней абсолютной ошибки. Антитела, имевшие наибольшее значение аффинности на каждой итерации алгоритма, становятся клетками памяти – шаблонами, описывающие анализируемый ряд.

5. Результаты сравнительного анализа

В ходе экспериментальных исследований было проведено краткосрочное прогнозирование рядов, используемых в M3-Competition [9], и сравнение результатов, полученных с помощью рассмотренного подхода (в таблице 1 – метод ClonAlg) с приведенными в [9] результатами прогнозирования неискаженных рядов с помощью экспоненциального сглаживания (Exp.Smooth), модели Хольта (HoltWinters), модели ARIMA. Симметричные средние абсолютные ошибки прогнозирования приведены в таблице 1, для ряда Meteo (более 20 тыс. значений среднесуточной температуры) указана средняя абсолютная ошибка (MAE, °C).

Таблица 1

Ошибка прогнозирования при использовании различных методов и разном количестве аномальных значений

Метод	N736 (44)	N1366 (63)	N2830 (104)	N2841 (104)	Meteo
Exp.Smooth	12,11	0,42	2,47	0,5	4,56
Exp.Smooth (10%)	12,5	0,45	2,48	0,59	4,7
HoltWinters	10,68	1,04	3,27	0,39	2,9
HoltWinters (10%)	12,04	1,18	3,51	0.44	3,13
Box–Jenkins	7,35	0,57	2,45	0,5	2,99
Box–Jenkins (10%)	7,85	0,65	2,48	0,53	3,35
ClonAlg	7,65	0,41	1,83	0,14	2,44
ClonAlg (10%)	9,74	0,91	2,34	0,21	2,49
ClonAlg (10%), cntrl	8,45	0,41	1,83	0,23	2,44
ClonAlg (20%)	8,26	0,56	3,31	0,53	3,77

Аномальные значения для экспериментов вносились путем увеличения или уменьшения случайно выбранных 10% значений ряда на 20% от $(z_{\max} - z_{\min})$. Для прогноза при помощи гибридного подхода анализировалось два варианта расположения аномалий – во всем ряду (пример на рис. 1), и только в контрольной части (для случая, когда была возможность перед использованием создать и обучить ИИС на неповрежденных данных). Кроме того, для предложенного подхода были проведены эксперименты с внесением 20% аномальных значений во временной ряд.

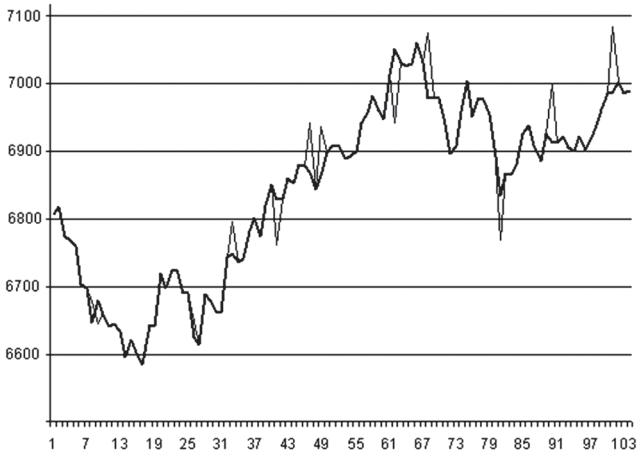

Рис. 1. Пример искаженного временного ряда (N2841)

Наличие аномальных значений ведет к увеличению ошибки на 5-8%, но если аномальные значения расположены только в контрольной части ряда (результаты для ClonAlg (10%), control), ошибка увеличивается только на 1-3% процента. Результаты прогнозирования подтверждают преимущество использования предложенного подхода на основе модели клонального отбора.

Вследствие выявления большинства аномалий, результаты, как правило, соответствуют результатам прогнозирования временных рядов, содержащих пропущенные значения [8]. Основной проблемой остается прогнозирование коротких временных рядов (результаты N736 и N1366), количество значений которых не позволяет создать популяцию антител достаточного размера, вследствие чего предложенный подход на основе модели клонального отбора уступает на таких рядах традиционным методам прогнозирования.

Выводы

Предложена гибридная модель краткосрочного прогнозирования искаженных временных рядов, содержащих аномальные значения, на основе модели клонального отбора. Данная модель отличается применением иммунных операторов для управления базой прецедентов, в качестве которых используются разнородные антитела, созданные на основе метода вывода по прецедентам и простейших методов прогнозирования.

В процессе предварительной обработки данных выполняется поиск аномальных значений прогнозируемого ряда при помощи подхода на основе модели клонального отбора. Возможна ее настройка как путем обучения с учителем, так и при помощи обучения без учителя, путем расчета матрицы аффинностей между антителами. В дальнейшем, в ходе получения прогноза, при определении аффинностей антител используются полученные на этапе поиска аномалий значения меры отклонения антитела. Данный подход позволяет выявлять аномалии во временном ряде в процессе решения задачи краткосрочного прогнозирования, что позволяет повысить точность получаемого прогноза.

Список литературы: 1. *Бучацкая, В. В.* Обработка аномальных значений уровней временного ряда как этап комплексной оценки информации в подсистеме прогнозирования для ситуационного центра [Текст] / В. В. Бучацкая // Вестник Адыгейского государственного университета, Сер.: Естественно-математические и технические науки. – 2013. – Вып. 3. – С. 98-102. **2.** *Литтл, Р. Дж. А.* Статистический анализ данных с пропусками / Р. Дж. А. Литтл, Д. Б. Рубин. – М. : Финансы и статистика, 1991. – 430 с. – ISBN 5-279-00443-X. **3.** *Черный, С. Г.* Применение case based reasoning для поддержки принятия решений [Текст] / С. Г. Черный // Вестник ХНТУ. – 2010. – № 2(38). – С. 336–342. **4.** *Кораблев, Н.М.* Гибридный метод краткосрочного прогнозирования временных рядов на основе модели клонального отбора [Текст] / Н.М. Кораблев, Г.С. Иващенко // 16-я всероссийская научно-техническая конференция с международным участием «Нейроинформатика-2014», 27-31 января 2014 р.: сборник научных трудов. – Москва, 2014. – Часть 1. – С. 79-89. **5.** *Кораблев, Н.М.* Обнаружение аномальных выбросов во временных рядах при помощи модели положительного отбора [Текст] / Н.М. Кораблев, Г.С. Иващенко, Т.В. Гайдамака // 2-я международная научно-техническая конференция «Проблемы информатизации», 12-13 апреля 2014 р.: тезисы докладов. – Киев, 2014. – С. 72. **6.** *Кораблев, Н.М.* Выявление аномальных значений во временных рядах при помощи модели клонального отбора, использующей вывод по прецедентам [Текст] / Н.М. Кораблев, Г.С. Иващенко // Сучасні проблеми правового, економічного та соціального розвитку держави, 22 листопада 2013 р.: тези доповідей. – Харьков, 2013. – С. 421-423. **7.** *Чучуева, И.А.* Модель экстраполяции временных рядов по выборке максимального подобия [Текст] / И. А. Чучуева // Информационные технологии. – 2010.– № 12.– С. 43–47. **8.** *Кораблев, Н.М.* Применение модели клонального отбора для прогнозирования временных рядов, имеющих пропущенные значения [Текст] / Н.М. Кораблев, Г.С. Иващенко // Электротехнические и компьютерные системы. – 2014. – Вып. 13 (89). – С. 170-177. **9.** *Makridakis, S.* The M-3 Competition: Results, Conclusions and Implications [Текст] / S. Makridakis, M. Hibon // International Journal of Forecasting. – 2000. – № 16.– P. 451–476.

Поступила в редколлегию 11.05.2015

УДК 004.89

Короткострокове прогнозування часових рядів, що містять аномальні значення, за допомогою моделей штучних імунних систем. / М.М. Корабльов, Г.С. Іващенко // Біоніка інтелекту: наук.-техн. журнал. – 2015. – № 2 (85). – С. 95–99.

У статті розглядається комбінований метод короткострокового прогнозування часових рядів за умов їх викривленості аномальними спостереженнями за допомогою штучних імунних систем. Запропоновано підхід на основі моделі клонального відбору та методу виведення з прецедентів. Був проведений порівняльний аналіз ефективності застосування запропонованої моделі та традиційних методів прогнозування часових рядів.

Табл. 1. Іл. 1. Бібліогр.: 9 найм.

UDK 004.89

Application of the models of artificial immune systems for short-term forecasting of time series containing abnormal values. /N.M. Korablev, G.S. Ivashchenko // Bionics of Intelligense: Sci. Mag. – 2015. – № 2 (85). – P. 95–99.

This paper proposes the combined method of short-term forecasting of time series containing abnormal values using artificial immune systems. A model of the prediction is based on the model of clonal selection and the case based reasoning method. Was performed a comparative analysis of the effectiveness of the proposed approach based on the immune model and the traditional methods of time series prediction.

Tab. 1. Fig. 1. Ref.: 9 items.