Научный вестник НГТУ. – 2013. – № 1(50)

УДК 519.95 + 004.93'14

Описание многомерных динамических процессов на языке иерархии концептов*

Н.Г. ЗАГОРУЙКО, И.А. БОРИСОВА, Д.А. ЛЕВАНОВ

Изучение структуры задач анализа данных, представленных таблицами типа «объект–свойство–время» (кубов данных) привело к заключению о целесообразности описания этих данных с помощью концептов разного иерархического уровня. После формирования словаря концептов появляется возможность порождать своего рода текст, описывающий сценарий динамического процесса. Для формирования концептов, их типизации и распознавании в потоке данных разработаны методы, основанные на использовании функции конкурентного сходства (FRiS-функции). Предложенные подходы тестировались на прикладной задаче распознавания одиннадцати состояний пациента по данным четырех датчиков, укрепленных на его теле. Результаты эксперимента показывают, что использование FRiS-функции позволяет надежно членить поток данных на участки, соответствующие концептам, и распознавать типы концептов.

Ключевые слова: динамические траектории, стационарные концепты, распознавание состояний, FRiS-функция.

ВВЕДЕНИЕ

Рассмотрение задач анализа данных, возникающих при изучении систем и процессов, описываемых таблицами типа «объект–свойство–время» (кубов данных) привело к заключению о целесообразности описания этих данных с помощью концептов разного иерархического уровня. Концептом здесь называется отрезок времени, на протяжении которого параметры динамического процесса подчиняются заданной закономерности. Для двумерных бинарных таблиц «объект–свойство» синонимом концепта является бикластер [1], т. е. набор строк и столбцов таблицы, на пересечении которых находятся единицы. В рассматриваемом нами случае задача усложняется тем, что признаки могут измеряться в более сильных шкалах и необходимо искать концепты с более сложными закономерностями, чем полное совпадение всех свойств у всех объектов. Переход к описанию некоторой динамической системы в терминах концептов позволяет представить ее поведение в упрощенном виде. Для формирования концептов, их типизации и распознавании в потоке данных были разработаны методы, основанные на использовании функции конкурентного сходства (FRiS-функции) [2].

Целесообразность такого перехода проверялась на задаче распознавания состояний системы. Такая задача часто возникает при анализе динамики объектов во времени. В ней каждый анализируемый объект может находиться в одном из нескольких различных состояний. На основе обучающей выборки, содержащей описания динамики одного объекта или ограниченного множества объектов, необходимо сформировать решающее правило, позволяющее различать исследуемые состояния и переходы объектов из одного состояния в другое. Эту задачу нельзя решать как классическую задачу распознавания, так как в ней нарушается предположение о независимости объектов обучающей выборки. Теория Марковских процессов к таким задачам также не всегда применима, так как в случае небольших объемов обучающих выборок, высокой оказывается погрешность при оценке параметров этих процессов.

Нами предлагается формальная постановка задачи распознавания состояний объектов, предлагаются простейшие подходы к ее решению, как с использованием предварительной разбивки потока данных на концепты, так и без. Эти подходы тестируются на прикладной задаче из области социальной медицины.

* *Статья получена 4 декабря 2012 г.*

Но прежде чем перейти к описанию методов, дадим краткое описание FRiS-функции – меры сходства, которая использовалась при выборе концептов и распознавании состояний.

1. ФУНКЦИЯ КОНКУРЕНТНОГО СХОДСТВА

Сходство в распознавании образов является категорией не абсолютной, а относительной. Чтобы ответить на вопрос «Насколько сильно объект z похож на объект a?», нужно знать ответ на вопрос «По сравнению с чем?». Адекватная мера сходства должна определять относительную величину сходства, зависящую от особенностей конкурентного окружения объекта z.

Все статистические алгоритмы распознавания учитывают конкуренцию между классами. В методе «k ближайших соседей» (kNN) новый объект z распознается как объект образа A, если среднее расстояние R_A до k ближайших объектов этого образа не только мало, но меньше, чем расстояние R_B до k ближайших объектов конкурирующего образа B. Оценка сходства в этом алгоритме делается в шкале порядка. Предложенная нами мера относительного сходства [2], которая оценивает сходство в абсолютной шкале, имеет следующий вид:

$$F(z, A|B) = [R(z, B) - R(z, A)]/[R(z, A) + R(z, B)]. \quad (1)$$

Здесь $R(z, A)$ и $R(z, B)$ – расстояния от объекта z до эталонов A и B, соответственно. Мы называем эту меру функцией конкурентного сходства FRiS (от Function of Rival Similarity). По мере передвижения объекта z от образа A к образу B можно говорить вначале о большом сходстве объекта z с образом A, об умеренном их сходстве, затем о наступлении одинакового сходства, равного 0, как с образом A, так и B. При дальнейшем продвижении z к B возникает умеренное, а затем и большое отличие z от A. Совпадение объекта z с эталоном образа B означает максимальное отличие z от A, что соответствует конкурентному сходству z с A равному –1.

Сходство в шкале порядка, используемое в методе kNN, отвечает на вопрос: «На эталон какого образа объект z похож больше всего?». Конкурентное сходство, измеряемое с помощью FRiS-функции, отвечает на этот вопрос и, кроме того, на такой вопрос: «Какова абсолютная величина сходства z с образом A в конкуренции с образом B?» Оказалось, что дополнительная информация, которую дает абсолютная шкала по сравнению со шкалой порядка, позволяет существенно улучшить методы анализа данных.

В случае, когда конкурирующий класс в явном виде не задан, применяется редуцированная функция конкурентного сходства (RFRiS) [3]. Идея RFRiS состоит в следующем. Объекты исходной выборки описаны N характеристиками. Добавляем к ним $(N + 1)$-ю характеристику, значения которой у всех объектах равны 0. Затем создаем виртуальную выборку B^* в виде клона исходной выборки с добавлением к нему $(N + 1)$-й характеристики, значения которой у всех объектов равны $r^* > 0$. Объявляем все объекты виртуальной выборки ее эталонами. В результате на расстоянии r^* от каждого объекта исходной выборки находится объект-конкурент из виртуальной выборки и можно вычислить редуцированное сходство F^* объекта z со своим эталоном A в конкуренции с виртуальным конкурентом

$$F^*(z, A|B^*) = [r^* - R(z, A)]/[r^* + R(z, A)]. \quad (2)$$

Редуцированная функция конкурентного сходства использовалась для решения задачи таксономии [4].

2. ФОРМИРОВАНИЕ ОДНОМЕРНЫХ КОНЦЕПТОВ

Пусть анализируется одномерный сигнал, представленный последовательностью значений амплитуды $a_0, a_1, \ldots, a_i, \ldots, a_T$ в моменты времени от 0 до T. Опишем процесс нахождения концепта стационарного типа. Под концептом стационарного типа будем понимать не-

расширяемый временной интервал $[t_l, t_r]$, на котором амплитуда объекта a изменялась незначительно, т. е. все реализации были похожи на какую-то типичную реализацию a_s, которую далее будем называть столпом концепта. При этом предполагается, что за пределами интервала изменения более существенны.

Формальное определение этого концепта через редуцированную функцию конкурентного сходства будет выглядеть следующим образом:

$$\sum_{i=t_l}^{t_r} F(a_i, a_s \mid r^*) \longrightarrow \max .$$

При этом максимизация ведется по всем t_l, t_r и $s \in [t_l, t_r]$, а r^* задает порог сходства.

Алгоритм поиска концептов, удовлетворяющих такому определению, выглядит следующим образом. На первом этапе каждая реализация a_i проверяется на роль столпа концепта, и параллельно определяются границы этого концепта.

1. Правая граница концепта t_r устанавливается в точке i, текущее качество $Q = 0$, рекорд качества $Q_{\max} = 0$.

2. Вычисляется конкурентное сходство F^* реализации a_{i+1} с a_i. Если эта величина больше 0, то она добавляется к счетчику $Q := Q + F(a_{i+1}, a_i|r^*)$, $Q_{\max} = Q$ и момент $(i + 1)$ считается текущей правой границей t_r концепта. Если F^* меньше 0, то процесс начинается со следующего момента времени.

3. Для всех последующих реализаций выполняется следующая процедура. Для k-ой реализации вычисляется текущее качество концепта $Q := Q + F(a_k, a_i|r^*)$ и, если оно оказывается больше рекорда качества, то рекорд обновляется. Если добавление этой реализации в концепт ухудшает текущее качество концепта, не более чем в α раз $(0 < \alpha < 1)$ относительно рекорда качества $Q_{\max}$, то процедура сдвига правой границы продолжается.

4. Если оказалось, что текущее качество $Q < \alpha Q_{\max}$, процедура расширения концепта вправо останавливается и правая граница возвращается в ту позицию, на которой был достигнут рекорд качества Q_r.

5. Аналогично в процессе движения влево от a_i, определяется качество левой части концепта Q_l и положение границы t_l.

6. Итоговое качество концепта, образованного вокруг реализации a_i, $D_i = Q_r + Q_l$.

После того, как все реализации проверяются на роль столпа концепта, в качестве столпов выделенных концептов берутся реализации концепта с максимальным качеством D. Из рассмотрения исключаются концепты, качество D или длина v которых ниже установленных порогов. Если два или более концептов пересекаются, то предпочтение отдается тому концепту, качество D которого больше.

Количество обнаруживаемых концептов зависит от величины r^*. Если минимальная и максимальная амплитуды сигнала равны $a_{\min}$ и $a_{\max}$, соответственно, то при $r^* \geq (a_{\max} - a_{\min})$ все отсчеты сигнала войдут в один концепт. При малых значениях r^* будет создаваться большое число концептов, в том числе концептов, незначительно отличающихся друг от друга средними значениями a и дисперсиями δ своих амплитуд. На этапе обучения следует выполнить серию экспериментов по сегментации сигнала и распознаванию концептов на данных обучающей выборки при разных значениях r^*. Критерием для выбора оптимального значения r^* служит высокое качество решения этих задач при минимальном числе концептов. В результате каждый стационарный концепт можно описать набором двух параметров a и δ.

При поиске динамических объектов с линейным ростом или падением амплитуд программа сначала определяет значения первых разностей d между соседними амплитудами: $d_i = (a_i - a_{i+1})/a_i$. При $d_i = 0$ будут выделяться стационарные концепты. Задавая разные значения d_i, можно формировать динамические концепты с разной крутизной подъема или снижения. После этого на последовательности значений разностей алгоритм работает так же, как и на амплитудах в стационарном случае.

При работе с контрольной выборкой принадлежность найденного концепта к тому или иному типу проверяется по функции конкурентного сходства его эталона с эталоном каждого типа в конкуренции с эталонами всех остальных типов.

Составление последовательности концептов и их иерархия. Фиксируется местоположение каждого обнаруженного концепта в сигнале. Те короткие участки сигнала, которые не вошли в состав концептов заданных типов, считаются принадлежащими неизвестному новому концепту. В зависимости от целей анализа, их можно игнорировать, считая выбросами, или использовать в качестве повода к пересмотру состава концептов.

В результате будет получена последовательность концептов, своего рода текст, описывающий сценарий динамического процесса на языке словаря концептов первого (лексического) уровня. Этот текст может использоваться для формирования концептов более высокого (синтаксического) уровня. Так, устойчивая последовательность динамических концептов «подъем»–«падение», повторенная три раза подряд, образует фигуру «голова и плечи», используемую при техническом анализе котировок на бирже. Для формирования концептов этого уровня нужно описать синтаксические классы в виде заданных последовательностей лексических концептов.

В свою очередь устойчивая последовательность синтаксических концептов позволяет формировать классы семантических концептов, которыми описывается более глубокий смысл наблюдаемых событий. Наконец, могут быть сформированы и классы прагматических концептов. Каждый такой класс связан с определенной реакцией на обнаруженную ситуацию.

Цепочка от лексических концептов к прагматическим может быть и более короткой. Например, обнаружение отдельного лексического концепта «скачок температуры» может служить сигналом к выполнению действий, связанных с тушением пожара.

Распознавание концептов в потоке данных. Для выделения и распознавания концептов в контрольном сигнале используется многоагентская процедура, аналогичная описанной выше. Каждый агент отслеживает появление концепта своего типа, опираясь на его эталон. Параллельно работают агенты, ответственные за обнаружение всех концептов всех предусмотренных иерархических уровней.

В результате на выходе системы будет появляться текст, описывающий последовательность состояний контролируемого объекта на языках всех уровней понимания. Лицо, принимающее решение, получив сигнал, требующий действий, имеет возможность увидеть концепты всех уровней и понять причины возникшей ситуации. Это обеспечивает прозрачность и дружественность системы контроля состояний динамических процессов.

3. ФОРМИРОВАНИЕ ДВУМЕРНЫХ И ТРЕХМЕРНЫХ КОНЦЕПТОВ

Примером потока данных, для описания которого применяются двумерные концепты, может служить протокол с записью N симптомов пациента в разные моменты времени; показания N геофизических приборов, измеряющих свойства грунта на разных глубинах; кросс-курсы N валют за T дней; изображение речевого сигнала на плоскости «частота–время» (видимая речь) и т. д.

В отличие от предыдущего случая, агенты при поиске концептов используют подмножества не отдельные отсчеты, а отдельные N-мерные вектора. Двумерный концепт представляет собой отрезок времени длиной не менее v сечений, на котором наблюдаются некоторые закономерные отношения между соседними векторами значений признаков. Эти закономерности могут описывать стационарные и переходные состояния наблюдаемого объекта. По сравнению с поиском двоичных бикластеров, в рассматриваемом нами случае задача усложняется процедурами измерения расстояний между признаками, измеренными в разных шкалах, и необходимостью искать более сложные концепты. С другой стороны имеется особенность динамических данных: здесь нельзя менять последовательность векторов во времени, что существенно понижает комбинаторную сложность задачи.

Процесс поиска концептов каждого из Q типов по каждому из N признаков ведется независимыми блоками программы (агентами) и состоит из тех же процедур, которые описаны для одномерных концептов. Но после того, как получены концепты первого лексического уровня, для формирования концептов следующего уровня начинают применяться другие решающие правила. Их закономерности имеют форму конъюнкций, включающей аргументы разных признаков: «Если по первому признаку имеет место лексический концепт q_{1i}, по второму – концепт q_{2j} и т. д., то на этом участке имеет место синтаксический концепт q_s».

Каждый агент должен отличать концепты своего типа от всех других концептов. Для этого бывает достаточно отслеживать не все N признаков, а лишь некоторое их подмножество, свое для каждого агента. Такие Q подсистем информативных признаков можно выбрать с помощью алгоритма FRiS–GRAD [4], решив Q задач распознавания типа «один против всех». В результате описание эталонов становится более кратким, что повышает надежность распознавания концептов и позволяет сократить машинные ресурсы, используемые агентами.

Местоположение каждого двумерного концепта в сигнале определено. В результате будет получена последовательность концептов, своего рода текст, описывающий сценарий динамического процесса на языке концептов второго (синтаксического) уровня. Устойчивые сочетания этих концептов можно считать классами следующего (семантического) уровня, которые в свою очередь можно объединить в концепты прагматического уровня, связанные с определенными реакциями на обнаруженную ситуацию.

Теперь рассмотрим систему, наблюдающую за динамикой состояний не одного, а нескольких многомерных объектов. При этом создается поток данных в виде куба «объект–признак–время». Концептом первого уровня будем называть отрезок времени из не менее, чем v сечений, с фиксированными для каждого из M объектов своими отношениями между значениями N признаков в соседних сечениях.

Для каждого из M объектов решается описанная выше задача обнаружения двумерных лексических и синтаксических концептов. Для поиска семантических концептов требуется сопоставление текущих синтаксических концептов всех объектов. По диагнозам заболевания отдельных пациентов без их сравнения между собой нельзя обнаружить эпидемию. По росту котировок отдельных акций без сравнения их друг с другом нельзя зафиксировать начало общего экономического подъема.

Семантические классы представляются набором правил такого типа: «Если в момент t состояние i-го объекта соответствует концепту C_1, состояние k-го объекта – концепту C_5, а состояние l-го объекта – концепту C_4, то состояние системы в целом соответствует семантическому классу S_1». На следующем уровне формируются концепты прагматического уровня, определяющие реакцию системы: «Если общее состояние S_1 следует после состояния S_7, то требуется выполнить такие-то действия». Например, помощь пациенту может требоваться, если дистанционные средства наблюдения обнаружили падение пациента и его лежание в неподходящем месте, его сидение неподвижно в неподходящем месте слишком долгое время и т. д.

4. РЕШЕНИЕ ПРИКЛАДНОЙ ЗАДАЧИ

В последние годы наблюдается увеличение интереса к созданию систем для дистанционного ухода за престарелыми пациентами (AAL) [5], включая умные дома, биомониторинг, и т. д. Проектируются системы, которые контролируют пользователя автономно с целью обнаружения критического положения, например, падения. Главная решаемая задача состоит в обнаружении аномальной ситуации в ежедневном поведении пациента.

Контроль за пациентом ведется по сигналам четырех датчиков, укрепленных на теле пациента: на груди, поясе и двух ногах. Каждый датчик фиксирует свое положение в трех ортогональных координатах X, Y и Z. Датчики опрашиваются поочередно с разной частотой – от 8 до 12 отсчетов в секунду. Их сигналы передаются по одному каналу связи в виде последовательности трехмерных векторов, порождаемых разными датчиками.

Этот поток данных можно представить в различных форматах. Можно считать, что наблюдаются четыре объекта (датчика) с тремя характеристиками или один объект (пациент) с двенадцатью характеристиками. По сигналам датчиков требуется различать одиннадцать состояний пациента: «ходит», «падает», «лежит», «ложится», «сидит», «садится», «садится на землю», «встает после сидения», «встает после сидения на земле», «встает после лежания», «стоит на четвереньках».

В экспериментах участвовало пять человек. Каждый из них по пять раз принимал последовательность состояний по заданному сценарию. Данные представлены двадцатью пятью файлами, в каждом из которых отражены сигналы, измеренные за время выполнения сценария (от трех до пяти минут на эксперимент) [6].

На этих данных ставились следующие задачи:

1. Сегментация потока данных на временные участки, соответствующие заданным концептам.
2. Выбор информативных признаков, построение решающих правил для решения задач распознавания концептов в потоке данных.

1. Метод сегментации сигнала на участки, соответствующие концептам, и распознавание типа концепта описаны выше. Экспериментальная проверка метода обнаружения стационарных концептов делалась на материале эксперимента с пациентом $A1$. В файле данных *A01* указаны границы между состояниями пациента. Величина параметра r^* определялась методом последовательных приближений. Было выбрано значение $r^* = 0.2$, при котором все стационарные состояния пациента («сидит», «лежит», «стоит на четвереньках») были покрыты концептами. Пример работы алгоритма приводится на рис. 1. Тонкой сплошной линией выделен исходный сигнал, а жирной – выделенные концепты. Между границами соседних концептов остались участки с динамическими концептами, которые должны обнаруживаться, если задать для них соответствующие значения параметра $d > 0$ или $d < 0$.

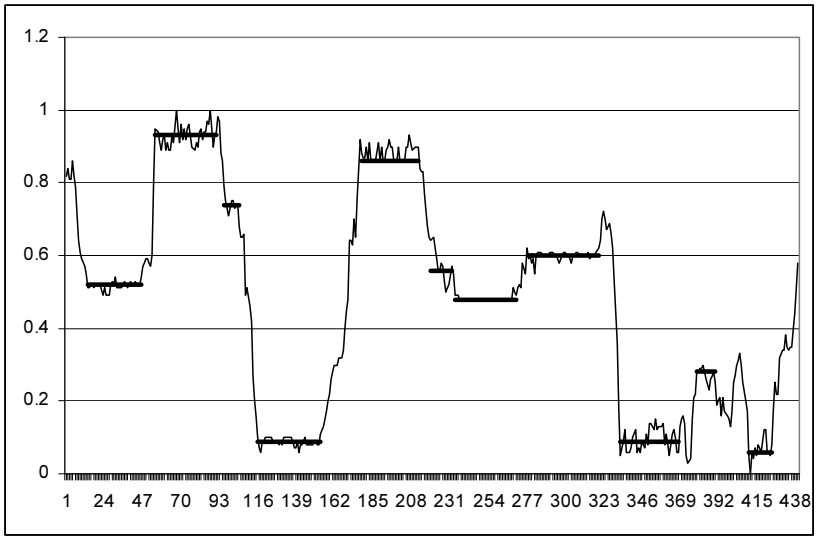

Рис. 1. Выделение стационарных концептов на одномерном сигнале

2. Вторая задача решалась с помощью алгоритма FRiS–GRAD, предназначенного для построения решающего правила с одновременным выбором информативных признаков. Была реализована идея многоагентской системы: каждый агент отвечает за обнаружение своего концепта. С учетом этого, для каждого концепта выбиралось подмножество информативных признаков, в пространстве которых строилось решающее правило, отделяющее заданное состояние от всех остальных. Рассматривались концепты следующих четырех стационарных концептов: «гуляет», «сидит», «лежит», «стоит на всех четырех».

В качестве обучающего материала использовался первый эксперимент пациента A1 (файл *A01*).

В табл. 1 результаты распознавания четырех стационарных состояний в контрольных файлах *A02* и *A03*, по решающим правилам в информативных подпространствах, выбранных на файле *A01*. Во втором столбце приводится число столпов (*S*) и количество признаков (*n*), выбранное алгоритмом FRiS–GRAD, в третьем – имена контрольных файлов, в четвертом – количество реализаций распознаваемого состояния (M_1) к количеству реализаций всех прочих состояний (M_2) в контрольной выборке, следующие три столбца содержат такие характеристики качества решающего правила, как чувствительность, специфичность и ожидаемая ошибка. В последней строчке результаты усредняются.

Таблица 1

Action	*S*/*F*	File	M_1/M_2	Sensivity	Specifity	Total Error
walking	21*S*/5*F*	A02	50/301	84%	95.35%	6.27%
		A03	33/265	78.79%	88.30%	12.75%
sitting	10*S*/6*F*	A02	67/284	100%	96.48%	2.85%
		A03	63/235	50.79%	97.02%	2.85%
lying	27*S*/8*F*	A02	137/214	89.05%	98.60%	5.13%
		A03	122/176	94.26%	97.16%	4.03%
on all fours	6*F*/6*S*	A02	15/336	100%	95.24%	4.56%
		A03	12/286	83.33%	99.65%	1.01%
Average				85%	96%	5%

ЗАКЛЮЧЕНИЕ

Разработан подход к анализу динамических кубов данных с использованием иерархии концептов. Для формирования концептов и их распознавания в потоке данных разработаны методы, основанные на использовании функции конкурентного сходства (FRiS-функции). Решена прикладная задача из области социальной медицины. Результаты эксперимента показывают, что использование FRiS-функции позволяет членить поток данных на участки, соответствующие концептам, и надежно распознавать типы стационарных концептов.

БЛАГОДАРНОСТИ

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проекты 11–01–00156, 12–01–31392).

СПИСОК ЛИТЕРАТУРЫ

[1] **Sara C. Madeira.** Biclustering Algorithms for Biological Data Analysis: A Survey / Sara C. Madeira, Arlindo L. Oliveira // IEEE/ACM Transactions on Computational Biology and Bioinformatics. – 2004. – Vol. 1. – № 1. – P. 24–45.

[2] **Борисова И.А.** Критерии информативности и пригодности подмножества признаков, основанные на функции сходства / И.А. Борисова, Н.Г. Загоруйко, О.А. Кутненко // Заводская лаборатория. Диагностика материалов. – Москва, 2008. – № 1. – Т. 74. – С. 68–71.

[3] **Борисова И.А.** Алгоритм таксономии FRiS-Tax / И.А. Борисова // Научный вестник НГТУ. – 2007. – № 3(28). – С. 3–12.

[4] **Zagoruiko N.G.** Methods of recognition based on the function of rival similarity / N.G. Zagoruiko at all // Pattern Recognition and Image Analysis. – 2008. – Vol. 18. – № 1. – P. 1–6.

[5] **Boštjan Kaluža.** Analysis of Daily-Living Dynamics / Boštjan Kaluža, Matjaž Gams // Journal of Ambient Intelligence and Smart Environments. – IOS Press – 2012. – Vol. 1. – P.1–51.

[6] Localization Data for Person Activity. – Режим доступа: http://archive.ics.uci.edu/ml/datasets/Localization+Data+for+Person+Activity

Загоруйко Николай Григорьевич, доктор технических наук, профессор, академик Российской академии естественных наук (РАЕН), заведующий лабораторией анализа данных Института математики СО РАН. Основное направление научных исследований: искусственный интеллект, анализ данных, распознавание образов. Имеет более 226 публикаций, в том числе 13 монографий.

Борисова Ирина Артемовна, кандидат технических наук, старший научный сотрудник лаборатории анализа данных ИМ СОРАН. Основное направление научных исследований: анализ данных, распознавание образов. Имеет более 40 публикаций.

Леванов Дмитрий Александрович, аспирант ИМ СО РАН. Основное направление научных исследований: анализ данных, системное программирование. Имеет 3 публикации.

N.G. Zagoruiko, I.A. Borisova, D.A. Levanov

Description of multivariate dynamic processes with concepts hierarchy

Analyzing the structure of Data Mining tasks, described by data set “object-feature-time” (data cube), gives an idea to describe such data with concepts hierarchy. After the concepts vocabulary forming we can processing a dynamic process into text scenario of the process. For concepts extraction, clustering and recognition the methods, based on function of rival similarity, were proposed. These methods were tested on the task of patient’s activity recognition, where eleven states of the person should be detected on the four sensors indications. Results of the experiments proved what the methods determined bounds of stationary concepts and recognized they types.

Key words: dynamic trajectory, stationary concepts, states recognition, FRiS-function.