УДК 5019.7

МЕТОДЫ АНАЛИЗА И СИНТЕЗА ИНФОРМАЦИОННОЙ СИСТЕМЫ РАСПОЗНАВАНИЯ ОБРАЗОВ

© Амиргалиев Е.Н., Амиргалиева С.Н.

Казахстан, Алматы, КазНТУ им. К.И.Сатпаева, КБТУ

e-mail: *amir_ed@mail.ru*

Abstract. In the given work development of information system conceptual model of recognition and classification on the basis of mathematical methods and models of image recognition and classification, intended for the analysis and processing of multidimensional data is considered. The methodology of working out of information system is realised with use of the object-oriented approach and carried out in language of modelling of complex program systems – the unified language of modelling UML. The considered concept of construction of information system can be used in different subject domains, as the modern concept of working out of similar information systems. For realisation of design decisions it is used CASE-means Rational Rose.

Введение

В быстроразвивающейся сфере разработки объектно-ориентированных приложений становится все труднее и труднее создавать и поддерживать приложения, обладающим высоким качеством, укладываясь при этом в разумные временные рамки. Унифицированный язык моделирования (Unified Modeling Language, UML) появился как ответ на потребность в универсальном языке объектного моделирования, который могла бы использовать любая компания. UML – это технология моделирования сложных программных систем, принятая в современная технология в индустрии информационных технологий. Это метод детального описания архитектуры системы. С помощью нотации UML легче создавать и сопровождать систему, вносить в неё требуемые изменения и совершенствовать ее далее.

Язык UML, пришедший на смену многочисленным системам нотации и методикам проектирования, предложил нотацию для описания объектно-ориентированных моделей, которая стала промышленным стандартом. Однако для эффективного применения нотации UML необходимо сочетать ее с каким-либо методом объектно-ориентированного анализа и проектирования.

В описываемом методе сочетаются прецеденты использования, статическое моделирование, и диаграммы последовательности событий, которые встречаются в нескольких методах. Применяемая нотация основана на UML. В ходе моделирование прецедентов определяется функциональное требование к системе в терминах актеров и прецедентов. Статическая модель предлагает статический взгляд на информационные аспекты системы. Класс определяется в терминах своих атрибутов и взаимоотношений с другими классами. Результатом динамического моделирования является динамический взгляд на систему. Уточняются сформулированные ранее прецеденты с целью показать взаимодействие объектов, участвующих в каждом из них. Разрабатываются диаграммы кооперации и последовательности, отражающие кооперацию

объектов в каждом прецеденте. Зависящие от состояния аспекты системы описывается с помощью диаграмм состояний, причем для каждого объекта составляется своя диаграмма.

В качестве математического обеспечения разрабатываемой системы рассмотрены математический аппарат решения задачи распознавания и классификации, задачи групповых классификации и оптимизационные модели, использующие различные виды функциоалов качества.

1. Постановка задачи классификации Z_k

Пусть задана начальная информация I, $\{S\}$ – множество допустимых объектов и каждый объект $S_i \in \{S\}$, $i=1,..., m$ характеризуется n-мерным вектором, координаты которого называются признаками, взятыми из алфавита признаков, т.е.

$Si=(\alpha i1,\alpha i2,....,\alpha in)$, $i=1,..., m$, n – число признаков, $\alpha ij \in Mj$.

Требуется построить алгоритм классификации A для множеств $\{I(S)$, $S\}$, классифицирующий исходное множество объектов S на ряд непересекающихся классов (кластеров), Kj, $j=1,..., l$, так чтобы объекты, принадлежащие к одному классу (кластеру) были сходными (близкими), в то время как объекты, принадлежащие различным классам, были в некотором определенном смысле непохожими (удаленными):

$$A\{I(S), S\} = \bigcup_{j=1}^{l} K_j, K_i \bigcap K_j = \emptyset,$$

если $i \neq j$, $Ki \neq \emptyset$, i, $j=1$, $2,..., l$.

Построение кластеров можно рассматривать как задачу распознавания образов без учителя, учитывая, что на заданном множестве объектов, как правило, отсутствует всякая информация, касающаяся числа классов и структуры классов.

При построении алгоритмов по принципу минимума расстояния, отыскание кластеров и определение эталонов являются вопросами первостепенной важности. При построении таких алгоритмов, как правило, используются два подхода. Один из них – эвристический, и в его основе лежит интуиция и опыт. Второй подход предусматривает минимизацию или максимизацию некоторого выбранного показателя качества классификации.

В моделях групповых классификации введена метрика, используемая в пространстве классификаций. Рассмотрим задачу синтеза групповой классификации Z_C. Пусть $A_1, ...,\ A_m \in \{A\}$ – исходный набор алгоритмов решения задачи классификации Z_k для множества объектов $M = \{S_1, ..., S_n\}$.

Результатом применения алгоритмов A_i к множеству $(M,\ J(M))$ являются классификации $K_i(M) \in \Re(M)$, $\Re(M)$- пространство классификаций конечного множества объектов M, элементами которого являются отдельные классификации. Пусть определена метрика $d(K', K")$ в $\Re(M)$ и $\varphi(K) = \sum_{i=1}^{m} d(K, K_i)$, $K_i = K_i(M)$, $K \in \Re(M)$.

Тогда основная задача группового синтеза (групповых классификации) Z_C состоит в следующем. Найти классификацию $K^*(M) \in \Re(M)$, минимизирующую функционал $\varphi(K)$, т.е.

$$\varphi(K^*) = \min\ \varphi(K),\ \ K \in\ \Re(M).$$

Рассмотрим пространство классификаций $\Re(M)$ множества M. Известно, что по любой классификации $K \in\ \Re(M)$ можно построить соответствующее ей бинарное отношение R, которое является отношением эквивалентности на множестве M. Также, любому отношению эквивалентности $R(M)$ на множестве M однозначно соответствует классификация $\Re(M)$ для M.

Будем рассматривать стандартное представление классификаций $K(M)$ 87 $\Re(M)$ в виде конечного множества классов $K_i(M)$ – подмножеств множества M, т.е. $K(M) = \{K_1(M), K_2(M), ..., K_l(M)\}$ и значит для описания классификации достаточно перечисления номеров объектов, попавших в каждый из классов. Причем, как будет показано ниже, способ нумерации классов может быть произвольным для полученной любой классификации, и не влияет на результат их сравнения. При таком представлении классификаций для решения задачи Z_C нужно иметь метрику в $\Re(M)$, которая бы достаточно полно отражала реально существующие в данном пространстве расстояния, и исследовать свойства пространства $\Re(M)$, являющегося структурой.

Обозначим через $K^l(M)$ множество классификаций M на l, $1 \le l \le n$ классов, т.е. $\bigcup_{l=1}^{n}\ K^l(M) = \Re(M)$. Пусть $K_n(l)$- произвольная классификация из $K^l(M)$. Зададим отображение $d : K(M) \times K(M) \to Z$ с помощью следующей формулы:

$$d(K_n^u(l_u), K_n^\nu(l_v)) = 2n - \sum_{j=1}^{l_u} \max_{l\le i\le l_v} \left\{\ \left| K_{n,j}^u(l_u) \bigcap K_{n,i}^v(l_v)\ \right|\ \right\} -$$
$$- \sum_{i=1}^{l_\nu} \max_{l\le j\le l_u} \left\{\ \left|\ K_{n,i}^v(l_v) \bigcap K_{n,j}^u(l_u)\ \right|\ \right\}\ ,$$

где $K_n^t(l_t) = \left\{\ K_{n,1}^t(l_t), ..., K_{n,l_t}^t(l_t)\ \right\},\ \ 1 \le l_t \le n,\ \ t \in \{u, v\}$.

Введенная метрика обладать свойствами метрики[3].

Концептуальная модель информационной системы распознавания и классификации. Концептуальная схема разработанной информационной системы распознавания и классификации показана на рисунке 1.

Кратко опишем функциональные назначения подсистем, входящих в состав системы: Подсистема ***Справка-Help*** представлена как справочная система. Показывает справку о функциональных возможностях, как отдельных подсистем, так и системы в целом; ***Управление*** – подсистема представляет программный интерфейс процесса управления проектируемой системы; ***База данных*** –Совокупность ряда таблиц для хранения данных необходимых для системы; Подсистема ***предварительной обработки*** предназначена для предварительной обработки исходных данных: определения незаполненных данных и определение оптимальных подсистем признаков в описании объектов; определение информативных признаков;

Модели и алгоритмы классификации осуществляет работу алгоритмов классификации, составляющих базовый набор для системы; **Модели группового синтеза** представляет работу алгоритмов группового синтеза, реализованных в рамках системы; **Визуализация результатов** представляет результаты работы,

Рис. 1. Концептуальная модель системы распознавания и классификации

как подсистем, так и системы в целом для пользователей в нужном формате; **Анализ и оценка результатов** предназначена для анализа и оценки результатов на основе показателей выбранных видов функционалов качества и представляет основу для выработки рекомендации об использовании различных вычислительных схем; Подсистема ***ГИС моделирование*** представляет возможности использования разработанных методов, алгоритмов группового синтеза в рамках геоинформационного моделирования.

На рисунке 2 приведена функциональная модель информационной системы распознавания и классификации с применением методологии функционального анализа и проектирования системы SADT – (Structured Analysis & Design Technique). На диаграмме указана схема передачи управляющих параметров входных и выходных данных для каждой подсистемы.

В рамках информационной системы, в зависимости от постановок задач пользователя, можно осуществить оптимизационные процедуры, используя конкретный вид функционалов качества. Схема проведения оптимизационных процедур осуществляет подсистема *Анализ и оценка* (Рисунок 3).

В рамках данной схемы возможны следующие модели оптимизации (применяются различные комбинации алгоритмов, моделей группового синтеза, функционалов качества): одноуровневая модель $M1\ \{(M\ k,\ Aj,\ F\ t):\ k=1,\ \ldots,\ N;\ j=1,\ldots,\ M;\ t=1,\ldots,\ N.\}$; двухуровневая модель $M2\ \{(M\ k,\ Aj,\ Z\ ci\ ,\ F\ t):\ k=1,\ \ldots,\ N;\ j=1,\ldots,\ M;\ i=1,\ldots,\ T;\ t=1,\ldots,\ K.\}$.

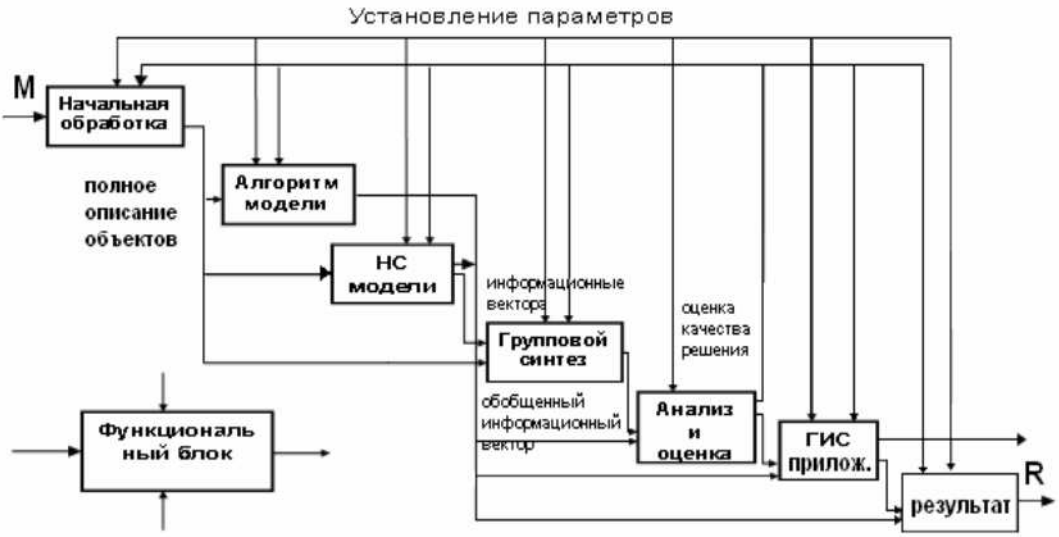

Рис. 2. Функциональная модель системы распознавания и классификации

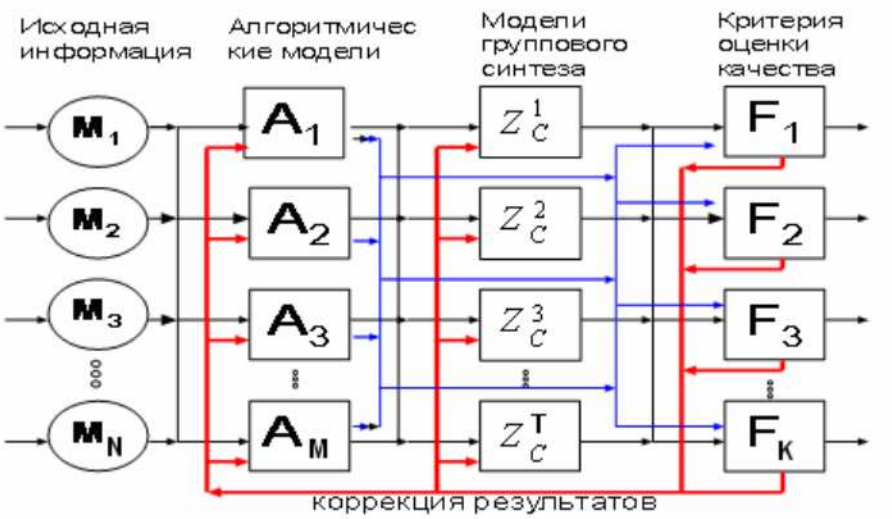

Рис. 3. Схема оптимизационных процедур (двухуровневая модель). Здесь: MN – множество исходных данных, AM – алгоритмы классификации, Z_C^T- модели группового синтеза, FK – функционалы качества.

В процессе моделирования системы использованы следующие виды диаграмм (нотации) UML: *диаграммы прецедентов* (*use case diagram*) представляет функциональность и поведение разрабатываемой системы, позволяет определить требования к системе, определить действующие в системе объекты и основные задачи, выполняемые этими объектами (отображение сценариев); *диаграммы классов (class diagram)* обеспечивают статическое проектное представление системы; *диаграммы кооперации*

(collaboration diagram) описывают взаимодействие объектов, абстрагируясь от последовательности передачи данных, отражаются все принимаемые и передаваемые сообщения конкретного объекта и типы этих сообщений; *диаграммы последовательности (sequence diagram)* позволяют определить последовательность передачи сообщений между объектами, показывают поток сообщений; *диаграммы состояний (state diagram)* предназначены для отображения состояний объектов системы, имеющих сложную модель поведения.

Для управления функционированием системы создается, так называемый *сеанс обработки,* диаграмма прецедентов которой показана на рис. 4.

Сеанс обработки – выбор вычислительного процесса и формирование совокупности входных и выходных данных для определенной подсистемы и сам процесс обработки данных. *Входные данные* – исходные данные, список алгоритмов и параметров. *Исходные данные* – подмножество входных данных подсистемы, которое является общим для всех алгоритмов этой подсистемы. Структура исходных данных регламентируется подсистемой. *Список алгоритмов* – подмножество алгоритмов подсистемы, которые выполняют обработку в данном сеансе. *Параметры* – подмножество входных данных подсистемы, которое не является общим для всех ее алгоритмов. Параметры группируются по алгоритмам. Каждый алгоритм объявляет свой набор параметров и их структуры. *Выходные данные* – совокупность кодов возврата и результатов. Каждый выполнявший обработку алгоритм ассоциируется с кодом возврата и результатом. *Код возврата* – числовое значение, обозначающее статус окончания алгоритма.

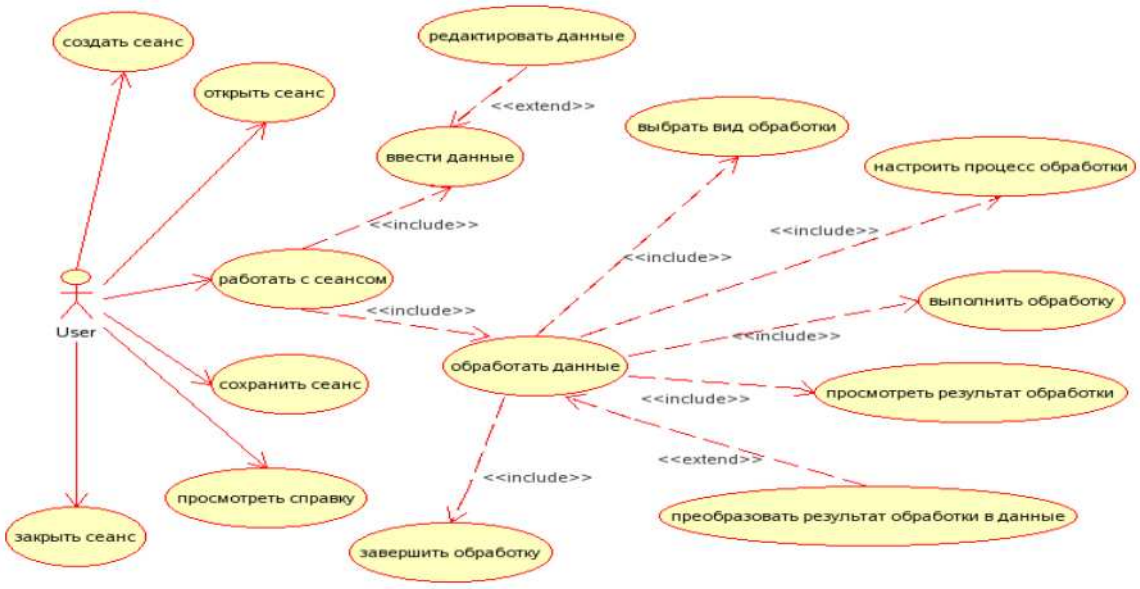

Рис. 4. Диаграмма прецедентов для сеанса обработки

Представим описания некоторых прецедентов, приведенных на рисунке 4.

Создать сеанс – *с*оздание нового сеанса. *Предусловие* – система незапущена или сеанс находится в состояний *новый. Основной поток событий: п*ользователь, запуская программу, или, используя элементы управления пользовательского интерфейса запущенной программы, создает *новый сеанс .пользователя. Постусловие* – система находится в состоянии *Новый сеанс.*

Открыть сеанс – открытие предварительно созданного сеанса**.** *Предусловие-с*еанс в состояний *Новый* или *Рабочий. Основные потоки событий:* 1)пользователь выбирает файл сеанса; 2) файл открывается для чтения; 3)сеанс инициализируется данными из файла; 4) файл закрывается. *Альтернативный поток событий-р*азличные ошибки ввода-вывода. Выводится соответствующее сообщение. *Постусловие* – *с*еанс переходит в состояние *Рабочий.*

Работать с сеансом – работа с сеансом. *Предусловие- с*еанс в состоянии *Новый* или *Рабочий. Основные потоки событий:*1) вводит данные; 2) обрабатывает данные; 3) просматривает результаты обработки. *Альтернативный поток событий – р*азличные ошибки при работе соответствующих подсистем. Выводится соответствующее сообщение. *Постусловие* – сеанс переходит в состояние *Работающий*, а затем в *Рабочий.*

Также по указанной схеме описываются и другие прецеденты сеанса обработки, указанные на диаграмме прецедентов (Рисунок 4).

Заключение

Таким образом, в данной работе мы показали некоторые аспекты моделирования информационной системы распознавания и классификации с применением унифицированного языка моделирования. Для реализации проектных решений использованы инструментарий CASE- средства Rational Rose.

Разработанная информационная система применена для решения реальных прикладных задач из области гидрогеологии, экологического мониторинга с использованием данных дистанционного зондирования и геологии.

Список литературы

1. *Г. Буч* Объектно-ориентированный анализ и проектирование. – М.: «Издательство Бином», 1999.
2. *Уэнди Боггс* Rational Rose & UML. – М.: «Издательство Лори», 1999, – 480с.
3. *Айдарханов М.Б., Амиргалиев Е.Н.* Алгоритмические основы построения систем классификации. – Алматы, 1998. – 100с.
4. *Вендров А.М.* Проектирование программного обеспечения ЭИС. – М.: 2000, - 290 с

Статья поступила в редакцию 27.04.2008