Применение методов выявления закономерностей для классификации химических соединений

Г.Т. Маракаева

Аннотация. Целью статьи является постановка задачи классификации неизвестных химических соединений. С помощью классификации выявляются признаки, характеризующие группу, к которой принадлежит тот или иной объект. Это делается посредством анализа уже классифицированных объектов и формулирования некоторого набора правил. Целью различных алгоритмов классификации является построение классификационной модели, называемой *классификатором*, которая будет предсказывать класс для заданного примера на основании имеющихся значений атрибутов. Как правило, классификация рассматривается, как задача Data Mining, что по-русски язык означает “обнаружение знаний в базах данных”, “выявление закономерностей”.

В рамках исследования задачи совместно с экспертами предметной области были проанализированы возможности формализации экспертных знаний в системе. Кроме того, были изучены формат данных и значения атрибутов химических соединений в нескольких лабораториях. Результатами этих работ стали выводы о возможности применения различных алгоритмов классификации для определения классов химических соединений. Эти выводы являются предметом следующей статьи.

1. Введение

Целью статьи является постановка задачи классификации химических соединений. Работа лаборатории заключается в проведении экспериментов над различными пробами по определению их свойств (например, содержание серы, вязкость и т.д.), а также по определению класса пробы (например, питьевая вода, природная вода и т.д.). Каждый полученный результат фиксируется в лабораторном журнале. В данном случае ключом записи является идентификационный номер пробы, значения ее параметров и класс пробы. После некоторого времени в лаборатории накапливается достаточно большое число записей о пробах, содержащих значения параметров и класс. Каждая проба может исследоваться, во-первых, на значение атрибутов, а во-вторых, на принадлежность к одному из классов химических проб.

Задача классификации состоит в определении класса новой пробы по неполному набору значений атрибутов. В лабораторию поступает новая проба

неизвестного класса. Задача сотрудника – определить класс пробы. В общем случае сотрудник должен по внешним факторам экспертно определить возможный класс и выполнить полный набор экспериментов, призванных подтвердить или опровергнуть его гипотезу. Если гипотеза не подтверждается, то проводится следующая серия экспериментов для исследования другой гипотезы.

Среди экспертов химической области наблюдается больший интерес к использованию информационных систем в своей работе. С привлечением все большего числа специалистов будут накапливаться базы информации об исследуемых соединениях, пробах и т.д. и, следовательно, возникнет необходимость анализа этих данных. Так как описание проб с помощью признаков является унифицированным для всех экспертов, то накапливаемая база может использоваться повсеместно.

Классификация очень часто рассматривается, как один из способов выявления закономерностей в большом объеме данных (Data Mining) [1, с. 19]. Под понятием «выявления закономерностей» понимается итеративный автоматический или ручной процесс изучения данных. [2, с.2].

Выявление закономерностей делятся на следующие категории [2, с.2;1, с.19].

- *Классификация (classification)* – с помощью классификации выявляются признаки, характеризующие группу, к которой принадлежит тот или иной объект. Это делается посредством анализа уже классифицированных объектов и формулирования некоторого набора правил [1, с.20]. Целью различных алгоритмов классификации является построение классификационной модели [4, с.29] называемой *классификатором*, которая будет предсказывать класс для заданного образца на основании имеющихся значений атрибутов [2, с.139]. Другими словами, классификация – это процесс определения ярлыка с дискретным значением (класс) для непомеченной записи, а классификатор – это модель (результат классификации), которая предсказывает один атрибут – класс образца – если заданы некоторые другие атрибуты этого образца. [2, с. 139].
- *Кластеризация (clustering)* – очень близка к определению классификации, за исключением того, что изначально классы предметной области не заданы
- *Ассоциация (dependency model)* – выявление правил взаимосвязей между атрибутами или между значениями атрибутов во всем множестве данных или в его подмножестве.
- *Последовательность* – выявление цепочек событий, связанных во времени.
- *Прогнозирование (predictability)* – изучение данных с целью прогнозирования реальных значений.

- *Определение изменений и отклонений (change and deviation detection)* – исследование наиболее значимых отклонений во множестве данных.

Основные термины, участвующие в постановке задачи приводятся в разделе 2. В этом разделе приводятся как формальные термины для задачи классификации, так и специфические понятия для заданной предметной области, а также их соответствие между собой. В разделе 3 приводится общая постановка задачи классификации, использующей вышеописанные термины. Раздел 4 содержит описание девяти алгоритмов классификации, наиболее часто встречающихся в литературе. В разделе 5 приведен подход к классификации пакетов и описаны наиболее распространенные системы. Раздел 6 содержит краткое описание выбранного алгоритма для классификации химических проб. В "Заключении" подведен итог работы и кратко описаны следующие статьи по данной тематике.

2. Терминология

Введем некоторые термины:

Признаком назовем пару $x = \langle имя, множество\_значений \rangle$. Множество значений признака обозначим через *T*. Множество значений может быть задано типом признака: целое или вещественное, перечисление или диапазон.

Словарь признаков представляет собой множество всех признаков $x_i = \langle n_i, T_i \rangle, i = 1,..., N$, где n_i – имя, T_i – множество значений для признака x_i.

Пространство признаков представляет собой декартово произведение множеств значений признаков T_i, $i=1,..,N$: $\tau = T_1 \times T_2 ... \times T_N$

Объект ω – представляет собой пару $\omega = \langle имя, (x_1,..., x_N) \rangle$, где $(x_1,..., x_N)$ – точка в признаковом пространстве. Множество всех объектов обозначим через *W*. Множество *W* является конечным или счетным, и все объекты могут быть занумерованы.

Пусть на множестве объектов *W* определено *m* классов $\Omega_1,..., \Omega_m$. Для каждого класса Ω существует характеристическая функция:

$$\chi(\omega) = \begin{cases} -1, \omega \in \Omega \\ 1, \omega \notin \Omega \end{cases} \qquad (1)$$

Таким образом, класс представляет собой пару $\Omega = \langle имя, \chi \rangle$, где χ – характеристическая функция, определенная формулой (1).

Если для всех *j=1,..,m* заданы характеристические функции χ_j, то для каждого объекта $\omega \in W$ можно определить его класс Ω_j. Таким образом, множество χ_j определяет классы объектов.

Обучающая выборка – это множество объектов $V = \{\omega_1,..., \omega_{l_1}, \omega_{l_1+1},..., \omega_{l_2},..., \omega_{l_{m-1}},..., \omega_{l_m}\}$, для которых известны включающие их классы: $\omega_1,..., \omega_{l_1} \in \Omega_1, \omega_{l_1+1},..., \omega_{l_2} \in \Omega_2,..., \omega_{l_{m-1}},..., \omega_{l_m} \in \Omega_m$

Помимо характеристических функций χ_j будем рассматривать функции $f_j(x,V)$, про которые известно, что $f_j(x,V)=-1$, если $x \in \Omega_j$, $f_j(x,V)=1$, если $x \notin \Omega_j$.

Такие функции будем называть *разделяющими функциями*. Если задана обучающая выборка, то по ней можно построить разделяющие функции.

Классификация – это сопоставление каждому объекту определенного класса, то есть определение множества пар $\langle объект, класс \rangle$

Тестовый объект – объект с некоторым заданным набором значений признаков и неизвестным классом.

Задача классификации. *Заданы множество имен классов и обучающая выборка. Требуется построить разделяющие функции для этих классов*.

Далее следуют пояснения к терминам рассматриваемой предметной области и их соответствие вышеописанным терминам:

Проба – некоторое количество (вообще говоря, неизвестного) химического соединения, достаточное для проведения экспериментов по выявлению количественных и качественных характеристик соединения, которые требуются для идентификации соответствующего химического соединения. В алгоритмах классификации проба – это объект предметной области. Идентификация химического соединения состоит в определении *класса*, которому принадлежит проба. Пробе соответствует объект.

Атрибут пробы – какое-либо химическое или физическое свойство соединения. У каждого атрибута пробы имеется тип, описывающий его возможные значения. Атрибут пробы является признаком объекта.

Обучающая выборка в терминах предметной области – это набор проб, для которых известны классы (вообще говоря, с некоторой степенью вероятности).

3. Постановка задачи классификации

Для тестовой пробы, или, другими словами, *тестового объекта* класс неизвестен. Поэтому для программиста задача ставится как нахождение класса для тестового объекта по некоторым заданным значениям атрибутов этого объекта. Для нахождения класса строится система, другими словами, модель, или классификатор. Обучение системы происходит с помощью обучающей выборки, то есть множества объектов, для которых значения их классов достоверно известны. Кроме того, в систему должны вводиться экспертные знания о классификации проб с помощью шаблонов, то есть система должна аккумулировать множество экспертных шаблонов. Будет считаться, что все данные, используемые на стадии обучении системы, являются истинными.

У задачи классификации для химической предметной области имеется ряд важных особенностей, которые существенно влияют на выбор решения. Во-первых, данные для построения и обучения системы носят не только экспериментальный, но и экспертный характер. В системе необходимо учитывать экспертные шаблоны. Во-вторых, типы многих атрибутов являются числовыми, поэтому многие шаблоны задаются с диапазонами значений атрибутов, а не перечислением. Шаблоны же, полученные на стадии обучения, в предикатных условиях которых содержится знак равенства, должны каким-либо образом быть преобразованы в шаблоны, задающие диапазонные предикаты. В-третьих, множество классов является бесконечным. Поэтому система никогда не будет окончательно сформировавшейся, она всегда будет обучаться новыми пробами.

4. Обзор алгоритмов решения задач классификации

Существует достаточно много различных алгоритмов классификации. Каждый из них может давать хорошие результаты для одного класса задач и плохие результаты для другого класса задач. Например, работа классификатора на основе нейронной сети будет давать хороший результат при достаточно большом объеме данных, на которых проводится обучение, и при наличии примерно одинакового числа тестовых объектов в каждом классе. Если же во входных данных находится очень много объектов одного класса, то результат работы сети будет часто склоняться именно к этому классу, независимо от значимости остальных объектов.

Для оценки методов классификации можно использовать следующие критерии:

- *точность и правильность прогнозирования* – способность корректно определять класс для новых данных;
- *скорость работы* – параметр вычислительных затрат на использование модели;
- *робастность* – способность делать корректный прогноз при “зашумленных” входных данных или данных с неполным набором атрибутов;
- *масштабируемость* – способность поддерживать эффективность метода при увеличении объема подаваемой на вход информации;
- *интерпретируемость* – характеристика метода, отвечающая за уровень понимания и способность проникновения в суть.

Ниже перечислены основные алгоритмы классификации, упоминаемые в литературе и применяемые в промышленных системах.

Каждый из этих алгоритмов обладает определенными свойствами, позволяющими делать предположения об их применимости на конкретном классе задач. Кроме того, у каждого из алгоритмов может иметься множество разновидностей, как, например, в п. 4). Как правило, в литературе приводятся общие принципы к построению классификатора, тогда как для реализации

требуется подробный анализ предметной области и правильный подбор детализированных алгоритмов.

4.1. Статистический метод

Статистические заключения – это наилучшая форма анализа данных, имеющих причинные связи. Теория статистических заключений состоит из методов, таких что на основе одного из них можно сделать заключение или обобщение о всем пространстве объектов. Такие методы могут быть разбиты на две основные группы: *вычисления* и *проверка гипотез*.

4.2. Дерево решений

Дерево решений – это определенный вид дерева, в котором каждый внутренний узел представляет собой контрольный критерий для атрибута, каждая ветвь представляет собой результат теста, а лист – это класс или подкласс. Самый верхний узел дерева – корень.

Для классификации нового образца используется тестирование значений атрибутов этого образца в узлах дерева. Дерево решений может быть легко трансформировано в классификационные правила.

На рис. 1 представлен пример графического представления дерева решений:

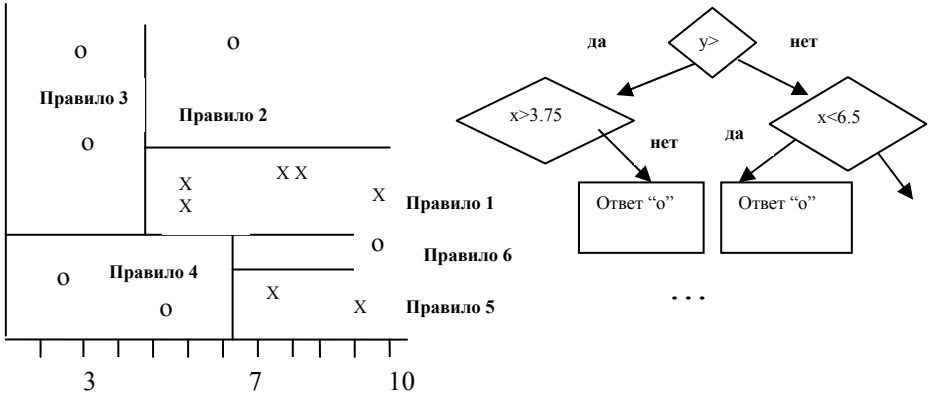

Рис. 1. Графические представления дерева решений

4.3. Алгоритмы обратного распространения (нейронные сети)

Алгоритм обратного распространения является одним из наиболее популярных. Для него требуется задание набора параметров (например, топологии сети), определение которых лучше всего производить эмпирически.

Нейронные сети критикуются за низкую способность к интерпретации. [3, с.27] Плюсами является высокая толерантность к зашумленным данным, а

также способность классифицировать данные, которые ранее не участвовали в обучающей выборке. Кроме того, уже разработано несколько алгоритмов, способных извлекать из нейронных сетей закономерности или правила.

4.4. Моделирование подмножества данных

Примеры вариантов моделирования подмножества данных представлены на Рис. 2.

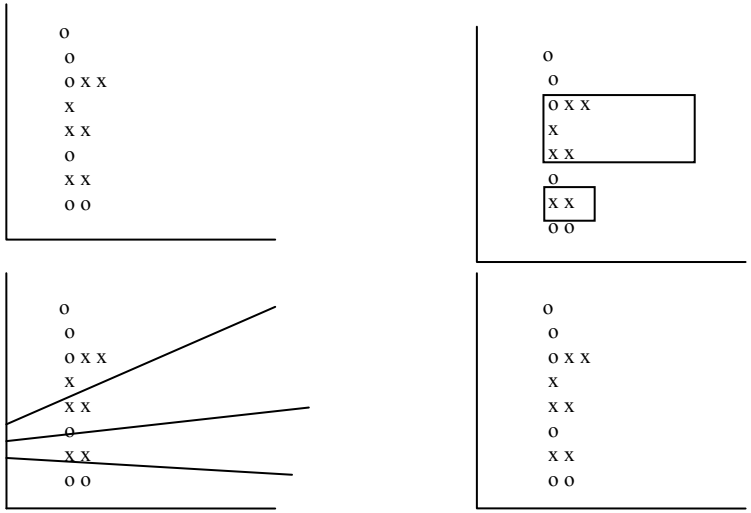

Рис. 2. Моделирование подмножества данных

4.5. Ассоциативные правила

Существует несколько разновидностей классификаторов, основанных на ассоциативных правилах.

Первый метод извлекает ассоциативные правила, основанные на кластеризации, и затем использует эти правила для классификации. В ARCS (Association Rule Clustering System) используется метод ассоциативных правил по формуле `A(quan1) Λ A(quan2) => A(cat)`, где `A(quan1)` и `A(quan2)` – это тесты над множеством значений атрибутов (где значения определяются динамически), и `A(cat)` – символизирует класс для атрибутов, представленных в тестовых данных. Ассоциативные правила наносятся в двухмерную картинку. Алгоритм проходит по этой картинке в поисках правил, составляющих прямоугольные кластеры. Примыкающие наборы появляющихся числовых атрибутов могут быть скомбинированы в кластерные правила. Кластеризованные ассоциативные правила, полученные в результате работы алгоритма ARCS, могут применяться для классификации.

Второй метод – ассоциативная классификация. Метод характеризуется правилами в форме `condset => y`, где `condset` – это множество наборов

пар “атрибут-значение”, а `y` – это класс. Правила, требующие минимальных предопределенных условий, называются повторяющимися. Правило, требующее минимальной специфики, называется точным. Метод ассоциативной классификации состоит из двух шагов. На первом шаге ищется набор всех вероятных правил для точных и повторяющихся правил. Используется итеративный метод, где знания от предыдущих шагов используются для упрощения поиска правила. На втором шаге для конструирования классификатора используется эвристический метод, в котором полученные правила располагаются в порядке значимости, основанной на их значениях точности и повторяемости. Алгоритм может требовать нескольких прогонок над множеством данных, в зависимости от длины самого длинного найденного правила. Когда классифицируется новый образец, для классификации используется первое найденное правило. Классификатор также содержит правило по умолчанию, имеющее наименьшую значимость и определяющее "типичный" класс для образца, класс которого не был рассмотрен в обучающем наборе данных.

В третьем методе, называемом CAEP (Classification by Aggregating and Emerging Patterns), используются знания об объектах данных для получения “выявленных шаблонов”, на основе которых в дальнейшем конструируется классификатор. Алгоритм CAEP является достаточно хорошим (по параметру точности) по сравнению с другими алгоритмами. Он также показывает хороший результат, когда в наборе данных основной класс содержит мало тестовых наборов по сравнению с другими классами.

4.6. Классификация по k-ближайшему соседу

Классификатор по ближайшему соседу основан на изучениях аналогий. Обучающая выборка описывается в `n`-мерном пространстве атрибутов. Каждый образец представляется в виде точки этого пространства. Если классификатору на вход дается неизвестный объект, то он ищет `k` ближайших к нему объектов из обучаемого множества. Степень близости определяется понятиями евклидового пространства. Для объектов `X=(x1, x2,…,xn)` и `Y=(y1,y1,…,yn)` $d(X,Y) = \sqrt{\sum (x_i-y_i)^2}$. Неизвестному объекту назначается класс, наиболее часто встречающийся среди `k` ближайших соседей.

Такой алгоритм является “ленивым”, так как начинает строить классификатор, только когда дается неизвестный пример. В то же время, например, алгоритм с построением дерева, строит структуру независимо от того, надо классифицировать что-то или нет. Поэтому метод `k`-ближайших соседей может быть очень долгим для большого числа данных и большого заданного `k`. Для ускорения процесса можно вводить индексы.

4.7. Классификация, основанная на прецедентах

В отличие от предыдущего алгоритма, где данные представлялись в виде точек Евклидова пространства, этот алгоритм преобразует данные или “случаи” как

сложные символические описания. Алгоритм используется для решения проблем, относящихся к обслуживанию клиентов, когда, например, случаи описывают диагностическую проблему, связанную с продуктом. Также он применяется в таких областях, как инженерия и право, где случаями являются технический проект или законы соответственно.

Когда для классификации подается новый случай, сначала проверяется, не существует ли уже точно такая же ситуация (прецедент). Если подобная ситуация найдена, то в качестве решения выдается решение прецедента. Если не находится ни одного прецедента, то алгоритм пытается найти ситуации, в которых повторяются компоненты. Концептуально, эти обучающие ситуации могут рассматриваться как ближайшие соседи для новой ситуации. Если ситуации представляются в виде графа, то поиск ведется по подграфам, похожим на подграфы неизвестной ситуации. Затем алгоритм пытается сопоставить все решения “соседей”, чтобы выдать решение по новой ситуации. Если по какому-либо решению появляется несовместимость, то делается откат в поисках других решений.

В этом алгоритме могут использоваться какие-либо дополнительные знания и стратегия решения задачи, чтобы выдавались правдоподобные комбинированные решения.

Сложными задачами в данном алгоритме является поиск хорошей метрики для сравнения подграфов, разработка эффективной техники индексирования обучающих ситуаций и методы комбинирования решений.

4.8. Генетические алгоритмы. Обучение нейронной сети

Подход генетических алгоритмов [3, с.27] является попыткой объединить идеи природной эволюции. Начальная популяция создается по случайным правилам. Каждое правило может быть представлено набором или строкой битов. В качестве простейшего примера можно рассмотреть тестовую выборку с двумя классами `C1` и `C2`, представленную двумя атрибутами `A1` и `A2`. Тогда правило “если `A1` и не `A2`, то `C2`” может быть закодировано строкой “100”, где два левых бита представляют атрибуты `A1` и `A2`, а правый – класс. Аналогично, правило “если не `A1` и не `A2`, то `C1`” представляется в виде “001”. Вообще, если у атрибута имеется 2^k различных значений, то для кодирования значения этого атрибута могут быть использованы `k` бит.

Основываясь на правиле сохранения подходящих данных, новая популяция формируется из правил, подходящих для текущей популяции, а также “отпрысков” (объектов, порожденных этими правилами) этих правил. Обычно подходящие правила проверяются на пригодность на основе использования набора тестовых данных.

“Отпрыски” создаются путем применения генетических операторов перехода и мутации. В переходе меняются местами подстроки из пары правил, образуя

новую пару правил. В мутации инвертируются (меняются местами) произвольно выбранные биты строки.

Процесс формирования новой популяции, основанный на правилах предшествующих популяций, продолжается до тех пор, пока популяция `P` “развивается” и каждое правило из `P` удовлетворяет предопределенным подходящим условиям.

Генетический алгоритм легко распараллеливается и используется как для классификации, так и для решения задач оптимизации. При поиске закономерностей он может использоваться для оценки пригодности других алгоритмов.

4.9. Классификация для неточных множеств

Теория неточных множеств может использоваться для классификации и выявления структурных связей в нестрогих или зашумленных данных. Этот подход применяется для атрибутов, имеющих дискретные значения. Если атрибуты принимают не дискретные значения, то они должны быть дискретизированы.

Теория неточных множеств основана на установлении эквивалентных классов для заданной обучающей выборки. Все образцы множества, относящееся к эквивалентному классу, являются неразличимыми, т.е. эти образцы считаются идентичными, несмотря на наличие различий в значениях атрибутов. Для множества данных из реальной жизни часто оказывается, что классы не могут быть различаться в терминах возможных значений атрибутов. Неточные множества могут использоваться для неточных или “грубых” классов.

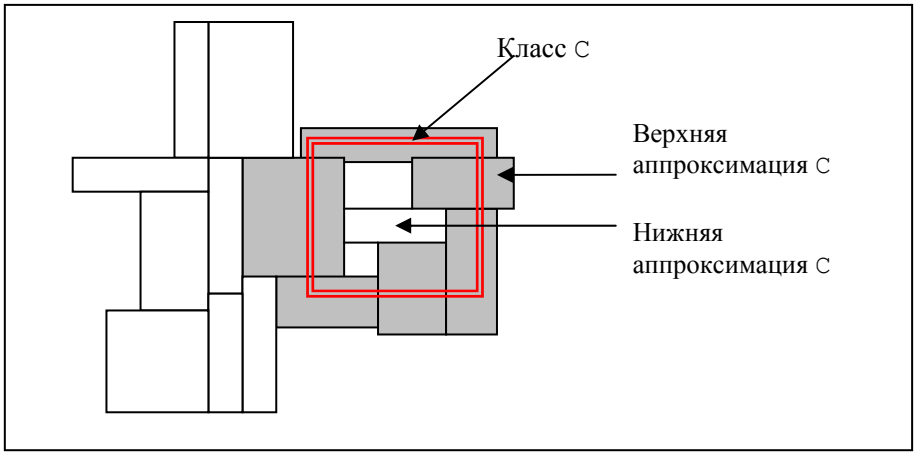

Рис. 3. Графическое представление верхней и нижней аппроксимаций

Неточные множества могут использоваться для будущего восстановления (когда могут быть выявлены и удалены атрибуты, не способствующие классификации) и анализа (когда значения каждого атрибута определяются с учетом поставленной задачи). Задача нахождения минимального подмножества (предела) атрибутов, которые могут описать все концепции имеющегося множества, является NP-сложной.

Пример. У классификаторов, основанных на правилах, имеется тот недостаток, что они строго разделяют непрерывные атрибуты. Рассмотрим, например, приложение для подтверждения потребительского кредита. Изначально правило говорит, что приложение подтверждает кредитоспособность клиентов, работающих более двух лет и обладающих достаточно высоким доходом (например, 50000 в год):

Если {стаж_работы ≥ 2} & {доход ≥ 50000}, то кредит = "подтвержден".

В соответствии с этим правилом клиент получит кредит, если он работает более двух лет и получает, например, 50000, но никак не 49000. Такой строгий порог может выглядеть несправедливым. Нечеткая логика допускает более "мягкие" правила или границы. Вместо того, чтобы строго нарезать множества по категориям, нечеткая логика использует истинностные значения от 0.0 до 1.0 для представления степени принадлежности определенного значения данной категории. Таким образом, при использовании нечеткой логики мы можем пропустить и доход в 49000, но не с такой высокой степенью достоверности, с какой будет одобрен доход в 50000.

Нечеткая логика применима для систем поиска закономерностей, выполняющих классификацию. Это позволяет использовать преимущества работы с высоким уровнем абстракции. В общем случае, использование нечеткой логики в системах, основанных на правилах, предполагает следующее.

- Значения атрибутов переводятся в нестрогую форму. Они разбиваются на дискретные категории: высокую, среднюю и низкую. Как правила, системы, основанные на нечеткой логике, сопровождаются графическими инструментами для пользователей, чтобы они могли задавать эти категории.
- Для нового объекта могут применяться несколько нестрогих правил. Каждое подходящее правило способствует нахождению этого объекта в нужной категории. Обычно истинностные значения для каждой предполагаемой категории суммируются.
- Суммы, полученные на предыдущем шаге, комбинируются в выходное значение. Этот процесс может проводиться путем взвешивания истинностных сумм для каждой категории путем умножения на среднюю величину правдивости категории.

Нестрогая логика используется в ряде систем классификации, например для здоровья, и финансовой сфере.

5. Существующие пакеты для классификации

В химической инженерии развитые модели используются для описания реакций и взаимодействий различных химических процессов. Современные средства разрабатываются также для визуализации этих структур и процессов [2, с.336].

Одной из разработок, поддерживающей подобные функции, является Paviolion Technologies Process Insights, в которой комбинируются нейронные сети, нечеткая логика и статистические методы. Это приложение успешно используется Eastman Kodak и другими компаниями для разработок в химическом производстве и контроле приложений для уменьшения потерь, улучшения качества продукции и повышения производительности.

Это показывает большой интерес экспертов химической области к использованию в своей работе информационных систем. С привлечением все большего числа специалистов будут накапливаться базы информации об исследуемых соединениях, пробах и т.д. и, следовательно, возникнет необходимость анализа этих данных. Поскольку описание проб с помощью признаков является унифицированным для всех экспертов, накапливаемая база может использоваться повсеместно.

Для химической области можно провести некую аналогию с фармацевтикой. На сегодняшний день в медицинской и фармацевтической областях используется классификация. В фармацевтике накоплен огромный объем знаний, разработаны системы для предоставления доступа пользователей к этой информации и организована подписка на обновление данных. Таким образом, пользователи во всем мире могут использовать экспертные знания.

Если говорить о пакетных решениях в области выявления закономерностей, то, как правило, они делятся на три категории: профессиональные, универсальные и специализированные. Профессиональные пакеты рассчитаны на пользователя-специалиста в области статистики. Все универсальные пакеты имеют много пересечений по составу статистических процедур. Специализированные пакеты ориентированы на одну предметную область или несколько смежных областей, их алгоритмы заточены под определенные данные и наиболее актуальные задачи предметной области. В некоторых пакетах реализованы алгоритмы классификации [1, с. 35]. Описания некоторых пакетов приведены ниже:

Наиболее старым продуктом на рынке считается система SAS. Этот пакет включает более 20 различных программных продуктов. Традиционно сложилось, что основными пользователями системы в России являются крупные предприятия, ВПК, государственные структуры. Для классификации в SAS представлены следующие модули:

- BASE SAS – ядро системы со встроенным языком программирования 4GL и поддержкой языка работы с базами данных SQL, средств управления данными, индексов баз данных, возможностей доступа к широкому набору форматов данных, процедур описательной статистики и генерация отчетов;
- FSP обеспечивает полноэкранный доступ к данным, ввод, редактирование, преобразование данных, генерацию отчетов;
- GRAPH поддерживает деловую, научную, рекламную графику, различные шрифты и карты;
- STAT включает в себя многофункциональный набор статистических процедур анализа данных

Основным достоинством SAS считают мощное интеллектуальное ядро, поддержку архитекторы клиент-сервер, возможность доступа и интеграции данных из любых источников. Главный недостаток системы – ее громоздкость, трудности в освоении, высокие требования к квалификации пользователя. Система считается универсальной.

Пакет SPSS является представителем универсальной категории, хотя в первую очередь предназначен для работы статистиков-профессионалов [1, с.38]. Пакет обладает весьма полным набором статистических (более 60) и графических процедур, а также процедур создания отчетов. Кроме того, пакет отличается простым и удобным интерфейсом, и отличается высокой точностью вычислений.

Пакет STATISTICA также ориентирован на пользователей-профессионалов. В нем имеется широкий спектр функциональных алгоритмов. По своей структуре STATISTICA состоит из нескольких связанных между собой “мини-пакетов”. Эти пакеты взаимодействуют между собой, поддерживая один и тот же формат системных файлов.

6. Как будет решаться поставленная задача

В качестве основы для реализации классификации химических проб был выбран алгоритм классификации на основе разделяющей функции. В алгоритме используются знания объектов данных (обучающей выборки) для построения разделяющих функции, которые в дальнейшем применяются для конструирования классификатора.

Описание алгоритма в литературе дает лишь общие принципы работы, а как его реализации могут существенно различаться. Ниже представлено описание реализации алгоритма для классификации.

Работа классификатора делится на два больших этапа: обучение системы и классификация тестового объекта.

Рассмотрим первый этап, на котором строится классификационная модель. Входными данными для этого этапа служит обучающая выборка. Проводится построение разделяющих функций, которые по набору значений атрибутов

класса выдают ответ о принадлежности объекта с такими атрибутами к классу соответствующей разделяющей функции. Построение функций для каждого класса производится независимо друг от друга. Объекты обучающей выборки разбиваются на группы, принадлежащие одному классу и содержащие «подряд идущие значения атрибутов». Если рассмотреть отрезки, в которых содержатся объекты группы, то в них разделяющая функция определяется порядковым номером класса этого объекта. Граничные точки будут определяться по некоторой формуле от тестового объекта и обучающей выборки.

Определение класса неизвестного объекта (пробы) производится с помощью разделяющих функций. Объект принадлежит тому классу, который выдает истинное значение на значениях атрибутах тестового объекта. Если разделяющие функции для всех классов не дают положительного результата, то считается, что тестовый объект принадлежит неизвестному классу.

7. Заключение

В данной статье поставлена задача классификации, сделан обзор алгоритмов и решений по этой тематике, а также описан алгоритм классификации химических проб.

В последующих работах будет представлено обоснование алгоритма, описанного в разделе 5, а также будет описана реализация этого алгоритма и результаты тестирования на реальных данных предметной области.

Литература

1. Дюк В., Самойленко А. «Data Mining: учебный курс». СПб: Питер, 2001.
2. Mehmed Kantardzic, “Data Mining. Concepts, Models, Methods and Algorithms”. Wiley-Interscience, 2003
3. Роберт Калан, “Основные концепции нейронных сетей”. Вильямс, 2003
4. Ю.А. Шрейдер, А.А. Шаров “Системы и модели”, Москва «Радио и связь», 1982.