OSTIS-2012
(Open Semantic Technologies for Intelligent Systems)

УДК 004.942

ИСПОЛЬЗОВАНИЕ ОНТОЛОГИЙ ПРИ ПОСТРОЕНИИ СИСТЕМ РАСПОЗНАВАНИЯ ОБРАЗОВ

Жукевич А.И., Олизарович Е.В., Родченко В.Г.

Гродненский государственный университет имени Янки Купалы,
г. Гродно, Республика Беларусь

san@grsu.by

e.olizarovich@grsu.by

rovar@mail.ru

При построении систем распознавания образов предусматривается возможность выполнения процедуры обучения в автоматическом режиме на основе анализа исходных данных, предварительно формируемых в виде классифицированной обучающей выборки (КОВ). Процесс формирования КОВ начинается с определения алфавита классов и априорного словаря признаков. Создание необходимых онтологий позволит автоматизировать как процесс построения алфавита классов и априорного словаря признаков, так и представление результатов распознавания.
Ключевые слова: алфавит классов, онтология, система распознавание образов, словарь признаков.

Введение

При решении целого ряда фундаментальных и прикладных задач в области естественнонаучных, социально-экономических и гуманитарных дисциплин приходится работать с многомерными объектами, которые обычно характеризуются большим числом признаков. В этом случае применение традиционного математического аппарата оказывается весьма затруднительным, и в качестве альтернативы более эффективным оказывается использование подходов и методов теории распознавания образов [Васильев, 1989].

Построение системы распознавания образов на основе наблюдаемых данных предполагает выполнение двух основных этапов, связанных с реализацией процедуры обучения и процедуры принятия решения. Процесс начинается с определения исходного алфавита класса и априорного словаря признаков.

Соответствующий алфавит классов формируется на основе предварительного анализа требований, предъявляемых к системе распознавания. В априорный словарь должны включаться такие признаки, которые отражают наиболее характерные особенности распознаваемых системой классов. Очевидно, что задача выбора наиболее четко разделяющих классы признаков, является нетривиальной, поскольку, с одной стороны, состояния сложной системы обычно характеризуются большим числом разнообразных по своей природе признаков, а, с другой стороны, часто оказывается невозможным детерминировать закономерности функционирования сложных систем.

При реализации систем распознавания важно ориентироваться на использование комплексного подхода, который предусматривает учет и анализ признаков, характеризующих разнообразные аспекты исследуемых объектов.

Для построения систем распознавания разработан ряд алгоритмов, которые позволяют в автоматическом режиме выполнить как процедуру обучения, так и процедуру принятия решения. Наиболее “узким” местом, с точки зрения автоматизации, остаются шаги алгоритма, связанные с формированием алфавита классов, априорного словаря признаков и представлением результатов распознавания [Родченко, 2008].

Использование онтологий является одним из перспективных направлений автоматизации как выполнения подготовительной работы по формированию алфавита классов и априорного словаря признаков, так и реализации заключительного этапа, связанного с представлением результатов. Таким образом, при построении систем распознавания образов могут быть использованы START-онтологии (S-онтологии) и RESULT-онтологии (R-онтологии).

Описание алгоритма построения системы распознавания

Построение системы распознавания можно осуществить на основе использования универсального алгоритма, предусматривающего выполнение семи шагов.

Первый шаг построения системы распознавания связан с формированием алфавита классов $A=\{A_1,A_2,...,A_k\}$ и априорного словаря признаков $P=\{P_1,P_2,...,P_n\}$. На основе тесного взаимодействия специалистов в соответствующей прикладной области и в области компьютерного анализа данных, путем проведения экспертных оценок и заключений формируются, во-первых, перечень соответствующих классов, который в данном случае будет представлять собой алфавит классов, и, во-вторых, априорный словарь признаков.

На втором шаге построения системы распознавания формируется классифицированная обучающая выборка. Результаты измерений значений всех признаков из априорного словаря для каждого экземпляра класса представляют собой вектор-столбец $x^T = (x_1,x_2,...,x_n)$. Если для каждого j-го класса состояний m_j значений всех соответствующих векторов-столбцов записать в виде таблицы, то результат описания объектов этого класса будет представлять прямоугольную матрицу, состоящую из n строк и m_j столбцов. В результате для каждого класса $A_j \subset A$ формируется соответствующая матрица X_j размерности n x m_j, где m_j – число измерений экземпляра j-го класса. Матрица X_j будет представлять собой формальное описание j-го класса в многомерном априорном признаковом пространстве. Объединенная прямоугольная матрица, построенная на множестве $X=\{X_1,X_2,...,X_k\}$ будет представлять собой классифицированную обучающую выборку.

Третий шаг предназначен для сепарирования признаков из априорного словаря с целью исключения из дальнейших исследований малоинформативных признаков. Природа этих признаков такова, что они “размывают” образы эталонов классов и в итоге создают помехи при выполнении заключительной процедуры принятия решения. В результате выполнения формируется уточненный рабочий словарь, содержащий только признаки, которые наиболее четко отражают особенности каждого класса из априорного словаря.

Отметим, что задача формирования наилучшей системы признаков относится к разряду наиболее сложных с технической и методологической точек зрения. В реальных системах далеко не все признаки, которые первоначально включаются в априорный словарь, пригодны для выполнения непосредственно процедуры принятия решений [Вакульчик и др., 2005]. Ошибка в выборе признаков может приводить к содержательно ложной классификации, даже если при этом она будет формально обоснованной.

Для решения задачи построения наилучшей системы признаков традиционно предлагается воспользоваться эвристическими алгоритмами, которые базируются:

- на полном переборе вариантов и максимизации некоторого критерия информативности признака или подсистемы признаков [Барабаш, 1983];
- на основе случайного поиска с адаптацией, когда наиболее информативная подсистема признаков обнаруживается с помощью случайного перебора подсистем с "поощрением" и "наказанием" отдельных признаков [Загоруйко и др., 1985];
- на применении метода экстремальной группировки признаков или метода корреляционных плеяд [Айвазян и др., 1989].

Для сепарирования признаков по уровню информативности, предлагается воспользоваться алгоритмом, который основывается не на полном или частичном переборе подсистем признаков, а на анализе и учете статистических характеристик выборок значений признаков. Такой подход позволяет на основе априорного словаря автоматически сформировать уточненный рабочий словарь.

Признаки из исходного априорного словаря сепарируются на три вида. К первому виду будут относиться такие признаки, значения которых фактически подчиняются одному и тому же закону распределения во всех классах $A=\{A_1,A_2,...,A_k\}$. Эти признаки не несут разделяющей разные классы функции, а потому будут “размывать” образы классов, как на этапе обучения системы, так и при выполнении процедуры принятия решения.

Ко второму виду будут относиться те признаки, для которых в результате сопоставления всех пар выборок значений этого признака из разных классов оказалось, что ни разу не выполнился соответствующий критерий однородности. Признаки такого вида будут обеспечивать разделение формальных образов классов в многомерном пространстве решений. Именно они и включаются в рабочий словарь информативных признаков, на основе которого в дальнейшем строятся компактные и разделенные в многомерном пространстве решений эталоны классов.

Признаками же третьего вида являются те, которые в процессе выполнения процедуры сепарирования не были отнесены ни к первому, ни ко второму виду. Природа этих признаков такова, что они не отражают какие-либо четко выраженные межклассовые различия.

На основании вышеизложенного, сепарирование признаков из априорного словаря выполняется по следующему сценарию: анализируется содержимое матрицы X, которая получается путем объединения матриц X_1, X_2, ..., X_k. Матрица X будет содержать n строк и m столбцов, причем значение m будет представлять собой сумму количества столбцов во всех матрицах соответственно X_1, X_2,...,X_k, т.е. $m=m_1+m_2+...+m_k$. На основе соответствующего исследуемому признаку P_i статистического

критерия однородности последовательно исследуются все признаки из априорного словаря $P=\{P_1, P_2,...,P_n\}$. В результате они разбиваются на три вида $P^{(1)}=\{P_1^{(1)},P_2^{(1)},...,P_{n1}^{(1)}\}$, $P^{(2)}=\{P_1^{(2)},P_2^{(2)},...,P_{n2}^{(2)}\}$, $P^{(3)}=\{P_1^{(3)}, P_2^{(3)},...,P_{n3}^{(3)}\}$, где $P= P^{(1)}\cup P^{(2)}\cup P^{(3)}$ и $n1+n2+n3=n$.

Отнесение очередного признака P_i к одному из трех видов производится по следующему правилу:

- если для всех возможных пар классов подтвердились гипотезы о статистической однородности выборок значений этого признака для двух сравниваемых классов, то признак P_i относится к первому виду;
- если для всех возможных пар классов оказалось, что выборки значений признака P_i для двух сравниваемых классов подтвердили гипотезу об их неоднородности, то этот признак P_i относится ко второму виду;
- если для признака P_i не выполнилось ни одно из двух предыдущих условий, то он относится к признакам третьего вида.

В рабочий словарь включаются только признаки второго вида $P^{(2)}=\{P_1^{(2)},P_2^{(2)},..., P_{n2}^{(2)}\}$. Отметим, переход к следующему шагу алгоритма происходит только в случае непустого рабочего словаря (т.е. когда $n2\neq0$), а иначе необходимо возвращаться к началу и формировать новый вариант априорного словаря.

Четвертый шаг алгоритма предусматривает проведение аттестации признаков из словаря $P^{(2)}$ и проверки достоверности распознавания на основе использования этих признаков. Из матриц $X_1,X_2,...,X_k$ исключаются строки, содержащие значения признаков первого $P^{(1)}$ и третьего $P^{(3)}$ видов, а все значения признаков второго вида нормируются к единичному интервалу по формуле $y_i=(x_i-x_{min})/(x_{max}-x_{min})$. В итоге получаются матрицы $Y_1,Y_2,...,Y_k$ размерности n_2 x m_i. Для аттестации признаков матрицы, множество $Y=\{Y_1,Y_2,...,Y_k\}$ распределяются на два подмножества таким образом, что $Y_i=Y_i^{(1)}+Y_i^{(2)}$. Матрица $Y_i^{(1)}$ имеет размерность n2 x $m_i^{(1)}$ (где $m_i^{(1)}= [m_i/2]$ - количество объектов, включенных в матрицу $Y_i^{(1)}$), а матрица $Y_i^{(2)}$ будет размерности n2 x $m_i^{(2)}$ (где $m_i^{(2)}=m_i-m_i^{(1)}$- количество объектов, включенных в матрицу $Y_i^{(2)}$). На основе столбцов матрицы $Y^{(1)}=\{Y_1^{(1)},Y_2^{(1)},...,Y_k^{(1)}\}$ строятся эталоны $E^{(1)}=\{E_1^{(1)},E_2^{(1)},...,E_k^{(1)}\}$ для каждого класса и задается пороговое значение допустимости ошибочных классификаций Q.

Далее проводится классификация объектов из множества $Y^{(2)}=\{Y_1^{(2)},Y_2^{(2)},...,Y_k^{(2)}\}$ и вычисляется число ошибочных классификаций G. Затем подмножества $Y^{(1)}=\{Y_1^{(1)},Y_2^{(1)},...,Y_k^{(1)}\}$ и $Y^{(2)}=\{Y_1^{(2)},Y_2^{(2)},..., Y_k^{(2)}\}$ меняются ролями, и на основе $Y^{(2)}$ строятся эталоны классов, а объекты из $Y^{(1)}$ подвергаются процедуре контрольного распознавания. Если значение G не превышает пороговое значение Q, то контрольная аттестация прошла успешно, а иначе необходимо вернутся к началу алгоритма и сформировать новый вариант априорного словаря.

На пятом шаге алгоритма выполняется построение эталонов классов. Один из вариантов реализации этого шага предусматривает, что на основе матриц $Y=\{Y_1,Y_2,...,Y_k\}$ формируются эталоны классов $E=\{e_1,e_2,\ldots,e_k\}$ для “компактных” образов, где эталон i-го класса вычисляется как “центр тяжести” для Y_i, или в более общем виде $E=\{E_1,E_2,\ldots,E_k\}$ для эталонов произвольной сложности.

Шестой шаг связан с выполнением процедуры принятия решения. Распознаваемый объект на основе использования признаков из рабочего словаря формально представляться в виде матрицы Y_{k+1}. Измеряются характеристики взаимного размещения образов эталонов $E=\{E_1,E_2,\ldots,E_k\}$, и образа распознаваемого объекта.

Завершающий седьмой шаг алгоритма связан с окончательной интерпретацией полученных результатов и выработкой заключения. Если в полученном наборе доминируют индексы одного и того же j-го класса, и к Y_{k+1} ближе всего расположен эталон E_j, то это говорит о том, что исследуемый объект относиться к j-ому классу.

О применении онтологий

Онтология позволяет определить единый словарь для специалистов, которые совместно используют информацию в соответствующей предметной области. В нашем случае S-онтология должна включать машинно-интерпретируемые формулировки необходимых базовых понятий предметной области и отношения между ними, тогда как R-онтология ориентирована на машинно-интерпретируемое представление результатов выполнения процедуры распознавания. Онтологии могут быть представлены в виде иерархии и в свою очередь состоят из взаимозависимых онтологий, которые могут быть декомпозированы на составляющие.

Опыт реализации систем распознавания подсказывает, что разработка онтологий потребует затрат разнообразных ресурсов, а потому существует вопрос о реальной потребности в построении онтологий [Олизарович и др., 2010].

Наличие S-онтологии при построении системы распознавания образов предоставляет возможность автоматизировать процедуру формирования алфавита классов, содержащих определение диагностируемых состояний, и процедуру определения априорного словаря признаков, который в общем случае представляют собой выборку из генерального словаря. Построение S-онтологии осуществляется на основе использования семантических сетей.

По результатам проведенного анализа содержимого классифицированной обучающей выборки может быть построена R-онтология,

которая будет использоваться для оперативного выполнения процедуры принятия решений. Специфика R-онтологии такова, что она базируется на использовании матричного формата представления образов типа “объект-свойство” и на формализованном представлении знаний в виде кластерных структур.

Опыт построения онтологии для практического их использования свидетельствуют о том, что любая онтология представляет собой сложную систему, которая в полной мере отвечает пяти признакам, сформулированным Г.Бучем [Буч, 2000].

Выбор того, какие компоненты в онтологии считаются элементарными, относительно произволен и в большей степени остается на усмотрение специалистов. Внутрикомпонентная связь сильнее, чем между компонентами, а сами онтологии состоят из немногих типов структурных компонентов.

Любая работающая онтология является развитием более простой, а разработка онтологии представляет собой итеративный процесс, который обычно продолжается в течение всего жизненного цикла онтологии.

Заключение

Анализ существующих подходов к построению онтологий свидетельствует об отсутствии некой универсальной методологии их разработки. Таким образом, на сегодняшний день не существует некоторого единственно правильного способа моделирования предметной области. Обеспечение доступа всех заинтересованных специалистов к разработке онтологии потенциально должно обеспечивать формирование наиболее качественного варианта и здесь в полной мере и проявляются преимущества проекта OSTIS.

При использовании онтологий в процессе построения аналитических систем предоставляется возможность отделения системных знаний предметной области от оперативных, случайных данных. Таким образом, в процессе построения систем распознавания можно независимо разработать онтологию обобщенного алфавита классов и онтологию обобщенного словаря признаков, а затем на их основе реализовать их оперативное наполнение с целью формирования классифицированной обучающей выборки.

Построение специализированных S-онтологий, в значительной мере ориентировано на использование результатов развития смежных онтологий. Если одна группа специалистов улучшает смежную онтологию, то и за счет этого может происходить развитие S-онтологии и соответственно повышается качество и универсальность соответствующей системы распознавания.

Построение и расширение R-онтологий способствует получению новых знаний об исследуемых сложных системах, поскольку процедура обучения направлена на выявление новых признаков классификации.

Разработка онтологий для анализа систем на основе методов распознавания образов, связана с реализацией как совместного использования специалистами, или программными агентами, так и совместной разработки, что в конечном итоге обеспечивает гораздо более качественное понимание структуры информации.

Библиографический список

[Айвазян и др., 1989] Айвазян, С.А. Прикладная статистика. Классификация и снижение размерности: Справ. изд. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1989.

[Барабаш, 1983] Барабаш, Ю.Л. Коллективные статистические решения при распознавании / Ю.Л.Барабаш – М.: Радио и связь, 1983.

[Буч, 2000] Буч, Г. Объектно-ориентированный анализ и проектирование с примерами приложений на C++ / Г. Буч – М.; СПб.: БИНОМ – Невский диалект, 2000.

[Вакульчик и др., 2005] Вакульчик В.Г. Об одном методе построения математической модели исследования патологических процессов: диагностика острого аппендицита у детей / В.Г. Вакульчик, Ю.В. Макаревич, В.Г. Родченко // Известия Гомельского государственного университета имени Франциска Скорины. – 2005. –№ 5(35). – С.16-19.

[Васильев, 1989] Васильев, В.И. Проблема обучения распознаванию образов / В.И. Васильев – К: Выща шк. Головное изд-во, 1989.

[Загоруйко и др., 1985] Загоруйко, Н.Г. Алгоритмы обнаружения эмпирических закономерностей /Н.Г. Загоруйко, В.Н. Елкина, Г.С. Лбов – Новосибирск: Наука. – 1985.

[Олизарович и др., 2010] Олизарович, Е.В. Метод построения систем диагностики компьютерных сетей на основе применения аппарата прикладной статистики / Е.В. Олизарович, В.Г. Родченко // Известия Гомельского государственного университета имени Франциска Скорины -2010.-№5(62).-С.84-88.

[Родченко, 2008] Родченко, В.Г. Об одном методе формирования пространства решений при построении систем распознавания образов / В.Г. Родченко // Известия Гомельского государственного университета имени Франциска Скорины - 2008. - № 5(50). - С.95-99.

USE OF ONTOLOGY FOR THE CONSTRUCTION OF PATTERN RECOGNITION SYSTEMS

Zhukevich A.I., Olizarovich E.V., Rodchenko V.G.

Yanka Kupala State University of Grodno, Grodno, Republic of Belarus

san@grsu.by

e.olizarovich@grsu.by

rovar@mail.ru

Construction of systems for pattern recognition provides for the possibility of teaching procedures implementation in automatic mode based on the analysis of initial data pre-generated in the form of classified training set (CTS). The process of CTS forming begins with identifying of the alphabet of classes and a priori features dictionary. Formation of the necessary ontology will allow to automate the process of building of the alphabet of classes and a priori features dictionary, as well as presentation of recognition results.