ПРИМЕНЕНИЕ МЕТОДОВ НЕЧЕТКОГО ВЫВОДА В ЗАДАЧАХ КЛАССИФИКАЦИИ И КЛАСТЕРИЗАЦИИ ДЛЯ СИСТЕМЫ КОНТРОЛЯ КАЧЕСТВА ЖИДКИХ НЕФТЕПРОДУКТОВ

Максимова А.Ю.

Институт прикладной математики и механики НАН Украины, Донецк

Системы нечеткого вывода были предложены и практически применены для задач управления сложно формализуемыми процессами. Нечеткое управление применяют, например, в задачах нефтехимической и металлургической промышленности [1]. В основе методов нечеткого управления лежит теория нечетких множеств, которая свое дальнейшее развитие получила в задачах распознавания образов [2].

В задачах контроля качества жидких нефтепродуктов выполняется анализ образцов с целью определения марки и производителя. Особенности технологического процесса, транспортировки и хранения продукции приводят к невозможности принятия однозначного решения для части образцов. Таким образом, возникает задача классификации образцов с неоднозначностью результата. В работе предлагается метод построения нечеткого классификатора на базе нечетких портретов, которые являются интегральными характеристиками классов образов. Нечеткие классификаторы использовались для решения задач классификации в медицине, промышленности [3-5].

В узком смысле задача классификации (распознавания образов) рассматривается как задача машинного обучения по выборки прецедентов. Необходимо разработать и настроить алгоритм, который будет принимать решение о принадлежности рассматриваемого элемента определенному классу образов. Под классами образов понимают поименованные группы объектов. В общем случае приходиться подстраиваться под существующие данные, т.е. под обучающую выборку, в предположении, что она репрезентативна. При этом гарантировать высокое качество распознавания на новых данных нельзя. Данная проблема может быть решена только при условии, что каждая группа объектов обладает общими свойствами, и информация о классе образов может быть представлена в виде его интегральной характеристики. В режиме эксплуатации системы появление нового элемента, принадлежащего неизвестному классу образов, приводит к отказу от распознавания либо к ошибочному отнесению его к одному из существующих классов.

Усугубляют ситуацию различные факторы нечеткости, которые возникают в задачах распознавания образов [6], например неоднозначность, выраженная пересечением классов образов. В ситуациях, когда происходит касание на границах, все еще можно применять стандартные четкие методы распознавания, которые, однако, не будут давать стопроцентной точности распознавания. В ряде задач классы образов обладают свойством априорной неразделимости:

$\exists V_i, V_j : V_i \bigcap V_j \neq \emptyset$, причем $|V_i \bigcap V_j|$ сопоставима с $|V_i|$ либо $|V_j|$: классы образов могут частично пересекаться либо содержаться один в другом. В таких случаях невозможно однозначно отнести рассматриваемый объект к одному из классов образов. Данная проблема решается методами нечеткой математики. Результатом работы алгоритма распознавания будет нечеткий информационный вектор $\tilde{\alpha}(\omega) = (\tilde{\alpha}_1(\omega),...,\tilde{\alpha}_k(\omega))$, где $\tilde{\alpha}_i(\omega) \in [0,1] \bigcup \{\Delta\}$, где Δ – означает отказ от распознавания, а численное значение $\tilde{\alpha}_i(\omega)$ определяет степень уверенности, с которой рассматриваемый образец можно отнести к классу V_j.

Задача контроля качества жидких нефтепродуктов. Предлагаемый в работе метод построения классификатора разрабатывается для задачи анализа жидких нефтепродуктов в лаборатории контроля качества. В данной задаче классы образов представимы в виде иерархической структуры. Существуют различные группы нефтепродуктов, например бензины и дизтопливо, для которых существуют различные марки. Необходимо установить марку и производителя для исследуемого образца. Несколько производителей могут выпускать образцы, имеющие одинаковые показатели, что является фактором неопределенности в рассматриваемой задаче. Особенности транспортировки и хранения продукции приводят к зашумленности образцов за счет присутствия смесей, что также приводит к возникновению неточностей.

В данной задаче возможно с течением времени изменение сформированной иерархической структуры классов образов. В связи с этим актуальным является разработка online метода обучения при пополнении обучающей выборки новыми данными.

Другим аспектом рассматриваемой прикладной задачи является исследование нестандартных образцов, поступающих в лабораторию. Применение методов интеллектуального анализа данных необходимо для получения новых данные, что позволит выявить, например, новые классы образов, по которым отсутствовала информация. Для решения этой задачи используют такие методы анализа данных, как нечеткая кластеризация.

Особенности технологического процесса производства жидких нефтепродуктов приводят к периодическим изменениям некоторых показателей для представителей определенного класса образов, что приводит к возникновению подгрупп внутри классов. Внутриклассовая структура также может быть восстановлена методами интеллектуального анализа данных.

Исходя из нечеткости результатов поставленной задачи, актуальной становиться задача анализа качества работы построенного классификатора. Сравнить качество работы предлагаемого метода классификации с другими возможно только условно, так как большинство алгоритмов тестировались на задачах с четким разделением классов образов. В рассматриваемой прикладной задаче фактор неоднозначности результата закладывает

большую долю ошибки при попытках четкой интерпретации полученного результата, т.к. классы образов имеют пересечения.

Постановка нечеткой задачи распознавания образов с иерархической структурой классов образов. Рассмотрим формальную постановку описанной выше задачи. Дано множество W объектов ω. Объекты заданы векторами значений некоторых признаков $\bar{x} = (x_1, x_2, ..., x_m)$. На всем множестве W задано разбиение на классы $V = \{V_j = (v_1, ..., v_i, ..., v_p) \mid v_i \in L_i, j = \overline{1,k}\}$, где k – количество классов образов, p – количество компонент, определяющих класс образов, L_i - множество возможных лингвистических значений компонент, определяющих класс образов. Исходное множество W формально задано обучающей выборкой $Y = \{\bar{y}^{(i)} = (\bar{x}^{(i)}, V^{(i)}) \mid \bar{x}^{(i)} \in X, V^{(i)} \in V, i = \overline{1,N}\}$, где $\bar{x}^{(i)} = (x_1^{(i)}, ..., x_m^{(i)})$ – вектор информативных признаков i-ого объекта обучающей выборки, $X \subset \{R, \Delta\}^m$. Значение любого признака может быть неопределенным, т.е. априори неизвестным либо ошибочным: в результате анализа данных была выявлена ошибка, которую невозможно исправить. Обозначим такую неопределенность символом Δ. Использование неопределенности для некоторых значение признакового вектора позволяет сохранить данный образец в рассмотрении.

Множество векторов V_j задает иерархическую структуру классов образов. На рис. 1 приведен пример такой иерархии представления классов.

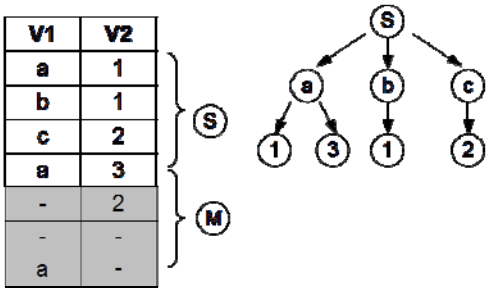

Рис. 1. Иерархическая структура классов образов.

Следует отметить, что в общем случае в исходных данных может возникнуть неопределенность отнесения образца одному из классов. Такие данные выделим в отдельную группу M, содержащее множество нестандартных образцов. Для таких нестандартных образцов выполняется автоматическая классификация.

В классической постановке задачи распознавания классы образов не имеют иерархической структуры. В связи с этим в качестве первого

базового подхода естественным образом в классы образов выбираются группы образцов, которым соответствуют уникальные векторы V_j. Поставим в соответствие каждому уникальному вектору V_j новый идентификатор класса V'_j .

В качестве альтернативного подхода предлагается сохранить иерархическую структуру классов и принимать решение поэтапно. На каждом очередном этапе принимать промежуточное решение относительно принадлежности образца к группам, доступным на текущем уровне структуры классов образов. Множество классов образов $\{V'_j \mid j = 1, k'\}$ на последнем уровне совпадает с аналогичным множеством для случая преобразования классов к единому уровню. На рис. 2. представлен способ перехода к одноуровневой структуре классов для первого подхода.

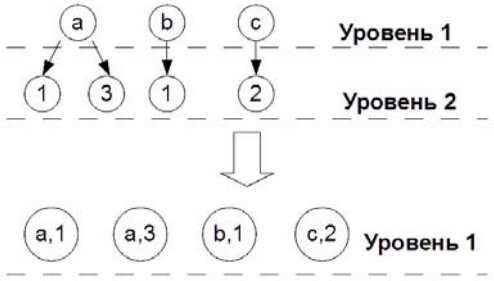

Рис. 2. Переход к одноуровневой структуре

Метод построения нечеткого классификатора. Нечеткие классификаторы строят правила нечетких продукций с использованием лингвистических термов для решения задач распознавания образов. Лингвистические правила для решения задач классификации содержат лингвистические условия в части антецедентов и метки классов образов в качестве консеквентов, которые могут быть представлены нечеткими множествами либо точным значением. Для каждого класса образов может существовать несколько правил в базе правил.

Рассмотрим элементарный нечеткий классификатор, который используется в классической ситуации без учета возможной иерархической структуры классов образов. В основе нечеткого классификатора лежит база правил нечетких продукций. Принятие решений по данной базе правил осуществляется методами нечеткого вывода.

База правил системы нечеткого вывода предназначена для формального представления эмпирических знаний или знаний экспертов о конкретной проблемной области. На формирование базы правил нечеткого вывода оказывают влияние факторы, которые определяются спецификой решаемой задачи и используемым алгоритмом нечеткого вывода [7].

Чтобы задать базу правил необходимо определить:

• множество правил нечетких продукций R вида:
RULE_j: IF $Condition_j$ THEN $Conclusion_j$ (w_j) (1),

где $\{Condition_j\}$ набор подусловий правил нечетких продукций, $\{Conclusions\}$ – набор подзаключений правил нечетких продукций;

• множество лингвистических переменных, которые используются в нечетких высказываниях подусловий (входных лингвистических переменных);
• множество лингвистических переменных (либо четких переменных, либо функциональных зависимостей), которые используются в нечетких высказываниях подзаключений.

Рассмотрим элементарный нечеткий классификатор (Base Fuzzy Clissifier - BFC), который предложен в [8-12]. Множество входных лингвистических переменных, которые используются в подусловиях, соответствует множеству информативных признаков. Количество правил вывод либо соответствует количеству классов образов, либо для случаев, когда выполнен анализ внутриклассовой структуры, соответствует общему количеству классов и подклассов. Подробнее данный аспект рассмотрен ниже.

Продукционное правило нечеткого классификатора имеет вид:

$$ЕСЛИ\ «x_1\ есть\ V_i»\ и\ ...\ и\ «x_m\ есть\ V_i»\ ТО\ «V_i»|\ \tilde{\alpha}_i^j = f(\bar{\mu}),$$

где $\bar{\mu} = (\mu_1^i(x_1),...,\mu_m^i(x_m))$, f - функция, которая рассчитывает степень соответствия рассматриваемого образца $\bar{x} = (x_1,...,x_m)$ классу V_i, $\tilde{\alpha}_i^j$ - степень соответствия $\bar{x}$ классу V_i по данному правилу. Элемент нечеткого вектора определяется следующим образом:

$$\tilde{\alpha}_i = \max_{\forall j}(\tilde{\alpha}_i^j) \quad (1)$$

Операция нечеткого вывода осуществляется следующим образом:

1. Для каждого нечеткого высказывания $«x_u\ есть\ V_i»$ вычисляется значение функции принадлежности $\mu_u^i(x_1)$.
2. Для каждого продукционного правила вычисляется значение степени соответствия классу V_i по формуле $\tilde{\alpha}_i^j = f(\bar{\mu})$, где:

$$f(\bar{\mu}) = \begin{cases} 0, \exists u : \mu_u(x_u) = 0 \\ \log((\mu_1(x_1)+1)(\mu_2(x_2)+1)\cdot...\cdot(\mu_m(x_m)+1))/2^m, \mu_u(x_u) > 0 \end{cases}$$

3. Значение нечеткого информационного вектора вычисляется по формуле (1).

В работе предлагается способ организации иерархического нечеткого вывода, при котором каждому узлу в дереве структуры классов ставиться в соответствие элементарный нечеткий классификатор. Для каждого

классификатора определен свой набор информативных признаков. Если $P = \{R_1,...,R_n\}$ - множество идентификаторов информативных признаков, а $\bar{c} = \{1,0\}^m$ то $P_{pr}^{\bar{c}}$ - набор информативных признаков рассматриваемого классификатора. Таким образов для каждого элементарного узла дерева задан параметр $\bar{c}$, а также определенно множество идентификаторов подгруппы v_i, которые определены по постановке задачи.

Для иерархической системы нечеткого вывода для каждого уровня меткой нечеткого классификатора будет идентификатор узла дерева иерархической структуры. Если значение полученного для данного узла элемента нечеткого информационного вектора больше некоторого порога λ (например $\lambda = 0.5$), то нечеткий классификатор активизируется.

Анализа внутриклассовой структуры и формирование нечетких портретов классов образов. Как было рассмотрено при анализе практической задачи контроля качества нефтепродуктов, классы образов могут состоять из нескольких подгрупп. Удобным инструментом анализа данных для данной задачи являются методы нечеткой и возможностной кластеризации.

Результатом кластеризации является нечеткое с-разбиение в случае нечеткой кластеризации либо нечеткой с-покрытие в случае возможностной кластеризации. Рассмотрим семейство нечетких множеств $R(X) = \{A^l \mid l = 1,...,c\}$, где $A_l = \{(x, \mu_l(x)) \mid \mu_l(x) \in [0,1]]\}, \forall x \in X -$ нечеткие множества, а c – количество кластеров, которые описываются нечеткими множествами. $R(X)$ является нечетким с-разбиением, если для него выполняется условие:

$$\sum_{l=1}^{c} \mu_l(x_i) = 1, i = 1,..., n, \mu_l \in [0,1] \quad (2)$$

или нечетким покрытием, если для него выполняется условие:

$$\sum_{l=1}^{c} \mu_l(x_i) \geq 1, i = 1,..., n, \mu_l \in [0,1] \quad (3)$$

Для каждого множества V_j, $j = 1,...,k$ выполним процедуру кластеризации. В результате получим множество кластеров в виде $R_j(X)$, $j = 1,...,k$. После выполнения кластеризации множеству Y поставим в соответствие множество $R_S = \{R_j(X) \mid j = 1,k\}$ и получим k' нечетких кластеров:

$$k' = \sum_{j=1}^{k} | R_j(X) | \quad (4)$$

Фактически происходит подмена множества классов образов новым расширенным набором классов образов за счет деления на подгруппы, которые определяются методами нечеткой кластеризации [6,13 с. 187-203]. Для каждого полученного кластера строиться свое правило вывода по предложенному в [8-10] алгоритму.

В основе рассмотренного алгоритма лежит понятия нечеткого портрета. Нечетким портретом S_j класса V_j будем называть тройку объектов (S_j, Rs_j, F), где S_j набор значений лингвистических переменных, соответствующий кластерам R_j, Rs_j – множество продукционных правил, соответствующих классу V_j, F – функций расчета степеней соответствия в консеквентах.

Нечеткий портрет определяет нечеткую область n-мерного пространства, аппроксимирующую все элементы рассматриваемого класса образов.

Достоинства и причины использования кластеризации классов образов в задачах обучения с учителем:

- позволяют определить параметры настройки функций принадлежности (ширину скользящего окна и шаг скольжения для алгоритма, предложенного в [8,9]);
- для каждого кластера строиться своя область (граница) подкласса (кластера), что позволяет в случае добавления нового класса избежать переобучения, а построить классификатор для отдельного класса.

Адаптация нечеткого классификатора. Существуют различные подходы к настройки параметров нечеткого классификатора. Предлагаемый в работах [4, 14] метод автоматически генерирует нечеткие продукционные правила. Опираясь на простую нечеткую сетку [14] строятся все возможные правила, а на следующем этапе генетическим алгоритмам строиться минимальный набор правил, при котором ошибка классификации минимальна. Фактически, в данном случае генетический алгоритм позволяет сократить перебор всех правил, которых для задач классификации при больших размерностях признакового пространства $(p+1)^m$, где p - количество элементов нечеткой сетки, а m - количество атрибутов. Достоинством такого подхода является фиксированная треугольная форма функций принадлежности, которая в процессе настройки алгоритма не меняется.

Для нечеткого классификатора, предложенного в [3] количество и вид правил определены изначально, настраиваются лишь веса правил и (или) коэффициенты функций принадлежности. В качестве функций принадлежности используются стандартные треугольные либо трапециевидные функции принадлежности.

В работах [8,9] предложен подход, в котором количество правил определено количеством классов образов, количество антицедентов

соответствует количеству признаков. Однако вид функций принадлежности зависит от частоты встречаемости значений и настраивается в процессе адаптации алгоритма.

В задачах, когда с течением времени возникают новые классы образов, а старые могут исчезнуть из рассмотрения, добавление в систему нового класса образов необходимо автоматизировать. Так как для любого классификатора неизбежным является процесс настройки его параметров (в большинстве случаев для этого решается задача минимизации некоторого функционала), то возникает проблема постоянного дообучения классификатора в процессе эксплуатации, так называемое online обучение.

В зависимости от метода и свойств классификатора может возникает две стратегии дообучения. Рассмотрим два подхода к построению классификатора. В первом случае ищем любую разделяющую границу, во втором случае строим «портреты» классов образов, по существу определяющие границы класса [8-12]: область многомерного пространства, в которую помещаются все классы. Первый подход реализован в линейных классификаторах, методе потенциальных функций, некоторых типах нейронных сетей. При росте количества классов образов эффективность данных методов падает.

При первом подходе построения классификатора, когда строиться любая граница, неизбежно приходиться обучать систему заново при добавлении новой информации. Однако если сразу построить точную границу области класса образов в m-мерном признаковом пространстве, дообучение не понадобиться. Если новый класс образов будет пересекаться с существующими, это отразиться появлением высоких степеней соответствия нескольким классам в результирующем информационном векторе. Если он будет попадать в свободную область, то использование неизбыточных областей будет гарантировать нам наилучшее качество рассматриваемого классификатора.

Апробация результатов исследования. Предложенные в работе методы реализованы в виде программной системы, которая на данном этапе находиться в опытной эксплуатации. Программные модули системы реализованы на языке C++. Разработана структура базы данных для предлагаемой предметной области. Предварительная настройка системы осуществлена на данных последних трех лет – 2008-2010 гг. В табл. 1 приведены результаты тестирования системы без учета on-line обучения. Для оценки качества алгоритма применялся 10-кратный скользящий контроль [15]. Рассчитывался функционал 10-кратного скользящего контроля по следующей формуле:

$$Q_{CV}^{10-fold}(A,X)=\frac{1}{10}\sum_{n=1}^{10}v(A(X^l),X^k) \qquad (5)$$

где A – алгоритм распознавания, $v(A,X)$ – частота ошибок алгоритма A на обучающей выборке X.

Табл. 1 Результаты тестирования.

Год	Количество элементов выборки	$Q_{CV}^{10-fold}$
2008	790	0,96
2009	660	0,93
2010	1546	0,92
2008-2010	2996	0,94

Выводы. В работе выполнена формальная постановка нечеткой задачи классификации для задач с иерархической структурой представления классов образов.

Предлагается метод решения поставленной задачи с учетом возникающих факторов нечеткости. Результатом работы алгоритма является нечеткий информационный вектор, определяющий степени соответствия рассматриваемого образца каждом их листовых классов образов. Высокое качество распознавания достигается за счет применения методов нечеткой кластеризации для анализа внутриклассовой структуры.

Для каждого класса образов строятся так называемые нечеткие портреты, которые аппроксимируют классы образов нечеткими областями в n-мерном признаковом пространстве.

Предложенный подход обеспечивает online настройку классификатора без переобучения при добавлении новых классов образов. Достичь такого результата позволяют методы построения нечетких портретов с использованием анализа внутриклассовой структуры.

Полученная в результате адаптации метода база правил представляет в явном и доступном для понимания и интерпретации экспертами виде причинно-следственные связи рассматриваемого множества данных. Кроме того, нечеткий классификатор отличается высокой обобщающей способностью.

Также в работе рассматривается два способа организации нечеткого вывода. Предложен метод иерархического нечеткого вывода, успешно решающий задачу распознавания образов с иерархической структурой классов образов.

Апробация результатов исследования выполнена в разработанной программной системе.

ЛИТЕРАТУРА

1. M. Sugeno, An introductory survey of fuzzy control // Infom. Sci. – 36, 1985. – P. 59-83.
2. Ishibuchi Hisao, Nakashima Tomoharu, Nii Manabu. Classification and Modeling with Linguistic Information Granules. – Spriger, 2005. – 307 p.
3. Штовба С.Д. Порівняння критеріїв навчання нечіткого класіфікатора. // Вісник Вінницького політехнічного інституту: Інформаційні технології та комп`ютерна техниіка. – 2007. – №6. – С. 84-91.

4. Ishibuchi Hisao, Nakashima Tomoharu, Murata Takahiko Performance Evaluation of Fuzzy Classifier Systems for Multidimensional Pattern Classification Problems // Systems, Man, and Cybernetics, Part B: Cybernetics. – 1999. – Vol.29, 5. – P. 601-618.
5. Rotshtein A. Design and Tuning of Fuzzy Rule-Based System for Medical Diagnosis. In «Fuzzy and Neuro-Fuzzy Systems in Medicine». – Boca-Raton: CRC-Press, 1998. – P. 243. – 289.
6. Вятченин Д.А. Нечеткие методы автоматической классификации. – Мн.: УП «Технопринт», 2004. – 219 с.
7. Леоненков А.В. Нечеткое моделирование в среде MATLAB и fuzzyTECH. – СПб.: БХВ-Петербург, 2003. – 736 с.
8. Козловский В.А., Максимова А.Ю. Алгоритм распознавания, основанный на нечетком подходе // Искусственный интеллект. – 2008. – №4. – С.594-599.
9. Козловский В.А., Максимова А.Ю. Решение задачи распознавания образов по нечетким портретам классов // Искусственный интеллект. – 2010. – № 4. – С. 221-228.
10. Козловский В.А., Максимова А.Ю. Нечеткая система распознавания образов для решения задач классификации жидких нефтепродуктов // Научные работы ДонНТУ, серия «Информатика, кибернетика и вычислительная техника» – 2011. – №13 (185). – С. 200-205.
11. Kozlovskii V.A. ,Maksimova A. Yu. Algorithm of Pattern Recognition with ntra-class clustering. // Proceedings of 11th International Conference PRIP`2011 «Pattern Recognition and Information Processing», 18-20 May, Minsk, Belarus. – 2011. – P. 54-57.
12. Козловский В.А., Максимова А.Ю. Построение нечетких характеристик классов образов по выборке прецедентов в задачах распознвания образов // 15 Всероссийская конференция, г. Петрозаводск, 11-17 сентября 2011 г.: Сборник докладов. – М.: МАКС Пресс, 2011. – С. 135-137.
13. Pedrycz , W., Skowron, A.,Kreinovich, V. (eds.): Handbook of Granular Computing. – New York : John Wley & Sons, 2007. – 1116 p.
14. Ishibuchi, H., Nozaki, K., Tanaka, H. Distributed representation of fuzzy rules and its application to pattern classification. // Fuzzy Sets and Systems. – 1992. – Vol 52. – P. 21-32.
15. Воронцов К.В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. Вып. 13: сборник статей. – М.: Физматлит, 2004. – С. 5-36.