УДК 004.93

КЛАССИФИКАТОРЫ НА ОСНОВЕ Z-МОДЕЛИ ИДЕНТИФИКАЦИИ

М.М. ТАТУР, Д.Н. ОДИНЕЦ, В.В. ОСТРОВСКИЙ, Д.А. ЛАВНИКЕВИЧ

Белорусский государственный университет информатики и радиоэлектроники
П. Бровки, 6, Минск, 220013, Беларусь

Поступила в редакцию 10 октября 2010

Предлагается обобщенная модель идентификации, которая демонстрирует гибкую трансформацию в рамках общепринятых парадигм путем изменения настроек. Данная модель может применяться для синтеза различных классификаторов с использованием априорной информации о прикладной задаче идентификации. На базе данной модели представлен подход к решению проблемы генерации репрезентативных обучающих последовательностей и корректного сравнительного анализа классификаторов.

Ключевые слова: адаптивный классификатор, математическая модель идентификации, обработка данных, обработка знаний.

Введение

Эффективность системы распознавания и принятия решений напрямую зависит от классификатора и алгоритма обучения, которые являются интеллектуальным ядром всей системы в целом. В современной теории распознавания известен целый ряд методов классификации (по минимуму расстояния, нейронные сети, машины опорных векторов и др.) с соответствующими алгоритмами обучения, а также многочисленные примеры их приложений [1–5]. При разработке классификатора для какой-либо прикладной задачи исследователь должен выполнить следующие действия:

- выбрать метод классификации (структуру классификатора);
- сформировать репрезентативную обучающую выборку.
- найти и применить эффективный метод обучения.
- оценить эффективность технического решения.
- реализовать классификатор (soft + hardware).

Задача классификации обычно рассматривается как разделение образов в гиперпространстве информативных признаков, а алгоритм обучения – как построение разделяющей гиперповерхности. Наиболее «простые» задачи являются линейно-разделимыми, т.е. для их решения достаточно поверхности, описываемой линейным полиномом. А «сложные» задачи требуют для своего решения разделяющие поверхности полиномов более высоких порядков. И здесь практически единственным подходом к решению задачи остаются многослойные нейронные сети, с достаточно развитым математическим аппаратом. Однако, при всей популярности нейронных сетей сохраняются присущие им недостатки: проблемы сходимости обучения, проблемы соответствия структуры нейронной сети решаемой задаче, а также интерпретация процессов классификации и обучении. Методы логического и нечеткого выводов, деревья решений, которые также применяются для задач классификации и принятия решений оперируют с извлеченными знаниями и поэтому являются хорошо интерпретируемыми.

На практике, часто многоклассовые задачи представляют совокупностью задач идентификации. Это создает определенные удобства в оценке их эффективности, формулировке критериев принятия решения и интерпретации результатов классификации. Поэтому, в исследованиях, часто теоремы, утверждения и т.п. доказываются, иллюстрируются для задач идентификации, а затем обобщаются для многоклассовых задач.

Обучение классификатора, в широкой трактовке, распространяется на все методы классификации, и включает различные аспекты этого вопроса, такие как алгоритмы обучения, выбор критериев принятия решений, формирование репрезентативных обучающих последовательностей, оценка эффективности обучения. Для проведения корректных сравнительных оценок завершенных систем распознавания используют единые базы данных (базы данных лиц FERET Data, базы данных текстур, медицинских диагнозов и др.) [6–8]. Для тестирования и оценки абстрактных классификаторов необходимы численные базы абстрактных данных. В дальнейшем, под обучающей последовательностью, будем понимать базу численных векторов, с известными откликами. В настоящее время нет общепризнанных числовых баз данных с едиными методиками постановки экспериментов для тестирования моделей классификаторов и алгоритмов обучения.

Общий подход

Для пояснения предлагаемого подхода введем понятие идеального классификатора. Под идеальным классификатором будем понимать функцию принятия решения в задаче идентификации

$$Z = F(X, Y), \tag{1}$$

где X – вектор входных информативных признаков; Y – вектор констант – настроек классификатора для решения конкретной задачи, причем функционал F и настройки Y – априори известны.

В классической постановке задачи на разработку классификатора, необходимо подобрать функционал Ψ, с настройками U, такими, чтобы обеспечить минимальную ошибку d (среднеквадратическую ошибку) на обучающей либо тестовой последовательности.

$$Z^* = \Psi(X, U), \tag{2}$$

$$d = \sqrt{\frac{1}{m-1}\left(\frac{\sum z_i - z^*_i{}^2}{\sum z_i^2}\right)}, \tag{3}$$

где m – число тестовых векторов.

В значительном числе прикладных задач функция классификатора может быть записана с известным функционалом и неточно заданными параметрами настройками $Y`$ [9].

$$Z` = F(X, Y`), \tag{4}$$

тогда, при разработке классификатора отпадает необходимость выбора математического метода либо разработки структуры классификатора (структуры нейронной сети). Обучение классификатора будет сводиться к настройке либо уточнению параметров $Y`$.

В настоящей работе предлагается оригинальная модель идентификатора, с условным названием Z-модель. Она напоминает модель формального нейрона и с общих методологических позиций объединяет в себе основные свойства классического и нечеткого прототипов. Применение данной модели позволит синтезировать классификаторы адекватно прикладным задачам, решить проблемы генерации репрезентативных обучающих последовательностей и корректной сравнительной оценки конечных технических решений.

Математическая модель идентификации

Предлагаемая модель идентификации содержит логическую и арифметическую компоненты [10]. В простейшем случае логическая составляющая модели L может быть представлена функцией арифметического минимума для информативных признаков, заданных множеством N'. Дальнейшее развитие модели может осуществляться в направлении реализации нечеткого вывода (в настоящем варианте не рассматривается).

$$L = \min_{i=1}^{n} \varphi'(x_i, a_i, b_i, c_i, d_i), \tag{5}$$

где $\varphi'(x_i, a_i, b_i, c_i, d_i) = \begin{Bmatrix} \varphi(x_i, a_i, b_i, c_i, d_i), i \in N' \\ 1, i \notin N' \end{Bmatrix}$

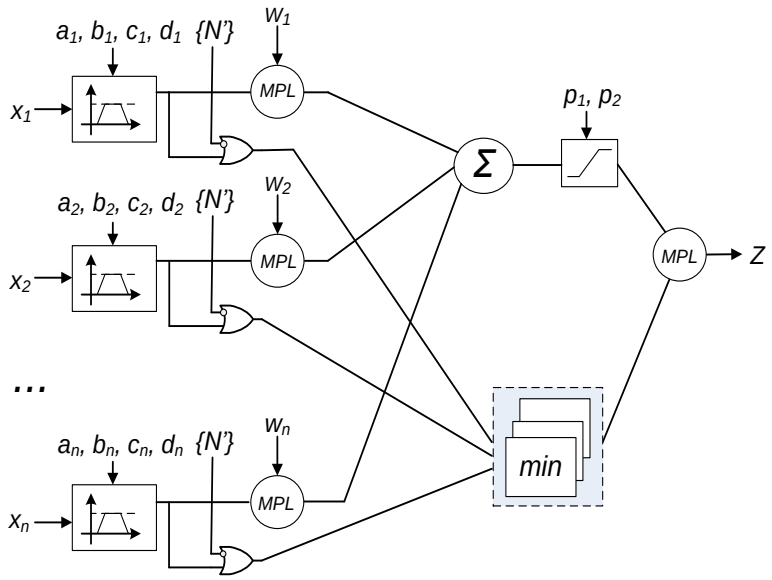
Параметризация информативных признаков $\varphi(x_i,a_i,b_i,c_i,d_i)$, а также функция активации Q линейно аппроксимированы с целью упрощения аппаратной реализации.

$$\varphi(x_i, a_i, b_i, c_i, d_i) = \begin{cases} 0, & a_i < x_i, x_i \geq d_i \\ \dfrac{d_i - x_i}{d_i - c_i}, & c_i \leq x_i \leq d_i \\ 1, & b_i \leq x_i \leq c_i \\ \dfrac{x_i - a_i}{b_i - a_i}, & a_i \leq x_i \leq b_i \end{cases} \tag{6}$$

$$Q = \begin{cases} 1, & S(X) > p_2 \\ \dfrac{S(X) - p_1}{p_2 - p_1}, & p1 \leq S(x) p_2 \\ 0, & S(X) < p_1 \end{cases}$$

где $S(X) = \sum_{i=1}^{n} w_i \varphi(x_i, a_i, b_i, c_i, d_i)$.

На основе приведенных выше формул обобщенная модель элементарного классификатора будет иметь следующий вид:



Графическое представление обобщенной модели элементарного классификатора

Методологическое значение модели состоит в том, что она сочетает как четкую, так и нечеткую, как арифметическую, так и логическую составляющие, способные работать как независимо, так и совместно. В целом, модель связывает два направления известных как обработка данных и обработка знаний. Путем сочетания настроек параметров a_i, b_i, c_i, d_i, w_i, N', p_1, p_2 модель демонстрирует плавную трансформацию качественных различий, позволяя тем самым, формировать типовые функции, адекватные прикладным задачам распознавания:

- при $p_1=p_2$ – реализация четкого принятия решения;
- при $p_1 \neq p_2$ – реализация нечеткого принятия решения;

- при w_i=1 – реализация отсутствия взвешивания признаков для всех i;

- при $w_i \neq 1$ – реализация взвешенного суммирования признаков;

- при a_i=0, b_i=c_i=d_i=1 – реализация линейного масштабирования (отсутствия параметризации) информативных признаков;

- при a_i=b_i, c_i=d_i – реализация четких границ параметров для каждого информативного признака;

- при $a_i \neq b_i$, $c_i \neq d_i$ – реализация нечетких границ параметров для каждого информативного признака;

- при a_i=b_i=c_i=d_i – реализация арифметического или логического совпадения информативного признака с параметром-эталоном;

- если $N' \in \varnothing$ – логические условия не используются;

- если $N' \notin \varnothing$ – реализация логических условий по ключевым признакам, в зависимости от параметризации по каждому информативному признаку;

- при p_1=p_2=0, $N' \notin \varnothing$ – реализация только логических условий.

Генерация «идеальных» тестовых баз

Предлагаемый подход позволяет генерировать численные базы данных $X \rightarrow Z$ для типовых задач классификации. Эти базы могут применяться для тестирования эффективности классификаторов и алгоритмов обучения, независимо от математического метода реализации. Термин «идеальная база» использован потому, что точные значения функции $Z=F(X,Y)$ вычисляются, задав параметры настройки модели – Y и сгенерировав последовательность входных данных – X. А традиционно применяемые базы, как правило, относятся к предметным областям и получены на основе экспертных оценок.

При создании базы, имеется возможность целенаправленно задавать:

- число информативных признаков – n;

- число тестовых векторов – m;

- формат и точность представления входных и выходных данных;

- диапазон нормировки данных;

- вид и уровень шумов.

Тестовые базы могут быть использованы в двух различных приложениях.

В первом приложении:

- для верификации и оценки эффективности конкретного классификатора и алгоритма обучения, разработанных для конкретной системы распознавания;

- для тестирования устойчивости конкретного классификатора и алгоритма обучения к «шумам эксперта».

В этом приложении необходимо тестовую базу генерировать под конкретную типовую задачу распознавания. При постановке экспериментов могут быть использованы частные методики сравнительного анализа.

Во втором приложении:

- для создания единой системы объективной оценки и проведения состязательного отбора классификаторов и алгоритмов обучения;

- для создания библиотеки классификаторов и алгоритмов обучения, отражающей современный уровень знаний в данной области.

В этом приложении создается единая база тестовых данных и единая методика тестирования. Причем, тестовая база охватывает не одну, а «весь» спектр типовых задач.

Разработан прототип тестовой базы под три типовые задачи [10]. Каждая из задач включает четкий и нечеткий способы принятия решений. Тестовая база представляет собой таблицу, в которой сформировано шесть вариантов задач, ранжированных по степени возрастания сложности («интеллектуальности»). В таблице содержатся входные вектора информативных признаков и значения функции классификатора $X \rightarrow Z$. База не зашумлена, следовательно, не рассчитана на проведение экспериментов по анализу устойчивости классификаторов к шумам.

Остальные характеристики базы-прототипа:

- число информативных признаков n=5;

- число тестовых векторов m=100;
- входные данные нормированы в диапазоне [0, 10];
- точность представления входных и выходных данных 0,01. [11]

Участники тестирования должны адаптировать, настроить свои классификаторы на количественные параметры единой базы и осуществить исследования своих программ и моделей по единой методике [12]. Для проведения подобного глобального исследовательского мероприятия тестовая база должна быть верифицирована на репрезентативность, в качестве организатора должна выступать авторитетная научно-исследовательская организация [6,12].

Soft и Hard реализация классификаторов

Предложенная модель идентификации объединяет в себе следующие составные части: модель формального нейрона и одно правило нечеткого вывода. При построении классификаторов для многоклассовых задач, необходимо выполнить тривиальное тиражирование данной модели. Функциональная полнота (способность к нелинейному разделению классов) обеспечивается каждым процессорным элементом. В отличие от многослойных нейронных сетей, которые в каждом конкретном случае имеют уникальную топологию связей, классификаторы на базе предложенной модели будут иметь неизменную топологию аналогичную однослойному персептрону, где число входов равно числу информативных признаков, а число выходов – числу классов. Это обстоятельство позволяет получить ощутимые преимущества при аппаратной реализации ядра классификатора в качестве сопроцессора.

Функции предобработки, обучения, ввода-вывода данных, принятия решения и общего управления возлагаются на основной компьютер, роль которого может выполнить обычный PC. Сопроцессор предназначен для того, чтобы быстро, за счет аппаратного распараллеливания, выполнить K задач идентификации. Соответственно увеличение производительности по сравнению с обычным PC будет тем больше, чем больше число информативных признаков и больше классов имеет прикладная задача. Сопроцессор реализуется на базе ПЛИС и соединяется с основным компьютером посредством одного из стандартных интерфейсов.

Архитектура такого вычислительного комплекса позволяет наращивать вычислительную мощность ядра классификатора в зависимости от решаемой задачи, но при этом оставаться простой в применении конечным пользователем. Предполагается, что стоимость комплекта будет сравнимой со стоимостью серийно выпускаемых нейрокомпьютеров на базе DSP.

Выводы

В настоящей работе предложена *Z*-модель идентификации, функциональность модели включает как операцию взвешенного суммирования, присущую классическому формальному нейрону, так и элементы нечеткого вывода, присущие нечетким нейронным сетям. Предлагается рассматривать эту модель как базовый компонент при построении каскадных классификаторов для произвольного числа классов. Тогда структура классификатора, не зависимо от реализуемой функции, будет одноуровневой и однородной.

В модели достигается возможность гибко учитывать функционал прикладной задачи в зависимости от настроек. На основе *Z*-модели и представленных примеров типовых задач идентификации можно синтезировать классификаторы, адекватные прикладным задачам (при условии априори известного функционала). Если функционал априори неизвестен, то задача построения классификатора по сложности сопоставима с задачей выбора структуры многослойной нейронной сети. В любом случае, структура предложенной модели по сравнению с многослойной нейронной сетью может адекватно интерпретироваться исследователями и разработчиками, может быть использован опыт экспертов для ее целенаправленного развития и обучения.

Путем задания настроек a_i, b_i, c_i, d_i, w_i, N', p_1, p_2 предоставляется возможность получать идеальные значения функции $F(X,Y)$. Таким образом, модель позволяет создавать (генерировать) тестовые численные базы данных для типовых задач классификации. С использованием таких «идеальных» баз (не зашумленных) открывается возможность осуществлять корректный сравнительный анализ классификаторов и алгоритмов обучения, выбирать подходящие для

конкретной задачи методы классификации, разрабатывать и верифицировать алгоритмы обучения. Универсальность модели и возможность гибкой функциональной адаптации путем настроек, позволяют рассматривать ее как математическую основу для реализации массовых операций при разработке перспективных архитектур нейрокомпьютеров.

CLASSIFIERS BASED ON IDENTIFICATION Z-MODEL

M.M. TATUR, D.N. ADZINETS, V.V. OSTROVSKY, D.A. LAVNIKEVICH

Abstract

In this paper we propose a generalized model of identification which shows flexible transformation within the limits of generally accepted paradigms by changing tunings. The application of this model enables to synthesize various classifiers using a priori information about applied tasks of identification. So, we describe the approach to the solution of the problem of generation of representative training sequences and correct comparative evaluation of classifiers.

Литература

1. *Бобов М.Н., Конопелько В.К.* Обеспечение безопасности информации в телекоммуникационных системах. М., 2002.
2. *Кухарев Г.А.* Биометрические системы: Методы и средства идентификации личности человека. СПб., 2001.
3. *Пугачев В.С.* Введение в теорию вероятностей. М., 1968.
4. *Yu J.* // Intelligent Systems. 2005. Vol. 20.
5. *Krisnapuram B., Hartemink A.J., Carin L., Figneiredo M.A* // Pattern Analysis and Machine Intelligence. 2004. Vol. 26. P. 1105–1111.
6. *Melnik O.* // Pattern Analysis and Machine Intelligence. 2004. Vol. 26. P. 973–981.
7. *Chen L., Pan Y., Xu X.* // Parallel and Distributed Systems. 2004. Vol. 25. P. 975–983.
8. *Theodoridis S., Koutroumbas K.* Second Edition. Academic Press an imprint of Elsevier. USA, 2003.
9. http://www.ics.uci.edu/~mlearn/MLRepository.html
10. http://www.grappa.univ-lille3.fr/~torre/Recherche/Experiments/Datasets/
11. http://face.nist.gov/frvt/feret/feretdocuments.htm
12. *Кохонен Т.* Самоорганизующиеся карты. М. 2008.