Proc. Workshop Konnektionismus, St.Augustin,
Arbeitspapiere der GMD, 1988

Mustererkennung mit verteiltem,assoziativem Speicher

Dr.R.Brause, Universität Frankfurt

Abstrakt

Verbindet man alle Eingabeleitungen in einem neuronalen Netzwerk systematisch mit allen Neuroneneinheiten, so lassen sich die Gewichte der Verbindungen in eine Verbindungsmatrix eintragen. Erscheint an den Ausgängen der Neuroneneinheiten direkt die Summe aller gewichteten Beiträge der Inputleitungen, so ist der Output (Menge der Ausgänge) eine lineare Funktion des Inputs. Der Beitrag zeigt, wie durch Einführung einer Schwelle, die für eine Ausgabe bei den Neuroneneinheiten überschritten werden muß, die Menge aller möglichen Inputmuster in Untermengen (Klassen) unterteilt wird. Jeder abgespeicherte Vektor wird zum Stellvertreter (Prototyp) einer ganzen Klasse von Inputvektoren.. Die ursprünglich lineare Operation des assoziativen Auslesens des Speichers wird damit durch die Einführung einer Schwelle zu einer parallelen Mustererkennungsoperation; jedes Inputmuster wird auf den ihm ähnlichst en Klassenprototypen abgebildet.

1.0 Einleitung

In dem folgenden Beitrag werden die Eigenschaften des verteilten, assoziativen Speichers beschrieben, wenn als Ausgabefunktion eine nichtlineare Schwellwertfunktion zugrunde gelegt wird. Im Unterschied zu dem lokalistischen Ansatz wird hier jedes Ereignis nicht durch eine zugeordnete neuronale Einheit codiert, sondern durch einen Zustand aller neuronalen Einheiten. Dadurch ist nicht nur die Zahl der möglichen, codierbaren Ereignisse wesentlich höher, sondern der Ausfall einer Einheit bewirkt nicht automatisch ein "Vergessen" des Ereignisses. Diese Proportionen eines "verteilten" Speichers entsprechen eher den Beobachtungen, die man bei der Suche nach dem Ort einer Erinnerung im menschlichen Gedächtnis machte. Durch Tests an Kriegsverletzten und Gehirnoperierten (Tumore), bei denen Teile des Gehirns funktionsunfähig wurden, erkannte man, daß man zwar ungefähre Gegenden angeben konnte, die für bestimmte Fähigkeiten und Erinnerungen notwendig sind, aber trotz Ausfall erheblicher Gehirnteile die Erinnerungen nicht signifikant eingeschränkt waren und damit nicht streng lokalisiert werden kann.

Die historischen Wurzeln des verteilten, assoziativen Speichers reichen von Ashby (1952) über die Steinbuch'sche Lernmatrix (1961), Willshaw (1964) und Longuet-Higgins (1968) bis Kohonen (1972). Im Unterschied zu dem historischen Ansatz von McCulloch und Pitts (1949), die Neuronenaktivität formal mit binären Signalen, reellwertigen Gewichten und einer Schwelle zu modellieren, beschränkten sich die Modelle des Assoziativen Speichers auf die lineare Funktion zwischen Eingabe- und Ausgabemuster (s. unten). Zwar wurden die potentiellen Mustererkennungs-Möglichkeiten ansatzweise erkannt (z.B. Grossberg /GRO/,p.125; Kohonen /KOH/p.165), aber nicht ausführlich untersucht.

Dieser Beitrag untersucht die Mustererkennungseigenschaften genauer und bestimmt den Wert der Schwelle für eine optimale Klassifikation.

Die Tatsache, daß ähnliche, vom gespeicherten Muster abweichende Eingabemuster die selbe, korrekte Ausgabe bewirken läßt sich auch als *Toleranz gegenüber fehlerhaften Daten* interpretieren. Diese *Software-Fehlertoleranz* ist eng mit der Toleranz gegen Ausfall von Verbindungen (*Hardware-Fehlertoleranz*) verknüpft; in /BRA/ ist dies näher untersucht worden.

Da die Fehlertoleranz-Eigenschaften ohne zusätzliche Maßnahmen und ausschließlich aus den funktionellen Proportionen des Speichermodells resultieren, wird dies als *inhärente Fehlertoleranz* bezeichnet.

Betrachten wir nun das Modell des verteilten, assoziativen Speichers etwas näher.

2.0 Der lineare, assoziative Speicher

Seien die Eingabemuster (Ereignisse, Reize) durch einen reellen Vektor $x = (x_1,...,x_n)$ und die dazu assoziierten Ausgabemuster durch reelle $y=(y_1,...,y_m)$ beschrieben, so läßt sich die Verknüpfung zwischen beiden linear mittels der Matrix $M = (M_{ij})$ modellieren:

$$y = M\,x$$

Hardwaremäßig in einer Implementierung entsprechen den Matrixkoeffizienten M_{ij} die Stärke der Verbindungen zwischen den Eingabeleitungen x und den Ausgabeleitungen y in Abbildung 1.

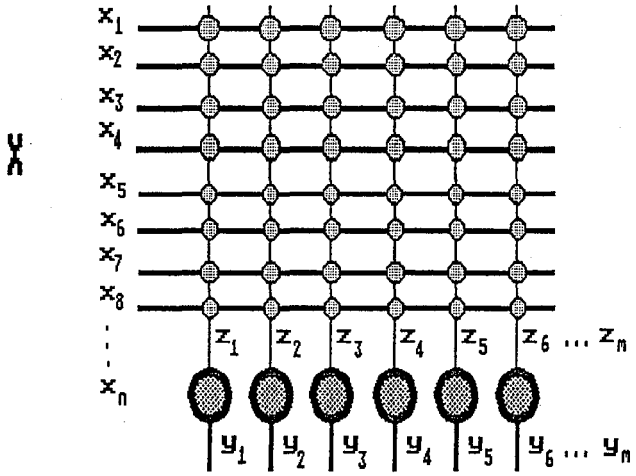

Abb.1 Hardwaremodell

Die Aktivitäten der einzelnen Komponenten von x summieren sich gewichtet in z_i (z.B. als elektr. Ströme) und erzeugen ein Ausgabesignal in y_i.
Zum Speichern eines Paares (x,y) werden gleichzeitig x und y an den Ein- bzw. Ausgängen präsentiert und die Gewichte an den Kreuzungspunkten. verändert, beispielsweise nach der Hebb'schen Regel

$$M_{ij} \sim y_i x_j \qquad \text{Hebb'sche Regel}$$

Nach dem Anlegen von p Mustern $x^1,...,x^p$ resultiren die Gewichte

$$M_{ij} = \sum_k c_k y_i^k x_j^k$$

mit der Proportionalitätskonstanten c_k.
Wird diesem System ein bereits gespeicherter Vektor x^r erneut präsentiert, so ergibt sich als Ausgabe

$$y = M\,x = z =$$

$$= c_r\, y^r x^r x^r + \sum_{k\#r} c_k\, y^k x^r x^k \qquad (2a)$$

ass.Antwort + Übersprechen von anderen Mustern

Normieren wir die Gewichte mit $c_k := 1/(x^k x^k)$ [mit der Notation vw für das Produkt zweier Vektoren v und w] und verwenden ein System von orthogonalen Speichervektoren x^k, so ergibt sich mit $x^i x^j = 0$, i#j wieder die zu x^r assoziierte Antwort y^r.

Fehlerhafte Eingabe

Was geschieht nun, wenn wir ein von x^r abweichendes Muster x dem linearen System präsentieren ?
Sei die Abweichung zu x mit $x' := x - x^r$ bezeichnet, so resultiert als Ausgabe

$$y = M\,x$$
$$= M(x^r + x') =: \quad y^r \quad + \quad y'$$

Orignal + Störterm

die Überlagerung aus der zu x^r assoziierten Antwort und einem "Störterm", der aus einer Linearkombination aller y^k gebildet wird. Ist x' nicht zu groß, so läßt sich dies zur Mustererkennung verwenden. Seien m (= dim(y)) Klassen mit den Ausgabemustern $y^1=(1,0,..,0)$, $y^2=(0,1,0,..0)$, etc , so ist bei Eingabe eines x bei dem Ausgabemuster y der Index der stärksten Komponente die Nummer der Klasse, zu der x gerechnet werden muß.
Ein Beispiel der Bilderkennung mit orthogonal codierten Bildern ist in Kohonen /KOH/, p.124 gezeigt.
Da bei dieser Mustererkennung nur die stärkste Komponente gefragt ist, kann man die korrekte Ausgabe (und damit die Klasse) automatisch dadurch erhalten, daß man alle Komponenten von y vor der Ausgabe einer Schwellwertoperation unterwirft. Ist die Schwelle geeignet gewählt, so wird nur eine Komponente sie überschreiten und das korrekte y^r produzieren.

3. Der lineare Speicher mit Schwellwert

Bisher betrachteten wir reelle Vektoren x und y. Identifizieren wir die reellen Werte mit den Spikefrequenzen der Neuronen, so beschränkt sich der Wertebereich der x und y auf positive Zahlen, da keine negativen Spikefrequenzen existieren.
Betrachten wir nun den Fall beliebiger x^k und orthogonaler y^k (orthogonale Projektion der x auf y), so gibt es nur ein y^{k_i}, bei dem die Komponente y_i ungleich null ist. Damit vereinfacht sich die Gleichung (2a) zu

$$z_i = y_i^{k_i} c_{k_i} x x^{k_i} \qquad (3a)$$

mit k_i aus 1..p, i aus 1..m

Verwenden wir beliebige und nicht, wie Kohonen, orthogonale x^k, so resultiert also als Ausgabevektor z ein Vektor, der in jeder Komponente das innere Produkt aus dem Eingabevektor und dem einzigen Speichervektor, der in dieser Komponente ungleich null ist. Damit ist das Ausgabemuster eine Funktion des Vektorprodukts, das die Kreuzkorrellation zwischen Inputmuster x und einem der Speichermuster x^{k_i} darstellt:

$$y_i = f(z_i) = f(x x^{k_i})$$

Welche Anforderungen werden an die Funktion $T(z_i)$ gestellt?
Bei $x = x^{k_i}$ ist die Kreuzkorrellation und damit z_i besonders groß. Um bei Eingabe von x^{k_i} eine Ausgabe von $y_i^{k_i}$ zu bewirken, muß die Funktion bei besonders hoher Kreuzkorrellation die Komponente $y_i^{k_i}$ ausgeben, sonst Null.
Die Entscheidung für den Vektor x^{k_i} in dem neuronalen Element i stellt somit eine Mustererkennung dar; die Menge aller auf x^{k_i} abgebildeten x bildet eine Klasse, repräsentiert durch den Klassenprototypen x^{k_i}. Das Ähnlichkeitskriterium, mit dem über die Einordnung entschieden wird, ist die Korrellation mit dem Klassenprototypen.
Betrachten wir die geometrische Veranschaulichung dieser Mustererkennung. Für jedes Muster x wird diejenige Klasse r gewählt, bei der für alle anderen Klassenprototypen x^k gilt

$$x x^k < x x^r$$

Aus dieser Relation folgt

$$x x^k - x x^r < 0$$

und erweitert

$$1/2\, x(x^k - x^r) - 1/2\, x^r(x^k - x^r) < -1/2\, x^r(x^k - x^r)$$

und somit

$$(x - x^r)\, 1/2\, (x^k - x^r) < 1/2\, x^r(x^k - x^r)$$

Mit den Definitionen

$$r := 1/2\, (x^k - x^r), \quad s := -x^r, \quad x' := x + s = x - x^r$$

wird obige Relation zu

$$x' r < r s = \text{const}$$

Betrachten wir dazu als geometrische Verdeutlichung im 3-dimensionalen Musterraum die folgende Abbildung 3b:

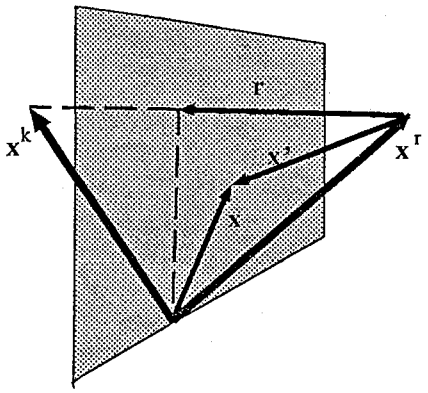

Abb. 3b Klassentrennung im 3-dim Musterraum

Das Produkt rs definiert eine Ebene, die den Raum zwischen x^r und x^k auf halbem Wege zerteilt. Sind x^k und x^r gleich lang, so steht r senkrecht auf der Ebene (Hessesche Normalform).

Für alle x' diesseits der Ebene gilt

$x' r < r s$ x aus Klasse r

und jenseits

$x' r > r s$ x nicht aus Klasse r

Damit ist eine Entscheidung für eine Klasse mit einem konstanten Schwellwert rs ohne explizite Kenntnis der anderen Klasse möglich.

4. Die optimale Schwelle

Wir wir sahen, ist es sinnvoll, die Entscheidungsfunktion f() als Schwellwertfunktion zu wählen.
Es gibt nun verschiedene, sinnvolle Möglichkeiten für f(). Ein Beispiel ist die für biologische Neuronen verwendte Modellierung

$$y_i = (z_i - t_i)T(z_i - t_i) \qquad (4a)$$

mit der Schwelle t_i und der Schwellwertfunktion

$$T(z - t) = \begin{cases} 0 & z \le t \\ 1 & z > t \end{cases}$$

Nach dem Überschreiten einer Schwelle ist in Gleichung (4a) die Antwort wieder linear.
Manche Neurophysiologen sind der Ansicht, daß nicht die Linearität, sondern der Schwellwert den wichtigsten Beitrag zum Ausgabemuster liefert. Hierfür läßt sich die Ausgabe als konstante Schwellwertfunktion (Aktivität/ keine Aktivität) formulieren

$$y_i = y_i^{k_i} T(z_i - t_i) \qquad (4b)$$

Modellieren wir die Neuronenaktivität mit binären Signalen, wie sie in Computern verwendet werden, so sind x_i, y_i aus {0,1} und (4b) wird zu

$$y_i = T(z_i - t_i) \qquad (4c)$$

In allen drei Fällen (4a),(4b) und (4c) ergeben sich ähnliche Mustererkennungseigenschaften. Um die Grundeigenschaften zu verdeutlichen, wählen wir uns das einfachste Modell (4c) aus. Die Gleichung (4c) läßt sich auch schreiben als

$$y_i = \begin{cases} 0 & x x^{k_i} \le t_i \\ 1 & x x^{k_i} > t_i \end{cases} \qquad (4d)$$

mit $y_i^{k_i} = 1$

Wie läßt sich der unbekannte Schwellwert t_i für (4d) ermitteln?
Betrachten wir dazu den Abstand $d(x^k,x^r)/2$ vom Klassenprototypen zur Klassengrenze. Im binären Fall gilt ebenfalls die Dreiecksungleichung

$$d(x,x^r) + d(x,x^k) \ge d(x^r,x^k)$$

Das Betragsquadrat des Abstandes $(x-x^k)^2$ ist im binären Fall gerade die Zahl der nicht übereinstimmenden Stellen, also der Hammingabstand d_H, so daß sich ergibt

$$d_H(x,x^k) = (x-x^k)^2 = |x|^2 - 2xx^k + |x^k|^2$$

Dabei ist $|x|^2$ die Zahl der Einsen in x.
Setzen wir dies in die Dreiecksungleichung ein, so ergibt sich

$$xx^k + xx^r - |x|^2 \le x^k x^r$$

Mit der maximalen Korrelation (Zahl der überlappenden Stellen) K^r_{max} des Klassenprototypen x^r mit allen anderen Klassen gilt

$$x^k x^r \le K^r_{max} := \max_l x^l x^r$$

und somit

$$xx^k + xx^r - |x|^2 \le K^r_{max}$$

An der Klassengrenze ist $xx^k = t = xx^r$, so daß gilt

$$xx^k \le 1/2(K^r_{max} - |x|^2)$$

Die Gleichheit gilt dabei für die Klassengrenze zur Klasse desjenigen Klassenprototypen, der die maximale Korrelation mit x^r aufweist.
Dies ist aber gerade die gesuchte Relation in (4d), so daß sich der obere Ausdruck als für die Unterdrückung der störenden Assoziationen (Übersprechen) notwendige und hinreichende Schwelle verwenden läßt:

$$\boxed{t_i = 1/2\,(K^r_{max} - |x|^2)} \qquad (4e)$$

Bei konstanter Aktivität (konstanter Länge $|x^k|^2 := a$) ist

$$K^r_{max} = 1/2(2a - d^r_{min}) = a - d^r_{min}/2$$

so daß sich die Schwelle auch als Funktion des minimalen Hammingabstandes d^r_{min} der Klasse r zu seinen Nachbarn ausdrücken läßt

$$\boxed{t_i = 1/2\,(a + |x|^2 - d^r_{min}/2)}$$

Hierbei ist die Klassengrenze bei $d^r_{min}/2$, was gut mit der Kodierungstheorie (Fehlererkorrektur bei Blockcodes) übereinstimmt.

Ist nur eine, überall gleiche, konstante Schwelle möglich, so muß von den p Schwellwerten der größte geneommen werden, um fehlerhafte Aktivität zu verhindern. Stellte die obige Festlegung (4e) durch die maximale Korrellation bereits eine Einschränkung der Klassengröße dar, so ist die Festlegung auf K_{max} mit

$$K_{max} = \max_{r} K^{r}_{max}$$

eine weitere Beschränkung der Klassengrößen, die nur bei einer Kodierung mit

$$K^{1}_{max} = K^{2}_{max} = \ldots = K_{max}$$

auf das notwendige Maß erreicht, damit jede neuronale Einheit i für sich entscheiden kann, ohne die Information der anderen Einheiten zu benötigen.

Betrachten wir den Schwellwert in (4e), so geht hier außer konstanten Größen noch der Betrag des Vektors x ein. Das ursprüngliche Hardware-Modell aus Abbildung 1 muß deshalb für die Einordnung von Mustern mit schwankender Aktivität |x| verändert werden. In Abbildung 4 ist dies durch eine zusätzliche Summierung zur Ermittlung von |x| erreicht; die zusätzlichen Knoten bestehen im einfachsten Fall aus Dioden, so daß die Regelspannung für die Schwellwertkomparatoren über den Summenstrom ermittelt werden kann.

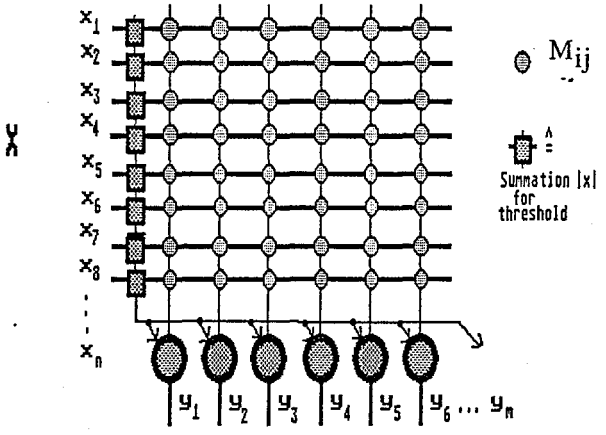

Abb.4 korrigiertes Hardware-Modell

Zusammenfassung

Dieser Beitrag zeigt, wie durch Einführung einer Schwelle aus dem linearen Neuronennetz ein System wird, das auf den Eingabedaten schnell (in einem Zyklus) und parallel eine Mustererkennung und Klassifizierung durchführt. Damit läßt sich diese Operastion als Grundmechanismus einer Stufe von mehrstufigen Kategorisierungs- und Abstraktionsmechanismen verwenden. Als Beispiel ist die Anwendung für die kategoriale Sprachwahrnehmung denkbar.

Diese Arbeit wurde von der Stiftung Volkswagenwerk unterstützt.

Referenzen

/KOH/ T.Kohonen
Self-Organization and Associative Memory, Springer Verlag 1984

/GRO/ S.Grossberg
Adaptive Pattern Classification
Biological Cybernetics 23,
Springer Verlag 1976

/BRA/ R.Brause
Fault-Tolerance Properties of Distributed Associative Memory
submitted to
IEEE Conf. FTCS-18, Tokyo 1988

Adresse des Authors:

Dr. R. Brause
J.-W. Goethe Universität
Fachbereich Informatik, VSFT
Postfach 11 19 32
D- 6000 Frankfurt 11