5. Μέθοδος οπισθοδιάδοσης (backpropagation) του λάθους

Σκοπός:

Προσδοκώμενα αποτελέσματα:

Λέξεις Κλειδιά:

Η μέθοδος οπισθοδιάδοσης του λάθους είναι η πιο δημοφιλής μέθοδος σήμερα για την εκπαίδευση ενός δικτύου που αποτελείται από πολλά επίπεδα, και έχει χρησιμοποιηθεί στις πιο πολλές εφαρμογές. Ιστορικά, πρώτα αναπτύχθηκαν δίκτυα ενός μόνο επιπέδου, όπως ο στοιχειώδης αισθητήρας, τα οποία όμως γρήγορα φάνηκε ότι έχουν μεγάλους περιορισμούς ως προς τις ικανότητες που είχαν, και έτσι σύντομα εγκαταλήφθηκαν. Έτσι φυσιολογικά ακολούθησαν τα δίκτυα πολλών επιπέδων που αναπτύχθηκαν αργότερα, και για τα οποία αρχικά δεν υπήρχαν θεωρητικοί τρόποι για την εκπαίδευσή τους, μέχρι που εμφανίστηκε η μέθοδος οπισθοδιάδοσης. Η μέθοδος αυτή αναπτύχθηκε ανεξάρτητα σε διάφορες παραλλαγές από τους Bryson και Ho (1969), P. Werbos (1974), D. Parker (1982), αλλά διαφημίσθηκε πολύ και προωθήθηκε από το έργο "Parallel distributed processing" των D.E. Rumelhart και J.L. McClelland (1986), το οποίο άνοιξε πολλές εφαρμογές και νέα πεδία, και ανακίνησε μεγάλο ενδιαφέρον σε όλη την περιοχή των νευρωνικών δικτύων. Ως μέθοδος βασίζεται σε καθαρά μαθηματική θεώρηση με δικαιολογημένες αποδείξεις. Το νευρωνικό δίκτυο στο οποίο εφαρμόζεται είναι αρκετά πιο περίπλοκο από τον στοιχειώδη αισθητήρα. Είναι ένα δίκτυο πολλαπλών επιπέδων.

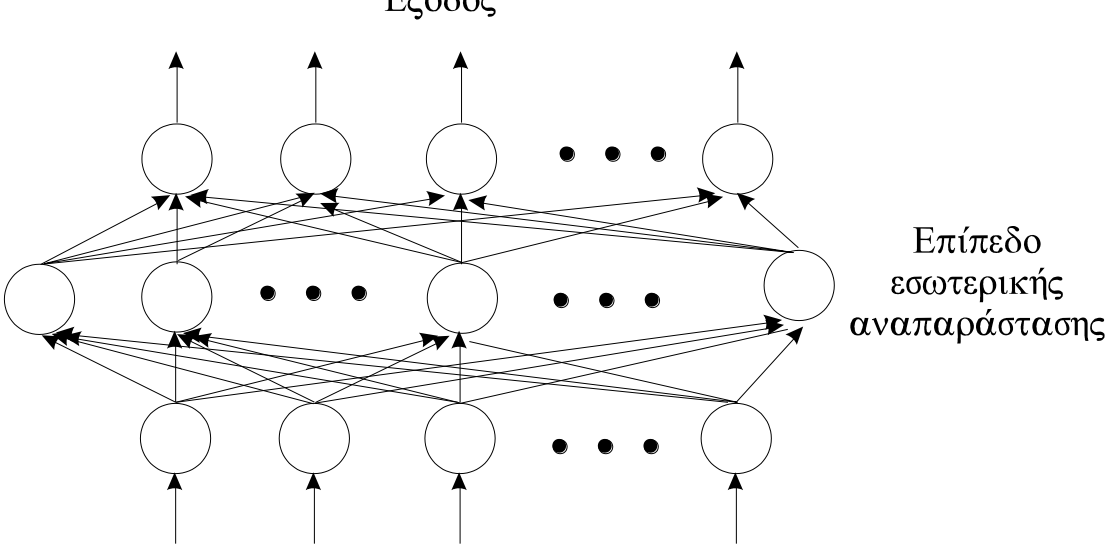
Η κεντρική ιδέα είναι αρκετά απλή: το δίκτυο ξεκινά την διαδικασία μάθησης από τυχαίες τιμές των βαρών του. Εάν δώσει λάθος απάντηση (που είναι και το πιο πιθανό) τότε τα βάρη διορθώνονται έτσι ώστε το λάθος να γίνει μικρότερο. Η ίδια διαδικασία επαναλαμβάνεται πολλές φορές έτσι ώστε σταδιακά το λάθος ελαττώνεται μέχρις ότου

γίνει πολύ μικρό και ανεκτό. Στο σημείο αυτό λέμε ότι το δίκτυο έχει μάθει τα παραδείγματα που του διδάξαμε με την ακρίβεια που θέλαμε να μάθει.

Έχουμε δει λεπτομερώς την δομή του μοντέλου του αισθητήρα, όπου τα εισερχόμενα σήματα στο δίκτυο φθάνουν στο επίπεδο εισόδου, επεξεργάζονται στους νευρώνες, και από εκεί οδηγούνται κατ' ευθείαν προς στο επίπεδο εξόδου. Τέτοια δίκτυα δεν έχουν εσωτερική αναπαράσταση. Αυτό σημαίνει ότι οποιαδήποτε κωδικοποίηση δίδεται στο σήμα εισόδου, ότι είναι αρκετή, καθ' όσον τα πρότυπα που εισάγονται στην είσοδο και αυτά που παράγονται στην έξοδο είναι του ίδιου τύπου. Αυτό επιτρέπει στα δίκτυα αυτά να κάνουν λογικές γενικεύσεις και να βρίσκουν πρότυπα τα οποία ποτέ δεν έχουν δει.

Ο περιορισμός όμως του ότι οι είσοδοι και έξοδοι πρέπει να είναι του ίδιου τύπου δεν τους επιτρέπει να λύσουν πιο γενικά προβλήματα. Αυτό συμβαίνει γιατί τα δίκτυα αυτά δεν έχουν εσωτερική αναπαράσταση. Στο γνωστό πρόβλημα του X-OR βλέπουμε ότι δύο πρότυπα που είναι τελείως διαφορετικά πρέπει να δώσουν ίδια απάντηση. Η λύση στην δυσκολία αυτή βρίσκεται με το να δώσουμε στο δίκτυο μια διαφορετική δομή και να αποκτήσει έτσι μία καινούρια ικανότητα. Προσθέτουμε τώρα και ένα τρίτο επίπεδο, μεταξύ του επιπέδου εισόδου και εξόδου, που ονομάζεται κρυμμένο επίπεδο, και το οποίο τώρα μπορεί να δημιουργήσει την εσωτερική αναπαράσταση των σημάτων εισόδου.

Μετά τις πολλές εργασίες που έγιναν με το μοντέλο του αισθητήρα φάνηκε ότι όταν υπάρχει ένα κρυμμένο επίπεδο τότε δημιουργείται πάντοτε ένας τρόπος αναπαράστασης στο κρυμμένο επίπεδο, το οποίο τώρα μπορεί να ξεπεράσει τον περιορισμό που υπήρχε προηγουμένως περί της ομοιότητος εισόδου-εξόδου. Αρκεί να έχουμε αρκετές μονάδες (νευρώνες) στο κρυμμένο επίπεδο και να βρούμε τα σωστά βάρη w με μια κατάλληλη διαδικασία. Ένα τέτοιο δίκτυο πολλαπλών επιπέδων φαίνεται στο σχήμα 5.1.



ΣΧΗΜΑ 5.1

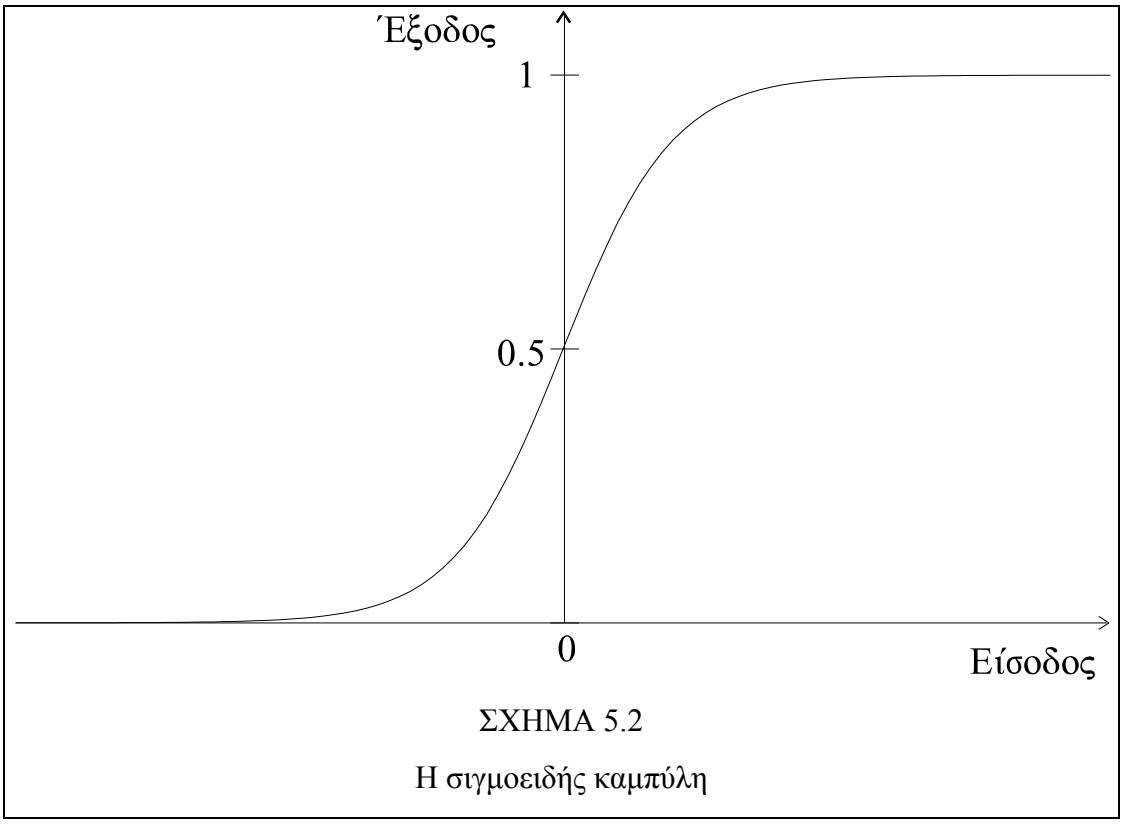
Ένα δίκτυο με πολλαπλά επίπεδα

Πρώτα, υπάρχει ένα επίπεδο εισόδου το οποίο αποτελείται από μία ομάδα νευρώνων οι οποίοι δεν κάνουν ουσιαστικά τίποτα αλλά παρά να δέχονται το σήμα εισόδου. Κατόπιν υπάρχει ένας αριθμός εσωτερικών επιπέδων, κάθε ένα από τα οποία έχει έναν αριθμό νευρώνων, και τα οποία δέχονται το σήμα από το επίπεδο εισόδου, το επεξεργάζονται, και κατόπιν το προωθούν προς την έξοδο. Τέλος υπάρχει ένα επίπεδο εξόδου, το οποίο έχει επίσης έναν αριθμό νευρώνων, οι οποίοι δέχονται σήμα από τα εσωτερικά επίπεδα αλλά δεν κάνουν καμμία επεξεργασία. Απλώς δίδουν αυτό που δέχονται ως έξοδο του δικτύου. Γενικά δεν υπάρχει κανόνας ως προς τον αριθμό τόσο των εσωτερικών επιπέδων όσο και ως προς τον αριθμό των νευρώνων που περιλαμβάνει κάθε επίπεδο (εισόδου, εξόδου, ή εσωτερικό). Η απάντηση σ' αυτό είναι διαφορετική σε κάθε πρόβλημα. Όπως φαίνεται και στο σχήμα οι νευρώνες των διαφορετικών επιπέδων είναι συνδεδεμένοι μεταξύ τους με μία γραμμή. Και στο σημείο αυτό δεν υπάρχει ένας γενικός κανόνας, δηλ. πόσοι και ποιοί νευρώνες είναι συνδεδεμένοι με ποιούς. Σε μία περίπτωση θα μπορούσε κάθε νευρώνας να είναι συνδεδεμένος με όλους τους άλλους νευρώνες, όλων των επιπέδων (μέγιστος αριθμός συνδέσεων). Σε άλλη περίπτωση θα μπορούσε

κάθε νευρώνας να συνδέεται με έναν μόνο άλλο νευρώνα, (ο ελάχιστος αριθμός των συνδέσεων που μπορεί να έχει). Στις ενδιάμεσες περιπτώσεις συνήθως υπάρχουν μερικές συνδέσεις μεταξύ των νευρώνων. Όπως είναι προφανές ο αριθμός των συνδέσεων, ιδίως για την πλήρη συνδεσμολογία είναι πολύ μεγάλος. Αν έχουμε N νευρώνες, τότε ο αριθμός των συνδέσεων σε πλήρη συνδεσμολογία είναιN(N-1)/2. Γιατί;

Η διαδικασία εκπαίδευσης είναι παρόμοια με τον αισθητήρα, αλλά έχει μερικές ουσιώδεις διαφορές. Το σήμα έρχεται σε κάθε νευρώνα του επιπέδου εισόδου (το πρώτο επίπεδο). Πολλαπλασιάζεται επί το αντίστοιχο βάρος w κάθε σύναψης. Σε κάθε νευρώνα αθροίζονται τα καταφθάνοντα γινόμενα και υπολογίζεται το S, όπως και στο μοντέλο του αισθητήρα. Εδώ υπάρχει μία ουσιαστική όμως διαφορά. Ενώ στον αισθητήρα το άθροισμα συγκρίνεται με το θ, εδώ γίνεται κάτι διαφορετικό. Υπάρχει μία συνάρτηση αναπαράστασης (ενεργοποίησης), η οποία είναι χαρακτηριστική του δικτύου, και η οποία χρησιμοποιείται κάθε φορά για να υπολογισθεί ποια θα είναι η τιμή της εξόδου. Έστω ότι η τιμή της εξόδου θα είναι ο. Μία συχνά χρησιμοποιούμενη συνάρτηση είναι η

$$o = \frac{1}{1 + e^{-S}} \qquad (1)$$



ΣΧΗΜΑ 5.2

Η σιγμοειδής καμπύλη

Η συνάρτηση αυτή φαίνεται στο σχήμα (5.2), και έχει τα εξής χαρακτηριστικά. Η τιμή του ο είναι πάντοτε 0<o<1, για οποιαδήποτε τιμή της εισόδου S. Αυτό είναι σημαντικό, διότι έτσι είμαστε βέβαιοι ότι δεν θα υπάρχουν περιπτώσεις που η έξοδος παίρνει μεγάλες τιμές ή απειρίζεται. Η καμπύλη αυτή ονομάζεται σιγμοειδής, λόγω του σχήματος που έχει. Είναι ιδανική συνάρτηση, γιατί συμπεριφέρεται καλά για όλα τα μεγέθη τιμών. Για μικρές τιμές του S η κλίση είναι μεγάλη, και έτσι η έξοδος δεν είναι σχεδόν 0. Ανάλογα, για μεγάλες τιμές του S η κλίση είναι κανονική, ούτως ώστε να μην μπορεί το δίκτυο να κορεσθεί. Μία άλλη ιδιότητα της συνάρτησης αυτής είναι ότι και η παράγωγός της συμπεριφέρεται καλά, κάτι που είναι απαραίτητο στην διαδικασία εκπαίδευσης, παρακάτω. Εύκολα δείχνουμε ότι:

$$\frac{\partial o}{\partial S} = o(1-o) \qquad (2)$$

Μία άλλη ονομασία της συνάρτησης ο είναι συμπιέζουσα συνάρτηση, διότι συμπιέζει οποιαδήποτε τιμή του S στο διάστημα μεταξύ 0 και 1. Παρατηρούμε επίσης ότι η συνάρτηση αυτή είναι μη γραμμική, μία απαραίτητη προϋπόθεση για να μπορεί το δίκτυο να δημιουργήσει αναπαράσταση των σημάτων.

Η συνολική αυτή διαδικασία αποτελεί ένα κύκλο, δηλ. ένα πέρασμα, εισόδος-έξοδος-είσοδος, και συνοψίζοντας περιλαμβάνει τα εξής βήματα:

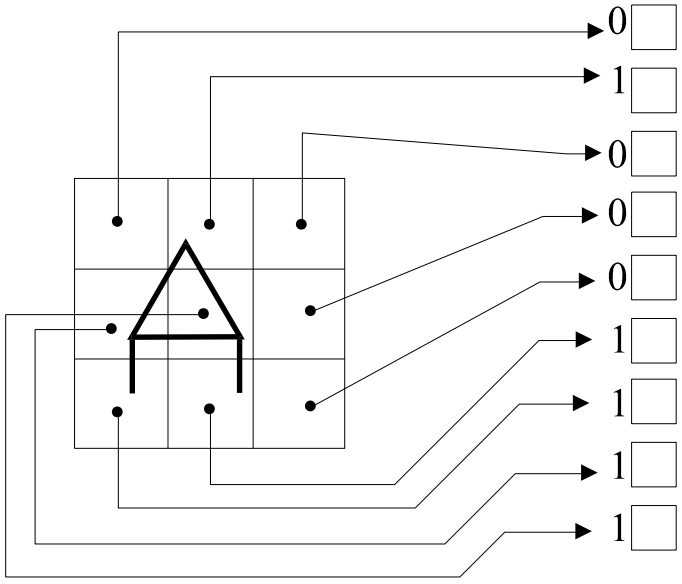
- Παίρνουμε ένα πρότυπο από τα πολλά. Το εισάγουμε στο επίπεδο εισόδου.
- Υπολογίζουμε την έξοδο.
- Το προωθούμε με τον ίδιο τρόπο σε ολα τα επίπεδα μέχρι το τελικό επίπεδο εξόδου.
- Υπολογίζουμε το σφάλμα.
- Μεταβάλλουμε τα βάρη, επιστρέφοντας από την έξοδο προς την είσοδο, ένα-ένα, και επίπεδο-προς-επίπεδο.
- Προχωρούμε στο επόμενο πρότυπο.

Μετά το τέλος ενός κύκλου επαναλαμβάνουμε την διαδικασία για πολλούς κύκλους, όσους χρειάζεται, έως ότου διαδοχικά το σφάλμα φθάσει να είναι αρκετά μικρό. Η ανοχή για το σφάλμα δίδεται εκ των προτέρων, και τυπικές τιμές είναι μερικές % μονάδες, όπως π.χ. 2 ή 5 %.

Ένα παράδειγμα ζεύγους προτύπου-στόχου δίδεται στο σχήμα 5.3, όπου το γράμμα Α έχει σχεδιασθεί σε ένα πλέγμα. Αν οποιαδήποτε γραμμή περνάει μέσα σε ένα τετραγωνάκι, τότε η είσοδος στον αντίστοιχα νευρώνα είναι 1. Διαφορετικά η είσοδος είναι 0. Ως έξοδος μπορεί να είναι ένας αριθμός που παριστάνει το Α, ή ένα άλλο σύνολο από 0 και 1. Για ολόκληρη την αλφαβήτα θα χρειαζόμασταν 24 ζεύγη εκπαίδευσης του δικτύου, ένα ζεύγος για κάθε γράμμα.

Η μέθοδος εκπαίδευσης της οπισθοδιάδοσης του σφάλματος χρησιμοποιεί τις ίδιες γενικές αρχές όπως και ο κανόνας Δέλτα. Το σύστημα πρώτα παίρνει τις εισόδους του πρώτου προτύπου κα με την διαδικασία που περιγράφτηκε προηγουμένως παράγει την

έξοδο. Την τιμή εξόδου την συγκρίνει με την τιμή του στόχου. Εάν δεν υπάρχει διαφορά μεταξύ των των δύο δεν συμβαίνει τίποτα και προχωράμε στο επόμενο πρότυπο. Εάν υπάρχει διαφορά τότε αλλάζουμε τις τιμές των w με τέτοιο τρόπο ώστε η διαφορά αυτή να ελαττωθεί.



ΣΧΗΜΑ 5.3

Αναγνώριση προτύπου

5.1 Η μέθοδος εκπαίδευσης για γραμμικούς νευρώνες

Η μέθοδος αυτή ελαχιστοποιεί το τετράγωνο της διαφοράς μεταξύ της εξόδου που λαμβάνεται και της επιθυμητής τιμής (στόχος), για όλους τους νευρώνες εξόδου και για όλα τα πρότυπα. Αυτό σημαίνει ότι η παράγωγος του σφάλματος ως προς κάθε βάρος w είναι ανάλογος προς την μεταβολή της τιμής του βάρους, όπως δίδεται από τον κανόνα Δέλτα, με αρνητική σταθερά αναλογίας. Αυτό είναι ανάλογο με την διαδικασία της πιο απότομης καθόδου (steepest descent) πάνω στην επιφάνεια που βρίσκεται μέσα στον χώρο των βαρών και στον οποίο χώρο το ύψος είναι ίσο με την τιμή του σφάλματος. Τα παραπάνω ισχύουν για γραμμικές μονάδες νευρώνων.

Έτσι έχουμε:

$$E_p = \frac{1}{2}\sum_j (t_{pj} - o_{pj})^2 \quad (3)$$

όπου E_p είναι το σφάλμα (διαφορά εισόδου-εξόδου) στο πρότυπο p, t_{pj} και o_{pj} είναι ο στόχος και η έξοδος του νευρώνα j για το πρότυπο p. Το συνολικό σφάλμα E είναι:

$$E = \sum_p E_p \quad (4)$$

Για γραμμικές μονάδες εφαρμόζουμε τον κανόνα Δέλτα, και ουσιαστικά έχουμε μία επικλινή κάθοδο (gradient descent) στο E.
Θα δείξουμε ότι:

$$\frac{-\partial E_p}{\partial w_{ji}} = \delta_{pj} x_{pi} \quad (5)$$

που είναι ποσότητα ανάλογη του $\Delta_p w_{ji}$.. Όταν δεν υπάρχουν κρυμμένες μονάδες τότε η παράγωγος υπολογίζεται αμέσως. Χρησιμοποιούμε τον κανόνα αλυσίδας και γράφουμε την παράγωγο ως γινόμενο δύο άλλων παραγώγων: μία παράγωγο του σφάλματος ως προς την έξοδο του νευρώνα επί μία παράγωγο της εξόδου ως προς το βάρος.

$$\frac{\partial E_p}{\partial w_{ji}} = \frac{\partial E_p}{\partial o_{pj}} \frac{\partial o_{pj}}{\partial w_{ji}} \quad (6)$$

Η πρώτη παράγωγος μας λέει πως αλλάζει το σφάλμα ως προς την έξοδο του j νευρώνα, ενώ το δεύτερο τμήμα μας λέει πόσο η μεταβολή του w_{ji} αλλάζει αυτήν την έξοδο. Έτσι υπολογίζουμε κατ' ευθείαν τις παραγώγους:

(7)

Η συνεισφορά του νευρώνα j στο σφάλμα είναι ανάλογη του δ_{pj}. Καθ' ότι έχουμε γραμμικές μονάδες:

$$o_{pj} = \sum_i w_{ji} x_{pi} \tag{8}$$

από το οποίο καταλήγουμε ότι:

$$\frac{\partial o_{pj}}{\partial w_{ji}} = x_{pi} \tag{9}$$

Αντικαθιστώντας στην εξίσωση (χχ) βλέπουμε ότι:

$$-\frac{\partial E}{\partial w_{ji}} = \delta_{pj} x_{pi} \tag{10}$$

όπως ακριβώς θέλουμε. Συνδιάζοντας την τελευταία αυτή εξίσωση με την παρατήρηση ότι

$$\frac{\partial E}{\partial w_{ji}} = \sum_p \frac{\partial E_p}{\partial w_{ji}} \tag{11}$$

μας οδηγεί στο συμπέρασμα ότι η μεταβολή στο w_{ji} μετά από ένα πλήρη κύκλο, όπου παρουσιάζουμε όλα τα πρότυπα, είναι ανάλογη προς στην παράγωγο αυτή, και ως εκ τούτου ο κανόνας Δέλτα εφαρμόζει μία επικλινή κάθοδο στο Ε. Κανονικά τα w δεν πρέπει να αλλάζουν κατά την διάρκεια του κύκλου που παρουσιάζουμε τα διάφορα πρότυπα, ένα-ένα, αλλά μόνο στο τέλος του κύκλου. Αν όμως ο ρυθμός εκπαίδευσης είναι μικρός δεν δημιουργείται μεγάλο σφάλμα και ο κανόνας Δέλτα δουλεύει σωστά. Τελικά θα βρούμε τις τιμές των w που ελαχιστοποιούν την συνάρτηση σφάλματος.

5.2 Η μέθοδος εκπαίδευσης για μη-γραμμικούς νευρώνες

Δείξαμε πως ο κανόνας Δέλτα επιφέρει επικλινή κάθοδο στο τετράγωνο του αθροίσματος του σφάλματος για γραμμικές συναρτήσεις ενεργοποίησης. Στην περίπτωση που δεν έχουμε κρυμμένα επίπεδα, η επιφάνεια σφάλματος είναι σαν ένα μπωλ με ένα μόνο ελάχιστο, και έτσι η επικλινής κάθοδος πάντοτε θα βρίσκει τις καλύτερες τιμές για τα βάρη w. Στην περίπτωση με τα κρυμμένα επίπεδα δεν είναι προφανές πως υπολογίζονται οι παράγωγοι και η επιφάνεια σφάλματος δεν είναι κοίλη προς τα πάνω, και έτσι υπάρχει η πιθανότητα να βρεθούμε σε ένα τοπικό ελάχιστο. Θα δείξουμε παρακάτω ότι υπάρχει ένας αποτελεσματικός τρόπος για τον υπολογισμό των παραγώγων, καθώς επίσης και ότι το πρόβλημα των τοπικών ελαχίστων συνήθως δεν επηρρεάζει την εκπαίδευση του δικτύου.

Χρησιμοποιούμε εδώ δίκτυα με δομές πολλαπλών επιπέδων και στα οποία το σήμα διαδίδεται πάντοτε στην ίδια κατεύθυνση, από το επίπεδο εισόδου προς το επίπεδο εξόδου (feedforward). Το σήμα έρχεται στο επίπεδο εισόδου, στο πιο χαμηλό επίπεδο, επεξεργάζεται από το δίκτυο και προωθείται στα κρυμμένα επίπεδα. Τα κρυμμένα επίπεδα το επεξεργάζονται και το προωθούν στο επίπεδο εξόδου. Η επεξεργασία γίνεται πάντοτε επίπεδο προς επίπεδο, σε κάθε νευρώνα χωριστά. Υπολογίζεται σε κάθε νευρώνα η ενεργοποίηση, χρησιμοποιώντας την μη-γραμμική συνάρτηση, παίρνοντας ως είσοδο την έξοδο του προηγούμενου επιπέδου, και δίδοντας ως έξοδο προς το παραπάνω επίπεδο την υπολογιζόμενη τιμή. Για μια τέτοια, μη γραμμική συνάρτηση η έξοδος είναι:

$$S_{pj} = \sum_i w_{ji} o_{pi} \qquad (12)$$

όπου o_{pi} είναι το σήμα εισόδου του νευρώνα i. Έτσι θα πρέπει:

$$o_{pi} = f_j(S_{pj}) \qquad (13)$$

όπου f είναι διαφορίσιμη και αυξάνουσα συνάρτηση. Γραμμικές συναρτήσεις εδώ δεν επαρκούν, διότι η παράγωγός τους είναι άπειρη στο κατώφλι, και μηδέν σε άλλες τιμές. Θεωρούμε ότι:

$$\Delta_p w_{ji} \approx -\frac{\partial E_p}{\partial w_{ji}} \tag{14}$$

όπου Ε είναι η συνάρτηση σφάλματος (άθροισμα τετραγώνων). Θέτουμε και εδώ την παράγωγο αυτή ως γινόμενο δύο παραγώγων: Μία που δίδει την μεταβολή του σφάλματος ως προς την μεταβολή στην τιμή εισόδου, και μία που δίδει την μεταβολή στην τιμή εισόδου ως προς την μεταβολή του βάρους. Έτσι:

$$\frac{\partial E_p}{\partial w_{ji}} = \frac{\partial E_p}{\partial S_{pj}} \frac{\partial S_{pj}}{\partial w_{ji}} \tag{15}$$

Βλέπουμε ότι:

$$\frac{\partial S}{\partial w_{ji}} = \frac{\partial}{\partial w_{ji}} \sum_k w_{jk} o_{pk} = o_{pi} \tag{16}$$

Ορίζουμε ότι:

$$\delta_{pj} = -\frac{\partial E_p}{\partial S_{pj}} \tag{17}$$

Ο ορισμός αυτός είναι ανάλογος με τον ορισμό της εξίσωσης (χχ), όπου $\delta_{pj}=(o_{pj}-t_{pj})$, καθ' όσον $o_{pj}=S_{pj}$ όταν οι νευρώνες είναι γραμμικοί. Η εξίσωση (χχ) γίνεται τώρα:

$$-\frac{\partial E_p}{\partial w_{ji}} = \delta_{pj} o_{pi} \tag{18}$$

Αυτό δηλώνει ότι για να εφαρμόσουμε την επικλινή κάθοδο ως προς Ε θα πρέπει να κάνουμε τις αλλαγές στα w ως εξής:

$$\Delta_p w_{ji} = \eta \delta_{pj} o_{pi} \tag{19}$$

όπως ακριβώς και στον συνήθη κανόνα Δέλτα. Τώρα πρέπει να υπολογίσουμε τα σωστά δ_{pj} για κάθε νευρώνα του δικτύου. Θέτουμε και εδώ την παράγωγο αυτή ως γινόμενο δύο παραγώγων: μία που δίνει την μεταβολή του σφάλματος ως συνάρτηση της εξόδου, και μία που δίδει την μεταβολή της εξόδου ως συνάρτηση της μεταβολής της εισόδου. Έτσι έχουμε:

$$\delta_{pj} = -\frac{\partial E_p}{\partial S_{pj}} = -\frac{\partial E_p}{\partial o_{pj}} \frac{\partial o_{pj}}{\partial S_{pj}} \tag{20}$$

Αλλά από την εξίσωση (χχ) έχουμε ότι:

$$\frac{\partial o_{pj}}{\partial S_{pj}} = f'_j(S_{pj}) \tag{21}$$

που είναι η παράγωγος της συνάρτησης ενεργοποίησης για τον νευρώνα j, υπολογιζόμενη στο σήμα εισόδου S_{pj} στον νευρώνα αυτό. Τώρα υπολογίζουμε την πρώτη παράγωγο στην εξίσωση του δ_{pj}. Εδώ χρειάζεται προσοχή. Τον παράγοντα αυτόν τον υπολογίζουμε διαφορετικά αν ο νευρώνας είναι στο επίπεδο εξόδου ή εσωτερικός. Στην περίπτωση που είναι στο επίπεδο εξόδου τότε:

$$\frac{\partial E_p}{\partial o_{pj}} = -(t_{pj} - o_{pj}) \tag{22}$$

που είναι το ίδιο αποτέλεσμα όπως με τον συνήθη κανόνα Δέλτα. Αντικαθιστώντας τους δύο παράγοντες στην εξίσωση (χχ) παίρνουμε:

$$\delta_{pj} = (t_{pj} - o_{pj}) f'_j(S_{pj}) \tag{23}$$

για νευρώνες που είναι στο επίπεδο εξόδου. Για νευρώνες που είναι εσωτερικοί υπάρχει το πρόβλημα ότι δεν έχουμε κανένα t_{pj}, δηλ. δεν έχουμε τιμές των στόχων. Στην περίπτωση αυτή έχουμε:

$$\sum_k \frac{\partial E_p}{\partial (S_{pk})} \frac{\partial (S_{pk})}{\partial o_{pj}} = \sum_k \frac{\partial E_p}{\partial (S_{pk})} \frac{\partial}{\partial o_{pj}} \sum w_{ki} o_{pi} =$$
$$= \sum_k \frac{\partial E_p}{\partial (S_{pk})} w_{kj} = -\sum_k \delta_{pk} w_{kj} \qquad (24)$$

Αντικαθιστώντας παρομοίως στην εξίσωση (χχ) παίρνουμε:

$$\delta_{pj} = f'_j(S_{pj}) \sum_k \delta_{pk} w_{kj} \qquad (25)$$

Οι εξισώσεις (xy και yy) δίνουν τον τρόπο με τον οποίο υπολογίζονται όλα τα δ, για όλους τους νευρώνες στο δίκτυο, και τα οποία χρησιμοποιούνται για να υπολογίσουμε την μεταβολή στα w σε όλο το δίκτυο. Η διαδικασία αυτή θεωρείται ότι είναι ένας γενικευμένος κανόνας Δέλτα.

Ως περίληψη, η παραπάνω διαδικασία μπορεί να συνοψισθεί σε τρεις εξισώσεις. Πρώτα, εφαρμόζουμε τον γενικευμένο κανόνα Δέλτα με τον ίδιο τρόπο όπως και τον γενικό κανόνα. Το w σε κάθε επίπεδο αλλάζει κατά μία ποσότητα που είναι ανάλογη του σήματος σφάλματος δ, και ανάλογος επίσης της εξόδου ο. Δηλαδή,

$$\Delta_p w_{ji} = \eta \delta_{pj} o_{pi} \qquad (26)$$

Οι άλλες δύο εξισώσεις δίδουν το σήμα του σφάλματος. Η διαδικασία του υπολογισμού του σήματος αυτού είναι μία κυκλική διαδικασία που ξεκινάει από το επίπεδο εξόδου. Για ένα νευρώνα στο επίπεδο εξόδου το σφάλμα είναι:

$$\delta_{pj} = (t_{pj} - o_{pj}) f'_j(S_{pj}) \qquad (27)$$

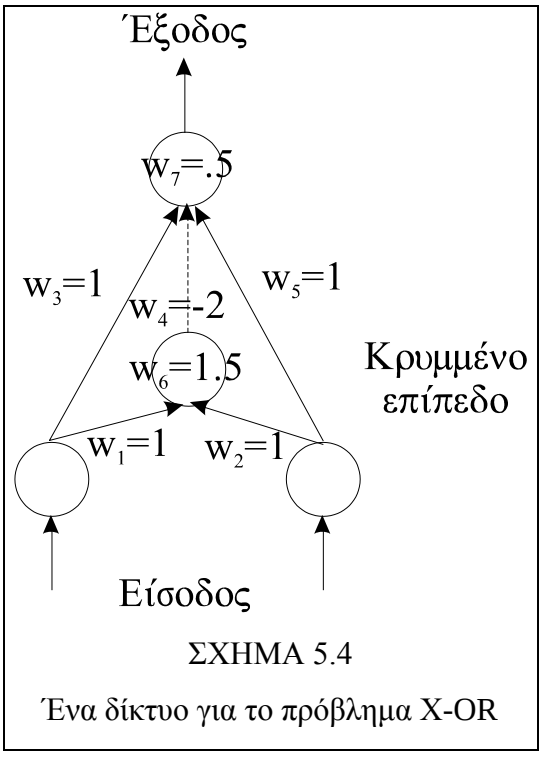
όπου $f'_j(S_{pj})$ είναι η παράγωγος της συνάρτησης ενεργοποίησης. Για νευρώνες στα κρυμμένα επίπεδα δίδεται από:

$$\delta_{pj} = f'_j(S_{pj}) \sum_k \delta_{pk} w_{kj} \qquad (28)$$

Οι τρεις αυτές εξισώσεις αποτελούν έναν κύκλο. Το σύστημα επαναλαμβάνει τόσους κύκλους όσους χρειάζεται για να εκπαιδευτεί.

5.3 Προσομοίωση του προβλήματος X-OR

Το γνωστό πρόβλημα του X-OR τώρα θα το λύσουμε με ένα δίκτυο που περιέχει ένα κρυμμένο επίπεδο, όπως περιγράφηκε παραπάνω και φαίνεται στο σχήμα 5.4. Η δομή



ΣΧΗΜΑ 5.4

Ένα δίκτυο για το πρόβλημα X-OR

του δικτύου περιλαμβάνει δύο νευρώνες

στο επίπεδο εισόδου, ένα νευρώνα στο κρυμμένο επίπεδο και ένα στο επίπεδο εξόδου. Οι συνδέσεις είναι όπως στο σχήμα. Είναι απαραίτητο οι είσοδοι να πηγαίνουν και στο κρυμμένο επίπεδο, και στην έξοδο κατ' ευθείαν. Το πρόβλημα αυτό έχει διδακτική σημασία, καθ' όσον η λύση του περιλαμβάνει όλες τις λεπτομέρειες της τεχνικής της οπισθοδιάδοσης, και γι' αυτό θα σκιαγραφήσουμε την λύση του με μέθοδο προσομοίωσης, ακολουθώντας ένα-ένα τα βήματα και εξισώσεις.

Χρησιμοποιούμε την εξίσωση (1) για τον υπολογισμό των εξόδων από κάθε νευρώνα, την σιγμοειδή συνάρτηση, ως:

$$o_{pj} = \frac{1}{1+e^{-(\sum_i w_{ji} o_{pi} + \theta_j)}}$$

όπου θ_j είναι η παράμετρος προδιάθεσης (ή προδιάθεση) και που παίζει κατά κάποιο τρόπο τον ρόλο του κατωφλίου. Οι τιμές της προδιάθεσης, θ_j θα διδαχθούν στο δίκτυο, όπως και οι τιμές των άλλων βαρών w. Δηλώνει το βάρος w μιας μονάδος (νευρώνα) που πάντα είναι ενεργός. Η εξίσωση της παραγώγου είναι ίδια με την εξίσωση (2). Επίσης οι εξισώσεις των δ είναι οι ίδιες. Η παράγωγος $o_{pj}(1-o_{pj})$ έχει μέγιστο για $o_{pj}=0.5$, και πλησιάζει το ελάχιστό της όταν το o_{pj} πλησιάζει το 0 ή 1, καθ' ότι $0 \le o_{pj} \le 1$. Η μεταβολή σε ένα w είναι ανάλογη της παραγώγου του, και έτσι τα w θα μεταβάλλονται περισσότερο για την περίπτωση που έχουμε μια μεσαία τιμή, δηλ. ούτε ακόμα ενεργό, ούτε μη-ενεργό. Αυτό το χαρακτηριστικό δίνει την σταθερότητα της λύσης του συστήματος.

Πρέπει επίσης να σημειώσουμε ότι όταν ελέγχουμε για τιμές εξόδου 0 ή 1, είναι αδύνατο να πάρουμε ακριβώς τις τιμές αυτές, παρά μόνο αν έχουμε w που τείνουν στο άπειρο. Συνήθως είναι αρκετό όταν παίρνουμε 0.1 και 0.9, έστω και αν αναφέρουμε 0 και 1. Η ακρίβεια αυτή φθάνει.

Η εξίσωση της μεταβολής των w έχει μία σταθερά, το η, που αντιπροσωπεύει τον ρυθμό εκπαίδευσης του δικτύου. Όσο μεγαλύτερο είναι το η, τόσο μεγαλύτερες είναι οι μεταβολές στα w, και τόσο γρηγορότερα το δίκτυο εκπαιδεύεται. Αν όμως το η γίνει πολύ μεγάλο, τότε αυτό οδηγεί σε ταλαντώσεις, και έτσι αναγκαζόμαστε να μην μπορούμε να το αυξήσουμε πολύ. Ένας τρόπος να αυξήσουμε τον ρυθμό εκπαίδευσης και να αποφύγουμε τις ταλαντώσεις είναι να περιλάβουμε και έναν όρο ακόμα που δηλώνει την ορμή του συστήματος. Έτσι η εξίσωση (χχ) γίνεται:

$$\Delta w_{ji}(n+1)=\eta\, \delta_{pj} o_{pi} + \alpha \Delta w_{ji}(n) \qquad (29)$$

όπου το n δηλώνει τον κύκλο, η τον ρυθμό εκπαίδευσης, και α είναι η σταθερά που λαμβάνει υπ' όψιν τις προηγούμενες μεταβολές των w όταν υπολογίζει την νέα μεταβολή. Αυτό είναι μία μορφή ορμής του συστήματος που ουσιαστικά φιλτράρει μεταβολές

υψηλής συχνότητας στην επιφάνεια σφάλματος. Συνήθως παίρνουμε μία τιμή του α που είναι α =0.9

Θεωρούμε αρχικά ότι τα βάρη $w_1=w_2=w_3=w_5=1$. Η τιμή $w_4=-2$ από τον κρυμμένο νευρώνα στον νευρώνα εξόδου καθιστά τον νευρώνα εξόδου μη-ενεργό όταν και οι δύο είσοδοι ταυτόχρονα είναι ενεργοί. Οι αριθμοί μέσα στους νευρώνες είναι οι τιμές του θ. Στο κρυμμένο επίπεδο θ=1.5 διότι έτσι ο νευρώνας αυτός θα πυροδοτεί μόνον όταν και οι δύο νευρώνες του πρώτου επιπέδου είναι ενεργοί. Η τιμή θ=0.5 στον νευρώνα εξόδου καθιστά τον νευρώνα αυτόν ενεργό μόνον όταν λαμβάνει θετικό σήμα μεγαλύτερο από 0.5. Από την πλευρά του νευρώνα εξόδου ο νευρώνας του κρυμμένου επιπέδου φαίνεται ως μια ακόμα μονάδα εισοδου. Τον βλέπει δηλ. σαν να υπήρχαν τρεις τιμές εισόδου. Βέβαια οι τιμές αυτές είναι καθαρά ενδικτεικές, και όχι απαραίτητες για την λύση του προβλήματος.

Εφαρμόζοντας τώρα την μέθοδο προσομοίωσης, λύνουμε το πρόβλημα του X-OR, ώστε το δίκτυο να μπορεί να βρει με επιτυχία τα πρότυπα του Πίνακα 3.χ. Ο αριθμός των κύκλων που χρειάζεται για να γίνει αυτό δίδεται στον Πίνακα 5.χ. Στην λύση που δίνεται παρακάτω ξεκινούμε με τυχαίες τιμές των w που κυμαίνονται στο διάστημα $-0.3<x<0.3$.

Πίνακας 5.χ

η	α	αριθμός κύκλων
0.1	0	82000 ± 24000
0.9	0	8000 ± 2200
0.1	0.1	8100 ± 2200
0.9	0.9	870 ± 230

Οι τιμές στον πίνακα 5.χ. είναι μέσοι όροι από δέκα πραγματοποιήσεις, με διαφορετικές τιμές των w κάθε φορά. Παρατηρούμε ότι η τυπική απόκλιση έχει μεγάλη τιμή, πράγμα που δείχνει ότι η λύση είναι άμεσα εξαρτώμενη από τα αρχικά w.

Συνοψίζοντας λοιπόν την λύση του προβλήματος αυτού δίνουμε το διάγραμμα ροής, καθώς και τις εξισώσεις στα σχήματα 5.8 και 5.9, παρακάτω.

Υπολογισμός των δ

$\delta[4]=der(o[4])\cdot(t-o[4])$

$\delta[3]=der(o[3])\cdot(\delta[4]\cdot w[4])$

$\delta[2]=der(o[2])\cdot(\delta[4]\cdot w[5]+\delta[3]\cdot w[2])$

$\delta[1]=der(o[1])\cdot(\delta[4]\cdot w[3]+\delta[3]\cdot w[1])$

όπου: der(o[i]) η παράγωγος της αντίστοιχης εξόδου

Υπολογισμός των Δw

$\Delta w[1]=\eta\cdot\delta[3]\cdot o[1]$

$\Delta w[2]=\eta\cdot\delta[3]\cdot o[2]$

$\Delta w[3]=\eta\cdot\delta[4]\cdot o[1]$

$\Delta w[4]=\eta\cdot\delta[4]\cdot o[3]$

$\Delta w[5]=\eta\cdot\delta[4]\cdot o[2]$

$\Delta w[6]=\delta[3]$

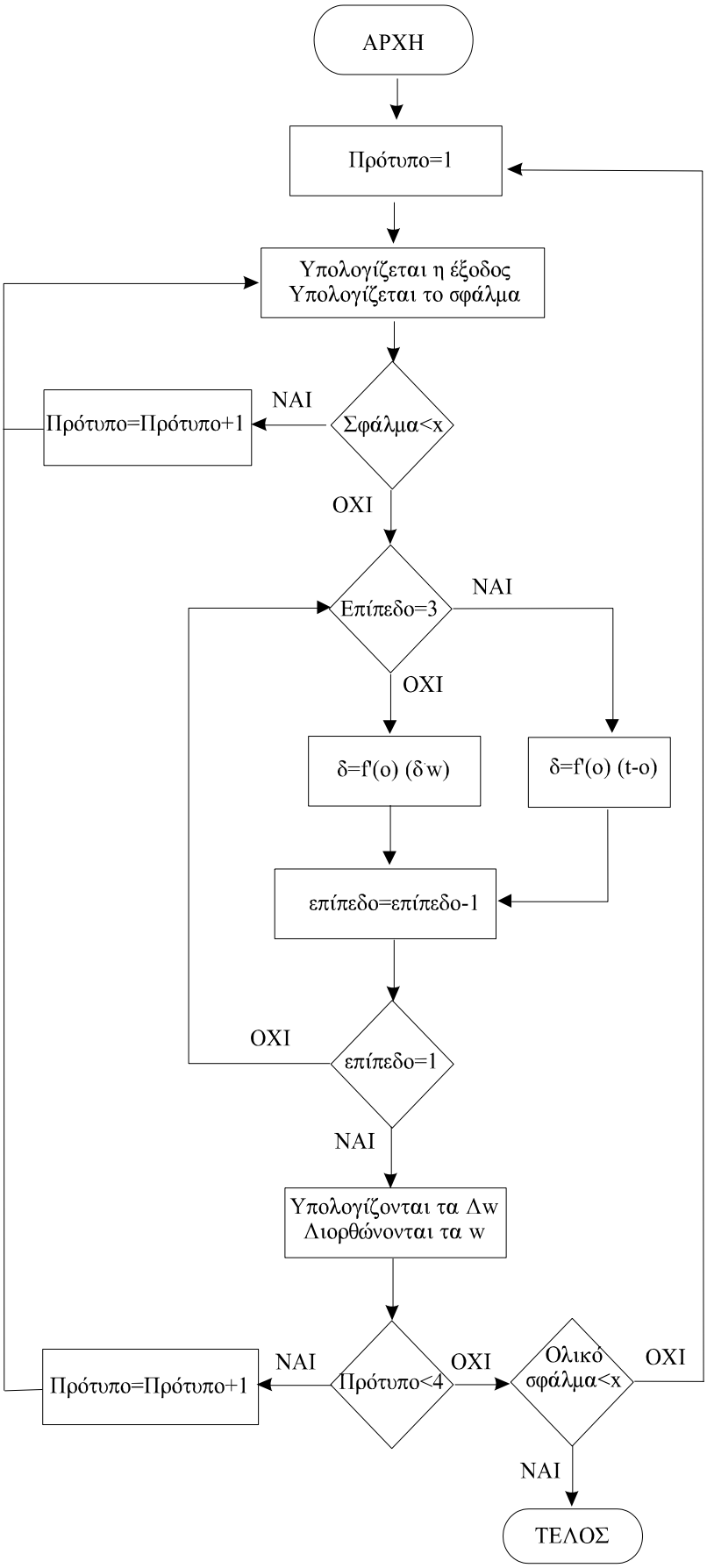
$\Delta w[7]=\delta[4]$

Αλλαγή των w

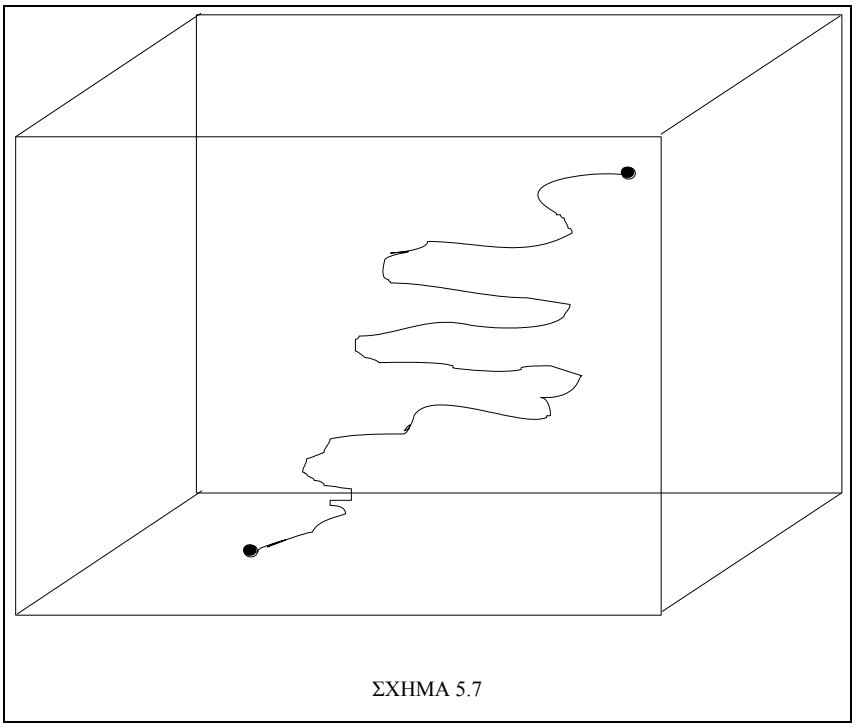
$w_i[n+1]=w_i[n]+\Delta w_i[n]+p.\Delta w_i[n-1]$

όπου i = 1..7

Σχήμα 5.8 Περίληψη εξισώσεων εκπαίδευσης δικτύου


ΑΡΧΗ
Πρότυπο=1
Υπολογίζεται η έξοδος
Υπολογίζεται το σφάλμα
Σφάλμα<x
ΝΑΙ
Πρότυπο=Πρότυπο+1
ΟΧΙ
Επίπεδο=3
ΝΑΙ
ΟΧΙ
δ=f'(o) (δ·w)
δ=f'(o) (t-o)
επίπεδο=επίπεδο-1
επίπεδο=1
ΟΧΙ
ΝΑΙ
Υπολογίζονται τα Δw
Διορθώνονται τα w
Πρότυπο<4
ΝΑΙ
Πρότυπο=Πρότυπο+1
ΟΧΙ
Ολικό σφάλμα<x
ΟΧΙ
ΝΑΙ
ΤΕΛΟΣ

Σχήμα 5.6 Διάγραμμα ροής του αλγορίθμου επίλυσης του XOR



ΣΧΗΜΑ 5.7

5.4 Μειονεκτήματα και προβλήματα

Παρά την μεγάλη επιτυχία της μεθόδου της οποσθοδιάδοσης, εν τούτοις υπάρχουν και περιπτώσεις που η μέθοδος αποτυγχάνει, ή δεν δουλεύει άμεσα με επιτυχία. Σε τέτοιες περιπτώσεις συνήθως χρειάζεται να αλλάξουμε τιμές παραμέτρων, αρχικές συνθήκες, κ.λ.π. μέχρις ότου διορθωθεί το πρόβλημα.

Μερικές φορές ο χρόνος εκπαίδευσης είναι υπερβολικά μεγάλος. Χρειάζονται π.χ. πολλά εκατομμύρια κύκλοι διόρθωσης μέχρις ότου το σύστημα συγκλίνει, ή μπορεί να μην συγκλίνει ποτέ. Σε τέτοιες περιπτώσεις πρέπει να αλλάξουμε το μέγεθος του βήματος. Αυτό συμβαίνει διότι τα βάρη μπορεί να πάρουν μεγάλες τιμές. Αυτό σημαίνει ότι πολλοί νευρώνες δίδουν μεγάλη τιμή εξόδου σε περιοχές όπου η παράγωγος της συνάρτησης εξόδου είναι πολύ μικρή. Καθ' όσον το σφάλμα που επιστρέφει από την έξοδο προς το κρυμμένο επίπεδο μέσα στο δίκτυο είναι ανάλογο της παραγώγου αυτής, μπορεί τότε η διαδικασία εκπαίδευσης να "κωλύσει". Τότε μικραίνουμε το μέγεθος του βήματος, αλλά αυτό έχει ως αποτέλεσμα να μεγαλώσει ο χρόνος εκπαίδευσης.

Ένα άλλο συχνό πρόβλημα είναι αυτό των τοπικών ελαχίστων. Η μέθοδος αυτή, όπως είδαμε παραπάνω, χρησιμοποιεί την μαθηματική τεχνική της επικλινούς καθόδου. Μία εικονική αναπαράσταση της καθόδου αυτής δίδεται στο σχήμα 5.7, όπου βλέπουμε ότι το σφάλμα στην αρχή είναι μεγάλο, αλλά σιγά-σιγά βρίσκει το ελάχιστο μέσα στον κύβο. Ακολουθείται η κλίση της επιφάνειας σφάλματος προς τα κάτω, μεταβάλλοντας συνεχώς τα βάρη μέχρι το σύστημα να φθάσει στο ελάχιστο. Το ελάχιστο αυτό όμως πρέπει να είναι το παγκόσμιο ελάχιστο. Η επιφάνεια μπορεί να έχει πολλά βουνά, λόφους, κοιλάδες, φαράγγια, χαράδρες, κ.λ.π. Αυτό σημαίνει ότι υπάρχουν πολλά τοπικά ελάχιστα, που είναι ψηλότερα από το παγκόσμιο ελάχιστο και στα οποία μπορεί εύκολα να παγιδευθεί το δίκτυο στην προσπάθειά του να βρει το παγκόσμιο ελάχιστο. Καθ' όσον το σύστημα θέλει να πάει πάντα προς τα κάτω, αν πέσει σε ένα τοπικό ελάχιστο δεν έχει τρόπο να απο-παγιδευθεί, και να συνεχίσει το δρόμο του. Συνήθως χρησιμοποιούμε στατιστικές μεθόδους εκπαίδευσης, για να αποφεύγεται το πρόβλημα αυτό.

Το μέγεθος του βήματος επίσης παίζει σημαντικό ρόλο στην ταχύτητα εκμάθησης. Εάν είναι πολύ μικρό, τότε η εκπαίδευση αργεί υπερβολικά, και πρέπει να το αυξήσουμε. Και εδώ η πιο σωστή και ιδανική λύση βρίσκεται με trial-and-error, δηλ. με πολλαπλές δοκιμές μέχρις ότου βρούμε την ιδανική τιμή.

Τέλος, θα πρέπει να θυμίσουμε ότι στην διαδρομή της εκπαίδευσης θα πρέπει να παρουσιάσουμε ταυτόχρονα όλα τα πρότυπα. Οι αλλαγές των βαρών θα πρέπει να γίνονται ταυτόχρονα σε όλα τα πρότυπα. Αν όμως το δίκτυο βρίσκεται σε ένα περιβάλλον το οποίο συνεχώς αλλάζει πρότυπα, τότε η εκπαίδευση του δικτύου δεν θα συγκλίνει ποτέ, και το δίκτυο θα εκπαιδεύεται άσκοπα. Βλέπουμε λοιπόν στο σημείο αυτό ότι η μέθοδος αυτή δεν μιμείται τα βιολογικά συστήματα.

5.5 Εφαρμογές

Όπως αναφέρθηκε παραπάνω η μέθοδος της οπισθοδιάδοσης είναι η πιο κοινή και ευρέως χρησιμοποιούμενη μέθοδος σήμερα για εκπαίδευση νευρωνικών δικτύων. Υπάρχουν πολλές εφαρμογές της, όπως οπτικής αναγνώρισης χαρακτήρων, λήψης αποφάσεων, κ.λ.π.

Η εταιρία Caere έχει σήμερα ένα πρόγραμμα για PC που ονομάζεται Omnipage, που αναπτύχθηκε το 1994, που διαβάζει τυπωμένα κείμενα με scanner και τα μετατρέπει σε χαρακτήρες ascii. Μάλιστα το πρόγραμμα αυτό δουλεύει, ικανοποιητικά έστω και αν τα γράμματα είναι μερικώς κατασταμμένα, όπως π.χ. από σελίδες fax.

Ένα άλλο παρόμοιο πακέττο, το NetTalk, ανπτύχθηκε από τους Sejnowski και Rosenberg (1987) που μετατρέπει με μεγάλη επιτυχία κείμενα Αγγλικών κατ' ευθείαν σε ομιλία.

Υπάρχουν επίσης και προσπάθειες και προγράμματα για την πιο δύσκολη διαδικασία, την αναγνώριση χειρογράφων κειμένων. Οι χαρακτήρες κανονικοποιούνται πρώτα ώστε να έχουν όλοι το ίδιο μέγεθος, μετά τοποθετούνται σε ένα πλέγμα, και γίνονται οι προβολές των γραμμών στα τετράγωνα του πλέγματος. Οι προβολές αυτές είναι οι τιμές εισόδου για το δίκτυο. Η μέθοδος αυτή αναπτύχθηκε από τον Burr (1987) και έχει >99% επιτυχία.

Παρόμοιο πρόγραμμα έχει αναπτύξει και η εταιρία υπολογιστών NEC με ακρίβεια >99% αλλά η αναγνώριση γίνεται με άλλες μεθόδους. Το νευρωνικό δίκτυο οπισθοδιάδοσης χρησιμοποιείται για να δώσει επιβεβαίωση των άλλων μεθόδων, αλλά βρέθηκε ότι ο συνδιασμός αυτός έχει μεγαλύτερο ποσοστό επιτυχίας._

Συμπεράσματα:

Βιβλιογραφία

1. Bryson and Y.C.Ho(1969). Applied optimal control, New York, Blaisdell.
2. Werbos (1974), Beyond regression: new tools for prediction and analysis is beharioral sciences, Ph.D thesis, Harvard University.
3. Parker (1982), Learning logic. Invention report s 81-64, File 1, Office of technology licensing, Stanford University, Stanford, California.
4. Rumelhart and J.L. McClelland (1986), Parallel distributed processing, Volumes 1 and 2, MIT Press, Cambridge, Mass.