

Dipl.-Math. Marcus Pfister, Herzogenaurach

Learning Algorithms for Feed-forward Neural Networks – Design, Combination and Analysis

Reihe **10**: Informatik/
Kommunikationstechnik Nr. **435**

Contents

1	Introduction	1
1.1	Neural Networks and Large Scale Applications	1
1.2	The Problem of Learning in Neural Networks	2
1.3	Objective and Structure of this Thesis	3
2	Neural Networks	5
2.1	Biological Motivation	6
2.1.1	Neurons and the Brain	6
2.1.2	Perceptrons and Artificial Neural Networks	6
2.2	Feed-forward Neural Networks	9
2.2.1	Neural Networks as Function-Approximators	9
2.2.2	Multilayered Neural Networks	12
2.2.3	Layered Neural Networks and Kolmogorov's Theorem	13
2.3	Definitions and Notations	15
2.4	Learning in Neural Networks	16
2.4.1	Estimating the Output Error in Feed-forward Networks	17
2.4.2	The Backpropagation Algorithm	18
2.4.3	Second Order Backpropagation	23
3	Learning Algorithms and Special Hardware	26
3.1	Parallel Hardware for Neural Networks	27
3.1.1	The SPERT Node	27
3.1.2	The SYNAPSE Computer	28

3.1.3	The CNAPS Neurocomputer	30
3.1.4	The Local Network: Also a Supercomputer	32
3.2	Mapping Backpropagation Networks to SIMD Arrays	32
3.2.1	The Feed-forward Computation	33
3.2.2	The Backpropagation Computation	34
3.2.3	The SIMD Computations: A small Example	35
3.2.4	The Second Order Backpropagation Computation	37
3.3	Performing Special Computations	37
3.3.1	The Evaluation of the Sigmoid Function	38
3.3.2	Dividing Integers	38
3.4	Implementation Details	48
3.4.1	Integer Scaling	48
3.4.2	Preparation of the Input Data	49
3.4.3	The Sigmoid Lookup Table	51
3.4.4	Mapping Feed-forward Networks to the CNAPS	51
3.4.5	Maximum Network Sizes for the CNAPS	52
3.4.6	Speed-up of the CNAPS over Sequential Computers	53
3.5	Conclusions	54
4	Fast Learning Algorithms for Neural Networks	55
4.1	Why Backpropagation-Accelerations?	56
4.2	Backpropagation-Variations: A Classification	56
4.3	Variations of the Standard Algorithm	59
4.3.1	Initializing the Network Weights	59
4.3.2	When should the learning Process be stopped?	60
4.3.3	Improving and Estimating Generalization Errors	61
4.3.4	More than one Pattern to learn: Batch vs. On-Line	64
4.3.5	Introduction of a Momentum Term	65
4.3.6	Using bipolar- instead of binary Vectors	66
4.3.7	Handling flat Spots of the Error Function	67

4.3.8	Escaping local Minima	69
4.3.9	Handling Redundancy in the Training Set	70
4.3.10	Decorrelation of the Training Set	71
4.4	Adaptive Step Algorithms	77
4.4.1	The Adaptation of one global Learning Rate	77
4.4.2	Local Adaptation of individual Learning Rates	82
4.5	Second Order Methods	86
4.5.1	Quasi-Newton Methods	88
4.5.2	Secant Methods	89
4.6	The Method of Conjugate Gradients	94
4.6.1	Original Motivation of the Conjugate Gradient Method	94
4.6.2	Solving nonquadratic Problems with the CG-Method	96
4.6.3	The Method of Fletcher and Reeves	96
4.6.4	The Scaled Conjugate Gradient (SCG) Algorithm	97
4.7	Alternative Approaches	100
4.7.1	The Cascade Correlation Algorithm	100
4.7.2	Relaxation Methods	102
4.7.3	Nonlinear Least Squares Methods	103
4.8	Evaluation of Backpropagation Variations	104
4.8.1	Evaluation of Adaptive Step and Second Order Methods	105
4.8.2	Evaluation of the Standard Variations	107
4.9	Conclusions	110
5	Dynamic Combination of Learning Strategies	112
5.1	Critique of the Common Backpropagation Accelerations	113
5.2	Hybrid Learning Algorithms	116
5.2.1	QRprop	117
5.2.2	DERprop	121
5.3	Related Work	124
5.3.1	Muted (Quasi-) Newton Methods	125

5.3.2	Combination of Steepest Descent and Newton's Method	126
5.3.3	Extended Quickprop	127
5.3.4	Finding the Global Minimum of the Error Function	127
5.4	Conclusions	128
6	Parallel Implementations	129
6.1	Comparing Learning Algorithms	130
6.2	Benchmarks for Learning Algorithms	131
6.2.1	Artificial Data Sets	131
6.2.2	Problems Arising from Real World Applications	132
6.3	Details of the Algorithms compared	136
6.4	Integer- vs. Floating-point Arithmetic	139
6.5	Runtime Comparison	143
6.5.1	Effect of the chosen Standard Variations	144
6.5.2	Results for the Sonar Signals Discrimination Problem	147
6.5.3	Results for the English Vowel Recognition Problem	147
6.5.4	Results for NETtalk	149
6.5.5	Results for the Protein Recognition Problem	150
6.5.6	Results for the Handprinted Digits Recognition Problem	151
6.6	Discussion of the Results	151
6.6.1	The Computational Complexity of the Algorithms	157
6.7	Conclusions	159
6.8	Results of the Runtime Comparison (Tables)	161
7	Convergence Properties of Hybrid Algorithms	164
7.1	An idealized Model of the Error Function	164
7.2	Convergence Properties of Rprop _{1D}	167
7.3	Convergence Properties of QRprop _{1D}	171
7.4	Convergence Properties of DERprop _{1D}	178
7.5	An <i>N</i> -Dimensional Convergence Proof	182
7.6	Conclusions	188

8	Conclusions	189
8.1	Outlook	191
A	Description of the Sequential Experiments	192
A.1	The Benchmarks	192
A.2	Tables for the Sequential Experiments	194