Redes Neurais Artificiais: Revisão da Literatura

Victor São Paulo Ruela
Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
Email: victorspruela@ufmg.br

***Resumo*— Este trabalho tem como objetivo apresentar uma revisão da literatura de redes neurais artificiais, com enfoque nas evoluções desenvolvidas a partir dos principais trabalhos clássicos que estabeleceram os fundamentos desta enorme área de pesquisa.**

I. Introdução

A Rede Neural Artificial (RNA) é uma classe de modelos muito popular em problemas de classificação, reconhecimento de padrões, regressão e predição [1]. Inspirado pelas características do cérebro humano, elas possuem como elementos básicos neurônios artificiais capazes de executar operações matemáticas, representando desta forma modelos de neurônios biológicos. Através de sua organização em diferentes estruturas de rede, tais modelos são capazes de se adaptar e representar funções matemáticas bastante complexas.

Diferentes representações estão presentes na literatura, as quais são classificadas de acordo com o seu nível de complexidade e requisitos coputacionais de implementação. Hipóteses básicas para regras de aprendizado de associações entre neurônios podem ser encontradas em trabalhos bastante antigos, como abordado no livro de William James em 1982 [2]. Entretanto, um grande marco desta área de pesquisa ocorreu na década de 40 após a introdução do modelo de McCulloch and Pitts (MCP) [3], o qual é a adotado atualmente nos principais modelos de RNAs.

O modelo MCP tem como saída a soma das ativações dos neurônios anteriores ponderados pelos pesos das conexões entre eles. Originalmente, uma função de ativação do tipo degrau é aplicada sobre sua saída, configurando modelo de soma-e-limiar originalmente descrito pelos autores. Este trabalho apresentou a configuração de diversas redes de neurônios MCP, com enfoque na implementação de funções lógicas. Vale a pena notar que os primeiros computadores digitais estavam surgindo nesta época, motivando esta aplicação. Entretanto, as estruturas apresentadas eram estáticas e não houve a sugestão de algum método de aprendizado para adaptá-las.

O aprendizado surgiu de forma mais concreta com o postulado de Hebb [4], originalmente publicado em 1949. De acordo com o autor, a eficiência de uma determinada sinapse que conecta dois neurônios é proporcional à co-ocorrência de ativação entre eles. Portanto, o princípio de aprendizado Hebbiano visa reforçar as conexões relevantes para as diferentes saída da rede, guiado pela correlação entre os neurônios. Considerando o neurônimo MPC, as primeiras estruturas de rede e algoritmos de treinamento descritos na literatura são o *Adaline* [5], em 1960, e o Perceptron simples, em 1957 [6].

Após um período de euforia com a introdução destes últimos dois modelos, a área de pesquisa de RNAs sofreu um descrédito e frustação até o início dos anos 80. Isso decorreu do trabalho de Minsky e Papert [7], o qual generalizou as limitações destes modelos para problemas considerados fundamentais, como o do OU-exclusivo (XOR). O interesse só foi reativado após o re-descobrimento do algoritmo *back-propagation* para treinamento de redes de múltiplas camadas [8], as quais são capazes de superar as limitações até então existentes das redes de camada única. Destacam-se também a introdução dos mapas de Kohonen [9] e redes recorrentes de Hopfield [10] para aprendizado não-supervisionado. Além disso, nesta época surgiram as primeiras conferências e periódicos dedicados à area de RNAs [12].

A partir destes princípios elementares, a área de RNAs evoluiu bastante nas últimas décadas. Após a introdução das primeiras regras de aprendizado, este tipo de modelo ganhou maior visibilidade e aplicabilidade para problemas reais, sendo possível encontrar uma enorme quantidade de aplicações publicadas [13]. Além disso, o aumento dos recursos computacionais disponíveis fomentou o desenvolvimento de novas técnicas para aprendizado e o aprimoramento das existentes, além de propostas de novas estruturas redes complexas capazes de lidar com problemas de grande dificuldade.

Portanto, o objetivo deste trabalho é apresentar a uma revisão da literatura contendo os principais trabalhos e entender a evolução dos diferentes modelos de redes neurais utilizadas atualmente. Partindo das referências clássicas, diferentes abordagens propostas serão analisadas de forma cronológica com o intuito de se entender a evolução desta área de pesquisa até o tempo presente. Este trabalho será divido da seguinte forma: as Seções II e III apresentam uma revisão da literatura com uma análise crítica dos principais trabalhos, para técnicas de aprendizada supervisionando e não-supervisionado, respectivamente. Finalmente, é feita uma conclusão deste trabalho.

II. Aprendizado supervisionado

A. Perceptron

Proposto inicialmente por Rosenblatt [6], este é um modelo geralmente utilizado para a solução de problemas de classificação lineares. No seu trabalho original, o autor descreve formas de adaptação dos parâmetros, ou pesos, da rede com o objetivo de reduzir a discrepância entre as saídas esperadas e

estimadas e aprender associações entre os neurônios, o que é a base da indução para diversos algoritmos atuais. Este trabalho é considerado um marco na literatura por diversos autores. Embora descrito como uma rede de duas camadas, originalmente seu treinamento só considerava uma camada. Por esse motivo, o Perceptron simples é comumente descrito na forma de somente um neurônio MCP. Sua regra de aprendizado é bem direta e consiste em alterar iterativamente os pesos da rede adicionado o erro total entre as saídas medidas estimadas ponderada pelo vetor de entradas.

Se considerarmos uma função de ativação contínua e diferenciável, os pesos da rede poderão ser inferidos de forma explícita, através do cálculo da pseudo-inversa, ou pelo algoritmo do gradiente descendente [14]. Exemplos de funções de ativação com esta característica frequentemente empregadas na literatura são a função logística, tangente hiperbólica e linear [1]. Vale a pena ressaltar que a covergência destas abordagens está condicionada aos dados utilizados para treinamento serem linearmente independentes [14].

Rosenblatt provou a convergência da regra de aprendizado original, porém a mesma só é garantida para problemas linearmente separáveis, o que constitui a principal limitação deste modelo. O trabalho de Minsky e Papert [7] evidenciou bastante esta limitação e através da aplicação do Perceptron a diversos problemas considerados fundamentais, levou ao descrédito deste modelo pela comunidade científica. Após este trabalho, Rosenblatt avaliou diferentes arquiteturas de rede tentando superar esta limitação, mas não conseguiu chegar ao desenvolvimento do aprendizado para múltiplas camadas. Por conta disso, o Perceptron foi pouco estudado pelos próximos de 20 anos [14].

O interesse pelo Perceptron retornou na década de 80 com a descrição do método de aprendizado conhecido como *back-propagation*, o qual é capaz de aprender os pesos de redes de múltiplas camadas de forma eficiente [8]. Aliado a isso, o Perceptron de múltiplas camadas é capaz de descrever superfícies de separação não-lineares, superando a pricipal limitação do trabalho de Rosenblatt. Uma descrição mais completa desta família de modelos é feita na seção II-C.

B. Adaline

O Adaline foi inicialmente desenvolvido por Widrow em 1960 [5], sendo principamente aplicado em problemas de regressão lineares. Assim como o Perceptron, originalmente este modelo considera somente um neurônio MCP em sua formulação, entretando sua função de ativação é a identidade. Seu treinamento é formulado como um problema de otimização com custo quadrático, onde originalmente foi utilizado o algoritmo do gradiente descendente para sua solução.

Para este agoritmo, em cada iteração é dado um passo na direção oposta ao gradiente da função objetivo, resultanto em uma convergência gradual para o mínimo do problema. Este gradiente pode ser calculado de forma análitica para a estrutura de rede do Adaline, o qual é no fim proporcional à diferença entre os valores estimados e reais [5], bastante similar ao Perceptron simples. É fácil notar que o treinamento também pode ser realizado de forma direta através do cálculo da pseudo-inversa dos dados de entrada, já que este é um problema de mínimos quadrados [15].

Uma extensão proposta deste modelo é conhecida como Madaline, o qual é caracterizada por uma rede composta por vários Adalines. Existem duas principais regras para seu treinamento, conhecidas por MRI [16] and MRII [17]. É interessante notar que a MRI surgiu bem antes do algoritmo *back-propagation*, podendo ser considerada uma estrutura primitiva de uma rede de múltiplas camadas. Os leitores são referidos à [18] para uma descrição mais completa destas regras e suas aplicações.

C. Perceptron de múltiplas camadas

O Perceptron de múltiplas camadas (MLP) é uma rede neural com uma ou mais camadas escondidas, ou seja, localizadas entre as entradas e saídas do modelo. Além disso, são caracterizadas por um alto grau de conectividade e por aplicar funções de ativação não-lineares e diferenciáveis ao modelo dos neurônios [15]. Estas camadas adicionais funcionam como detectores de características, aplicando transformações não-lineares sequenciais aos dados de forma que estes sejam mais facilmente separados nesse novo espaço. Portanto, a introdução das camadas escondidas permite modelar superfícies de decisão não-lineares, sendo consideradas aproximadores universais de funções se a função de ativação é contínua, limitada e não-constante [19].

O treinamento de redes de camada única vistas nas seções anteriores é bem direto pois podemos facilmente derivá-las analiticamente. Entretando, ao incluir camadas escondidas e alta conectividade, analisar teoricamente o comportamento do Perceptron torna-se mais difícil. Aliado a isso, o seu treinamento se torna mais complexo justamente por haver uma maior quantidade de estruturas de rede possíveis para representar os dados de entrada. O primeiro algoitmo eficiente de treinamento de tais foi formalizada por Rumelhart em 1985 [8], conhecido como *back-propagation*.

O *back-propagation* é uma técnica de aprendizado online (ou estocástica), na qual os pesos da rede são ajustados amostra-a-amostra. Ou seja, em cada época de treinamento, os dados de entrada são apresentados individualmente para a rede objetivando a minimização do erro das saídas estimadas e desejadas. O algoritmo pode ser dividido em duas etapas: na primeira os dados são apresentados à rede mantendo os pesos fixos, e calculada a sua respectiva saída; na segunda, o sinal de erro em relação à saída esperada é calculado e propagado no sentido inverso da rede, onde ajustes sucessivos são realizados. A atualização dos pesos da rede é feita com base na técnica do gradiente descendente, cuja derivação completa será omitida deste trabalho por brevidade. O leitor pode encontrá-la em [8], [15], por exemplo.

Embora bastante eficiente, sua convergência pode ser lenta se o algoritmo não for usado corretamente [20]. É recomendado realizar a normalização das entradas para equalizar a taxa de atualização dos pesos entre as camadas, além de remover variávies altamente correlacionadas [21]. Ainda de acordo

com [21], é sugerido o uso de sigmoides simétricas, como a tangente hiperbólica, as quais geralmente possuem maior velocidade de convergência. Além disso, é importante que os valores desejados estejam dentro dos limites da sigmoide escolhida. Outro fator importante é a taxa de treinamento: [22] apresenta o estudo de algumas técnicas para adaptação da taxa de treinamento presentes na literatura, mostrando que o seu uso é bastante benéfico.

É interessante notar que os problemas descritos anteriormente são similiares aos encontrados para a otimização de funções não-lineares e não-convexas. Ou seja, é possível analisar o treinamento do MLP como um problema de otimização e aplicar diferentes algoritmos e heurísticas disponíveis da literatura. De fato, isso é explorado por diversos autores, os quais aplicam o método de Newton [23], gradiente conjugado [24], Gauss-Newton [25], Levenberg-Marquardt [26] e Quasi-Newton [27] para o treinamento do MLP.

Métodos de segunda ordem possuem o atrativo da convergência acelarada, mas em contra-partida é necessária a estimativa da Hessiana, o que exige mais recursos computacionais e está sujeito a problemas numéricos adicionais, sendo limitado a redes pequenas e a usar aprendizado por batelada. Portanto, este são fatores que devem ser levados em conta na hora de escolher o otimizador para realizar o treinamento de uma rede. Conforme argumentado por [21], o uso de informações de segunda ordem nem sempre é necessário em alguns problemas, para os quais a técnica do gradiente estocástico bem ajustada é dificilmente superado para problemas de larga escala.

O estado da arte para treinamento de RNAs são o algoritmo Adagrad [28], Adadelta [29] e o Adam [30], publicados nessa ordem. O Adagrad icorpora um ajuste automático da taxa de aprendizado, ao invés de mantê-las fixas como o *back-propagation* original, eliminando a necessidade de ajuste deste parâmetro. Este algoritmo possui como limitação um decaimento muito acentuado da taxa de aprendizado, compromentendo sua eficiência. O Adadelta é uma extensão do Adagrad que amemiza este problema pela restrição da janela de acumulação dos gradientes passados. O Adam é considerado um dos algoritmos mais robustos e eficientes [31], o qual aumenta a robustez do algoritmo Adadelta pela introdução de um decaimento da média dos gradientes passados.

Mais recentementes, alguns autores começaram a propor o uso de algoritmos evolucionários (EAs) para o treinamento do MLP. Eles são atrativos pelo fato de convergirem para o ótimo global se um tempo de treinamento suficiente estiver disponível. Entretanto, o custo computacional poderá ser bastante alto, além do ajuste dos hiper-parâmetros ser trabalhoso. Em [32], os autores propõe o uso do algotimo Differential Evolution, o qual concluem que ele não apresenta desempenho superior ao *back-propagation*. Já [33] propôs uma técnica híbrida de algoritmos genéticos e backpropagation, a qual se mostrou menos suscetível a ficar presa em mínimos locais durante o treinamento. O leitor pode ser referir a [34] para um estudo mais completo de EAs e RNAs.

D. Redes de funções de base radial (RBF)

As redes RBF foram inicialmente introduzidas por [35] e são caracterizadas por um aprendizado que envolve duas etapas: (i) aplicar uma transformação aos padrões para um espaço onde a probabilidade de serem linearmente separáveis é alta (ii) encontrar os pesos usando o estimador mínimos quadrados usado no Perceptron simples. Essa estrutura pode ser representada por um rede de três camadas, onde sua camada escondida é reponsável pela transformação não-linear das estradas para o novo espaço, geralmente para uma dimensão muito alta.

Essa transformação é justificada pelo teorema de Cover sobre a separabilidade de padrões [36], o qual diz que um problema de classificação complexo projetado não-linearmente para um espaço de alta dimensão é mais provável de ser separável do que em um espaço de baixa dimensão, desde que o espaço não seja densamente povoado. Boa parte da teoria, que é relacionada ao campo de interpolação multivariável, considera um kernel baseado na função Gaussiana, que é uma classe importante de RBFs. Teoricamente, as redes RBF podem ser consideradas um aproximador universal de funções contínuas se a RBF é selecionada apropriadamente [37], [38], [39]. Uma condição apropriada é dada pelo teorema da interpolação de Micchelli [40], ou quando uma determinada classe de RBFs é contínua e diferenciável [38], por exemplo.

No seu treinamento, além de tratamentos especiais na presença de ruídos nos dados [41], uma etapa importante é a estimativa dos parâmetros das unidades Gaussianas, além dos pesos da rede. Isso pode ser feito distribuindo os centros uniformemente, por exemplo. De forma mais geral, pode-se selecionar aleatoriamente subconjutos dos padrões de entrada se este são representativos e grandes o suficiente para o apredizado [42]. Para aproximação de funções, uma heurística possível é colocar os centros nos extremos da derivada de segundo grau da função e também mais densamente em áreas cujo valor absoluto e maior [43], o que possui resultados melhores em relação à distribuição uniforme.

Outra abordagem bastante utilizada para a definição dos centros é o uso de técnicas de agrupamento (*clustering*) dos dados [44], [45]. Uma escolha popular são o algoritmo não-supervisionado *k-means* [46] e redes neurais baseadas em memórias associativas [47]. Também podem ser utilizadas técnicas de clustering supervisionadas [48], as quais podem ser mais eficientes para redes RBF [42]. Após determinados os centros, define-se as matrizes de covariância das RBFs como a covariância dos dados de cada *cluster* [42].

Com os centros e covariâncias definidos, o aprendizado dos pesos é feito pela minimização do erro quadrático médio. Para problemas simples, pode-se utilizar o algoritmo de mínimos quadrados ou gradiente descendente, similar ao aprendizado do Perceptron simples. O treinamento via gradiente descendente é equivalente ao do MLP [49], onde pode-se também utilizar as mesmas abordagens por otimização irrestrita [44]. Uma descrição de diferentes abordagens para treinamento de RBFs por otimizaçao podem ser encontradas em [42], como o uso

de *back-propagation* seletivo [50] e programação linear [51].

No caso dos mínimos quadrados, problemas complexos que requerem um número elevado de RBFs na camada escondida estão sujeitos a problemas numéricos na estimativa da pseudo-inversa. Para lidar com isso, uma abordagem comum é o uso de técnicas de ortogonalização, como as decomposições SVD e QR [52]. Em algumas condições, é possível também resolver esse problema usando a transformada de Fourier da rede RBF, a qual requer menos esforço computacional [53]. Outra alternativa eficiente é aplicar a ortogonalização de Gram-Schimidt (GSO) à RBF [54], a qual possui menor requisito de armazenamento e pode sem implementado de forma paralela.

Outra forma eficiente para aprendizado da rede consiste no uso da técnica de mínimos quadrados ortogonal (OLS) [55], [56], [57], [58]. Nesta abordagem, não só os pesos podem ser determinados, mas também a quantidade e centros das RBFs. O GSO é inicialmente aplicado para contruir um conjunto de vetores ortogonais no espaço criado pela unidade escondida, e na sequência um novo centro RBF é selecionado de acordo com a sua minimização do erro quadrático médio de treinamento. Esse algotimo pode ser executado de forma sequencial ou reversa, ou seja, podemos iniciar com uma rede vazia ou com o máximo de unidades escondidas, por exemplo. Existem também versões recursivas deste algoritmo na literatura, conhecidas por ROLS [59], as quais são aplicadas a problemas envolvendo sistemas de múltiplas entradas e saídas, por exemplo.

E. Máquinas de aprendizado extremo

Inicialmente proposto por [68], as máquinas de aprendizado extremo (ELM) são redes neurais com uma única camada escondida, as quais possuem o atrativo de poucos parâmetros a serem ajustados, generalização maior ou similar e redução do tempo de treinamento das redes em relação aos métodos convencionais. Seu treinamento é baseado na teoria de minimização de risco empírico, necessitando de somente uma iteração para este processo, evitando múltiplas iterações e algoritmos de otimização local [69].

ELMs são capazes de definir adaptivamente o número neurônios da rede e aleatoriamente escolher os pesos das entradas e viéses da camada escondida [70]. Isso faz com o que a rede possa ser considerada como um sistema linear, o qual pode ser calculado de forma analítica através de uma operação de inversão da matriz de pesos da camada de saídas [70]. Essa característica permite uma drástica redução do esforço computacional do treinamento, geralmente de 10 vezes ou mais [71].

Apesar do apelo no ganho de tempo de treinamento, essa abordagem permite menor adaptabilidade ao conjunto de dados, além de problemas numéricos para o cálculo da inversão por mínimos quadrados e problemas de robustez se os dados forem ruidosos [69]. Além disso, a aleatoriedade na escolha de pesos e viéses pode levar a uma quantidade maior de neurônios na camada escondida, bem como tornar o sistema linear não solucionável e reduzir sua acurácia [72].

Embora tenha sido proposto bem recentemente (2004), têm recebido uma atenção considerável na literatura a nível téorico e de aplicações [73]. Do ponto de vista da teoria, o foco dos autores é divido entre duas frentes: (1) aprimorar sua capacidade universal de aproximação com camadas escondidas aleatórias e (2) propor novas técnicas de aprendizado mais rápidas e robustas. De acordo com uma revisão feita em 2011 por Huang et. al [73], destacam-se o ELM baseado em kernel, com aprendizado online, incremental, *ensemble* de ELM e ELM com poda.

F. Generalização

Uma das suas principais características do modelo RNA é sua habilidade de generalização, que é sua capacidade em estimar corretamente dados ainda não vistos. Em outras palavras, algoritmos de aprendizado de RNA possuem como objetivo encontrar um modelo que consiga caputar as principais características do conjunto de dados de treinamento, mas que também seja capaz de prever com precisão um conjunto de teste ainda não visto.

Este conceito é introduzido pela primeira vez por Geman et. al em 1992 [60], o qual é o bastante conhecido dilema viés-variância. Através deste trabalho, os autores introduzem o conceito de que existe uma competição entre duas fontes de erro que impede o modelo de generalizar além dos dados de treinamento, o viés e a variância. Além disso, eles mostra que existe um compromisso entre estas grandezas, de forma que um modelo com viés máximo possuirá mínima variância, e vice-versa.

O erro de viés está relacionado a suposições incorretas sobre o modelo, podendo ser defindo como a diferença entre o valor previsto médio e o esperado. Um alto viés leva ao fenômeno de *under-fitting*, onde o modelo possuei uma estrutura simples e não é capaz de representar os principais padrões presentes nos dados. Já o erro de variância indica a sensibilidade do modelo a flutuações no conjunto de treinamento. Uma variância elevada leva ao fenômeno de *over-fitting*, onde o modelo é bastante complexo e se ajusta perfeitamente aos dados de treinamento. Um exemplo deste comportamente pode ser visto na Figura 1. É importante notar que estes dois erros podem contribuir para um desempnho ruim, ou seja, mesmo um modelo não-enviésado com variância alta pode resultar em um erro de treinamento maior que o esperado [60].

Existem algumas formas de se balancear o viés e variância, as são divididas no caso da conjuntos amostras finitas e asintóticamente infinitas. Por serem menos comuns em aplicações práticas, as últimas serão omitidas deste trabalho e o autor pode se referir a [60] para mais detalhes. Para conjuntos de dados finitos, destacam-se as técnicas de regularização [62]. A regularização é uma metodologia proposta por Tikhonov [63] em 1963 para o tratamento de problemas mal-condicionados. Esta técnica sugere que o objetivo de treinamento deva ser minimizar uma combinação convexa entre o erro de aproximação e um termo que expressa a complexidade do modelo.

A regressão Ridge [64] utiliza o conceito de Tikhonov, acrescentando uma regra que penaliza a norma ℓ-2 dos pesos

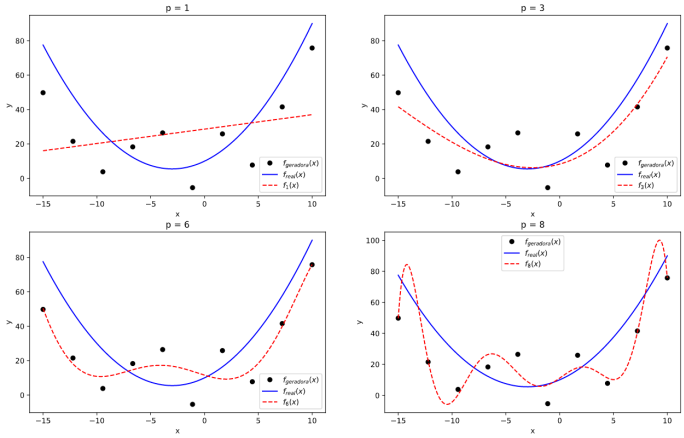

Fig. 1. Exemplo de *under-fitting* ($p = 1$) e *over-fitting* ($p > 3$) para problema de regressão polinomial sobre dados uma função quadrática.

ao treinamento. Outra possibilidade é a regressaão Lasso, [65] cuja única diferença está em penalizar utilizado a normal ℓ-1 dos pesos. Esta é caracteriza por obter soluções onde os pesos encontrados são mais esparsos, sendo útil para, pro exemplo, aplicar alguma forma de seleção de atributos. Outra técnica bastante utilizada é conhecida por *Elastic Net*, a qual realiza uma combinação das duas abordagens apresentadas anteriormente [66]. Uma dificuldade destas técnicas consiste no ajuste dos parâmetros que controlam a penalização, a qual é geralmente tratada pela técnica da curva-L [67] ou validação cruzada [61].

O problema de generalização deu origem a técnicas de aprendizado que incluem em sua formulação formas de se balancear o viés e a variância automaticamente, geralmente via regularização. Entre eles, destacam-se as máquinas de vetores suporte (SVM) e o aprendizado multiobjetivo. SVMs incorporam uma forma de regularização em seu treinamento, enquanto que no aprendizado multiobjetivo é proposta uma nova metodologia para obtenção de soluções balanceadas, inspirada na teoria de otimização.

G. Máquinas de vetores suporte

Introduzido por Vapnik em 1992 [74] as máquinas de vetores suporte (SVM) são considerados um dos algoritmos mais robustos e poderosos para aprendizado supervisionado até os dias atuais. No trabalho de Vapnik, o autor apresenta de forma bem completa os fundamentos teóricos deste modelo, apresentando justificativas para sua ótima capacidade de generalização, bem como os limites para a sua validade. Seu princípio de treinamento está na maximização da margem entre os padrões de treino e a superfície de decisão, que é uma representação convexa do compromisso entre viés e variância. Estas propriedades são uma consequência do uso dos chamados vetores de suporte no seu treinamento, que são os sub-conjuntos de dados mais próximos da superfície de decisão, ou seja, os mais difícies de classificar e relevantes para sua otimalidade. O problema então consiste em encontrar o hiperplano que maximiza a margem de separação entre os padrões, que é equivalente a minimizar a norma Euclidiana do seu vetor de pesos [15].

O problema de encontrar o vetor de pesos que define o hiper-plano foi convenientemente formulado com base na teoria de otimização convexa. Isso é muito importante pois esta classe de problemas tem sua otimalidade garantida e bem definida, não estando sujeita aos problemas observados no algoritmo do gradiente descendente, por exemplo. A função objetivo é descrita como a norma Euclideana dos pesos acrescida de um segundo termo que limita a quantidade de erros de treinamento, já que os dados podem não ser perfeitamente serparáveis por um hiper-plano. Este problema é então transformado utilizando a técnica de multiplicadores de Lagrange, a qual faz com que seja descrito em função somente dos dados de treinamento.

O segundo termo pode ser ajustado pelo usuário por meio de uma constante, a qual irá controlar o nível de regularização aplicada ao problema, o que resulta em sua característica de generalização. Um valor alto poderá resultar em *over-fitting* se os dados forem ruidosos, o que sugere cuidado na sua escolha. Sua definição pode ser feita experimentalmente usando técnicas de validação cruzada [61], por exemplo.

Através da discussão anterior, é fácil notar que o SVM é aplicável somente a problemas linearmente separáveis se são for realizada nenhuma transformação não-linear aos dados. Portanto, um conceito importante utilizado por este algoritmo são os *kernels* [75]. A partir da escolha de um kernel apropriado [76], [77], o problema poderá ser resolvido por um modelo linear. Note que isto é bem similiar ao princípio de funcionamento das redes RBF. Em relação ao MLP, ele possui a vantagem de controlar a complexidade independente de sua dimensionalidade, ou seja, é possível ter uma quantidade muito grade de neurônios na camada escondida, conforme o teorema de Cover [36], e ainda assim obter uma boa generalização [78].

Embora muito poderoso, esse algoritmo é caracterizado por um elevado custo computacional de sua implementação [79]. Como sua complexidade pode ser proporcional ao quadrado da quantidade de amostras de treinamento, sua aplicação em problemas de grande porte pode se tornar proibitiva. Outra limitação está relacionada à solução do problema de otimização, uma vez que solvers disponíveis possuem suas próprias limitação em relação ao número de variáveis de decisão suportadas. O problema de otimização é ainda mais dificultado pela esparsidade da solução do SVM, pois o vetor de pesos a ser encontrado possuirá poucos elementos não-nulos [15], resultando em problemas númericos que impedirão sua correta solução. Diversas melhorias foram propostas para melhorar esta situação, as quais são classificadas entre seleção de dados [80], decomposição [81], geometria [82], implementação paralela [83] e heurísticas [84].

Outra limitação deste algoritmo está relacionada a dados não-balanceados. Técnicas propostas para lidar com essa situação realizam o balanceamento dos dados antes do treinamento, ou modificam a estrutura do modelo para torná-los menos sensitivos. As técnicas mais comuns são o *under* e *over-sampling*, e o SMOTE [85]. Esta última é mais indicada para

SVMs [86]. É importante ressaltar também que o SVM foi originalmente formulado problemas de classificação binária, de forma que este precisou de ser posteriormente extendido para ser aplicação à problemas com mais de uma classe [87], [88].

Em [89], os autores apresentam um revisão bem completa da literatura de SVMs para classificação. Embora utilizado com sucesso em diversas aplicaçõos reais, o SVM ainda possui pouca adoção para problemas com grandes volumes de dados, bem como baixo desempenho para dados não-balanceados. Como tendências de trabalhos futuros, os autores citam além de melhorias nestes problemas clássicos, avanços em treinamento on-line, seleção automática de kernel e parâmetros com menor custo, além de aplicações para aprendizado semi-supervisionado.

H. Aprendizado multi-objetivo

Na seção II-F, ficou evidente pela discussão que o treinamento de RNAs é um problema multi-objetivo, no qual precisamos encontrar uma solução de compromisso entre o viés e a variância. Portanto, em um dos extremos teremos um conjunto de pesos que resultam em um viés máximo (*overfitting*) e no outro variância máxima (*underfitting*). A partir desta observação, alguns autores comeaçaram a investigar a formulação do treinamento de MLPs como um problema de otimização multi-objetivo, com o intuito de melhorar a generalização. Esta abordagem foi inicialmente proposta por [90].

Para a solução de um problema multi-objetivo, o primeiro passo consiste na obtenção do conjunto pareto ótimo, o qual conterá todas as soluções eficientes. Em seguida, é necessária uma estratégia de decisão para escolher a solução mais apropriada deste conjunto pareto ótimo [91]. Embora complementares, estas duas tarefas originam de teorias distintas: na primeira, estamos interessados em como amostrar de forma eficiente o conjunto pareto ótimo [92]; na segunda, o objetivo é a tomada de decisão propriamente dita, dado o conjunto pareto anterior [93].

Em [90], os autores formalizam o problema considerando dois objetivos: minimização da norma dos pesos e do erro quadrático médio de treinamento. A técnica proposta busca uma solução com boa generalização utilizando uma variação da abordagem ϵ-restrito para obtenção do conjunto pareto ótimo. Embora tenha apresentado resultados similares ao SVM, a técnica proposta tem um grande custo computacional associado, uma vez que é necessário estimar muitos pontos do conjunto pareto ótimo, o que pode ser tornar proibitivo computacionalmente para um volume de dados de treinamento grande.

A partir do trabalho de [90], diferentes melhorias foram propostas para consolidar esta abordagem. Em [94], foi proposta uma variação do algoritmo de *Levenberg–Marquardt* restrito pela complexidade da RNA. O autor de [95] propôs o uso do algoritmo da seção áurea como forma de reduzir o custo computacional da estimativa da fronteira pareto. Já em [96], [97], são propostas abordagens eficientes para treinamento de redes RBF multiobjetivo. Recentemente, [98] propôs representar a rede neural de forma esférica, a qual foi capaz de obter desempenho superior em relação às técnicas clássicas.

Os trabalhos anteriores possuem maior enfoque na estimativa da fronteira pareto. Do ponto de vista da etapa tomada de decisão, destacam-se o trabalho recente de [99], onde é proposta uma nova técnica para a tomada de decisão em problemas de classificação binária. Os autores propõe o uso do conhecimento prévio sobre o problema representado pela medida de imprecisão do processo de obtenção dos dados, ao invés de se basear em um conjunto de validação. O método se mostra eficiente por deixar o processo de decisão independente de uma amostra dos dados, além de poder considerar todo o conjunto de dados para o treinamento. Isso evita problemas de generalização do modelo por uma escolha indevida do conjunto de treino e teste, o que também é importante para problemas com poucos dados disponíveis.

III. Aprendizado não-supervisionado

A. Aprendizado Hebbiano

Baseado no postulado de Hebb [4], o aprendizado Hebbiano pode ser definido como uma forma de plasticidade (habilidade de adaptação a novos dados) sináptica dependente de atitividade, na qual a eficência da conexão entre dois neurônios é reforçada à co-ocorrência de ativação, ou correlação. Esta teoria é traduzida em linguagem matemática para o aprendizado de RNAs através da regra de Hebb, que diz que o ajuste dos pesos é proporcional ao produto entre entradas e saídas da rede [100]. Estudos mostram que regras de aprendizado baseados nesse princípio resultam em uma evolução dos campos neuronais recepivos e mapas organizados topologicamente [101]. Uma rede Hebbiana representa uma estrutura básica de memórias associativas, que é um dos princípios básicos de modelos mais gerais, como as redes de Hopfield [10].

O ajuste do pesos pela regra de Hebb clássica pode ser matematicamente resolvido pela pseudoinversa do produto entrada-saida da rede, o que a caracteriza como uma regra baseada em correlação. Uma limitação conhecida deste princípio está no fato de que os pesos irão crescer ilimitadamente, uma vez que ele será incrementado em todas as atualizações [102]. Esse fenômeno é conhecido como inferência cruzada, ou *crosstalk*, o qual causa o deslocamento das estimativas da rede em relação ao esperado. O *crosstalk* será nulo somente quando os dados de entrada forem ortogonais, o que nem sempre pode ser garantido para dados reais. Ele é inerente ao aprendizado Hebbiano, e irá depender de como os padrões estão espacialmente relacionados entre si, sendo possível de ser controlada, geralmente por técnicas de limites adaptativos [103] ou normalização de pesos [104].

O tamanho do conjunto de dados também possui influência sobre o *crosstalk*, o qual pode aumentar a probabilidade do seu crescimento ilimitado, pelo acréscimo de mais interferências presentes na rede. Além disso, teremos também um aumento na quantidade de produtos entre entradas-saídas, contribuindo

para este mesmo efeito. Portanto, pode-se perceber que memórias associativas baseadas na regra de Hebb possuem capacidade de armazenamento limitada [1]. Estudos teóricos destes limites são geralmente encontrados em modelos baseados na regra de Hebb, como para as redes de Hopfield, cujo número máximo de padrões armazenados não ultrapassará 15% da quantidade de dados [105].

Uma das primeiras descobertas de aplicação de aprendizado Hebbiano foi feita por Oja [106]. Neste trabalho, o autor mostra que aplicando um fator de normalização à regra de Hebb, os pesos irão convergir para a primeira componente principal das entradas. Este foi um resultado muito interessante, uma vez que mostra correspondência direta com o método estatístico de análise de componentes principais (PCA). Este resultado foi posteriomente generalizado por Sanger [107], onde o autor evolui a regra de Oja de forma a conseguir estimar todas as componentes principais das entradas. Em [108], podem ser vistos diferentes contextos em que os princípios de apredizado Hebbiano podem ser aplicados.

As redes de Hopfield [10] são caracterizadas por uma arquitetura de rede recorrente, na qual as saídas de um neurônio são re-alimentadas na própria rede. Seu treinamento é basedo em uma função de energia, a qual sempre irá decrescer e atingir um ponto de energia mínima à medida em que o estado da rede evolui de acordo com uma dinâmica pré-estabelecida. Este tipo de RNA é muito popular em memórias associativas e para a solução de problemas de otimização combinatória, especialamente o problema do caixeiro viajante [109], [110]. Entretanto, o seu uso é bastante dificultado pela necessidade de ajuste de diversos parâmetros da rede, como a forma de se reescrever o problema de otimização através da função de energia. Entretanto, esta é uma tarefa difícil de ser realizada na prática, gerando soluções infactíveis com muita frequência. Solucionar de forma eficiente um problema combinatório com redes de Hopfield ainda é um trabalho em aberto na literatura [111].

B. Mapas auto-organizáveis

Diferente do aprendizado Hebbiano, onde múltiplos nós da rede podem estar ativas ao mesmo tempo, o aprendizado competitivo introduz o conceito de competição para que somente um nó esteja ativo em um determinado momento. Este comportamento resulta no agrupamento (*clustering*) dos dados de entrada, obtido automaticamente baseado nas correlações existentes. Este princípio deu origem ao algoritmo VQ (*vector quantization*), bastante utilizado para processamento de sinais [112].

O algoritmo VQ é capaz de particionar o espaço de entradas em um número finito de regiões, representadas de forma ótima por um único vetor modelo, tipicamente construídos de acordo com a distância euclidiana. Se o conjunto de dados for finito, o VQ se torna equivalente ao algoritmo k-médias [113]. Em 1982, Kohonen [9] propôs uma nova forma de projeção não-linear, chamada de mapas auto-organizáveis (SOM), as quais são capazes de produzir um mapeamento que remete ao VQ, porém com a característica importante de preservar a topologia do espaço de entradas, de forma que a localização física dos nós no mapa representam a similaridade relativa entre os pontos no espaço original.

A estrutura de SOMs é tipicamente composta de uma ou duas camadas entrelaçadas de neurônios, formando uma malha de duas dimensões. Esta malha é o mapa que se auto-organiza em cada iteração em função das entradas. O mapa é construído de forma que modelos similares estarão associados a nós mais próximos na malha, enquanto modelos menos similiares estarão situados gradualmente mais longes. Em cada iteração, o nó com a menor distância em relação à entrada apresentada é o vencedor, o qual irá ajustar o seu peso e o da sua vizinhança [114].

Kohonen [9] originalmente propôs o ajuste de forma similar ao gradiente desecente, o qual é computacionalmente eficiente, produz bons resultados mesmo em alta dimensionalidade e têm sua convergência avaliada teoricamente [115]. Entretanto, tal estratégia requer um ajuste de uma quantidade muito grande de parâmetros para se ter uma convergência satisfatória, de forma que é sugerido pela literatura o uso de sua versão de treinamento em batelada para uma melhor experiência de uso desta RNA [113].

Uma coleção dos principais avanços de SOMs podem ser encontradas em [116], [117], [118]. Embora um pouco defasadas, elas foram escritas de forma bem interessante: modelos SOM foram treinados sobre uma lista de trabalhos publicas e posteriormente categorizados de acordo com a relevância ao tema. Este sistema é conhecido por WEBSOM [119], sendo bastante utilizada para mineração de texto. Anualmente é realizado um Workshop discutindo os principais avanços desenvolvidos nessa área de pesquisa, chamado WSOM. O leitor pode encontrar as discussões de 2019 em [120], por exemplo.

IV. Conclusões

Neste trabalho foi feita uma revisão bibligráfica de alguns dos principais trabalhos sobre redes neurais aritificiais. Realizando a divisão entre modelos para aprendizado supervisionado e não-supersvisionado, os conceitos básicos dos modelos estudados foram apresentados para contextualização, bem como uma breve análise das principais evoluções e aplicações propostas na literatura. Cada um destes modelos possui uma enorme quantidade de trabalhos publicados, portanto é de se esperar que publicações importantes tenham sido omitidos.

Um aspecto não muito abordado neste trabalho foram as aplicações de RNAs, dado que o foco deste trabalho foi em entender um pouco mais de sua teoria. Em [13] está disponível uma lista das diferentes áreas em que RNAs são comumente aplicadas. Conforme observado durante a realização deste trabalho, grande partes das evoluções visam aprimorar eficiência dos algoritmos treinamento. Isso também é observado por [13], o qual considera uma tendência trabalhos futuros visando aprimorar este aspecto.

Referências

[1] Anil K Jain, Jianchang Mao, and K Moidin Mohiuddin. Artificial neural networks: A tutorial. *Computer*, 29(3):31–44, 1996.

[2] William James. *Psychology, briefer course*, volume 14. Harvard University Press, 1984.

[3] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[4] Donald Olding Hebb. *The organization of behavior: a neuropsychological theory*. J. Wiley; Chapman & Hall, 1949.

[5] Bernard Widrow and Marcian E Hoff. Adaptive switching circuits. Technical report, Stanford Univ Ca Stanford Electronics Labs, 1960.

[6] Frank Rosenblatt. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[7] Marvin Minsky and Seymour Papert. An introduction to computational geometry. *Cambridge tiass., HIT*, 1969.

[8] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[9] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

[10] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[11] Gail A Carpenter and Stephen Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing*, 37(1):54–115, 1987.

[12] Bohdan Macukow. Neural networks–state of art, brief history, basic models and architecture. In *IFIP international conference on computer information systems and industrial management*, pages 3–14. Springer, 2016.

[13] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 4(11):e00938, 2018.

[14] John Hertz, Anders Krogh, Richard G Palmer, and Heinz Horner. Introduction to the theory of neural computation. *PhT*, 44(12):70, 1991.

[15] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice-Hall, Inc., 2007.

[16] Bernard Widrow. Generalization and information storage in network of adaline'neurons'. *Self-organizing systems-1962*, pages 435–462, 1962.

[17] Capt Rodney Winter and B Widrow. Madaline rule ii: A training algorithm for neural networks. In *Secondf Annual International Conference on Neural Networks*, pages 1–401, 1988.

[18] Bernard Widrow and Michael A Lehr. 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, 78(9):1415–1442, 1990.

[19] Kurt Hornik, Maxwell Stinchcombe, Halbert White, et al. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[20] Yann LeCun. Efficient learning and second-order methods. *A tutorial at NIPS*, 93:61, 1993.

[21] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[22] George D. Magoulas, Michael N. Vrahatis, and George S Androulakis. Improving the convergence of the backpropagation algorithm using learning rate adaptation methods. *Neural Computation*, 11(7):1769–1796, 1999.

[23] Sue Becker, Yann Le Cun, et al. Improving the convergence of back-propagation learning with second order methods. In *Proceedings of the 1988 connectionist models summer school*, pages 29–37, 1988.

[24] Erik M Johansson, Farid U Dowla, and Dennis M Goodman. Backpropagation learning for multilayer feed-forward neural networks using the conjugate gradient method. *International Journal of Neural Systems*, 2(04):291–301, 1991.

[25] Roberto Battiti. First-and second-order methods for learning: between steepest descent and newton's method. *Neural computation*, 4(2):141–166, 1992.

[26] Martin T Hagan and Mohammad B Menhaj. Training feedforward networks with the marquardt algorithm. *IEEE transactions on Neural Networks*, 5(6):989–993, 1994.

[27] B Robitaille, B Marcos, M Veillette, and G Payre. Modified quasi-newton methods for training neural networks. *Computers & chemical engineering*, 20(9):1133–1140, 1996.

[28] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[29] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[31] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[32] Jarmo Ilonen, Joni-Kristian Kamarainen, and Jouni Lampinen. Differential evolution training algorithm for feed-forward neural networks. *Neural Processing Letters*, 17(1):93–105, 2003.

[33] Shifei Ding, Chunyang Su, and Junzhao Yu. An optimizing bp neural network algorithm based on genetic algorithm. *Artificial intelligence review*, 36(2):153–162, 2011.

[34] Seyedali Mirjalili. Evolutionary algorithms and neural networks. *Studies in Computational Intelligence*, 2019.

[35] D. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Syst.*, 2, 1988.

[36] Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334, 1965.

[37] Tomaso Poggio and Federico Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.

[38] Jooyoung Park and Irwin W Sandberg. Universal approximation using radial-basis-function networks. *Neural computation*, 3(2):246–257, 1991.

[39] Yi Liao, Shu-Cherng Fang, and Henry LW Nuttle. Relaxed conditions for radial-basis function networks to be universal approximators. *Neural Networks*, 16(7):1019–1028, 2003.

[40] Charles A Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive approximation*, 2(1):11–22, 1986.

[41] David S Broomhead and David Lowe. Radial basis functions, multivariable functional interpolation and adaptive networks. Technical report, Royal Signals and Radar Establishment Malvern (United Kingdom), 1988.

[42] Yue Wu, Hui Wang, Biaobiao Zhang, and K-L Du. Using radial basis function networks for function approximation and classification. *ISRN Applied Mathematics*, 2012, 2012.

[43] V David Sánchez A. Second derivative dependent placement of rbf centers. *Neurocomputing*, 7(3):311–317, 1995.

[44] Ke-Lin Du and Madisetti NS Swamy. *Neural networks in a softcomputing framework*. Springer Science & Business Media, 2006.

[45] K-L Du. Clustering: A neural network approach. *Neural networks*, 23(1):89–107, 2010.

[46] John Moody and Christian J Darken. Fast learning in networks of locally-tuned processing units. *Neural computation*, 1(2):281–294, 1989.

[47] Teuvo Kohonen. *Self-organization and associative memory*, volume 8. Springer Science & Business Media, 2012.

[48] C-L Chen, W-C Chen, and F-Y Chang. Hybrid learning algorithm for gaussian potential function networks. In *IEE Proceedings D (Control Theory and Applications)*, volume 140, pages 442–448. IET, 1993.

[49] Dietrich Wettschereck and Thomas Dietterich. Improving the performance of radial basis function networks by learning center locations. In *Advances in neural information processing systems*, pages 1133–1140, 1992.

[50] Mohammad-Taghi Vakil-Baghmisheh and Nikola Pavešić. Training rbf networks with selective backpropagation. *Neurocomputing*, 62:39–64, 2004.

[51] Asim Roy, Sandeep Govil, and Raymond Miranda. An algorithm to generate radial basis function (rbf)-like nets for classification problems. *Neural networks*, 8(2):179–201, 1995.

[52] Gene H Golub and Charles F Van Loan. An analysis of the total least squares problem. *SIAM journal on numerical analysis*, 17(6):883–893, 1980.

[53] Yoshinori Abe and Y Figuni. Fast computation of rbf coefficients for regularly sampled inputs. *Electronics Letters*, 39(6):543–544, 2003.

[54] Wladyslaw Kaminski and Pawel Strumillo. Kernel orthonormalization in radial basis function neural networks. *IEEE Transactions on Neural Networks*, 8(5):1177–1183, 1997.

[55] Sheng Chen, Colin FN Cowan, and Peter M Grant. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on neural networks*, 2(2):302–309, 1991.

[56] S Chen, PM Grant, and CFN Cowan. Orthogonal least-squares algorithm for training multioutput radial basis function networks. In *IEE Proceedings F (Radar and Signal Processing)*, volume 139, pages 378–384. IET, 1992.

[57] S Chen and J Wigger. Fast orthogonal least squares algorithm for efficient subset model selection. *IEEE Transactions on Signal Processing*, 43(7):1713–1715, 1995.

[58] Xia Hong and SA Billings. Givens rotation based fast backward elimination algorithm for rbf neural network pruning. *IEE Proceedings-Control Theory and Applications*, 144(5):381–384, 1997.

[59] DL Yu, JB Gomm, and D Williams. A recursive orthogonal least squares algorithm for training rbf networks. *Neural Processing Letters*, 5(3):167–176, 1997.

[60] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

[61] Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

[62] Federico Girosi, Michael Jones, and Tomaso Poggio. Regularization theory and neural networks architectures. *Neural computation*, 7(2):219–269, 1995.

[63] Andrei Nikolaevich Tikhonov. On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk*, volume 151, pages 501–504. Russian Academy of Sciences, 1963.

[64] Arthur E Hoerl and Robert W Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.

[65] Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

[66] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

[67] Per Christian Hansen and Dianne Prost O'Leary. The use of the l-curve in the regularization of discrete ill-posed problems. *SIAM journal on scientific computing*, 14(6):1487–1503, 1993.

[68] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: a new learning scheme of feedforward neural networks. In *2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541)*, volume 2, pages 985–990. IEEE, 2004.

[69] Shifei Ding, Han Zhao, Yanan Zhang, Xinzheng Xu, and Ru Nie. Extreme learning machine: algorithm, theory and applications. *Artificial Intelligence Review*, 44(1):103–115, 2015.

[70] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[71] Wan-Yu Deng, Qing-Hua Zheng, Lin Chen, and Xue-Bin Xu. Research on extreme learning of neural networks. *Chinese Journal of Computers*, 33(2):279–287, 2010.

[72] Yuguang Wang, Feilong Cao, and Yubo Yuan. A study on effectiveness of extreme learning machine. *Neurocomputing*, 74(16):2483–2490, 2011.

[73] Guang-Bin Huang, Dian Hui Wang, and Yuan Lan. Extreme learning machines: a survey. *International journal of machine learning and cybernetics*, 2(2):107–122, 2011.

[74] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[75] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, illustrated edition edition, 2004.

[76] James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.

[77] Richard Courant and David Hilbert. *Methods of Mathematical Physics*, volume 1. Wiley, New York, 1989.

[78] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series, 2018.

[79] Léon Bottou and Chih-Jen Lin. Support vector machine solvers. *Large scale kernel machines*, 3(1):301–320, 2007.

[80] Jigang Wang, Predrag Neskovic, and Leon N Cooper. Training data selection for support vector machines. In *International Conference on Natural Computation*, pages 554–564. Springer, 2005.

[81] Jian-xiong Dong, Adam Krzyzak, and Ching Y Suen. Fast svm training algorithm with decomposition on very large data sets. *IEEE transactions on pattern analysis and machine intelligence*, 27(4):603–618, 2005.

[82] Zhi-Qiang Zeng, Hua-Rong Xu, Yan-Qi Xie, and Ji Gao. A geometric approach to train svm on very large data sets. In *2008 3rd International Conference on Intelligent System and Knowledge Engineering*, volume 1, pages 991–996. IEEE, 2008.

[83] Hans Graf, Eric Cosatto, Leon Bottou, Igor Dourdanovic, and Vladimir Vapnik. Parallel support vector machines: The cascade svm. *Advances in neural information processing systems*, 17:521–528, 2004.

[84] Dennis DeCoste and Kiri Wagstaff. Alpha seeding for support vector machines. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 345–349, 2000.

[85] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[86] Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery*, pages 107–119. Springer, 2003.

[87] Jason Weston and Chris Watkins. Multi-class support vector machines. Technical report, Citeseer, 1998.

[88] Yi Liu and Yuan F Zheng. One-against-all multi-class svm classification using reliability measures. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 849–854. IEEE, 2005.

[89] Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, and Asdrubal Lopez. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215, 2020.

[90] Roselito de Albuquerque Teixeira, Antônio Pádua Braga, Ricardo HC Takahashi, and Rodney R Saldanha. Improving generalization of mlps with multi-objective optimization. *Neurocomputing*, 35(1-4):189–194, 2000.

[91] Kalyanmoy Deb. Multi-objective optimization. In *Search methodologies*, pages 403–449. Springer, 2014.

[92] Yann Collette and Patrick Siarry. *Multiobjective optimization: principles and case studies*. Springer Science & Business Media, 2004.

[93] RO Parreiras and JA Vasconcelos. Decision making in multiobjective optimization problems. *Real-world multi-objective system engineering*, pages 29–52, 2005.

[94] Marcelo Azevedo Costa, Antônio de Pádua Braga, and Benjamin Rodrigues de Menezes. Improving generalization of mlps with sliding mode control and the levenberg–marquardt algorithm. *Neurocomputing*, 70(7-9):1342–1347, 2007.

[95] Roselito A Teixeira, Antônio P Braga, Rodney R Saldanha, Ricardo HC Takahashi, and Talles H Medeiros. The usage of golden section in calculating the efficient solution in artificial neural networks training by multi-objective optimization. In *International Conference on Artificial Neural Networks*, pages 289–298. Springer, 2007.

[96] Illya Kokshenev and Antonio Padua Braga. An efficient multi-objective learning algorithm for rbf neural network. *Neurocomputing*, 73(16-18):2799–2808, 2010.

[97] Gladston JP Moreira, Elizabeth F Wanner, Frederico G Guimarães, Luiz H Duczmal, and Ricardo HC Takahashi. Lmi formulation for multiobjective learning in radial basis function neural networks. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2010.

[98] Honovan P Rocha, Marcelo A Costa, and Antônio P Braga. Training multi-layer perceptron with multi-objective optimization and spherical weights representation. In *Proceedings of the European symposium on neural networks*, pages 131–136, 2015.

[99] Talles Henrique de Medeiros, Honovan Paz Rocha, Frank Sill Torres, Ricardo Hiroshi Caldeira Takahashi, and Antônio Pádua Braga. Multi-objective decision in machine learning. *Journal of Control, Automation and Electrical Systems*, 28(2):217–227, 2017.

[100] Wulfram Gerstner and Werner M Kistler. Mathematical formulations of hebbian learning. *Biological cybernetics*, 87(5-6):404–415, 2002.

[101] Richard Kempter, Wulfram Gerstner, and J Leo Van Hemmen. Hebbian learning and spiking neurons. *Physical Review E*, 59(4):4498, 1999.
[102] Yoonsuck Choe. *Hebbian Learning*, pages 1–5. Springer New York, New York, NY, 2013.
[103] Elie L Bienenstock, Leon N Cooper, and Paul W Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2(1):32–48, 1982.
[104] Kenneth D Miller and David JC MacKay. The role of constraints in hebbian learning. *Neural computation*, 6(1):100–126, 1994.
[105] ROBERTJ McEliece, Edwardc Posner, EUGENER Rodemich, and SANTOSHS Venkatesh. The capacity of the hopfield associative memory. *IEEE transactions on Information Theory*, 33(4):461–482, 1987.
[106] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.
[107] Terence D Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks*, 2(6):459–473, 1989.
[108] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.
[109] John J Hopfield and David W Tank. "neural" computation of decisions in optimization problems. *Biological cybernetics*, 52(3):141–152, 1985.
[110] Kate A Smith. Neural networks for combinatorial optimization: a review of more than a decade of research. *INFORMS Journal on Computing*, 11(1):15–34, 1999.
[111] Huaguang Zhang, Zhanshan Wang, and Derong Liu. A comprehensive review of stability analysis of continuous-time recurrent neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 25(7):1229–1262, 2014.
[112] Stanley C Ahalt, Ashok K Krishnamurthy, Prakoon Chen, and Douglas E Melton. Competitive learning algorithms for vector quantization. *Neural networks*, 3(3):277–290, 1990.
[113] Teuvo Kohonen. Essentials of the self-organizing map. *Neural networks*, 37:52–65, 2013.
[114] Melody Y Kiang. Extending the kohonen self-organizing map networks for clustering analysis. *Computational Statistics & Data Analysis*, 38(2):161–180, 2001.
[115] Marie Cottrell, Jean-Claude Fort, and Gilles Pagès. Theoretical aspects of the som algorithm. *Neurocomputing*, 21(1-3):119–138, 1998.
[116] Samuel Kaski, Jari Kangas, and Teuvo Kohonen. Bibliography of self-organizing map (som) papers: 1981–1997. *Neural computing surveys*, 1(3&4):1–176, 1998.
[117] Merja Oja, Samuel Kaski, and Teuvo Kohonen. Bibliography of self-organizing map (som) papers: 1998–2001 addendum. *Neural computing surveys*, 3(1):1–156, 2003.
[118] Matti Pöllä, Timo Honkela, and Teuvo Kohonen. Bibliography of self-organizing map (som) papers: 2002-2005 addendum. *Neural Computing Surveys*, 2009.
[119] Samuel Kaski, Timo Honkela, Krista Lagus, and Teuvo Kohonen. Websom–self-organizing maps of document collections. *Neurocomputing*, 21(1-3):101–117, 1998.
[120] Alfredo Vellido, Karina Gibert, Cecilio Angulo, and José David Martín Guerrero. Advances in self-organizing maps, learning vector quantization, clustering and data visualization. In *Conference proceedings WSOM*, page 6. Springer, 2019.