Learning and Nonlinear Models (L&NLM) – Vol. 10, Iss. 2, pp. 119–126, 2012.

IDENTIFICAÇÃO E PREVISÃO DE SÉRIES TEMPORAIS UTILIZANDO LS-SVM OTIMIZADO PELO ALGORITMO DE CARDUMES

Leonardo Trigueiro Dos Santos
Departamento de Engenharia Elétrica, Universidade Federal do Paraná
Pós-Graduação em Engenharia Elétrica, Av. Cel. Francisco H. dos Santos, 210
Centro Politécnico, CEP 81530-970, Curitiba, PR, Brasil
leo.trigueiro@hotmail.com

Edgar Leite dos Santos Filho
Grupo Produtrônica, Pós-Graduação em Engenharia de Produção e Sistemas (PPGEPS)
Pontifícia Universidade Católica do Paraná (PUCPR)
Rua Imaculada Conceição, 1150, CEP 80215-901, Curitiba, PR, Brasil
edgarlsfilho@gmail.com

Leandro dos Santos Coelho
Departamento de Engenharia Elétrica, Universidade Federal do Paraná
Pós-Graduação em Engenharia Elétrica, Av. Cel. Francisco H. dos Santos, 210
Centro Politécnico, CEP 81530-970, Curitiba, PR, Brasil

Grupo Produtrônica, Pós-Graduação em Engenharia de Produção e Sistemas (PPGEPS)
Pontifícia Universidade Católica do Paraná (PUCPR)
Rua Imaculada Conceição, 1150, CEP 80215-901, Curitiba, PR, Brasil
leandro.coelho@pucpr.br

Resumo – A máquina de vetor suporte (SVM) é uma técnica relativamente recente. A SVM tem se mostrado muito eficiente quando aplicada à identificação e previsão de séries temporais, um importante problema no campo da engenharia. Uma variante deste método, a máquina de vetor suporte à mínimos quadrados (LS-SVM) possui as mesmas características básicas de sua predecessora e possui a vantagem de ser mais adequada ao processamento computacional. A fim de refinar o processo de identificação realizado pela LS-SVM o algoritmo de otimização por cardumes (FSS) foi escolhido dado suas características de adequação a problemas de difícil delimitação e alta dimensionalidade do espaço de busca, como no presente artigo. Os resultados das simulações baseados no uso combinado do LS-SVM com o FSS são promissores em termos de precisão e custo computacional quando aplicados ao índice EPEA/ESALQ (Centro de Estudos Avançados em Economia Aplicada/Escola Superior de Agricultura Luiz Queiroz) da soja.

Palavras-chave – Dança da chuva, jogar sal nas nuvens, incêndio em florestas, modelo não-linear, adaptação e aprendizagem.

Abstract – Support vector machine (SVM) is a relatively recent technique. SVM has been shown very powerful when applied to the time series forecasting, important problems in engineering field. A variant of this method, the least square support vector machine (LS-SVM) has the same basics characteristics of the SVM and has the advantage of being more suitable for computational processing. To improve the identification task by the LS-SVM approach, the optimization algorithm called Fish School Search (FSS) was chosen by adaptation characteristics of high dimensionality search spaces in this paper. Simulation results based on LS-SVM combined with FSS are promising in terms of accuracy and computational cost when applied to CEPEA/ESALQ index of soy.

Keywords – forecasting, time series, support vector machine, fish school search.

1 Introdução

A máquina de vetor suporte (SVM) foi originalmente idealizada para aplicações de classificação de dados binária, onde um hiperplano ótimo separador divide o conjunto de dados em dois maximizando a margem de separação dos conjuntos. Para tal utiliza o conceito de Funções de Núcleos ($Kernels$). Posteriormente foi extrapolada a utilização da máquina de vetor suporte para a regressão não linear, onde através do artifício de aumento da dimensionalidade do espaço, torna os dados linearmente separáveis facilitando a delimitação de uma função que melhor realiza a separação dos dados de entrada em seu espaço original.

[1] Uma das principais vantagens da SVM é a grande capacidade de generalização da função de regressão obtida pelo método, isto ocorre pois a influência da complexidade da hipótese é dada pela margem de separação dos dados. [2] Outro grande diferencial da máquina de vetor suporte é a consideração da minimização do risco estrutural e não somente do risco empírico na construção da função de regressão. [3] Sendo o risco estrutural vinculado a capacidade de generalização da função de regressão e o risco empírico vinculado a adequação desta função de regressão aos dados de entrada. [4] A base biológica na qual os algoritmos de inteligência de enxames foram baseados é uma característica comportamental dos enxames, proveniente de um mecanismo natural denominado auto-organização [5], onde pequenas regras determinam o comportamento e a iteração entre o grupo que, quando seguidas por todos os indivíduos da colônia, possibilita ao enxame a resolução de problemas que um único indivíduo não seria capaz de realizar. [6] Para o caso específico dos algoritmos de cardumes esse problema é a busca por comida. A contribuição do presente é a avaliação do uso de uma abordagem evolutiva, bio-inspirada no comportamento de cardumes, com características propíscias a aplicação em problemas de difícil delimitação e de alta dimensionalidade na otimização da função de regressão. [7] O presente artigo visa demonstrar a utilização conjunta destasduas ferramentas na identificação de séries temporais, aqui representada por uma série financeira histórica doi preço da soja no Paraná. O restante do artigo está organizado da seguinte forma. Na seção 2 são descritos os fundamentos da LS-SVM e do algoritmo de cardumes. Na seção 3 a aplicação utilizando a série temporal e na seção 4 a conclusão é apresentada.

2 Métodologia

Nesta seção serão detalhados os conceitos da máquina de vetor suporte e do algoritmo de otimização por cardumes.

2.1 LS-SVM

Como alternativa ao uso da SVM original muitas outras abordagens foram desenvolvidas (citar algumas ouitras formas do SVM), e dentre estas formulações a Máquina de Vetor Suporte à Mínimos Quadrados (LS-SVM) que foi idealizada substituíndo as restrições baseadas em inequações da SVM original por restrições de igualdade e tendo como função objetivo a soma do erro médio quadrático [8]. Com isto foi possível manter as características da SVM original e reduzir o custo computacional para realização das operações do método. [3] A Máquina de vetor suporte aproxima a relação entre a saída e a entrada pela equação

$$y = \omega * \phi(x) + b \tag{1}$$

onde b é um limite escalar ($threshold$), ω um coeficiente de ponderação e $\phi(x)$ uma não linearidade mapeada a partir da entrada. Os coeficientes ω e b serão estimados pelo algoritmo de otimização tendo como função objetivo a minimização da função de risco J, dada por

Minimizar

$$J = \frac{1}{2} \parallel W \parallel^2 + \gamma \frac{1}{2} \sum_{i=1}^{N} \epsilon(y_i, f(x_i)) \tag{2}$$

tal que,

$$\epsilon(y_i, f(x_i)) = \begin{cases} 0, \parallel y_i, f(x_i) \parallel \leq \epsilon \\ \parallel y_i, f(x_i) - \epsilon \parallel, \text{outros} \end{cases} \tag{3}$$

onde W é o vetor de ponderações e γ o parâmetro de regularização que estabelece um equilíbrio entre a complexidade e o erro de treinamento do meodelo.

A primeira parte da equação 2 faz a normalização dos pesos, convergindo para valores menores. Este procedimento é adotado para reduzir a variação imposta no modelo por pesos demasiadamente grandes o que deteriora a capacidade de generalização do modelo acentuando o problema de sobre ajuste da função aos dados de entrada. A segunda parte da equação 2 representa os erros de regressão para o conjunto de dados de treinamento. A restrição de igualdade imposta pela equação 3 fornece a definição do erro de regressão. Quando aplicado a padrões não linearmente separáveis, são adicionadas variáveis de folga, ξ_i e $\xi^*{}_i$ sendo então possível transformar a equação 3 em uma função objetivo primal dada por

Minimizar

$$J = \frac{1}{2} \parallel W \parallel^2 + \gamma \frac{1}{2} \sum_{i=1}^{N} \epsilon(\xi_i + \xi_i^*) \tag{4}$$

sujeito a

$$y_i - W * \phi(x_i) - b \leq \epsilon + \xi_i \tag{5}$$

$$W * \phi(x_i) + b - y_i \leq \epsilon + \xi_i^* \tag{6}$$

onde $i = 1, ..., N$ e $\xi_i, \xi_i^* \geq 0$

Introduzindo-se os vetores de multiplicadores de lagrange α_i e α_i^*, denominados vetores de suporte, a função núcleo e maximizando-se a função dual da equação 4, a função de regressão dada pela equação 1 apresenta a seguinte forma explícita

$$f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) * K(x_i, x_j) + b \tag{7}$$

onde $K(x_i, x_j)$ é a função núcleo. Os vetores α_i são obtidos através da solução do sistema linear de equações, seguindo as condições de Karush-Kuhn-Tucker. O valor de $K(x_i, x_j)$ determina o produto interno de dois vetores x_i e x_j no espaço característico, $\phi(x_i)$ e $\phi(x_j)$, logo $K(x_i, x_j)$=$\phi^T(x_i)\phi(x_j)$. Utilizar a função núcleo tem como objetivo o cálculo de $\phi(x_i)$ e $\phi(x_j)$, que apresenta uma complexidade muito alta, de uma forma aproximada e mais simples. Esta função núcleo gera um mapeamento entre o espaço de entrada e o espaço de alta dimensionalidade, dito característico. O hiperplano gerado pela máquina de vetor suporte no espaço característico, quando mapeada de volta para o espaço original dos dados de entrada, torna-se uma superfície não linear. Assim sendo, o hiperplano de separação passa a ser não mais uma função linear dos vetores de entrada, mas uma função linear dos vetores do espaço característico. [1] Neste artigo a função núcleo adotada foi uma função de base radial (RBF, Radial Basis Function) que é dada por

$$K(x_i, x_j) = exp(\frac{\| x_i - x_j \|^2}{2 * \sigma^2}) \tag{8}$$

Onde σ é a largura das gaussianas do núcleo.

2.2 O algoritmo de Cardumes, FSS

O comportamento do cardume durante sua busca por comida pode ser descrita em três passos distintos [9], um primeiro movimento feito de maneira individual onde cada peixe se desloca de maneira aleatória e com movimentos de pequena amplitude, essa amplitude reduzida é devido ao fato de estarem em um cardume o que restringe sua mobilidade individual. Um segundo movimento instintivo realizado pelos peixes, na direção da maior quantidade de comida, quando percebem que outros indivíduos foram mais eficientes na busca pelo alimento no passo anterior e, por fim, o terceiro movimento, mais bem descrito como a vontade coletiva do cardume, que se dá pela continuação do deslocamento do cardume no sentido da maior quantidade de comida.

2.2.1 Alimentação

Este operador é responsável pelo cálculo do peso atual de cada peixe do cardume, através da fórmula

$$W_i(t+1) = W_i(t) + \frac{f[x_i(t+1)] - f[x_i(t)]}{MAX\| f[x_i(t+1)] - f[x_i(t)] \|} \tag{9}$$

onde $W_i(t)$ é o peso do peixe i e $f[x_i(t)]$ a adaptabilidade na função objetivo, na inspiração biológica a adaptabilidade é análoga a quantidade de comida em determinada região do aquário.

2.2.2 Natação

O operador de natação descreve matematicamente os três passos observados, descritos anteriormente [10] no comportamento do cardume durante sua locomoção, são eles. Movimento individual, A direção do movimento é aleatória, permitindo uma varredura completa do espaço debusca. Instinto coletivo,

$$x_i(t+1) = x_i(t) + \frac{\sum_{i=1}^{N} \delta x_{ind_i} \{f[x_i(t+1)] - f[x_i(t)]\}}{\sum_{i=1}^{N} \{f[x_i(t+1)\} - f[x_i(t)]\}} \tag{10}$$

Quando alguns peixes encontram uma fonte mais abundante de comida, o instinto do cardume é de mover-se na direção desta comida.

E por fim a Vontade coletiva, esse pode ser um movimento de contração ou expansão do cardume, seria um pequeno deslocamento posterior ao instinto coletivo, a intenção do cardume se deslocar na direção em que obteve mais sucesso na obtenção da comida.

Contração:

$$x_i(t+1) = x_i(t) - step_{vol} * rand * [x_i(t) - Bari(t)] \tag{11}$$

Expansão:

$$x_i(t+1) = x_i(t) + step_{vol} * rand * [x_i(t) - Bari(t)] \tag{12}$$

Sendo δx_{ind_i} a média dos movimentos individuais realizadas no passo individual. $Step_{vol}$ o tamanho do passo dado durante o movimento de expansão ou contração do cardume. $Bari$ o centro de massa, baricentro, do cardume. E $rand$ um número gerado aleatoriamente com distribuição uniforme contido no intervalo [0,1] [9].

O fluxograma de execução do algoritmo é apresentada na Figura 1

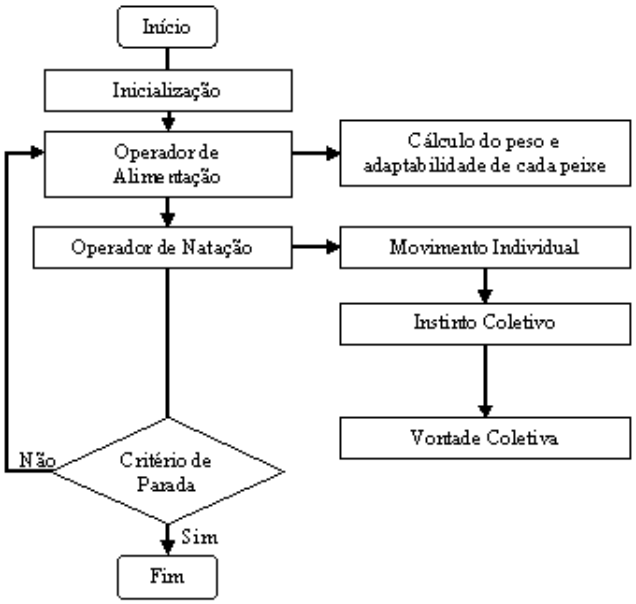

Figura 1: Fluxograma de execução do algoritmo de cardumes

3 Aplicações

3.1 Estudo de Caso

O Índice CEPEA / ESALQ da soja, anteriormente chamado ESALQ / BM & FBovespa, representa uma média ponderada dos preços no estado do Paraná para a soja em grão, tipo exportação, de acordo com os termos do Concex: até 14 por cento de umidade, até 2 por cento do impurezas, um limite máximo de 8 por cento para grãos danificados e de 30 por cento de grãos quebrados.

O Índice de soja tem sido publicado diariamente pelo Cepea desde agosto de 1997 e é usado como referência para negócios.

Embora o Índice CEPEA / ESALQ considere apenas os preços no Paraná - um importante produtor, processador e exportador de soja, a equipe do Cepea também captura os preços nos Estados do Rio Grande do Sul, Mato Grosso, Mato Grosso do Sul, Goiás, Minas Gerais e São Paulo.

3.2 Simulações

A faixa de valores factíveis aos parâmetros da LS-SVM, passada ao algoritmo de cardumes compreende uma faixa de valores entre 0.01 e 100. Os tamanhos dos passos individual e final foram alterados para demonstrar a robustez do algoritmo e a influência destes parâmetros na convergência do algoritmo. A tabela 1 apresenta os valores escolhidos para os parâmetros em cada execução, métodos propostos, do algoritmo. O algoritmo rodou um total de 30 experimentos onde cada experimento compreendia um total de 50 iterações do algoritmo de cardumes, considerando uma população de 10 indivíduos, peixes.

Tabela 1: Parâmetros do FSS

Método	LS-SVM(1)	LS-SVM(2)	LS-SVM(3)
STEPind	5	10	100
STEPindfim	0.0001	0.0001	0.0001
STEPvol	1	5	10
STEPvolfim	0.01	0.01	0.01

Tabela 2: Análise de convergência do LS-SVM usando FSS

Método	LS-SVM(1)	LS-SVM(2)	LS-SVM(3)
Custo Mínimo	0,620503	0,620502	0,620502
Custo Máximo	0,622424	0,622423	0,62244
Custo Médio	0,622237	0,621912	0,621758
Desvio Padrão de Custo	0,000562	0,000859	0,000921

O melhor conjunto de parâmetros testados foi o do segundo método proposto, pois obteve um custo mínimo inferior ao do primeiro método e um menor desvio padrão em relação ao terceiro método. A tabela 2 apresenta um custo mínimo para a otimização da SVM de 0,620502 onde os valores ótimos encontrados para os parâmetros foram; σ = 100,000000 e γ = 0,926702. Com o tempo médio de processamento para uma iteração do algoritmo de cardumes sendo de 16,7357 segundos. O algoritmo da máquina de vetor suporte, durante o processo de identificação, avalia dois fatores para determinar a função que melhor representa os dados de entrada. São eles, maximizar a aderência da função obtida aos dados iniciais da série da soja e também a adequação da função a dados desconhecidos. O valor do erro apresentado na Figura 2 é resultado dessa consideração conjunta, aumentando um pouco o erro relativo aos dados de entrada mas garantindo uma melhor confiabilidade a dados desconhecidos. O retângulo tracejado vermelho na Figura 2 ressalta a região de maior discrepância, com um maior erro absoluto, entre a série real e a estimada pela SVM.

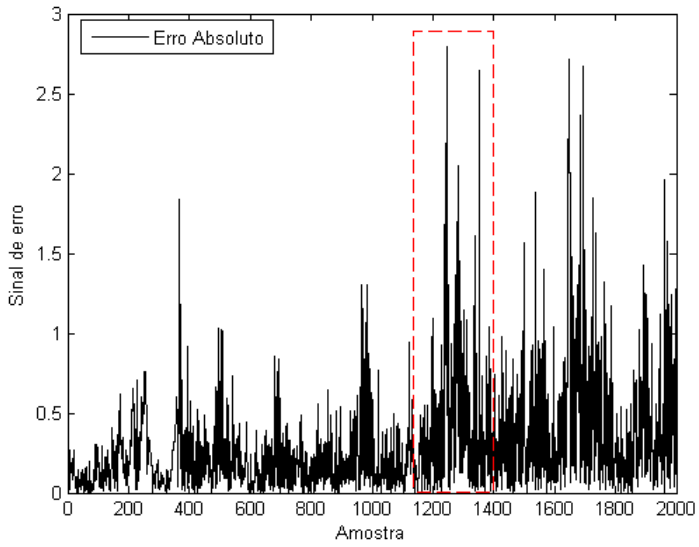

Figura 2: Erro Absoluto entre a saída real e estimada

Analisando os gráficos conjuntos da saída real e estimada da série fica clara a capacidade do algoritmo de identificação de mapear todas as características da série, o que possibilita uma aproximação mais que satisfatória dos dados de entrada pela máquina de vetor suporte. A região onde o erro absoluto apresentou maior discrepância na Figura 3 foi representada por um retângulo tracejado. A Figura 4 representa uma ampliação da maior discrepância entre as séries reais e estimadas, onde podemos analisar a influência do erro na função de saída estimada pelo SVM. Com a magnitude do erro não ultrapassando três em nenhuma amostra, a característica mais perceptível é um atraso da saída estimada em relação a saída real, em torno de uma amostra em atraso na região de maior discrepância entre as funções.

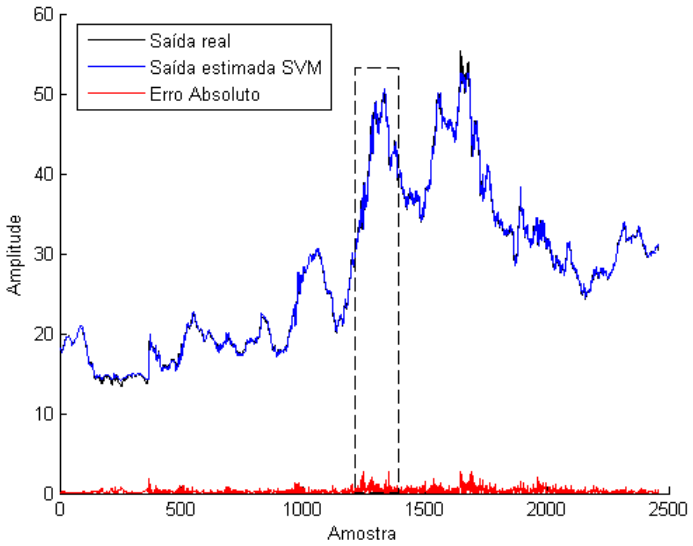

Figura 3: Erro Absoluto entre a saída real e estimada

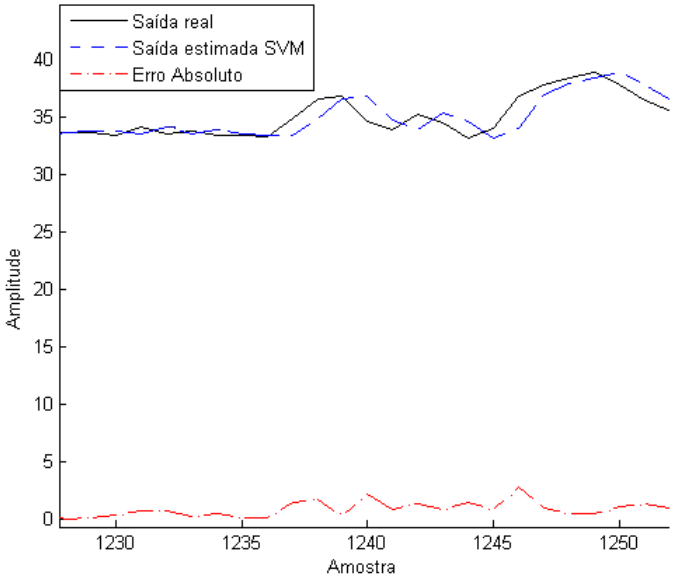

Figura 4: Erro Absoluto entre a saída real e estimada

4 Conclusão

O presente artigo tem como intuito demonstrar o uso das técnicas de otimização e identificação em conjunto para a obtenção de resultados mais confiáveis na previsão de séries temporais.

- O algoritmo de cardumes foi utilizado pelas suas características de adequação a problemas de difícil delimitação do espaço de busca e de alta dimensionalidade;
- A máquina de vetor suporte à mínimos quadrados é um algoritmo de identificação versátil e com um custo computacional baixo o que faz dele uma ótima opção para ser usada em conjunto com os algoritmos de otimização, uma vez que durante a otimização o processo de identificação será executado por várias vezes o que poderia aumentar demasiadamente o custo computacional ou até mesmo tornar inviável o uso das técnicas em conjunto;
- Outra característica da Máquina de Vetor Suporte a ser salientada é o fato de que durante o processo de identificação o erro considerado para a escolha da função que melhor representa os dados de entrada é uma ponderação dos erros Empírico e de Estrutural, ou seja, a melhor função será aquela que além de ser aderente aos dados de entrada também garantirá um erro mínimo na generalização da função obtida aos dados desconhecidos;
- O uso conjunto das técnicas de otimização e identificação possibilita um ajuste mais preciso dos parâmetros do algoritmo de identificação a um problema específico o que resultará em uma previsão mais confiável;

Para demonstrar a robustez do algoritmo de cardumes e a influência da variação dos parâmetros do FSS na velocidade e precisão da convergência do algoritmo de otimização, três configurações distintas do FSS foram testadas com os parâmetros variando de acordo com a tabela 1.

Uma análise dos dados apresentados na tabela 2 de convergência do algoritmo permite afirmar que o algoritmo de cardumes atingiu as expectativas e conseguiu com uma boa precisão e robustez nos três testes realizados, i.e. baixo desvio padrão ao otimizar os parâmetros da LS-SVM.

Quanto a influência do tamanho do passo na convergência do algoritmo de cardume, nota-se que até determinado ponto o aumento do passo permite que o algoritmo se aproxime da região ótima com mais rapidez, com isso consegue um resultado mais preciso, contudo esse aumento no passo também ocasiona um espalhamento maior do cardume o que aumenta o desvio padrão mas não de forma significativa.

Outro ponto importante na análise é o custo computacional reduzido deste conjunto atuante, uma iteração do algoritmo de cardumes teve um tempo de processamento médio de 16,7357 segundos. Sendo que para cada iteração do algoritmo de cardumes cada indivíduo, peixe, foi testado na função objetivo, LS-SVM, pelo menos três vezes a cada iteração. Com isso tem-se que se foram utilizados dez indivíduos na população são executados ao menos 30 testes na função de identificação além da execução dos passos internos do algoritmo de cardumes durante este tempo de processamento.

Um próximo passo na utilização conjunta da máquina de vetor suporte e algoritmos de otimização como o FSS seria a implementação de um método recursivo para que os parâmetros da LS-SVM fossem estimados em tempo real juntamente com a identificação de uma nova função que contemple os novos dados de entrada da última aquisição realizada.

Referências

[1] O. Bousquet, S. Boucheron, stephane.boucheron and G. Lugosi. *Introduction to Statistical Learning Theory*. Springer Berlin Heidelberg, 2004.

[2] D. Boswell. "Introduction to Support Vector Machines". Technical report, California Institute of Technology, 2002.

[3] J. Suykens, J. V. Gestel, T. D. Brabanter, J. D. Moor and B. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, 2002.

[4] T. Van Gestel, J. Suykens, D.-E. Baestaens, A. Lambrechts, G. Lanckriet, B. Vandaele, B. De Moor and J. Vandewalle. "Financial time series prediction using least squares support vector machines within the evidence framework". *Neural Networks IEEE Transactions on*, vol. 12, no. 4, pp. 809–821, July 2001.

[5] D. J. Futuyma. *Evolutionary Biology*. Sinauer Associates, 1997.

[6] R. Eberhart. *Swarm Intelligence*. Morgan Kaufmann, 2001.

[7] X. Chen, J. Wang, D. Sun and J. Liang. "Time Series Forecasting Based on Novel Support Vector Machine Using Artificial Fish Swarm Algorithm". *International Conference on Natural Computation*, vol. 2, pp. 206–211, 2008.

[8] A. Gilsberts, G. Metta and L. Rothkrantz. *Evolutionary Optimization of Least-Squares Support Vector Machines*. Springer US, 2009.

[9] L. A. N. A. Filho C., Neto F. de Lima and L. M. "A novel search algorithm based on fish school behavior". In *In Systems Man and Cybernetics 2008 SMC 2008 IEEE International Conference on*, pp. 2646–2651, 2008.

[10] S. M. P. M. Bastos Filho C., Lima Neto F. and M. S. "On the influence of the swimming operators in the Fish School Search algorithm". In *In Systems Man and Cybernetics 2009. SMC 2009 IEEE International Conference on*, pp. 5012–5017, 2009.