Classificação semi-supervisionada de dados utilizando competição e cooperação entre partículas.

Rafael de Souza Ferreira, Fabricio Aparecido Breve, Unesp - Rio Claro, Instituto de Geociências e Ciências Exatas, Ciências da Computação, rsferreira11@gmail.com, FAPESP.

Palavras Chave: *Aprendizado de Máquina, Inteligência Artificial, Classificação de Base de Dados Numéricas.*

Introdução

Neste trabalho foi realizada a classificação semi-supervisionada de bases de dados numéricas utilizando um modelo de aprendizado de máquina baseado em competição e cooperação entre partículas[1]. O modelo foi estudado, implementado e aplicado em algumas bases de dados, como parte de um projeto para estendê-lo para posterior aplicação em fluxos de dados. O modelo utiliza algumas amostras pré-rotuladas para extrair conhecimento e classificar os demais dados da base. Para tanto, ele constrói uma rede a partir dos dados originais, gerando um grafo em que cada nó se conecta aos seus *k* vizinhos mais próximos. Após agrupar os dados, cada nó rotulado recebe uma partícula que atacará nós vizinhos para tentar rotulá-los com a classe que ela representa. Após certo número de ataques o programa verifica todos os nós e classifica as amostras correspondentes na base de dados.

Material e Métodos

Para classificação das bases de dados, foi utilizado um método de Aprendizado de Máquina semi-supervisionado que utiliza competição e cooperação entre partículas em uma rede[1]. O modelo envolve diversos conceitos matemáticos e computacionais, como distância Euclidiana, grafos, manipulação de matrizes e vetores, etc. Para avaliar os resultados foi preciso extrair alguns dados com o uso de métodos estatísticos. Para a implementação do programa foi utilizada a Linguagem C, compilada com o GCC (GNU Compiler Collection) no sistema operacional Linux Ubuntu. Para a ordenação de dados foi utilizado o método BubbleSort. O modelo foi aplicado às bases de dados "Wine" e "Iris" retiradas do Repositório de base de dados UCI[2].

Resultados e Discussão

Os resultados dos testes são muito satisfatórios, chegando a 98% de acerto em base de dados nas bases testadas, conforme mostram as Figuras 1 e 2. Como os dados rotulados são selecionados aleatoriamente, existe a possibilidade de serem *outliers* e nesses casos a taxa de classificação correta é menor, o que explica a variância observada nos resultados.

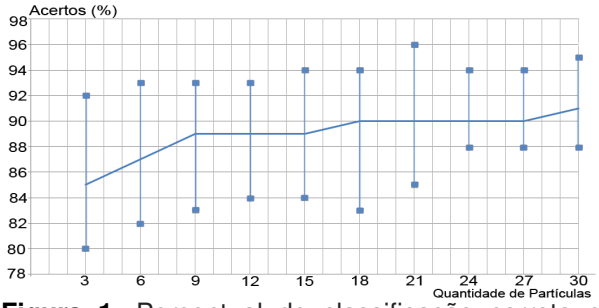

Figura 1. Percentual de classificação correta em dados da base Wine[2], variando a quantidade de partículas.

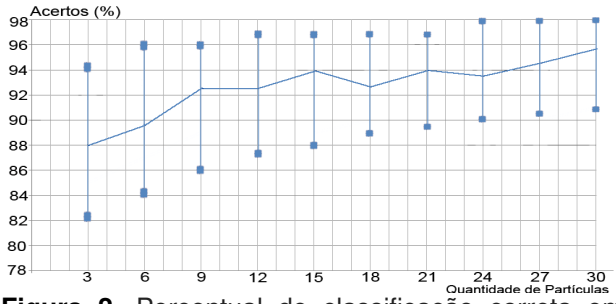

Figura 2. Percentual de classificação correta em dados da base Iris[2], variando a quantidade de partículas.

Conclusões

Com a utilização de algoritmos de Aprendizagem de Máquina conseguimos bons resultados. Por ser um processo estocástico, dependendo de eventos aleatórios, eventualmente ocorrem resultados inesperados, porém na maioria dos casos, temos resultados excelentes na classificação de bases de dados. Espera-se que esses resultados também ocorram na extensão do algoritmo para aplicação em fluxos de dados, que será o próximo passo do projeto.

Agradecimentos

Os autores gostariam de agradecer à FAPESP e à Fundunesp pelo auxílio financeiro.

---

[1] Breve F. A.; Zhao L.; Quiles M. G.; Pedrycz W.; Liu J. "Particle competition and cooperation in networks for semisupervised learning". *IEEE Transactions on Knowledge and Data Engineering (PrePrints)*, 2012. DOI 10.1109/TKDE.2011.119.

[2] Frank, A.; ASUNCION, A. "UCI Machine Learning Repository", **2010**. Disponível em: <http://archive.ics.uci.edu/ml>.