ÁREA: CV ( ) CHSA ( ) ECET (X)

ESTUDO COMPARATIVO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA NA CLASSIFICAÇÃO DE USUÁRIO DE UMA REDE SOCIAL

Bruno Vicente Alves de Lima (Bolsista PIBIC/UFPI), Vinicius Ponte Machado (Orientador, Departamento Informática e Estatística/UFPI)

Introdução

Com o objetivo de contribuir para que ocorra facilidade a troca de informações entre os pesquisadores propõe-se uma rede social chamada Scientia.Net. Essa rede permite uma melhor interação entre pesquisadores. Os usuários serão classificados de acordo com seus respectivos perfis, fazendo com que os usuários com perfis compatíveis possam interagir e possam compartilhar experiências científicas.

Com isso torna-se necessário, para a classificação dos usuários do Scientia.Net, que esta Rede Social Online trabalhe com Algoritmos de Aprendizagem de Máquina. E que ao mesmo tempo tenha bons resultados em classificar usuários, artigos e eventos científicos, realizando seu procedimento de maneira satisfatória.

Para garantir a eficiência do Scientia.Net neste trabalho foram estudados e comparados outros algoritmos de aprendizado de máquina, a Rede de Kohonen, Algoritmo K-Means e Máquina de Vetor de Suporte, com o intuito de escolher aquele com melhor desempenho para trabalhar no Scientia.Net.

Metodologia

A primeira etapa do plano de trabalho consiste em enumerar os algoritmos de aprendizagem de máquina e escolher 5 desses algoritmos para classificação automática de usuários. Isso levando em consideração o seu uso na comunidade científica.

Segunda etapa foi realizada uma pesquisa, procurando por artigos onde continham trabalhos relacionados à classificação de usuários utilizando esses algoritmos. Para a execução dos algoritmos foi utilizada a base de dados do Scientia.Net que possui um total de 2000 usuários. Devido à importância no perfil acadêmico dos usuários foram levados em consideração os seguintes campos: Graduação, Mestrado, Sub Área do Mestrado, Doutorado, Sub Área do Doutorado, Pós Doutorado, Sub Área do Pós Doutorado, Área de Interesse.

Posteriormente replicou-se a base de dados do Scientia.Net em mais 9 cópias. Sendo que cada uma possui os mesmos registros, mas em ordem distinta. Após as execuções os resultados foram analisados e colocados em forma de matriz de confusão para facilitar na comparação. Para a execução dos algoritmos supervisionados (Redes Neurais Multicamadas e Máquina de Vetor de Suporte) utilizou-se um total de 80% dos dados para treinamento e 20% para o teste da classificação.

Após a execução obteve-se, para cada algoritmo, um total de 10 matrizes de confusão, pois cada execução gera uma matriz. Para mostrar o resultado foram calculadas as médias das matrizes de confusão de cada algoritmo. Os testes foram executados na máquina com 2 GB de memória RAM, HD de 160 e Processodor Intel Quad Core 2 GHz.

.

Resultados e Discussão

Os tempos de execução mostrados na tabela 1, como dito anteriormente, são a média dos

tempos de 10 execuções, considerando tempo de treinamento e tempo de classificação ou clusterização.

Tabela 1. Tempos de Execuções dos Algoritmos.

Algoritmo	Tempo de Execução (Minutos)
Redes Neurais Multicamada	3,58
Máquina de Vetor de Suporte	0,0096
Rede de Kohonen	5,77
Algoritmo K-means	0,012
Algoritmo Cobweb	0,26

As Redes Neurais Multicamadas (MLP) e a Máquina de Vetor de Suporte (SVM) foram treinadas utilizando um total de 1800 usuários e 200 para testes. Com isso as MLPs obtiveram uma taxa de acerto de 98,5% e as SVMs obtiveram uma taxa de erro média de 0,1%.

A Rede de Kohonen e o Algoritmo K-means não necessitam dividir os dados para treinamento e outros para classificação. Pois neste trabalho é utilizado os grupos gerados por parte destes algoritmos. Levando em consideração a homogeneidade dos grupos gerados a Rede de Kohonen obteve uma taxa de erro médio de 3,22% e o Algoritmo K-means uma taxa média de 13,26%.

O Algoritmo Cobweb não apresentou resultados satisfatórios em relação os outros algoritmos, apresentando uma média de 89,7 grupos em 10 execuções. O Cobweb a cada execução gerou quantidade de grupos distintos, dificultando a criação da matriz de confusão como foi feita com os outros algoritmos. Considerando os grupos inválidos gerados o Cobweb apresentou uma taxa de acertos de 98% em um tempo de 0,26 min (Tabela 1).

.

Conclusão

Analisando os resultados dos algoritmos de aprendizado supervisionado, pode-se perceber que a Máquina de Vetor de Suporte sobressaírem-se melhor em relação às Redes Neurais Multicamadas, pois obtiveram um melhor desempenho no quesito de classificação, ou seja, apresentou o menor erro médio (tabela 1).

Dos algoritmos não-supervisionados, o algoritmo Cobweb apresentou resultados insatisfatórios, pois o comportamento do algoritmo levou a criação de mais grupos do que o esperado, assim dificultado a comparação com os outros algoritmos.

Em relação os outros algoritmos, a Rede de Kohonen obteve um tempo de execução médio bem maior em relação ao Algoritmo K-means, porém obteve uma menor taxa de erro em relação à classificação dos usuários.

Considerando as duas formas de aprendizado abordadas neste trabalho, a forma não-supervisionada torna-se mais adequada à classificação de usuários dentro do ScientiaNet, pois na classificação, utilizando algoritmos supervisionados, torna-se necessário conhecer previamente as classes em que se deseja classificar para a execução do treinamento. Os algoritmos não-supervisionados comportam-se gerando grupos de acordo com a similaridade dos usuários, sem a necessidade de atribuição prévia de classes.

Portando levando em consideração todos os fatores citados a Rede de Kohonen torna-se o

método mais adequado, pois é o método não-supervisionado que apresenta resultado satisfatórios e mais coerentes com o propósito do ScientiaNet.

Apoio: Universidade Federal do Piauí.

Referências

Fonseca, F.C.S.; Beltrame, W. A. R. ; Aplicações Práticas dos Algoritmos de Clusterização Kmeans e Bisecting K-means.

MONARD, M.C.; BARANAUSKAS, J. A. Conceitos de Aprendizagem de Máquina. In: REZENDE, S.O. (Ed). Sistemas Inteligentes: fundamentos e aplicações. São Carlos: Manole, 2003. P. 89-114. Cap. 4.

O. L. Júnior e E. Montgomery, Redes Neurais: Fundamentos e Aplicações com Programas em c. Rio de Janeiro, Brasil, 2007.

P. Braga, A. P. L. F. Carvalho e T. B. Ludermir, Redes Neurais Artificiais; Teoria e Aplicações. 2ed, Rio de Janeiro, Brasil, 2007.

S. Russel e P. Norving, Inteligência Artificial. Rio de Janeiro: Elsevier, 2004.

CORTES, C. e VAPNIK, V. Support-vector network. Machine Learning, pages 273 297. 1995.

Palavras-chave: Classificação.Aprendizado de Máquina. Perfil.