Bare Demo of IEEEtran.cls for IEEECS Conferences

Alan S. Castro
Universidade Federal de São Carlos
UFSCar
Sorocaba, Brasil
email@gmail.com

Cássia C. Monteiro
Universidade Federal de São Carlos
UFSCar
Sorocaba, Brasil
kssia.cm@gmail.com

Lucas Renan A. Nunes
Universidade Federal de São Carlos
UFSCar
Sorocaba, Brasil
email@gmail.com

Thamires C. Luz
Universidade Federal de São Carlos
UFSCar
Sorocaba, Brasil
thamiluz@gmail.com

***Resumo*—The abstract goes here. DO NOT USE SPECIAL CHARACTERS, SYMBOLS, OR MATH IN YOUR TITLE OR ABSTRACT.**

***Keywords*-web spam; classificador;**

I. Introdução

Com a internet, é comum usarmos sites de busca para encontrarmos resultados como páginas, arquivos, imagens, músicas ou etc . Os resultados exibidos são elencados, através de search engines que geram um rank com os melhores resultados de acordo com o conteúdo relevante na página. O search engine começou por volta de 1994, com o World Wide Web Worm, e tinha por volta de 110.000 páginas web associadas. Após essse ano, isso foi crescendo cada vez mais [**?**]. Com o crescimento do uso de sites de busca, surgiu o web spam que segundo [**?**] é definido pela ação de alterar páginas da internet com a intenção de enganar os search engines, de maneira que o conteúdo inserido para elencar a página no rank, não tenha relação com a página em questão. [**?**], relata que web spam é um dos maiores problemas dos search engines, porque degrada a qualidade dos resultados oferecidos, onde muitas pessoas ficam frustadas ao acessarem páginas que não tem relação com o termo buscado. Ainda relata que o web spam impacta também na área econômica, pois o retorno financeiro de propagandas são aumentadas de acordo com uma boa colocação no rank de busca. Com o crescimento desse problema, estudos se tornam cada vez mais necessários para detectar se uma página usa de métodos fraudulentos para se beneficiar na posição do rank. Com o uso de técnicas de aprendizado de máquina, podemos classificar as páginas como sendo legítimas ou spam. Na sessão II, é realizada a descrição da base de dados utilizada nesse trabalho, e citados métodos definidos. Na sessão III, é descrito os métodos utilizados com maior detalhe.

II. Descrição do trabalho

Para esse trabalho foi disponibilizado uma base de dados com aproximadamente 4 mil amostras, ou seja páginas, e 137 atributos além do atributo classe. Os atributos são dados contínuos, e descrevem características para páginas que são ou não são spam. A base está balanceada tendo metade das amostras classificadas como spam, e a outra metade classificada como não spam. De acordo com isso, é proposto a comparação de três métodos de aprendizado de máquina, sendo definido a utilização de Naive Bayes, Regressão Logística e Redes Neurais Artificiais, descritas com detalhe na próxima seção. Foi utilizado o software livre Octave para implementação e testes necessários.

III. Métodos

A. Naive Bayes

De acordo com [**?**], o classificador bayeseano é uma técnica probabilística baseada no teorema de Thomas Bayes, denominado naive Bayes, onde é considerado um algoritmo ”ingênuo”, devido assumir que todos os atributos possuem relações independentes entre si, o algoritmo de naive Bayes, pode ser descrito como o produtório das probabilidades de cada atributo pelas classes existentes, onde a amostra é classificada para determinada classe, caso o valor calculado de produtório seja superior ao das outras classes. Comparativos entre algoritmos demonstram que o naive Bayes, obteve resultados compatíveis com os métodos de árvore de decisão e redes neurais. Devido a sua simplicidade e o alto poder preditivo, é um dos algoritmos mais utilizados [**?**].

B. Regressão Logística

A regressão logística, é um classificador que através da função sigmoidal classifica um grupo de amostras. [**?**], retrata que o método é robusto, flexível e de fácil utilização, com o objetivo de encontrar o melhor modelo que descreva os atríbutos em uma determina classe.

C. Redes Neurais Artifíciais

IV. Conclusion

The conclusion goes here. this is more of the conclusion

V. section

teste thamires