

# Bare Demo of IEEEtran.cls for IEEECS Conferences

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

Lucas Renan A. Nunes  
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

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 esse 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 frustradas 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, nesse trabalho comparamos três métodos para classificação de web spam. Na sessão II, é realizada a descrição da base de dados utilizada nesse trabalho, e citados os métodos definidos. A sessão III contém uma explicação mais detalhada dos métodos utilizados,

modelo de comparação entre os algoritmos, e as escolhas de parametros.

## 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 bayesiano 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 produto das probabilidades de cada atributo pelas classes existentes, onde a amostra é classificada para determinada classe, caso o valor calculado de produto 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 sigmoideal 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 atributos em uma determinada classe.

### C. Redes Neurais Artificiais

Redes Neurais Artificiais é um método que pode ser usado como classificador, foi inicialmente desenvolvido para funcionar como o sistema neural humano, desde então uma grande variedade de modelos de redes foram desenvolvidos, são compostas por conexões entre neurônios que em conjunto determinam o comportamento da rede, a escolha do modelo da rede depende do problema a ser resolvido, segundo [?] o mais utilizado é o de multilayer perceptrons.

## IV. METODOLOGIA

Para a implementação dos métodos já citados, foi utilizado o Octave, e outras bases de dados mais simples. Para o naive bayes, foi utilizado uma base de dados com atributos discretos e duas classes booleanas, com essa base de dados foi encontrado uma acurácia de 92.3% sem utilizar suavização, onde com a suavização o resultado obtido não era satisfatório sendo em torno de 60% . Regressão logística apresentou a acurácia de 100% em uma base de dados linear com dois atributos e duas classes booleanas, para uma base de dados não linear com a mesma quantidade de atributo e classe, o método apresentou uma acurácia de 83% . Com os algoritmos criados, foram feitas alterações e ajustes nos algoritmos de forma que tivesse melhor aproveitamento para a base de dados desse trabalho. As alterações, dificuldades e escolhas de parâmetros são descritas nas sub-sessões abaixo.

### A. Naive Bayes

A principal dificuldade encontrada em alterar o naive bayes originalmente criado com uma base de dados discreta, foi encontrar resultados satisfatórios com uma base de dados com atributos contínuos, onde existe uma grande dispersão de dados, tornando-se lento o cálculo de probabilidade dos elementos. Verificado que o naive bayes não tem um bom resultado com dados contínuos, para resolver esse problema, foi escolhida a solução de cestas, isto é, uma maneira de discretizar os dados contínuos, agrupando-os em cestas de equivalência, uma descrição mais detalhada pode ser encontrada em [?]. Para determinar a escala dos atributos foi utilizado o quartil das amostras de treino, sendo utilizado essa mesma escala para as amostras de teste. Após definida a escala, as amostras foram agrupadas em cinco cestas de valores múltiplos com relação a escala, onde o valor cinco foi escolhido através de testes realizados através da análise do comparativo entre desempenho e acurácia obtida.

### B. Regressão Logística

A implementação do método de regressão logística originalmente tinha como teste uma base de dados com dois atributos, sendo portanto rápida e precisa. Ao testar o

algoritmo implementado com a base de dados desse trabalho, foi constatado que teve um aumento significativo de desempenho, então foi decidido utilizar a técnica de análise de componentes principais (PCA) para reduzir a dimensionalidade de atributos [?]. Com o uso do PCA, foi calculado o parâmetro  $k$  com a matriz de autovalores  $S$  retornada da função  $SVD$ , de modo que a soma dos valores até  $k$  calculado sobre o total de atributos, resultasse em uma variância de 99%, desse modo, os atributos foram reduzidos e método ficou eficaz.

### C. Redes Neurais Artificiais

Explicar a implementação do RN e como foi feito as escolhas dos parâmetros, neurônios etc

### D. Medida de desempenho

Para validar os resultados obtidos de todos os algoritmos implementados, utilizamos a validação cruzada, ou *cross-validation* que consiste em separarmos amostras classificadas da base de dados em  $k$  partes, após essa separação, fazemos  $n$  repetições de maneira que cada parte seja utilizada como teste de classificação, e as outras partes sejam utilizadas para o treinamento do algoritmo. De acordo com [?], a escolha de  $k$  é geralmente igual a dez.

## V. CONCLUSION

The conclusion goes here. this is more of the conclusion

## VI. SECTION

teste thamiere