COMPARATIVO ENTRE ALGORITMOS DE CLASSIFICAÇÃO EM IMAGENS DE POSSÍVEIS METEOROS

Leonardo Gomes Nunes, *114829* Ronaldo Akira Fujigaki Kinoshita, *112258*

***Resumo*—De posse de um conjunto de dados contendo características de imagens com possíveis meteoros, foi feita uma análise sobre o desempenho dos seguintes algoritmos de aprendizado de máquina a fim de classificá-las como meteoros ou não-meteoros: Redes Bayesianas, Perceptron Multicamadas, Árvores de Decisão e K-Vizinhos mais próximos.**

I. INTRODUÇÃO

Meteoro é um fenômeno astronômico da passagem de um objeto sólido pela atmosfera terrestre, sendo o mesmo proveniente do espaço. Após o contato com o atmosfera, é formado um efeito luminoso que pode ser visto e também captado. A detecção desses eventos são de grande importância para inúmeros propósitos científicos e educacionais. Dessa forma, trabalhamos com um conjunto de dados baseado em uma série de fotos que foram tiradas do céu pelo Observatório de Astronomia da UNIVAP (Universidade do Vale do Paraíba) com a intenção de identificar meteoros. As fotos foram feitas por câmeras de baixo custo e podem incluir alguns elementos que dificultam nesta identificação como: insetos, pássaros, aviões, raios, chuva e os próprios ruídos da foto. Foi utilizado um software nas fotos em áreas de interesse, com a ideia de extrair algumas características e com base nas mesmas classificar as imagens em "Meteoro" e "Não-Meteoro". A proposta deste trabalho é utilizar quatro algoritmos diferentes (Redes Bayesianas, Perceptron Multicamadas, Árvores de Decisão e K-Vizinhos mais Próximos) para resolver esse problema de classificação, analisando seus desempenhos e os comparando.

Figura 1. Exemplo de um meteoro

Figura 2. Exemplo de um não-meteoro

II. TRABALHOS RELACIONADOS

Embora a computação e os métodos de inteligência artificial sejam usados para diversas tarefas, ainda não foram realizados muitos estudos sobre algoritmos para classificação de meteoros. Porém existem trabalhos sobre o uso de algoritmos para classificação de imagens.

No artigo de [1] foi feito uma comparação com o uso de redes neurais e com o método de Máxima Verossimilhança Gaussiana e o uso de Redes Neurais para classificação de imagens de satélite para sensoriamento remoto.

Waleska Nishida [2] também fez um estudo sobre o uso de redes neurais artificiais para a classificação de imagens para sensoriamento remoto em seu Trabalho de Mestrado, no qual ela chega a conclusão que o uso de redes neurais pode ser satisfatório, quando comparado com os métodos estatísticos, pelo fato de obterem resultados parecidos, porém não necessitam de informação estatística prévia.

Álisson Krug et al. [3] também usaram inteligência artificial para a classificação de imagens. Eles usaram redes neurais artificiais para classificar sinais na íris e assim auxiliar na iridologia.

Os romenos Victor Ştefan Roman e Cătălin Buiu [4] em seu artigo, Automatic detection of meteors in spectrograms using artificial neural networks, usam do algoritimo de redes neurais para automatizar a detecção de meteoros usando gravações de radio salvadas como espectrogramas.

Denis Vida, Emil Siladi e Emmanuel Karlo Nyarko [5] publicaram um artigo sobre usar redes neurais e máquinas de vetores de suporte com o objetivo de minimizar o número de falsos-positivos na classificação de meteoros

III. METODOLOGIA

A. Pré-processamento

Para a análise dos algoritmos foi realizado um pré-processamento dos dados, sendo que o conjunto foi separado

em teste e treino, com uma porcentagem de 35% e 65%, respectivamente. Após a separação em teste e treino foi realizado uma remoção nas colunas de **"Id"** e **"Class"** que não serão usadas como atributos para os algoritmos. Sendo assim, o conjunto possui a seguinte configuração:

Tabela I
CONJUNTO DE DADOS BRUTO

Conjunto	Linhas	Atributos
Treino	80	3451
Teste	42	3451

Foi realizada a normalização dos dados e uma remoção nos atributos com variação nula. Foi observado também que existe um valor **"nan"** que precisava ser retirado, então ele foi substituído pela média do atributo no conjunto.

Tabela II
CONJUNTO DE DADOS APÓS REMOÇÃO DE DADOS COM VARIAÇÃO NULA

Conjunto	Linhas	Atributos
Treino	80	1306
Teste	42	1306

Para as análises que se seguem, o conjunto acima (após remoção de atributos com variação nula) será entendido como o novo conjunto de dados bruto.

B. Técnicas

1) Seleção de atributos: Com o intuito de diminuir o número de atributos dos conjuntos de dados, foram aplicados dois algoritmos de seleção de atributos, nos quais os mesmos estão explicitados abaixo:

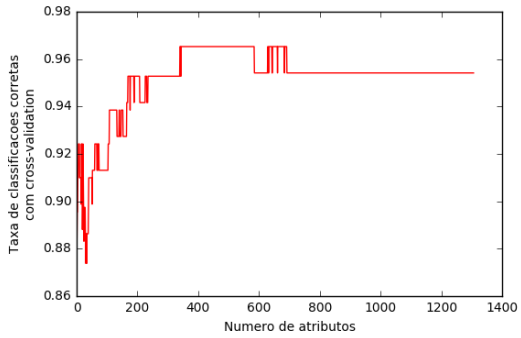
Seleção de atributos baseada em árvores de decisão (TFS): Essa técnica de seleção de atributo faz uso de uma Árvore de Decisão para atribuir pesos (*leia-se importância*) aos atributos do conjunto de dados. Os atributos com maiores pesos são selecionados para representar o conjunto final. Após sua utilização, o conjunto foi reduzido de 1306 atributos para 96.

Tabela III
CONJUNTO DE DADOS APÓS **TFS**

Conjunto	Linhas	Atributos
Treino	80	96
Teste	42	96

Seleção de atributos recursiva com cross-validation (RFE): De forma recursiva, e por meio de uma SVM (Support Vector Machine), essa técnica utiliza cada coluna para classificar o conjunto por meio de validação cruzada. A cada interação é calculada a acurácia dessa classificação e após o final da mesma é gerada um conjunto resultante com o número de colunas que obteve a maior acurácia.

Figura 3. Taxa de classificações corretas por num. de atributos no RFE



O número de atributos que obteve a maior acurácia foi 691, de tal forma que permanecemos com o seguinte conjunto:

Tabela IV
CONJUNTO DE DADOS APÓS **RFE**

Conjunto	Linhas	Atributos
Treino	80	691
Teste	42	691

2) Algoritmos de classificação:

KNN - K Vizinhos mais próximos: A ideia do KNN é determinar o rótulo de classificação de uma amostra baseado nas amostras vizinhas advindas de um conjunto de treinamento, a grande vantagem do KNN é sua abordagem simples de ser compreendida e implementada.

Redes bayesianas: Em resumo, redes bayesianas, também conhecidas como redes de opinião, redes causais e gráficos de dependência probabilística, são modelos de representação do conhecimento que trabalham com o conhecimento incerto e incompleto por meio do Teorema de Bayes, onde os nós representam as variáveis (discretas ou contínuas), e os arcos representam conexões diretas entre eles.

Decision Tree - Árvore de decisão: Uma árvore de decisão é um mapa dos possíveis resultados de uma série de escolhas relacionadas. Permite que um indivíduo compare possíveis ações com base em seus custos, probabilidades e benefícios. Uma árvore de decisão geralmente começa com um único nó, que se divide em possíveis resultados. Cada um desses resultados leva a nós adicionais, que se ramificam em outras possibilidades. Assim, cria-se uma forma de árvore.

Perceptron: O perceptron é um algoritmo de aprendizado de classificadores binários. Ele é capaz de resolver apenas problemas linearmente separáveis. Por meio do processo de treinamento, o algoritmo aprende a classificar as entradas em dois grupos diferentes.

MLP - Perceptron multicamadas: O perceptron multicamadas (MLP) é uma rede neural semelhante ao perceptron simples, porém possui mais de uma camada de neurônios. Em casos em que não há a possibilidade de uma única reta separar os elementos, há o uso da MLP que, gera mais de uma reta classificadora.

IV. Análise experimental

Conforme foi mostrando anteriormente, para a análise experimental foram utilizados três conjuntos de dados, sendo eles o conjunto **Bruto**, após **TFS** e após **RFE**. Para cada conjunto foi feita a classificação utilizando os algoritmos descritos e comparadas suas medidas de qualidades e desempenho.

A. Conjunto de Dados

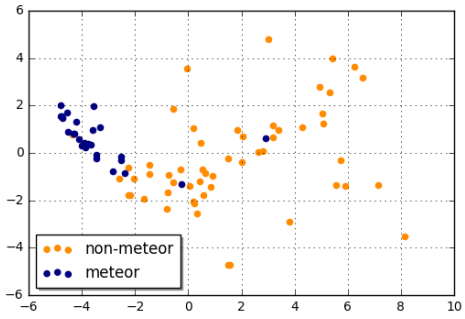
A distribuição de meteoros e não-meteoros nos conjuntos de dados é a seguinte:

Tabela V
Conjunto de Dados

Conjunto	Meteoro	Não-Meteoro	Total
Treino	26	54	80
Teste	15	27	42

Como forma de visualizar os dados do conjunto de treino de maneira gráfica, foi aplicado um algoritmo de redução de dimensionalidade, resultando assim em duas dimensões.

Figura 4. Gráfico PCA de duas dimensões



B. Configurações do algoritmo e ambiente computacional

1) ***Ambiente****:* Sistema operacional - Ubuntu 16.04 Desktop (64-bit), CPU - Intel Core i5-5200U CPU @2.20GHz x4, Memória RAM - 8GB DDR3.

2) ***Linguagem de programação****:* Python 3.5.2, GCC 5.4.0.

3) ***Pacotes****:* Pandas 0.23.1, SciKit-Learn 0.19.1, SciKit-Plot, NumPy e SciPy.

C. Critérios de análise

Foram utilizadas as seguintes métricas para analisar a qualidade e desempenho dos algoritmos de classificação:

1) Acurácia
2) Log Loss
3) Precisão
4) Matriz de Confusão
5) Média
6) Mediana
7) Desvio Padrão

V. Resultados e discussões

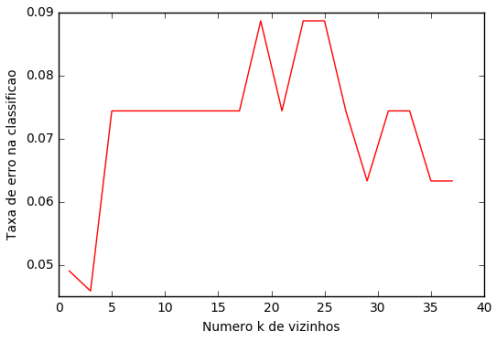
Como primeira análise, utilizamos os algoritmos de classificação no conjunto de dados bruto, e os resultados estão descritos na tabela a seguir:

Tabela VI
Conjunto de Dados **Bruto**: Acurácia, Log Loss e Precisão

Resultados	Acurácia	Log Loss	Precisão
KNN	0,9286	2,4671	1,0000
MLP	0,9048	3,2894	1,0000
DECISION TREE	0,8810	4,1118	1,0000
REDES BAYESIANAS	0,7381	9,0459	0,83333

Especialmente para o KNN, foi feita uma 10-fold cross-validation para encontrar um melhor número de vizinhos K possível, a fim de maximar sua acurácia. Isso consiste em calcular a taxa de erro em um range especifico de K vizinhos, permanecendo com aquele que obtiver a menor taxa e o utilizando na classificação do conjunto de teste. Essa técnica foi utilizada nos três conjuntos e o resultado para esse conjunto é o seguinte:

Figura 5. Melhor K encontrado: 3



Como forma de analisar o desempenho do conjunto utilizado, foi calculada algumas métricas para a acurácia:

Tabela VII
Conjunto de Dados **Bruto**: Análise da acurácia de todos os algoritmos

Resultados	Acurácia
MÉDIA	0,8631
MEDIANA	0,8929
DESVIO PADRÃO	0,0741

Para que possamos visualizar melhor quais são efetivamente as classificações que os algoritmos estão realizando, foram calculadas matrizes de confusão para cada um:

Tabela VIII
Matriz de confusão KNN - Conjunto de dados bruto

KNN	NEGATIVO	POSITIVO
NEGATIVO	27	0
POSITIVO	3	12

Tabela IX
MATRIZ DE CONFUSÃO MLP - CONJUNTO DE DADOS BRUTO

MLP	NEGATIVO	POSITIVO
NEGATIVO	27	0
POSITIVO	4	11

Tabela X
MATRIZ DE CONFUSÃO DECISION TREE - CONJUNTO DE DADOS BRUTO

Decision Tree	NEGATIVO	POSITIVO
NEGATIVO	27	0
POSITIVO	4	11

Tabela XI
MATRIZ DE CONFUSÃO REDES BAYESIANAS - CONJUNTO DE DADOS BRUTO

Redes Bayesianas	NEGATIVO	POSITIVO
NEGATIVO	26	1
POSITIVO	10	5

Dessa forma, podemos notar que o algoritmo de classificação que se saiu melhor foi o KNN, com uma acurácia pouco maior que 92%, seguido por MLP, Árvore de Decisão e Redes Bayesianas. Em questão de precisão, todos os algoritmos, exceto a Rede Bayesianas, obtiveram 100%, significando que os mesmo classificaram corretamente todas as imagens nas quais efetivamente não continham um meteoro, como é possível observar na Matriz de Confusão de cada um deles.

Para os próximos dois conjuntos de dados será suprimida a visualização da Matriz de Confusão.

A seguir, temos os resultados para o conjunto de dados após a seleção de atributos baseada em árvores de decisão (TFS):

Tabela XII
CONJUNTO DE DADOS APÓS **TFS**: ACURÁCIA, LOG LOSS E PRECISÃO

Resultados	Acurácia	Log loss	Precisão
KNN	0,8571	4,9341	0,90909
MLP	0,8810	4,1118	0,91667
DECISION TREE	0,9047	3,2894	1,0000
REDES BAYESIANAS	0,7619	8,2236	0,77778

Dessa vez, o algoritmo que obteve a maior acurácia foi a Árvore de Decisão, pouco maior que 90%, e também a única a obter uma precisão de 100%. Essa medida é inferior a maior acurácia do conjunto anterior, obtida pelo KNN, e exatamente igual à MLP, também do conjunto anterior. Medidas referentes a acurácia do conjunto de dados como um todo também foram calculadas:

Tabela XIII
CONJUNTO DE DADOS APÓS **TFS**: ANÁLISE DA ACURÁCIA DE TODOS OS ALGORITMOS

Resultados	Acurácia
MÉDIA	0,8571
MEDIANA	0,8810
DESVIO PADRÃO	0,0558

Por ultimo, temos as medidas referentes a seleção de atributos recursiva com cross-validation (RFE):

Tabela XIV
CONJUNTO DE DADOS APÓS **RFE**: ACURÁCIA, LOG LOSS E PRECISÃO

Resultados	Acurácia	Log loss	Precisão
KNN	0,88095	4,11178	0,91667
MLP	0,9048	3,2894	1,0000
DECISION TREE	0,8810	4,1118	1,0000
REDES BAYESIANAS	0,8333	5,7565	1,0000

Tabela XV
CONJUNTO DE DADOS APÓS **RFE**: ANÁLISE DA ACURÁCIA DE TODOS OS ALGORITMOS

Resultados	Acurácia
MÉDIA	0,8750
MEDIANA	0,8810
DESVIO PADRÃO	0,0259

VI. CONCLUSÃO

Com a análise experimental pode-se concluir que o algoritmo com melhor acurácia geral, foi o KNN, com o conjunto de dados bruto. Porém calculando-se as acurácias médias dos 3 conjuntos analisados, a maior média de acurácia foi obtida com o conjunto após RFE. De uma forma geral, todos os algoritmos conseguiram uma acurácia satisfatória, porém isso não pode ser uma medida precisa devido ao fato do conjunto de dados ser pequeno.

Para trabalhos futuros pode-se testar os algoritmos em conjuntos maiores, e procurar padrões nos atributos, assim encontrando-se os atributos mais relevantes para a classificação de meteoros. Também poderá ser feito um estudo mais aprofundado nos algoritmos usados para extração de dados sobre as imagens.

REFERÊNCIAS

[1] A. G. R. A. T. G. Rossana B. Queiroz, Priscila A.da R. Severino, "Redes neurais: Um comparativo com máxima verossimilhança gaussiana na classificação de imagens cbers 1," Artigo, 2004.

[2] W. NISHIDA, "Uma rede neural artificial para classificaÇÃo de imagens multiespectrais de sensoriamento remoto," Tese de Mestrado, 2 1998.

[3] F. L. M. N. e. A. S. M. Állison Bohnert Krug, Adriane Parraga, "Análise e reconhecimento de padrões usando processamento de imagens e inteligência artificial," Revista de Iniciação Científica da ULBRA, 2008.

[4] V. Ştefan Roman e Cătălin Buiu, "Análise e reconhecimento de padrões usando processamento de imagens e inteligência artificial," 2015 IEEE 10th Jubilee International Symposium on Applied Computational Intelligence and Informatics, 5 2015.

[5] D. V. e. E. K. N. Emil Silađi, "Video meteor detection filtering using soft computing methods," Proceedings of the IMC, 8 2015.