**UM MÉTODO PARA CLASSIFICAÇÃO DE DADOS BASEADO NOS K-VIZINHOS**

**MAIS PRÓXIMOS PARA O RECONHECIMENTO DE CARACTERES**

**RESUMO**

Fonte de intensa pesquisa, o reconhecimento de padrões ainda apresenta vários problemas, seja no reconhecimento de caracteres manuscritos, caracteres impressos ou imagens. As dificuldades mais evidentes estão relacionadas ao grande volume de dados associados às abordagens tradicionais de extração de informação ou à natureza complexa de dados em conjuntos de bases menores. Diante desse cenário exemplificador podemos observar o clássico algoritmo de classificação k-NN. Uma nova abordagem para o algoritmo k-NN é proposta com o objetivo de tornar a classificação de dados para reconhecimento de caracteres mais eficaz. Esta abordagem pode trazer benefícios tanto para sociedade no uso em aplicações tecnológicas, como para a comunidade científica.

1. **INTRODUÇÃO**

A análise dessas bases de dados tem fornecido informações relevantes para a comunidade científica, assim como para o mercado corporativo. Atender à crescente demanda por extração, interpretação e/ou relacionamentos de informações, torna-se necessária a utilização do processo de descoberta de conhecimento em bases de dados (KDD, do inglês Knowledge Discovery in Database).

O processo de descoberta de conhecimento em bases de dados foi definido como “processo não trivial, formado por várias etapas, interativo e iterativo, para identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, em grandes conjuntos de dados”. Trata-se de um processo que pode utilizar conceitos e técnicas de diversas disciplinas como aprendizado de máquina, banco de dados, estatística e inteligência artificial.

O KDD é considerado um processo interativo por necessitar de especialistas em sua condução, e iterativo por eventualmente ser necessário retornar ou avançar etapas. Basicamente, as etapas operacionais do processo de KDD são organizadas em três grandes grupos, a saber: pré-processamento, mineração de dados e pós-processamento. Problemas reais de diversos domínios podem ser analisados e estudados para que conhecimentos sejam adquiridos através de aplicações de técnicas de mineração de dados. Nesse sentido, pode-se citar: Processamento de Imagens, Bioinformática, Mineração de Texto e Análise de Dados para previsão de Abalos Sísmicos. Dentre as tarefas de mineração de dados existentes na literatura, a Classificação consiste em associar objetos a uma classe, tendo por base uma coleção de classes predefinidas. O presente trabalho propõe a pesquisa e o estudo relacionados à Classificação e considera o problema de Identificação de Caracteres Manuscritos como um estudo de caso.

1. **RECONHECIMENTO DE CARACTERES MANUSCRITOS E SUAS ABORDAGENS**

Atualmente ainda há dificuldades com o reconhecimento de caracteres, seja ele um caráter manuscrito ou impresso.

Por conta disso, o desafio atual é criar um algoritmo que tenha maior eficácia com o reconhecimento de caracteres, fazendo que com isso, reduza os números de incidentes/erros da leitura dos mesmos. Já existem aplicações reais de sucesso sobre o assunto, como por exemplo, a identificação de placas de automóveis.

Apesar de tal dificuldade, não se trata de um tema novo e com a solução deste problema pode ser adaptado e aplicado em diversas áreas.

1. **A CLASSIFICAÇÃO DE DADOS E SEUS ALGORITMOS**

A mineração de dados nada mais é que uma das etapas do KDD, a qual aplica algoritmos para análise e descoberta de padrões ou modelos de dados. Dentro desta etapa de mineração de dados, existem as tarefas como: extração de regras de associação, o agrupamento de dados(clustering) e a classificação de dados.

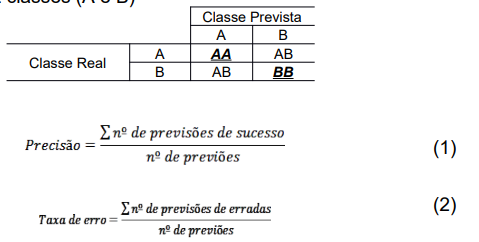
A classificação de dados consiste em um processo que mapeia um objeto (exemplar ou tupla) em uma ou classe (categoria) predefinida com base em um conjunto de objetos previamente classificados.

Em síntese, um modelo de classificação pode ser definido como a tarefa de aprendizado de uma função que mapeie um conjunto de objetos em classes (rótulos) pré determinadas. Este modelo é construído através de uma técnica de classificação, ou seja, uma abordagem sistemática a partir de um conjunto de dados de entrada. Para a execução dessas tarefas existem diversas técnicas ou algoritmos que podem ser considerados, destacando-se, dentre eles: redes neurais, classificadores baseados em regras e algoritmos que atuam em árvores de decisão.

Já a avaliação de um modelo de classificação é baseada nas contagens de registros de testes previstos, as quais apresentam, pelo modelo, respostas corretas e incorretas. Estas contagens são tabuladas em uma tabela conhecida como matriz de confusão.

A Tabela 1 apresenta a matriz confusão de uma classificação binária em que os resultados obtidos satisfatoriamente pelo algoritmo (acertos) são apresentados na diagonal principal (em destaque). Além disso, embora a matriz confusão forneça informações importantes, apresentar tais resultados como um único índice pode ser mais conveniente. Nesse sentido, podem ser consideradas como métricas de desempenho tanto a Acurácia (Precisão) (Equação 1) quanto a Taxa de erro (Equação 2).

**Tabela 1**: Matriz confusão para um problema de 2 classes (A e B).



Neste contexto, o presente trabalho propõe um método de classificação que considera o clássico algoritmo k-NN e conceitos de processamento de imagens no pré processamento.

1. PRÉ-PROCESSAMENTO

Está etapa e aplicada com extrema relevância no processo KDD e seu objetivo e preparar os dados de maneira apropriada para mineração. O tratamento desses dados são realizados diversas técnicas e podem se destacar: amostragem, redução de dimensionalidade.

Essa etapa de pós-processamento e importante devido à grande diversidade de formas em que os caracteres manuscritos podem ser encontrados. Uma de suas abordagens no reconhecimento de um caractere manuscrito é através da digitalização.

O pré processamento passa por uma etapa bem conhecida ao ser efetuado: Limpeza de ruídos, Centralização do caractere, Ajuste da imagem, Matriz binária monocromática (preto e branco), Corte com base na quantidade de pixels pretos e brancos.

Amostragem: técnica de escolha de amostras adequadas para um analise de um todo.

Redução de dimensionalidade: existem 2 métodos a ser feito a extração de características e seleção de características, a extração cria novas características a partir de transformações ou combinações do conjunto de características original. A seleção selecionam, segundo determinado critério, o melhor subconjunto do conjunto de características original

1. MÉTODO PROPOSTO

O algoritmo k-NN atua na classificação de objetos desconhecidos, ou seja, que ainda estão sem classe, baseando-se na comparação com objetos similares que foram classificados previamente. Para cada objeto ainda não classificado, o algoritmo verifica as classes dos k objetos mais próximos (mais semelhantes ou menos diferentes) e aloca esse objeto a uma classe. Para a análise de similaridade entre dois objetos é possível utilizar várias métricas, como a distância de Manhattan ou a Distância Euclidiana.

Um algoritmo é proposto e traz por etapas a execução do k-NN. O algoritmo calcula a distância para cada objeto z ainda não classificado e todos os objetos da classe de treinamento. Em seguida, as distâncias de z aos seus k-vizinhos mais próximos são ordenadas, e z recebe o rotulo (é alocado a classe) de maior frequência dentre os k objetos mais próximos.

O método proposto tem como objetivo tornar mais robusta a identificação de caracteres manuscritos com base no clássico algoritmo k-NN, ao acrescentar etapas de binarização da imagem através de diferentes limites de uma matriz de inteiros intermediária, que se chama corte de pixels, e diferentes grupos de atributos ao algoritmo k-NN tradicional.

Antes do início da identificação da imagem através do método proposto, é necessário realizar algumas etapas de pré-processamento. Em primeiro lugar é preciso converter a imagem do formato RGB, vermelho, verde e azul, para o formato monocromático. Na segunda etapa, aplicar um filtro para eliminação de ruídos. Na terceira etapa, foi necessário atuar na centralização de cada caractere no espaço. Por fim, a quarta e tapa consiste em dimensionar a imagem para 300x300 pixels. Após terminado as etapas de pré-processamento, o método proposto deverá ser executado e então ocorrerá a identificação do caractere.

A primeira etapa do método proposto consiste na extração de atributos com base na matriz binária B com 32 linhas e 32 colunas. Para isso a imagem pré-processada serve como origem para preencher uma matriz intermediária de inteiros I. Cada célula dessa matriz possui a quantidade x de pixels pretos existente em sua área correspondente na imagem original.

Com o intuito de gerar diferentes matrizes binárias, valores de corte foram submetidos. Para esse trabalho foi utilizado os valores 25%, 50%, 75% e 90%, que já foram identificados em experimentos preliminares.

Para tornar o método mais robusto, uma base de dados com os resultados coletados é formada. São armazenados os k objetos mais próximos aplicados a cada matriz, para cada grupo de instâncias. O método retorna à classe com maior ocorrência dentro os k objetos semelhantes.

1. computacionais

Para fins técnicos, especificações para realizar o trabalho são documentadas, tais como o algoritmo proposto neste trabalho que foi implementado em linguagem Java. Também foi necessário um computador utilizando de um processador Intel Celeron, 1,5 Ghz, com 2 GB de memória RAM e Sistema Operacional Windows 10 Pro. A avaliação dos resultados foi realizada em duas etapas. Primeiro, a validação das bases utilizando k-fold cross validation, que atua na verificação da precisão das bases, através de sua acurácia. Em segundo lugar, experimentos foram realizados com o método proposto e com algoritmos da ferramenta Weka.

Uma das formas de validar a generalização de um modelo para uma base de dados é o método de validação cruzada. Consiste em selecionar, vindo da base, várias amostras aleatórias simples e de igual tamanho, que serão destinadas a conjuntos de treino e teste. É feito o comparativo através da aplicação da expressão que permite calcular o percentual do erro fornecido. O método K-fold cross validation consiste em estabelecer segmentos, amostras, de mesmo tamanho removidos da base. E para cada iteração, essas amostras são utilizadas como base de treinamento ou de teste.

Dentro da complexidade que é lidar com o processo do KDD, são utilizadas ferramentas com intuito de realizar tarefas operacionais e de controle. Essas ferramentas auxiliam na necessidade de lidar com volume alto de dados e dificuldades de integração e comparação de algoritmos específicos. A ferramenta Weka possui implementação de diversos dos algoritmos mais conhecidos para mineração de dados, e foi considerada nos experimentos computacionais.

A Weka é uma ferramenta de código aberto, customizável e expansível. Desenvolvida em Java e possui entradas de arquivos em ARFF, CSV e C45. A extensão de arquivo ARFF se trata de um formato ASCII para definir atributos e valores, como numa base de dados.

Para realizar a avaliação e validação das bases de dados, foi utilizado a abordagem k-fold cross-validation. Utilizando-se das bases binárias do repositório UCI, testes foram realizados com o método proposto e os algoritmos de classificação k-NN e redes neurais, que se encontram na ferramenta Weka. Os resultados do algoritmo k-NN do Weka foram processados através do classificador denominado IBK.

O próximo experimento consistiu em executar o método proposto e os algoritmos do Weka no grupo de atributos de 1 para k = 3. O método proposto se destacou obtendo os melhores resultados em oito das dez instâncias. Quanto ao k-NN, o método proposto produziu os melhores resultados em nove das dez bases.

1. CONSIDERAÇÕES FINAIS

O presente trabalho trouxe a proposta de um método para classificação de caracteres, considerando conceitos de vizinhança mais próxima. Foram realizadas atividades do processo de KDD, em que: (i) no pré-processamento foi realizada limpeza, transformação e organização de imagens e matrizes com dados de entrada; (ii) na Mineração de dados ocorreu execução do método proposto e de algoritmos do Weka para obtenção dos resultados; (iii) por fim, no pós-processamento, houve a apresentação dos resultados comparativos com algoritmos da literatura bem como apresentação de uma matriz confusão. A partir dos resultados, constatou-se que o presente trabalho apresentou um ganho considerável em acurácia, especialmente em caracteres que se diferenciam do padrão comum. Naturalmente, o método proposto consome maior tempo de processamento do que o k-NN devido utilização deste método várias vezes, sendo mais rápido que o algoritmo de Redes Neurais para testes com instâncias de mesmo tamanho.

Assim, com base nos resultados obtidos, o método proposto pode se constituir como uma alternativa para a resolução do problema abordado. Para a realização de novas pesquisas seguem como sugestões para trabalhos futuros: (i) Considerar, de maneira aleatória, subconjuntos de atributos. Dessa maneira o método pode se tornar heurístico, buscando a maximização da acurácia por meio da seleção de subconjuntos de atributos; (ii) Pesquisar sobre maneiras de calibrar automaticamente o valor de corte para a geração de matrizes binárias; (iii) Usar outras instâncias maiores tanto em quantidade de caracteres quanto em classes.