Classificação de Texto: Acelerando o kNN com uma Estrutura de Árvore

Anônimo[1]

[1]Instituição

Abstract. *This meta-paper describes the style to be used in articles and short papers for SBC conferences. For papers in English, you should add just an abstract while for the papers in Portuguese, we also ask for an abstract in Portuguese ("resumo"). In both cases, abstracts should not have more than 10 lines and must be in the first page of the paper.*

Resumo. *Vive-se atualmente na Era da Informação, graças a popularização da internet, que permite aos usuários criar dados dos mais diversos tipos. Com tantos dados espalhados, torna-se imprescindível criar-se mecanismos para seleção de informações relevantes para cada tipo de usuário, isto é, classificar os dados para seu público alvo. Este artigo apresenta uma versão mais rápida do algoritmo de classificação, kNN. Aceleramos esse algoritmo por meio da implementação de uma estrutura dados em Árvore, onde os vizinhos mais próximos são encontrados ao se percorrer pelos nós da árvore. Desta forma, observamos uma melhoria na velocidade do processo de classificação, habilitando-o para o processamento de grandes volumes de dados.*

1. Introdução

Em muitos dos processos e atividades realizados diariamente por pessoas, é utilizado, mesmo que implicitamente, o ato de classificar. Desde um médico diagnosticando uma doença, até um buscador retornando o resultado de uma pesquisa, tudo isso, em alguma etapa, passa por um processo de classificação.

Esse processo é usado de forma explícita em setores que lidam com uma grande quantidade de dados, pois possibilita a recuperação de informação a partir deles, o que auxilia na tomada de decisão por parte do interessado (OLIVEIRA, OLIVEIRA e CIARELLI, 2013). Um professor passa por essa mesma situação: há atividades e provas (dados) de diversos alunos, as quais ele precisa atribuir uma nota (classificar). Esse caso é mais extremo em turmas de ensino a distância, como no Ambiente Virtual de Aprendizagem (AVA) Moodle, onde pode haver um número massivo de alunos pelos quais um professor é responsável. O uso de uma ferramenta capaz de realizar esse processo de forma automática iria auxiliar e acelerar consideravelmente o trabalho do professor. Em turmas de Programação, por exemplo, os códigos-fonte de atividade submetidas por alunos poderiam ser classificados quanto a sua qualidade, para que assim uma nota seja estimada automaticamente, poupando o árduo trabalho de correção do professor [citar o artigo do Elias].

Já em navegações na Web, usuários se veem inundados com tamanho conteúdo disponível para consumo, e dos mais diversos tipos, como filmes, músicas, jogos etc. Isso também é um problema para os produtores de conteúdo, que precisam saber que tipo de

conteúdo está em alta, e o perfil dos usuários que o consomem. Torna-se imprescindível, então, tanto para os usuários, quanto para os produtores de conteúdo, que sejam desenvolvidas maneiras eficientes de classificar os tipos de usuários e os tipos de conteúdo, para que cada pessoa receba o que é de seu interesse.

Uma maneira de lidar com ambos os problemas apresentados é o uso de classificadores automáticos [colocar referência]. Classificadores, no campo de *Machine Learning*, são algoritmos focados em reconhecimento de padrões, que realizam classificação baseado no que foi aprendido em sua fase de treino [citar Baeza], e podem ser aplicados nos mais diversos contextos. Um algoritmo tradicional e muito estudado, é o kNN, que percorre toda a lista de documentos fornecida em sua fase de treino, para classificar um item baseado nas classes que seus vizinhos mais próximos possuem [citar o artigo]. Porém, tal algoritmo tem seu uso prático limitado pelo seu alto custo computacional, ocasionado pela necessidade de percorrer toda essa lista a cada elemento a ser classificado [citar o kdir 15]. Este artigo propõe uma nova versão do kNN, chamada NOME DO ALGORITMO, que faz o uso de Árvore de Busca para tentar diminuir a quantidade de elementos verificados, e assim atenuar os problemas de desempenho da versão tradicional.

O artigo está subdivido em 6 seções. Na Seção 2 discutimos outros trabalhos com a mesma proposta. Já a Seção 3 descreve e explica o problema da classificação de documentos. O algoritmo proposto, assim como alguns conceitos no qual se baseia, são descritos na Seção 4. A Seção 5 é onde serão relatados os experimentos feitos, seguida pela Seção 6, que contém as conclusões obtidas.

2. Trabalhos Relacionados

3. A Classificação de Documentos

4. O Método

4.1. O *k*NN Tradicional

No caso da indexação de fontes de códigos, isso não é diferente. Também extraímos características, que neste caso são as palavras-chave da linguagem, ou linguagens, alvo.

Para que seja possível realizar a classificação de documentos, almente pelo processo de indexação, que pode ser resumido de forma simplificada como extrair as características importantes de um texto, de forma a representar as palavras como índices e assim tornar viável o processamento das mesmas. O modelo matemático escolhido para representação dos documentos foi o Modelo Vetorial (SALTON, WONG e YANG, 1975), no qual cada documento é representado como um vetor em um hiperespaço – o número de dimensões é o número de termos na base de dados.

Cada componente desse vetor será um dos termos presentes nos documentos, sendo representando por um peso atribuído (depende da métrica escolhida) ao termo, e esses pesos indicarão a relevância de tal. No algoritmo utilizado, cada termo presente nos textos é armazenado junto de, dentre outras coisas, a quantidade de aparições desse termo e a quantidade de documentos nos quais ele apareceu.

O kNN (k Nearest Neighboors) (FUKANAGA e HOSTETLER, 1975), que tem como proposta classificar um documento de acordo com as classes que seus vizinhos mais próximos possuem – lembrando que essa proximidade é vetorial, considerando cada documento como um vetor –, com o número de vizinhos mais próximos a serem considerados determinado pela variável k. O documento vai ser classificado com a classe que

tiver mais representantes próximos a tal documento. Para isso, precisamos fornecer inicialmente para o algoritmo uma base de dados com documentos e suas respectivas classes, chamada base de treino. Este é o pseudocódigo do algoritmo kNN:

[Colocar a imagem do pseudocódigo]

4.2. Árvore de Buscas

4.3. Algoritmo

5. Experimentos

6. Conclusão

Vimos que utilizar estruturas de dados do tipo Árvore é uma estratégia promissora para se acelerar o kNN, por evitar ter que percorrer toda a base de treinos durante as classificações. Com os experimentos validando o algoritmo proposto, comprovamos que essa é uma alternativa viável ao algoritmo kNN tradicional, e uma escolha a si considerar como algoritmo de classificação no geral. Ele é mais rápido, por realizar um menor número de comparações, e os resultados das classificações foram razoavelmente bons, comparados ao kNN, nas métricas utilizadas.

Como trabalho futuro, poderia ser realizada uma análise aprofundada do comportamento da árvore durante o processo de classificação, para tentar compreender as peculiaridades ~~no~~ seu funcionamento. ~~Seria~~ também de grande importância um estudo mais detalhado sobre a influência dos parâmetros ~~no resultado~~ da classificação, e, além disso, como determinar de forma ~~teórica~~ as boas escolhas de parâmetro para cada base de dados.

Tendo um algoritmo como o NOME DO ALGORITMO, apto para se trabalhar com grandes bases de dados, seu uso prático poderia ser explorado em diversos problemas pertinentes, principalmente na área de *Data Mining*, para classificação de documentos. Durante a Era da Informação, com uma quantidade quase infinita de dados disponíveis, as aplicações de algoritmos de classificação são inúmeras, e podemos ver no NOME DO ALGORITMO, uma real possibilidade de escolha por parte dos estudiosos e profissionais da área.

7. Referências