Avaliação do Desempenho do Algoritmo ML-kNN em Classificação de Textos Livres de Atividades Econômicas

Elias Oliveira
Depto. de Ciência da Informação
elias@lcad.inf.ufes.br

Patrick Marques Ciarelli
Depto. de Engenharia Elétrica
pciarelli@lcad.inf.ufes.br

Felipe Pedroni
Depto. de Informática

Wallace Favoreto Henrique
Depto. de Informática
whenrique@lcad.inf.ufes.br

Lucas Veronese
Depto. de Informática

Universidade Federal do Espírito Santo
Av. Fernando Ferrari s/n
29060-970 - Vitória,ES Brasil

ABSTRACT
Automatic text classification is still a challenging in the literature, specially for multi-label classification. In this work we evaluate the performance of the Multi-Label k-Nearest Neighbor algorithm for a multi-labeled dataset with more than 1,000 possible labels to be assigned to each one of the documents in the dataset. The results are promising.

Categories and Subject Descriptors
H.3.1 [**Information System**]: Content Analysis and Indexing

General Terms
Algorithms, Business Activities Classification

Keywords
Text classification, Machine Learning, Business Activities Classification

RESUMO
A classificação automática de textos é ainda um problema desafiador na literatura, especialmente quanto à classificações multi-rotuladas. Neste trabalho avaliamos o desempenho do algoritmo *Multi-Label k-Nearest Neighbor* quando aplicado a uma base de dados com mais de 1000 categorias possíveis de serem associadas a cada um dos documentos da base de dados. Os resultados obtidos são promissores.

Palavras-Chave
Classificação de Texto, Aprendizado de Máquina, Classificação de Atividades Econômicas

1. INTRODUÇÃO
A classificação automática de textos é em geral um problema desafiador na literatura. Essa classificação pode ser aplicada a bases de dados que possuem duas características distintas: uma onde os documentos são classificados em uma única categoria, e outra onde os documentos podem ser classificados em um número indeterminado de categorias.

Neste trabalho estudaremos o segundo tipo, também conhecido como classificação multi-rotulada. Por ter obtido resultados superiores a outros algoritmos propostos para resolução desse mesmo tipo de problema [5], o algoritmo *Multi Label k-Nearest Neighbor (ML-kNN)* foi selecionado para ser aplicado ao problema proposto. No entanto, diferentemente do que foi feito nos experimentos em [5], onde a quantidade máxima de categorias foi de 40 categorias, neste trabalho aplicaremos tal algoritmo a uma base de dados da ordem de centenas de categorias. Assim avaliaremos como o algoritmo *ML-kNN* se comporta num domínio onde exista um elevado número de categorias possíveis de serem atribuídas a cada documento.

A característica de uma base de dados multi-rotulada cujos documentos podem estar relacionados a uma grande quantidade de categorias é encontrada em um conjunto de documentos que representam descrições de atividades econômicas de empresas. Para cada atividade econômica de empresa é associada uma ou mais categorias de acordo com as definições das atividades pré-definidas na tabela Classificação Nacional de Atividades Econômicas (CNAE) [1].

Essa tabela é definida em 5 níveis distintos: seção, divisão, grupo, classe e subclasse, nesta ordem. Neste estudo, os documentos serão classificados de acordo com o último nível, o qual possui 1183 sublcasses, o que é incomum na literatura [4]. A classificação dessas empresas tem como objetivo a arrecadação de tributos das mesmas, além de servir para realizar análises estatísticas do setor econômico do país. Além disso, existe uma grande demanda de classificação anual (aproximadamente 1,5 milhões) [2], pois muitas empresas são abertas ou alteram suas atividades econômicas ao longo de cada ano.

Este trabalho está organizado na seguinte estrutura. Na Seção 2 iremos detalhar o funcionamento do algoritmo *ML-kNN*. Na Seção 3 será comentado quais foram os experimentos realizados e os resultados obtidos. Por fim, na Seção 4, a conclusão do trabalho é apresentada.

2. ALGORITMO ML-KNN

O *ML-kNN* é um classificador multi-label construído baseado no popular método *kNN* [5]. Para cada d_j elemento de teste, o *ML-kNN* inicialmente encontra seus k vizinhos mais próximos no conjunto de treino usado, isto é, encontra os k primeiros elementos ordenados pelo valor de similaridade com d_j de forma decrescente, usando por exemplo, a distância Euclidiana. Mais tarde, o algoritmo identifica quantos exemplares de cada categoria existem dentre os k vizinhos mais próximos de d_j, que chamaremos de k_i ($i \in \{1, 2, \ldots, |C|\}$), onde $|C|$ é o número de categorias. Seja H_1^i o evento em que d_j possui o rótulo i e H_0^i o evento em que d_j não possui o rótulo i. E mais, seja E_j^i o evento em que existem j vizinhos mais próximos de d_j pertencentes à categoria i. Assim, temos:

$$y_{d_j}(i) = argmax_{b \in 0,1} P(H_b^i) P(E_j^i | H_b^i) \quad (1)$$

Na Equação 1, $y_{d_j}(i)$ é a probabilidade da amostra d_j pertencer à categoria i. $P(H_b^i)$, onde ($i \in \{1, 2, \ldots, |C|\}$ e $b \in \{0, 1\}$) e $P(E_j^i | H_b^i)$ ($j \in \{0, 1, \ldots, k\}$) são, respectivamente, a probabilidade *priori* e *posteriori* da categoria i. Essas probabilidades são calculadas na etapa de treino. Primeiramente é estimada a probabilidade *priori* de cada categoria usando a seguinte equação:

$$P(H_1^i) = \frac{\delta + N_i}{2\delta + N} \qquad P(H_0^i) = 1 - P(H_1^i) \quad (2)$$

N_i denota, na Equação 2, o número de exemplares da categoria i no conjunto de treino e N denota o número total de exemplares. δ é um parâmetro para suavizar a probabilidade.

Após isso, o *ML-kNN* faz o seguinte. Para cada exemplar w_y no conjunto de treino, onde ($y \in \{1, 2, \ldots, N\}$), o algoritmo encontra seus k vizinhos mais próximos e calcula o número total de votos que cada categoria recebe dos k vizinhos mais próximos. Em outras palavras, seja k_i o número de votos que cada categoria i recebeu de w_y, se o exemplar w_y pertence à categoria i, então será adicionado 1 a $L_{k_i}^i$, senão será adicionado 1 a $\overline{L_{k_i}^i}$. $L_{k_i}^i$ e $\overline{L_{k_i}^i}$ indicam quantos exemplares de treino estão relacionados com a categoria i, e respectivamente, não relacionados com a categoria i. Finalmente com essas informações, as probabilidades *posteriori* são calculadas como descrito nas Equações 3 e 4:

$$P(E_j^i | H_1^i) = \frac{\delta + L_j^i}{\delta(k+1) + \sum_{o=0}^{k} L_o^i} \quad (3)$$

$$P(E_j^i | H_0^i) = \frac{\delta + \overline{L_j^i}}{\delta(k+1) + \sum_{o=0}^{k} \overline{L_o^i}} \quad (4)$$

O *ML-kNN* precisa apenas de dois parâmetros: o número k de vizinhos mais próximos e a suavização δ da probabilidade. Nas Equações 2, 3 e 4 o valor de δ modifica levemente as probabilidades *priori* e as probabilidades *posteriori*.

3. EXPERIMENTOS E RESULTADOS

A base de dados utilizada no experimento é composta de 3281 documentos que representam descrições de atividades econômicas de empresas da cidade de Vitória-ES e também é composta por 1183 definições de subclasse da tabela CNAE. Houve uma preparação inicial na base de dados antes que os experimentos pudessem ser realizados. Nessa preparação cada documento da base de dados foi submetido a um processo de *stemming*, fazendo com que as palavras nos documentos ficassem sem gênero, número e grau e também a um processo de retirada de *stop words*, isto é, retirada de artigos, preposições, conjunções, números e outras palavras que apenas prejudicariam a caracterização do documento. Após isso, cada documento foi representado como um vetor no espaço R^n, onde n é o número total de termos encontrados no conjunto de documentos, como feito em [3].

Os experimentos foram divididos em duas etapas: a etapa de validação e a etapa de teste. Na etapa de validação foram utilizadas 1183 definições de subclasse juntamente com 820 descrições de atividades econômicas no treinamento do algoritmo. Após o treinamento a validação do algoritmo (obtenção do melhor valor do parâmetro k) foi realizada utilizando outras 820 descrições de atividades econômicas. A seguir foi realizada a etapa de teste, onde o *ML-kNN* foi treinado com 1183 definições de subclasses e com 1640 documentos e foi testado com outros 1641 documentos.

Os resultados foram analisados utilizando as métricas *Coverage*, *One Error*, *Average Precision* e *Ranking Loss* definidas em [5]. As métricas *One Error*, *Ranking Loss* e *Average Precision* são definidas no intervalo de 0 a 1. A métrica *Coverage* possui limite inferior igual a 0 e limite superior igual a $|C|$ - 1. No entanto, para uma melhor representação gráfica essa métrica foi normalizada na faixa de 0 a 1. Para o entendimento dos resultados obtidos é importante destacar que quanto menor o valor encontrado para as métricas *Coverage*, *One Error* e *Ranking Loss* melhor é o resultado. Já em relação a métrica *Average Precision* quanto maior o valor melhor o resultado. Os resultados obtidos são mostrados a seguir na Figura 1.

Considerando o grande número de categorias presentes na base de dados, foram obtidos resultados expressivos nas métricas de desempenho utilizadas. Analisando os resultados obtidos e comparando com o resultado apresentado em [3], muito embora as métricas utilizadas sejam diferentes, intuitivamente percebe-se que o *ML-kNN* retornou resultados melhores que o algoritmo *Vizinho Mais Próximo*.

4. CONCLUSÃO

Este trabalho teve como objetivo avaliar o desempenho do algoritmo *ML-kNN* quando submetido a classificação de uma base de dados com uma grande quantidade de categorias, e foi constatado que ele pode ser aplicado à resolução de tal tipo de problema. Como trabalhos futuros é previsto realizar não apenas a comparação exata entre o algoritmo *ML-kNN* e o algoritmo *Vizinho Mais Próximo*, mas também pretendemos fazer comparações do algoritmo *ML-kNN* com alguns outros algoritmos de classificação multi-rotulada.

No entanto, apesar do *ML-kNN* suportar uma grande quantidade de categorias, um estudo que também é muito perti-

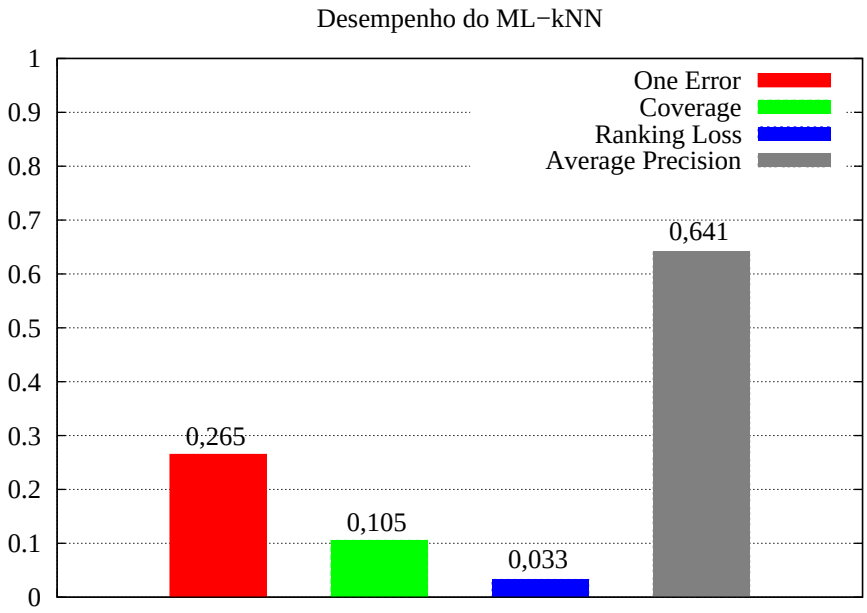

Figura 1: Resultados experimentais obtidos com o ML-kNN.

nente diz respeito a incorporar ao *ML-kNN* algumas técnicas para viabilizar a classificação de uma base de dados independentemente do número de documentos de que ela seja composta.

5. AGRADECIMENTOS

Gostaríamos de agradecer a Andréa Pimenta Mesquita, coordenadora da classificação CNAE no setor da cidade de Vitória, por disponibilizar a base de dados usada neste trabalho. Gostaríamos de agradecer também a Min-Ling Zhang por toda ajuda com a ferramenta de categorização ML-kNN. Este trabalho é parcialmente financiado pela *Receita Federal do Brasil*, o CNPq através das bolsas (308207/2004-1, 471898/2004-0, 620165/2006-5), *Financiadora de Estudos e Projetos*—FINEP-Brasil (bolsas CT-INFRA-PRO-UFES/2005, CT-INFRA-PRO-UFES/2006) e *Fundação de Apoio à Ciência e Tecnologia do Espírito Santo*—FAPES-Brasil (bolsa 41936450/2008).

6. REFERÊNCIAS

[1] CNAE. Classificação nacional de atividades econômicas - cnae 1.0 / cnae-fiscal 1.1. Technical report, Instituto Brasileiro de Geografia e Estatística (IBGE), 2003.

[2] DNRC. *Ranking* das Juntas Comerciais Segundo Movimento de Constituição, Alteração e Extinção e Cancelamento de Empresas, 2008. Ministério do Desenvolvimento, Indústria e Comércio Exterior – Secretaria do Desenvolvimento da Produção, Departamento Nacional de Registro do Comércio (DNRC).

[3] E. Oliveira, P. M. Ciarelli, W. F. Henrique, L. Veronese, F. Pedroni, and A. F. D. Souza. Intelligent classification of economic activities from free text desciptions. *V Workshop em Tecnologia da Informação e da Linguagem Humana - TIL*, 2007.

[4] F. Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[5] M.-L. Zhang and Z.-H. Zhou. ML-KNN: A Lazy Learning Approach to Multi-Label Learning. *Pattern Recogn.*, 40(7):2038–2048, 2007.