CATEGORIZAÇÃO AUTOMÁTICA DE ARTIGOS CIENTÍFICOS DA ENGENHARIA DE PRODUÇÃO UTILIZANDO MÉTODOS DE APRENDIZAGEM DE MÁQUINA

FERNANDO JOSE FERREIRA ANDINOS JUNIOR (UCAM)
ferandinos@gmail.com
GEOGIA REGINA RODRIGUES GOMES (UCAM)
georgia@ucam.campos.br

Este artigo apresenta três métodos tradicionais de aprendizagem de máquina (Naïve Bayes, k-NN e SVM) e propõe um método de grupo para realizar a categorização de artigos da área de Engenharia de Produção, com o objetivo de auxiliar alunos ee professores na escolha da melhor área para submissão de trabalhos. Para isso, os métodos utilizados baseiam-se no conteúdo textual do documento, tendo como insumo de aprendizagem, artigos previamente publicados em anais de dois dos principais congressos de Engenharia de Produção, o ENEGEP e o SIMPEP. Baseado nos resultados experimentais apresentados, o método de grupo proposto obteve melhor desempenho nas métricas definidas (acurácia, precisão e abrangência) que os métodos tradicionais isoladamente. Os principais fatores para a elaboração desse trabalho foram a dificuldade exposta por alunos e professores em algumas vezes escolher a área de submissão mais adequada para seus trabalhos, somado ao crescimento observado no número de artigos publicados nesses congressos nos últimos anos. Espera-se que este trabalho contribua para o crescimento, organização e qualidade da produção científica em Engenharia de Produção no Brasil.

Palavras-chaves: categorização de documentos, mineração de texto, gestão do conhecimento

1. Introdução

O Brasil atualmente possui 486 cursos de graduação em Engenharia de Produção reconhecidos pelo MEC (NUPENGE, 2012) e 58 cursos de pós-graduação *strictu-senso* recomendados pela CAPES (CAPES, 2012). Além de atender a demanda crescente do mercado de trabalho, boa parcela desses indivíduos contribui para produção científica, gerada principalmente por professores e alunos de pós-graduação.

A escolha da melhor área para submissão de artigos científicos em congressos de uma área abrangente e multidisciplinar como a Engenharia de Produção, que atualmente divide-se em 11 áreas de conhecimento, subdivididas em 58 subáreas conforme ABEPRO (2012), pode não ser trivial.

Diante disso, professores e alunos em alguns momentos demonstram dificuldade em decidir a área mais adequada. Então, se existisse uma ferramenta que baseada no conteúdo textual, os auxiliasse sugerindo a área mais apropriada para submissão do artigo, a probabilidade de aceitação aumentaria, pois seriam direcionados a avaliadores mais indicados. Além disso, uma vez aprovado e categorizado na área mais aderente ao seu conteúdo, o trabalho teria melhor divulgação e atingiria o público esperado pelos autores.

O objetivo deste trabalho é utilizar técnicas de aprendizagem de máquina (*machine learning*), ramo da inteligência artificial responsável por desenvolver métodos que permitam ao computador aprender, para que a partir de artigos previamente categorizados, consiga-se predizer a categoria de novos artigos, auxiliando os autores na escolha da melhor área de submissão em congressos da Engenharia de Produção.

Este trabalho está organizado em cinco seções seguidas de bibliografia. Na seção 2, é apresentado o conceito de categorização automática de textos. A seção 3 descreve a metodologia com os passos realizados no experimento e as métricas de desempenho utilizadas. Na seção 4, os resultados são apresentados e analisados e a seção 5 apresenta as conclusões.

2. Categorização de textos

A categorização de textos é a atribuição de documentos escritos em linguagem natural em categorias pré-definidas de acordo com o seu conteúdo (SEBASTIANI, 2002). Apesar do estudo da categorização automática de textos ter iniciado nos anos 60 com Maron e Kuns (1961), a partir da década de 90 que esse campo vem se desenvolvendo devido ao crescente número de documentos digitais, viabilizado pelo surgimento da *World Wide Web*, gerando a necessidade de organizá-los para facilitar seu acesso e manuseio.

Existem duas principais abordagens para a categorização de textos: uma é conhecida como engenharia do conhecimento (*knowledge engineering*), onde o próprio especialista codifica o sistema através de regras que definem cada categoria da coleção de documentos, como a que foi utilizada no desenvolvimento da ferramenta CADWeb (CADWeb, 2012) por Gomes e Moraes Filho (2011), e outra, utilizada neste trabalho, que usa técnicas de aprendizagem de máquina. Nessa abordagem, o classificador é construído automaticamente, aprendendo as

propriedades das categorias a partir de um conjunto de documentos de treinamento previamente classificados (FELDMAN; SANGER, 2007). No conceito de aprendizagem de máquina, esse processo é chamado de aprendizado supervisionado.

Segundo Sebastiani (2002), as vantagens dessa abordagem são: precisão comparável às atingidas pelos especialistas com consideráveis economias de mão-de-obra, pois não existe a necessidade de intervenção humana para a construção do classificador ou adaptação para outro domínio de conhecimento. Existem diversos algoritmos classificadores utilizados na tarefa de categorização de textos, este trabalho utilizará três dos principais: Naïve Bayes, k-NN (*k-nearest Neighbor*) e SVM (*Support Vector Machines*), devido tratar-se de algoritmos com resultados comprovadamente satisfatórios e utilizar métodos distintos para tratar o problema de categorização. Propõe-se também, um método de grupo, combinando os métodos anteriores em um esquema de votação. A seguir descrevem-se cada um deles.

2.1 Naïve Bayes

O Naïve Bayes, é um classificador probabilístico baseado no teorema de Bayes, definido na equação (1).

$$P(c_i \mid \vec{d}\,) = P(c_i)\frac{P(\vec{d} \mid c_i)}{P(\vec{d}\,)} \tag{1}$$

Esse tipo de classificador computa a probabilidade de um documento $\vec{d}$ pertencer à classe c_i, assumindo que a presença de um termo em uma categoria não está condicionada a presença de qualquer outro. Devido à independência dos termos, apenas as variações para cada classe necessita de ser determinada, e não a matriz de covariância completa (ZHANG, 2004). Segundo Domingos e Pazzani (1997), a independência de termos na maioria dos casos não prejudica a eficiência do classificador.

2.2 k-NN

O k-NN, é a base dos algoritmos conhecidos como preguiçosos (*lazy algorithms*). Ele armazena todo conjunto de treinamento e empenha todo o esforço em direção à generalização indutiva até o momento da classificação (WETTSCHERECK; AHA; MOHRI, 1997). Esse classificador representa cada exemplo como um ponto de dado em um espaço *d*-dimensional, onde *d* é o número de atributos. Dado um exemplo de teste, calcula-se a proximidade com o resto dos pontos de dados no conjunto de treinamento usando uma função de proximidade (TAN; STEINBACH; KUMAR, 2009). Exemplos de função de proximidade são: correlação, distância euclidiana, medida de semelhança de Jaccard e co-seno, sendo as duas últimas mais indicadas para lidar com dados de alta dimensionalidade, que é o caso de documentos.

2.3 SVM

O SVM constitui uma técnica baseada na teoria do aprendizado estatístico, baseado no princípio de minimização do risco estrutural introduzido por Vapnik (2000). O objetivo desse algoritmo é encontrar o hiperplano de separação linear ótimo entre duas classes, maximizando

a margem entre seus pontos mais próximos. O hiperplano de classificação é escolhido durante a fase de treinamento como o único que separa as instâncias positivas conhecidas das instâncias negativas com a margem máxima entre elas (FELDMAN; SANGER, 2007). Os exemplos mais próximos do hiperplano são chamados vetores de suporte (*support vectors*). A Figura 1 ilustra esses conceitos apresentando um exemplo de duas classes linearmente separáveis.

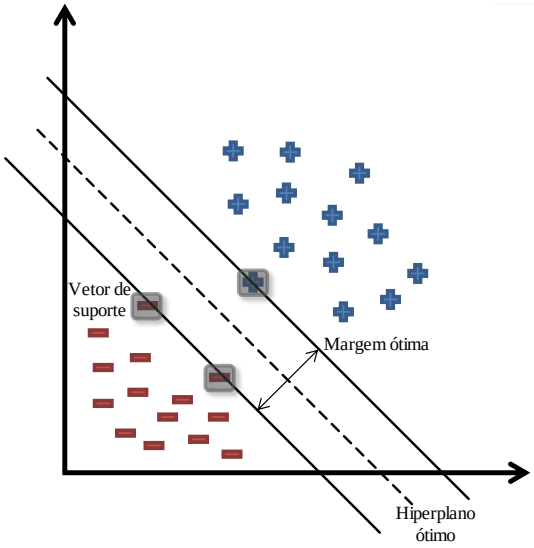

Figura 1 – Classes separadas linearmente em um espaço bi-dimensional. Os vetores de suporte, em cinza, definem a margem de maior separação entre as categorias
Fonte: Adaptado de Cortes e Vapnik (1995)

Para lidar com casos onde os exemplos de treinamento não são completamente separáveis e um pequeno erro de classificação é permitido, utiliza-se o conceito de margens suaves (*soft margins*), que introduz um parâmetro de custo *C*, especificado pelo próprio usuário que determina o nível aceitável de tolerância a erros (BERRY; KOGAN, 2010). Outro parâmetro importante a ser configurado é o critério de parada ε (epsilon), para evitar um loop infinito na busca pelo hiperplano ótimo. Neste trabalho, utiliza-se a implementação LIBSVM, criada por Chang e Lin (2011).

2.4 Método de Grupo

O objetivo desta técnica é melhorar o desempenho da classificação agregando a previsão de múltiplos classificadores. Segundo Feldman e Sanger 2007, para obter bons resultados, os classificadores devem ser significativamente diferentes, seja na representação dos documentos ou no método de aprendizagem. Neste trabalho, é proposto um método de grupo utilizando-se dos três métodos descritos anteriormente.

Em cada método, gera-se um valor de confiança *Conf*, para cada par $(\vec{d}, c_i)$, sendo $\vec{d}$, o documento e c_i, a categoria. A confiança é um valor normalizado dentro do intervalo [0,1] que representa o nível de certeza de uma determinada predição. Em outras palavras,

representa o nível de pertinência de um documento $\vec{d}$ à categoria c_i, segundo um classificador em particular. Esse valor é calculado de forma distinta para cada método, não cabendo ao escopo deste trabalho aprofundar neste processo.

No método proposto, a categoria atribuída ao documento é aquela com maior soma das confianças nos três métodos *m*, utilizados. Para cada método, a confiança é multiplicada por um peso *w*, proporcional à posição do classificador no ranking gerado após a etapa de otimização de parâmetros e avaliação preliminar. Essa possibilidade de combinação de classificadores é mencionada em Feldman e Sanger 2007. A equação (2) demonstra matematicamente o cálculo da pontuação C_i, de cada categoria c_i e a equação (3), a atribuição da categoria que obtiver maior pontuação ao documento $\vec{d}$.

$$C_i = \sum_{m=1}^{3} \left(Conf(\vec{d}, c_i) \cdot w_m \right) \tag{2}$$

$$c(\vec{d}) \leftarrow \max(C_1, C_2, ..., C_{11}) \tag{3}$$

3. Metodologia

O experimento foi realizado através de cinco etapas descritas nas próximas subseções. O software *open source Rapidminer*, criado por Mierswa et al (2006), foi utilizado como ferramenta principal ao longo do trabalho.

Utilizou-se neste trabalho, 4336 artigos em língua portuguesa publicados em edições anteriores do ENEGEP (ABEPRO, 2011) e SIMPEP (SIMPEP, 2011), sendo 3408 para treinamento dos classificadores e 928 para testes. A Tabela 1 apresenta a distribuição dos documentos nas onze categorias.

Categorias	Nº Artigos
1 GESTÃO DA PRODUÇÃO	741
2 GESTÃO DA QUALIDADE	489
3 GESTÃO ECONÔMICA	331
4 ERGONOMIA E SEGURANÇA DO TRABALHO	397
5 GESTÃO DO PRODUTO	297
6 PESQUISA OPERACIONAL	358
7 GESTÃO ESTRATÉGICA E ORGANIZACIONAL	544
8 GESTÃO DO CONHECIMENTO ORGANIZACIONAL	532
9 GESTÃO AMBIENTAL	310
10 EDUCAÇÃO EM ENGENHARIA DE PRODUÇÃO	194
11 ENG. PROD., SUSTENTABILIDADE E RESPONSABILIDADE SOCIAL	143
Total	**4336**

Tabela 1 – Distribuição dos documentos por categoria

3.1 Pré-processamento dos documentos

O objetivo dessa etapa é representar os documentos de forma que eles possam ser processados pelos algoritmos de aprendizagem. Primeiramente todos os documentos adquiridos em formato PDF foram transformados em texto simples, através do software *Some PDF to TXT converter v1.0* (FREE ...2011). A motivação foi o ganho em desempenho na execução dos algoritmos, na ordem de vinte vezes, aproximadamente. Após essa conversão, se manteve apenas o conteúdo textual dos documentos. Figuras e opções de formatação foram automaticamente ignoradas.

Em seguida, foi necessário excluir o tema do congresso para os documentos onde o mesmo foi identificado no corpo do texto, pois são termos que não representam com fidelidade seu conteúdo.

Como os algoritmos de aprendizagem de máquina são incapazes de processar documentos em seu formato original, durante essa etapa, realizou-se a representação dos documentos em vetores de características. O tipo mais comum de representação é chamado *bag of words* (saco de palavras), que utiliza todos os termos do documento como características. Dessa forma, a dimensão do espaço de características é igual ao número de termos diferentes encontrados em todos os documentos. Existem várias formas de atribuir pesos aos termos. Neste trabalho utiliza-se a frequência do termo normalizada, modelada matematicamente nas equações (4) e (5):

$$F_i = \frac{O_i}{T} \tag{4}$$

$$FN_i = \frac{F_i}{\sqrt{(F_i)^2 + (F_{i+1})^2 + \ldots + (F_{i+n})^2}} \tag{5}$$

Onde: F_i : Frequência do termo i;

O_i : Ocorrências do termo i;

T : Número de termos no documento;

FN_i : Frequência do termo normalizada.

Antes de efetivamente gerar o vetor de características para cada documento, cinco processos são executados sequencialmente com o objetivo de reduzir a dimensão do espaço de representação dos documentos:

- *Case folding*: Esse processo é responsável por transformar todas as letras dos termos em minúsculas;
- Remoção de *stopwords*: O objetivo deste processo é remover termos que não apresentam um conteúdo semântico significativo no contexto em que se apresentam no documento. Geralmente trata-se de palavras auxiliares ou conectivas (por exemplo: a, de, aos, com), que não fornecem nenhuma informação que venha a representar

conteúdo dos documentos. A exclusão se dará com base em um arquivo texto com a lista dos termos;

- *Prunning*: Especifica critérios de eliminação. Neste trabalho, obtiveram-se melhores resultados ignorando termos que ocorrem em menos de 4% dos documentos e em mais de 99% dos documentos. Foram removidas também as palavras com menos de cinco letras;
- *Stemming*: O objetivo deste processo é a remoção do sufixo e prefixo dos termos que possam vir a representar uma variação verbal ou plural, gerando apenas os radicais de acordo com as regras gramaticais da língua utilizada. Por exemplo: os termos computação, computador e computar são transformados em *comput*. A principal finalidade desse processo é a redução do espaço dimensional. Neste trabalho, utilizou-se o algoritmo de Porter adaptado para a língua portuguesa na linguagem *snowball*, criada pelo próprio Porter. Informações sobre ao algoritmo podem ser obtidas em Willet (2006) e sobre a linguagem *snowball* em Porter (2011).

A Figura 2 ilustra a sequência de processos modeladas no *RapidMiner*.

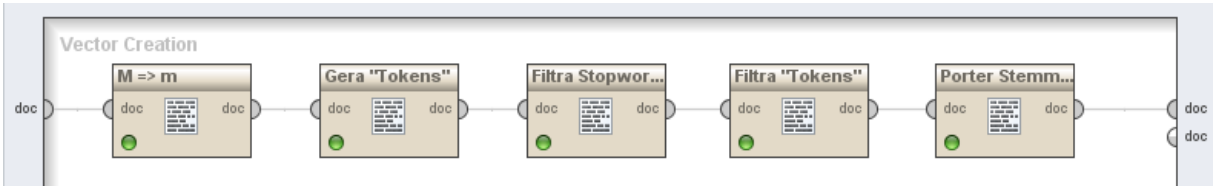

Figura 2 – Etapas do pré-processamento modeladas no *Rapidminer*

Ao término dessa etapa, obtiveram-se todos os 4336 representados no modelo *bag of words*, com uma redução do espaço dimensional na ordem de 50,3%.

3.2 Otimização de parâmetros e avaliação preliminar dos algoritmos

Primeiramente, definiu-se-se as métricas utilizadas na avaliação do desempenho dos métodos: acurácia (*accuracy*), precisão (*precision*), abrangência (*recall*) e F_1. A acurácia, *a*, é a medida mais básica de eficiência do classificador, sendo a fração de documentos corretamente categorizados. Abrangência, *r*, é definida como a fração dos documentos de uma categoria corretamente classificados. Precisão, *p*, é definida como a fração de documentos corretamente classificados dentre todos os documentos atribuídos pelo classificador a uma categoria. Portanto, uma abrangência perfeita é alcançada caso todos os documentos da categoria em questão sejam nela classificados, independentemente se outros documentos de outras categorias sejam também atribuídos a ela. Por outro lado, uma boa precisão é alcançada ao evitar que documentos provenientes de diferentes categorias sejam atribuídos a uma só. Em virtude da variedade de aspectos de avaliação, uma abordagem mais usual para avaliar o desempenho da categorização é F_1, uma combinação entre precisão e abrangência, dada pela

média harmônica dessas duas métricas. As equações (6), (7), (8) e (9) definem as métricas citadas anteriormente:

$$a = \frac{DC}{TD} \tag{6}$$

$$p = \frac{VP}{VP + FP} \tag{7}$$

$$r = \frac{VP}{VP + FN} \tag{8}$$

$$F_1 = \frac{2}{\left(\left(\frac{1}{p}\right) + \left(\frac{1}{r}\right)\right)} \tag{9}$$

Onde: DC : documentos corretamente categorizados;

TD : total de documentos;

VP : verdadeiro-positivos;

FP : falso-positivos;

FN : falso-negativos.

Com os documentos devidamente representados no modelo *bag of words*, o próximo passo foi utilizar os 3408 documentos separados inicialmente para otimização dos parâmetros e treinamento dos algoritmos de forma a maximizar sua acurácia, estimando assim o desempenho do classificador quando apresentados ao conjunto de teste. A busca pelos parâmetros ótimos deu-se de forma empírica utilizando duas técnicas em conjunto para auxiliar neste processo: A busca por força bruta (*grid search*), basicamente uma forma automática de variar um parâmetro dentro de uma faixa pré-estabelecida de valores incrementada por alguma função, e a validação cruzada, que consiste em dividir o conjunto de treinamento em x subconjuntos de igual tamanho, testando sequencialmente cada subconjunto no classificador treinado com os elementos dos subconjuntos $x-1$ restantes (HSU; CHANG; LIN, 2010). Neste trabalho realizou-se a busca por força bruta utilizando validação cruzada com $x=10$ e a acurácia obtida foi armazenada. Esse processo foi utilizado na escolha do valor de k (número de vizinhos) para o algoritmo k-NN, e do valor de C (nível de tolerância a erros) e ε (critério de parada) para o SVM. O classificador Naïve Bayes não necessita de customização de parâmetros.

Para o algoritmo k-NN, o objetivo foi encontrar o valor de *k* que maximizasse a acurácia do modelo. Para isso, realizou-se uma busca por força bruta com 50 valores de *k* em escala logarítmica na faixa de 1 a 100 utilizando a validação cruzada para avaliação de cada iteração. Ao final do processo, chegou-se ao número de 23 vizinhos. A Figura 3 demonstra o resultado desse processo.

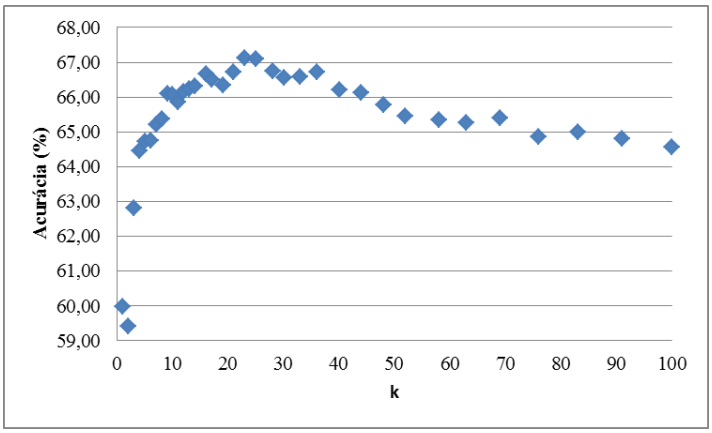

Figura 3 – Resultado do processo de busca pelo valor de *k* do algoritmo k-NN

Diferentemente do processo de busca por força bruta utilizado para encontrar o melhor valor de *k*, do classificador k-NN, para o classificador SVM não se utilizou todo o conjunto de treinamento devido ao alto custo computacional e tempo exigido para a conclusão do processo. Por esta razão, optou-se por utilizar 10% do conjunto de treinamento, respeitando o balanceamento entre as categorias. Conforme sugerido por Hsu, Chang e Lin, 2010, os valores de C e ε variaram exponencialmente da seguinte forma: $C = \{2^{-5}, 2^{-3}, \ldots, 2^{9}\}$ e $\varepsilon = \{2^{-15}, 2^{-13}, \ldots, 2^{1}\}$. Apesar do resultado da busca apontar para os valores $C = 2$ e $\varepsilon = 0{,}00003$, quando se utilizou todo o conjunto de treinamento, a acurácia foi menor que a obtida com os valores padrões ($C = 0$ e $\varepsilon = 0{,}001$), portanto, esses últimos foram utilizados. A tabela com o resultado deste processo de busca pode ser consultado no APENDICE 1.

3.3 Geração dos modelos de classificação

Superada a otimização e avaliação preliminar, gerou-se o modelo de cada classificador com os parâmetros ótimos e com os 3408 documentos de treinamento servindo como base de aprendizagem. A Figura 4 ilustra esse processo modelado no *Rapidminer*.

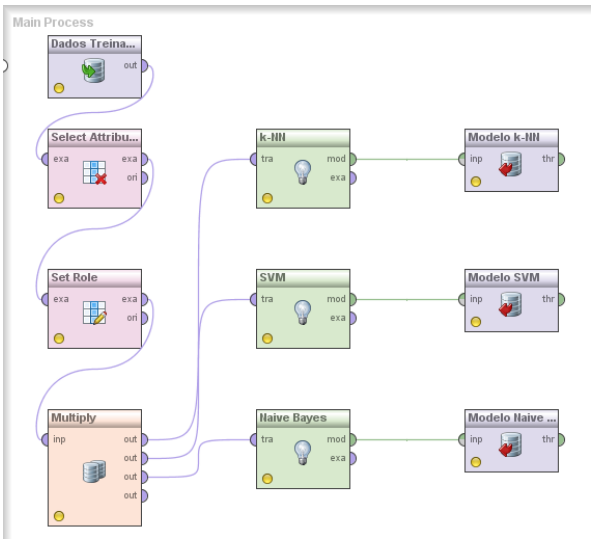

Figura 4 – Modelagem feita no *RapidMiner* para geração dos classificadores

Observa-se que o mesmo conjunto de treinamento é utilizado em todos os classificadores.

3.4 Testes

Na etapa de testes, realizou-se a classificação propriamente dita, onde os 928 documentos separados para essa finalidade foram submetidos aos três classificadores, e o resultado de cada classificação, incluindo a confiança para cada par $(\vec{d}, c_i)$, gravado em um arquivo. Vale ressaltar que esses documentos não foram utilizados na etapa de otimização e avaliação preliminar dos algoritmos. Na Figura 5, o modelo construído para o teste de classificação é apresentado.

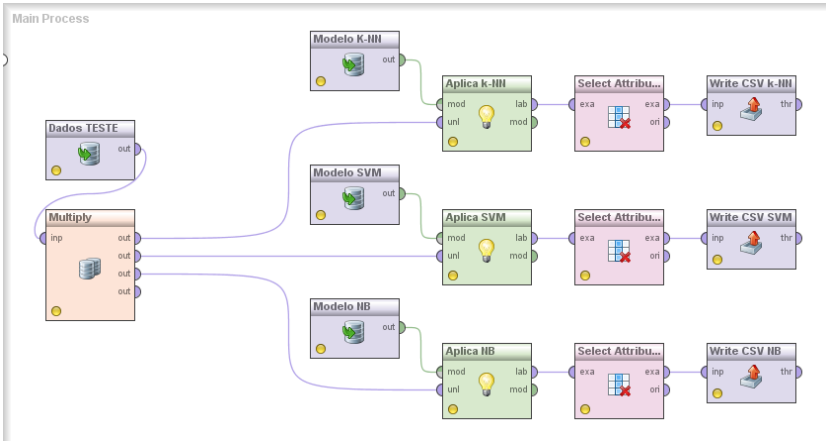

Figura 5 – Modelagem feita no *RapidMiner* para classificação dos documentos de teste

Além da classificação, realizou-se um experimento para verificar o grau de similaridade entre os documentos das onze categorias da Engenharia de Produção servindo como insumo de avaliação do desempenho dos classificadores, utilizando a medida de proximidade similaridade do co-seno. Neste experimento, combinaram-se documentos de cada área para representá-la, de forma a utilizar-se de pelo menos um documento de cada subárea. Como base de comparação, o mesmo experimento foi realizado com três artigos de outras áreas do conhecimento completamente distintas: Odontologia, Direito e Informática.

4. Resultados e Discussões

Esta seção apresenta os resultados obtidos desde a etapa de otimização de parâmetros e avaliação preliminar até os resultados de cada classificador quando confrontados com o conjunto de teste.

Na Tabela 2, é apresentado o resultado de cada classificador na etapa de otimização de parâmetros e avaliação preliminar.

Categorias	SVM			k-NN			Naïve Bayes		
	Precisão	*Abrangência*	F_1	*Precisão*	*Abrangência*	F_1	*Precisão*	*Abrangência*	F_1
1 GESTÃO DA PRODUÇÃO	0,63	0,60	**0,62**	0,59	0,57	**0,58**	0,58	0,54	**0,56**
2 GESTÃO DA QUALIDADE	0,75	0,75	**0,75**	0,69	0,79	**0,74**	0,68	0,75	**0,72**
3 GESTÃO ECONÔMICA	0,68	0,71	**0,70**	0,69	0,66	**0,68**	0,68	0,66	**0,67**
4 ERGONOMIA E SEGURANÇA DO TRABALHO	0,88	0,90	**0,89**	0,86	0,88	**0,87**	0,89	0,84	**0,87**
5 GESTÃO DO PRODUTO	0,77	0,74	**0,75**	0,66	0,69	**0,68**	0,65	0,72	**0,68**
6 PESQUISA OPERACIONAL	0,72	0,78	**0,75**	0,77	0,68	**0,72**	0,75	0,64	**0,69**
7 GESTÃO ESTRATÉGICA E ORGANIZACIONAL	0,65	0,63	**0,64**	0,63	0,62	**0,62**	0,66	0,51	**0,58**
8 GESTÃO DO CONHECIMENTO ORGANIZACIONAL	0,68	0,73	**0,70**	0,61	0,75	**0,67**	0,51	0,80	**0,62**
9 GESTÃO AMBIENTAL	0,71	0,73	**0,72**	0,72	0,68	**0,70**	0,76	0,54	**0,64**
10 EDUCAÇÃO EM ENGENHARIA DE PRODUÇÃO	0,72	0,68	**0,70**	0,72	0,53	**0,61**	0,75	0,54	**0,63**
11 ENG. PROD., SUSTENTABILIDADE E RESP. SOCIAL	0,58	0,38	**0,46**	0,70	0,27	**0,39**	0,48	0,43	**0,45**
Acurácia		**71,10%**			**68,12%**			**65,61%**	

Tabela 2 – Resultado do processo de otimização e avaliação preliminar dos classificadores

De acordo com os resultados obtidos nessa etapa, o classificador SVM obteve melhor acurácia, com 71,10%, o k-NN foi o segundo colocado com 68,12% e o Naïve Bayes o terceiro com 65,61%. A partir desses resultados definiram-se os pesos w_m para o método de grupo proposto como sendo: $w = 2$ para o SVM, $w = 1{,}5$ para o k-NN e $w = 1$ para o Naïve Bayes.

Observa-se na Tabela 3, que de forma geral se obteve um desempenho inferior dos três classificadores em todas as métricas quando apresentados ao conjunto de teste comparado aos valores obtidos na etapa de otimização e avaliação preliminar. Pequenas diferenças entre o desempenho estimado e o real são comuns na maioria dos casos. O classificador SVM, apesar de manter-se como o de melhor desempenho, apresentou a maior queda na acurácia entre as etapas (5,48%). O classificador k-NN obteve a menor variação entre o desempenho estimado e o real.

Categorias	SVM			k-NN			Naïve Bayes		
	Precisão	*Abrangência*	F_1	*Precisão*	*Abrangência*	F_1	*Precisão*	*Abrangência*	F_1
1 GESTÃO DA PRODUÇÃO	0,87	0,52	**0,65**	0,83	0,54	**0,65**	0,82	0,50	**0,62**
2 GESTÃO DA QUALIDADE	0,64	0,77	**0,70**	0,61	0,77	**0,68**	0,57	0,73	**0,64**
3 GESTÃO ECONÔMICA	0,60	0,73	**0,66**	0,53	0,66	**0,59**	0,57	0,66	**0,61**
4 ERGONOMIA E SEGURANÇA DO TRABALHO	0,69	0,89	**0,78**	0,69	0,81	**0,75**	0,72	0,87	**0,79**
5 GESTÃO DO PRODUTO	0,53	0,69	**0,60**	0,49	0,85	**0,62**	0,40	0,65	**0,50**
6 PESQUISA OPERACIONAL	0,40	0,53	**0,46**	0,51	0,47	**0,49**	0,44	0,40	**0,42**
7 GESTÃO ESTRATÉGICA E ORGANIZACIONAL	0,61	0,70	**0,65**	0,59	0,71	**0,65**	0,59	0,54	**0,56**
8 GESTÃO DO CONHECIMENTO ORGANIZACIONAL	0,66	0,80	**0,72**	0,64	0,83	**0,72**	0,51	0,87	**0,64**
9 GESTÃO AMBIENTAL	0,63	0,83	**0,71**	0,63	0,78	**0,70**	0,72	0,72	**0,72**
10 EDUCAÇÃO EM ENGENHARIA DE PRODUÇÃO	0,44	0,71	**0,55**	0,47	0,47	**0,47**	0,42	0,47	**0,44**
11 ENG. PROD., SUSTENTABILIDADE E RESP. SOCIAL	0,72	0,39	**0,51**	0,81	0,37	**0,51**	0,71	0,52	**0,60**
Acurácia		**65,62%**			**64,87%**			**61,53%**	

Tabela 3 – Resultado dos classificadores com o conjunto de teste

A Tabela 4 apresenta o resultado do classificador de grupo proposto, que obteve melhor desempenho em todas as métricas utilizadas, superando o desempenho individual do método SVM.

Categorias	Método de Grupo		
	Precisão	*Abrangência*	F_1
1 GESTÃO DA PRODUÇÃO	0,91	0,60	**0,72**
2 GESTÃO DA QUALIDADE	0,68	0,87	**0,76**
3 GESTÃO ECONÔMICA	0,72	0,89	**0,80**
4 ERGONOMIA E SEGURANÇA DO TRABALHO	0,77	0,96	**0,86**
5 GESTÃO DO PRODUTO	0,65	0,77	**0,70**
6 PESQUISA OPERACIONAL	0,57	0,60	**0,59**
7 GESTÃO ESTRATÉGICA E ORGANIZACIONAL	0,71	0,81	**0,76**
8 GESTÃO DO CONHECIMENTO ORGANIZACIONAL	0,68	0,88	**0,76**
9 GESTÃO AMBIENTAL	0,71	0,87	**0,78**
10 EDUCAÇÃO EM ENGENHARIA DE PRODUÇÃO	0,52	0,65	**0,58**
11 ENG. PROD., SUSTENTABILIDADE E RESP. SOCIAL	0,88	0,46	**0,60**
Acurácia		**73,71%**	

Tabela 4 – Resultado do classificador de grupo proposto com o conjunto de teste

Observando os resultados da Tabela 4 é possível obter-se algumas conclusões. Com exceção das categorias 6, 10 e 11, todas obtiveram um valor de F_1 acima de 0,70. A categoria 4 obteve melhor valor de F_1, isto é, tem o melhor desempenho de classificação combinando a precisão e a abrangência (0,86). Além disso, foi a categoria que atingiu o maior nível de abrangência, com 96% dos documentos pertencentes à categoria 4 corretamente classificados. Na prática isso se traduz em um alto número de verdadeiro-positivos. Pode-se afirmar, que os documentos pertencentes a essa categoria, possuem uma grande quantidade de termos que pesam em sua representação de forma a diferenciá-la bastante das demais. A categoria que atingiu o maior nível de precisão foi a categoria 1, com 91%. Isto representa um baixo número de falso-positivos. A categoria 11 demonstrou-se como a de menor abrangência e a 10 de menor precisão.

Pelos dados da Tabela 5, observa-se que as categorias 11 e 9, possuem o maior grau de similaridade, que se traduz na prática como documentos que compartilham grande quantidade de termos com pesos equivalentes, sugerindo publicação em ambas áreas, mas contribui para a baixa abrangência apresentada pela categoria 11 no experimento realizado. Na Tabela 6 apresenta-se como referência, o cálculo de similaridade de documentos de áreas totalmente distintas obtidas em um experimento de apoio para esta finalidade.

Categoria 1	Categoria 2	similaridade
11	9	0,540
10	2	0,397
5	7	0,330
3	8	0,318
11	7	0,275

Tabela 5 – Os cinco pares de categorias com maior grau de similaridade dentre as onze categorias da Engenharia de Produção

Categoria 1	Categoria 2	similaridade
Direito	Odontologia	0,007
Direito	Informática	0,008
Odontologia	Informática	0,008

Tabela 6 – Teste de similaridade com artigos de Direito, Odontologia e Informática

Para ilustrar um caso prático, a escolha da área de submissão do presente artigo (8.Gestão do Conhecimento Organizacional) foi realizada utilizando o classificador proposto. A Tabela 7 apresenta o resultado da votação que determinou essa escolha.

Categorias	1	2	3	4	5	6	7	8	9	10	11
Pontuação	0,30	0,70	0,43	0,10	0,09	1,02	0,03	**1,24**	0,06	0,50	0,02

Tabela 7 – Resultado da categorização do presente artigo pelo método proposto

5. Conclusão

Considerando o alto grau de similaridade entre documentos de algumas categorias comprovado experimentalmente, e o fato de não existir na literatura um valor mínimo estipulado para determinar se os valores das métricas: acurácia, precisão e abrangência são satisfatórios, trazendo essa subjetividade aos especialistas do domínio estudado, conclui-se que o classificador proposto neste trabalho pode ser utilizado em uma ferramenta de apoio a professores e alunos da área de Engenharia de Produção, de forma a auxiliá-los no processo de escolha da melhor área para publicação do seus artigos.

Além disso, pelos resultados obtidos neste trabalho, sugere-se utilizar o mesmo modelo adaptando-o para realizar o segundo nível de classificação, determinando a subárea de publicação do artigo.

Enfim, espera-se que este trabalho contribua para o crescimento, organização e qualidade da produção científica em Engenharia de Produção no Brasil.

Referências

ABEPRO (Brasil) (Org.). *ANAIS ENEGEP***.** Disponível em: <http://www.abepro.org.br/publicacoes/>. Acesso em: 19 fev. 2011.

ABEPRO (Rio de Janeiro). *Áreas e Sub-áreas para envio de artigos***.** Disponível em: <http://www.abepro.org.br/internasub.asp?m=1061&ss=42&c=1104 >. Acesso em: 08 abr. 2012.

BERRY, Michael W.; KOGAN, Jacob. *Text Mining Applications and Theory.* Wiley, 2010. 223 p.

CADWeb, Disponível em: <http://www.net.ucam-campos.br/>. Acesso em: 04 abr. 2012.

CAPES (Brasil). *Relação de Cursos Recomendados e Reconhecidos***.** Disponível em: <http://conteudoweb.capes.gov.br/conteudoweb/ProjetoRelacaoCursosServlet?acao=pesquisarIes&codigoArea=30800005&descricaoArea=ENGENHARIAS+&descricaoAreaConhecimento=ENGENHARIA+DE+PRODU%C7%C3O&descricaoAreaAvaliacao=ENGENHARIAS+III>. Acesso em: 19 mar. 2012.

CHANG, Chih-chung; LIN, Chih-jen. *LIBSVM: A library for support vector machines.* Acm Trans. Intell. Syst. Technol., New York, p.1-27, 2011. Disponível em: <http://doi.acm.org/10.1145/1961189.1961199>. Acesso em: 20 maio 2011.

CORTES, Corinna; VAPNIK, Vladimir. *Support-Vector Networks*. Machine Learning, v. 20, p.273-297, 1995.

DOMINGOS, P.; PAZZANI, M. *On The Optimality of the Simple Bayesian Classifier Under Zero-one Loss.* Machine Learning, 29 (2/3), 103, 1997.

FELDMAN, Ronen; SANGER, James. *THE TEXT MINING HANDBOOK: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press, 2007. 422 p.

FREE PDF to TXT Converter, Disponível em: <http://www.somepdf.com/some-pdf-to-txt-converter.html>. Acesso em: 20 mar. 2011.

GOMES, Georgia Regina Rodrigues; MORAES FILHO, Rubens de Oliveira. *CADWeb – Categorização automática de documentos digitais*. Ci. Inf.**,** Brasília, v. 1, n. 40, p.68-76, jan. 2011.

HSU, Chih-wei; CHANG, Chih-chung; LIN, Chih-jen. *A Practical Guide to Support Vector Classification.* Bioinformatics, v. 1, p.1-16, 2010. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.6.3096&rep=rep1&type=pdf>. Acesso em: 20 maio 2011.

MARON, M. E.; KUHNS, J. L.. *On Relevance, Probabilistic Indexing and Information Retrieval.* Journal Of The Acm (jacm), New York, v. 8, n. 3, p.216-244, jul. 1961.

MIERSWA, Ingo et al. *YALE: Rapid Prototyping for Complex Data Mining Tasks.* Proceedings Of The 12th Acm Sigkdd International Conference On Knowledge Discovery And Data Mining: KDD , Philadelphia, p.935-940, 2006. Disponível em: <http://rapid-i.com/component/option,com_docman/task,doc_download/gid,25/Itemid,62/>. Acesso em: 02 maio 2011.

NUPENGE (Brasil). *CURSOS DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO***.** Dados organizados pelo NUPENGE (Núcleo de Estudos e Pesquisas sobre Formação e Exercício Profissional em Engenharia da UFJF) com base nos dados coletados do site http://emec.mec.gov.br. Revisado em julho de 2011. Apoio: ABEPRO. Disponível em: <http://www.ufjf.br/proengprod/files/2010/05/cursosEP.xls>. Acesso em: 19 mar. 2012.

PORTER, Martin F.. *Snowball: A language for stemming algorithms.* Disponível em: <http://snowball.tartarus.org/texts/introduction.html>. Acesso em: 20 maio 2011.

SEBASTIANI, Fabrizio. *Machine learning in automated text categorization*. Acm Computing Surveys, v. 34, n. 1, p.1-47, 2002.

SIMPEP (Brasil). *ANAIS SIMPEP***.** Disponível em: <http://www.simpep.feb.unesp.br/anais.php>. Acesso em: 19 mar. 2011.

TAN, Pang-ning; STEINBACH, Michael; KUMAR, Vipin. *Introdução ao DATA MINING Mineração de Dados.* Rio de Janeiro: Ciência Moderna Ltda, 2009. 900 p.

VAPNIK, Vladmir. *The Nature of Statistical Learning Theory.* 2. ed. New York: Springer, 2000. 314 p.

WETTSCHERECK, Dietrich; AHA, David W.; MOHRI, Takao. *A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms.* Rtificial Intelligence Review**,** Springer Netherlands, v. 11, n. 1, p.273-314, 01 fev. 1997. Disponível em: <http://dx.doi.org/10.1023/A:1006593614256>. Acesso em: 04 abr. 2012.

WILLETT, Peter. *The Porter stemming algorithm: then and now*. Program: Electronic Library And Information Systems, v. 40, n. 3, p.219-223, 2006.

ZHANG, H.. *The optimality of naive bayes.* Proceedings Of The Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, p.562-567, 2004. Disponível em: <http://www.springerlink.com/content/51t4233286xn76rr/fulltext.pdf>. Acesso em: 23 mar. 2012.

ANEXO

C	ε	acurácia (%)
2,00	0,00003	70,94
0,50	0,00700	70,14
0,50	0,00012	69,96
2,00	0,00200	69,79
0,50	0,00003	69,63
0,50	0,03125	69,63
32,00	0,00012	69,62
2,00	0,50000	69,36
8,00	0,12500	69,32
512,00	0,12500	69,31
128,00	0,00700	69,30
512,00	0,00003	69,17
2,00	0,00012	69,17
8,00	0,03125	69,15
128,00	0,00200	69,14
2,00	0,12500	69,13
32,00	0,00003	69,12
0,50	0,00200	68,98
2,00	0,00700	68,98
0,50	0,12500	68,80
128,00	0,00012	68,79
512,00	0,00700	68,79
2,00	0,03125	68,79
32,00	0,00200	68,66
8,00	0,50000	68,65
8,00	0,00003	68,64
0,13	0,00200	68,50
8,00	0,00012	68,49
8,00	0,00200	68,48
0,50	0,50000	68,48
128,00	0,00003	68,35
8,00	0,00700	68,32
32,00	0,03125	68,31
32,00	0,00700	68,30
512,00	0,00012	68,29
0,13	0,00003	68,14
512,00	0,00200	68,13
32,00	0,12500	68,00
32,00	0,50000	67,83
512,00	0,50000	67,65
128,00	0,12500	67,49
0,13	0,12500	67,31
128,00	0,50000	67,30
0,13	0,03125	67,16
128,00	0,03125	67,16
0,13	0,00700	67,16
0,13	0,50000	67,15
0,13	0,00012	66,68
512,00	0,03125	66,36
512,00	2,00000	65,71
2,00	2,00000	65,54
32,00	2,00000	65,53
8,00	2,00000	65,20
0,50	2,00000	65,18
128,00	2,00000	65,02
0,03	0,03125	64,54
0,03	0,00700	63,85
0,03	0,00003	63,19
0,03	0,12500	63,19
0,03	0,00012	63,04
0,03	0,00200	62,72
0,03	0,50000	62,54
0,13	2,00000	60,22
0,03	2,00000	56,06

APÊNDICE 1 – Resultado do processo de busca por força bruta dos parâmetros *C* e ε do classificador SVM.