Predição de Rankings de Artistas por Meio de Regressão

Alternative Title: Prediction of Artists' Rankings by Regression

Felipe L. M. Faria
Departamento de Ciência da Computação
Universidade Federal de Ouro Preto
Ouro Preto - MG - Brasil
felipelmfaria@hotmail.com

Álvaro R. Pereira Jr.
Departamento de Ciência da Computação
Universidade Federal de Ouro Preto
Ouro Preto - MG - Brasil
alvaro@iceb.ufop.br

Luiz H. C. Merschmann
Departamento de Ciência da Computação
Universidade Federal de Ouro Preto
Ouro Preto - MG - Brasil
luizhenrique@iceb.ufop.br

RESUMO

A construção de *rankings* consiste na ordenação de resultados recuperados de acordo com um determinado critério. Os *rankings* podem fornecer informações relevantes para analistas de diferentes setores da indústria. Na indústria fonográfica, os *rankings* possibilitam a compreensão de como os estilos musicais e a popularidade dos artistas e suas músicas evoluem com o tempo, permitindo análises de históricos de execuções e de tendências. Devido à importância da construção de *rankings* no escopo musical, técnicas de mineração de dados têm sido utilizadas para predizer os *rankings* através de informações contidas em mídias sociais. Este trabalho avalia modelos de regressão para a predição de *rankings* de artistas utilizando-se de dados históricos (*rankings* diários de artistas) extraídos do *website* Vagalume. Três técnicas de regressão (k-Nearest Neighbors - k-NN, Regressão Linear Múltipla - RLM e Random Forests - RF) foram avaliadas neste trabalho considerando-se diversos cenários. Os resultados obtidos por meio de experimentos mostraram que predições com baixas taxas de erros podem ser obtidas, indicando que técnicas de mineração de dados podem ser utilizadas para obtenção de informações que auxiliem a indústria fonográfica na tomada de decisões.

Palavras-Chave

Mineração de Dados, Regressão, Mídia Social

ABSTRACT

The construction of rankings consists of ordering retrieved results according to certain criteria. Rankings can provide relevant information to analysts from different sectors of industry. For the music industry, rankings enable understanding how musical genres and popularity of artists and their songs evolve over time, allowing analyses of history data and trends. Due to the importance of building rankings in the musical scope, data mining techniques have been used to predict rankings by using information from social media. This work evaluates regression models for prediction of artists' rankings using historical data (daily rankings of artists) extracted from website Vagalume. Three regression techniques (k-Nearest Neighbors - k-NN, Multiple Linear Regression - MLR and Random Forests - RF) were evaluated in this study considering different scenarios. Results obtained from experiments showed that predictions with low error rates can be obtained, indicating that data mining techniques can be used to obtain information to assist the music industry in decision making.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining

General Terms

Experimentation

Keywords

Data Mining, Regression, Social Media

1. INTRODUÇÃO

Segundo os autores do trabalho [11], a construção de *ranking* consiste na ordenação de resultados recuperados de acordo com um determinado critério. No contexto fonográfico, a construção de *rankings* possibilita a compreensão de como os artistas, estilos musicais e a própria música evoluem ao longo do tempo, podendo nos auxiliar no entendimento do processo criativo de fazer música [3]. Além disso, os *rankings* podem auxiliar críticos de música a analisarem tendências a partir do posicionamento dos artistas nos mesmos.

O volume de vendas de músicas está perdendo seu papel como indicador de popularidade de artistas. Embora os números de vendas e execuções em rádios possam ter sido muito utilizados para medir a popularidade de um artista nas décadas de 50 e 60, essas variáveis tornaram-se cada vez mais pobres como indicadores de popularidade em detrimento ao rápido crescimento da divulgação da música em meio digital [9]. No meio digital, as mídias sociais têm sido muito utilizadas para geração de *rankings* de artistas. Um exemplo disso é o *ranking* "Social 50" divulgado na revista Billboard, o qual é construído a partir de dados coletados em mídias sociais.

Com o aumento da quantidade de informação disponível na *web*, a realização de *rankings* tornou-se um problema

importante a ser estudado [11]. Desse modo, no contexto musical, diversas técnicas de mineração de dados têm sido utilizadas para predição de *rankings* de artistas. Essas técnicas usam dados históricos para gerar modelos que posteriormente são utilizados para inferir novos *rankings*.

O Vagalume[1], um *website* colaborativo brasileiro de música, é uma mídia social bastante popular. De acordo com o *ranking* divulgado no *website* Alexa[2], que leva em consideração o número de acessos a um *website*, o Vagalume tem se mantido ao longo do tempo bem classificado quando comparado com outras mídias brasileiras, como o Letras[3], o Kboing[4] e a versão brasileira do LastFM[5].

O Vagalume nos permite ouvir músicas, visualizar notícias do meio musical, criar *playlists* (listas de músicas), enviar letras, responder enquetes, comentar sobre os artistas, seguir o que outro usuário está fazendo nessa mídia social, entre outras atividades. Além disso, ele disponibiliza um *ranking* diário de artistas, o qual é construído a partir da popularidade do artista no *website*.

Assim, este trabalho utiliza os dados históricos (*rankings* diários de artistas) fornecidos pelo *website* Vagalume para avaliar a capacidade de modelos de regressão na predição de *rankings* de artistas. Três técnicas (*k-Nearest Neighbors* – *k*-NN, Regressão Linear Múltipla – RLM e Random Forests – RF) foram utilizadas para a construção dos modelos de regressão avaliados. A geração de modelos precisos é importante para auxiliar a indústria fonográfica brasileira nas análises de tendências, por exemplo, indicando quais artistas estarão no topo do *ranking* numa data futura.

As demais seções deste trabalho estão organizadas como descrito a seguir. A Seção 2 contém um breve resumo de alguns trabalhos relacionados. Em seguida, a metodologia utilizada neste trabalho é apresentada na Seção 3. A descrição dos experimentos e os resultados obtidos são mostrados na Seção 4. Por fim, a Seção 5 apresenta as conclusões e sugestão de trabalho futuro.

2. TRABALHOS RELACIONADOS

Os autores do trabalho [9] ressaltam a importância da utilização das mídias sociais para a classificação de artistas por meio da construção de *rankings*. Assim, devido à importância desse assunto, vários autores têm realizado esforços na geração de *rankings* confiáveis e precisos para atender diversas necessidades de usuários do meio fonográfico.

No trabalho [5], os autores propuseram uma abordagem, que utiliza técnicas de mineração de dados em texto, para medir a popularidade e construir *rankings* de artistas a partir da análise dos comentários dos ouvintes na rede social MySpace[6], uma mídia musical popular. Eles demonstraram ter obtido resultados mais próximos àqueles que seriam gerados manualmente por usuários (estudantes universitários), do que os que são gerados a partir dos métodos utilizados pela revista Billboard.

Autores do trabalho [3] construíram um *ranking* de artistas com base no *website WhoSampled*[7], um *website* de informações sobre o meio musical. Para isso, utilizaram várias características de artistas e gêneros musicais e adotaram a métrica PageRank para permitir a interpretação e descrição de padrões de influência musical, tendências e características de músicas.

Indo além de métodos que utilizam as mídias digitais para a construção de *rankings*, outros trabalhos já foram realizados para auxiliar o meio fonográfico em suas diversas necessidades. Nesses trabalhos, a utilização de técnicas de regressão no auxílio à construção de *rankings* relacionados ao meio fonográfico tem gerado resultados satisfatórios. Alguns desses trabalhos são apresentados a seguir.

No trabalho [14], os autores construíram um sistema de reconhecimento de emoções que utiliza valores de valência (polaridade das emoções) e de excitação (intensidade das emoções) para recomendar músicas para usuários. De acordo com a emoção escolhida por um usuário, o sistema retorna músicas caracterizadas com valores de valência e excitação mais próximos àqueles associados à emoção escolhida pelo usuário. As emoções são tratadas como variáveis contínuas cujos valores são computados por meio de funções que combinam valores de valência e de excitação. Técnicas de regressão foram utilizadas para predizer os valores de valência e de excitação das músicas a partir de características (timbre do cantor, intensidade dos agudos, sonoridade) dos segmentos das mesmas. Dentre os resultados obtidos, foi constatado que realizar a predição utilizando técnicas de regressão é melhor do que realizá-la a partir de técnicas de classificação, onde haveria a dificuldade para se distinguir entre as emoções "feliz" e "satisfeito". Além disso, os autores afirmaram que os resultados de acurácia preditiva obtidos superaram aqueles apresentados na literatura.

Os autores do trabalho [4] propuseram a recomendação de artistas a partir das características musicais dos artistas e das preferências dos usuários. Nesse trabalho, um artista foi caracterizado por parâmetros do seu modelo acústico correspondente. A partir disso, esses modelos acústicos juntamente com os *rankings* de preferência dos usuários foram utilizados como vetores de entrada para gerar um modelo de regressão que posteriormente foi usado para recomendar outros artistas para os usuários. Os resultados experimentais desse trabalho demonstraram que essa abordagem superou outras em termos da qualidade da recomendação.

No trabalho [13], os autores propuseram classificar músicas por meio de características denominadas de conceito, que podem ser gêneros, instrumentação, características vocais e outros. Para isso, treinaram um modelo de conceitos através de uma técnica de regressão ordinal, que foi utilizada visando minimizar o problema de desbalanceamento de classes da base de dados utilizada nos experimentos. Com essa abordagem, em uma avaliação que foi realizada sobre um conjunto de dados, os autores mostraram que foram capazes de melhorar a precisão da detecção de conceitos em 10,89% em média em relação a outros trabalhos apresentados na literatura.

Diferentemente do foco dos trabalhos descritos anteriormente, neste trabalho utilizam-se técnicas de regressão para predizer as posições dos artistas no *ranking* de uma determinada data a partir de suas posições em datas anteriores por meio de dados obtidos em uma mídia social. Espera-se, que assim como nos trabalhos citados anteriormente, a utilização da tarefa de regressão gere resultados satisfatórios.

[1]http://www.vagalume.com.br
[2]http://www.alexa.com
[3]http://www.letras.mus.br
[4]http://www.kboing.com.br
[5]http://www.lastfm.com.br
[6]http://www.myspace.com
[7]http://www.whosampled.com

3. METODOLOGIA

Nesta seção, é apresentada a maneira como as bases de dados foram construídas, na Seção 3.1. Em seguida, na Seção 3.2, é realizado um estudo dessas bases de dados.

3.1 Bases de Dados

O *website* Vagalume foi a mídia social escolhida para a coleta dos dados utilizados na construção das bases de dados adotadas neste trabalho. Esse *website* disponibiliza *rankings* de artistas que são construídos de acordo com a popularidade deles nessa mídia levando-se em consideração a quantidade de acessos realizados por usuários. Para cada artista, a rede social disponibiliza um histórico contendo sua posição diária no *ranking* ao longo do tempo. Esse histórico da posição dos artistas no *ranking* foi coletado utilizando-se uma API (*Application Programming Interface*) disponibilizada pelo Vagalume. A título de ilustração desse histórico, a Figura 1 apresenta um gráfico disponibilizado pelo *website*. Nesse gráfico, o eixo x representa a data e o eixo y a posição do artista no *ranking*.

Fonte: Adaptado do *website* Vagalume

Figura 1: Exemplo do *Ranking* diário de um artista no *website* Vagalume.

As bases de dados utilizadas neste estudo foram construídas escolhendo-se 478 artistas de 63 gêneros diferentes de música. Como há mais de 478 artistas no *ranking* gerado pelo Vagalume, para garantir que os artistas selecionados fossem classificados entre as posições 1º e 478º, uma reclassificação foi realizada respeitando-se a ordem da posição dos artistas no *ranking* original do Vagalume. O período considerado para coleta de dados foi de 13 de janeiro a 03 de abril de 2014.

A seguir, a Figura 2 apresenta a estrutura das bases de dados construídas. Cada instância das bases representa um artista que é caracterizado pelas suas posições nos *rankings* gerados em diferentes dias. Com o objetivo de avaliar a predição em diferentes datas, os dias 21 (classe 21), 41 (classe 41) e 81 (classe 81) foram escolhidos como atributo classe para as diferentes bases de dados geradas. Além disso, para cada atributo classe, quatro bases de dados foram construídas variando-se a quantidade de dias, os atributos preditores. Foram construídas bases com a quantidade de 10, 20, 40 e 80 atributos preditores.

Bases com as características mencionadas anteriormente foram geradas para diferentes quantidades de artistas de acordo com o gênero musical dos mesmos, a saber: Sertanejo, Rock, Pop, MPB e todos os gêneros juntos. Sendo assim, foram construídas bases contendo somente artistas do gênero Sertanejo (41 instâncias), somente artistas do gênero Rock (134 instâncias), somente artistas do gênero Pop (109 instâncias), somente artistas do gênero MPB (63 instâncias) e o total de artistas de todos os gêneros (478 instâncias).

Artistas/Dias	Dia 1	Dia 2	Dia 3	...	Dia n
Artista A	82	82	96	...	...
Artista B	83	87	81	...	...
Artista C	84	84	86	...	...
...	...	...	...	...	...
Artista Z	86	90	92	...	...

Figura 2: Exemplo da base de dados construída do *Ranking* diário de artistas.

A construção dessas bases considerando-se diferentes gêneros foi realizada com intuito de se avaliar diferentes cenários. Os gêneros Pop e Sertanejo representam os gêneros mais acessados na *web*. Em 2013, por exemplo, segundo a lista das expressões mais acessadas do Google[8], o gênero Pop foi um dos mais buscados, e no Youtube Brasil[9], dentre os vídeos mais acessados, estão os gêneros Pop e Sertanejo. Pesquisa realizada pela empresa Crowley[10], especializada em monitoração eletrônica de *broadcast* de áudio no Brasil desde 1997, aponta que nos últimos 14 anos os gêneros Sertanejo e Pop seguem em alta, sendo o primeiro e o segundo mais executados ao longo dos anos nas rádios, respectivamente. Por outro lado, o gênero MPB e Rock vêm decaindo em termos do número de execuções em rádios.

3.2 Análise de Dados

Um estudo nas bases construídas foi realizado com o intuito de se observar como a posição dos artistas no *ranking* varia ao longo do tempo. Desse modo, o gráfico da Figura 3 apresenta o cálculo da média da variação de posições para os artistas classificados em faixas do *ranking* para uma das bases (base com 10 atributos preditores e a classe 81). As faixas são: 1º ao 10º, 11º ao 20º, 21º ao 30º e assim por diante até o intervalo 471º ao 478º. O gráfico em questão mostra essa variação média de posições para cada um dos intervalos mencionados anteriormente considerando os artistas de todos os gêneros musicais. Vale ressaltar que para todas as bases construídas, o comportamento dos gráficos foi semelhante. No gráfico da Figura 3, o eixo x apresenta os intervalos de posições de 10 em 10 e o eixo y mostra a variação média de posições.

No cenário apresentado no gráfico da Figura 3 pode-se observar que para alguns intervalos as variações de posições em dias subsequentes são menores, indicando que a posição dos artistas no *ranking* é mais estável ao longo do tempo. Por outro lado, intervalos com variações maiores indicam uma maior oscilação de posições ao longo dos dias. Por exemplo, no gráfico da Figura 3 pode-se verificar que a variação média de posições no *ranking* para artistas classificados entre 421º e 430º é igual a 10. Isso significa que, entre dias subsequentes, a posição de um artista classificado nesse intervalo varia em média 10 posições. De modo geral, essa análise mostrou que as variações menores estão associadas aos intervalos das posições iniciais (entre 1º e 50º) e finais do *ranking* (entre 431º e 478º), enquanto as maiores variações ocorrem nos intervalos que representam as posições intermediárias (entre

[8]http://www.google.com/trends/topcharts#geo=BR&date=2013

[9]http://www.youtube.com/playlist?list=PLoeZWzNXxmy8FAz3OUQln9CTWKq5lCs5i

[10]http://www.crowley.com.br

Figura 3: Variação média de posições considerando-se 10 atributos preditores e a classe 81.

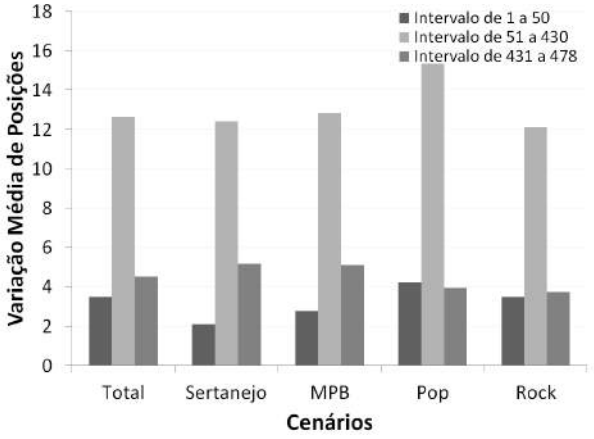

Figura 4: Variação média de posições considerando-se 10 atributos preditores e a classe 81 para os artistas de todos os gêneros e os artistas dos gêneros MPB, Pop, Rock e Sertanejo separadamente.

51º e 430º) do *ranking*.

O padrão de variação de posições observado para os artistas presentes separadamente nos gêneros MPB, Pop, Rock e Sertanejo é semelhante àquele apresentado no gráfico da Figura 3 (para os artistas de todos os gêneros musicais). Isso pode ser observado no gráfico da Figura 4, que apresenta para todos os cenários estudados a variação média de posições dos artistas no *ranking* ao longo do tempo para três intervalos de posições: intervalo das posições iniciais (entre 1º e 50º), intervalo das posições intermediárias (entre 51º e 430º) e intervalo das posições finais do *ranking* (entre 431º e 478º). Nesse gráfico, o eixo x apresenta os diferentes cenários e o eixo y mostra a variação média de posições.

Essa análise será útil para ajudar a explicar os resultados obtidos a partir dos modelos de regressão, os quais serão apresentados na Seção 4.2.

4. EXPERIMENTOS COMPUTACIONAIS

Nesta seção, será apresentada a configuração experimental, mostrando as técnicas e as métricas utilizadas, na Seção 4.1. Em seguida, na Seção 4.2 serão apresentados os resultados experimentais obtidos no estudo realizado neste trabalho.

4.1 Configuração Experimental

Para os estudos conduzidos neste trabalho foram escolhidas três técnicas de regressão bastante difundidas na literatura: k-NN (*k-Nearest Neighbors*) [1], RF (*Random Forests*) [2] e RLM (Regressão Linear Múltipla) [7]. Os experimentos foram realizados utilizando-se as implementações contidas no *software* R [10] (para as técnicas k-NN e RF) e na ferramenta Weka [6] (para a RLM).

Para a técnica k-NN, avaliou-se modelos de regressão com os seguintes valores do parâmetro k: 1, 3, 5, 7 e 10. Para a técnica RF dois parâmetros foram ajustados: para o *ntree*, que define o número de árvores da floresta, foram avaliados os valores 500 e 1000 (recomendado em [2]); para o *mtry*, que define o número de atributos que serão considerados na construção de cada árvore, foram avaliados os seguintes valores: o número de atributos preditores divididos por 3, 1/6 da quantidade de atributos preditores e 2/3 da quantidade de atributos preditores (sugeridos em [2]).

Para avaliar a precisão preditiva dos modelos de regressão utilizou-se a técnica de validação cruzada denominada *Leave-One-Out* [8]. Nessa técnica, considerando-se uma base de dados históricos com n instâncias, n modelos de regressão são construídos a partir de bases de treinamento e testados a partir de bases de teste. Para isso, a base de dados históricos é subdividida em treinamento e teste n vezes. Em cada subdivisão, das n instâncias que compõem a base de dados históricos, $n-1$ são utilizadas na formação da base de dados de treinamento e a instância restante compõe a base de dados de teste. Vale ressaltar que cada uma dessas n instâncias da base de dados históricos é utilizada uma vez para compor a base de teste. Para medir a acurácia preditiva dos modelos utilizou-se a métrica Desvio Médio Absoluto (DMA), que computa a diferença entre os valores reais e os valores preditos. A Equação 1 apresenta o cálculo do DMA.

$$DMA = \frac{\sum_{t=1}^{n} |e_t|}{n}, \quad (1)$$

onde e_t é a diferença entre os valores real e predito e n é o

número de instâncias da base.

Além do DMA, outra métrica de avaliação foi utilizada: o Coeficiente de Determinação (R^2) [12]. O R^2 representa o poder explicativo do modelo de regressão, que tem por objetivo avaliar o quanto os atributos preditores conseguem explicar o atributo classe. Quanto mais próximo o R^2 for de 100% melhor é o modelo, enquanto que resultados próximos a 30% são considerados fracos.

4.2 Resultados Experimentais

Apesar de diversos resultados terem sido gerados a partir de diferentes valores de parâmetros das técnicas de regressão avaliadas (ver Seção 4.1), apenas os melhores resultados serão apresentados a seguir. Esses resultados foram gerados com os seguintes valores de parâmetros: k igual a 3 (para a técnica k-NN), $ntree$ igual a 1000 e $mtry$ igual a 2/3 da quantidade de atributos preditores (para a técnica RF).

As Tabelas 1, 2, 3, 4 e 5 apresentam os resultados de DMA e R^2 para os diferentes cenários avaliados. Nessas tabelas, a primeira coluna refere-se à quantidade de atributos preditores (10, 20, 40 e 80 atributos); a segunda coluna contém as classes utilizadas (classe 21, classe 41 e classe 81) e as três colunas seguintes contêm os resultados (DMA e R^2) das três técnicas avaliadas (k-NN, RF e RLM). Os melhores e piores valores de DMA estão em negrito. Vale observar que para determinadas combinações de "Quant. Atributos Preditores" e "Atributo Classe" os resultados estão marcados com o símbolo –, indicando que para aquele atributo classe não existia na base de dados históricos aquela quantidade de atributos preditores para geração do modelo de regressão.

A Tabela 1 apresenta os resultados de DMA para a base que contém artistas de todos os gêneros (478 instâncias). Dentre os diversos cenários avaliados (variando-se a quantidade de atributos preditores e a classe a ser predita), obteve-se melhor resultado para a classe 81, em que a aplicação da técnica RLM na base com 10 atributos preditores resultou num DMA de 7,5. O pior resultado foi obtido para a classe 41, em que a técnica k-NN para a base com 40 atributos preditores apresentou um DMA de 15,41. Comparando as três técnicas, os melhores resultados de DMA sempre foram obtidos pela técnica RLM e os piores pelo k-NN, permanecendo o RF com desempenho intermediário entre a RLM e o k-NN.

Tabela 1: DMA considerando todos os artistas

Quant. Atributos Preditores	Atributo Classe	k-NN (DMA/R^2)	RF (DMA/R^2)	RLM (DMA/R^2)
10	classe 21	14,18/0,970	12,25/0,977	11,11/0,979
	classe 41	13,69/0,978	12,55/0,981	11,18/0,984
	classe 81	10,21/0,986	8,96/0,989	**7,50**/0,991
20	classe 21	14,49/0,968	12,26/0,977	11,50/0,978
	classe 41	14,69/0,974	12,59/0,981	10,61/0,984
	classe 81	11,39/0,982	8,89/0,990	7,61/0,991
40	classe 21	–	–	–
	classe 41	**15,41**/0,967	12,51/0,992	11,61/0,980
	classe 81	12,36/0,981	9,00/0,989	8,05/0,991
80	classe 21	–	–	–
	classe 41	–	–	–
	classe 81	13,56/0,975	9,08/0,989	8,78/0,989

A Tabela 2 apresenta os resultados de DMA para os artistas do gênero Sertanejo. Novamente o melhor resultado (10,84 de valor de DMA) foi obtido para a classe 81 utilizando-se a técnica RLM na base com 10 atributos preditores. O pior resultado (85,96 de DMA) foi obtido para a classe 41 com a técnica RLM na base com 40 atributos preditores. De modo semelhante à análise anterior, na maioria dos casos, a técnica RLM alcançou os melhores resultados e o k-NN os piores, ficando o RF com os resultados intermediários.

Tabela 2: DMA considerando o gênero Sertanejo

Quant. Atributos Preditores	Atributo Classe	k-NN (DMA/R^2)	RF (DMA/R^2)	RLM (DMA/R^2)
10	classe 21	19,79/0,957	18,79/0,965	15,70/0,973
	classe 41	16,78/0,972	15,76/0,977	11,02/0,988
	classe 81	15,02/0,975	13,10/0,982	**10,84**/0,985
20	classe 21	19,54/0,966	19,95/0,966	22,81/0,949
	classe 41	16,85/0,970	15,98/0,974	14,40/0,989
	classe 81	13,87/0,980	12,13/0,981	13,48/0,978
40	classe 21	–	–	–
	classe 41	16,60/0,970	16,31/0,972	**85,96**/0,266
	classe 81	14,15/0,979	14,33/0,973	47,67/0,798
80	classe 21	–	–	–
	classe 41	–	–	–
	classe 81	16,19/0,964	17,78/0,961	12,61/0,977

A Tabela 3 apresenta os resultados de DMA para os artistas do gênero Pop. O melhor resultado (9,19 de DMA) foi obtido pela técnica RLM para a classe 21 utilizando-se a base com 10 atributos preditores. O pior resultado (DMA de 25,89) foi encontrado para a classe 41 com a técnica RLM para a base que contém 40 atributos preditores. Para os artistas do gênero Pop, na maioria das vezes, a RLM supera as duas outras técnicas para as bases contendo 10 e 20 atributos preditores. Já para as bases com 40 e 80 atributros preditores, a técnica RF supera as demais.

Tabela 3: DMA considerando o gênero Pop

Quant. Atributos Preditores	Atributo Classe	k-NN (DMA/R^2)	RF (DMA/R^2)	RLM (DMA/R^2)
10	classe 21	18,83/0,953	16,69/0,962	**9,19**/0,991
	classe 41	18,02/0,964	17,13/0,968	14,60/0,970
	classe 81	11,93/0,986	10,17/0,990	9,20/0,991
20	classe 21	19,62/0,951	17,56/0,957	19,34/0,957
	classe 41	21,10/0,954	17,78/0,966	16,35/0,964
	classe 81	12,16/0,984	10,68/0,990	10,70/0,984
40	classe 21	–	–	–
	classe 41	22,13/0,955	18,23/0,965	**25,89**/0,876
	classe 81	13,80/0,981	11,85/0,987	12,41/0,980
80	classe 21	–	–	–
	classe 41	–	–	–
	classe 81	16,15/0,973	13,28/0,983	21,62/0,952

A Tabela 4 apresenta os resultados de DMA para os artistas do gênero MPB. Observa-se que o melhor resultado (11,27 de DMA) foi alcançado pela técnica RLM para a classe 81 com a base que possui 10 atributos preditores. Já o pior resultado (DMA de 41,10) também foi obtido pela técnica RLM para a classe 81 com a base que contém 80 atributos preditores. Na maioria dos casos a técnica RLM apresenta os melhores resultados e o k-NN os piores, ficando o RF numa posição intermediária.

Tabela 4: DMA considerando o gênero MPB

Quant. Atributos Preditores	Atributo Classe	k-NN (DMA/R^2)	RF (DMA/R^2)	RLM (DMA/R^2)
10	classe 21	20,66/0,930	17,32/0,956	18,07/0,950
	classe 41	17,90/0,966	18,04/0,966	13,80/0,979
	classe 81	12,06/0,984	11,36/0,986	**11,27**/0,983
20	classe 21	21,68/0,924	18,23/0,952	21,36/0,941
	classe 41	18,99/0,940	17,46/0,950	12,18/0,985
	classe 81	12,59/0,984	12,42/0,984	12,24/0,985
40	classe 21	–	–	–
	classe 41	17,01/0,970	18,40/0,918	16,66/0,969
	classe 81	19,48/0,955	15,65/0,970	16,37/0,973
80	classe 21	–	–	–
	classe 41	–	–	–
	classe 81	20,37/0,947	17,31/0,958	**41,10**/0,913

Por fim, a Tabela 5 apresenta os resultados de DMA para os artistas do gênero Rock. Mais uma vez o melhor resultado (DMA de 7,8) foi alcançado pela técnica RLM para a classe 21 para a base com 10 atributos preditores. O pior resultado (DMA de 23,62) foi também alcançado para a RLM, mas para a classe 81 utilizando-se a base com 80 atributos preditores. Considerando-se todos os cenários avaliados, verifica-se que a RLM supera as demais técnicas na maioria dos casos e que o k-NN geralmente apresenta o pior desempenho, ficando novamente o RF com os resultados intermediários.

Tabela 5: DMA considerando o gênero Rock

Quant. Atributos Preditores	Atributo Classe	k-NN (DMA/R^2)	RF (DMA/R^2)	RLM (DMA/R^2)
10	classe 21	10,89/0,987	9,24/0,991	**7,80**/0,993
	classe 41	16,75/0,963	14,54/0,970	13,00/0,974
	classe 81	14,26/0,968	11,71/0,977	9,07/0,982
20	classe 21	10,52/0,987	9,79/0,989	8,43/0,991
	classe 41	14,59/0,971	14,89/0,972	13,76/0,970
	classe 81	13,36/0,972	11,56/0,976	10,29/0,980
40	classe 21	–	–	–
	classe 41	15,75/0,969	15,49/0,971	16,13/0,950
	classe 81	14,34/0,969	11,64/0,976	13,57/0,971
80	classe 21	–	–	–
	classe 41	–	–	–
	classe 81	16,86/0,957	12,54/0,973	**23,62**/0,929

Nos diferentes cenários avaliados até aqui para bases formadas por artistas de cada um dos quatro gêneros analisados separadamente (Sertanejo, Pop, Rock e MPB) e para uma base que reúne todos os artistas de todos os gêneros musicais, pode-se verificar que a técnica RLM se destacou em relação às demais, apresentando na maioria das vezes os melhores resultados de DMA e R^2. Além disso, para todos os casos analisados, o melhor resultado sempre foi obtido a partir da base formada por 10 atributos preditores. É possível observar também que consideráveis variações de desempenho podem ocorrer quando diferentes atributos classe são considerados. No caso dos experimentos aqui realizados, de modo geral, os melhores desempenhos foram alcançados para a classe 81.

Comparando-se os resultados entre os quatro diferentes gêneros (Sertanejo, Pop, Rock e MPB), percebe-se que os melhores resultados foram encontrados para os artistas do gênero Rock. Esse resultado pode ser explicado pelo fato de esse gênero geralmente ter apresentado as menores variações de posições no *ranking* ao longo do tempo (ver gráfico da Figura 4). Dentre todos os cenários avaliados, a base considerando os artistas de todos os gêneros foi a que alcançou o melhor resultado (7,5 de DMA).

As tabelas apresentadas até o momento mostraram apenas o DMA considerando-se o *ranking* completo em cada um dos cenários avaliados. Para viabilizar uma análise do desempenho dos modelos de regressão para diferentes faixas do *ranking*, gerou-se o gráfico da Figura 5, que apresenta os valores de DMA para todos os cenários em diferentes intervalos de posições: iniciais (entre 1º e 50º), intermediárias (entre 51º e 430º) e finais (entre 431º e 478º). Apenas os resultados de DMA para a RLM são mostrados, já que essa foi a técnica que proveu os melhores resultados. Nesse gráfico, o eixo x apresenta cada um dos cenários analisados e o eixo y os valores de DMA. Pode-se observar que, exceto para os artistas do gênero MPB no intervalo de 1º e 50º, para todos os demais casos, os resultados de DMA apresentados no gráfico da Figura 5 estão de acordo com as variações de posições no *ranking* apresentadas no gráfico da Figura 4, ou seja, quanto maior a variação de posição, pior o desempenho preditivo dos modelos de regressão e vice-versa.

Vale ressaltar que, no gráfico da Figura 5, o gênero Sertanejo apresentou o valor de DMA igual a 0 no intervalo de posições 431º a 478º porque não houve artistas desse gênero nesse intervalo de posições do *ranking* considerando-se o atributo classe analisado.

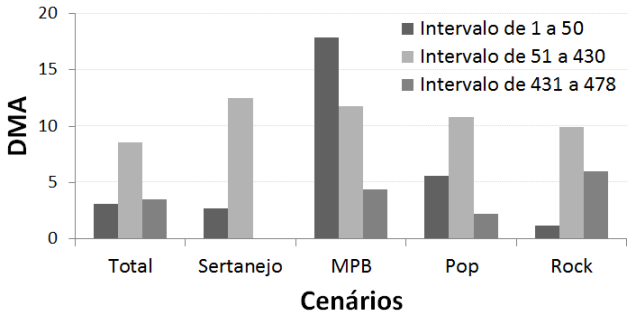

Figura 5: DMA considerando três intervalos, quantidade de atributos preditores de 10 dias e a classe 81.

Na maioria dos cenários os melhores resultados de DMA foram obtidos para o intervalo de posições iniciais (entre 1º e 50º). Ao levar em consideração que os artistas melhores posicionados nos *rankings* são geralmente os mais visados pela indústria fonográfica, os bons resultados preditivos para esses artistas mostram-se relevantes no auxílio à tomada de decisões nesse meio.

A Tabela 6 apresenta o desempenho da técnica RLM de acordo com a segunda métrica utilizada neste trabalho, o Coeficiente de Determinação R^2, considerando-se os intervalos de posições iniciais, intermediárias e finais. Pode-se observar que os resultados obtidos para essa métrica estão de acordo com os resultados de DMA apresentados nos gráficos e tabelas anteriores, ou seja, modelos de regressão gerados a partir da RLM se ajustaram muito bem aos dados ($R^2 \geq 0,722$) disponíveis em todos os cenários (exceto

para os artistas do gênero MPB no intervalo de posições iniciais, onde $R^2 = 0,12$, e para os artistas do gênero Sertanejo considerando-se o *ranking* completo, utilizando-se a classe 41 com a base que possui 40 atributos preditores, onde $R^2 = 0,266$ – ver Tabela 2).

Tabela 6: R^2 para os intervalos de posições iniciais, intermediárias e finais

Intervalos	Total	Sertanejo	Pop	Rock	MPB
Intervalo Inicial (1 a 50)	0,838	0,931	0,722	0,995	0,120
Intervalo Intermediário (51 a 430)	0,983	0,974	0,981	0,972	0,973
Intervalo Final (431 a 478)	0,906	–	0,975	0,883	0,886

5. CONCLUSÃO

Os dados contidos nas mídias sociais têm sido utilizados como indicadores da popularidade de artistas e, por isso, vêm sendo utilizados para geração de *rankings*. Um exemplo disso é o *ranking* "Social 50" divulgado na revista Billboard. A geração de *rankings* é importante por permitir aos analistas compreender a dinâmica do meio fonográfico e identificar tendências. Assim, a realização de *rankings* é um problema importante de ser estudado.

Devido à importância da realização de *rankings* no escopo musical, este trabalho avalia técnicas de mineração de dados na predição de *rankings*. Mais especificamente, a partir de dados históricos de posições de artistas em *rankings*, utilizam-se técnicas de regressão para predizer as posições de artistas no *ranking* em momentos futuros.

As técnicas de regressão k-NN, RF e RLM foram avaliadas com diferentes parâmetros, número de instâncias e quantidade de atributos preditores. De acordo com as métricas DMA e R^2, a técnica que gerou os melhores modelos de regressão foi a RLM. Para todos os casos analisados, o melhor resultado sempre foi obtido a partir de uma base de dados formada por 10 atributos preditores. Além disso, consideráveis variações de desempenho foram observadas quando diferentes atributos classe (associados a diferentes datas) foram utilizados. Verificou-se também a variação de desempenho preditivo dos modelos quando os artistas contidos nas bases de dados são separados de acordo com o gênero musical. Dentre os gêneros avaliados, os melhores desempenhos foram obtidos para o Rock, que está entre os gêneros onde os artistas apresentam as menores variabilidades de posição no *ranking* ao longo do tempo.

Por fim, verificou-se que os modelos de regressão fornecem predições próximas aos valores reais, ou seja, com baixas taxas de erro. Por exemplo, considerando a base que contempla artistas de todos os gêneros, o desvio médio absoluto (DMA) foi de 7,5. O bom desempenho dos modelos gerados pode ser confirmado a partir da métrica R^2, que na maioria dos casos se aproximou de 100%. Considerando os resultados preditivos por intervalos de posições no *ranking* (iniciais – do 1º ao 50º, intermediárias – do 51º ao 430º e finais – do 431º ao 478º), os melhores resultados (1,14 de DMA para a base com artistas do gênero Rock) foram obtidos para os artistas classificados no intervalo de posições iniciais do *ranking*. Esse resultado é interessante pelo fato de a indústria fonográfica geralmente estar mais interessada nos artistas melhores posicionados no *ranking*.

Apesar de outros trabalhos na literatura já terem abordado o problema da predição de *rankings*, os resultados obtidos por eles não são comparáveis aos aqui apresentados porque os dados utilizados para realizar a predição são diferentes daqueles adotados neste trabalho.

Como trabalho futuro sugere-se uma experimentação que considere a sazonalidade específica por datas comemorativas como, por exemplo, natal e carnaval, gerando modelos preditivos a partir de bases de dados históricos que contemplem artistas de gêneros musicais relacionados com essas datas comemorativas.

6. AGRADECIMENTOS

Os autores agradecem à Universidade Federal de Ouro Preto, CAPES, FAPEMIG e CNPq por apoiarem o desenvolvimento desta pesquisa. Agradecem também aos revisores anônimos por seus comentários construtivos, contribuindo para uma versão melhor deste artigo.

7. REFERÊNCIAS

[1] D. W. Aha. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies*, 36(2):267–287, 1992.

[2] L. Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[3] N. J. Bryan and G. Wang. Musical influence network analysis and rank of sample-based music. In *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR*, pages 329–334, Miami, Florida, 2011.

[4] Z.-S. Chen, J.-S. Jang, and C.-H. Lee. A kernel framework for content-based artist recommendation system in music. *IEEE Transactions on Multimedia*, 13(6):1371–1380, 2008.

[5] J. Grace, D. Gruhl, K. Haas, M. Nagarajan, C. Robson, and N. Sahoo. Artist ranking through analysis of online community comments. Technical report, IBM, 2007.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009.

[8] Han, J. and Kamber, M. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2006.

[9] N. Koenigstein and Y. Shavitt. Song ranking based on piracy in peer-to-peer networks. In *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR*, pages 633–638, Kobe, Japan, 2009.

[10] R Development Core Team, R: A Language and Environment for Statistical Computing, Viena, Austria, 2009. [Online]. Available: http://www.R-project.org/

[11] Y. Sun, H. Li, I. G. Councill, J. H. 0002, W.-C. Lee, and C. L. Giles. Personalized ranking for digital libraries based on log analysis. In *Proceedings of the 10th ACM International Workshop on Web Information and Data Management*, pages 133–140, Napa Valley, California, 2008.

[12] G. L. Toledo and I. I. Ovalle. *Estatística Básica*, page 251. Atlas, São Paulo, 2 edition, 1983.

[13] Y.-H. Yang, Y.-C. Lin, A. Lee, and H. Chen. Improving musical concept detection by ordinal regression and context fusion. In *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR*, pages 147–152, Kobe, Japan, 2009.

[14] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen. A regression approach to music emotion recognition. *IEEE*, 16(2):448–457, 2008.