Avaliação de Técnicas de Aprendizado de Máquina em uma Base de Dados da Netflix

Willian Soares Girão

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)
Niterói – RJ – Brazil

williansg@id.uff.br

Resumo. *Este trabalho destina-se a aplicar algoritmos de aprendizado de máquina na base de dados da Netflix, com o objetivo de prever a nota (rating) que um usuário dará à um filme não avaliado utilizando notas para filmes que o usuário avaliou no passado, bem como as avaliações que usuários semelhantes deram aos filmes. Dois algoritmos para clusterização, K-Means e K-Nearest Neighbors, foram utilizados para agrupar usuários semelhates. A etapa de predição acotence após o processo de agrupamento e consiste de uma média dos n pontos mais próximos ao usuári cuja predição deseja-se estimar. Apresentamos e discutimos os resultados obtidos.*

Netflix-prize, Netflix, Ratings, Clusterização, Predição.

1. Introduction

A base de dados escolhida para este trabalho foi da empresa de serviço de streaming Netflix. Trata-se de uma base de dados contendo as avaliações de cerca de 480 mil usuários anônimos da plataforma para mais de 17 mil filmes, gerando em torno de 100 milhões de registros de avaliações. A coleta foi realizada entre outubro de 1998 e dezembro de 2005, e reuniu avaliações feitas aos filmes contidos na plataforma feitas pelos usuários que a utilizaram neste período. As avaliações são feitas em forma de estrelas que podem ser atribuídas aos filmes, variando em quantia de 1 a 5, expressando a opinião do usuário em relação a qualidade do filme. Quanto maior o número de estrelas mais o usuário gostou do filme.

A base de dados foi apresentada na forma de mais de 17.000 arquivos, cada um correspondendo a um filme da plataforma. Cada arquivo contendo registros no formato: "*ID do usuário*" "*Estrelas*", "*Data da avaliação*".

O objetivo deste trabalho é realizar o agrupamento de usuários semelhantes para, posteriormente, prever a avaliação que um usuário dará a determinado filme baseado nas avaliações que usuários semelhantes a ele deram a filmes semelhantes. Foram feitas clusterizações utilizando dois algoritmos, um supervisionado (Knn) e outro não-supervisionado (K-Means). Predições e comparações de desempenhos foram feitas utilizando um subset da base original a fim de se colaborar com novas perspectivas sobre o comportamento de diferentes algoritmos sobre este tipo de base, assim como interpretações dos resultados e aspectos visuais dos dados.

O artigo é construído conforme segue: na seção 2 são apresentados alguns trabalhos relacionados. A seção 3 apresenta os concecitos teóricos que servirão de pressupostos

para classificar os dados. Na seção 4 são apresentados os pré-processamentos necessários, a partir dos pressupostos estabelecidos na seção anterior. Por fim, a seção 8 traz os resultados da aplicação dos conceitos na base; seguinda pela seção **??**, onde são feitas algumas discussões sobre os resultados e pela seção 10 que conclui o artigo com alguns tópicos para trabalhos futuros.

2. Trabalhos Relacionados

O principal trabalho relacionado é o de [Bennett et al. 2007], que é o trabalho mais citado na literatura, e trabalhou com a base de dados da Netflix com o objetivo de bater a precisão do sistema de recomendação da própria Netflix. Para isso foi necessária a formação do seu conjunto de treino, o qual contém subconjuntos de testes que foram selecionados aleatoriamente por usuários que forneceram pelo menos 20 avaliações em determinado período. Em um outro trabalho [Hong and Tsamis 2006] utilizando a mesma base de dados da Netflix, utilizaram a premissa que usuários semelhantes classificam filmes semelhantes de maneira similar. Para isso, utilizaram o KNN por ser um dos algoritmos mais populares quando se trata de sistemas de recomendação, porém, os autores tiveram limitações de tempo, memória e recursos, e com isso não foi muito viável alcançar resultados muito bons.

3. Aspectos Teóricos

Nesta seção serão explicados os pressupostos usados na classificação. Para isso serão expostos os conceitos de *semelhança entre filmes* e *semelhança entre usuários* que foram usados na clusterização, e que, mais tarde, dá ensejo para a predição. Essa seção contém uma subseção tratando de dados faltantes.

Suponha que a Netflix conta com três filmes (*F1*, *F2* e *F3*) e três cliente (*Alice*, *Bob* e *John*). Cada um desses três clientes já assistiram a três desses filmes e deram avaliações conforme mostra a tabela 1. Note que *Alice* e *John* avaliaram os filmes *F1*, *F2* e *F3* de forma semelhante. Quando comparamos *John* com *Bob* vemos que os gostos parecem diferir, dado que as notas para os filmes estão mais distantes.

Para estipular o método para predição, partimos do pressupost que "*usuários semelhantes avaliam filmes de forma semelhante*".

Métodos de clusterização procuram agrupar dados semelhantes em grupos distintos: elementos dentro de um grupo são mais parecidos entre si do que com elementos de outro grupo, e essa semelhança entre indivíduos de cada grupo é estimada a partir do conjunto de características que descrevem esses elementos. Em nossa abordagem, os elementos a serem agrupados são os usuários da base, o cunjunto de características que descrevem esses usuários são seus padrões de avaliações, e os agrupamentos a serem encontrados são os de usuários com gostos parecidos.

Tabela 1. Primeira avaliação dos clientes da Netflix

clientes	F1	F2	F3
Alice	5	1	5
Bob	2	3	2
John	4	1	4

Tabela 2. Avaliações do novo filme *F4*

Filmes	Alice	Bob	John
F1	5	2	4
F2	1	3	1
F3	5	2	4
F4	1	4	**?**

O processo de predição consiste em inferir o rating que o usuário alvo dará com base nos *n* usuários mais próximos. A tabela 2 exemplifica esse processo: A avaliação estimada para *John* para o filme *F4* seria *1*, dado que *Alice* é a usuária com gosto mais similar ao de *John*.

4. Pré-processamentos

O objetivo desta seção é trazer o contexto dos dados: como foram recebidos; e sob quais tratatementos de pré-processamento foram submetidos.

4.1. Contexto dos dados

A base de dados do Netflix, fornecida no desafio Netflix Prize, conta com mais de 100 milhões de avaliações, de mais de 17 mil filmes, de 480 mil clientes anônimos da empresa, escolhidos aleatoriamente. Os dados coletados datam no período de outubro de 1998 a dezembro de 2005, contendo todos os ratings fornecidos durante esse período. As avaliações ("*ratings*") estão em uma escala de 1 a 5, em inteiro. Com o intuito de preservar a privacidade de cada cliente, a empresa submeteu o *ID* de cada cliente a uma função de dispersão, substituindo este valor por um *ID* aleatório. Por fim, também foram fornecidos a data de cada avaliação e o título e ano de lançamento de cada filme.

As avaliações foram organizadas por filmes. A base original disponibiliza 17770 arquivos com o nome "*mv_<ID do filme>.txt*". O conteúdo de cada um destes contém, na primeira linha, o *ID* do filme e nas demais os dados são organizados seguindo o *ID* do cliente, a avaliação deste no filme, e a data em que a tal avaliação foi fornecida.

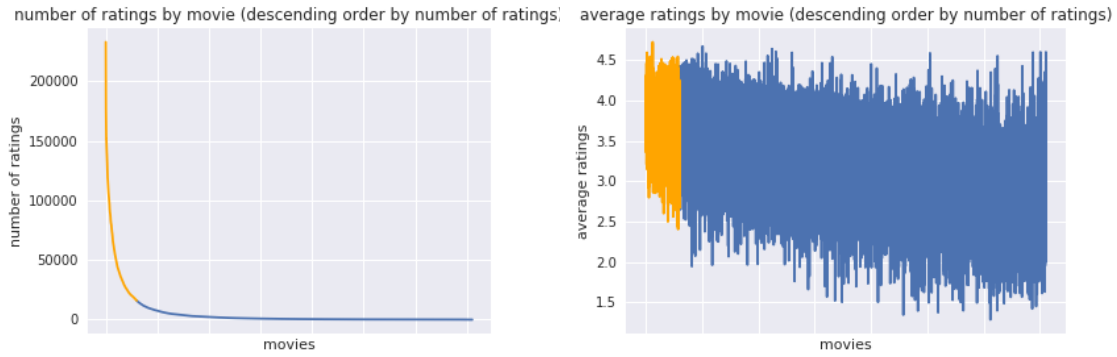
4.2. Pré-processamento

O primeira observação feita sobre os dados foi sobre contagem de quantidade de avaliações recebidas para cada filme, assim como sua avaliação média. Para melhor visualização, apresentamos ambas métricas na figura 1.

O que se percebe é que a base contém muitos filmes com poucas avaliações. Os autores destacam em laranja os 1,5 mil filmes melhor avaliados (com mais de 10.000 avaliações).

Como pode-se ver pelos gráficos, a maior parte dos filmes não possuem informação útil: dado o problema objetivo desse trabalho e a modelagem de sua solução, para evitar pontos esparsos, filmes que não possuem um número expressivo de avaliações foram removidos e somente os 1.500 mais avaliados foram levados em consideração.

O objetivo dessa seleção dos filmes mais avaliados é tornar a matriz de pontos (usuários), que será fornecia aos métodos de clustering, menos esparsa e com a maior



(a) Número de avaliações por filme. **(b) Média de avaliações por filme.**

Figura 1. Contagem de avaliações (a) por filme e avaliação média (b). Cor laranja representa os 1.500 filmes mais avalaidos.

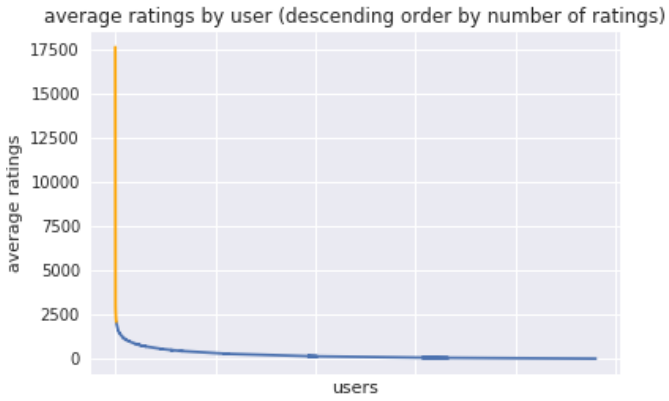


Figura 2. Usuários da base original e suas respectivas contagem de votos. 1.500 usuários mais ativos destacados em laranja.

quantidade de informação útil (filmes avalaidos) possível. Com a mesma finalidade, o mesmo processo de seleção foi feito para os usários. Dentre os usuários da base, selecionamos somente os 1.000, representados pela porção laranja do gráfico na figura 2, que mais avaliaram filmes (usuários mais ativos).

Esses subgrupos de filmes e usuáios constituem a base final que será utilizada para agrupamento e posteriom predição. O total de votos abrangidos por esse dataset é de 1,033,708.

5. Modelagem

Como já explicado, a semelhança que buscamos está em como usuários parecidos avaliam de forma parecida, assim, colocamos cada um desses 1.000 usuários mais ativos como linhas de um matriz onde as colunas são suas respectivas avaliações para cada um dos 1.500 filmes mais votados. O mapa de calor dessa matriz é apresentado na figura 3.

Podemos ver na figura 3 que tanto as linhas quanto as colunas da matriz estão bem preenchidas, o que torna a matriz mais densa. Ao fazer isso, diminuimos a probabilidade de um usuário dentro da matriz não ter um votado em algum dos filmes (característica que

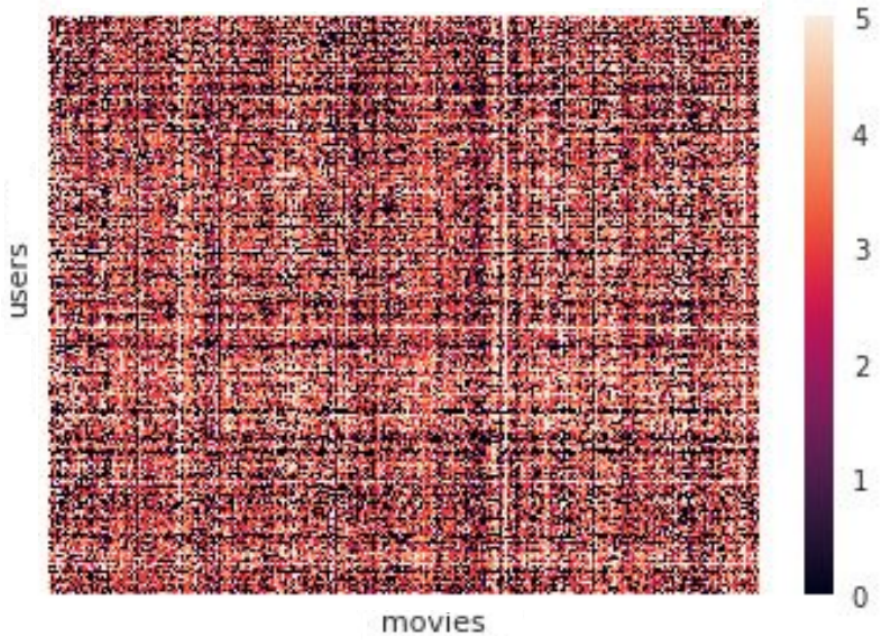


Figura 3. Mapa de calor da matriz de treino.

usaremos para agrupar) e aumentamos a probabilidade do mesmo usuário ter votado nos filmes que o usuário alvo votou.

De modo a explorar melhor a modelagem desenvolvida para verificar se a mesma poderia ser aplicada a outro tipo de predição, um dos testes realizados consiste em inverter a matriz de dados para agrupamento, de modo que os filmes sejam representados como as linhas da matriz. A ideia por trás desse teste é verificar se filmes com conteúdo semelhante podem ser agrupados utilizando as mesmas características de votos.

6. Algoritmos Utilizados

Os algoritmos utilizados no trabalho foram **K-Means** e **DBSCAN**. Ambos são técnicas populares de aprendizado de máquina não supervisionados geralmente utilizados para realização de agrupamento (cluster). Já o **KNN** que é uma técnica de aprendizagem supervisionada. O KNN é usado para classificação e regressão de dados conhecidos onde geralmente o atributo/variável de destino é conhecido de previamente.

6.1. K-Means (K - Médias)

O K-Means (também conhecido como K-Médias) é um algoritmo de clusterização bem simples, ele particiona o conjunto de dados em k clusters, onde o valor de k é fornecido pelo programador. [O. Duda et al. 2001] O algoritmo K-Means fornece uma clusterização de instâncias de acordo com os próprios dados. Este cluster, como será mostrado a seguir, é baseado em análise e comparações entre os valores numéricos das instâncias. Desta maneira, o algoritmo automaticamente fornece uma clusterização automática sem a necessidade de nenhuma supervisão humana, ou seja, sem nenhuma pré-classificação existente. Por causa desta característica, o K-Means é considerado como um algoritmo de mineração de dados não supervisionado, normalmente utilizado em cenários de compre-

ensão de população demográfica, segmentação de mercado, tendências de mídia social, detecção de anomalias, etc., onde os clusters são préviamente desconhecidos.

Neste trabalho foi utilizado o K-Means através do Scikit-learn para realizar agrupamento dos usuários mais semelhantes baseado nas avaliações feitas por eles aos filmes presentes no dataset, para posteriormente ser feita predição da avaliação que este usuário dará a um filme "M"de acordo com as avaliações feitas pelos "N"usuários mais semelhantes a ele a filmes semelhantes.

6.2. KNN (K — Nearest Neighbors)

O KNN(K — Nearest Neighbors) é um dos muitos algoritmos (de aprendizagem supervisionada) usados no campo de data mining e machine learning, ele é um classificador onde o aprendizado é baseado “no quão similar” é um objeto (um vetor) do outro. O algoritmo classifica um novo objeto com base nos k elementos do conjunto de treinamento mais próximos a ele. [Mucherino et al. 2009]

7. Experimentos

Dois tipos de experimentos foram realizados neste trabalho. O primeiro grupo de experimentos consiste em avaliar a performance da abordagem desenvolvida para o problema de predição de voto para um dado filme. O segundo grupo de experimentos consiste em verificar se a mesma modelagem pode ser utilizada para encontrar filmes similares.

7.1. Predição de Voto

Para cada um dos algoritmos utilizados, dois testes com diferentes composições de usuários foram realizados. O primeiro teste consiste em avaliar a predição que 100 usuários testes darão a um filme. Para esse teste, foram selecinados usuários que tiveram contagem de votos entre 99 - 380 (considerando todos os filmes da base original). O segundo teste também consiste em avaliar a predição que 100 usuários testes darão a um filme, com o diferencial desses 100 usuário serem mais ativos, tendo contagem total de votos variando entre 998 - 1000 (considerando todos os filmes da base original).

7.2. Predição de Filme Similar

Para avaliar a capacidade de nossa bordagem em encontrar filmes semelhantes, três filmes aleatórios (que não se encontram nos dados de treino) foram selecionados. Para cada um desses filmes, utilizamos o método K-Means para encontrar grupos de filme avaliados de forma semelhante. O teste consiste em descobrir a qual dos grupos o filme alvo pertence e, então, encontrar o filme dentro do mesmo grupo o filme que esteja mais próximo e verificar se ambos são semelhantes.

Para esse grupo de testes, o valor do parâmetro K para o algoritmo K-Means escolhido foi 10.

7.3. Heurísticas de Predição e Parâmetros

A heurística básica de predição de voto em conjunto com os dois algorítmos consiste em utilizar informação dos n pontos mais próximos ao usuário alvo.

7.3.1. K-Menas

Nos testes com o K-Means, o a avaliação predita consistes da média aritmética dos votos dos *n* pontos mais próimos ao usuário alvo para o filme que se quer predizer. Caso essa média seja 0, isso significa que nenhum desses *n* pontos votou no filme alvo da predição e, nesse caso, um número randômico entre 1 e 5 é gerado como predição feita.

O valor de K utilizado nos esperimentos foi **6**. Para estimarmos esse parâmetro do algorítmo, fizemos uma análise do mapa de calor, assim como das métricas que nos levaram a selecionar os subgrupos de uruário e filme finais. A **distância euclidiana** foi a métrica de distância escolhida.

7.3.2. Knn

Nos testes com o Knn, o a avaliação predita consistes da moda (valor mais frequente) do valor de avaliação dos *n* pontos mais próimos ao usuário alvo para o filme que se quer predizer. Caso essa média seja 0, isso significa que nenhum desses *n* pontos votou no filme alvo da predição e, nesse caso, um número randômico entre 1 e 5 é gerado como predição feita. A **distância euclidiana** foi a métrica de distância escolhida.

8. Resultados

Nesta sessão, são apresentados os resultados dos testes que foram realizados tendo a base processada como dados de treino. Comentários e interpretações serão apresentados na próxima sessão.

8.1. Predição de Voto

Na tabela 3 aprensentamos os resultados para os experimentos executados.

Como se pode observar, embora o algoritmo K-Means tenha tido resultados de acurácia melhor que os alcançados pelo Knn em ambos as configurações de teste, ambos os modelos de predição desenvolvidos possuem desempenho não satisfatório. Possíveis motivos e interpretações serão discutidos na próxima sessão.

8.2. Predição de Filme

Na tabela 4 aprensentamos os resultados para os experimentos executados.

Enquanto o filme *Something's Gotta Give* é uma comédia romântica e *Rudy* é um filme biográfico de sports, conteúdos aparentemente disassociados, tanto *Duplex (Widescreen)* quanto *Much Ado About Nothing* são comédias românticas com mesmo ano de

Tabela 3. Resultados dos experimentos

Knn	Total de predições	Acurácia
Teste 1	100	0,15
Teste 2	100	0,26
K-Means	Total de predições	Acurácia
Teste 1	100	0,26
Teste 2	100	0,38

Tabela 4. Resultados dos experimentos

Filme Alvo	Ano	Similar	Ano
Lilo and Stitch	2002	*Spy Kids 2: The Island of Lost Dreams*	2002
Something's Gotta Give	2003	*Rudy*	1993
Duplex (Widescreen)	1993	*Much Ado About Nothing*	1993

lançamento, enquanto *Lilo and Stitch* e *Spy Kids 2: The Island of Lost Dreams* são ambos filmes infantis (também com o mesmo ano de lançamento).

9. Comentários e Interpretações

9.1. Predição de Voto

Ambos os métodos utilizados nesse trabalho se mostraram ineficazes para o problema de predição de voto. Mas alguns dois devem ser destacados.

Em primeiro lugar, pelo mapa de calor apresentado na figura 4, o algoritmo K-Means aparenta ter agrupado de forma adequada os usuários, como podemos verificar pelas distintas mudanças de padrão de coloração no eixo dos usuários, assim como as linhas com distribuição de cor homogêneas no eixo dos filmes, o que significa os usuários daquele grupo votaram de forma muito semelhante entre si. Isso pode significar que, ao tentar clusterizar um novo usuário, esse usuário não possui informações de ratings suficientes para fazer uma boa estimativa de qual clustes ele pertence, o que parece condizer com fato de que menos de 1% de todos os usuários da base original são usuários ativos. Para testar essa cenário, fizemos o segundo conjunto de testes de predição utilizando um grupo de 100 usuários com maior contagem de voto quando comparados aos usuários do primeiro teste. De fato, ambos os métodos obtiveram melhoria considerável em suas acurácias. Isso pode significar que métodos de predição baseados em clustering não obtêm bom desempenho quando trabalham com vetores de características de dimensão muito alta e cujos novos pontos a serem agrupados necessitem da presença da maior parte dessas características para poderem ser agrupados de forma viável.

O segundo ponto a ser notado é a métrica de distância utilizada: é conhecido que, para espaços de dimensão muito alta, a distância euclidiana não é a recomendada. Para a dimensão do vetor de características estipulado em nossa modelagem talvez resultados melhores possam ser obtidos utilizando a métrica *semelhança cosseno*.

9.2. Predição de Filme

Apesar de, dentre as três predições feitas, duas tenham aparentemente acertado ao indicar filmes semelhantes, há a necessidade de serem executados não só um maior número de testes, mas também avaliar a real semelhança entre os filmes para além do gênero.

Uma possível interpretação de como os pontos estejam sendo agrupados está no quão famoso um filme é (em quantia de votos recebidos) e os usuários mais ativos: talvez muitos usuários da base de teste, por serem mais ativos, tenham gêneros de filmes melhor definidos e isso esteja refletido em seus votos, portanto uma possível relação encontrada pelo método seja justamente um agrupamento de bons representantes de alguns gêneros de filmes (um exemplo de bom representante para o gênero *sci-fi* seria o filme *Matrix*, enquanto que *Senhor dos Anéis* seria um bom representante do gênero *medieval*), que são

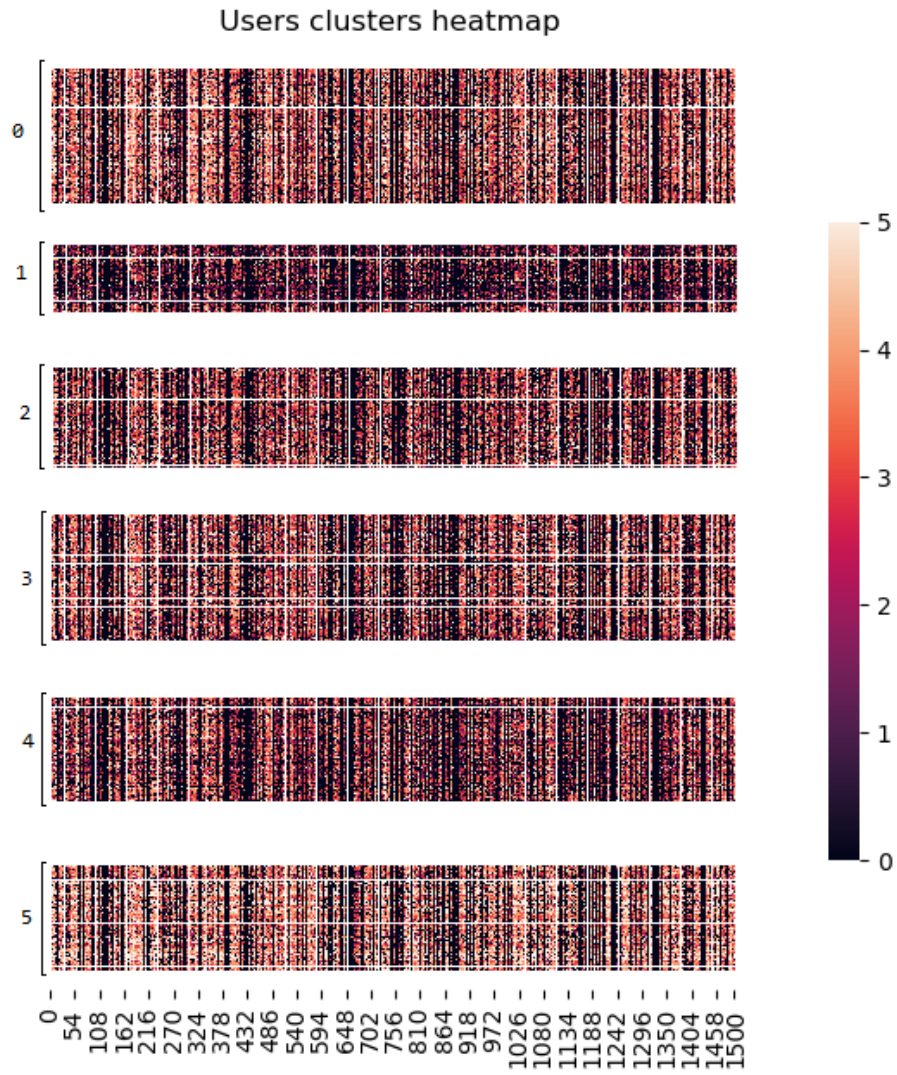


Figura 4. Mapa de calor dos grupos encontrados pelo algoritmo K-Means.

justamente filmes muito avaliados. Essa seria uma possível interpretação para as duas predições "corretas"desse experimento.

10. Conclusão

Neste trabalho foram aplicados algoritmos de aprendizado de máquina sobre os dados da avaliação que clientes davam aos filmes da Netflix. Para isso, os dados foram submetidos a uma fase de clivagem, com o intuito de obter informações mais relevantes, e, posteriormente, foram agrupados com o objetivo de refletir padrões de preferência.

Apesar de resultados iniciais apontarem que as técnicas aplicadas não são adequadas para os dois tipos de predição que buscamos (pelo menos quando utilizamos votos como características), informações visuais e resultados empíricos (grupo de usuários mais ativos nos experimentos de predição de voto) apontam que talvez a baixa qualidade de resultados esteja na qualidade dos pontos utilizados para teste, ou seja, tentativas de agrupamentos de usuários que não possuem muita informação útil (não votaram para muitos dos 1,500 filmes selecionados) para comparação.

Novos testes com outras bases de dados e mudanças nas métricas de distâncias devem ser executados para reavaliar valores de acurácia e validar as interpretações aqui colocadas.

Referências

Bennett, J., Lanning, S., and Netflix, N. (2007). The netflix prize. In *In KDD Cup and Workshop in conjunction with KDD*.

Hong, T. and Tsamis, D. (2006). Use of knn for the netflix prize. *CS229 Projects*.

Mucherino, A., Papajorgji, P. J., and Pardalos, P. M. (2009). *k-Nearest Neighbor Classification*, pages 83–106. Springer New York, New York, NY.

O. Duda, R., E. Hart, P., and G.Stork, D. (2001). *Pattern Classification*, volume xx.