Censo 1994 – Quem ganha mais de 50 mil?

Celso M. Araujo Filho, Renan F. Almeida, Vinnícius F. Silva
Departamento de Computação (DComp)
Universidade Federal de São Carlos (UFSCar)
18052-780, Sorocaba, São Paulo, Brasil
celsofilho95@gmail.com, renanfalmeida@outlook.com, vinniciusfs@hotmail.com

***Resumo*—Na área de Aprendizado de Máquina, a utilização de censos demográficos em problemas de classificação tem se mostrado cada vez mais interessante e extremamente benéfica para pesquisas. A proposta desse trabalho é apresentar a análise de diferentes métodos de aprendizado de máquina utilizados em problemas de classificação para determinar quais indivíduos possuem renda maior a 50 mil, com base nos diversos atributos coletados para cada indivíduo. Para os experimentos, foi utilizado um censo real, realizado em 1994 nos Estados Unidos. Os métodos de classificação foram validados com a técnica de validação cruzada com 5 partições. Os métodos de Redes Neurais Artificiais e Máquinas de Vetores de Suporte apresentaram o melhor desempenho.**

***Palavras-chave*-censo; renda; classificação.**

I. Introdução

Censos são instrumentais para o desenvolvimento do conhecimento humano, consistindo na aquisição e recordação de informações sobre membros de uma dada população, tais como censos de agricultura, indústria e motoristas. Seu uso é variado, podendo ser usado em pesquisas, *marketing* e planejamento de negócios.

Através de um censo, é possível identificar características entre os indivíduos da população, e usar esse conhecimento várias maneiras, como estabelecer uma relação causal entre certos atributos.

Um dos censos mais realizados é o de renda, familiar ou *per capita*. Através destes, é possível criar índices de desigualdade de renda entre os habitantes da região estudada pelo senso.

De tal maneira, esse trabalho propõe a aplicação de diferentes métodos de aprendizado de máquina para classificação, a fim de identificar, a partir de uma série de atributos, quais indivíduos possuem uma renda anual superior a 50 mil, utilizando os métodos k-vizinhos mais próximos, regressão logística, redes neurais artificiais e máquinas de vetores de suporte, que são métodos consagrados na área de problemas de classificação.

Dentre as técnicas utilizadas, temos a busca em *grid*, também conhecida como otimização de hiperparâmetros, o *holdout*, uma simples divisão dos dados originais em conjuntos de treinamento e de teste seguindo uma proporção definida, e a validação cruzada com k partições. O número escolhido para esse projeto foi $k = 5$, um número adequado para bases de dados de tamanho pequeno a médio.

II. Base de Dados

A base de dados utilizada no trabalho é um censo realizado nos Estados Unidos, em 1994, pelo *United States Census Bureau*[1]. Trata-se de uma amostra reduzida dos dados reais, parte do repositório de Aprendizado de Máquina da UCI, disponível no *website* Kaggle[2].

Há 32561 amostras na base, distribuídas em duas classes: quem ganha menos que 50 mil (75,9%) e aqueles que ganham mais de 50 mil (24,1%). Todas as amostras possuem modestos 14 atributos com dados demográficos. Destes, 6 atributos são numéricos, e os outros 8 são nominais. Os atributos são descritos a seguir.

- Idade, com valores de 17 até 90 anos;
- Setor de trabalho: setor privado, nunca trabalhou, servidor público, etc.;
- *Final Weight*: Um peso calculado pelo *U.S. Census Bureau*, indicando o grupo social do indivíduo.
- Educação e número de educação: O grau de educação do cidadão, de pré-escola até doutorado, e o respectivo número, de 1 a 16;
- Estado Conjugal: solteiro, casado, viúvo, etc.;
- Ocupação: o cargo do indivíduo;
- Relacionamento familiar;
- Grupo étnico;
- Sexo: Masculino ou Feminino;
- Excedente de capital;
- Déficit de capital;
- Número de horas de trabalho por semana
- País Nativo.

Os métodos de classificação utilizados não foram ajustados para compatibilidade com atributos nominais. Assim, foi realizada a transformação dos atributos nominais em atributos numéricos através da atribuição de um número inteiro para cada valor possível, a partir do número 1.

Para o atributo de Educação, a transformação de Educação em números seria redundante com o atributo Número de Educação. Com isso, decidiu-se descartar o atributo Educação, visto que o Número de Educação é suficiente para os propósitos de informação.

[1] www.census.gov/
[2] www.kaggle.com/uciml/adult-census-income

Há, no entanto, amostras com valores desconhecidos, indicados por '?', totalizando 2399 (7,37%) amostras incompletas. Entre as opções de deleção de amostras e preenchimento dos valores, optou-se por atribuir um número diferente dos outros para os valores desconhecidos.

As transformações foram feitas através de substituição de texto. Um arquivo alternativo à base original de dados está na pasta do projeto. Para mais detalhes sobre os valores atribuídos a cada campo, ver o arquivo "Campos e possíveis valores.txt".

III. Metodologia Experimental

Neste trabalho foram avaliados os seguintes métodos de classificação: *k*-vizinhos mais próximos (KNN), regressão logística (RL), redes neurais artificiais (ANN) e máquinas de vetores de suporte (SVM).

Todos os métodos citados têm desempenho sensível ao ajuste de seus parâmetros. Para uma escolha adequada ao conjunto de dados utilizado, foi utilizada a técnica de busca em *grid*, com variações pré-determinadas nos parâmetros. Nesta fase, foi feito um *holdout* dos dados, criando uma divisão em conjunto de treinamento (70% dos dados) e conjunto de teste (30% dos dados).

Por questões de simplicidade e confiabilidade de escolha, a escolha de amostras no *holdout* não é feita aleatoriamente, mas para as mesmas amostras toda vez que o programa é executado, considerando que o conjunto de dados não seja alterado.

O critério de escolha dos parâmetros é descrito a seguir.

- O valor de K no método k-vizinhos mais próximos, testado com valores ímpares de 1 até 51.
- O valor de λ, o fator de regularização do método de regressão logística, variando de 10^{-10} até 10^{10}, com incrementos de 1 na potência.
- No método Redes neurais, o fator de regularização λ, com valores de 0 a 10, incrementados de 1 unidade, e o número de nós na camada oculta, de 6 a 11, com incrementos de 1 unidade.
- No método SVM, foram testados os parâmetros C, para valores de 0,5 até 8, com cada iteração dobrando o valor anterior, e valores para γ de 0,5 a 8, seguindo o mesmo crescimento de C Cada valor de C foi testado com todos os valores de γ, totalizando 27 testes.

Realizada a busca em *grid*, foram selecionados e utilizados os seguintes parâmetros.

- k-vizinhos mais próximos – $K = 25$;
- Regressão logística – $\lambda = 10^4$;
- Redes neurais – $\lambda = 2$, *hidden layer size* = 6;
- SVM – $C = 2$, $\gamma = 0,5$.

Para a implementação do método de máquina de vetores auxiliares, foi utilizada a biblioteca LIBSVM para MATLAB/Octave.

Na fase experimental, não é utilizado o *holdout*, mas sim uma validação cruzada com 5 partições. Assim, cada execução treina e testa cinco vezes e toda amostra é utilizada no conjunto de treino e, em outro momento, utilizada no conjunto de teste.

Considerando o uso da validação cruzada, a medida final de desempenho será a média de acurácia – a porcentagem de amostras corretamente classificadas – de cada uma das cinco partições de conjunto de teste utilizadas.

Tabela I
Acurácia do Método de Regressão Logística

Partição	1	2	3	4	5	**Média**
Acurácia (%)	55,4	81,6	80,9	80,9	80,8	**75,9**

Tabela II
Acurácia do Método de Redes Neurais Artificiais

Partição	1	2	3	4	5	**Média**
Acurácia (%)	81,0	74,6	74,7	79,0	74,7	**76,8**

IV. Resultados

O método k-vizinhos mais próximos não pôde ser corretamente avaliado. Por problemas na implementação, o algoritmo está ineficiente e não parece produzir resultados para conjuntos que não sejam bastante pequenos.

A acurácia do método de regressão logística, com $\lambda = 10^4$ pode ser observada na Tabela I.

Para a validação do método de redes neurais artificiais, com fator de regularização $\lambda = 2$ e número de nós na camada oculta = 8, a acurácia pode ser vista na Tabela II.

Os resultados de acurácia do método SVM podem ser notados na Tabela III, com os parâmetros $C = 2$ e $\gamma = 0,5$.

Considerando o contexto do problema, o desempenho é satisfatório, mas para aplicações mais cruciais, onde o erro é severamente penalizado, seriam necessários ajustes nos algoritmos e nos dados.

V. Conclusões

Dentre os métodos corretamente testados, o melhor foi o SVM, com acurácia 5% maior que os outros métodos testados, o que o torna a melhor escolha para uma alta acurácia de classificação.

Se a qualidade desejada é consistência, o método recomendado é o de Redes Neurais Artificiais. Sua acurácia total não excepcional comparada com a de outros métodos, mas é o método mais consistente, com menor desvio nos valores, e sua acurácia para bases maiores é mantida, ainda que o desempenho sofra.

Por outro lado, dada a simplicidade do método de regressão logística e seu desempenho sólido, é também uma escolha recomendável para o problema

Uma possível medida para se melhorar o desempenho seria uma transformação mais criteriosa dos atributos descritivos em numéricos. Atributos nominais não possuem ordem entre si – como 'azul', 'branco' –, então atribuir

Tabela III
ACURÁCIA DO MÉTODO MÁQUINAS DE VETORES AUXILIARES

Partição	1	2	3	4	5	**Média**
Acurácia (%)	55,4	81,6	80,9	81,0	80,8	**80,8**

Tabela IV
ACURÁCIA DO MÉTODO K-VIZINHOS MAIS PRÓXIMOS

Partição	1	2	3	4	5	**Média**
Acurácia (%)	0	0	0	0	0	**0**

números sequenciais a estes atributos pode causar hipóteses fracas.

Atributos com muitos valores possíveis, como País, poderiam antes ser agrupados em blocos – como por regiões –, e depois codificados em cadeias de *bits*, de maneira que a distância de Hamming entre todas as amostras fosse 1.

APŁNDICE

Para a utilização do projeto, basta executar o arquivo Grupo_05.m. Imediatamente, o programa imprime algumas mensagens, indicando o carregamento e pré-processamento dos dados. Em seguida, imprime-se o menu principal, onde o usuário deverá escolher uma das opções seguintes.

- Teste de Parâmetros;
- Classificação dos Dados;
- Sair.

As opções são representadas pelos números 1, 2 e 0, respectivamente. Enquanto o usuário não digitar um desses números, uma entrada válida continuará sendo pedida.

A opção 1 refere-se à escolha dos parâmetros de cada método, sendo realizada uma busca em *grid*, como detalhado anteriormente. Ao selecionar essa opção, o usuário deverá escolher um dos quatro métodos de classificação. A acurácia do método para diferentes valores será imprimida na tela, e o usuário será levado de volta ao menu principal.

Ao escolher a opção 2, o usuário deverá escolher um dos quatro métodos de classificação, que serão utilizados para classificar o conjunto de dados. A acurácia final do método será exibida na tela, e o usuário volta para o menu principal. É importante notar que a opção 2 é independente da opção 1. A escolha dos parâmetros para a classificação já foi codificada com base nos experimentos.

Satisfeito com os testes realizados, selecionar a opção Sair do menu principal finalizará a execução do programa.

REFERŁNCIAS

[1] J. M. Campanha, J. V. Lochter e T. A. Almeida, *Detecção automática de spammers em redes sociais*, Anais do XI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC '14), São Carlos, Brasil, 2014