Daniel Pereira Cinalli

Estudo de métodos de ensembles com seleção de classificadores dinâmica

Projeto de Graduação em Computação submetido à Universidade Federal do ABC para a obtenção dos créditos na disciplina Projeto de Graduação em Computação I do curso de Ciência da Computação

Orientador: Prof. Dr. Thiago Ferreira Covões

Universidade Federal do ABC

19 de Novembro de 2020

RESUMO

Ensembles são utilizados pela sua maior capacidade de generalização e também melhor acurácia, quando comparados à utilização de um classificador apenas para a mesma tarefa. Uma abordagem possível é a seleção dinâmica dos classificadores de um ensemble, em que é estimado quais classificadores possuem um melhor desempenho, que então são utilizados para realizar a classificação. Dois métodos de seleção dinâmica serão estudados. O primeiro escolhe os classificadores baseado na acurácia ao classificar objetos conhecidos próximos do objeto que se deseja classificar. O segundo utiliza o conceito de similaridade de decisão, escolhendo uma porcentagem dos classificadores que mais concordam na classificação de uma quantidade escolhida de objetos, gerados aleatoriamente na proximidade do objeto que se quer classificar.

SUMÁRIO

CAPÍTULO 1

INTRODUÇÃO

Em aprendizado de máquina, modelos preditivos são gerados à partir de um conjunto de dados, para que se possa fazer predições. Um modelo pode ser um classificador, caso se deseje realizar uma predição categórica, ou um regressor, quando o que se quer é uma predição numérica [Zhou 2012]. O foco deste estudo será em classificação, que de acordo com [Tan, Steinback e Kumar 2006], é a tarefa de se aprender uma função f que mapeia cada conjunto de atributos x para uma das classes predefinidas y.

O desempenho de generalização de um classificador é definido como seu desempenho em classificar objetos não utilizados durante a fase de treinamento. Uma motivação do uso de *ensembles* é sua maior capacidade de generalização, se comparado à utilização de um único classificador para a mesma tarefa. Para obter isto é feita a combinação das predições dos diversos classificadores do *ensemble*, o que reduz o efeito daqueles classificadores que apresentam um desempenho insatisfatório [Polikar 2006].

Entre os métodos de combinação para predições numéricas, pode-se citar os métodos de média simples e média ponderada. Para predições nominais, alguns métodos são voto da maioria, em que é necessário que pelo menos um dos resultados possíveis tenha recebido mais de metade dos votos; voto plural, onde simplesmente é escolhido aquele com mais

votos; votação ponderada, entre outros. Existem, no entanto, diversos outros modos de se combinar os resultados de um *ensemble* [Zhou 2012].

Dado que um *ensemble* considera a predição de diversos classificadores, é desejável que este conjunto de classificadores possua alta diversidade. Caso os classificadores individuais sejam similares, não haverá um ganho grande ao se combinar seus resultados. Ou seja, os classificadores individuais que fazem parte do *ensemble* devem ter uma baixa correlação entre si. Classificadores com alta correlação entre si em geral diminuem o erro do *ensemble* menos se comparado a um *ensemble* onde estes possuem baixa correlação entre si, e se reduz ainda mais no caso de uma correlação negativa [Kuncheva e Whitaker 2003].

Uma alta diversidade portanto deve levar a uma melhor acurácia do *ensemble*, mas mensurar esta diversidade é difícil, e diversos métodos existentes não chegam a uma medida de diversidade que apresente boa correlação com a acurácia do *ensemble*. É entendido que esses métodos não sejam bons indicadores de diversidade, mas que a busca por diversidade ainda é importante [Zhou 2012].

Em classificação, existem também técnicas de seleção dinâmica de classificadores. Em seleção dinâmica, é feita uma estimação daqueles classificadores com melhor desempenho, que então são utilizados para realizar a classificação [Cruz, Sabourin e Cavalcanti 2018]. Neste trabalho, será estudado o desempenho de dois métodos onde os classificadores de um ensemble são escolhidos dinamicamente, que serão introduzidos a seguir.

O primeiro método escolhe os melhores classificadores de acordo com a acurácia local, com base nos objetos de seu conjunto de treinamento e validação. Para isso, se escolhe os objetos mais próximos do objeto que se deseja classificar, e se determina a acurácia com base nestes. Com base nesses resultados, se determina quais classificadores serão utilizados para fazer a classificação.

O segundo método será baseado no artigo [Kurvers et al. 2019], onde foi analisada a ideia de se utilizar a similaridade de decisão para se encontrar indivíduos ou grupos com alto desempenho. Neste, a similaridade de decisão entre diferentes indivíduos foi usada para predizer a acurácia destes. Em [Cruz, Sabourin e Cavalcanti 2018], é proposto que enquanto é necessário ter diversidade entre os classificadores disponíveis em todo o ensemble,

se deve ter consenso, e não diversidade, entre os classificadores selecionados para a tarefa de predição.

A similaridade de decisão é medida apresentando um conjunto de perguntas que cada indivíduo deve responder. Para cada indivíduo, se calcula a porcentagem de concordância em relação a cada outro indivíduo, ou seja, o quanto concordam nas respostas às perguntas. A similaridade de decisão de cada indivíduo será a média das porcentagens de concordância calculadas para este. Tanto na análise matemática quanto na empírica, se concluiu que a abordagem é viável, encontrando grupos com melhor desempenho [Kurvers et al. 2019].

Diferentemente do método anterior, que busca objetos próximos já existentes, os objetos serão gerados aleatoriamente, próximos ao objeto que se quer classificar, e a decisão de quais serão os classificadores escolhidos será baseado na similaridade de decisão.

Embora estes métodos estejam escolhendo apenas alguns classificadores para realizar a predição, o objetivo é identificar bons classificadores localmente, e ainda é importante que o *ensemble* como um todo possua alta diversidade.

CAPÍTULO 2

JUSTIFICATIVA

O presente trabalho busca explorar duas técnicas em que se realiza a seleção dinâmica de classificadores em um ensemble. Essa técnica de seleção dinâmica tem como objetivo aumentar a acurácia de uma predição, selecionando por algum critério quais classificadores do ensemble serão utilizados para a predição sendo realizada para um dado objeto. Deseja-se determinar se as técnicas estudadas são viáveis como uma ferramenta para se obter uma melhor acurácia para um ensemble.

A técnica que utiliza o critério de similaridade de decisão foi baseada em [Kurvers et al. 2019]. No artigo se mostra que este critério consegue obter uma maior acurácia selecionando um subconjunto de todos os indivíduos de um grupo para se fazer uma predição, se comparado a utilizar o conhecimento combinado de todos. Dado o resultado positivo apresentado no artigo, a técnica será implementada em um ensemble para determinar seu desempenho no contexto de aprendizado de máquina.

CAPÍTULO 3

FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os fundamentos de *ensembles* e validação cruzada, assim como os classificadores que serão utilizados como base para os *ensembles*.

3.1 Validação cruzada

Em validação cruzada, um conjunto de dados é particionado em k partes de mesmo tamanho. Uma dessas partições será o conjunto de teste enquanto as outras $k-1$ partes serão usadas para treinamento. Isso é repetido até todas as k partições terem sido utilizadas como conjunto de teste [Tan, Steinback e Kumar 2006].

3.2 Ensembles

Ensembles consistem de um conjunto de classificadores base treinados no conjunto de treinamento. A classificação realizada pelo *ensemble* é realizada por votação das predições feitas pelos classificadores base. Esse método tem como objetivo aumentar a acurácia de classificação, se comparado ao utilizar um de seus classificadores base apenas. Para se conseguir isso, há duas condições necessárias:

- Os classificadores base devem ser independentes entre si;
- Os classificadores base devem ter desempenho melhor que uma escolha aleatória.

Na primeira condição, dois classificadores serem independentes significa que seus erros não são correlacionados entre si. Na segunda condição, se implica que cada classificador precisa ter uma taxa de erros menor que 50% [Tan, Steinback e Kumar 2006].

3.3 Árvore de decisão

Uma árvore de decisão é composta de nós de decisão e de nós folha. Em cada nó de decisão, é implementada uma função de teste com saídas discretas, em que o resultado do teste determina qual o próximo nó escolhido. Isso ocorre até se encontrar um nó folha, que é um nó em que se encontra uma das possíveis saídas. Para se gerar uma árvore de decisão para classificação, se quantiza o quão boa é uma divisão criada por um nó de decisão usando uma medida de impureza. Uma medida possível é a função de entropia, que para um nó m e exemplo i é dada por [Alpaydin 2010]:

$$I_m = -\sum p_m^i log_2 p_m^i \tag{3.1}$$

Uma divisão pura, por exemplo, é uma em que a distribuição das classes após a divisão será (0, 1). A construção da árvore é feita se escolhendo a decisão com entropia mínima. Isso é feito recursivamente até que todas as divisões sejam puras. O tamanho do espaço de busca faz com que seja inviável de se encontrar uma árvore ótima, e por isso algoritmos de construção de árvores de decisão usualmente utilizam uma estratégia gulosa que faz decisões localmente ótimas [Tan, Steinback e Kumar 2006].

3.4 Perceptron

Um perceptron possui n entradas e 1 saída. Para cada entrada x_j, está associado um peso w_j, e também um termo de *bias* w_0. O caso mais simples para se calcular a saída é a média ponderada, apresentada abaixo:

$$y = \sum_{j=1}^{n} w_j x_j + w_0 \tag{3.2}$$

O termo w_0 pode ser visto como se fosse associado a um termo x_0, com valor sempre igual a 1. A saída do perceptron pode então ser representada pelo produto escalar abaixo:

$$y = \mathbf{w}^T \mathbf{x} \tag{3.3}$$

em que $\mathbf{x} = [1, x_1, ..., x_n]$ e $\mathbf{w} = [1, w_1, ..., w_n]$. Durante o treinamento, se aprendem os pesos **w**. Em treinamento online inicia-se com pesos aleatórios, a saída y é calculada para a primeira instância da entrada, os pesos são atualizados e o processo é repetido até que toda instância tenha passado pelo perceptron. Abaixo é mostrado o valor da atualização dos pesos, onde α é a taxa de aprendizado, $y^{(t)\prime}$ é a predição, $y^{(t)}$ é o valor real, e $x_j^{(t)}$ é a entrada, para a instância t:

$$\Delta w_j^{(t)} = \alpha(y^{(t)} - y^{(t)\prime}) x_j^{(t)} \tag{3.4}$$

3.5 Classificador Naïve Bayes

O teorema de Bayes é utilizado para se determinar a probabilidade de um evento Y ocorrer dado que se conhece que o evento X já ocorreu, e é dado pela seguinte equação:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{3.5}$$

Para se utilizar o teorema de Bayes para classificação, é necessário aprender as probabilidades a posteriori $P(Y|\mathbf{X}$, em que Y é a classe e $\mathbf{X}$ é o conjunto de atributos, à partir dos dados de treinamento. Sabendo essas probabilidades e sendo $\mathbf{X}'$ o objeto que quer se classificar, se encontra a classe Y' que maximiza a probabilidade a posteriori $P(Y'|\mathbf{X}'$. O teorema de Bayes é utilizado então para se expressar esta probabilidade em função de $P(Y)$ (probabilidade a priori), $P(\mathbf{X}|Y$ (probabilidade condicional de classe), e $P(X)$ (evidência) [Tan, Steinback e Kumar 2006].

Em uma classificador Naïve Bayes, se supõe que os atributos são condicionalmente independentes. Com essa suposição se tem a seguinte relação:

$$P(\mathbf{X}|Y = y) = \prod_{i=1}^{n} P(X_i|Y = y) \tag{3.6}$$

Com essa suposição, não é necessário computar a probabilidade condicional de classe para cada combinação de **X**, apenas computando a probabilidade condicional para cada X_i, dado um Y. Com isso, para se classificar um objeto, se usa a seguinte equação para se obter a probabilidade a posteriori de cada classe:

$$P(Y|\mathbf{X}) = \frac{P(Y)\prod_{i=1}^{n} P(X_i|Y)}{P(\mathbf{X}} \tag{3.7}$$

É suficiente se maximizar o numerador para se encontrar a classe do objeto, já que $P(\mathbf{X})$ será igual para todo Y.

3.6 Classificador kNN

O classificador kNN (*k-nearest neighbours*) computa a distância entre o objeto **X** que se deseja classificar e cada exemplo do conjunto de treinamento, e então seleciona os k mais próximos dentre estes, gerando o conjunto denotado por $D_{\mathbf{X}}$. O processo para se escolher a classe consiste então de uma votação majoritária, onde a classe em maior número em $D_{\mathbf{X}}$ é a predição para o objeto **X** [Tan, Steinback e Kumar 2006].

CAPÍTULO 4

METODOLOGIA

Neste capítulo serão apresentados em maior detalhe como os métodos estudados funcionam.

4.1 Método de seleção por acurácia

O primeiro método escolhe os melhores classificadores de acordo com a acurácia local, com base nos objetos de seu conjunto de treinamento e validação. Possui dois parâmetros: N, o número de objetos próximos que serão considerados; e $Acc_{\min}$, a acurácia mínima necessária para se aceitar um classificador. O processo para se classificar um objeto P é descrito abaixo:

1. Escolhe-se os N objetos mais próximos do objeto que se quer classificar;
2. Calcula-se a acurácia de cada classificador com base nesses objetos;
3. Se determina quais classificadores serão utilizados na classificação, filtrando aqueles com acurácia menor que $Acc_{\min}$;
4. Pesos são dados a cada classificador, proporcionais à sua acurácia;

5. É feita uma média ponderada das classificações de cada classificador escolhido, e o resultado é normalizado.

4.2 Método de seleção por similaridade

Como explicado anteriormente, neste método os objetos serão gerados aleatoriamente, próximos ao objeto que se quer classificar, e a decisão de quais serão os classificadores escolhidos será baseado na similaridade de decisão. Os parâmetros para este método são: N, a quantidade de objetos aleatórios que serão gerados; M, a quantidade de classificadores que serão escolhidos. O processo para se classificar um objeto Pé o seguinte:

1. São gerados N objetos próximos a P;
2. Cada classificador então classifica estes objetos;
3. Cada classificador tem sua similaridade de decisão calculada;
4. Os M classificadores com maior similaridade de decisão são escolhidos;
5. O objeto Pé classificado considerando apenas estes classificadores.

O processo para se gerar objetos próximos a P é explicado a seguir. Se sabe o desvio padrão original σ_i para cada atributo i, obtidos do conjunto de treinamento. Para se gerar um dos objetos aleatórios próximo a P, cada atributo i seguirá uma distribuição normal, com média igual ao valor para aquele atributo em P, e com desvio padrão igual a 10% de σ_i.

CAPÍTULO 5

CRONOGRAMA

Neste capítulo é apresentado o cronograma das atividades, na Tabela 5.1 abaixo.

(A) - Levantamento Bibliográfico - Pesquisa e leitura do material necessário para a fundamentação teórica;

(B) - Implementação - Desenvolvimento dos dois métodos;

(C) - Aplicação - Aplicação dos métodos em conjuntos de dados;

(D) - Análise de Resultados- Analisar os resultados obtidos em relação com o esperado do projeto, e discussão de propostas futuras;

(E) - Escrita do Relatório Técnico - Escrita do texto;

(F) - Elaboração da apresentação - Apresentação que será feita ao final do projeto.

Tabela 5.1: Cronograma

Atividades	**Fev**	**Mar**	**Abr**	**Mai**	**Jun**	**Jul**	**Ago**	**Set**	**Out**	**Nov**	**Dez**
A	X	X	X	X	X	X	X				
B			X	X	X	X	X				
C					X	X	X	X			
D								X	X	X	X
E	X	X	X	X	X	X	X	X	X	X	X
F										X	X

REFERÊNCIAS

[Alpaydin 2010]ALPAYDIN, E. *Introduction to Machine Learning*. [S.l.: s.n.], 2010.

[Cruz, Sabourin e Cavalcanti 2018]CRUZ, R. M. O.; SABOURIN, R.; CAVALCANTI, G. D. C. Dynamic classier selection: Recent advances and perspectives. *Information Fusion*, v. 41, p. 195–216, 2018.

[Kuncheva e Whitaker 2003]KUNCHEVA, L.; WHITAKER, C. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, v. 51, p. 181207, 2003.

[Kurvers et al. 2019]KURVERS, R. H. J. M. et al. How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances*, American Association for the Advancement of Science, v. 5, n. 11, 2019. Disponível em: <https://advances.sciencemag.org/content/5/11/eaaw9011>.

[Polikar 2006]Polikar, R. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, v. 6, n. 3, p. 21–45, 2006.

[Tan, Steinback e Kumar 2006]TAN, P.-N.; STEINBACK, M.; KUMAR, V. *Introduction to Data Mining*. [S.l.: s.n.], 2006.

[Zhou 2012]ZHOU, Z.-H. *Ensemble methods: Foundations and algorithms*. [S.l.]: Taylor Francis Group, LLC, 2012.