02

Vizualizando as árvores de decisão

Transcrição

Agora tentaremos entender um pouco mais das árvores de decisão.

Para isso, escolheremos outro algoritmo em "*Choose* > 'trees' > *J48*" vamos rodar esse algoritmo e a saída dele será um pouco inferior ao algoritmo anterior, o *RandomForest*, mas ele permitirá enxergar um texto estranho no "*Classifier output*". Esse texto estranho se trata de uma série de subdivisões, uma árvore criada com uma série de ramificações para dividir os clientes entre os que iriam fazer o depósito e clientes que não fariam o depósito. Ele é feito com base nos atributos que foram coletados.

Clicando com o botão direito do mouse em "*Result list*", em cima de "*tress J48", veremos uma série de opções, e elas serão diferentes para cada algoritmo. No J48 poderemos enxergar essa árvore de uma forma diferente, com diversas ramificações. O primeiro parâmetro utilizado foi a duração. Para uma duração maior que 206 os clientes serão divididos com base em alguns atributos.Se ela for menor do que 206, a base serão outros atributos.

Para aumentar a tela da árvore, clicaremos com o botão direito do mouse sobre ela e escolheremos a opção "*Fit to screen*". A tela será um pouco ajustada, mas ainda não será o suficiente. Clicando com o botão direito do mouse novamente e escolhendo "*Auto Scale*" a árvore será totalmente aberta. Então, teremos que clicar com o botão esquerdo do mouse sobre a tela para segurar e arrastar. Assim conseguiremos enxergar as subdivisões da árvore. Como dissemos, a divisão será efetuada de acordo com atributos, até chegar aos atributos que revelarão se o cliente fez ou não um depósito.

Tentaremos entender a árvore com base num problema fictício, inspirado pelo caso que já temos, em que a **duração** foi o fator importante. Os clientes que assistiram a campanha por mais de 206 segundos fazem parte de um grupo, que ainda não bastará para haver a divisão entre os que fizeram depósito ou não.

Então, levaremos em consideração a informação de **contato**. Será importante dividir os clientes de acordo com contato móvel e fixo. Se ele tem o contato móvel, após assistir a campanha, sempre fará o depósito. Se tem contato fixo, ainda precisaremos da informação de **empréstimo** para decidir.

Se o cliente fez um empréstimo, ele não fará um depósito. Se ele não fez, nesse caso fará um depósito. Para os clientes que assistiram a campanha por um tempo menor, cairemos direto na informação empréstimo. Tendo feito o empréstimo, o cliente fará o depósito. Não tendo feito, ele também não fará o depósito.

Sendo assim, a informação de empréstimo funcionará de forma diferente para os clientes que assistiram a campanha por menos tempo, ou seja, para um determinado número de clientes poderão ser usados alguns atributos que não servirão para outros. A forma como as subdivisões e os atributos que utilizaremos são escolhidos varia de algoritmo para algoritmo e dependem da árvore de decisão, mas normalmente são escolhidos primeiramente os atributos que melhor dividem as classes de clientes finais.

Sabemos que a duração dividirá muito bem os clientes que fizeram ou não um depósito, pois temos uma quantidade de clientes que fizeram um depósito maior para os que assistiram a campanha por mais de 206 segundos.

Veremos, então, a nomenclatura utilizada para árvores de decisão. Para o primeiro nodo de todos que vai acima e que divide melhor a árvore, o nome será ***Root Node*** ou **Raiz**. Também teremos nodos internos ou intermediários. Eles terão teclas apontando para eles a partir do nodo Raíz e flechas apontando para fora deles também. Por fim, haverá os ***Leaf Nodes*** ou Folhas. Esses nodos-folha terão apenas flechas que apontam para eles.

Retornando ao Weka, modificaremos um pouco as opções do J48. Clicaremos no nome do algoritmo ao lado da opção "*Chose*" para ver as opções. Escolheremos o parâmetro "*minNumObj*", ou menor número de folhas. Se colocarmos o mouse em cima do parâmetro, veremos a descrição de que se trata desse número mínimo de folhas que se pode ter. Nesse caso, o número é 2, então teremos pelo menos duas folhas por nodo no nodo final Mas podemos aumentar, colocando 4 folhas ao final.

Como a tela é pequena, clicaremos na aba de tarefas e optaremos por visualizar Janelas em cascata para conseguir clicar em "Ok". Abriremos novamente a tela e rodaremos o algoritmo J48 com a opção com mais folhas por nodo. Isso fará com que haja menos subdivisões, teremos menos folhas por nodo, então haverá mais clientes que disseram sim ou não no nodo final.

Visualizaremos a árvore novamente e veremos que ele subdividiu um pouco melhor, pois teremos uma árvore mais enxuta. Com essa opção, quase conseguiremos visualizá-la totalmente apenas com o "*Fit to screen*", mas selecionaremos o "*Auto Scale*" e tentaremos entender um pouco melhor o que aconteceu.

Teremos que ter pelo menos 4 clientes por folha, será a tentativa. Como teremos mais clientes por folhas, teremos menos subdivisões desnecessárias, por isso a árvore fica mais simplificada para entender. Tivemos 14 clientes que disseram "Sim", eles foram classificados corretamente. 116 clientes disseram "Não", mas desses clientes, 25 foram classificados de forma incorreta. Então, a subdivisão não foi perfeita.

Em alguns casos, não teremos um número de clientes suficiente para um determinado parâmetro de qualquer forma para haver pelo menos 4, o que foi o caso das subdivisões por mês.

Mas esse processo já nos ajudou, já que em comparação com o anterior, terá aumentado um pouco o número de clientes classificados corretamente, pois não tivemos aquelas subdivisões desnecessárias. Esse será um parâmetro que teremos que escolher por tentativa e erro.