04

Filtrando e visualizando os dados

Transcrição

Até o momento, tentamos trabalhar com os algoritmos e suas opções para melhorar os resultados da nossa taxa de acerto na classificação.

Agora, começaremos a mexer nos dados. Clicaremos na aba "*Visualize*" e nela teremos uma matriz de gráficos. Vamos em "*PointSize*" puxaremos a barra para a direita aumentando o número do ponto, clicando em "*Update*" depois. Levando a barra em "*Jitter*" para a direita, será aumentado o número de pontos.

Feito isso, teremos os gráficos, e na última coluna deles teremos o depósito de acordo com os outros atributos. O depósito pelo depósito, pelo status do pagamento, pela idade, e assim por diante. Clicaremos na idade e para ver o gráfico aumentado, aumentaremos em "*Jitter*" para ver melhor os pontos. Parece haver uma relação entre a idade e o depósito, quando as pessoas são mais velhas, tendem mais a falar "Sim".

Iremos visualizar outro atributo como exemplo, o status de pagamento, e à primeira vista parece não haver uma relação tão forte entre ele e o depósito quanto havia na idade. Se aumentarmos o número de pontos parece até haver uma relação, mas ainda assim não tão forte, ou seja, o depósito não é tão dependente desse parâmetro de status de pagamento.

Existe outra forma de visualizar gráficos. Voltando na aba "*Classifier*", clicaremos com o botão direito sobre o último "*trees J48*", último teste que fizemos, poderemos clicar em "*Visualize classifying errors*". Veremos um gráfico do mesmo tipo que mostrará os erros de classificação em forma de quadrados. Onde houver quadrados, há erros, a classificação foi feita incorretamente. Nos "X", ela estará correta.

Clicaremos em um quadrado para visualizar a característica do cliente classificado errado. Pode ser que assim tenhamos uma noção do porquê isso aconteceu. Mas nesse caso conseguimos notar que não existe uma relação tão forte entre um determinado parâmetro, o do depósito, e o do status de pagamento.

Feito isso, podemos pensar em excluir esse atributo, a coluna do status de pagamento, para ver se com isso nossa classificação melhorará. Para fazer a exclusão, voltaremos à aba "*Preprocess*". Nela, escolheremos um filtro em "*Filter > 'filters' > 'unsupervised'*" e clicaremos num filtro para os atributos dos nossos dados. Clicaremos em "*Remove*" ao lado de "*Choose*" e poderemos escolher quais parâmetros remover. Deveremos digitar qual o índice de nosso atributo, o número que aparece em "*Attributes*" na aba "*Preprocess*". O número 16 será o referente ao status de pagamento.

Assim, se clicarmos em "*Apply*" esse status será removido. Se fizermos o treinamento da classificação novamente obteremos um resultado diferente, um pouco superior. Passaremos de 84 para 85% de taxa de acerto. Portanto, acertamos em excluir esse parâmetro. Isso dependerá de uma análise de parâmetro para parâmetro num determinado problema.

Se tivéssemos obtido um resultado ruim, poderíamos ter voltado. Clicando em *"Undo"*, esse procedimento será desfeito e o status de pagamento voltará. Clicaremos novamente em "*Choose*" para escolher um novo tipo de filtro. Se formos em tipos de "contato" veremos que há as alternativas do celular, do telefone fixo e um tipo desconhecido, "*unknown*". Poderemos julgar, por exemplo, que esse parâmetro desconhecido não será interessante por não ter tantas informações desse cliente, e poderá ser excluído.

Entretanto, não queremos excluir as informações dos clientes que tem celular e telefone, então deverá ser excluído um tipo determinado de instância, uma linha do nosso arquivo. Vamos novamente em "*Choose > Filter > 'filters' >*

'unsupervised'", mas dessa vez não iremos em filtros de atributos e sim nos de instância ("*instance*").

Procuraremos por "*RemoveWithValues*". Clicaremos nas opções dele e nesse caso o atributo "contato" terá como índice o número 9 e queremos remover o número 1 de "nominalindices", o índice do parâmetro desconhecido do atributo contato. Clicaremos em "*Apply*" e se voltarmos ao contato, veremos que ele continua lá, mas "*unknown*" estará com contagem 0, ou seja, teremos 0 contatos desconhecidos na nossa base de dados.

Agora poderemos fazer uma nova classificação. Lembrando que voltamos com o atributo que tinha sido removido anteriormente, o de status de pagamento. Porém, veremos que o atributo "*unknown*", que foi removido, tinha sua importância, uma vez que agora a nossa taxa de acerto terá diminuído para 84% novamente. Ela estará inclusive mais baixa do que a anterior à remoção do atributo de status do pagamento.

Nesse caso realmente teremos que retornar os procedimentos, desfazendo o filtro clicando em "*Undo*" para o tipo de contato "*unknown*" voltar para nossa base de dados, pois essa é uma informação que não deveria ter sido excluída.