Mineração de Dados – Classificação
Preparação da base de microdados do ENEM 2012

Jonathan Coutinho Luz de Queiroz[1]
Guilherme Lima Bernal[1]

[1]Instituto de Matemática – Universidade Federal da Bahia

jonathanqueiroz@dcc.ufba.br, ufba@lbguilherme.com

1. Introdução

Nessa atividade a base de dados do ENEM 2012 foi preparada para ser utilizada em um algorítmo de classificação. O objetivo é formar 4 clusters dentre os participantes da prova: Os que tiveram nota na redação "ruim", "regular", "boa" e "ótima". O intervalo de valores para a nota da redação que se qualifica em cada um desses grupos é baseado na divisão do dataset original em quartils, sendo "ruim" o primeiro quartil e "ótima" o último. Para passar pela qualificação todas as colunas precisam ser normalizadas para números entre 0 e 1.

2. Normalização da base de dados

As seguintes operações foram realizadas com o intuito de normalizar a base de dados:

- Foi criada a coluna ***UF_INSC_IGUAL_PROVA***, que guarda "1" se ***UF_INSC*** for iqual ao ***UF_MUNICIPIO_PROVA***, indicando que o candidato não mudou de estado para fazer a prova. "0" caso contrário.
- ***UF_INSC*** foi expandido para um total de 26 colunas, cada uma contendo "1" ou "0" para indicar se o candidato fez a prova nesse estado ou não. Nenhum candidato poder ter mais de um estado marcado como "1".
- De forma similar ***TP_SEXO*** foi expandido para duas colunas: ***GENERO_HOMEM*** e ***GENERO_MULHER***, com apenas uma delas sendo marcada como "1".
- As colunas ***NOTA_CN***, ***NOTA_CH***, ***NOTA_LC*** e ***NOTA_MT*** passaram a guardar a nota do candidato em cada uma das provas. Para que o valor seja consistente entre 0 e 1, essa nota foi dividida por 1000.
- Várias colunas que guardavam as opções marcadas pelo usuário, gabarito, identificação da prova e afins foram removidas por não acrescentar informação relevante para obtenção do resultado. Todas as informações daqui de alguma forma implicam na nota do candidato, que está mantida vide item anterior.
- Várias das colunas que guardavam informações refernetes a cidade, localização da prova, estado civil, escolha de idioma, ano de conclusão do ensino médio e afins foram removidas por acrescentarem pouca relevancia e por possuir pouca variabilidade, gerando uma amostra ruim para classificação.

- A idade foi normalizada entre “0” e “1” através de uma divisão por 100. Dessa forma a menor idade passa a ser 0.12 e a maior 0.70.
- ***TP_COR_RACA*** foi dividido em 5 colunas, uma com cada cor/raça. Cada uma possui “1” se ela é verdadeira para o candidato. Nenhum candidato tem duas dessas colunas como verdadeiras.
- ***ST_CONCLUSAO*** foi dividido em 4 colunas seguindo o mesmo princípio.
- Por fim, a nota da redação foi dividida em um dos 4 grupos definidos anteriormente.