CLASSIFICAÇÃO DE FATORES QUE INFLUENCIAM NO CRESCIMENTO, DESENVOLVIMENTO E PRODUTIVIDADE DA CULTURA DE SOJA

Silvia Ribeiro Mantuani (UEPG) E-mail: silviamantuani@gmail.com
Edson de Almeida (UEPG) E-mail: edsonuepg@gmail.com
José Carlos Ferreira da Rocha (UEPG) E-mail: jrocha@uepg.br

Resumo: A cultura de soja é desenvolvida em diversas partes do Brasil. Por isso diferentes fatores podem influenciar no crescimento, desenvolvimento e produtividade dessa cultura. A classificação em bases de dados volumosas é custosa, a mineração de dados com o uso de algoritmos classificadores destaca-se como uma técnica viável para acelerar este processo. Este trabalho descreve o funcionamento do classificador bayseano desenvolvido para trabalhar com a base de dados de soja soybean-Small e soybean-large, disponibilizada pela UCI. O classificador probabilístico adotado é baseado no teorema de Bayes – Naive Bayes. Os índices de acertos foram de 94% para a base soybean-large e 97% para soybean-Small.

Palavras-chave: Classificação de plantas. Probabilidade. Algoritmo Bayesiano. Naive Bayes.

CLASSIFICATION FACTORS AFFECTING THE GROWTH, DEVELOPMENT AND PRODUCTIVITY OF SOYBEAN CULTURE

Abstract: The soybean crop is developed in different parts of Brazil. So different factors can influence the growth, development and productivity of this crop. The classification in large databases is costly, data mining using classifiers algorithms stands out as a viable technique to speed up this process. This paper describes the operation of bayseano classifier developed to work with soybean database soybean-Small and soybean-large, provided by UCI. The adopted probabilistic classifier is based on Bayes' theorem - Naive Bayes. The hit rates were 94% for soybean-large basis and 97% for soybean-Small.

Keywords: Plant classification. Probability. Bayesian algorithm. Naive Bayes.

1. INTRODUÇÃO

A cultura de soja é desenvolvida em vários estados brasileiros, sendo responsável por 57,02% da área cultivada do país. A estimativa de crescimento de 3,6%, com 32,1 milhões de hectares cultivados em 2014/15, para 33,2 milhões na atual safra (CONAB,2016).

Muitos fatores podem influenciar no crescimento, desenvolvimento e produtividade da cultura de soja, tais como: sensibilidade ao fotoperíodo e a temperatura do ar, ao acamamento, retenção foliar, excesso ou falta de água, número de plantas por área, legumes por plantas, grãos por legumes, peso do grão. (MUNDSTOCK; THOMAS, 2005). Classificar que características presentes na planta ou no espaço externo exerce efeito em relação ao crescimento, desenvolvimento e produtividade dessa cultura em base de dados volumosas é caro. A técnica de mineração de dados com o uso de algoritmos de classificação apresenta-se como uma opção viável para potencializar o processo de aquisição de conhecimentos novos e uteis em bases de dados.

Os classificadores Bayesianos destacam-se como classificadores estatísticos que rotulam um objeto numa determinada classe baseando-se na probabilidade deste objeto pertencer a esta classe (ROUSSEL, 2004). O processo de classificação é rápido e de acurácia confiável quando aplicados a grandes volumes de dados. Segundo Amo (2014), os resultados produzidos por esta técnica possuem acurácia maior quando comparados aos produzidos por árvores de decisão e redes neurais.

O objetivo deste artigo é classificar as plantas de soja de acordo com os atributos presentes na base de dados da soja disponibilizada pela UCI por meio do classificador *Nave Bayes*.

2. FUNDAMENTAÇÃO TEÓRICA

Um dos tipos de informação obtida com o uso das técnicas de mineração de dados é a classificação, onde é apresentado um modelo capaz de descrever ao qual grupo um item pertence por meio dos itens já classificados e pela inferência de um conjunto de regras. A Mineração de Dados (MD) é uma tecnologia disponível que pode auxiliar na análise dos dados, pois dispõe de algoritmos sofisticados para processar dados de alto volume (FAYYAD, 1996). Dentre estes algoritmos há os classificadores estatísticos onde estão presentes os classificadores bayesianos, os quais classificam um objeto numa determinada classe baseando-se na probabilidade deste objeto pertencer a esta classe (WITTEN; FRANK, 2005).

Os métodos probabilísticos bayesianos assumem que a probabilidade de um evento A, que pode ser uma classe (por exemplo, uma determinada doença em uma planta), dado um evento B, que pode ser um valor para um atributo de entrada (por exemplo, coloração alterada de uma folha da planta), não depende apenas da relação entre A e B, mas também da probabilidade de observar A independentemente de observar B(MITCHELL; AGLE;WOOD,1997). Assim, para a determinação do evento A dado que B foi observado utiliza-se a probabilidade a priori da classe, P(A), a probabilidade de observar vários objetos que pertençam à classe, P(B|A), e a probabilidade de ocorrência desses objetos, P(B).

O espaço amostral ou espaço de resultados P(E) deve satisfazer os três axiomas de *Kolmogorof* (PESTANA;VELOSA, 2002) onde i) P(E) é maior ou igual a 0, ii) a soma dos eventos é igual a 1 e iii) se A e B são eventos disjuntos então a probabilidade de A com união em B é igual a probabilidade de A mais a probabilidade de B. A partir destes axiomas pode-se derivar a lei da probabilidade total, onde, se B1, B2, ..., Bn formam uma partição no espaço de eventos, então para qualquer evento A, tem-se que: P(A) = $(\sum_{i=1..} P(A|B_i) * \mathrm{P}(B_i))$

Um classificador Bayesiano simples, ou *Naive Bayes*, pressupõe que a atribuição de valores a um atributo que não é o atributo meta independe da atribuição dos outros atributos. Assim, o valor de um atributo não influencia o valor dos outros. Esta premissa do classificador simples tem como objetivo reduzir a complexidade dos cálculos envolvidos na classificação.

Todas as probabilidades necessárias para a obtenção do classificador *Naive Bayes* são computadas a partir dos dados de treinamento. Para calcular a probabilidade a priori de uma determinada classe, é necessário manter um contador para cada classe. Para calcular a probabilidade condicional de observar um valor de um atributo dado que o exemplo pertence a uma classe, é necessário distinguir entre atributos nominais e atributos contínuos.

Os atributos nominais, o conjunto de possíveis valores é um conjunto enumerável. Para calcular a probabilidade condicional, basta manter um contador para cada valor de atributo por classe. Para os atributos contínuos, quando o número de possíveis valores é infinito, há duas possibilidades. A primeira é assumir uma distribuição particular para os valores do atributo, e geralmente é assumida a distribuição normal. A segunda alternativa é discretizar o atributo em uma fase de pré-processamento. No entanto, a discretização apresenta resultados piores que a primeira possibilidade apresentada (DOMINGOS;PAZZANI, 1997).

A superfície de decisão de um classificador *Naive Bayes* em um problema de duas classes definido por atributos booleanos é um hiperplano, ou seja, a superfície de decisão é linear. Todas as probabilidades exigidas pela Equação 5.4 podem ser calculadas a partir do conjunto de treinamento em uma única passagem. O processo de construir o modelo é bastante

eficiente. Outro aspecto interessante do algoritmo é que ele é fácil de implementar de uma forma incremental. Além de apresentar bom desempenho em grande variedade de domínios, incluindo muitos em que há dependências entre os atributos.

3. METODOLOGIA

Foram escolhidas as bases de dados *soybean-large* e *soybean-small*, disponibilizadas pela *UCI Machine Learning Repository*, sendo um conhecido repositório de banco de dados, teorias de domínio e geradores de dados que são usados pela comunidade de aprendizado de máquina para a análise empírica dos algoritmos de aprendizado de máquina. A base *soybean-small* possui 48 registros, com 35 atributos. Já a *soybean-large* possui 35 atributos e 307 registros. Estes dados descrevem a planta, desde tratamento, amarelecimento de folhas, tamanho dos grãos, entre outros.

A função para o classificador bayesiano foi desenvolvida utilizando o software R. Este é um *software* voltado para computação estatística e gráficos, além de compila e roda em uma ampla variedade de plataformas, sendo um *software* livre.

O classificador bayesiano fará a classificação das plantas em uma das 4 classes (D1, D2, D3 e D4) através da probabilidade, descrita na sequência. Os mesmos procedimentos serão aplicados para a base *soybean-large.* Inicialmente foi realizado o treinamento com a base de dados da soja. Na sequência, foi desenvolvida uma função para calcular a probabilidade da planta ser classificada em uma das classes: D1, D2, D3 ou D4 para *soybean-small* e as classes apresentadas na Tabela 1 para *soybean-large.*

Tabela 1 - Classes da base de dados Soybean-large, Fonte: UCI Machine Learning Repository

Dados Soybean-Large	
1	2-4-d-injury
2	alternarialeaf-spot
3	anthracnose
4	bacterial-blight
5	bacterial-pustule
6	brown-spot
7	brown-stem-rot
8	charcoal-rot
9	cyst-nematode
10	diaporthe-pod-&-stem-blight
11	diaporthe-stem-canker
12	downy-mildew
13	frog-eye-leaf-spot
14	herbicide-injury
15	phyllosticta-leaf-spot
16	phytophthora-rot
17	powdery-mildew
18	purple-seed-stain
19	rhizoctonia-root-rot

Após realizado o treinamento da base de dados, ainda no *software R* foram realizados os testes, ou seja, foram executados procedimentos para avaliar a probabilidade de informações/parâmetros na base de dados para classificar as plantas corretamente.

4. RESULTADOS

A equação desenvolvida para probabilidade pode ser visualizada por meio do código gerado no software R conforme mostra a Função 1, em relação a utilização de classificadores bayesianos – Naive Bayes, determinado pela equação 1 abaixo:

Equação 1 - Probabilidade Utilizando Classificadores Bayseanos

$$\mathbf{P(Y|X)} = \mathbf{P(X|Y)} * \mathbf{P(Y)}$$

Função 1 - Função de Probabilidade

```
probabilidade<-function(nbayes,atributos,valores){
 numClasses=length(nbayes[[1]]$valores)
 evidencia<-c( rep(0,length(atributos)) )
 for(j in 1:length(atributos)){
  if(valores[j]=="não sei"){
   next
  }else{
   for (i in 2:length(nbayes) ){
    if(nbayes[[i]]$nome == atributos[j]){
     evidencia[j]<-which(nbayes[[i]]$valores==valores[j])
     break
    }
   }
  }
 }
 resultados<-c(rep(1,numClasses ))
 for (i in 2:(length(evidencia)+1)){
  for (j in 1:numClasses){
   if(evidencia[i-1] != 0){
    pBDadoA<-nbayes[[i]]$tabela[j,evidencia[i-1]]
    pA<-nbayes[[1]]$tabela[j]
    resultados[j]<-resultados[j]*(pBDadoA*pA)#ponto chave
   }
  }
 }
 ##normalizando os dados
 soma<-sum(resultados)
 resultados<-  resultados/soma
 #fim nomalizando os dados
 print("========== RESULTADO FINAL ==========")
 maiorResult<-which(resultados == max(resultados))
 for (i in 1:length(maiorResult)){
  print("_____")
```

```
    print(nbayes[[1]]$valores[maiorResult[i]])
    print(resultados[maiorResult[i]])
    print("_____")
  }
  for(i in 1:numClasses){
    print("######")
    print("====classe====")
    print(nbayes[[1]]$valores[i])
    print(resultados[i])
  }
  return(resultados)
}
```

O classificador obteve 94% para a base *soybean-large* e 97% para *soybean-Small*. A base de dados para *soybean-Small* não tem falta de dados porém possui menos registros para treinamento do classificador. Já a base *soybean-large* possui ausência de dados que são compensadas com um número maior de registros para treinamento.

Nota-se que a utilização dos Classificadores Bayesianos possibilitou a descoberta de diversos padrões, associando este resultado será possível perceber um crescimento, desenvolvimento e produtividade melhor em relação a culturas de soja.

5. CONSIDERAÇÕES FINAIS

Desenvolver modelos de classificação para bases de dados utilizando os conceitos de Classificadores Bayesianos é bastante trabalhoso, contanto após o modelo ser gerado, uma das vantagens é a fácil inserção de novas variáveis no modelo.

Um aspecto importante do algoritmo *Naive Bayes* é a sua eficácia elaborar estimativas de probabilidade e não somente classificações, dessa forma o classificador pode gerar uma estimativa de o novo objeto pertencer à mesma classe antes mesmo de rotulá-la.

Com a realização desta pesquisa é possível verificar que uso de técnicas probabilísticas é de aplicável ao processo de classificação, já que os resultados obtiveram um nível de confiabilidade satisfatório, pois mais de 94% das plantas foram classificadas corretamente mesmo com dados ausentes em alguns registros.

REFERÊNCIAS

AMO, S. Curso de Data Mining. Universidade Federal de Uberlândia, 2014.

MUNDSTOCK, C. M.; THOMAS, A. L. Soja fatores que afetam o crescimento e rendimento de grãos. Porto Alegre: Departamento de Plantas de Lavoura da universidade Federal do Rio Grande do Sul. Evangraf, 2005, 31 p.

CONAB – Companhia Nacional de Abastecimento. Disponível em: < http://www.conab.gov.br/OlalaCMS/uploads/arquivos/16_09_09_15_18_32_boletim_12_setembro.pdf> Acesso em 02 de Setembro de 2016.

DOMINGOS, P.; PAZZANI, M. Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. Machine Learning, 1997, 103-130 p.

FAYYAD, U.M.; PIATETSKI-SHAPIRO, G.; SMYTH, P; UTHURUSAMY, R. Advances in Knowledge Discovery and Data Mining. Menlo Park: AAAI Press, 1996, p. 10-36.

MITCHELL, R.; AGLE, B.; WOOD, D. Toward a Theory of Stakeholder Identification and Salience: Defining the Principle of Who and What Really Counts. *The Academy of Management Review*, v.22, n 4 , 1997, 853-886 p.

PESTANA, D.D.; VELOSA, S.F. Introdução à Probabilidade e à Estatística, Vol. I Fundação Calouste Gulbenkian, 2002.

RUSSEL, S., NORVIG, P. Inteligência Artificial. Tradução da 2ª edição. Rio de Janeiro: Editora Campus, 2004.

SOFTWARE R. Disponível em: < https://www.r-project.org/> Acesso em: 25 de Julho de 2016.

UCI Machine Learning Repository. Disponível em: <http://archive.ics.uci.edu/ml/> Acesso em: 25 de Julho de 2015.

WITTEN, I. H.; FRANK, E. Data Mining: Practical Machine Learning Tools and Techniques. 2ª edição. Nova Zelândia: Morgan Kaufmann Publishers, 2005, 525 p.