DESCOBERTA DE CONHECIMENTO ATRAVÉS DE MÉTODOS DE APRENDIZAGEM DE MÁQUINA SUPERVISIONADOS APLICADOS AO AMBIENTE VIRTUAL DE APRENDIZAGEM SIGAA/UFPI

TERESINA/PI MAIO/2017

ALINE MONTENEGRO LEAL SILVA - UNIVERSIDADE FEDERAL DO PIAUÍ - alineleal5@yahoo.com.br

FRANCISCO DAS CHAGAS IMPERES FILHO - UNIVERSIDADE FEDERAL DO PIAUÍ - fcoimperes@hotmail.com

JOSELITO MENDES DE SOUSA JUNIOR - UNIVERSIDADE FEDERAL DO PIAUÍ - joselitojunior10@hotmail.com

VINÍCIUS PONTE MACHADO - UNIVERSIDADE FEDERAL DO PIAUÍ - vinicius@ufpi.edu.br

Tipo: INVESTIGAÇÃO CIENTÍFICA (IC)

Natureza: PLANEJAMENTO DE PESQUISA

Categoria: MÉTODOS E TECNOLOGIAS

Setor Educacional: EDUCAÇÃO SUPERIOR

RESUMO

O presente trabalho exibe um processo de descoberta de conhecimento através de métodos de Aprendizagem de Máquina (AM) supervisionados aplicados à base de dados do SIGAA da Universidade Federal do Piauí (UFPI), cujos registros foram coletados a partir do segundo semestre de 2014. Neste processo de descoberta de conhecimento, procurou-se a identificação de perfis dos alunos do curso de Licenciatura em Computação na modalidade a distância, para descobrir a correlação entre o IRA (Índice de Rendimento Acadêmico) e os aspectos sociais de tais alunos. Foram utilizados três algoritmos de AM supervisionados com diferentes paradigmas: J48 (simbólico), Naive Bayes (estatístico) e IBK (baseado em exemplos). Os perfis descobertos podem auxiliar os gestores do sistema de educação superior a distância na tomada de decisões em relação a melhorias no processo de ensino-aprendizagem, como por exemplo identificar se os menores IRA são provenientes de pessoas de um determinado sexo, raça, estado civil ou ano de conclusão do ensino médio. A partir dessas informações, é possível definir-se estratégias diferenciadas em relação a esses alunos, como um acompanhamento presencial por parte dos tutores nos polos de apoio do sistema de educação a distância. Observou-se que o J48 obteve a melhor performance dentre os algoritmos utilizados, exibindo regras de produção bastante concisas.

Palavras-chave: Descoberta de Conhecimento, Aprendizagem de Máquina, Educação a Distância

AGRADECIMENTOS

Agradecimento especial ao Centro de Educação Aberta e a Distância (CEAD/UFPI), na pessoa do professor Dr. Gildásio Guedes Fernandes.

1.Introdução

As transformações mais marcantes ocorridas na Educação a Distância (EaD) refletem a mudança de um ambiente informacional para um ambiente de conhecimento, onde a figura do professor é considerada como mediador da aprendizagem e o facilitador do acesso ao conhecimento com base no diálogo e na interação. Dessa forma, essa modalidade de educação propicia a criação de um ambiente educacional em que o aluno precisa ser um indivíduo capaz de mostrar autonomia e comprometimento com a aquisição de conhecimento, estimulando assim, o processo ensino-aprendizagem, já que ele ocorre em lugares e/ou tempos diferentes (Farias, 2013).

Na EaD, os ambientes de gestão da aprendizagem (LMS, do inglês *Learning Management System)* ou ambientes virtuais de aprendizagem (AVAs), a exemplo do SIGAA (Sistema Integrado de Gestão de Atividades Acadêmicas), abrangem funcionalidades para armazenar, distribuir e gerenciar conteúdos de aprendizado, de forma interativa e gradativa. Os LMS são desenvolvidos para permitir abordagens didáticas que auxiliem a promoção do ensino e da aprendizagem em situações de mediação virtual ou semi presencial e acumulam muitos dados também (Carvalho et.al., 2012).

Atualmente é alarmante a distância crescente entre a geração de dados e a capacidade de analisá-los e compreendê-los. À medida que o volume de dados aumenta, a proporção dos dados que é analisada e entendida pelas pessoas diminui e escondido entre todo este volume de dados está a informação potencialmente útil (Batista, 2003). Existe, portanto, a necessidade de uma nova geração de técnicas e ferramentas que possibilite os analistas humanos compreenderem estas grandes bases de dados, as quais são objetos de estudo de uma área de pesquisa chamada de Descoberta do Conhecimento em Base de Dados (DCBD) (Facelli et.al., 2011).

O presente trabalho exibe um processo de descoberta de conhecimento através de métodos de Aprendizagem de Máquina (AM) supervisionados aplicados à base de dados do SIGAA/UFPI, cujos registros foram coletados a partir do segundo semestre de 2014. Neste processo de descoberta de conhecimento, procurou-se a identificação de perfis dos alunos do curso de Licenciatura em Computação na modalidade a distância, para descobrir a correlação entre o IRA (Índice de Rendimento Acadêmico) e os aspectos sociais de tais alunos. Foram utilizados três algoritmos de AM supervisionados com diferentes paradigmas: *J48* (simbólico), *Naive Bayes* (estatístico) e *IBK* (baseado em exemplos). Os perfis descobertos podem auxiliar os gestores do sistema de educação superior a distância na tomada de decisões em relação a melhorias no

processo de ensino-aprendizagem, como por exemplo identificar se os menores IRA são provenientes de um determinado sexo, de pessoas com determinada raça, estado civil ou ano de conclusão do ensino médio. A partir dessas informações, é possível definir-se estratégias diferenciadas em relação a esses alunos, como um acompanhamento presencial por parte dos tutores nos polos de apoio do sistema de educação a distância. Observou-se que o *J48* obteve a melhor performance dentre os algoritmos utilizados, exibindo regras de produção bastante concisas.

2. Descoberta de Conhecimento em Base de Dados

Existem diversas definições para DCBD, mas uma das mais utilizadas é a proposta por (Fayyad et.al., 1996), que define Descoberta do Conhecimento em Base de Dados como um processo não trivial para identificar padrões válidos, novos, potencialmente úteis e compreensíveis em dados existentes. Tais padrões devem ser compreensíveis aos analistas humanos, se não imediatamente, ao menos após algum pós-processamento (Araujo, 2014).

Esse é o processo de DCBD considerado neste trabalho, que encontra-se estruturado em 5 (cinco) fases: a) coleta: obtenção do conjunto de dados, b) pré-processamento: realiza a remoção de ruídos (dados inconsistentes) e balanceamento de classes, c) transformação: formatação dos dados para a aplicação dos algoritmos de Aprendizagem de Máquina, d) mineração de dados: aplicação dos algoritmos de AM, e) avaliação e interpretação dos resultados: corresponde à descoberta do conhecimento adquirido.

3. Aprendizagem de Máquina (AM)

A Aprendizagem de Máquina, do inglês *machine learning*, é utilizada na fase de mineração de dados do processo de descoberta do conhecimento em base de dados e surgiu da percepção de criar programas computacionais que aprendam um determinado comportamento ou padrão automaticamente, a partir de exemplos ou observações (Russel and Norvig, 2009).

Existem várias estratégias de aprendizagem que podem ser utilizadas por um sistema computacional como, por exemplo, o aprendizado por hábito, por instrução, por dedução, por analogia e por indução.O Aprendizado Indutivo pode ser dividido basicamente em Aprendizado Supervisionado e Não Supervisionado (Monard and Baranauskas, 2003). No primeiro, busca-se a criação de um modelo preciso em relação à predição de valores para novos dados enquanto que no segundo o objetivo é encontrar características que podem resumir os dados. A diferença básica entre esses

dois modos de aprendizagem, é a presença ou não do atributo que rotula os exemplos do conjunto de dados. No Aprendizado Supervisionado esse rótulo é conhecido, ao passo que no aprendizado não supervisionado os exemplos não estão previamente rotulados. Adicionalmente, existe um terceiro modo de aprendizagem, conhecido como aprendizado semi-supervisionado, no qual somente poucos exemplos encontram-se rotulados.

3.1 Algoritmos de Aprendizagem de Máquina

Algoritmos de AM podem ser vistos como mecanismos que extraem um padrão de comportamento a partir de experimentações (Machado, 2011). A seguir são descritos alguns dos algoritmos de aprendizado supervisionados propostos na literatura, de acordo com os principais paradigmas desse aprendizado.

- *J48:* algoritmo de paradigma simbólico, surgiu da necessidade de recodificar o algoritmo C4.5 para a linguagem *Java*. Ele tem a finalidade de gerar uma árvore de decisão baseada em um conjunto de dados de treinamento, sendo este modelo usado para classificar as instâncias no conjunto de teste.
- *Naive Bayes*: algoritmo de paradigma estatístico, baseia-se na aplicação da teoria *Bayesiana* para o cálculo das probabilidades necessárias para a classificação. Tem como função predizer a probabilidade de um exemplo pertencer a uma determinada classe. Esse classificador considera a hipótese de que todos os atributos são independentes, dado a variável classe.
- *IBK*: algoritmo de paradigma baseado em exemplos, fundamenta-se no algoritmo *K-Nearest Neighbors* (K-NN), o qual parte da ideia de que em um espaço n-dimensional, um ponto P e os seus K vizinhos mais próximos pertencem à mesma classe. A proximidade entre dois pontos nesse espaço é calculada por meio da distância euclidiana entre eles.

4**. Trabalhos Relacionados**

Alguns autores têm relatado a maneira que diferentes conteúdos de Inteligência Artificial, mais especificamente a Aprendizagem de Máquina, têm sido abordados atualmente. Em Araujo (2014), foi realizado o aprendizado automático do processo de regulação médica/odontológica para reduzir os custos das operadoras de plano de saúde e aumentar sua eficiência na tomada de decisões, favorecendo assim a redução de custos também para os beneficiários, além de gerar mais agilidade durante a realização de exames / procedimentos / tratamentos. Para isso, retratou métodos de AM supervisionados, incluindo os paradigmas simbólico, estatístico, baseado em exemplos

e conexionista.

Em Viterbo et.al. (2016), foram avaliadas quatro ferramentas de apoio ao Aprendizado de Máquina em cursos de graduação, dentre elas a *Weka* que é utilizada nesse trabalho, devido ao crescente uso de tais ferramentas no meio acadêmico e o resultado dessa avaliação pode auxiliar como um guia na escolha de ferramentas para ensino de AM. Constatou-se que a *Weka* é uma das ferramentas mais utilizadas em cursos de graduação e pós-graduação no Brasil, pela sua facilidade de uso. A avaliação da usabilidade dessas ferramentas pode aumentar o potencial de cada uma delas, utilizando o conhecimento do usuário e ajudando-o a obter resultados cada vez mais expressivos das bases de dados.

5. Metodologia

Neste trabalho, a metodologia empregada considerou o esquema de Descoberta do Conhecimento em Base de Dados proposto por Fayyad (Holzinger et.al., 2014). Após as etapas de coleta, pré-processamento e transformação os dados obtidos foram utilizados, juntamente com técnicas de aprendizagem de máquina, para gerar o conhecimento desejado.

Inicialmente, durante a fase de coleta, constatou-se que os dados encontravam-se disponibilizados em uma tabela com 19 atributos e 322 instâncias, provenientes da base real do SIGAA fornecida pela UFPI. Ao longo da fase de pré-processamento, observou-se que alguns atributos não tinham relevância para a pesquisa em questão, pois o objetivo era obter uma correlação entre o IRA dos alunos do curso de Licenciatura em Computação na modalidade a distância e os aspectos sociais de tais alunos. Portanto, utilizou-se o ganho de informação, que é uma técnica de seleção de atributos relevantes da ferramenta *Weka*, que calculou a razão de ganho para cada um dos atributos e aqueles com razão de ganho nulo foram eliminados. Com isso, 13 atributos foram desconsiderados de um total de 19, restando apenas 6 (Polo, Sexo, Raça, Estado_civil, Ano_conclusao e IRA).

Notou-se também, ainda nessa fase, que um dos atributos (Ano_conclusão), no caso se refere à conclusão do Ensino Médio, possuía 19 instâncias preenchidas de forma default, que após o tratamento de valores desconhecidos, foram ignoradas. A base resultante passou a ter 303 instâncias. Considerando ainda o pré-processamento, percebeu-se que algumas classes encontravam-se desbalanceadas, sendo assim utilizado o filtro *Resample* da ferramenta *Weka*, que balanceia um conjunto de dados por meio de uma amostragem com reposição, mantendo o número de exemplos no conjunto

de treinamento constante.

Já na fase de transformação, houve a normalização do atributo Estado_civil em Solteiro(a) e Não_Solteiro(a) visando realizar um equilíbrio dos dados, pois inicialmente tínhamos Solteiro(a), Casado(a), Divorciado(a), Separado Consensualmente, Separado Judicialmente e Outro, sendo que os quatro últimos campos citados foram encontrados em poucas instâncias. Também nessa fase, fez-se a discretização dos valores do campo IRA, dando lugar ao atributo nominal Faixa_IRA, já que os algoritmos de classificação trabalham com atributos classe nominais.

O Faixa_IRA foi escolhido para ser o atributo classe, visto que o trabalho visa correlacioná-lo com os aspectos sociais dos alunos. A etapa seguinte consistiu na execução da fase de mineração de dados, onde algoritmos supervisionados foram utilizados para a classificação de padrões. Visando a obtenção da extração do conhecimento para a identificação dos perfis, foram escolhidos três algoritmos de paradigmas diferentes: *J48, Naive Bayes e IBK.* Para avaliar a qualidade da classificação foi utilizada a acurária, que mede a quantidade de instâncias corretamente classificadas (taxa de acerto).

6. Avaliação dos Resultados

Nesta seção, são apresentados os resultados de três classificadores com diferentes paradigmas *J48, Naive Bayes e IBK*, aplicados à base de dados do SIGAA/UFPI. Para chegarmos ao resultado obtido, foi feita uma sequência de cinco testes (T1, T2, T3, T4 e T5) na tentativa de encontrar resultados significativos para a base em questão.

6.1 Classificador de Paradigma Simbólico

Inicialmente, os testes foram realizados com o algoritmo *J48*. No primeiro momento, a discretização do atributo classe foi realizada considerando as faixas de valores: Baixo, Regular, Bom e Ótimo (Teste T1).

Em T1, a acurácia obtida foi de 53.4653%. Tentando reduzir a perda de informação gerada pela discretização, já que cada faixa de valores em T1 estava grande e isso dificultava o conhecimento sobre a nota exata do aluno, realizou-se um novo teste considerando as seguintes faixas de valores: Insuficiente, Baixo, Regular, Bom e Ótimo (Teste T2). Nesse segundo teste, as 303 instâncias foram divididas em 5 grupos e a acurácia obtida foi de 32.0132%. Uma nova sequência de testes foi realizada (Teste T3), dessa vez dividindo o total de instâncias em 10 grupos, na tentativa de reduzir ainda

mais a perda de informação. A acurácia alcançada através do Teste T3 foi de 21.4521%. Como pode-se observar, a medida em que aumentamos o conjunto de dados na discretização do atributo classe, a acurácia tende a diminuir. Então como a classificação foi piorando, decidiu-se voltar a faixas de valores mais restritas.

Novos testes foram feitos (Teste T4), dessa vez considerando menos conjuntos de dados para o Faixa_IRA, com as seguintes faixas de valores: Baixo, Regular e Bom. Dessa vez, a acurácia obtida por meio de T4 foi de 54.4554%. Então, finalmente, resolveu-se realizar a discretização com um conjunto menor de dados, dentre todos já testados. As faixas de valores consideradas foram: Exame Final ou Aprovativo e Reprovativo.

Com esse novo teste (T5), conseguimos a maior acurácia alcançada até o momento para o atributo classe Faixa_IRA. A acurácia foi de 67.6568%. A partir desse resultado, resolveu-se considerar um subconjunto aleatório para teste, dessa vez com 200 instâncias, na tentativa de encontrar um conjunto mais representativo. O resultado ultrapassou a acurácia do teste anterior, obtendo o valor de 73% para acurácia. Portanto, optou-se por não realizar mais nenhuma discretização em nenhum atributo para que o resultado pudesse refletir exatamente a realidade do curso em questão, no caso o de Licenciatura em Computação da UFPI na modalidade a distância, que possui níveis de aprovação elevados em determinados polos de apoio presencial do sistema de Educação a Distância, em contraposição a níveis baixos em outros polos.

Permitiu-se apenas aplicar o filtro *Resample*, que faz uma subamostragem estratificada do conjunto de dados, o que gerou uma acurácia de 94,5%. Verificou-se que os resultados após o balanceamento tiveram uma melhora significativa.

6.2 Classificador de Paradigma Estatístico

O *Naive Bayes* foi o algoritmo de paradigma estatístico utilizado para classificar o conjunto de dados. A acurácia obtida por esse algoritmo através da discretização do atributo Faixa_IRA (teste T5) e considerando o mesmo subconjunto aleatório obtido pelo *J48*, que também indicou ser o melhor dentre todos os testes realizados pelo *Naive Bayes*, foi de 70,5\%. Posteriormente, observou-se que a aplicação do filtro *Resample* à base de dados resultou em uma acurácia de 79%. Constatou-se que os resultados alcançados pelo *Naive Bayes* foram um pouco inferiores ao do melhor teste do J48, tanto antes como após o balanceamento.

6.3 Classificador de Paradigma Baseado em Exemplos

Utilizou-se o *IBK* como algoritmo de paradigma baseado em exemplos para a classificação de padrões. A acurácia atingida por esse classificador através do teste T5 e em seguida considerando o mesmo subconjunto aleatório obtido pelo *J48*, que também foi o melhor dentre todos os testes realizados pelo *IBK*, foi de 68%, para K=5. Em um segundo momento, notou-se que o filtro *Resample* quando aplicado à base de dados, gerou uma acurácia de 83,5%. Observou-se que o *IBK* obteve resultados inferiores ao do melhor teste do *J48*, tanto antes quanto depois do balanceamento, mas superiores ao resultado do *Naive Bayes* após o balanceamento.

7. Interpretação dos Resultados

Considerando que o J48 apresentou os melhores resultados dentre os classificadores considerados, segue o resultado das regras de produção geradas por ele.

Pôde-se observar que no polo de Bom Jesus, os alunos que obtiveram um melhor desempenho acadêmico foram aqueles que concluíram o ensino médio até 2009. Já no polo de Inhuma, o pior desempenho foi dos alunos cujo ano de conclusão do ensino médio foi a partir de 2013. Em Marcos Parente, constatou-se que as mulheres com ano de conclusão do ensino médio até 2001 tiveram um desempenho abaixo dos homens, considerando essa mesma época. Em Pio IX, os alunos com ano de conclusão do ensino médio até 2011 e raça branca, se sobressaíram nos estudos em relação aos negros e pardos nesse mesmo período. Já em Piripiri, por exemplo, observou-se que as pessoas não-solteiras possuem maiores dificuldades para estudar. No polo de São João do Piauí, os alunos com pior desempenho foram aqueles que concluíram o ensino médio até 2004. Finalmente, em Teresina, as mulheres que concluíram o ensino médio até 1997 tiveram um desempenho acadêmico superior ao dos homens nessa mesma época.

8. Conclusões

Este trabalho apresentou um processo de descoberta de conhecimento através de métodos de Aprendizagem de Máquina (AM) supervisionados aplicados à base de dados do SIGAA/UFPI, cujos registros foram coletados a partir do segundo semestre de 2014. Neste processo de descoberta de conhecimento,procurou-se a identificação de perfis dos alunos do curso de Licenciatura em Computação na modalidade a distância, para descobrir a correlação entre o IRA (Índice de Rendimento Acadêmico) e os aspectos sociais de tais alunos. Foram utilizados três algoritmos de AM supervisionados com diferentes paradigmas: *J48* (simbólico), *Naive Bayes* (estatístico) e *IBK* (baseado em exemplos).

Observou-se que o algoritmo *J48* apresentou uma performance melhor em relação aos outros classificadores, tanto antes quanto depois do balanceamento, mostrando regras de produção bastante concisas. Portanto, optou-se por utilizar as regras obtidas através do *J48* para o reconhecimento de padrões sobre os perfis. Constatou-se ainda que a acurácia obtida por meio do *J48* antes do balanceamento não foi satisfatória, no caso foi de 73%, mas optou-se por gerar um resultado que pudesse refletir exatamente a realidade do curso em questão, que possui níveis de aprovação elevados em determinados polos de apoio presencial do sistema de Educação a Distância, em contraposição a níveis baixos em outros polos. Após o balanceamento, a acurácia passou a ser de 94,5%.

Percebeu-se, por exemplo, que no polo de Bom Jesus os alunos que obtiveram um melhor desempenho acadêmico foram aqueles que concluíram o ensino médio até 2009. Já em Inhuma, o pior desempenho foi dos alunos cujo ano de conclusão do ensino médio foi a partir de 2013. Em Marcos Parente, notou-se que as mulheres com ano de conclusão do ensino médio até 2001 tiveram um desempenho abaixo dos homens, considerando essa mesma época.

Conforme visto, os perfis descobertos podem auxiliar os gestores do sistema de educação superior a distância do curso de Licenciatura em Computação na tomada de decisões em relação a melhorias no processo de ensino-aprendizagem, como identificar se os menores IRA são provenientes de um determinado sexo, raça, estado civil ou ano de conclusão do ensino médio. A partir dessas informações, é possível definir-se estratégias diferenciadas em relação a esses alunos, como um acompanhamento presencial por parte dos tutores nos polos de apoio do sistema de educação a distância.

Por fim, acredita-se que uma correlação entre o desempenho do trabalho do professor tutor em cada polo de apoio presencial do sistema de Educação a Distância e a performance dos alunos daquele polo constitui potenciais pontos para uma avaliação futura.

9. Referências

Araujo, F.H.D. (2014). Descoberta de Conhecimento em Base de Dados para o Aprendizado da Regulação Médica/Odontológica em Operadora de Plano de Saúde. Dissertação de Mestrado. Maio.

Batista, G.E.A.P.A. (2003). Pré-processamento de Dados em Aprendizado de Máquina Supervisionado. Tese de doutorado, Instituto de Ciências Matemáticas e de

Computação - ICMC/USP.

Carvalho, R.S, Filho, I.J.M., Vidal, T.C., Melo, R.M., Gomes, A.S. (2012). Integração entre o sistema de gestão acadêmica e o sistema de gestão da aprendizagem: identificando necessidades e prototipando requisitos favoraveis a pratica docente. Revista Brasileira de Computação Aplicada, v.4, páginas 81-91,2012.

Facelli, K., Lorena, A.C., Gama, J., Carvalho, A.C.P.L.F. (2011). Inteligência Artificial: Uma abordagem de aprendizado de maquina. LTC.

Farias, S.C. (2013). Os benefícios das tecnologias de informação e comunicação (TIC) no processo de educação a distância(EaD). Rev. digit. bibliotecon. cienc. inf., Campinas, SP, v.11, n.3, páginas 15-29, set/dez, 2013. ISSN 1678-765X.

Fayyad, U., Piatetsky-Shapiro, G., Smith, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. Second International Conference on Knowledge Discovery and Data Mining (KDD-96), pages 82-88.

Holzinger, A., Dehmer, M., Jurisica, I. (2014). Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directionsk. BMC Bioinformatics,15 Suppl 6:I1. doi: 10.1186/1471-2105-15-S6-I1.

Machado, V.P. (2011). Inteligencia Artificial. Editora EDUFPI.

Monard, M.C. and Baranauskas, J.A. (2003). Conceitos sobre Aprendizado de Maquina. Volume 1, Rezende,1.edition.

Russel, S.J. and Norvig, P. (2009). Artificial Intelligence: A Modern Approach. 3rd. Prentice Hall.

Viterbo, J., Boscarioli, C., Bernardini, F., Teixeira, M.F. (2016). Avaliação de Ferramentas de Apoio ao Ensino de Técnicas de Mineração de Dados em Cursos de Graduação. Anais do CSBC 2016. Páginas 2006-2015.