USO DA MINERAÇÃO DE DADOS EM GENES LIGADOS A DIABETES UTILIZANDO A FERRAMENTA WEKA EM BASE DE DADOS BIOLOGICAS.

Lidiane Ferreira Silva Silveira[1*],André Bevilaqua[2]

[1]Bacharelando em Sistemas de Informação, peloInstituto Luterano de Ensino Superior de Itumbiara-ILES/ULBRA, Itumbiara-GO, *deleonlidiane@hotmail.com,[2]Mestre em Ciências da Computação pela Universidade Federal de São Carlos, São Paulo-SP,andrebevilaquax@gmail.com.

RESUMO – A Inteligência Artificial está presente em diversos campos de estudo e suas técnicas podem ser usadas em diversas áreas para resolução de problemas. A Bioinformática é uma das áreas que utiliza de técnicas da Inteligência Artificial para a solução de problemas biológicos. O uso de algoritmos genéticos para solucionar problemas vem sendo bastante utilizado, juntamente com o uso da mineração de dados já que a mesma tem por objetivo transformar grandes quantidades de dados em conhecimento que será usada para resolver inúmeros problemas. Sendo assim este artigo tem como objetivo usar regras de mineração de dados na base de dados Dataset Diabetes com o objetivo de classificar qual algoritmo tem um melhor desempenho na classificação dos dados da base, desta forma, aplicou-se os métodos de aprendizado de máquina (algoritmos genéticos) J48, Naive Bayes e o IBK.

PALAVRAS-CHAVE: Mineração de Dados, Bioinformática, Aprendizado de Máquina.

INTRODUÇÃO

Segundo Cabena (2007) a mineração de dados ou *data mining* é definida, como um processo de descoberta de padrões em quantidades substanciais de dados, de forma automática ou, na maioria das vezes, semiautomática, para a extração de informação previamente desconhecida, valida e que gera ações úteis, em que os padrões descobertos são significativamente vantajosos para a tomada de decisões estratégicas.

De acordo com Han&Kamber (2001) a informação e o conhecimento obtidos no processo de mineração de dados podem ser utilizados para diversas aplicações, que vão do gerenciamento de negócios, controle de produção e análise de mercado ao projeto de engenharia e exploração cientifica.

A mineração de dados então pode ser usada em diversas áreas para resolver problemas, uma das áreas a ser usada é a da Bioinformática.

Dados biológicos estão sendo disponibilizados a uma taxa muito elevada, fazendo com que os bancos de dados atuais cresçam exponencialmente (Baldi e Brunak, 2001). Manipular e analisar os dados acumulados nessas bases de dados tornou-se um dos maiores desafios da Bioinformática.

A Bioinformática ou Biologia Computacional diz respeito à utilização de técnicas e ferramentas de computação para a resolução de problemas da Biologia (Baldi e Brunak, 2001). Dentre as diversas áreas da Biologia, aquele em que a aplicação de técnicas computacionais tem se mostrado mais promissora é a Biologia Molecular (Setúbal e Meidanis 1997). O termo expressão gênica, refere-se ao processo em que a informação codificada por um determinado gene é decodificada em uma proteína, manifestando assim, características particulares àqueles gene. Assim, a seleção de genes relevantes a uma determinada doença torna-se uma tarefa importantíssima, podendo num futuro próximo, ser aplicada no diagnostico médico.

Na busca destes pequenos conjuntos de genes preditores, técnicas advindas da Inteligência Artificial (IA), tais como, os Algoritmos Genéticos (AGs) e as Redes

Disponível em: www.anaissimpesquisa.wix.com/ilesulbraitumbiara
ISSN: 2319-0930

Neurais Artificiais (RNAs), são cada vez mais empregadas, devido a sua capacidade de aprender automaticamente a partir de grandes volumes de dados e produzir hipóteses úteis (Baldi&Brunak 2001).

Neste trabalho aplicaram-se os métodos J48, Naive Bayes e o IBK pertencentes ao Weka à base de dados Dataset Diabetes, onde cada método será analisado e com isso pontuado quais foram os melhores e os piores resultados. O Dataset possui características multivariadas, possui 100.000 instancias e 50 atributos.

METODOLOGIA

O Dataset utilizado neste estudo chama-se Dataset Diabetes e foi obtido do repositório de dados UCI - Machine Learning Repository(http://archive.ics.uci.edu/ml/datasets/Diabetes+130US+hospitals+for+years+1999-2008#).

O conjunto de dados representa 10 anos (1999-2008) de cuidados clínicos em 130 hospitais dos Estados Unidos e redes integradas. Ele inclui mais de 50 recursos que representam os resultados do paciente e do hospital. As informações foram extraídas do banco de dados paraencontros que preencham os seguintes critérios.

(1) É um encontro de internação (a internação).

(2) É um encontro diabético, isto é, durante o qual um qualquer tipo de diabetes foi introduzido no sistema como um diagnóstico.

(3) O tempo de permanência foi de pelo menos um dia e, no máximo, 14 dias.

(4) Os exames laboratoriais foram realizados durante o encontro.

(5) Os medicamentos foram administrados durante o encontro.

Os dados contêm atributos como número de paciente, raça, sexo, idade, tipo de admissão, tempo no hospital, especialidade médica de médico admitir, número de teste de laboratório realizados, o resultado do teste HbA1c, diagnóstico, número de medicamentos, medicamentos para a diabetes, número de ambulatório , em regime de internamento, e visitas de emergência no ano anterior à internação, etc

Os dados foram originalmente disponibilizados num ficheiro com extensão .diabetic_data, ao qual correspondia o formato ***commaseparatedvalues***(.csv). Adicionalmente foi disponibilizado um ficheiro (IDs_mapping) com a descrição dos atributos. Por forma a preparar os dados para utilização (mineração) na ferramenta WEKA, foi necessário colidir toda a informação num só ficheiro com extensão .arff.

Os algoritmos utilizados neste trabalho para classificação na base de dados Dataset Diabetes foram o J48, Naive Bayes e o IBK. A análise dos dados foi realizada utilizando os algoritmos selecionados em dois testes gerais: use training set e cross-validation (em 10 folds).

Os dados a analisar incluíram:

- Percentagem de acerto ou número de instâncias corretamente classificadas;
- Percentagem de erro ou número de instâncias incorretamente classificadas;
- Tempo para construção do modelo ou tempo de aprendizagem;
- Erro médio;
- Taxa de verdadeiros positivos
- Taxa de falsos positivos;
- Sensibilidade;
- Especificidade.
- Área de receiveroperatingcharacteristics (ROC)

RESULTADOS E DISCUSSÃO

Utilizando o algoritmo J48 com use training set foi possível observar que gastou-se 129,38 segundos para criação do modelo. O numero de instâncias corretamente classificadas foi de 62,62%, enquanto que as instancias incorretamente classificadas corresponde apenas a 37,38%. Já utilizando o algoritmo J48 com cross-validation (10 folds) os resultados obtidos foram tempo

Disponível em: www.anaissimpesquisa.wix.com/ilesulbraitumbiara
ISSN: 2319-0930

gasto para criar o modelo foi de 135,94 segundos. O numero de instâncias corretamente classificadas foi de 57,38%, enquanto que as instancias incorretamente classificadas corresponde apenas a 42,61%.

Utilizando o algoritmo Naive Bayes com use training set foi possível observar que gastou-se 1.26 segundos para criação do modelo. O numero de instâncias corretamente classificadas foi de 57,28%, enquanto que as instancias incorretamente classificadas corresponde apenas a 42,71%. Já utilizando o algoritmo Naive Bayes com cross-validation (10 folds) os resultados obtidos foram tempo gasto para criar o modelo foi de 135,94 segundos. O número de instâncias corretamente classificadas foi de 57,38%, enquanto que as instancias incorretamente classificadas corresponde apenas a 42,61%.

Utilizando o algoritmo IBK com use training set foi possível observar que gastou-se 0.2 segundos para criação do modelo. O número de instâncias corretamente classificadas foi de 100%, enquanto que as instâncias incorretamente classificadas corresponde apenas a 0%. Já utilizando o algoritmo IBK com cross-validation (10 folds) os resultados obtidos foram tempo gasto para criar o modelo foi de 0.26 segundos. O número de instâncias corretamente classificadas foi de 46,17%, enquanto que as instancias incorretamente classificadas corresponde apenas a 53,82%.

CONCLUSÕES

Através da mineração de dados usando a ferramenta weka na base de dados biológicas dataset diabetes foi possível concluir que o melhor algoritmo genético para classificar a base de dados no modo use training set foi o IBK com 100% das instâncias classificadas corretamente. Enquanto que o J48 obteve 62,62% das instancias corretamente classificadas e Naive Bayes apresentou o pior resultado com 57,28%. Já no modo cross-validation (10 folds) o melhor algoritmo genético para classificar a base de dados foi o J48 com 57,38%, seguido pelo Naive Bayes com 57,28% enquanto que neste modo o IBK obteve apenas 46,17%.

Os resultados obtidos neste trabalho servirão para uma compressão melhor da base de dados por parte dos especialistas podendo trazer um grande avanço no tratamento da diabetes.

REFERÊNCIAS BIBLIOGRÁFICAS

AMARAL, L. R.(2007). **Mineração de regras para classificação de oncogenes medidos por microarray utilizando algoritmos genéticos.**Master`sthesis, Universidade Federal de Uberlândia.

HAN, J; KAMBER, M. **Data Mining: Conceitos e Tecnicas.**Elsevier, 2006.

RODRIGUES, Fabrício Alves. **Mineração de Genes Ligado ao Câncer Utilizando a Ferramenta Weka em Base de Dados Biológicas.** Disponível em: www.informaticabiomedica.com.br/wibn. Acessado dia 02 de Setembro de 2014.

Tabela 1: Precisão detalhada da mineração do dataset diabetes na ferramenta weka, usando o algoritmo genético J48, no modo use training set.

PRECISÃO DETALHADA POR CLASSE – J48 – USE TRAINING SET							
	Tp Rate	Fp Rate	Precision	Recall	F-Measure	ROC Área	Class
	0.837	0.524	0.652	0.837	0.733	0.715	NO
	0.471	0.198	0.56	0.471	0.512	0.697	>30
	0.092	0.004	0.761	0.092	0.164	0.692	<30
Weighted Avg.	0.626	0.352	0.632	0.626	0.592	0.706	

Disponível em: www.anaissimpesquisa.wix.com/ilesulbraitumbiara
ISSN: 2319-0930

Tabela 2: Precisão detalhada da mineração do dataset diabetes na ferramenta weka, usando o algoritmo genético J48, no modo cross-validation (10 folds).

PRECISÃO DETALHADA POR CLASSE – J48 – CROSS-VALIDATION (10 FOLDS)							
	Tp Rate	Fp Rate	Precision	Recall	F-Measure	ROC Área	Class
	0.79	0.562	0.622	0.79	0.696	0.648	NO
	0.412	0.239	0.48	0.412	0.443	0.616	>30
	0.037	0.013	0.268	0.037	0.065	0.597	<30
Weighted Avg.	0.574	0.388	0.533	0.574	0.537	0.631	

Tabela 3: Precisão detalhada da mineração do dataset diabetes na ferramenta weka, usando o algoritmo genético naivebayes, no modo use training set.

PRECISÃO DETALHADA POR CLASSE – NAIVE BAYES – USE TRAINING SET							
	Tp Rate	Fp Rate	Precision	Recall	F-Measure	ROC Área	Class
	0.864	0.668	0.602	0.864	0.71	0.682	NO
	0.266	0.132	0.519	0.266	0.352	0.647	>30
	0.128	0.037	0.301	0.128	0.179	0.666	<30
Weighted Avg.	0.573	0.411	0.539	0.573	0.525	0.668	

Tabela 4: Precisão detalhada da mineração do dataset diabetes na ferramenta weka, usando o algoritmo genético naivebayes, no modo cross-validation (10 folds).

PRECISÃO DETALHADA POR CLASSE – NAIVE BAYES – CROSS-VALIDATION (10 FOLDS)							
	Tp Rate	Fp Rate	Precision	Recall	F-Measure	ROC Área	Class
	0.857	0.67	0.599	0.857	0.705	0.67	NO
	0.261	0.139	0.503	0.261	0.344	0.63	>30
	0.12	0.038	0.282	0.12	0.168	0.643	<30
Weighted Avg.	0.567	0.414	0.53	0.567	0.519	0.654	

Tabela 5: Precisão detalhada da mineração do dataset diabetes na ferramenta weka, usando o algoritmo genético IBK, no modo use training set.

PRECISÃO DETALHADA POR CLASSE – IBK – USE TRAINING SET							
	Tp Rate	Fp Rate	Precision	Recall	F-Measure	ROC Área	Class
	1	0	1	1	1	1	NO
	1	0	1	1	1	1	>30
	1	0	1	1	1	1	<30
Weighted Avg.	1	0	1	1	1	1	

Tabela 6: Precisão detalhada da mineração do dataset diabetes na ferramenta weka, usando o algoritmo genético IBK, no modo cross-validation (10 folds).

PRECISÃO DETALHADA POR CLASSE – IBK – CROSS-VALIDATION (10 FOLDS)							
	Tp Rate	Fp Rate	Precision	Recall	F-Measure	ROC Área	Class
	0.563	0.468	0.584	0.563	0.573	0.548	NO
	0.404	0.344	0.387	0.404	0.395	0.531	>30
	0.155	0.111	0.148	0.155	0.151	0.523	<30
Weighted Avg.	0.462	0.385	0.467	0.462	0.464	0.539	

Disponível em: www.anaissimpesquisa.wix.com/ilesulbraitumbiara
ISSN: 2319-0930