Estudo Comparativo entre os algoritmos de Mineração de Dados Random Forest e J48 na tomada de Decisão

Cassio Dal Castel Lorenzett[1], Alex Vinícios Telöcken[1]

[1]Curso de Ciência da Computação– Universidade De Cruz Alta (UNICRUZ)
Caixa Postal 838 – 98.005-972 – Cruz Alta – RS – Brasil

cassiolorenzett@gmail.com, telockenalex@unicruz.edu.br

Abstract. *This article presents the results obtained through a quantitative research by comparing two data mining algorithms along with their main characteristics, functions and importance for decision-making. The same concentrates on presenting the comparative study of the mining algorithms Random Forest and J48 belonging to the decision tree techniques on a database to be analyzed, thus providing check which algorithms have better performance, speed and accuracy the discovery of new standards.*

Resumo. *Este artigo apresenta os resultados obtidos através de uma pesquisa quantitativa por meio da comparação de dois algoritmos de mineração de dados junto com suas principais características, funções e importância para a tomada de decisão. O mesmo concentra-se em apresentar o estudo comparativo entre os algoritmos de mineração Random Forest e J48 pertencentes as técnicas de árvore de decisão diante de uma base de dados a ser analisada, com isto proporcionando verificar qual dos algoritmos tem melhor desempenho, velocidade e precisão na descoberta de novos padrões.*

1. Introdução

Entre os meios de armazenamento de dados existentes como banco de dados, nuvem(*cloud*), entre outras, torna-se necessário efetuar a mineração de informações, onde estas vão ser de grande importância para o ganho de conhecimento sobre o mercado onde a empresa ou negócio está aplicada, após este processo de estudo de mercado, são aplicações técnicas de mineração de dados (*Data Mining*) onde vão identificar padrões que se destacam e possibilitam novos caminhos para a tomada de decisão em uma base de dados. Por meio deste processo foi criado passos a serem seguidos para uma determinada empresa obter estes resultados vantajosos.

O seguinte trabalho aborda dois algoritmos que podem ser utilizados na mineração de dados referente à classificação das informações usando árvores de decisão. São eles: *Random Forest* e J48, onde é analisado o desempenho, performance e processamento de cada um na mineração de uma base de dados, sendo que o algoritmo de classificação conhecido como *Random Forest* foi proposto por Breiman, que consiste em uma técnica de agregação de classificadores do tipo árvore de decisão, construídos de forma que sua estrutura seja composta de maneira aleatória. Para determinar a classe de uma instância, o método combina o resultado de várias árvores de decisão, por meio de um mecanismo de votação. Ao final cada árvore dá uma classificação, ou um voto

para uma classe. A classificação final é dada pela classe que recebeu o maior número de votos entre todas as árvores da floresta. (DINIZ et.al, 2013).

O algoritmo J48 caracteriza-se por criar uma árvore de decisão a partir de uma base de dados para o ganho de conhecimento, e com isto formatar uma tomada de decisão. Este método possui o objetivo de construir uma árvore de decisão onde o atributo mais significativo é considerado a raiz da árvore. Conforme descrito acima cada algoritmo tem suas características e métodos para efetuar a mineração e classificação das informações para se tirar os objetivos necessários para maior conhecimento dos dados analisados. Com isto, este trabalho efetuou o estudo comparativo dos algoritmos *Random Forest* e J48 em uma base de dados para verificação de qual tem melhor performance, desempenho, precisão e velocidade na mineração.

Neste trabalho foi desenvolvido uma aplicação na linguagem de programação Java incorporando a API do *software* WEKA, e uma base de dados para efetuar a mineração dos mesmos através das duas técnicas descritas anteriormente, verificando qual tem melhor lógica, velocidade, desempenho e inteligência na mineração das informações, assim contribuindo cientificamente para futuros estudos e socialmente para empresas e instituições que busquem algoritmos para identificação de informações importantes em suas bases de dados.

2. Mineração de Dados

Com as informações disponíveis em mídias digitais, internet e bases de dados, o conhecimento se torna importante para o descoberta de dados chaves que se diferenciam dos outros, possibilitando novos caminhos para a exploração do conhecimento, a partir disto foi apresentada a mineração de dados no final da década de 80 onde caracteriza-se como um conjunto de técnicas automáticas de exploração e que possibilita o aprendizado necessário para a análise e formatação de grandes massas de informação de forma a descobrir novos padrões e relações que devido ao volume de dados, não seriam facilmente descobertas a olho nu pelo ser humano. (Amorim, 2006).

A mineração de dados é composta de várias técnicas criadas para facilitar a descoberta de algo novo que possa contribuir no conhecimento a partir de milhares de informações. Sendo assim serão citados brevemente a seguir os principais métodos que podem ser usados para aplicar a mineração de dados.

2.1. Classificação

Conforme o autor (CAMILO et.al, 2009), a classificação trata-se de uma tarefa de mineração de dados que tem o objetivo de identificar/classificar novas informações, dentro de grandes bases de dados ou outras informações para que possa se sobressair das demais criando novos padrões. A ideia é derivar uma regra que possa ser usada para classificar de forma otimizada, uma nova observação a uma classe já rotulada para isto, é usado outras métodos ou logicas desenvolvidas dentro da mesma para a geração do conhecimento necessário, que seriam: Árvores de Decisão, Redes Neurais, Raciocínio baseado em memória, Redes Bayesianas, entre outras.

2.1.1. Random Forest

Random Forest trata-se de um algoritmo classificador que faz uso do método de árvores de decisão criada por Breiman (2001) possibilitando a mineração dos dados passados a mesma. Esta técnica possui uma ideia um pouco diferente dos algoritmos de árvores de decisão, a qual pertence, enquanto uma árvore possui o objetivo de construção total de uma estrutura a partir de uma base de dados o Random Forest tem o objetivo de efetuar a criação de várias árvores de decisão usando um subconjunto de atributos selecionados aleatoriamente a partir do conjunto original, contendo todos os atributos e que estes possuem um tipo de amostragem chamado de bootstrap, a qual é do tipo com reposição, possibilitando assim melhor análise dos dados. (NETO, 2014).

Com a quebra das massas de dados e construção de vários subconjuntos, uma árvore de decisão é construída. Com este procedimento então a construção das árvores ocorre pela seleção de atributos aleatoriamente a partir dos subconjuntos, onde os mesmos são aplicados nos nós de cada uma das árvores criadas. Uma *Random Forest* ou floresta aleatória é um conjunto dessas árvores de decisão.

Após a criação dos conjuntos de árvores é possível efetuar a classificação de qual possui melhor ganho de conhecimento para a solução de determinado problema, para isto é necessário escolher um subconjunto de árvores de decisão que possui melhor lógica e vantagens para a tomada de decisão. Para cada subconjunto é dado um voto sobre qual classe o atributo chave deve pertencer, este voto possui um "peso" onde o mesmo é afetado pela igualdade entre as árvores, "sendo que quanto menor a similaridade entre duas árvores melhor, e pela força que cada árvore tem individualmente, ou seja, quanto mais precisa uma árvore for, melhor será sua nota." (NETO, 2014).

As *Random Forests*, conforme verificado anteriormente, possuem a característica de Dividir-para–Conquistar, e isto possibilita a mesma algumas características que se destacam referentes às outras técnicas, algumas delas são:

- Algoritmo mais poderoso do que comparado somente a uma árvore de decisão;
- Possui boa taxa de acerto quando testado em diferentes conjuntos de dados;
- Técnica exata;
- Evitam sobre ajuste (overfitting);
- Menos sensíveis a ruídos;
- Classificação aleatória das árvores sem intervenção humana.

A seguir é apresentado o funcionamento do método de classificação *Random Forest* através da Figura 1.

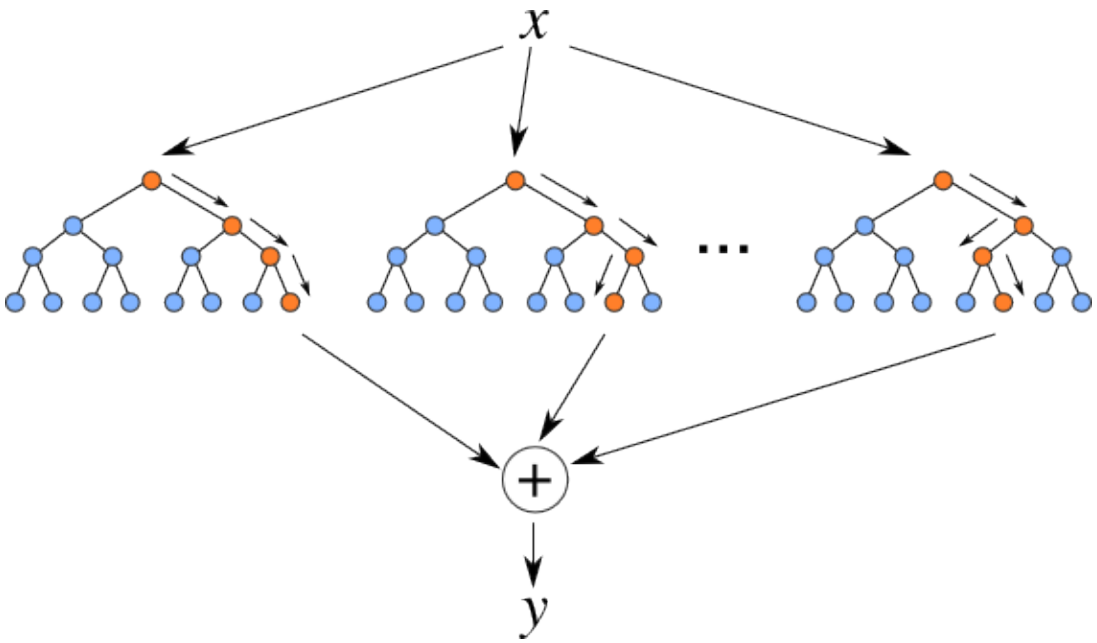

Figura 1. Ilustração da lógica por trás do algoritmo Random Forest

Na imagem anterior é possível verificar que partindo de um elemento X, no caso uma base de dados, gerou-se várias Random Forest, neste ponto cada uma gera várias regras e nelas a possibilidade de descoberta de novos padrões que poderão ser decisivos na tomada de decisão correta. Com as florestas criadas o próximo passo é calcular qual delas contém as regras mais exatas para a mineração. Com a escolha feita é aplicado na base de dados as mesmas e assim chegando a um resultado Y.

2.1.2. J48

Criado inicialmente por Ross Quinlan em 1993 trata-se de um algoritmo de mineração de dados pertencente a classe de classificação e árvores de decisão onde o mesmo possui o objetivo de dividir para conquistar, com isto o mesmo tem a vantagem de efetuar o ganho de conhecimento e a tomada de decisão a partir de uma construção de uma árvore com base em uma massa de dados. Este algoritmo é caracterizado pela a evolução de seus antecessores os quais são, ID3, C4.5, C5.0, onde os mesmo são implementados na linguagem C, e o J48 tem a mesma lógica só que este é implementado na linguagem JAVA. Uma das vantagens da aplicação deste algoritmo na tomada de decisão é que mesmo se mostra adequado para os procedimentos, envolvendo as variáveis qualitativas contínuas e discretas presentes nas bases de dados, e o mesmo permite a construção de árvores de decisão que classifica e apresenta em suas ramificações os atributos de maior relevância.

Este método usa a abordagem de dividir um problema complexo em partes menores onde proporciona a aplicação de uma estratégia recursiva para cada problema, com isto o mesmo divide o espaço definido pelos atributos em subespaços, associando-se a eles uma classe. Após a criação da árvore o algoritmo aplica técnicas de poda onde vão ser retiradas as “impurezas” da árvore deixando somente as informações importantes e que agregarão para a tomada de decisão. A poda deste método o corre de maneira a comparar taxas de estimativa de erro de cada subárvore do nó folha gerados, são processados sucessivos testes, a partir do nó raiz da árvore, de forma que, se a

estimativa de erro indicar que a árvore será mais precisa se os nós descendentes (filhos) de um determinado nós ”n” forem eliminados, então estes nós descendentes serão eliminados e o nó “n” passará a ser o novo nó (LIBRELOTTO, 2014).

J48 possui algumas características onde ele adquire conhecimento suficiente para a construção de uma árvore de decisão a partir de um banco de dados podendo assim possibilitar a tomada de decisão. Assim um seletor usado para a construção e análise das informações passadas ao mesmo é o information gain ratio ou taxa de ganho de informação, esta taxa é caracterizada pelo uso de uma métrica para ranquear todos os atributos de uma base de dados, a mesma é calculada utilizando o ganho de informação (Gain) de um atributo A contra o número de saídas que um teste com aquele atributo pode resultar (SplitInfo). O Gain Ratio tende a selecionar atributos com maior possibilidades de valores, mesmo que estes não sejam os mais relevantes. (NETTO, p. 11, 2013).

Para a construção de um nó a taxa de ganho é calculada para cada atributo e aquele que apresentar maior valor será usado no nó para dividir o conjunto de exemplos de treinamento. Se o subconjunto que deverá ser testado em um determinado nó só possua exemplos da mesma classe ou se todos os exemplos nele contidos apresentam a mesma taxa de ganho, então nenhum teste é proposto e uma folha é criada. (VASCONCELLOS, 2011). O algoritmo J48 tem a característica de testar um único atributo em cada nó, as árvores de decisão (AD) são consideradas univariantes ou axis-parallel (eixo-paralelo), as AD recebem este nome por ser testado um único atributo por nó, como descrito anteriormente. Este tipo de teste divide o espaço de atributos com um multiplano perpendicular ao eixo que representa o atributo em questão, ou seja, paralelo aos demais. Logo, a denominação axis-parallel refere-se a esta propriedade.

3. Escolha e formatação da base de dados utilizada na mineração

Para aplicação dos algoritmos Random Forest e J48, foi necessário escolher uma base de dados para minerá-la e com isto descobrir os padrões que cada uma gera para tomada de decisão. Para o início do trabalho foi selecionado as informações presentes na base de dados do SUS a DORS2013 onde a mesma possui todas as declarações de óbitos do ano de 2013 no estado do Rio Grande Do Sul, foi inicialmente aberta no software TabWin32 do próprio departamento de tecnologia do SUS o DATASUS, onde o mesmo proporciona manipular as informações presentes no arquivo possibilitando exportar a base para diferentes extensões como: sql, dbf, csv, xml entre outros.

Após de aberto a base no TabWin32 foi selecionado o arquivo DORS2013.dbf e salvo como extensão SQL. O próximo passo dado foi importar a base para o MySql usando a ferramenta HeidiSQL para execução do processo de KDD das informações presente na tabela. Neste estágio foi escolhido os principais atributos ou tuplas a serem minerados pela ferramenta, os atributos selecionados são os seguintes: Tipo_Óbito, Data_Nascimento, Data_Óbito, Local_Óbito, Assistencia_Médica, Estado_Civil, Raça_Cor e sexo sendo que cada um possui deferentes valores na sua composição, e a base de dados contém o valor total de 30000 (Trinta mil) registros cadastrados. A Tabela 1 demonstra cada valor pertencentes as suas tuplas.

Tabela 1. Atributos e seus possíveis valores da base de dados do SUS

Atributos/Tuplas	Valores
Tipo_Óbito	Não_Fetal
Data_Nascimento	Numérico
Data_Óbito	Numérico
Local_Óbito	Via_Publica , Domicilio , Hospital , Outro_Estab_Saúde , Outros , Ignorado
Assistência_Médica	Sem_Assistência , Ignorado , Com_Assistência
Estado_Civil	Ignorado , Viúvo , Casado , Solteiro , Separado_Judicialmente
Raça_Cor	Branca , Preta , Parda , Amarela , Indígena
Sexo	Masculino , Feminino

Na tablela anterior é possivel verificar que a base de dados formatada para a mineração ser inciada esta dividida em 8 grupos no caso os seguintes: Tipo_Óbito, Data_Nascimento, Data_Óbito, Local_Óbito, Assistência_Médica, Estado_Civil, Raça_Cor e Sexo e cada um deles com determinados valores, onde o sistema que vai minerar, classificará as informações cadastradas para a criações de um conhecimento e consequentemente tomar decisões apartir deste aprendizado gerado.

Com a formatação da tabela concluída para prosseguir com o desenvolvimento do projeto, foi exportada a mesma para o formato CSV e depois convertida para o formato ARFF onde vai ser através desta extensão que o sistema vai manipular e minerar as informações.

4. Resultados obtidos entre os algoritmos Random Forest e J48

Para início da comparação entre os algoritmos foi importado a base de dados do SUS, DORS2013 para WEKA (*Waikato Environment for Knowledge Analysis*), que trata-se de um software de mineração de dados pertencente a Universidade de Waikato (Nova Zelândia) que foi implementado pela primeira vez em sua forma moderna em 1997. Ele encontra-se licenciado ao abrigo da *General Public License* sendo portanto, possível estudar e alterar o respectivo código fonte. O software foi escrito na linguagem Java e contém uma GUI para interagir com arquivos de dados e produzir resultados visuais. Ele também tem uma API geral, assim é possível incorporar o WEKA, como qualquer outra biblioteca, a seus próprios aplicativos para fazer coisas como tarefas de mineração de dados automatizadas no lado do servidor.

O WEKA minerou as informações passadas a ele pelo arquivo gerado no caso a base do SUS, nos dois algoritmos escolhidos o Random Forest e J48, a partir deste

processamento o mesmo adquiriu conhecimento suficiente para tomar decisões na geração de novos padrões que podem ser facilmente analisados visualmente tanto em forma de gráficos ou em forma de texto pelo usuário final. No resultado final de cada mineração o software apresenta informações como tempo de execução, exatidão, desempenho, grau de certeza e incerteza sobre os dados, número de instancias, atributos entre várias outras características que o mesmo fornece. Para demonstração dos resultado alcançados inicialmente será mostrado os dados do algoritmo Random Forest e posteriormente os resultado do J48.

4.1. Random Forest

Ao término da mineração observou-se que o algoritmo teve 99.9633 % (por cento) de certeza com 29989 (Vinte e nove mil, novecentos e oitenta e nove) registros processados e 0.0367% (por cento) de incerteza na sua mineração com 11 (onze) registro processado, sendo que a base tem um valor de índices relativamente grande possuindo um total de 30000 (Trinta mil) registros incluídos. É possível verificar através de uma matriz gerada no resultado, chamada de matriz de confusão (Confusion Matrix) que dos registros presentes no arquivo 26491 (Vinte e seis mil, quatrocentos e noventa e um) foram minerados como de RACA_COR Branca. Dos 1406 (mil e quatrocentos e seis) registros minerados, 1402 (mil e quatrocentos e dois) foram classificados como de RACA_COR Preta, 3 (três) como Branca e 1 (um) como Parda. Dos 2015 (dois mil e quinze) registros da cor Parda, foram classificados 7 (sete) como Branca, e 2008(dois mil e oito) registros pertencentes a Parda. 28(vinte e oito) registros como Amarela e 60 (sessenta) como Indígena. Verificando, portanto que o algoritmo teve somente 3(três) erros ou 11 registros classificados errados, presente na RACA_COR Preta, minerando 3 registro como Branca e 1 como Parda, e na RACA_COR Parda minerando 7(sete) como Branca. Porem os demais não apresentaram nenhum erro em sua mineração.

Com relação ao tempo foi verificado que o mesmo tem variação entre a primeira execução do algoritmo no software e as demais que podem ser efetuadas, nos teste feitos foram executados 5 (cinco) tipos de resultados e cada um tendo um valor diferente em seu tempo de execução chegando em uma média final de 37.982. Figura 2 mostra os resultado obtidos em forma gráfica.

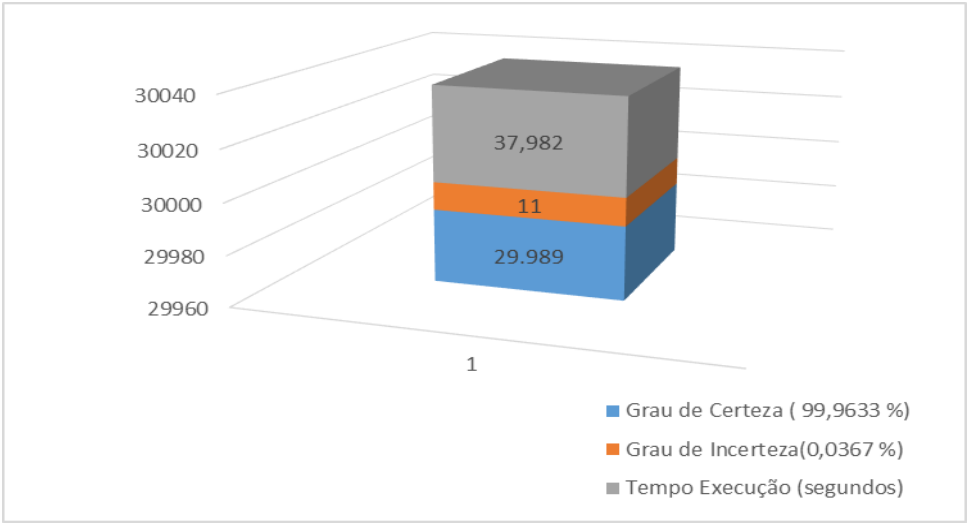

Figura 2. Gráfico relação grau certeza, incerteza e tempo execução Random Forest

O gráfico anterior presente na Figura 2 mostra os resultados obtidos durante a mineração da base de dados do SUS, nele é possível verificar que 99,9633% foi classificada corretamente o que equivale a 29.989 registros dos 30.000 presentes na base, e que 0,0367% foi classificado de forma errada correspondendo a somente 11 registros dos 30.000, tudo isto em um tempo médio de 37,982 segundos de execução.

4.2. J48

Ao término da mineração utilizando o algoritmo J48 observou-se que o mesmo teve 88.3033 % (por cento) de certeza com 26491 (vinte e seis mil, quatrocentos e noventa e um) registros processados e 11.6967 % (por cento) de incerteza na sua mineração com 3509 (três mil, quinhentos e nove) registro processado, sendo que a base de dados possui 30000 (trinta mil) registros armazenados. Na matriz de confusão verificou-se que dos registros presentes no arquivo, 26491 (vinte e seis mil, quatrocentos e noventa e um) foram minerados como de RACA_COR Branca, 1406 (mil e quatrocentos e seis) registros de cor Preta foram classificados como Branca, 2015 (dois mil e quinze) registros de cor Parda classificados como Branca, 28 (vinte e oito) de cor Amarela classificados como Branca e 60 (sessenta) de cor Indígena classificados também como Branca. Verificando, portanto que o algoritmo minerou corretamente somente os que pertenciam ao grupo de RACA_COR Branca e os demais o mesmo classificou de forma errada colocando os registros presentes para as RACA_COR Preta, Parda, Amarela e Indígena todas como se fossem Branca.

Com relação ao tempo foi verificado que o mesmo tem as mesmas variações presentes na execução do algoritmo Random Forest, mesmo assim foi aplicado a lógica anterior de calcular uma média dos tempos gerados durante 5 testes efetuados, chegando no resultado final de 1,378 segundos. Figura 3 mostra os resultado obtidos em forma gráfica.

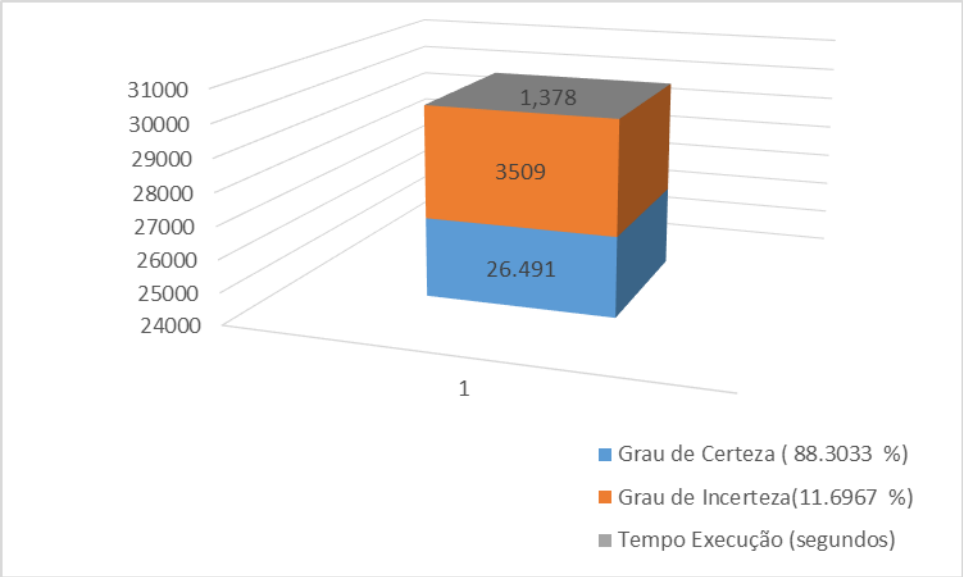

Figura 3. Gráfico relação grau certeza, incerteza e tempo execução J48

O gráfico anterior presente na Figura 3 mostra os resultados obtidos durante a mineração da base de dados do SUS, nele é possível verificar que 88.3033% foi

classificada corretamente o que equivale a 26.491 registros dos 30.000 presentes na base, e que 11.6967% foi classificado de forma errada correspondendo a 3.509 registros dos 30.000, isto em um tempo médio de 1,378 segundos de execução.

5. Conclusão

Neste Trabalho, foi proposto efetuar o estudo comparativo entre os algoritmos de mineração de dados pertencentes aos métodos de classificação chamados, Random Forest e J48 buscando contribuir na descoberta de qual das técnicas tem melhor desempenho, exatidão, agilidade entre outras características em sua implementação, juntamente com o desenvolvimento de uma aplicação onde fosse possível apresentar as informações da mineração para o usuário de uma forma de fácil identificação e com uma interface amigável.

Considerando que na literatura existem vários métodos e formas de os mesmos serem aplicados em uma base de dados para a classificação das informações, os motivos do desenvolvimento da pesquisa é poder concluir qual das técnicas estudadas geram melhores resultados em determinados ambientes, e com isto, possibilitar a tomada de decisão mais precisa.

A partir deste processo e dos resultados identificados anteriormente nas subseções 4.1 e 4.2 pode-se concluir, que o algoritmo de mineração de dados J48 é melhor em questão de performance em velocidade ou agilidade na mineração se comparada com a Random Forest, pois apresentou tempo significativamente baixo na análise da base de dados, mas em questão de desempenho para achar novos padrões existentes teve uma grande deficiência em sua execução.

Contudo, se comparado às duas técnicas em relação ao desempenho na mineração, o algoritmo *Random Forest* sobressai a J48, pois em uma base de dados relativamente grande obteve ótimos resultados em sua classificação, minerando 99,9636 % como instâncias relacionadas corretamente, porem em um tempo maior.

Portanto, primeiramente a principal contribuição científica observada no desenvolvimento da pesquisa se identifica na validação de uma teoria ou modelo que no campo de estudo atual ainda é discutida para obtenção de resultados definitivos. A segunda é que os algoritmos testados têm características próprias que os definem melhor em diferentes ocasiões, como é o caso do *Random Forest*, onde o mesmo é mais preciso e o J48 mais rápido na mineração. Com isto a aplicação destes métodos em bases de dados para o ganho de conhecimento depende principalmente se sua execução vai gerar os resultados satisfatórios ou não para a tomada de decisão em qual quer tipo de empresa ou instituição.

Referência

AMORIM, Thiago. **Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados.** Universidade Federal de Pernanbuco, 2006.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de Dados: Conceitos, Tarefas, Métodose Ferramentas.** Instituto de Informática, Universidade Federal de Goiás. 2009.

DINIZ, Fábio Abrantes ; NETO, Francisco Milton Mendes; JÚNIOR, Francisco das Chagas Lima Fontes, LAYSA Mabel de O. Fontes; **RedFace: um sistema de reconhecimento facial baseado em técnicas de análise de componentes principais e autofaces: comparação com diferentes classificadores.** Revista Brasileira de Computação Aplicada (ISSN 2176-6649), Passo Fundo, v. 5, n. 1, p. 42-54, abr. 2013.

LIBRELOTTO,Solange Rubert. **ANÁLISE DOS ALGORITMOS DE MINERAÇÃO J48 E APRIORI APLICADOS NA DETECÇÃO DE INDICADORES DA QUALIDADE DE VIDA E SAÚDE.** Universidade de Cruz Alta, 2014.

NETO, Cesare Di Girolamo**. Potencial de técnicas de mineração de dados para o mapeamento de áreas cafeeiras.** INPE, São José dos Campos, 2014.

NETTO, Oscar Picchi. **Um Filtro Interativo Utilizando Árvores de Decisão.** Universidade de São Paulo, 2013.

VASCONCELLOS, Eduardo Charles. **ÁRVORES DE DECISÃO APLICADAS AO PROBLEMA DA SEPARAÇÃO ESTRELA/GALÁXIA.** Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada – INPE, 2011.