Classificação Associativa Utilizando Seleção e Construção de Regras: um Estudo Comparativo

Gustavo E.A.P.A. Batista[1], Ronaldo C. Prati[1], Maria Carolina Monard[1], Rafael Giusti[1], Claudia R. Milaré[2]

[1] Laboratório de Inteligência Computacional (LABIC)
Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)
Caixa Postal 668 – 13560-970 – São Carlos – SP – Brasil

[2]Centro Universitário das Faculdades Associadas de Ensino (UNIFAE)
Caixa Postal 96 – 13870-377 – São João da Boa Vista – SP – Brasil

{gbatista,rcprati,mcmonard,rg}@icmc.usp.br, cmilare@fae.br

Resumo. *Classificação associativa é uma abordagem híbrida que tem se mostrado bastante competitiva com outros classificadores simbólicos. Nessa abordagem, regras de associação com o atributo classe como conseqüente são utilizadas como classificador. Uma limitação dessa abordagem é o grande número de regras geradas, sendo muitas delas redundantes. Para contornar essa limitação, foram propostos dois métodos de seleção de regras baseado na análise ROC:* ROCCER*, baseado em busca no espaço ROC; e* GARSS*, que aplica algoritmo genético para selecionar um subconjunto de regras que maximize a medida AUC. Neste trabalho, apresentamos o* MORLEA*, que utiliza um algoritmo genético multi-objetivo para aprimorar as regras em vez de selecioná-las. Resultados experimentais mostram que o* MORLEA *é capaz de induzir um classificador com número de regras inferior quando comparado com o classificador constituído de todas as regras de associação de classificação, e ao mesmo tempo que apresenta precisão semelhante a do algoritmo* C4.5

Abstract. *Associative classification is a hybrid approach that has been verified to achieve quite good results when compared to other symbolic classifiers. In this approach, association rules having the class attribute as its consequent are used as a classifier. However, the great number of rules (often including redundant ones) is a drawback of this approach. To overcome that problem, two rule subset selection algorithms based on ROC analysis have been proposed:* ROCCER*, which is based on geometric properties of the ROC space; and* GARSS*, which uses a genetic algorithm to select a subset of rules which maximises AUC. This work presents* MORLEA*, which uses genetic algorithm to evolve rules.* MORLEA *is capable of inducting classifiers with reduced number of rules, compared to classifiers with all class association rules, as well as achieving classification performance comparable to* C4.5.

1. Introdução

Aprendizado de regras a partir de conjuntos de dados é uma tarefa de fundamental importância em Mineração de Dados e Descoberta de Conhecimento (KDD). Por serem

facilmente compreensíveis por usuários não especialistas, regras podem prover valiosos *insights* a respeito dos dados. Tradicionalmente, o aprendizado de regras vem sendo tratado a partir de duas perspectivas diferentes: geração de regras descritivas e aprendizado de regras de classificação. A geração de regras descritivas explora principalmente conjuntos de dados não-rotulados (*i.e.*, dados que não contém o atributo classe) e o foco está em encontrar todas as regras que cumpram os requisitos mínimos estabelecidos pelo usuário, e que contenham associações e correlações entre os dados. Essas regras são geralmente chamadas regras de associação. Em contrapartida, o aprendizado de regras de classificação está relacionado a conjuntos de dados rotulados com o atributo classe e o objetivo é construir um conjunto não ordenado de regras ou uma lista de regras de decisão com um bom desempenho de classificação.

Nos últimos anos, uma abordagem híbrida vem ganhando destaque. Essa abordagem consiste na utilização de classificadores baseados em regras de associação, ou classificadores associativos [Liu et al. 1998, Yin and Han 2003]. Um classificador associativo é composto por todas as regras geradas por um algoritmo de regras de associação nas quais o conseqüente é o atributo classe. Essas regras são também conhecidas como regras de associação de classificação (*Class Associations Rules – CARs*). Por gerarem o conjunto completo de regras, o principal atrativo de um classificador associativo é a possibilidade de transpor uma possível limitação dos algoritmos de aprendizado de regras de classificação de desprezar regras potencialmente boas devido à busca heurística que esses algoritmos realizam [Domingos 1996]. Entretanto, essa facilidade geralmente vem acompanhada do fato que classificação por associação gera um enorme conjunto de regras. Alguns trabalhos recentes têm sugerido um passo adicional para a remoção de regras redundantes e irrelevantes, de maneira a tornar o classificador associativo mais compacto [Javanoski and Lavrač 2001]. Duas dessas abordagens baseadas na análise ROC foram propostas recentemente: o algoritmo ROCCER [Prati and Flach 2005] e o GARSS [Batista et al. 2006].

Uma nova abordagem para a geração de regras que vem ganhando a atenção da comunidade, principalmente com respeito à descoberta de conhecimento, é a indução de regras isoladas com propriedades específicas [Lavrac et al. 2004, Kavsek and Lavrac 2006]. A principal característica desses algoritmos é permitir uma maior flexibilidade relacionada aos parâmetros do algoritmo de maneira a ressaltar algumas propriedades de interesse com relação aos dados. Dentre essas abordagens, encontra-se a proposta de um algoritmo evolutivo multi-objetivo que pode combinar várias propriedades de interesse na função de aptidão multi-objetivo [Pila 2007]. O algoritmo MORLEA (*Multi-objective Rule Learning Evolutionary Algorithm*), tratado neste trabalho, é uma extensão desse algoritmo para gerar um modelo de classificação. A idéia é combinar esse algoritmo com uma estratégia de cobertura de conjunto, de modo que regras com propriedades específicas possam ser induzidas em todo o espectro do conjunto de treinamento. O objetivo principal não é criar um classificador altamente preciso, mas formado pelas regras geradas com propriedades específicas. O MORLEA foi comparado com os algoritmos ROCCER e GARSS em 4 conjuntos de dados da UCI. Resultados experimentais mostram que o MORLEA é capaz de induzir um classificador com número de regras inferior quando comparado com o classificador constituído de todas as regras de associação de classificação, e ao mesmo tempo que apresenta precisão semelhante a do algoritmo C4.5

Este trabalho está organizado da seguinte maneira: na Seção 2 são relatados alguns dos principais trabalhos relacionados a classificação associativa e a métodos de seleção de regras para esses classificadores; na Seção 3 são apresentados brevemente os algoritmos GARSS e ROCCER e é apresentado o algoritmo MORLEA; na Seção 4 são apresentados os resultados dos experimentos e, por fim, na Seção 5 são apresentadas as conclusões deste trabalho.

2. Trabalhos Relacionados

Algoritmos de aprendizado de regras de classificação normalmente efetuam busca gulosa por um conjunto de regras capaz de prever, de forma satisfatória, conjuntos de exemplos desconhecidos. De forma geral, tais algoritmos podem ser divididos em duas grandes famílias: separar-e-conquistar, tal como CN2 [Clark and Boswell 1991] e dividir-e-conquistar, tal como C4.5 [Quinlan 1993]. Algoritmos de regras de associação, por outro lado, efetuam uma busca global por todas as regras que satisfaçam certas restrições, como suporte e confiança mínimos. Embora algoritmos de regras de associação sejam principalmente utilizados para descrição de dados, *i.e.*, adotam uma abordagem de aprendizado não-supervisionado, é possível utilizá-los para predição, como foi proposto em [Liu et al. 1998]. Essa abordagem é conhecida como classificação associativa.

Muitos estudos mostraram que classificadores gerados por algoritmos de regras de associação podem ser tão competentes quanto classificadores induzidos por algoritmos de separação-e-conquista e de divisão-e-conquista [Yin and Han 2003]. No entanto, o número de regras de associação gerado pode facilmente exceder a capacidade de avaliação humana. Por isso, o uso dessa classe de algoritmos depende de técnicas para seleção das regras mais promissoras. Nesse sentido, algumas extensões têm sido propostas ao algoritmo de regras de associação Apriori [Agrawal et al. 1993] para possibilitar a construção de classificadores associativos mais compreensíveis. Um deles é o algoritmo Classification Based on Association – CBA – [Liu et al. 1998], o qual primeiramente avalia as regras para decidir quais delas serão adicionadas ao classificador, sendo essa avaliação baseada nas medidas de confiança e suporte das regras. Outra extensão foi implementada no algoritmo Classification Based on Multiple Association Rule – CMAR – [Li et al. 2001], o qual usa uma estratégia de poda em três estágios: as regras são podadas com base nas medidas de suporte e confiança, na correlação com outras regras e no número de exemplos de treinamento cobertos. Uma idéia semelhante foi explorada nos algoritmos AprioriC e AprioriSD [Javanoski and Lavrač 2001, Kavsek and Lavrac 2006], nos quais uma etapa de filtragem é aplicada para remover algumas das regras redundantes. Outras extensões foram implementadas, tais como o algoritmo Classification Based on Predictive Association Rule – CPAR – [Yin and Han 2003], o qual aplica busca gulosa para gerar um número de regras reduzido e um método proposto em [Veloso and Jr. 2005] para selecionar regras com base no custo causado por erros de classificação em vez da taxa de erro. Dois dos algoritmos utilizados para comparação neste trabalho, GARSS e ROCCER diferem dos trabalhos supracitados por usar AUC como métrica principal para seleção de regras e por utilizar algoritmos genéticos como método de busca.

3. Métodos

Nesta seção são descritos os três algoritmos comparados experimentalmente neste trabalho. Os algoritmos GARSS [Batista et al. 2006] e ROCCER [Prati and Flach 2005] se-

lecionam regras visando a maximização do AUC total, e o algoritmo MORLEA constrói regras com propriedades específicas a partir de um conjunto inicial de regras e maximiza uma função multi-objetivo relacionada à otimização dos dois critérios (*tpr* e *fpr*, descritos a seguir) utilizados no espaço ROC.

Em linhas gerais, um gráfico ROC é um gráfico que projeta a fração de exemplos positivos incorretamente classificados — taxa de falso-positivo (*fpr*) — no eixo x e a fração de exemplos positivos corretamente classificados — taxa de positivo-verdadeiro (*tpr*) — no eixo y. É possível representar tanto regras individuais, classificadores (compostos por conjuntos de regras ou não) ou mesmo classificadores parciais (compostos por subconjuntos de regras, por exemplo) num gráfico ROC. Uma curva no espaço ROC é um conjunto de pontos interligados, no qual cada ponto representa diferentes compromissos entre *tpr* e *fpr*. Dada uma curva ROC, é possível calcular a área abaixo da curva ROC. Essa área é um bom indicativo do desempenho do algoritmo para diferentes compromissos de *tpr* e *fpr* [Provost and Fawcett 2001].

3.1. GARSS

Genetic Algorithm for Rule Subset Selection – GARSS – é um algoritmo de seleção de regras que utiliza algoritmos genéticos para selecionar regras visando a maximização da métrica AUC. Algoritmos genéticos são métodos de busca baseados na seleção natural e na genética natural [Goldberg 1998]. Consistem de sucessivos conjuntos de soluções potenciais (população), codificadas como uma seqüência de bits ou números (indivíduos), os quais são obtidos com a aplicação de uma série de transformações (as mais utilizadas são os operadores de *crossover* e mutação), e pela avaliação da qualidade (aptidão) dos indivíduos como soluções para o problema em questão.

No GARSS, dada uma base de regras, uma chave primária é associada a cada regra da base. Assim, cada regra pode ser acessada de forma independente por meio da sua chave. Um vetor de chaves representa um indivíduo, *i*.e. um conjunto de regras que será interpretado como um classificador (solução potencial). Na implementação realizada, a população inicial é criada por um número arbitrário de indivíduos aleatórios, cada qual composto por uma quantidade arbitrária de elementos (regras). Para a avaliação experimental apresentada na Seção 4, a população inicial é construída com 30 indivíduos compostos por 30 regras. Como mencionado, a função de avaliação utilizada é a métrica AUC. O método de seleção é proporcional à função de aptidão, sendo o número de reproduções esperado para um indivíduo proporcional a essa função. Um operador de *crossover* em dois pontos foi aplicado com probabilidade 0.6 (a implementação permite probabilidade arbitrária). No *crossover* em dois pontos, dois indivíduos pais são selecionados da população e duas posições potencialmente distintas são aleatoriamente escolhidas. Cada posição é utilizada para separar um indivíduo-pai em duas partes, e estas partes são trocadas para gerar dois indivíduos filhos. Com *crossover* em dois pontos, os indivíduos filhos têm grandes chances de possuir tamanhos (quantidades de regras) distintos dos indivíduos pais. Assim, o GARSS pode convergir para conjuntos de regras com tamanhos distintos daqueles estipulados na população inicial. O operador de mutação altera, aleatoriamente, posições selecionadas de indivíduos, *i.e.*, este operador troca uma regra selecionada aleatoriamente por outra regra. Neste trabalho, a mutação foi aplicada com probabilidade 0.10, mas a implementação aceita valores arbitrários. As probabilidades de ocorrência de mutação e *crossover* foram escolhidas com base em nos-

sas experiências anteriores com algoritmos genéticos [Milaré et al. 2004]. Finalmente, um método de elitismo foi utilizado. De acordo com esse método, o melhor indivíduo de cada população é preservado ao ser copiado para a geração seguinte. Em contraste com o MORLEA (Seção 3.3), GARSS utiliza a representação *Pittsburgh*, na qual cada indivíduo da população representa um conjunto fechado de regras [Freitas 2002].

3.2. ROCCER

ROCCER [Prati and Flach 2005] é um algoritmo de seleção de regras baseado em algumas propriedades geométricas do gráfico ROC. Em [Fürnkranz and Flach 2005] é demonstrado como o aprendizado de regras, abordado por maximização da cobertura, pode ser visto como o traço de uma curva no espaço ROC. Para entender como isso é possível, assuma uma lista de regras vazia, representada pelo ponto $(0, 0)$ no espaço ROC. Adicionar uma nova regra R_j à lista de regras implica um deslocamento para o ponto $(\mathit{fpr}_j, \mathit{tpr}_j)$, onde fpr_j e tpr_j são a *tpr* e a *fpr* da lista parcial de regras (interpretada como uma lista de decisão) contendo todas as regras aprendidas até o momento, incluindo R_j. Uma curva pode ser traçada projetando-se todas as listas parciais $(\mathit{fpr}_j, \mathit{tpr}_j)$, com j variando de 0 ao número total de regras n na lista final de regras, na ordem em que elas são aprendidas. Uma regra padrão que sempre prevê a classe positiva pode ser adicionada ao final, conectando o ponto $(\mathit{fpr}_n, \mathit{tpr}_n)$ ao ponto $(1, 1)$.

No ROCCER, as regras são importadas de um conjunto externo maior de regras e o algoritmo realiza um passo de seleção baseado no fecho convexo atual do gráfico ROC. Essa abordagem é motivada pela observação de que classificadores ótimos, sob custos distintos de erro de classificação, posicionam-se no fecho convexo superior do espaço ROC [Provost and Fawcett 2001]. A idéia consiste em somente inserir uma regra na lista de regras se a sua inserção leva a um ponto fora do fecho convexo ROC atual (o fecho convexo ROC atual é o fecho convexo superior das regras que se encontram atualmente na lista de regras). Caso contrário, a regra é descartada.

3.3. MORLEA

O MORLEA é um algoritmo da família separar-e-conquistar implementado utilizando as facilidades da ECLE — Evolutionary Computing Learning Environment [Pila 2007]. ECLE é uma biblioteca de classes para executar e avaliar, sob diferentes cenários, um algoritmo evolutivo que tem como objetivo construir regras de conhecimento com propriedades específicas de forma isolada, ou seja, sem considerar o problema de interação entre as regras. Como na ECLE o objetivo é encontrar regras individuais de conhecimento com propriedades específicas, e não construir um classificador, a ECLE utiliza a representação *Michigan*, que é a mais apropriada para tal fim [Freitas 2002]. Nessa representação, cada indivíduo da população é uma regra na qual o algoritmo evolutivo atua diretamente até encontrar a melhor regra que satisfaça os critérios especificados.

O algoritmo evolutivo da ECLE utiliza uma rica estrutura para representar as regras (indivíduos), a qual possibilita utilizar uma grande variedade de operadores evolutivos e funções de avaliação. Quatro métodos de seleção (roda da roleta, torneio, *ranking* linear e *ranking* exponencial), três operadores de *crossover* (estrutural, de atributo e local) e dois operadores de mutação (estrutural e local) encontram-se atualmente implementados.

A ECLE pode ser executada com um ou vários operadores de *crossover* aplicados sucessivamente. O mesmo é valido para mutação. A função de avaliação pode ser

simples-objetivo ou multi-objetivo. No caso de simples-objetivo, encontram-se implementadas na ECLE uma vasta gama de medidas de avaliação de regras, tais como as propostas no *framework* de Lavrač [Lavrac et al. 1999], bem como outras medidas propostas na literatura. É possível compor uma função de avaliação multi-critério que considere qualquer conjunto dessas medidas de avaliação de regras, as quais são combinadas em uma função de avaliação simples-objetivo utilizando *rankings* [Pila et al. 2006]. Esse tipo de combinação é interessante pois, ainda que utilizando esse método não seja possível afirmar que o melhor indivíduo evoluído pertença a fronteira de Pareto, não há problemas de escala das medidas individuais já que elas são usadas somente para *rankear* os indivíduos, além de ser facilmente implementável.

Utilizando o algoritmo evolutivo implementado na ECLE, o MORLEA é o algoritmo de cobertura de conjunto padrão aplicado recursivamente para construir um classificador. Ou seja, iniciando com um classificador vazio, um conjunto de exemplos de treinamento, um conjunto de regras de conhecimento quaisquer relacionadas a esses exemplos e a função de avaliação a ser utilizada, a ECLE é ativada e retorna a melhor regra (indivíduo) evoluído. Essa regra é retirada da população da ECLE e adicionada ao classificador, todos os exemplos corretamente e incorretamente cobertos por essa regra são removidos e o processo é repetido até atingir o critério de parada. Finalmente, uma regra *default* do classificador é construída considerando a classe majoritária dos exemplos de treinamento que não foram cobertos pelas regras construídas pela ECLE. Diferentemente do ROCCER, e similar ao GARSS, o MORLEA gera conjuntos de regras não ordenadas.

Neste trabalho estamos interessados na comparação do MORLEA com os algoritmos GARSS e ROCCER que utilizam a AUC para construir o classificador. Assim, a função de avaliação multi-objetivo utilizada pela ECLE na execução do MORLEA tenta maximizar concomitantemente as medidas de *tpr* e 1 - *fpr* (*fpr* normalmente é minimizado no espaço ROC) da melhor regra. Entretanto, há uma diferença fundamental entre o MORLEA e os outros dois algoritmos: o MORLEA acha bons valores para essas duas medidas considerando as regras isoladamente, enquanto que o GARSS e o ROCCER consideram a interação dessas regras para maximizar a AUC. Diferentemente dos classificadores gerados pelo GARSS e ROCCER, que consistem de um subconjunto das regras de associação utilizadas, o MORLEA constrói novas regras utilizando as regras originais.

Nos experimentos apresentados neste trabalho, o MORLEA foi executado utilizando os seguintes parâmetros do algoritmo evolutivo da ECLE: seleção por torneio com tamanho de torneio igual a 5; *crossover* composto por *crossovers* estrutural, local e de atributo, cada qual com 60% de probabilidade de ocorrência e o fator $\alpha = 0.3$ para *crossover* local; mutação composta por mutação estrutural com 5% de probabilidade de ocorrência e mutação local com 25% de probabilidade de ocorrência. O critério de parada do algoritmo evolutivo para retornar ao MORLEA a melhor regra evoluída em uma iteração é que o desvio-padrão da função de avaliação para os últimos 20 indivíduos seja menor que 0.005, ou até um máximo de 50 iterações. O MORLEA termina a execução quanto o conjunto de treinamento fique vazio ou o ECLE gere em três iterações sucessivas regras que cobrem corretamente apenas um exemplo.

4. Resultados

Para avaliar experimentalmente o MORLEA, foi realizado um estudo utilizando quatro conjuntos de dados da UCI [Blake and Merz 1998]. Os conjuntos de dados empregados no estudo não possuem valores desconhecidos uma vez que o algoritmo Apriori não é capaz de lidar com esses valores. Na Tabela 1 são descritas algumas das principais características dos conjuntos de dados utilizados. Para cada conjunto de dados, é mostrado o número de atributos (#Atrib), número de exemplos (#Exemplos), e porcentagem de exemplos na classe majoritária. Embora o MORLEA seja capaz de manipular dados com mais de duas classes, os experimentos foram restringidos a conjuntos de dados com duas classes com o objetivo de facilitar o cálculo das curvas ROC e da medida AUC. A implementação de [Borgelt and Kruse 2002] do Apriori foi usada para gerar as regras de associação de classificação. O valor utilizado para o parâmetro confiança foi 50% e para o parâmetro suporte utilizou-se 1/3 da porcentagem da classe minoritária. O ROCCER seleciona regras a partir dessas regras geradas, enquanto que o MORLEA e o GARSS utilizam essas regras como um *pool* de regras para compor os indivíduos.

Conjunto de dados	#Atrib	#Exemplos	Classe Maj. %
Breast	10	683	65.00
Bupa	7	345	57.98
German	21	1000	70.00
Heart	14	270	55.55

Tabela 1. Conjuntos de dados do repositório UCI utilizados nos experimentos.

Nos experimentos realizados os valores AUC foram estimados utilizando 10-fold cross-validation estratificado. Ainda, para todos os algoritmos foram dados os mesmos conjuntos de treinamento e teste. Além do MORLEA, foram incluídos os resultados de outros quatro métodos nos experimentos: GARSS e ROCCER já foram utilizados para investigar formas de seleção de regras em um trabalho anterior [Batista et al. 2006]; C4.5 [Quinlan 1993] usado como uma referência para comparar os resultados, já que o algoritmo C4.5 é amplamente utilizado pela comunidade e reconhecidamente provê bons resultados; e All, um classificador composto por todas as regras de associação geradas pelo Apriori. De uma forma geral, é interessante criar um classificador que seja semelhante ou melhor em desempenho que C4.5 e que tenha um conjunto de regras com menos regras que o All.

Na tabela 2 são apresentados os resultados de desempenho de classificação dos métodos utilizados na comparação. Os resultados reportados são valores de AUC médios calculados sobre os 10 conjuntos de teste, e seus respectivos desvios padrão apresentados entre parênteses. Os resultados obtidos pelo MORLEA foram muito semelhantes aos obtidos pelo C4.5: para o conjunto de dados Bupa o MORLEA obteve valores médios AUC maiores que os valores médios obtidos pelo C4.5; para os conjuntos de dados Breast e German, o C4.5 foi ligeiramente melhor que o MORLEA; e, por fim, para o conjunto de dados Heart, os dois métodos, MORLEA e C4.5, obtiveram resultados semelhantes. Comparando o MORLEA com o GARSS e o ROCCER, pode-se notar que MORLEA obteve valor médio de AUC superior ao obtido pelo GARSS para o conjunto de dados Heart. Para todos os outros casos os resultados obtidos pelo GARSS e ROCCER foram superiores aos resultados obtidos pelo MORLEA.

Conjunto de dados	MORLEA	GARSS	ROCCER	C4.5	All
Breast	96,48(1,90)	99,06(0,46)	98,63(1,88)	97,76(1,51)	99,07(0,87)
Bupa	63,64(9,95)	64,65(3,96)	65,30(7,93)	62,14(9,91)	65,38(10,63)
German	69,04(6,23)	74,16(1,60)	72,08(6,02)	71,43(5,89)	73,37(4,84)
Heart	84,30(5,85)	82,86(3,36)	85,78(8,43)	84,81(6,57)	90,72(6,28)
Média	78,36	80,18	80,45	79,04	82,14

Tabela 2. Valores AUC médios e respectivos desvios padrão estimados utilizando 10-*fold cross-validation* estratificado.

Os valores médios AUC obtidos sobre todos os conjuntos de dados (última linha da Tabela 2) mostram que o MORLEA teve um desempenho médio muito parecido com o do C4.5, ficando um pouco abaixo. O ROCCER tem um desempenho médio um pouco melhor que o GARSS, e ambos os métodos foram melhor que o C4.5. Como esperado, os classificadores compostos por todas as regras de associação de classificação geradas pelo Apriori (All) desempenharam muito bem. Entretanto, esses classificadores são geralmente compostos por um número muito grande de regras limitando a compreensibilidade do classificador.

Na Tabela 3 é mostrados o número médio de regras induzidas para cada um dos métodos. Como mencionado anteriormente, o classificador All consiste de conjuntos com muitas regras. Pode-se observar que para os conjuntos de dados German e Heart, o número médio de regras é maior que o número de exemplos. Por outro lado, MORLEA, GARSS e ROCCER reduziram consideravemente o número de regras nos classificadores induzidos. Para o MORLEA, o número médio de regras para os conjuntos de dados Breast, Bupa, German e Heart representa 3,04%, 14,37%, 2,28% e 0,85% de todas as regras de associação (All) geradas para cada conjunto de dados, respectivamente. Por outro lado, a redução de desempenho de classificação considerando o classificador All (Tabela 2) foi de apenas 2,61%, 2,66%, 5,90% e 7,08% para os conjuntos de dados Breast, Bupa, German e Heart, respectivamente. De uma forma geral, para o número de regras geradas, MORLEA provê bons resultados nos conjuntos de dados Breast e Heart, gerando menos regras que GARSS e ROCCER. Embora o número médio de regras tenha uma alta variância para cada conjunto de dados, comparando o número médio de regras para cada conjunto de dados (última linha da Tabela 3), o MORLEA gerou conjuntos de regras um pouco menores que os demais métodos.

Conjunto de dados	MORLEA	GARSS	ROCCER	C4.5	All
Breast	15,30(3,32)	46,00(2,56)	48,40(2,32)	37,80(12,62)	502,10(8,96)
Bupa	42,10(11,17)	61,10(1,55)	3,90(0,99)	15,00(10,53)	292,80(21,57)
German	66,00(28,78)	34,90(5,88)	23,70(6,75)	78,20(18,50)	2886,10(577,30)
Heart	16,10(3,05)	43,90(1,89)	68,20(4,42)	13,20(4,49)	1875,60(91,90)
Média	34,87	46,48	36,05	36,05	1389,15

Tabela 3. Numero médio de regras e seus respectivos desvios padrão.

Para testar se os resultados obtidos são significativos, realizamos o teste de hipótese de Bonferroni-Dunn [Demšar 2006] utilizando o MORLEA como controle tanto nos valores da AUC quanto no número de regras. Quanto à AUC, somente a abordagem All supera o MORLEA, de acordo com esse teste, com 95% de confiança. Não foram encontradas diferenças significativas com relação aos outros métodos. Com respeito

ao número de regras, All se saiu pior que MORLEA, e também não foram encontradas diferenças significativas com relação aos outros métodos.

Considerando os resultados obtidos em relação ao número de regras e valores da medida AUC, o algoritmo C4.5 poderia ser considerado a melhor escolha, dado também que esse algoritmo é computacionalmente menos intensivo do que os demais. Entretanto, o conhecimento embutido nos classificadores gerados pelos métodos ROCCER, GARSS e MORLEA é potencialmente diferente do conhecimento gerado pelo C4.5, uma vez que as regras foram induzidas utilizando abordagens diferentes. Investigar essas possíveis diferenças será tema de trabalhos futuros.

5. Conclusão

Neste trabalho apresentamos o algoritmo para construção de conjuntos de regras MORLEA, que usa como base um algoritmo genético para a construção de regras com propriedades específicas. O MORLEA foi comparado com os algoritmos de seleção de regras GARSS e ROCCER, que selecionam um subconjunto de regras a partir de um conjunto maior de regras previamente construídas por algoritmos de regras de associação classificativa. Comparativamente com os métodos de seleção, o MORLEA obteve desempenho semelhante com esses algoritmos, gerando classificadores bastante compactos.

Agradecimentos

Trabalho realizado com auxílio do CNPq, FAPESP e FPTI – Fundação Parque Tecnológico Itaipu.

Referências

Agrawal, R., ImielinskiP., T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the International Conference on Management of Data*, SIGMOD, pages 207–216.

Batista, G. E. A. P. A., Milaré, C. R., Prati, R. C., and Monard, M. C. (2006). A comparison of methods for rule subset selection applied to associative classification. *Inteligencia Artificial*, (32):29–35.

Blake, C. L. and Merz, C. J. (1998). UCI repository of machine learning databases. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Borgelt, C. and Kruse, R. (2002). Induction of association rules: A priori implementation. In *15th Conf. on Computational Statistics*, pages 395–400. Physica-Verlag.

Clark, P. and Boswell, R. (1991). Rule induction with CN2: Some recent improvements. In *Proc. 5th European Conf. on Machine Learning*, volume 482 of *LNAI*, pages 151–163. Springer-Verlag.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Domingos, P. (1996). Unifying instance-based and rule-based induction. *Machine Learning*, 24(2):141–168.

Freitas, A. (2002). *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer-Verlag.

Fürnkranz, J. and Flach, P. (2005). ROC'n'rule learning – toward a better understanding of rule covering algorithms. *Machine Learning*, 58(1):39–77.

Goldberg, D. E. (1998). *Genetic Algorithms in Search, Optimization & Machine Learning*. Addison Wesley.

Javanoski, V. and Lavrač, N. (2001). Classification rule learning with Apriori-C. In *Proc. 10th Portuguese Conf. on Artificial Intelligence*, volume 2258 of *LNAI*, pages 44–52, Porto, Portugal. Springer-Verlag.

Kavsek, B. and Lavrac, N. (2006). Apriori-SD: Adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, 20(7):543–583.

Lavrac, N., Flach, P. A., and Zupan, B. (1999). Rule evaluation measures: A unifying view. In *International Workshop on Inductive Logic Programming*, pages 174–185.

Lavrac, N., Kavsek, B., Flach, P. A., and Todorovski, L. (2004). Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188.

Li, W., Han, J., and Pei, J. (2001). Cmar: Accurate and efficient classification based on multiple class-association rules. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 369–376. IEEE Computer Society.

Liu, B., Hsu, W., and Ma, Y. (1998). Integrating classification and association rule mining. In *Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining*, pages 80–86, New York, USA.

Milaré, C. R., Batista, G. E. A. P. A., Carvalho, A. C. P. L. F., and Monard, M. C. (2004). Applying genetic and symbolic learning algorithms to extract rules from artificial neural neworks. In *Proc. Mexican International Conference on Artificial Intelligence*, volume 2972 of *LNAI*, pages 833–843. Springer-Verlag.

Pila, A. D. (2007). Computação Evolutiva para a Construção de Regras de Conhecimento com Propriedades Específicas. Tese de Doutorado, ICMC-USP.

Pila, A. D., Giusti, R., Prati, R. C., and Monard, M. C. (2006). A multi-objective evolutionary algorithm to build knowledge classification rules with specific properties. In *6th International Conference on Hybrid Intelligent Systems (HIS 2006)*, Auckland, New Zealand. IEEE Computer Society. publicado em CD-ROM.

Prati, R. C. and Flach, P. A. (2005). ROCCER: An algorithm for rule learning based on ROC analysis. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI'2005)*, pages 823–828, Edinburgh, Scotland, UK.

Provost, F. and Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231.

Quinlan, J. R. (1993). *C4.5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Veloso, A. and Jr., W. M. (2005). Rule generation and rule selection techniques for cost-sensitive associative classification. In *20 Simpósio Brasileiro de Bancos de Dados*, pages 295–309.

Yin, X. and Han, J. (2003). CPAR: Classification based on predictive association rules. In *Proc. of the 3rd SIAM Int. Conf. on Data Mining*, San Francisco, CA. SIAM.