

UMA ANÁLISE DE MÉTODOS DE SELEÇÃO DE CARACTERÍSTICAS EM ENSEMBLES COM DIFERENTES MÉTODOS DE COMBINAÇÃO

Laura E. A. Santana, Diogo F de Oliveira, Anne M P Canuto and Marcílio C P de Souto

*Laboratório de Lógica e Inteligência Computacional, Departamento de Informática e Matemática Aplicada,
Universidade Federal do Rio Grande do Norte (UFRN)*

*E-mails: lauraemmanuella@yahoo.com.br, diogo.pow@gmail.com,
anne@dimap.ufrn.br, marcilio.souto@gmail.com*

Abstract— Feature selection methods are applied in ensembles in order to find subsets of features for the classifiers of the ensemble. The use of these methods aims to reduce the redundancy of the features as well as to reduce diversity of the classifiers of an ensemble. In this paper, a comparative analysis of six different feature selection methods is performed in ensembles using six different combination methods. The main aim of this paper is to investigate which combination methods are more affected by the use of feature selection methods.

Keywords— Ensemble, Feature Selection Methods, Combination Methods

Resumo— Métodos de seleção de características são aplicados em ensembles com o intuito de encontrar subconjuntos de características para os classificadores do ensemble. O uso desses métodos objetiva reduzir a redundância de características bem como reduzir diversidade dos classificadores do ensemble. Neste trabalho, uma análise comparativa de seis diferentes métodos de seleção de características é realizada nos ensembles usando seis métodos de combinação diferentes. O objetivo principal deste trabalho é investigar que métodos de combinação são mais afetados pelo uso de métodos de seleção de características.

Palavras-chave— Comitês, Métodos de Seleção de Características, Métodos de Combinação

1 Introdução

No passado, pesquisadores da área de reconhecimento de padrões focaram-se no projeto de únicos classificadores para resolver problemas de reconhecimento de padrões. Numa tentativa de melhorar o desempenho de reconhecimento dos classificadores base, uma abordagem comum é combinar vários classificadores, formando um Sistema Multi-Classificador (SMCs) [3,10]. SMCs, também conhecidos como ensemble ou comitês, exploram a idéia que um coleção de diferentes classificadores, referindo-os individualmente como classificadores base, pode oferecer informações complementares com relação aos padrões que serão classificados, melhorando a eficiência de todo o processo de reconhecimento [10]. Na literatura, *ensembles* têm sido amplamente utilizados para diversas tarefas dessa área. Na última década, por exemplo, um grande número de artigos propuseram a combinação de múltiplos classificadores para o projeto de sistemas classificadores de alto desempenho nas mais variadas áreas [3,6,12].

Neste contexto, um aspecto que tem sido reconhecido como muito importante é a diversidade [10]. Por exemplo, claramente não existe ganho algum na acurácia de um ensemble composto por um conjunto de classificadores bases idênticos. Uma forma de aumentar essa diversidade é usar ensembles compostos por classificadores base treinados com diferentes subconjuntos de características (entradas), isso é conhecido como seleção de características no ensemble. Há diversos métodos de seleção de características que podem ser usados por ensembles [2,4,9,13,14,15,16,18,19,20,21]. O objetivo principal deste trabalho é analisar a influência de importantes métodos de seleção de características em ensembles com seis métodos de combinação diferentes. Mais precisamente, o foco principal desse trabalho não são os métodos de seleção em si, mas os métodos de combinação

usados pelos ensembles, objetivando definir que métodos são mais efetivos para o uso de métodos de seleção de características.

2 Trabalhos Relacionados

Seleção de características ou de atributos tem sido um tópico de pesquisa tradicional, datando de pelo menos do início da década de 70. Este é um assunto amplo que atravessa disciplinas pesquisadas tais como estatísticas, aprendizado de máquina, fractais, conjuntos da teoria difusos (do inglês, *rough sets theory*) e programação matemática [5,8,17]. As vantagens da seleção de características são a redução da dimensionalidade do espaço de busca e remoção de dados redundantes, irrelevantes e ruidosos.

No contexto de ensembles, o objetivo dos métodos de seleção de características é reduzir o número de atributos de entrada presentes na base de classificadores e, ao mesmo tempo, lidar com problemas de dimensionalidade e diversidade. Recentemente, vários autores investigaram o uso desses métodos em ensembles, tais como em [2,4,9,13-16,18-21]. Por exemplo, o método do subespaço aleatório [9,19] depende de um procedimento pseudo-aleatório para selecionar um pequeno número de dimensões de um certo espaço busca. Em cada passo, essa seleção é feita e um subespaço é fixado.

Há também métodos de seleção de características não aleatórias, tais como [2,4,13,15,16,20,21]. Em [20], por exemplo, a correlação entre os erros dos classificadores é reduzido, separando os classificadores e treinando-os com subconjuntos de características de entrada diferentes. Isto difere do subespaço randômico, pois para cada classe de correlação entre cada característica e saída da classe é explicitamente calculada, e o classificador é treinado somente no subconjunto de características mais correlatas.

Recentemente, vários autores têm investigado métodos de otimização para projetar ensemble de classificadores, tais como pesquisa Tabu, arrefecimento simulado (do inglês, *simulated annealing*), algoritmos genéticos, entre outros [7,11]. Em [11], por exemplo, os autores sugerem duas formas simples de usar algoritmo genético para essa finalidade. Eles apresentam duas versões desse algoritmo. O primeiro usa apenas subespaços de atributos disjuntos, enquanto que outras possibilidades sobrepõem os subconjuntos de características. A função de adequabilidade empregada é a acurácia do ensemble. Além disso, em [7], os autores usam um algoritmo genético simples para explorar o espaço de todos os possíveis subconjuntos de atributos, a partir deles criam um ensemble. Nos seus experimentos, essa abordagem obteve melhor desempenho do que os métodos clássicos tais como *Bagging* e *Boosting*.

Em todos os trabalhos anteriormente mencionados, o foco principal é o método de seleção de característica. Entretanto, a acurácia de um ensemble também é fortemente afetada pelo método de combinação (regras para combinar as saídas dos classificadores base) usado pelo ensemble. Por exemplo, dois ensembles compostos pelos mesmos classificadores base, mas usando diferentes métodos de combinação, podem ter diferentes acurácias. Diferenciando-se dos trabalhos mencionados, como o previamente mencionado, o nosso foco não são os métodos de seleção de características, mas o impacto desses métodos em diferentes métodos de combinação usados pelos comitês.

3 Comitês de Classificadores

Como previamente mencionado, a meta de usar *ensembles* ou SMCs é melhorar o performance do sistema de reconhecimento de padrões com relação a uma melhor generalização e/ou ao aumento da eficiência e de um projeto nítido[3,10]. Há dois principais problemas no projeto de um comitê: os seus componentes e o seu método de combinação utilizado.

Com relação ao primeiro problema, a escolha correta do conjunto de classificadores base é fundamental para o desempenho de todo o comitê. Em uma situação ideal teremos um conjunto de classificadores base com erros não correlacionados – eles combinarão de tal maneira que os efeitos desses erros são minimizados. Em outras palavras, os classificadores bases devem ser diversos entre si. Diversidade pode ser alcançada por diferentes formas: variação dos parâmetros dos classificadores base (por exemplo, pesos iniciais e topologia de um modelo de rede neural), uso de diferentes exemplos da mesma base de dados como conjunto de treinamento, uso de diferentes tipos de classificadores base, entre outros.

Uma vez que o conjunto de classificadores tiver sido criado, o próximo passo será escolher uma forma eficaz de combinar suas saídas. Há um grande número de métodos de combinação na literatura [3,6,10,12]. De acordo com as suas funcionalidades, há na literatura, três estratégias principais de métodos de combinação: baseado em fusão, baseado em seleção e métodos híbridos.

- Métodos baseados em Fusão: Neste métodos, também conhecidos como métodos baseados em combinação, assumem que todos os classificadores base possuem experiências iguais em todo o espaço do conjunto. Sendo assim, as decisões de todos os classificadores são levadas

em conta para um certo padrão de entrada. Eles podem ser classificador de acordo com suas características como Linear ou Não-Linear [3]. Atualmente, a maneira linear mais simples de combinar a saída de múltiplos classificadores é a soma e a média das saídas. No caso do não-linear, alguns exemplos são: o voto majoritário (do inglês, Majority Voting), Dempster-Shafer, redes neurais e algoritmos genéticos, entre outros.

- Métodos baseados em Seleção: Nestes métodos, ao contrário dos baseados em fusão, somente um classificador é utilizado para classificar o padrão de entrada. Para fazermos isso, é importante definirmos um processo para escolher um membro do comitê que será responsável por fazer essa decisão, freqüentemente ele é baseado no padrão de entrada a ser classificado. A escolha do classificador usado para produzir a saída é feita durante a fase de operação. Essa escolha é tipicamente baseada no grau de certeza da decisão atual. Preferência é dada para o classificador com maior grau de certeza. Um dos principais métodos de seleção é a Seleção de Classificadores Dinâmicos (*Dynamic Classifier Selection – DCS*).
- Métodos Híbridos: Estes métodos são os que utilizam técnicas de seleção e fusão para prover uma saída mais adequada ao padrão de entrada. Normalmente, há um processo criterioso para decidir se será utilizado o método de seleção ou de fusão. Além disso, a idéia principal para usar seleção é se e somente se o melhor classificador for muito bom para classificar o padrão de teste. Caso contrário, o método de combinação é utilizado. Dois exemplos de métodos híbridos são: Seleção de Classificadores Dinâmicos baseados no comportamento de múltiplos classificadores (*Dynamic Classifier Selection based on multiple classifier behavior – Dcs-MCS*) e Seleção de Classificadores usando Templates de Decisão (*Dynamic classifier selection using also Decision Templates – Dcs-DT*) [10]. Embora esses métodos sejam considerados híbridos, eles usam o procedimento de seleção como primeira opção, sendo assim neste texto os consideraremos como métodos baseados em seleção.

4 Seleção de Características em Comitês

Com o intuito de reduzir redundância entre os atributos de uma padrão e aumentar a diversidade nos Comitês, distribuição de dados torna-se uma característica desejável em um comitê [4,13-15,18,21]. Distribuição de dados pode ser classificada como horizontal e vertical. Na horizontal, um classificador de um comitê conhece todas as características de um certo padrão, mas não de todos os padrões. Por outro lado, na distribuição vertical, um classificador de um comitê conhece algumas características de cada padrão, mas nenhum classificador conhece todas as características de um determinado exemplo [13]. Neste trabalho, a abordagem vertical é investigada, conhecida também como seleção de características em comitês.

No presente trabalho, dois tipos de distribuição vertical de características serão utilizados:

1. Distribuição completa: os classificadores de um comitê efetuam suas decisões baseadas em conjuntos completamente diferentes de atributos (subconjuntos de características disjuntas). Em outras palavras, não há sobreposição de atributos entre os classificadores base.
2. Distribuição parcial: os classificadores de um comitê têm suas decisões baseadas em conjuntos de atributos

parcialmente diferentes (sobreposição de subconjuntos de características). Neste trabalho, 50% de sobreposição entre conjuntos de atributos serão utilizados. Por exemplo, se um padrão de entrada tem 18 atributos, nove deles serão visto apenas por um único classificador e os nove restantes serão observados por todos os classificadores.

Além disso, os tipos de distribuição supracitados serão comparados com sistemas usando nenhuma distribuição de características entre os classificadores. Em outras palavras, todos os classificadores serão treinados com todas as características.

Como mencionado previamente, há vários métodos de seleção de características que podem ser usados pelo comitês [10], tais como seleção aleatória e não aleatória. No primeiro caso, cada classificador do comitê é construído sobre subconjunto de características escolhido aleatoriamente [9,19]. Em contrapartida, na seleção não aleatória, um determinado critério é usado para definir a importância de atributos. Baseado nessas informações, os atributos são distribuídos entre os classificadores. Há vários critérios que podem ser usados para ordenar os atributos, tais como variância e entropia.

Outra classe importante de métodos de seleção de atributos são as baseadas em algoritmos de otimização, tais como algoritmos genéticos [14,18]. Por exemplo, algoritmos genéticos fazem uma pesquisa aleatória guiada no espaço de busca de todos os subconjuntos de características. Isso pode nos levar a encontrar uma distribuição ótima das características entre os classificadores.

No presente trabalho, seis diferentes métodos de seleção de características serão utilizados, onde em um deles é baseado na seleção aleatória, quatro são baseados em seleção não-aleatória (Variância, Entropia, Erro e Correlação de Pearson) e o último é o algoritmo genético. Nas próximas seções, os métodos de seleção de características são descritos.

4.1 Métodos de Seleção Não Aleatória de Características

Todos os quatro métodos de seleção não aleatória de características usam os mesmos procedimentos gerais, mudando somente o critério para escolha dos atributos (parâmetros). O procedimento pode ser descrito como:

1. Ordenar todos os atributos T em ordem decrescente, baseado em um determinado critério (parâmetro) aplicado com relação ao conjunto de validação.
2. Colocar M ($M \leq T$) atributos na lista ordenada, na mesma ordem decrescente do passo anterior ($M=T$ para distribuição completa e $M = T/2$ para a parcial).
3. Para o primeiro N (normalmente $N =$ número de classificadores) atributos:
 - 3.1. Distribuir os atributos aleatoriamente entre os classificadores da forma que todos os classificadores sejam designados a um e somente um atributo.
4. Remova esses atributos da lista.
 - 4.1. Se a lista ordenada estiver vazia, pare a distribuição.
 - 4.2. Caso contrário, volte para o passo 3.

O uso deste procedimento anterior é uma tentativa de ser ter certeza que importantes atributos (baseado no critério escolhido) serão distribuídos entre todos os classificadores. Em aplicações onde o número de atributos não é igualmente dividido entre os números de classificadores, o número de atributos dos classificadores será diferente.

Como mencionamos previamente, a principal diferença entre eles é o critério (parâmetro) usado para ordenar os atributos. Neste trabalho, cinco diferentes critérios são usados na ordenação dos atributos:

- Variância estatística (método 1). Nos métodos de seleção de características, a ordenação de atributos baseado na variância é uma forma de medir a importância deles (atributos com variância alta são mais importantes para um classificador que atributos com variância baixa) [2].
- Entropia (mét 2). Entropia mede a média de informação necessária para identificar a classe de um padrão escolhido. Valores altos de entropia significam que um atributo é capaz de classificar corretamente o padrão de entrada, diminuindo a entropia do conjunto de padrões [4,13].
- Erro de classificação (mét 3). O parâmetro de erro define a relevância de um atributo para os classificadores, removendo esse atributo do conjunto de atributos e calculando a acurácia dos classificadores. Então, para uma base com N atributos, N classificadores são treinados e testados, casa um com $N-1$ atributos. Depois disso, a importância de cada atributo é verificado pelo erro apresentado pelo classificador que não tem esse atributo no conjunto de treinamento. Erro pode ser usado como parâmetro em alguns métodos de distribuição de dados [15,21].
- Correlação de Pearson (mét 4): Coeficiente de Correlação do de Perason (do inglês, *Pearson's Product Moment Correlation Coefficient* – PMCC) é um valor que indica a força de relacionamentos lineares entre variáveis [10]. Essa medida pode ser utilizada, por exemplo, para definir a correlação (dependência) das características de um padrão de entrada. Como um método de seleção de atributos, ele pode ser utilizado de forma que somente as menores características correlatas sejam colocadas juntas em um subconjunto de características alocadas para um classificador base. Dessa forma, todas as características são ordenadas baseadas na correlação em relação às outras características e distribuídas entre os classificadores.

4.2 Algoritmo Genético

Neste trabalho, o algoritmo genético (AG) padrão é usado para encontra o subconjunto ótimo de características para os classificadores de um comitê. O resultado do AG é o subconjunto de atributos alocado para cada classificador do comitê. A abordagem utilizada neste trabalho, que é similar ao descrito em [10], cada indivíduo da população representa todo o comitê. Nesse intuito, a função de adequabilidade é a acurácia do comitê representado pelo cromossomo. A representação dos indivíduos do AG depende do tipo de distribuição de características utilizada:

- Distribuição completa. Com o intuito de representar os subconjuntos disjuntos, o tamanho do vetor do cromossomo é n (número de características). Cada elemento do vetor usa inteiros de 0 a L (número de classificadores). O valor da posição i denotará que classificador usará a característica f_i e 0 significa que a característica f_i não será utilizada por nenhum classificador.
- Distribuição parcial. Para representar subconjuntos de sobreposição de características, os comitês podem ser representados por um cromossomo binário de tamanho $L \times n$. Os primeiros n bits representarão o subconjunto de características do primeiro classificador, seguindo pelos n bits para o segundo classificador, e assim por diante. É importante enfatizar que um cuidado especial deve ser

tomado para garantir que somente cromossomos viáveis sejam produzidos, onde 50% das características são compartilhadas por todos os classificadores e 50% das características são particulares para cada classificador.

5 Configuração dos Experimentos

Com o intuito de investigar o desempenho dos diferentes métodos de seleção de atributos aplicados a comitês usando diferentes métodos de combinação, um estudo empírico foi conduzido. Quatro das regas de combinação são métodos baseados em fusão (Soma, Voto, MLP e Classificador Naïve Bayes) e dois deles são métodos baseados em seleção (DCS-DT e DCS-MCB). Nessa investigação, o desempenho dos comitês usando a distribuição completa e distribuição parcial é comparada com comitês que não usam métodos de seleção de atributos (sem distribuição). Além disso, duas bases de dados diferentes são usadas:

- Base A – Imagens de Outdoor. Esse conjunto de dados foi pego do repositório UCI (segmentação) [1]. As 2.310 instâncias foram geradas aleatoriamente de uma base de dados de 7 imagens de outdoor. As imagens foram segmentadas a mão para criar uma classificação para cada instância, onde cada instância é uma região 3 x 3. 19 atributos foram extraídos de cada região.
- Base B – Proteínas. Essa base de dados representa uma classificação hierárquica, detalhada manualmente, estruturas, bem conhecidas, de proteínas. Elas são organizadas de acordo com seus relacionamentos evolucionário e estrutural. As principais classes de proteínas são all- α , all- β , α/β , $\alpha+\beta$ e small. É uma base desbalanceada, que contém um total de 582 padrões, os quais 111 padrões pertencem a classe all- α , 177 padrões para a classe all- β , 203 padrões da α/β , 46 da classe $\alpha+\beta$ e 45 da classe small.

Dois tamanhos diferentes de comitê serão utilizados nesta investigação. No primeiro, os comitês serão compostos por três classificadores. No segundo, os comitês serão compostos por nove classificadores base. Cinco tipos de classificadores base serão investigados: k-NN (vizinhos mais próximos), C4.5 (árvore de decisão), rede RBF (funções de base radial), rede neural MLP (multi-layer perceptron) e rede fuzzy-MLP. Para cada tamanho, 10 diferentes configurações serão utilizadas. Em todas essas configurações, classificadores bases diferentes e/ou topologias diferentes do mesmo classificador base serão utilizadas. Por questões de simplicidade, os valores mostrados neste trabalho representam a acurácia média de todas as configurações utilizadas.

6 Resultados e Análise

Antes de iniciarmos a investigação do desempenho dos comitês, é importante analisarmos a acurácia dos classificadores base para ambas as distribuições (completa e parcial). Como métodos de distribuição diferentes nos levam a diferentes subconjuntos de atributos distribuídos entre os classificadores base. Por exemplo, para a base de dados B, quando usamos comitês com 3 classificadores base, cada classificador terá 42 atributos. Por outro lado, para comitês com 9 classificadores base, cada classificador terá 14 atributos. Por causa disso, há duas colunas diferentes na Tabela 1, uma para o comitê com tamanho 3 e outro para o comitê com tamanho 9. Finalmente, a Tabela 1 mostra a acurácia média de todos os classificadores (k-NN, C4.5, MLP, RBF e Fuzzy MLP) para

cada configuração (método de seleção de atributos, distribuição de características e número de classificadores base).

Como pode ser observado da Tabela 1, a acurácia média dos classificadores base são muito similares para todos os métodos de seleção, com exceção do método de seleção aleatória de características. Essa variabilidade no resultado do método de seleção aleatória de características já era esperada já que a escolha é feita de forma aleatória. Para a base de dados A, a maior largura de acurácia média foi alcançada com o método de correlação de Pearson (PC). Em contrapartida, para a base de dados B, a maior largura de acurácia média foi obtida pelo método de Erro.

Tabela 1. Acurácia (ACC) e desvio padrão (SD) dos classificadores individuais para os dois tamanhos do ensemble (três e nove).

	Distribuição Completa		Distribuição Parcial	
	Três	Nove	Três	Nove
	ACC±SD	ACC±SD	ACC±SD	ACC±SD
Base de Dados A				
Ent	79.86±4.29	55.39±6.12	82.00±5.6 9	82.65±5.7 1
Error	77.53±4.65	55.28±5.95	84.66±6.4 1	85.04±5.1 5
GA	78.31±3.15	56.45±3.16	85.89±3.6 5	83.72±1.5 2
PC	79.14±8.97	58.89±6.70	89.12±4.8 8	87.87±4.8 3
Ran	72.44±5.19	53.69±6.26	85.50±5.6 0	84.23±5.1 0
Vari	79.24±4.81	55.75±7.32	88.42±4.8 5	86.84±5.6 3
Base de Dados B				
Ent	70.39±5.32	58.46±5.93	70.20±6.8 9	54.51±7.3 1
Error	69.64±6.59	60.41±5.86	72.45±7.9 7	70.22±8.2 5
GA	68.13±6.75	55.87±5.45	65.06±6.0 9	56.12±6.0 3
PC	64.42±5.44	55.60±6.34	60.39±6.7 4	57.33±5.9 4
Ran	63.84±9.47	53.94±7.02	65.45±12	64.82±9.8 8
Vari	64.64±6.32	54.61±6.13	68.57±5.6 6	69.22±5.8 2

Quando comparamos as acurácias dos classificadores base para comitês de tamanho 3 e 9, quando usamos o comitê de tamanho 3, os classificadores base sempre provêm um acurácia maior. Isto é um resultado importante, já que o número de características observadas para os classificadores base é menor para o comitê de 9 do que do comitê de 3. Por exemplo, para a base de dados A, os classificadores base do comitê de tamanho 3 (distribuição completa) tem 6 atributos, enquanto que os classificadores base para o comitê de tamanho 9 tem somente 2 atributos.

6.1 Comitês com três classificadores bases

Tabela 2 mostra a acurácia (AC) e o desvio padrão (SD) dos sistemas comitê quando nenhuma distribuição de características foi utilizada. Além disso, a Tabela 3 mostra a acurácia e o desvio padrão dos comitês usando três classificadores base quando eles foram aplicados as bases de dados A e B. Como mencionado anteriormente, seis diferentes métodos de seleção de características foram utilizados são usados por comitês usando seis diferentes métodos de combinação. Todos os comitês são aplicados nas distribuições completas e parciais. Valores mostrados nessas tabelas representam a acurácia média de todas as configurações utilizadas para cada tamanho do comitê.

Quando analisamos as Tabelas 2 e 3, nós podemos ver que as acurácias dos comitês foram maiores que os classificadores base (Tabela 1) para ambas bases. De maneira geral,

quando comparamos as primeiras e terceiras colunas da Tabela 2 com Tabela 3, como esperado, a acurácia do comitê diminuiu um pouco quando as características foram distribuídas entre os classificadores. Apesar de tudo, o uso de subconjuntos disjuntos de características (distribuição completa) não nos leva a diminuição na acurácia em alguns comitês para ambas as bases. Isto é um importante resultado já que mostra que quando atributos são bem distribuídos entre os classificadores a redundância pode ser diminuída e a acurácia dos comitês não é afetada pela diminuição do número de características observadas pelos classificadores.

Tabela 2. Acurácia (ACC) e desvio padrão (SD) dos seis métodos de seleção de características para comitês como três e nove classificadores base e usando nenhuma distribuição de atributos.

Base	A		B	
	Three	Nine	Three	Nine
	CM±SD	CM±SD	CM±SD	CM±SD
D_DT	94.21±3.1 5	95.18±2.5 5	80.79±5.3 0	84.11±6.0 2
D_MCB	96.97±2.1 2	97.44±1.9 4	80.04±4.2 9	83.43±4.4 9
Naive	96.52±2.2 6	95.95±3.0 9	80.45±4.2 5	80.75±5.1 1
NN	97.70±1.9 8	98.78±1.6 3	79.83±5.0 0	77.47±5.1 2
Sum	97.51±2.0 9	98.82±1.3 2	80.42±4.0 2	82.04±3.8 1
Voting	96.95±1.9 1	97.50±1.7 4	80.60±4.3 4	83.09±4.3 0

Dos métodos de combinação, as redes neurais proveram a mais alta acurácia de todos, para ambas as bases. Com o intuito de analisar o efeito dos métodos de seleção de características nos métodos de combinação, a menor acurácia de cada método de combinação quando usando um método de seleção de características é escolhida e comparada com acurácia do comitê quando nenhum método de seleção de características é utilizado. O principal objetivo dessa comparação é detectar a maior variação na acurácia de cada método de combinação quando usando e não usando um método de seleção de características. Quanto maior a variação, mais sensível para os métodos de seleção de características os métodos de combinação serão. Para a base de dados A, a menor variação foi alcançada por redes neurais (4.3), seguido pelo Classificador Naïve Bayes (6.2), Soma (7.61), DT (7.71), MCB (9.47) e Voto (9.93). Para a base de dados B, a menor variação foi alcançada por rede neural (6.83), seguida pelo Classificador Naïve Bayes (7.45), MCB (7.54), Soma (8.42), Voto (9.3) e DT (7.71).

Com o objetivo de avaliar se a variação na desempenho dado pelos métodos de combinação é significativa, o teste hipotético (t-test) compara, para todos os métodos de combinação, a menor acurácia de todos os métodos de seleção de características e quando nenhuma distribuição é usada, usando um nível de confiança de 95%. O uso de um nível de confiança de 95% significa que duas diferenças serão estatisticamente significativas se o valor p dessa comparação for menor que 0.05. Como um resultado do teste hipotético, podemos observar que todos os métodos de combinação têm variações no desempenho que são estatisticamente significativas para as duas bases.

6.2 Comitês com nove classificadores base

Tabela 4 mostra a acurácia e o desvio padrão do comitê usando nove classificadores base quando eles são aplicados nas bases de dados A e B. Todos os comitês são aplicados nas distribuições completas e parciais de características.

Analisando os resultados nas Tabelas 2 e 4, podemos ver que as acurácias dos comitês de 9 diminuiram muito pou-

co se compararmos com o comitê de 3, principalmente para a base de dados A. Em uma visão geral, quando comparamos as segundas e quartas colunas da Tabela 2 com a 4, como na seção anterior, a acurácia do comitê diminuiu um pouco quando as características são distribuídas entre os classificadores. Além de tudo, o uso da distribuição completa não nos levou a uma diminuição das acurácias de alguns comitês para ambas as bases de dados.

Entre os métodos de combinação, redes neurais proveram a maior acurácia total para a base de dados A. Por outro lado, o combinador Classificador Naïve Bayes proveu a maior acurácia para a base de dados B. Analisando a variação das acurácias dos métodos de combinação, como na seção anterior, para a base de dados A, a menor variação foi alcançada pela rede neural (5.78), seguida do Classificador Naïve Bayes (9.75), Soma (11.42), DT (14.88), Voto (15.1) e MCB (15.84). Para a base de dados B, a menor variação foi alcançada pela rede neural (3.77), seguida por Soma (9.94), Classificador Naïves Bayes (10.75), Soma (13.49), DT (14.01) e MCB (14.13).

Com objetivo de avaliar se a variação na acurácia alcançada pelos métodos de combinação é significativa, os testes hipotéticos (t-test) comparando a menor acurácia de todos os métodos de seleção de características e usando nenhuma distribuição, usando um nível de confiança de 95%, é realizado. Como um resultado do teste de hipótese, foi observado que a variação alcançada pelo combinador rede neural para a base de dados B não foi estatisticamente significativa (valor de $p = 0.054$). Todos os outros métodos combinadores têm variações em seus desempenhos que são estatisticamente significativas para ambas as bases.

7 Considerações Finais

Com objetivo de realizar uma investigação sobre a influência dos métodos de seleção de características em alguns métodos de combinação existentes, este trabalho analisou a influência de importantes métodos de seleção de características com seis diferentes tipos de métodos de combinação. Para fazermos isso, dois diferentes tipos de distribuição de atributos foram utilizados (parcial e completa). Além disso, dois tipos diferentes de tamanhos de comitês foram utilizados (com três e nove classificadores base).

Através dessa análise, podemos concluir que embora todos os métodos combinadores são afetados pelo uso de método de seleção de características, os métodos baseados em fusão são menos afetados que os baseados em seleção. O combinador de rede neural, por exemplo, foi o método menos afetado onde sempre proveu a menor variação na acurácia e essa variação não foi estatisticamente significativa em um dos casos analisados. Por outro lado, os métodos de combinação que proveram a maior variação foram sempre os métodos baseados em seleção (DT para comitês com tamanho 3 e MCB para comitê com tamanho 9). Baseado nesse experimento é possível afirmar que os métodos baseados em seleção são mais afetados pelo uso dos métodos de seleção que os baseados em fusão.

Referências Bibliográficas

- [1] Blake, C.L e Merz, C.J. UCI Repository of machine learning databases. Univ of California, Dept of Information and Computer Science. [http://www.ics.uci.edu/~mllearn/MLRepository.html]

- [2] Canuto, A.M.P., Santana, L.E.A. and Abreu, M.C.C. Analyzing the Performance of an Agent-based Neural System for Classification Tasks Using Data Distribution among the Agents. Int Joint Conf on Neural Networks (IJCNN), 2006.
- [3] Canuto, A.M.P. Combining neural networks and fuzzy logic for applications in character recognition. Tese de PhD, Univ of Kent, 2001.
- [4] Caragea, D., Silvescu, A. e Honavar, V. Decision tree induction from distributed, heterogeneous, autonomous data sources. In Conf on Int Systems Design and App (ISDA), 2003.
- [5] Chen, M., Han, J. E Yu, P. Data mining: an overview from database perspective. IEEE Trans. Knowledge and Data Engineering, 8(6):866–883, 1996.
- [6] Czyz, J., Sadeghi, M., Kittler, J. e Vandendorpe, L. Decision fusion for face authentication, Proc First Int Conf on Biometric Authentication, 686-693, 2004.
- [7] Gerra-Salcedo, C. e Whitley, D. Genetic approach to feature selection for ensemble creatin. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 236–243. Orlando-USA (1999)
- [8] Hall, M. A. Correlation-Based Feature Selection for Machine Learning. PhD thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, Apr. 1999.
- [9] Ho, T. K. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20(8), 832–844 (1998)
- [10] Kuncheva, L I. Combining Pattern Classifiers. Methods and Algorithms, Wiley, 2004.
- [11] Kuncheva, L. I. e Jain, L. C. Designing classifier fusion systems by genetic algorithms. IEEE Trans. Evol. Comput. 4(4), 327–336 (2000)
- [12] Lemieux, A e Parizeau, M. Flexible multi-classifier architecture for face recognition systems. The 16th International Conference on Vision Interface, 2003.
- [13] Modi, P. J. e Tae Kim P. W. Classification of Examples by multiple Agents with Private Features. Proc of IEEE/ACM Int Conf on Intelligent Agent Technology, 223-229, 2005.
- [14] Opitz, D. Feature selection for ensembles, in: Proc. 16th Nat. Conf. on Art. Intelligence, AAAI Press, pp. 379–384, 1999.
- [15] Provost, F.J. e Hennessy, D.N.. Scaling up: Distributed machine learning with cooperation. In Proc of the Thirteenth National Conference on Artificial Intelligence, 1996.
- [16] Rodriguez, J. J., Kuncheva, L. I. e Alonso, C.J. Rotation Forest: A new classifier ensemble method, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28 (10), 2006, 1619-1630.
- [17] Setiono, R. Neural network feature selector. IEEE Trans. Neural Networks, 8(3):654–662, 1997.
- [18] Tsymbal, A., Pechenizkiy, M. e Cunningham, P. Diversity in search strategies for ensemble feature selection. Information Fusion, Vol 6(1), pp. 83-98, 2005.
- [19] Tsymbal, A, Puuronen, S e Patterson, D. W. Ensemble feature selection with the simple Bayesian classification. Inf. Fusion 4, 87–100 (2003).
- [20] Tumer, K. e Oza, N. C. Input decimated ensembles. Pattern Anal. Appl. 6, 65–77 (2003).
- [21] Vaidya, J. e Clifton, C. Privacy preserving association rule mining in vertically partitioned data. In The Eighth Int Conference on Knowledge Discovery and Data Mining, 2002.
- [22] Woods, K. e Kegelmeyer, W. e Bowyer, K. Combination of Multiple Classifiers using Local Accuracy estimates, in: IEEE Trans on Patt Anal and Mach Intell, 19(4), 405-410, 1997.

Tabela 3. Acurácia e desvio padrão dos seis métodos de seleção de características para ensembles com três classificadores base e usando seis diferentes regras de combinação e com duas formas de distribuição de atributos (completa e parcial).

	Base de Dados A						Base de Dados B					
	DT	MCB	Naïve	NN	Sum	Voting	DT	MCB	Naïve	NN	Sum	Voting
Distribuição Completa												
Ent	91.9±4.3	92.7±2.3	94.8±3.0	96.2±2.3	94.1±2.4	91.8±2.6	75.5±5.0	75.9±5.3	76.9±4.7	73.2±4.6	76.3±5.1	75.9±5.2
Err	89.9±3.8	91.0±3.3	93.0±3.6	95.0±2.5	92.3±2.9	89.6±3.3	72.4±5.9	76.9±5.2	78.5±3.3	73.0±3.4	76.1±4.6	75.1±4.6
GA	89.8±6.2	89.4±7.9	94.3±6.6	94.6±3.4	90.1±7.9	90±4.0	71±15.4	74.8±11	77.6±6.9	74.1±4.0	73.5±8.5	72±19.6
PC	92.4±4.2	93.1±3.1	90.4±11.4	95.6±2.2	94.3±2.8	90.9±4.2	76.7±6.9	77.4±5.3	77.6±5.6	73.4±6.2	77.1±4.7	75.1±5.9
Ran	86.5±7.2	87.5±4.5	93.1±3.5	93.4±3.4	89.9±4.0	88.1±4.6	71.9±8.7	72.5±5.0	76.1±6.1	74.6±4.8	76.0±4.6	74.4±5.4
Vari	92.2±5.3	92.0±4.0	93.1±4.2	95.0±3.5	91.4±4.5	87.0±4.6	74.3±7.3	72.9±7.0	73.4±5.1	75.3±5.8	72.9±5.8	71.3±6.0
Distribuição Parcial												
Ent	92.9±4.6	93.1±3.3	93.9±4.1	95.9±2.7	93.6±3.0	90.1±3.9	75.0±6.5	75.1±6.0	76.3±4.8	73.8±4.5	76.1±5.2	75.7±5.2
Err	92.8±4.0	94.7±3.0	94.0±4.0	96.3±2.7	94.2±3.0	92.9±3.5	78.4±5.8	78.6±4.3	79.2±3.7	74.0±4.9	78.5±5.0	78.5±4.3
GA	95.0±3.1	96.0±3.1	94.3±6.2	95.1±2.6	91.7±3.3	90.2±2.4	77.1±4.1	79.1±5.1	80.1±3.1	73.8±4.5	72.0±7.0	78.4±3.6
PC	94.7±3.9	95.7±2.9	95.8±2.9	96.6±2.4	95.7±2.9	94.4±3.0	77.5±6.3	79.3±6.0	79.1±6.0	76.9±4.8	77.0±4.8	79.8±5.8
Ran	92.6±3.8	95.1±2.4	95.0±3.8	96.8±2.0	95.2±2.3	93.8±3.8	72.4±6.5	75.2±4.8	73±10.3	73.1±4.4	76.5±4.5	75.1±4.7

Tabela 4. Acurácia e desvio padrão dos seis métodos de seleção de características para ensembles com nove classificadores base e usando seis regras de combinação diferentes e com duas formas de distribuição de atributos (completa e parcial).

	Base de Dados A						Base de Dados B					
	DT	MCB	Naïve	NN	Sum	Voting	DT	MCB	Naïve	NN	Sum	Voting
Distribuição Completa												
Ent	80.3±11.7	81.8±4.4	86.2±17.2	94.4±2.7	88.4±4.3	83.7±4	72.3±3.3	75.7±6.6	75.3±6.2	76.1±5.6	77.9±5.5	76.5±6.3
Error	84.8±10.8	84.0±4.4	93.0±6.9	95.9±1.9	89.2±3.1	84.9±3.7	72.1±3.6	75.8±4.7	79.3±4.0	75.4±4.6	78.8±3.9	76.3±4.6
GA	85.8±8.5	84.5±4.5	89±10.1	94.0±2.3	90.3±2.4	86.2±2.3	74.1±6.8	75.6±3.1	85.1±6.3	74.5±3.0	72.1±8.3	72.7±8.2
PC	87.4±8.7	86.0±4.5	93.9±3.5	93.0±1.9	90.4±3.5	87.8±3	70.1±5.0	70.4±5.7	73.7±5.7	75.6±6.1	74.5±4.9	72.1±6.6
Ran	82.3±10.6	82.2±4.7	88±14.8	94.1±2.1	88.7±3.5	83.2±4.8	73.5±7.3	71.0±5.4	72.4±5.2	74.9±5.0	72.7±5.8	69.6±5.7
Vari	81.9±12.4	81.6±5.2	90.6±9.5	93.2±2.6	87.4±4.1	82.4±4.5	75.3±2.5	69.3±5.6	73.1±4.1	77.2±4.2	73.3±3.8	70.3±4.9
Distribuição Parcial												
Ent	92.1±7.6	94.1±2.4	86.2±15.8	97.8±2.5	95.7±2.1	94.0±2.3	71.0±5.8	70.7±6.6	70±10.4	74.3±5.1	75.1±6.5	70.9±7.3
Error	92.0±4.5	93.9±1.9	93.2±6.5	97.4±1.9	95.5±1.7	94.2±1.8	78.2±5.9	80.6±4.7	79.1±6.9	76.3±5.8	80.7±5.4	80.7±4.2
GA	92.0±3.1	93.4±0.9	92.8±3.6	96.4±4.9	98.6±1.0	97.2±4.5	81.1±6.1	82±6.5	81.1±7.8	78.1±2.6	80.4±4.5	81.1±6.5
PC	92.4±4.3	93.5±3.3	93.7±4.7	97.2±2.7	96.0±2.3	97.7±3.2	74.0±6.8	71.5±4.8	74.6±7.0	74.2±5.1	74.3±4.2	70.1±4.6
Ran	91.6±5.3	91.5±2.1	94.3±4.5	98.2±2.1	94.9±1.8	92.3±2.5	72.5±7.3	76.3±4.4	75.8±4.1	73.7±4.7	77.6±4.5	76.3±3.7