Tratamento de Características Ausentes via Subespaços Aleatórios e Imputação

Mirlem Rodrigues Ribeiro
Gerência de Serviços
Instituto Federal do Amazonas, IFAM
Manaus, Brasil
mirlem@ifam.edu.br

Eulanda Miranda dos Santos
Instituto de Computação
Universidade Federal do Amazonas, UFAM
Manaus, Brasil
emsantos@icomp.ufam.edu.br

***Abstract* - Databases with missing features are very frequent in processing and pattern recognition real applications, as well as in other fields such as data mining. The most frequent solution employed in the literature to deal with missing features is based on substituting missing values with meaningful estimates. This is the so-called imputation of missing values. In this paper, we propose to combine imputation methods with classifier ensembles generated by random subspace in order to reduce data corruption caused by imputation. We present experimental results obtained using different databases. These databases range from relatively high-dimensional feature spaces to small feature spaces. Our results show that classifier ensembles generated by random subspace help to reduce data corruption and lead to better performance.**

Keywords: Classification methods, ensemble of classifiers, random subspace, missing features, imputation methods.

I. INTRODUÇÃO

A ausência de características em aplicações reais é inevitável. Os desafios para esse tipo de problema giram em torno de questões como: valores desconhecidos, ou seja, incapazes de serem identificados ou que não puderam ser registrados por algum motivo técnico e/ou ambiental, ou mesmo por incapacidade de mensuração. As causas da ausência de características são diversas, por exemplo, pessoas que não respondem todas as questões de uma pesquisa, falha de pixels, falhas de sensores e/ou equipamentos, testes médicos que não podem ser realizados, ruído extremo, dados corrompidos, dentre outros [1].

Quando uma amostra a ser classificada não apresenta todas as características esperadas, uma ação comum é descartá-la, ignorando completamente essa ocorrência. Entretanto, essa não é uma solução adequada devido a pelo menos três razões: (1) a maioria das aplicações práticas demanda uma tomada de decisão, isto é, não é possível ignorar amostras; (2) se houver uma grande quantidade de amostras com características ausentes, a taxa de rejeição à tomada de decisão será elevada; e (3) em problemas em que cada amostra possui pelo menos uma característica ausente, nenhum dado será classificado. Além disso, valores ausentes podem de fato ter algum significado. Por outro lado, um classificador treinado normalmente não pode processar dados com características ausentes, uma vez que o processamento dos vetores de características de entrada torna-se impossível [2]. Portanto, é fundamental que as características ausentes sejam tratadas para que métodos de classificação sejam utilizados eficientemente. Além disso, o desempenho de um sistema de classificação está diretamente relacionado com as taxas de ausência de características. Segundo Nanni&Brahnam [1], 5% de características ausentes é geralmente uma taxa considerada tratável. Taxas entre 5% e 15% demandam métodos de tratamento *ad hoc*. Por fim, taxas de ausência de características acima de 15% são dificilmente tratáveis [1].

Dentre as principais soluções propostas nas últimas décadas para lidar com dados com valores ausentes em problemas envolvendo classificação ou regressão, destacam-se os métodos baseados em imputação, ou seja, o valor ausente é substituído por um valor estimado. A estimação pode ser feita por meio de técnicas estatísticas, como imputação pela média (dados quantitativos) ou moda (dados qualitativos) e regressão, ou por meio de técnicas de aprendizagem de máquina, como k *Nearest Neighbor* (k-NN) e Redes Neurais Auto Associativas [19]. Entretanto, métodos de imputação podem produzir distorções nos dados, principalmente quando ocorrem elevadas taxas de ausência de características [2].

Uma solução mais recente para tratar características ausentes é o uso de conjuntos de classificadores. Este método tem apresentado bons resultados na literatura [2], especialmente quando os membros dos conjuntos são gerados através do método de subespaços aleatórios [13]. Este método escolhe aleatoriamente *n* diferentes subconjuntos de características, a partir do espaço de características original, para treinar *n* classificadores. Portanto, como cada membro do conjunto é treinado com apenas uma porção das características, é possível classificar uma amostra com características ausentes ao selecionar os membros do conjunto treinados somente com as características disponíveis no dado a ser classificado [2].

Neste artigo é proposta a combinação entre métodos de imputação e conjuntos de classificadores gerados pelo método de subespaços aleatórios. Nossa estratégia baseia-se na hipótese de que essa combinação pode minimizar as desvantagens individuais de cada abordagem e elevar a taxa de acerto dos sistemas de classificação automática. É importante

destacar que investigamos a ausência de características apenas nos dados de teste.

O restante do artigo está organizado da seguinte forma. Na seção II é apresentada uma descrição resumida dos principais trabalhos relacionados encontrados na literatura e na seção III o protocolo experimental. Na seção IV os resultados são apresentados e discutidos. Por fim, a seção V descreve as conclusões e destaca as perspectivas para trabalhos futuros.

II. TRABALHOS RELACIONADOS

Os trabalhos discutidos nesta seção estão agrupados em quatro categorias, de acordo com os métodos empregados para tratamento de características ausentes: (1) uso de imputação e classificadores individuais; (2) uso de conjuntos de classificadores gerados por métodos diferentes de subespaços aleatórios; (3) uso de conjuntos de classificadores gerados por subespaços aleatórios; e (4) uso de imputação com conjuntos de classificadores.

Em [7], os autores investigaram quatro métodos de imputação: *k*-NN, imputação pela média ou moda, CN2 [15] e C4.5, sendo os dois últimos algoritmos de Árvores de Decisão. As bases usadas nos experimentos não possuem características ausentes. Portanto, a inserção dos dados ausentes foi completamente aleatória (MCAR) nos seguintes percentuais: 10%, 20%, 30%, 40%, 50% e 60% do total de amostras. Os autores mostram que *k*-NN obteve melhor desempenho entre os quatro métodos investigados. No entanto, apenas a técnica de Árvore de Decisão foi usada para medir o impacto dos métodos de imputação em seu desempenho. Além disso, as bases utilizadas são pequenas e possuem poucas características, pois, variam de 6 a 9 características.

Em [16], Branden e Verbone compararam cinco métodos de imputação: *k*-NN original, uma versão modificada de *k*-NN, um método de imputação de processo interativo, um método baseado em PCA (Análise dos Componentes Principais) Bayesiano [17] e uma técnica de imputação sequencial. Três bases de dados reais pequenas, porém com espaços de características relativamente elevados, em torno de 3226 características, foram usadas nos experimentos. Os autores também apresentaram a proposta de um método de imputação, ROB*impute*, pois, o objetivo do trabalho em [16] foi avaliar como os métodos de imputação se comportam diante da ocorrência de *outilers* em seu conjunto de dados. Foram testados sete diferentes percentuais de taxas de ausência introduzidas por MCAR: 1%, 3%, 5%, 10%, 15%, 20% e 30%.Os resultados mostram que o método proposto superou os demais métodos de imputação empregados, devido ao fato deste método ter sido projetado para ser robusto a *outliers*. De forma semelhante ao trabalho anterior, um único classificador (baseado em distância) foi usado nos experimentos.

Em [18], seis métodos de imputação de características ausentes foram comparados, sendo três estatísticos: média, *hot*-deck e imputação múltipla; e três métodos de aprendizagem de máquina: Rede Neural do tipo MLP (*MultilayerPerceptron*), Rede Neural do tipo SOM (*Self-OrganizingMap*) e *k*-NN. As características ausentes também foram inseridas usando MCAR, sendo utilizada apenas 30% de taxa de ausência. Os autores destacam que os métodos baseados em técnicas de aprendizado de máquina foram os mais bem sucedidos na imputação de valores ausentes. Novamente, apenas um classificador foi usado nos experimentos para medir o impacto dos métodos de imputação, neste caso, uma Rede Neural Artificial foi empregada.

Na mesma linha, Garcia-Laencina *et al*. [19] compararam quatro técnicas de imputação: *k*-NN, SOM, MLP e o algoritmo EM (*Expectation Maximization*). As seguintes taxas de ausência de características foram inseridas com base em MCAR: 5%, 10%, 20%, 30% e 40%. Os autores concluíram que não houve um único método superior em todas as bases testadas. Como nos trabalhos anteriores, apenas um classificador, Rede Neural Artificial, foi utilizado para medir o efeito de técnicas de estimativas de características ausentes.

Mais recentemente, em [20] foram comparados quatro grupos de métodos de imputação: *k*-NN , métodos baseados em probabilidade, métodos Bayesianos e imputação múltipla, aplicados ao problema de fusão biométrica. Para o conjunto de teste, as características foram removidas por MCAR em 10% e 25% de taxas de ausência. Os resultados indicam que *k*-NN foi melhor que todos os outros métodos investigados. É importante mencionar que a técnica de fusão *score-level*, ao invés de um classificador, foi usada para classificar amostras do conjunto de teste. Portanto, novamente apenas uma técnica foi usada para medir o desempenho dos métodos de imputação.

Técnicas envolvendo conjuntos de classificadores surgiram como alternativa aos métodos de imputação. Em [8], Juszczak *et. al* propõem combinar um conjunto de classificadores, em que cada membro do conjunto é treinado com uma única característica. O objetivo dos autores foi mostrar que um conjunto de classificadores é capaz de lidar com qualquer combinação de características ausentes, com o menor número de classificadores possível. Embora a estratégia seja simples, essa abordagem somente é eficaz quando as características individuais são suficientemente representativas para possibilitar a estimação razoável da superfície de decisão, fato que não ocorre com freqüência em problemas reais.

Ghannad-Rezaie *et al.* [3] propõem um método baseado na descoberta de padrões na classificação de valores ausentes em conjuntos de dados incompletos. Esta abordagem foi especialmente concebida para a classificação de conjuntos de dados com um pequeno número de amostras, elevado percentual de valores ausentes e com amostras de treino com valores ausentes. A proposta baseia-se em encontrar subconjuntos de amostras para as quais a maioria das características esteja disponível e treinar um classificador para cada subconjunto. Na fase de aplicação, para cada amostra de teste, um subconjunto de classificadores que não utilizam as características ausentes na referida amostra, é combinado para atribuir uma decisão final. O método proposto foi comparado a outras cinco estratégias: (1) exclusão; (2) CART; (3) EM; (4) imputação múltipla com EM; e (5) uma versão modificada de conjuntos de classificadores. Somente o método SVM foi usado para a classificação dos dados. Oito pequenas bases de

dados disponíveis no UCI *Repository of Machine Learning Datasets* foram usadas no experimento, em algumas as características foram removidas via MCAR e em outras, por meio da estratégia MAR (*Missing at Random*), nos seguintes percentuais: 10%, 20% e 30%. Os resultados experimentais mostram que a precisão na classificação do método proposto foi superior aos métodos comparados. Os autores também destacam que o nível de superioridade depende do padrão e do percentual de valores ausentes.

O terceiro grupo de métodos para tratamento de características ausentes envolve o uso de conjuntos de classificadores gerados por subespaços aleatórios. Provavelmente, a primeira proposta nessa linha de métodos é fundamentada no algoritmo de aprendizagem incremental *Learn*$^{++}$ [9]. Como evolução dessa abordagem, temos os trabalhos [10], [11] e [12], os quais também utilizam versões modificadas do *Learn*$^{++}$ para tratar o problema de características ausentes.

Esse método foi detalhado e comparado a outras estratégias em [2]. Trata-se do algoritmo *Learn*$^{++}$*.MF* (*Learn*$^{++}$*MissingFeatures*), onde um classificador é treinado com *l* características, dentre as *m* características disponíveis, escolhidas aleatoriamente. Em seguida, cada uma das *m* características recebe um peso, sendo que é atribuído um peso menor às *l* características escolhidas inicialmente, para reduzir as chances destas características serem selecionadas novamente para compor o subespaço do segundo classificador. Esse processo é repetido *n* vezes, e cada característica escolhida na etapa anterior, diminui a chance de ser sorteada na próxima execução. Na etapa de teste, as amostras com características ausentes são classificadas por voto majoritário dos classificadores cujos subespaços de treinamento não incluem as características ausentes. O desempenho de *Learn*$^{++}$ *.MF* foi comparado aos seguintes métodos: classificador *Naïve Bayes* individual com imputação média, uma abordagem que combina subespaços aleatórios com imputação média (descrito em mais detalhes no final desta seção) e a abordagem de conjunto proposta em [8]. Rede Neural do tipo MLP foi o classificador usado em todos os experimentos para medir o impacto dos métodos de tratamento de características ausentes na taxa de acerto do classificador. Foram usadas 11 bases de dados com vetores de características de tamanhos variados. As seguintes taxas de ausência de características foram inseridas via MCAR: 2.5%, 5%, 7.5%, 10%, 15%, 20%, 25% e 30%. Os autores mostram que *Learn*$^{++}$ *.MF* pode tratar uma quantidade substancial de características ausentes, com um aumento gradual da taxa de erro em função do aumento da taxa de valores ausentes. Entretanto, o número de classificadores candidatos, ou seja, válidos para o teste, diminui proporcionalmente em função do aumento do número de características ausentes. Esse problema é agravado quando o número de características originais é pequeno.

Por fim, dentre os trabalhos disponíveis na literatura, foram identificados dois artigos que combinam imputação com conjuntos de classificadores. Saar-Tsechansky e Provost [5] comparam diferentes técnicas de manipulação de características ausentes tais como, imputação pela média, baseada em distribuição (C4.5) e conjuntos de classificadores, chamado pelos autores de método híbrido, em que ocorre a combinação entre imputação e conjuntos de classificadores. Nesse método, conjuntos de Árvore de Decisão e Regressão Logística foram criados usando *bagging*. A escolha das amostras usadas para treinar cada membro do conjunto foi guiada pela probabilidade de ausência de características nas amostras de teste. Ou seja, os autores assumiram que as probabilidades de ocorrência de valores ausentes são conhecidas. Em algumas situações, os autores removeram características das amostras de treino para representar as probabilidades de ausência conhecidas. Na fase de teste, os classificadores que não dependiam das características ausentes foram selecionados para classificar cada amostra de teste. Quando não havia classificadores aptos, o método de imputação pela média foi usado para substituir os valores ausentes, seguindo uma política de imputação mínima. Nos experimentos realizados por Saar-Tsechansky e Provost [2007], foram usadas 15 bases de dados, algumas disponíveis no UCI. Algumas bases, especificamente 04, continham características ausentes. As características foram removidas aleatoriamente por MCAR das demais bases. Os resultados obtidos mostram que os modelos híbridos obtiveram melhor desempenho em relação à imputação baseada em distribuição (C4.5) e imputação pela média. Uma desvantagem desse método é a necessidade de conhecimento *a priori* das probabilidades de ocorrência de valores ausentes. Logo, quando ocorre ausência via MCAR, como neste trabalho, não há como criar os classificadores membros dos conjuntos.

Nanni *et al*. [1] desenvolveram uma nova abordagem usando imputação múltipla baseada em agrupamento (*clustering)* em que combinam imputação pela média com conjuntos de classificadores gerados por subespaços aleatórios. Os autores assumem a existência de características ausentes tanto na base de treino quanto na base de teste. O método inicia agrupando os dados de treino em um número fixo de grupos. Em seguida, os valores ausentes são imputados pela média calculada nas amostras que compõem o grupo. Um conjunto de 50 classificadores é gerado via subespaços aleatórios para cada grupo de dados. Na fase de teste, a amostra de teste é atribuída a um grupo, os valores ausentes desta amostra são imputados por meio da média dos valores dos membros de treinamento que compõem o grupo e as saídas dos classificadores gerados para o referido grupo são combinadas via regra da soma. Esse processo é repetido para todos os grupos de dados, ou seja, ocorre múltipla imputação dos valores ausentes.

A análise feita neste trabalho envolve duas técnicas de imputação: (1) pela média ou moda; e (2) baseado em aprendizagem de máquina, k-NN. São usadas 05 bases de dados cada uma variando entre 2.545 a 75.089 amostras e 11 a 180 características. Para a etapa de classificação, foram escolhidos dois tipos de classificadores, k-NN e Árvore de Decisão, usados individualmente e em conjuntos gerados por subespaços aleatórios. As principais diferenças entre nossa análise, e os trabalhos descritos nesta seção são:

1. Nossa análise verifica se o uso de conjuntos minimiza o ruído introduzido por métodos de imputação e se os métodos de imputação se comportam de forma diferente

quando se tem um classificador individual e um conjunto de classificadores.

2. Comparado com os trabalhos [1] e [2], em nossos experimentos são utilizados dois classificadores que trabalham muito bem em subespaços aleatórios, segundo a literatura.

3. Comparado com os trabalhos que analisam métodos de imputação, nossa investigação envolve o uso de dois métodos de classificação e não apenas um.

Os detalhes dos experimentos, assim como os resultados são apresentados na próxima seção.

III. PROTOCOLO EXPERIMENTAL

Nesta seção estão as informações relacionadas aos experimentos desenvolvidos.

A. *Bases de Dados*

Os experimentos foram realizados em 5 bases de dados com quantidades de características médias e reduzidas. As bases de dados são: DNA, Feltwell, NIST, Ship e Texture. A escolha dessas bases levou em consideração a disponibilidade e as variações em: quantidade de amostras, classes e principalmente as características. Além disso, o fato de serem investigadas bases com muitas amostras de dados diferencia este trabalho dos demais trabalhos existentes na literatura, pois, normalmente, são utilizadas bases com poucas amostras e poucas características.

A base de dados DNA está disponível através do Project Stalog em www.niaad.liacc.up.pt/old/statlog, *Texture* está disponível no UCI (http://archive.ics.uci.edu/ml/). A base *Feltwell* é uma base de imagens de sensoriamento remoto, enquanto a Ship é uma base composta por imagens de infravermelho de navios. A base NIST é formada por 10 classes de números manuscritos de 0 a 9. A Tabela I descreve as 06 (seis) bases utilizadas. A última coluna indica a quantidade de características utilizada para gerar subespaços aleatórios na geração de conjuntos de classificadores.

TABELA I. INFORMAÇÕES SOBRE AS BASES DE DADOS.

Base de Dados	Número de Amostras	Número de Classes	Número de Características	Subespaços Aleatórios
DNA	3.186	3	180	45
Feltwell	109.44	5	15	8
NIST	65.089	10	132	32
Ship	2.545	8	11	6
Texture	5.500	11	40	20

As bases de dados foram divididas em conjuntos de treino (±40% da base original), teste (±40% da base original), e validação (±20% da base original), de acordo com a estratégia *holdout validation*.

B. *Remoção Artificial das Características*

Apesar da ausência de características ser um problema comum em diversas áreas de aplicação, como bases para recomendação composta por quantidades significativas de amostras e características, inclusive ausentes, inclusive ausentes, optou-se pela remoção artificial de características nas 05 bases de dados públicas descritas na Tabela I, que não possuem ausência de características na sua composição original. É importante observar que esse procedimento é freqüentemente utilizado em trabalhos disponíveis na literatura.

As bases de dados foram divididas em treino e teste. Os classificadores foram treinados com as amostras sem características ausentes por meio do método de subespaços aleatórios de acordo com Ho [13,14]. As características foram removidas de forma artificial unicamente das amostras da base de teste, de modo completamente aleatório, dentro das seguintes faixas percentuais 0%, 2,5% 5%, 7,5%, 10%, 15%, 25%, 35% e 45%. As amostras com características ausentes foram, então, usadas para classificação.

C. *Métodos de Imputação*

Conforme mencionado anteriormente, dois métodos de imputação são investigados neste trabalho: (1) *mean* e (2) kNN. O primeiro denota o procedimento que substitui os valores ausentes das características da base de dados pela média (para características quantitativas) ou pela moda (para características qualitativas) das características presentes em todas as demais amostras. No segundo, para cada amostra com características ausentes, as k amostras da base de treino mais próximas desta, são identificadas, e a média dos valores das características correspondentes nas amostras mais próximas é utilizada para substituir os valores ausentes na respectiva amostra de teste. Em termos de método de imputação kNN, o valor k precisa ser definido pelo usuário. Em geral, o valor de k=5 foi o melhor para 05 bases em nossos experimentos.

D. *Conjuntos de Classificadores*

Foram treinados dois conjuntos com 100 classificadores homogêneos, combinados por voto majoritário: um conjunto com 100 árvores de decisão e um conjunto com 100 kNN. Essa quantidade de classificadores foi definida com base nos resultados obtidos por Ho [14]. No conjunto de kNN, foram treinados 100 classificadores 1-NN, sem ajuste do parâmetro k, pois o método de conjunto de classificadores assume que os membros do conjuntos são classificadores menos robustos, que ganham força ao combinar suas decisões. Também segundo Ho [14], metade das características originais, em média, é suficiente para treinar os membros de um conjunto gerado via subespaços aleatórios, a fim de garantir elevada taxa de acerto. Com isso, aproximadamente 50% das características foram selecionadas para compor subespaços aleatórios para os conjuntos de classificadores gerados com as bases Feltwell, Ship e Texture. Nas bases NIST e DNA, os subespaços foram gerados com apenas 25% das características, uma vez que essas bases possuem vetor de características relativamente elevado. Essas informações estão resumidas na Tabela I.

IV. RESULTADOS

Os experimentos realizados para comparar os dois métodos de imputação escolhidos foram divididos em duas séries: (1) considerando classificadores individuais; e (2) considerando os dois conjuntos de 100 classificadores homogêneos gerados por

subespaços aleatórios. Os resultados dessas duas séries de experimentos são apresentados nas próximas subseções.

A. Resultados da comparação entre métodos de imputação em classificadores individuais

Nesta série de experimentos, classificadores individuais foram utilizados para classificar as amostras de teste cujos valores de características ausentes foram imputados pelos dois métodos de imputação investigados. A Figura 1 apresenta a variação do erro em função da taxa de ausência de características para as 05 bases de dados.

Dado erro_ausência*%, o valor de erro obtido quando o classificador é usado para atribuir classes a amostras de teste com *% de valores ausentes (2,5%, 5%, 7,5%, etc.) e erro_ausência0%, a taxa de erro obtida sem ausência de características, a variação do erro é calculada da seguinte maneira:

$$variação\_erro = erro\_ausência^{*\%} - erro\_ausência^{0\%}$$

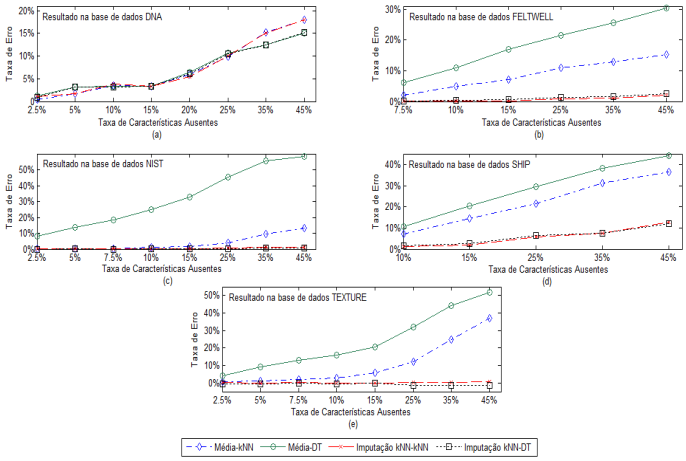

Fig. 1. Variação do erro nos classificadores individuais nas 05 bases.

Conforme pode ser observado na Figura 1, os resultados obtidos com os dois classificadores individuais k-NN e Árvore de Decisão mostram que o método de imputação k-NN é superior ao *mean*, em quatro das cinco bases investigadas. Os resultados do teste estatístico não paramétrico Kruskal-Wallis, apontam uma diferença significativa na variação da taxa de erro obtida com o uso dos dois métodos nas quatro bases consideradas. Apenas para a base DNA, os resultados foram diferentes.

B. Resultados da comparação entre métodos de imputação em conjuntos de classificadores via subespaços aleatórios

Neste segundo grupo de experimentos, os conjuntos de 100 kNN e 100 Árvores de Decisão são utilizados para classificar cada amostra de teste. Quando uma característica encontra-se ausente, o método de imputação é utilizado para estimar o valor ausente. A Tabela II apresenta os resultados obtidos com conjunto de 1-NN. Enquanto a Tabela III mostra os resultados obtidos com o conjunto de Árvores de Decisão. O símbolo "-" indica que não há amostras com características ausentes.

TABELA II. TAXAS DE ERRO OBTIDAS POR UM CONJUNTO DE 100 CLASSIFICADORES *1-NN* EM 04 BASES DE DADOS.

	Feltwell		Nist		Ship		Texture	
Taxa de ausência	*Mean*	*kNN*	*Mean*	*kNN*	*Mean*	*kNN*	*Mean*	*kNN*
0%	15,30	15,30	3,72	3,72	10,52	10,52	1,46	1,46
2,5%	-	-	3,76	3,83	-	-	1,51	1,29
5%	-	-	3,86	3,93	-	-	2,57	1,36
7,5%	16,00	15,71	3,99	3,99	-	-	2,80	1,36
10%	18,72	16,28	4,07	4,05	13,86	11,01	3,33	1,44
15%	19,72	16,26	4,29	4,30	19,17	12,29	5,75	1,44
25%	23,19	16,90	4,85	4,73	28,02	15,63	11,13	1,82
35%	24,79	17,47	6,71	5,40	35,59	18,19	22,34	1,74
45%	27,71	18,29	8,03	5,51	45,72	23,89	33,48	2,12

Em geral, pode-se afirmar que o método kNN é superior ao *mean*, mas essa superioridade é ainda maior no conjunto de Árvore de Decisão. A razão para esse comportamento é a instabilidade do classificador. Dado que o método *mean* insere muito mais ruído nos dados do que kNN, o classificador kNN , que é estável, sofre menos com esse efeito. Os classificadores instáveis têm comportamento oposto. Esse comportamento pode ser claramente observado na base NIST. Através da variação do erro dos dois conjuntos de classificadores é possível observar que ocorre uma diferença muito mais significativa entre os métodos de imputação no conjunto de Árvores de Decisão do que no conjunto de kNN .

TABELA III. TAXAS DE ERRO OBTIDAS POR UM CONJUNTO DE 100 CLASSIFICADORES ÁRVORE DE DECISÃO EM 04 BASES DE DADOS.

	Feltwell		Nist		Ship		Texture	
Taxa de ausência	*Mean*	*kNN*	*Mean*	*kNN*	*Mean*	*kNN*	*Mean*	*kNN*
0%	14,18	14,18	2,92	2,92	6,39	6,39	9,32	9,32
2,5%	-	-	3,12	2,98	-	-	9,77	9,24
5%	-	-	3,44	3,03	-	-	10,23	9,47
7,5%	15,11	14,34	3,74	3,05	-	-	10,68	9,17
10%	16,96	15,02	4,44	3,21	8,55	7,18	10,68	9,02
15%	19,63	15,48	6,26	3,46	17,11	8,26	14,32	9,32
25%	22,21	15,62	13,59	4,14	24,39	12,78	22,73	9,62
35%	26,37	16,64	26,95	4,92	35,10	14,75	35,45	9,24
45%	30,11	17,72	33,18	5,27	44,94	18,49	45,98	9,32

C. Comparação entre os resultados obtidos com classificadores individuais e conjuntos de classificadores

Os resultados apresentados na Tabela IV mostram o resumo da média dos erros obtidos para cada base. Portanto, os conjuntos de classificadores gerados por subespaços aleatórios foram melhores que os classificadores individuais quando todas as características estavam presentes e também quando as características estavam ausentes, isso porque conseguiram suavizar o efeito de ruídos tanto por meio do uso de um bom método de imputação, como *kNN*, ou com um método mais simples como *mean*. Porém, o método mais simples foi mais beneficiado porque introduz mais ruído nos

dados e, consequentemente, o conjunto suaviza mais o ruído gerado.

TABELA IV. MÉDIA DOS ERROS OBTIDOS PARA CADA UMA DAS 05 BASES.

Base	k-NN				Árvore de Decisão			
	Individual		Conjunto		Individual		Conjunto	
	Mean	kNN	*Mean*	kNN	*Mean*	kNN	*Mean*	kNN
DNA	21,79	21,86	18,61	18,39	28,51	28,41	16,83	**16,73**
Feltwell	24,14	17,19	20,78	16,60	33,44	18,44	20,66	**15,57**
Nist	7,96	4,99	4,81	4,38	38,91	10,53	10,85	**3,66**
Ship	30,91	17,29	25,48	15,26	34,68	15,92	22,75	**11,31**
Texture	11,12	1,63	9,38	**1,40**	30,88	9,06	18,80	9,30

Portanto, nossos resultados experimentais mostram que combinar imputação com conjunto de classificadores gerados por subespaços aleatórios reduz o ruído causado por imputação e ocasiona uma consequência ainda maior, a taxa de acerto geral do sistema melhora. Logo, a hipótese destacada na introdução deste trabalho foi confirmada pelos experimentos, isto é, a combinação de métodos de imputação com conjuntos de classificadores pode minimizar as desvantagens individuais de cada abordagem de tratamento de características ausentes e elevar a taxa de acerto de sistemas de classificação automática.

V. CONCLUSÃO E TRABALHOS FUTUROS

Neste artigo foi realizado um estudo comparativo focado em técnicas para tratar o problema de características ausentes. Os resultados das análises de todos os experimentos realizados mostram que o uso de conjuntos de classificadores minimiza o ruído introduzido pelos métodos de imputação, quando comparados a classificadores individuais. Embora o método de imputação *kNN* tenha superado o método *mean*, tanto com classificadores individuais quanto com conjuntos de classificadores, os experimentos mostram que os métodos de imputação se comportam de forma diferente quando se tem um classificador individual e um conjunto de classificadores. A diferença entre os dois métodos é maior em classificadores individuais do que em conjuntos.

Também é importante destacar que o método instável Árvore de Decisão é significativamente menos robusto a ruídos ocasionados por métodos de imputação do que o classificador *kNN*, tanto em conjunto quanto individualmente. Por outro lado, o conjunto de Árvores de Decisão alcançou as menores taxas de erro na maioria das bases investigadas. Com isso, pode-se concluir que conjuntos de classificadores gerados por subespaços aleatórios, especialmente conjuntos de Árvores de Decisão, podem elevar a taxa de acerto de sistemas de classificação automática em problemas com ausência de características.

Durante a execução deste trabalho foram identificadas algumas possibilidades de trabalhos futuros: (1) realizar experimentos com outros tipos de bases, levando em consideração características ausentes não só na base de teste, mas também na base de treino, semelhante aos trabalhos mais recentes encontrados na literatura e (2) combinar outros métodos de imputação com os classificadores individuais e os conjuntos usados neste trabalho. Além de acrescentar mediana, média ponderada e cruzamento como experimento fatorial.

REFERÊNCIAS

[1] L. Nanni, A. Lumini and S. Brahnam. “A classifier ensemble approach for the missing feature problem”, Artificial Intelligence in Medicine 55:37-50, 2012.

[2] R. Polikar, J. Pasquale, H. Mohammed, G. Brown and L. Kuncheva. “Learn++.MF: A random subspace approach for the missing feature problem”, Journal Elsevier: Pattern Recognition, vol. 43, pp. 3817-3832, 2010.

[3] M. Rezaie, H .Zadeh, H. Ying. and M. Dong. “Selection-fusion appoach for classification of datasets with missing values”, Journal Elsevier: Pattern Recognition, vol. 43, pp. 2340-2350, 2010.

[4] K. Kargupta, B. Park and H. Dutta. “Orthogonal decision trees”, IEEE Transactions on Knowledge and Data Enginneering, vol. 18, pp.1028-1042, 2006.

[5] M. Tsechansky and F. Provost. “Handling Missing Values when Applying Classification Models”, Journal of Machine Researh, vol. 8, pp. 1625-1657, 2007.

[6] T. Hastie, R. Tibshirani and J. Friedman. “The Elements of Statistical Learning”, Springer Verlag, New York, August 2001.

[7] G. Batista and M. Monard. “An analysis of four missing data treatment methods for supervised learning”, Applied artificial intelligence (2003), vol.17(5), pp. 519-533, 2003.

[8] P. Juszczak and R. Duin. "Combining One-Class Classifiers to Classify Missing Data," Multiple Classifier Systems MCS 2004, Lecture Notes in Computer Science, vol. 3077, pp. 92-101, 2004.

[9] S. Krause and R. Polikar. “An ensemble of classifiers approach for the missing feature problem” In Proceedings of the International Joint Conference on Neural Networks, pp. 553–550, 2003.

[10] H. Mohammed, N. Stepenosky and R. Polikar. “An ensemble technique to handle missing data from sensors”, IEEE Sensors Applications Symposium, Houston, Texas USA, 2006.

[11] J. Depasquale and R. Polikar. “Random feature subset selection for analysis of data with missing features”, Proceeding of International Joint Conference on Neural Networks, USA, 2007.

[12] J.Depasquale and R. Polikar. “Random feature subset selection for ensemble based classification of data with missing features”, Springer-Verlag berlin Heidelberg, 2007.

[13] T. Ho. “Nearest neighbors in random subspaces”, In Proceedings of the Second International Workshop on Statistical Techniques in Pattern Recognition, pp. 640–648, Sydney, Australia, 1998.

[14] T. Ho. “The random subspace method for constructing decision forests”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pp. 832–844, 1998.

[15] P. Clark and T. Niblett.The CN2 Induction Algorithm. Machine Learning, 3(4):261–283, 1989.

[16] K.V. Branden and S. Verboven Robust data imputation Computational Biology and Chemistry, 33(2009):07-13, 2009.

[17] C. M. Bishop. Bayesian PCA. In: Kearns, M.S., Solla, S.A., Cohn, D.A. (Eds.), Advances in Neural Information Processing Systems, vol. 11, pp. 382–388.1998.

[18] J. M. Jerez, I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence in Medicine, 50(2):105–15, 2010.

[19] P.J. Garcia-Laencina and J.L. Sancho-Gomez and A.R. Figueiras-Vidal Pattern Classification with missing data: a review. Neural Computing & Applications, 12(2):263-282, 2010.

[20] Y. Ding and A. Ross. Comparison of imputation methods for handling missing scores in biometric fusion Pattern Recognition, 45(2012):919-933, 2012.