Abordagens de Fusão de Algoritmos de Seleção de Atributos para Problemas de Classificação

Jhoseph Jesus, Daniel Araújo e Anne Canuto

***Resumo*—A grande quantidade de dados produzida pelas aplicações nos últimos anos precisam ser analisadas, a fim de extrair informações implícitas importantes. Algoritmos de aprendizado de máquina são ferramentas úteis para realizar essa tarefa, mas normalmente é necessário reduzir a complexidade dos dados usando algoritmos de seleção de atributos. Como sempre, muitos algoritmos foram propostos para reduzir a dimensão dos dados, cada um com suas vantagens e desvantagens. Baseado nisso, este trabalho propôs uma análise de duas abordagens distintas de combinação de algoritmos de seleção de atributos (fusão de decisão e fusão de dados). A análise foi realizada no contexto de classificação supervisionada, utilizando conjuntos de dados reais e artificiais. Resultados mostraram que uma das abordagens propostas (fusão de decisão) alcançou o melhor resultado para maioria dos conjuntos de dados.**

***Index Terms*—Seleção de Atributos, Comitês, Informação Mútua, Análise de Dados.**

I. Introdução

CENÁRIOS do mundo real tendem a possuir alta complexidade e, a fim de construir modelos mais aproximados, um grande número de variáveis (atributos) precisam ser usados. Problemas nos campos de Bioinformática, por exemplo, precisam de milhares de medidas de expressão gênica para descrever algumas dezenas de pacientes. Processamento de imagem, como segmentação e reconhecimento de padrões, usam pixels como atributos de imagens, resultando um número elevado de atributos para descrever uma única imagem.

Para lidar com esse problema, diversos métodos têm sido propostos nos últimos anos. A principal ideia de reduzir a dimensionalidade (número de atributos) de um conjunto de dados é encontrar um conjunto de atributos que possa representar todo conjunto de dados de forma que o problema possa ser tratado.

Uma alternativa comum utilizada pelos pesquisadores de aprendizado de máquina pode ser usada no contexto de seleção de atributos: abordagens de combinação ou comitê. Esse tipo de abordagem geralmente usa diversos métodos e combina suas saídas para produzir uma única solução, provavelmente melhor que as soluções individuais. Visando contribuir nessa importante área, este trabalho visa analisar duas abordagens distintas de combinação de múltiplos algoritmos de seleção de atributos.

J. Jesus and D. Araújo estão com o Instituto Metrópole Digital, Universidade Federal do Rio Grande do Norte, Natal/RN, Brazil, e-mail: jhoseph.kelvin@gmail.com, daniel@imd.ufrn.br

A. Canuto está com o Departamento de Informática e Matemática Aplicada, Universidade Federal do Rio Grande do Norte, Natal/RN, Brazil, e-mail: anne@dimap.ufrn.br

J. Jesus também está com o Departamento de Informática e Matemática Aplicada, Universidade Federal do Rio Grande do Norte

A primeira, combina as soluções produzidas por diferentes algoritmos de seleção de atributos usando um esquema de votação para criar uma única solução (fusão de dados). A segunda abordagem é baseada em comitês de classificadores treinados por conjuntos de dados reduzidos por algoritmos de seleção de atributos (fusão de decisão).

II. Abordagens de Fusão Propostas

A. Fusão de Dados

Um modo de combinar atributos obtidos por algoritmos de seleção de atributos é através do uso de um esquema de votação para escolher os atributos mais relevantes baseados na saída de cada algoritmo de seleção de atributos. Em outras palavras, essa abordagem provê a fusão dos atributos (dados) selecionados por diferentes técnicas de seleção de atributos e usa uma estratégia de votação para selecionar os atributos mais importantes. O esquema de votação é baseado na relevância que os atributos aparecem nas saídas de cada algoritmo. Uma visão geral da abordagem de fusão de dados pode ser vista na figura 1.

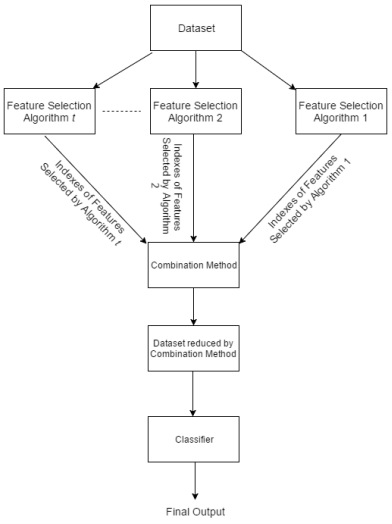

Figura 1. Fusão de Dados

B. Fusão de Decisão

A ideia é usar uma estrutura de comitê de classificadores como uma abordagem de fusão, onde a decisão de cada algoritmo de classificação é combinada por um método de combinação do comitê (Fusão de Decisão). Nesse contexto, a ideia dessa abordagem consiste em combinar algoritmos

de seleção de atributos usando um comitê de classificadores homogêneo. Logo, nós não combinamos as saídas geradas por cada algoritmo de seleção, mas a decisão provida pelos algoritmos de classificação treinados com os conjuntos de dados reduzidos pelos algoritmos de seleção de atributos. Uma visão geral da abordagem de comitê (Fusão de Decisão) pode ser vista na figura 2.

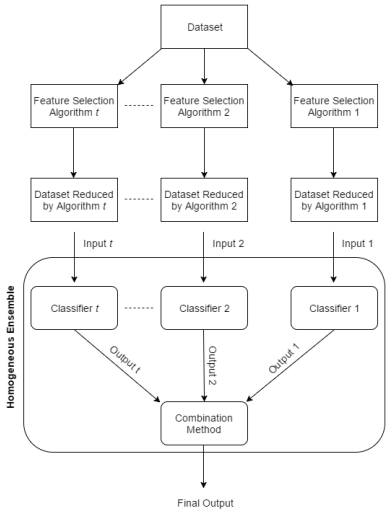

Figura 2. Fusão de Decisão

III. Materiais e Métodos

A. Algoritmos de Redução de Dimensionalidade

Selecionamos cinco algoritmos de redução baseados em Informação Mútua, utilizando a abordagem de [1] e [2]. Algoritmos baseados em informação mútua tem alto potencial de realizar seleção de atributos quando comparados com métodos tradicionais, isso se deve ao fato de que informação mútua, assim como outros descritores de Teoria da Informação, utilizam a informação dos próprios dados para quantificar e selecionar os atributos mais relevantes.

- Quadratic programming feature selection (QPFS);
- Spectral Relaxation Global Conditional Mutual Information (SPECCMI);
- Maximum Relevance Minimum Total Redundancy (MRMTR);
- Conditional Mutual Information Maximization (CMIM);
- Mutual Information Feature Selection (MIFS).

B. Algoritmos de Classificação

A fim de avaliar a performance das abordagens propostas, nós utilizamos três algoritmos de classificação, que são: Decision Tree [3], Naive B ayes Bayes [3] e k Nearest Neighbor (k-NN) [3].

C. Conjuntos de Dados

As principais características de cada conjunto de dados são apresentadas na tabela I, onde n é número de amostras, C é o número de classes e d é número de atributos (dimensionalidade).

Dataset	n	C	**Dist.of Classes**	d
LSVT	126	2	42,84	310
Lung Cancer	181	2	31,150	12533
Breast Cancer Diagnostic	569	2	212,357	30
Connectionist Bench	208	2	97,111	60
Ionosphere	351	2	126,225	32
St Jude Leukemia	248	6	15,27,64,20,40,79	985
Gaussian	60	3	20,20,20	600
Simulated	60	5	8,12,10,15,5,10	600
Friedman	1000	2	436,564	100
Colon Cancer	62	2	22,40	2000

Tabela I
Conjuntos de Dados

IV. Resultados e Discussão

Em resumo, baseado na análise empírica conduzida durante o projeto de pesquisa, podemos afirmar que a abordagem de fusão de decisão é o melhor método de seleção de atributos e, portanto, o melhor redutor de dimensionalidade, quando comparado a abordagem de fusão de dados e ao PCA, provendo uma performance superior para maioria das bases de dados. Bem como, a abordagem de fusão de decisão pode melhorar a performance, quando comparado ao conjunto original de dados (sem seleção de atributos/sem redução na dimensionalidade), para maioria das bases de dados.

V. Conclusões

Este trabalho de pesquisa gerou a criação de duas abordagens de combinação de múltiplos algoritmos de seleção de atributos. A primeira, combina soluções produzidas por diferentes algoritmos de seleção de atributos utilizando uma estratégia de voto para criar uma única solução. A segunda abordagem é baseada no sistema de comitês de classificadores treinados com os conjuntos de dados reduzidos pelos algoritmos seleção de atributos. Com objetivo de avaliar a performance das abordagens propostas, uma análise empírica foi conduzida. Nessa análise, a abordagem proposta utilizou três algoritmos de classificação diferentes (Árvore de Decisão, Naive Bayes e k-NN) e eles foram aplicados em dez bases de naturezas diferentes. Para efeito de comparação, foram aplicados os algoritmo de extração de características PCA e o conjunto original de dados (sem seleção de atributos). Através dessa análise, podemos afirmar que a abordagem de fusão de decisão é o melhor método de seleção de atributos quando comparada a abordagem de fusão de dados, PCA e ao conjunto de dados original, provendo melhores resultados para maioria das bases de dados.

Reconhecimento

Este trabalho foi parcialmente financiado pelo CNPq Universal Grant no 480.997 / 2013-6 e pelo programa de bolsa de estudos SmartMetropolis/UFRN.

Referências

[1] X. V. Nguyen, J. Chan, S. Romano, and J. Bailey, "Effective global approaches for mutual information based feature selection," in *Proceedings of the 20th ACM SIGKDD*. NY, USA: ACM, 2014, pp. 512–521.

[2] G. Brown, "A new perspective for information theoretic feature selection," in *Proceedings of the 20º AISTATS*, D. V. Dyk and M. Welling, Eds., vol. 5. Journal of Machine Learning Research, 2009, pp. 49–56. [Online]. Available: http://jmlr.csail.mit.edu/proceedings/papers/v5/brown09a/brown09a.pdf

[3] T. M. Mitchell, *Machine Learning*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997.