# Abordagem de Combinação de Algoritmos de Seleção de Atributos para Redução de Dimensionalidade

Daniel Araújo, Jhoseph Jesus, Adrião Dória Neto e Allan Martins,

***Resumo*—A quantidade de dados no mundo têm crescido exponencialmente devido ao número elevado de aplicações nos mais diversos contextos. Estes dados precisam ser analisados, a fim de extrair valiosas informações implícitas a partir deles. Técnicas de aprendizado de máquina são ferramentas úteis para realizar esta tarefa, mas a alta complexidade dos dados faz com que seja necessário o uso de métodos para reduzir tal complexidade. Redução de dimensionalidade (seleção de atributos) é um dos métodos mais utilizados para alcançar esse objetivo. Este trabalho propõe um método de combinação de algoritmos de seleção de atributos com o intuito de criar uma solução única e mais estável. Nós testamos essa abordagem usando conjuntos de dados reais e algoritmos de aprendizado de máquina. Resultados mostraram que nós podemos usar a solução gerada pela combinação com pouca ou nenhuma perda de acurácia de classificação. A abordagem de combinação pode ser utilizada como uma escolha estável quando se possui pouco conhecimento acerca do problema.**

***Index Terms*—Redução de Dimensionalidade, Informação Mútua, Teoria da Informação, Classificação, Seleção de Atributos.**

## I. Introdução

NOS últimos anos, o número de aplicações que geram dados têm crescido de forma tremenda. Esses dados precisam ser analisados e as técnicas de aprendizado de máquina são uma das opções para realizar essa análise, a fim de descobrir informações subjacentes valiosas. Entretanto, cenários do mundo real tendem a ter alta complexidade, e, com intuito de construir modelos mais realísticos, um grande número de variáveis (atributos) precisam ser considerados.

Dados provenientes de diversas aplicações, como por exemplo sensores utilizados em cidades inteligentes são, em geral, muito complexos e necessitam de um pré-processamento. Para tanto, fizemos a redução de dimensionalidade desses dados. Problemas em outros diversos campos, como na Bioinformática, por exemplo, necessitam de milhares de medidas de expressão gênica para descrever algumas dezenas de pacientes.

Com uma grande quantidade de atributos, a maioria dos algoritmos de aprendizado de máquina sofrem para encontrar boas soluções devido ao problema da maldição da dimensionalidade e obter mais amostras do problema nem sempre é possível. Então, uma possível solução é reduzir o número de atributos [1].

Baseado nisso, este trabalho se propôs a definir uma solução simples para combinar atributos selecionados de diversos algoritmos de seleção de atributos baseados em Informação Mútua com a intenção de combinar diferentes perspectivas dos algoritmos em uma única solução.

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

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

A. Dória Neto está com o Departamento de Computação e Automação, Universidade Federal do Rio Grande do Norte

A. Martins está com Departamento de Engenharia Elétrica, Universidade Federal do Rio Grande do Norte

## II. Abordagem de Combinação Proposta

### A. Método de Combinação

Uma maneira simples de combinar atributos obtidos por diversos algoritmos de seleção é usar um esquema de votação para escolher os atributos mais relevantes, baseado na saída dos diferentes algoritmos de seleção de atributos. Em outras palavras, essa abordagem provê a fusão dos atributos selecionados por diferentes algoritmos de seleção de atributos e usa uma estratégia de votação para selecionar os mais importantes. A técnica de votação é baseada na relevância em que os atributos aparecem nas saídas de cada algoritmo. A combinação de algoritmos de seleção de atributos têm sido utilizada com sucesso na literatura de reconhecimento de padrões, tais como em [2] e em [3]. Uma visão geral sobre o método de combinação pode ser visto na figura 1.

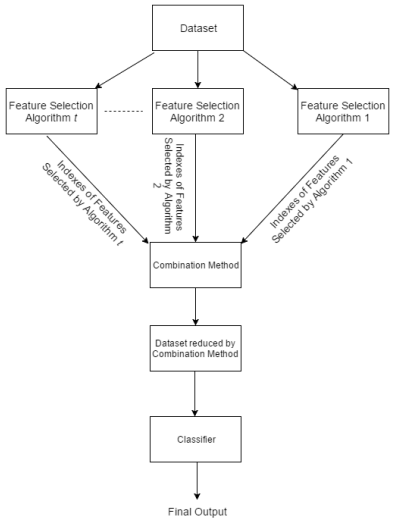


Figura 1. Método de Combinação

Utilizando os atributos selecionados **f**, nós contamos quantas vezes cada atributo aparece em cada solução encontrada pelos algoritmos de seleção, ponderado pela sua relevância. A relevância, nesse contexto, é inversamente proporcional a sua posição no vetor de atributos. Logo, a relevância do $f_i$ atributo pode ser definida como:

$$r_i = \frac{1}{j} \tag{1}$$

onde $j$ representa a posição do atributo no vetor de saída. Por exemplo, se um atributo é a primeira escolha de um algoritmo, sua relevância é igual a um. Se aparecer na quarta posição, então sua relevância será de 0.25. Usando essa estratégia nós consideramos não apenas a presença do atributos na saída dos algoritmos, mas sua importância em todo o processo.

## III. Materiais e Métodos

### A. Algoritmos de Redução de Dimensionalidade

A fim de testar nossa abordagem de combinação, nós seguimos a abordagem usada em [4] e selecionamos nove algoritmos de seleção de atributos baseados em Informação Mútua para nossa análise, como eles tornaram publica suas implementações no Matlab®[1] [2].

- Maximum relevance (maxRel);
- Minimum redundancy maximum relevance (MRMR);
- Minimum redundancy (minRed);
- Quadratic programming feature selection (QPFS);
- Mutual information quotient (MIQ);
- Maximum relevance minimum total redundancy (MRMTR);
- Spectral relaxation global Conditional Mutual Information (SPEC_CMI);
- Conditional mutual information minimization (CMIM);
- Conditional Infomax Feature Extraction (CIFE).

### B. Algoritmos de Classificação

A fim de avaliar a performance da abordagem proposta, nós utilizamos dois algoritmos de classificação, que são: Support Vector Machine (SVM) [1] e $k$ Nearest Neighbor ($k$-NN) [1].

### C. Conjuntos de Dados

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

| **Dataset** | $n$ | $C$ | **Dist.of Classes** | $d$ |
|---|---|---|---|---|
| LSVT | 126 | 2 | 42,84 | 310 |
| Lung Cancer | 181 | 2 | 31,150 | 12533 |
| Semeion Digits | 1593 | 10 | 161,158,162,159,159,161,159,161,158,155 | 256 |
| Connectionist Bench | 208 | 2 | 97,111 | 60 |
| Ionosphere | 351 | 2 | 126,225 | 32 |

Tabela I
Descrição dos Conjuntos de Dados

[1]avaliable at http://www.mathworks.com/matlabcentral/fileexchange/47129-information-theoretic-feature-selection

[2]avaliable athttp://www.mathworks.com/matlabcentral/fileexchange/26981-feature-selection-based-on-interaction-information

## IV. Resultados e Discussão

Em termos de impacto sob algoritmos de classificação, no contexto deste trabalho, nós podemos observar que não existe uma melhor técnica para realizar a seleção de atributos para todos os algoritmos, bases de dados ou dimensões alvo. Se um algoritmo de redução de dimensionalidade tem melhor desempenho para um conjunto de dados, é provável que não será o melhor para outro. Logo, escolher um algoritmo de seleção de atributos deve ser uma tarefa difícil se o pesquisador não tem as informações necessárias acerca do domínio do problema ou dos dados. Por outro lado, o método de combinação proposto produz soluções mais robustas para todos os conjuntos de dados. Se observarmos os resultados, podemos notar que, para maioria dos casos, a performance do método de combinação está nós três maiores valores. Isto é, se um pesquisador procura por uma solução mais estável ao invés de um bom resultado, o método de combinação pode ser uma escolha mais segura. Na tabela II nós mostramos a performance ranqueada para o método de combinação em todos os conjuntos de dados. Nós também calculamos a média (com desvio padrão), mediana e moda para análise deste ranking.

| CA | SVM | | | k-NN | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TD | 2D | 3D | sqrt | 2D | 3D | sqrt | | | | | |
| | Rank | Rank | Rank | Rank | Rank | Rank | Mean | Std | Median | Mode | |
| LSVT | 2 | 4 | 5 | 1 | 1 | 2 | 2.5 | 1.6 | 2.0 | 2.0 | 4/6 |
| Lung Cancer | 1 | 1 | 1 | 1 | 1 | 1 | 1.0 | 0.0 | 1.0 | 1.0 | 6/6 |
| Semeion Digits | 1 | 5 | 5 | 1 | 5 | 5 | 3.7 | 2.1 | 5.0 | 5.0 | 2/6 |
| Connectionist Bench | 1 | 3 | 4 | 1 | 1 | 6 | 2.7 | 2.1 | 2.0 | 1.0 | 4/6 |
| Ionosphere | 2 | 8 | 9 | 2 | 5 | 9 | 5.8 | 3.3 | 6.5 | 2.0 | 2/6 |
| | 5/5 | 2/5 | 1/5 | 5/5 | 3/5 | 2/5 | | | | | |

Tabela II
Ranking de performance do Método de Combinação.

## V. Conclusões

A abordagem de combinação proposta proveu resultados estáveis. Executar diversos algoritmos para combinar as soluções representa um custo computacional extra, pelo menos na primeira vez, quando o esquema de votação é criado (após isso o custo computacional é realmente baixo). Ao final, a abordagem e combinação pode ser uma escolha segura quando o pesquisador não possui informações suficientes acerca dos dados ou sobre os algoritmos de seleção de atributos.

## 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] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

[2] Q. Shen, R. Diao, and P. Su, "Feature selection ensemble," in *Turing-100. The Alan Turing Centenary*, ser. EPiC Series in Computing, A. Voronkov, Ed., vol. 10. EasyChair, 2012, pp. 289–306.

[3] R. C. Prati, "Combining feature ranking algorithms through rank aggregation," in *The 2012 IJCNN, Brisbane, Australia, June 10-15, 2012*, 2012, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/IJCNN.2012.6252467

[4] 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. [Online]. Available: http://doi.acm.org/10.1145/2623330.2623611