Grafo de Fluxo como Estrutura de Dados em Ambiente de Aprendizado de Máquina Supervisionado

Emilio Carlos Rodrigues[1,2], Maria do Carmo Nicoletti[1,3]

[1]Faculdade Campo Limpo Paulista (FACCAMP)
Campo Limpo Paulista – SP, Brasil

[2]Instituto Federal de São Paulo (IFSP) *campus* Bragança Paulista
Bragança Paulista – SP, Brasil

[3]DC – UFSCar. Universidade Federal de São Carlos.
São Carlos – SP, Brasil

emiliorc1986@gmail.com, carmo@cc.faccamp.br

Abstract. *This paper describes an ongoing research work that focuses on the investigation of flow graphs (FGs), which were proposed for representing a set of supervised training data, aiming at its analysis as well as use for extracting decision rules. The main goal of the work is to extend the original FGs proposal by means of discretization algorithms, so to make possible to induce GFs based on continuous data.*

Resumo. *Este artigo descreve um trabalho de pesquisa em andamento que tem por foco a investigação de grafos de fluxo (GFs), propostos com o objetivo de representar um conjunto de treinamento, com vistas à sua análise e uso para a extração de regras de decisão. O principal objetivo do trabalho é o de estender a proposta original por meio do uso de algoritmos de discretização, para viabilizar o uso de GFs em ambientes com dados contínuos.*

1. Introdução

O conceito de Grafo de Fluxo (GF) foi inicialmente proposto em 2003 por Z. Pawlak [Pawlak 2003], como parte de um formalismo matemático para a representação e exploração de características associadas a um conjunto $X = \{x_1, x_2, ..., x_N\}$ de N instâncias de dados. Tal conjunto, de acordo com a concepção de Pawlak, é tipicamente caracterizado como *conjunto de treinamento* em Aprendizado Supervisionado de Máquina (ASM) (ver [Mitchell 1997]), em que cada uma das instâncias $x_i \in X$ ($1 \leq i \leq N$) é descrita por valores associados a um conjunto fixo de atributos $\{a_1, a_2, ..., a_M\}$ e, também, por uma classe, dentre K classes $\{c_1, c_2, ..., c_K\}$. O formalismo baseado em GFs para a representação de um conjunto de treinamento X tem como principais objetivos facilitar e promover a análise da distribuição do fluxo dos valores de atributos das instâncias de dados representadas no GF, com relação às classes existentes, bem como, subsidiar a extração de regras de decisão, que podem ser usadas para a classificação de novas instâncias de dados que não possuem classe associada.

Algumas representações de dados (*e.g.,* aquelas fundamentadas na Regra de Bayes) têm uma interpretação probabilística, pois seus resultados estão fortemente ligados à probabilidade. Os resultados das análises realizadas em um GF, por sua vez, têm uma interpretação determinística, uma vez que refletem efetivamente o fluxo dos

dados e não apenas uma probabilidade sobre eles. Evidentemente, como qualquer abordagem indutiva para a representação de dados/conhecimento, os fatos representados pelas instâncias de dados do conjunto de treinamento terão profundo impacto no desempenho futuro de um GF, quando utilizado como uma estrutura que subsidia regras de decisão, para a classificação de novas instâncias de dados com classe desconhecida.

Por meio do levantamento bibliográfico conduzido até o momento ficou evidente que o formalismo que subsidia a indução de um GF e, na sequência, a extração de regras de decisão do GF induzido, não se popularizou tanto quanto outros formalismos de ASM, tais como Árvores de Decisão [Quinlan 1993], Redes Neurais [Bishop 2005], etc. Talvez uma das principais razões para GFs não serem tão populares se deve ao fato que os inúmeros trabalhos envolvendo GFs abordam apenas conjuntos de treinamento com valores discretos, o que, de certa forma, torna seu uso confinado a dados artificialmente gerados.

A pesquisa sendo desenvolvida reportada neste artigo está ainda em sua fase inicial. Pretende-se, ao longo da investigação sendo conduzida, investir na proposta de uma extensão de GFs, por meio da adequação dessa estrutura a domínios de dados contínuos, com vistas a ampliar suas possibilidades de uso como uma ferramenta para o aprendizado de classificadores, por meio do uso de algoritmos de discretização [García *et al.* 2013]. As próximas seções desse artigo contemplam a apresentação de conceitos básicos relacionados ao formalismo que subsidia GFs (Seção 2), a descrição detalhada da construção e uso de um GF em um conjunto de treinamento de tamanho reduzido (Seção 3) e um conjunto de atividades planejadas para a continuação e finalização dessa pesquisa, além de algumas conclusões parciais sobre o trabalho realizado até o momento (Seção 4). O texto termina com a apresentação das referências bibliográficas utilizadas.

2. Grafos de Fluxos (GFs) – Conceituação e Principais Características

A definição de Grafos de Fluxo (GFs) envolve a definição de grafos direcionados (ou dígrafos); é importante, pois, apresentar ambos os conceitos, grafos e dígrafos, antes da definição formal de GFs.

Como definido em [Nicoletti & Hruschka 2018], um grafo G = (V, E) consiste de dois conjuntos finitos em que (1) $V \neq \emptyset$ é o *conjunto de vértices* (ou *nós*) do grafo e (2) E é o *conjunto de arestas* do grafo. Cada aresta $a \in E$ representa um *par não ordenado* de nós de V, notado por (u,v), em que $u, v \in V$. O par não ordenado que representa a aresta *a i.e.*, *a* = (u,v), indica que a partir do nó u chega-se ao nó v por meio de *a* e, também, que a partir de v chega-se a u por meio de *a*. É importante notar que se o conjunto $E = \emptyset$, o grafo em questão é chamado de *grafo nulo*. Particularmente, se ao invés de pares não ordenados (u,v), arestas forem definidas por *pares ordenados* de nós, notados por <u,v>, elas são chamadas *arestas direcionadas* (ou arcos). Uma aresta direcionada a = <u,v> indica que *a* tem origem no nó u e destino no nó v, mas não vice-versa. Quando da modelagem de dados utilizando grafos, muitas vezes é necessário atribuir valores tanto a nós quanto a arestas (arcos). Particularmente quando são atribuídos valores a arestas (arcos), tais grafos (dígrafos) são também referenciados como grafos (dígrafos) ponderados.

Como definido em [Pawlak 2003, 2004a], um grafo de fluxo G é denotado por G = (N, β, φ), em que:

(1) $N \neq \varnothing$ é um conjunto finito de nós,

(2) $\beta \subseteq N \times N$ é um conjunto finito de arcos que interligam nós $x \in N$, e

(3) $\varphi: \beta \rightarrow R^+$ é uma função que associa a cada arco de G, um número real positivo.

Em um grafo de fluxo, um arco é representado pelo par ordenado <x,y>, em que x é o nó de origem e y é o nó de destino do arco, e ambos, x, y $\in$ N. Na terminologia associada a GFs, é dito também que, dado um arco <x,y>, o nó x é uma *entrada* do nó y e o nó y é uma *saída* do nó x.

Dado um conjunto de instâncias de dados X (conjunto de treinamento), como definido logo ao início da seção Introdução *i.e.*, um conjunto com N instâncias de dados, cada uma delas descrita por M atributos, A_1, A_2, ..., A_M, e uma classe associada (de um conjunto com k classes), para a construção de um GF que representa X, deve ser considerado que:

(1) a cada atributo que descreve as instâncias de dados de um conjunto de treinamento é associado, na representação do GF, um conjunto de nós distintos, cada um deles representando um possível valor que o atributo em questão assume em X.

Um conjunto de treinamento descrito por M atributos e uma classe associada será representado por um GF contendo M + 1 camadas. As M camadas, C_1, C_2, ..., C_M, estão associadas aos atributos A_1, A_2, ..., A_M, respectivamente. A camada associada à classe é notada por C_{M+1}. Cada uma das M camadas terá tantos nós quantos forem os diferentes valores do respectivo atributo que a camada representa. Também, a camada C_{M+1} terá tantos nós quanto for o número de classes representadas em X (no caso, k nós).

(2) considere que cada atributo A_i ($1 \leq i \leq M$) tenha como conjunto de valores associados $\{A_{i_1}, A_{i_2}, ..., A_{i_{j_i}}\}$. A camada C_i, que representa A_i ($1 \leq i \leq M$), terá pois $|\{A_{i_1}, A_{i_2}, ..., A_{i_{j_i}}\}|$ nós a ela associados. Então, para cada um dos valores de A_i representado no GF como um nó associado à camada C_i, é contabilizado o número de instâncias de X em que A_i assume tal valor. Tal contabilização é chamada de *fluxo do nó* (denotado por φ(nó)) em questão.

(3) em um GF, os nós de uma determinada camada C_i ($1 \leq i \leq M$) podem se relacionar com um ou mais nós da camada C_{i+1}, na dependência dos valores que os respectivos atributos que definiram ambas as camadas têm, nas instâncias de dados do conjunto de treinamento. Essa relação é representada por um arco com origem o nó associado à camada C_i e com destino ao nó associado à camada C_{i+1}.

(4) para cada arco construído em (3) é contabilizado o número de ocorrências da relação entre os dois valores de atributos que definem o arco no conjunto de treinamento. Tal número de ocorrências é chamado *fluxo do arco* (denotado por φ(arco)) e é associado ao arco como um rótulo de ponderação.

Uma vez construídos o conjunto de nós e o conjunto de arcos, atribuídos rótulos aos nós (fluxo do nó) e atribuídos pesos aos arcos (fluxo do arco), a construção da versão simplificada do GF está finalizada. Por fim, define-se o *fluxo do GF* G, notado por φ(G), como o número de instâncias (N) sumarizadas por G.

3. Exemplo de Criação de um Grafo de Fluxo a partir de um Conjunto de Instâncias de Dados

O exemplo apresentado e discutido nesta seção considera um conjunto de treinamento contendo 15 instâncias de dados, notado por $X = \{x_1, ..., x_{15}\}$, em que cada uma delas é descrita por quatro atributos, At_1, At_2, At_3 e At_4, sendo o valor associado ao último (At_4), a classe à qual a instância em questão pertence. A Tabela 1 mostra a descrição das 15 instâncias. Como pode ser visto na Tabela 1, os possíveis valores associados aos atributos At_1 e At_2 são booleanos *i.e.*, estão no conjunto {0,1}. Já o terceiro atributo é também um atributo discreto, com cinco possíveis valores, {C,D,F,S,V}. Os três possíveis valores do atributo At_4 estão no conjunto {1,2,3}. A Figura 1 apresenta o grafo de fluxo G obtido a partir do conjunto de treinamento mostrado na Tabela 1.

Instância	At_1	At_2	At_3	At_4
x_1	0	1	C	3
x_2	0	1	S	2
x_3	1	0	V	1
x_4	0	0	D	2
x_5	0	1	D	2
x_6	0	1	C	2
x_7	1	0	F	1
x_8	0	0	C	2
x_9	0	1	S	2
x_{10}	1	0	D	1
x_{11}	0	0	S	2
x_{12}	0	1	V	2
x_{13}	0	1	S	3
x_{14}	0	1	V	2
x_{15}	1	0	S	1

Tabela 1. Conjunto de treinamento com 15 instâncias {x_1, x_2, …, x_{15}}, cada uma delas descrita por três atributos (At_1, At_2 e At_3) e uma classe associada (At_4).

O GF da Figura 1 foi criado a partir do conjunto de treinamento descrito na Tabela 1. Como cada atributo do conjunto de treinamento define uma camada do GF, o GF sendo construído tem quatro camadas. Considerando o conjunto de treinamento, pode ser visto que nele o atributo At_1 comparece com apenas dois valores associados, {0,1}. Portanto, associados à camada At_1, são criados dois nós no GF, o nó 0 e o nó 1, como podem ser vistos sob o rótulo At_1, na Figura 1. Esse processo é repetido para todos os demais atributos do conjunto de treinamento, *i.e.*, At_2, At_3 e At_4; ao final do processo, estão definidos os nós do GF. Pode ser observado no conjunto de treinamento que existem três instâncias em que At_1=0 e At_2=0. Essa relação entre os valores dos atributos At_1 (=0) e At_2 (=0) define, no GF, um arco <0,0> com fluxo igual à 3 (*i.e.* φ<0,0>=3).

De forma análoga, há quatro instâncias no conjunto de treinamento em que At_1=1 e At_2=0 e, portanto, é definido um arco <1,0> com fluxo igual à 4 (*i.e.* φ<1,0>=4). Esse procedimento é repetido para todos os pares de nós associados a camadas adjacentes (primeiro nó do par na camada J e segundo nó do par na camada J+1) no GF. Com a finalização da construção do GF, pode ser observado que a estrutura construída, de certa forma, condensa o conjunto de treinamento fornecido e, assim, pode ser utilizada para análise dos dados, bem como para a extração de regras de decisão, entre outros.

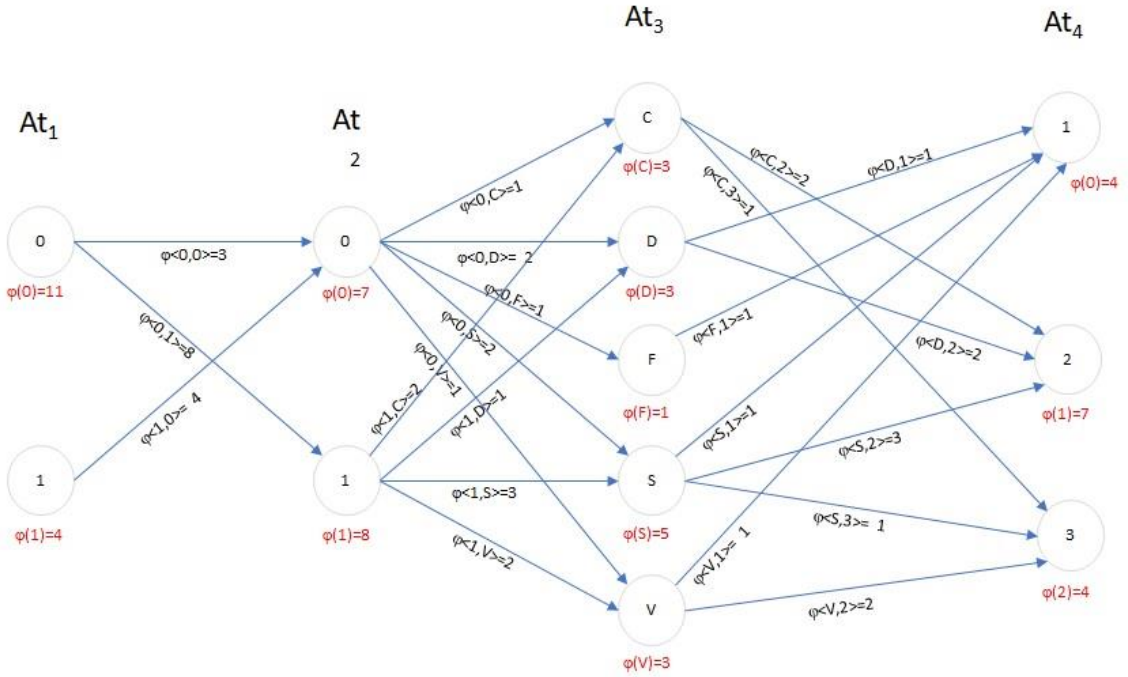

Figura 1. Grafo de fluxo induzido a partir das instâncias de dados da Tabela 1.

4. Atividades, Continuação do Trabalho e Conclusões Parciais

Durante o trabalho de pesquisa em andamento já foi realizado um levantamento bibliográfico inicial em que foram identificados inúmeros trabalhos com foco em GFs que, de certa forma, evidenciam um certo interesse e investimento em pesquisa na área. Surpreendentemente, entretanto, foi um número aquém do que se esperaria de um formalismo que se propõe a fornecer uma ferramenta matemática de análise de dados com base na distribuição de fluxo dos mesmos, de forma prática e com viés determinístico, auxiliando assim, na criação de algoritmos de classificação de instâncias, como anunciam algumas referências [Pawlak 2003, 2003a, 2004, 2004a]. Vários desses trabalhos, de autoria de Pawlak *i.e.*, o mesmo pesquisador que propôs e investiu na divulgação de seu formalismo, são refinamentos dos conceitos, definições e resultados inicialmente publicados, bem como refinamentos da notação empregada em suas respectivas descrições anteriores. Com base no levantamento bibliográfico inicial já foi visto que o formalismo de GF se estende além do apresentado nesse artigo, contemplando grafos de fluxo normalizados, grafos de fluxo inversos, estabelecimento de fatores de certeza, cobertura, força, extração de regras de decisão, análise de dependência entre nós, e alguns outros.

A pesquisa sendo realizada prevê a continuidade do acompanhamento de trabalhos publicados na área específica de GF, que envolvam tanto extensões do formalismo quanto o seu uso na modelagem e resolução de problemas, com foco em classificadores. Também está planejada uma extensão de GFs de maneira a adequar o formalismo para que possa tratar dados descritos por atributos contínuos, já que a proposta original de GF tem foco apenas em atributos discretos. Um sistema computacional que implementa a construção e uso de GFs como classificadores está sendo desenvolvido com o objetivo de disponibilizar um ambiente para a realização de avaliações de GFs como classificadores e como estruturas de sumarização de conjuntos

de treinamento. O estudo e investigação de GFs, até o momento, evidenciaram diversos problemas relacionados tanto com a falta de padronização na notação empregada em diversas referências, quanto algumas pequenas inconsistências com relação ao próprio formalismo. Tais problemas eventualmente serão abordados por meio de uma reescrita padronizada e rigorosa da conceituação envolvida, nos trabalhos a serem divulgados, relacionados à pesquisa sendo realizada.

Referências

[Bishop 2005] Bishop, C. M. Neural Networks for Pattern Recognition: Oxford University Press.

[García *et al.* 2013] García, S.; Luengo, J.; Sáez, J. A.; López, V.; Herrera, F. (2013) A survey of discretization techniques: taxonomy and empirical analysis in supervised learning, *IEEE Trans. on Knowledge and Data Eng.*, v. 25, no. 4, pp. 734–750.

[Mitchell 1997] Mitchell, T. M (1997) *Machine Learning*, USA: McGraw-Hill.

[Nicoletti & Hruschka Jr. 2018] Nicoletti, M. C.; Hruschka Jr., E. R. (2018) Fundamentos da Teoria dos Grafos para Computação. Editora GEN-LTC, 3a. Edição, 224 pgs.

[Pawlak 2003] Pawlak, Z. (2003) Flow graphs and decision algorithms, G. Wang *et al.* (Eds.) *Lecture Notes in Artificial Intelligence*, Berlin: Springer-Verlag, pp.1–10.

[Pawlak 2003a] Z. Pawlak (2003) Decision algorithms and flow graphs; a rough set approach, *Journal of Telecommunications and Information Technology* 3, pp. 98–101.

[Pawlak 2004] Pawlak, Z. (2004) Decision rules and flow networks, *European Journal of Operational Research*, v. 152, pp. 184–190.

[Pawlak 2004a] Z. Pawlak (2004) Flow graphs – a new paradigm for data mining and knowledge discovery, JAIST Forum 2004 – Technology Creation Based on Knowledge Science: Theory and Practice, jointly with The 5th International Symposium on Knowledge and Systems Science (Proc. of the KSS2004), pp. 147–153.

[Quinlan 1993] Quinlan., J. R. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.